Technische Universität München
TUM School of Social Sciences and Technology

Natural Language Processing for Violence Studies: Investigating Trauma and Online Aggression

Miriam Katharina Doris Schirmer

Complete reprint of the dissertation approved by the TUM School of Social Sciences and Technology of the Technical University of Munich for the award of the

Doktorin der Philosophie (Dr. phil.).

Chair:
Prof. Dr. Janina Steinert

Examiners:

1. Prof. Dr. Jürgen Pfeffer
2. Research Assistant Prof. Fred Morstatter, Ph.D.
3. Prof. Dr. Sarah Diefenbach

The dissertation was submitted to the Technical University of Munich on 2 September 2024 and accepted by the TUM School of Social Sciences and Technology on 16 October 2024.

Abstract

This dissertation explores the application of Natural Language Processing (NLP) to analyze and understand different forms and stages of violence through diverse perspectives. It makes three significant contributions: (1) applying advanced language models to uncover the nuanced representations of violence in text on a methodological level, (2) integrating computational NLP methods with theoretical frameworks from the social sciences, and (3) providing actionable recommendations for societal engagement and real-world interventions. This dissertation explores the opportunities and challenges associated with using NLP for violence studies, emphasizing its potential to detect, understand, and address various forms of violence in different online contexts.

The dissertation is structured around eight key studies: Study 1 integrates NLP with qualitative analyses to study trauma in genocide contexts; studies 2 and 3 focus on creating datasets and models for detecting violence and trauma in tribunal testimonies; study 4 discusses making these models accessible for public use; study 5 extends the analysis of trauma to online mental health forums. Studies 6 and 7 shift the focus to the active use of violent language, specifically examining how such language manifests in online forums and analyzing abusive behaviors on platforms like TikTok. Study 8 explores the synergy between human intelligence and AI in researching hate speech on social media, demonstrating the potential of combining thematic analysis with large language models to understand and address hate speech effectively.

Zusammenfassung

Diese Dissertation untersucht die Anwendung von Natural Language Processing (NLP) zur Analyse und zum Verständnis verschiedener Formen und Stadien von Gewalt aus unterschiedlichen Perspektiven. Sie leistet drei wesentliche Beiträge: (1) die Anwendung von Sprachmodellen unterschiedlicher Komplexität, um Gewalt in Texten auf methodischer Ebene nuanciert darzustellen, (2) die Integration von computergestützten NLP-Methoden mit sozialwissenschaftlichen Theorien und Fragestellungen und (3) Anreize zur Verknüpfung der Ergebnisse mit Interventionen in der Praxis für gesellschaftliches Engagement und Interventionen zu Gewaltprävention. Diese Dissertation zeigt Möglichkeiten und Herausforderungen bei der Nutzung von NLP für Gewaltforschung auf und betont ihr Potenzial, verschiedene Formen von Gewalt in diversen Online-Kontexten zu erkennen, zu verstehen und präventiv dagegen vorzugehen.

Die Dissertation ist um acht zentrale Studien strukturiert: Studie 1 integriert NLP mit qualitativen Analysen zur Untersuchung von Traumata im Kontext von Genoziden; Studien 2 und 3 konzentrieren sich auf die Erstellung von Datensätzen und Modellen zur Erkennung von Gewalt und Traumata in Zeug:innenaussagen vor Gericht; Studie 4 diskutiert, wie entsprechende Modelle für die Öffentlichkeit zugänglich gemacht werden können; Studie 5 erweitert die Analyse von Traumata auf Online-Foren für psychische Gesundheit. Studien 6 und 7 verlagern den Fokus auf den aktiven Gebrauch gewalttätiger Sprache, indem sie untersuchen, wie sich solche Sprache in Online-Foren, insbesondere innerhalb von Online-Communities, manifestiert und analysieren übergriffiges Verhalten auf Plattformen wie TikTok. Studie 8 untersucht die Synergie zwischen menschlicher Intelligenz und Large Language Models (LLMs) bei der Analyse von Hatespeech in sozialen Medien und zeigt, wie qualitative thematische Analysen LLM-unterstützt durchgeführt werden können, um Hatespeech effektiv zu verstehen und zu bekämpfen.

Publications

This thesis includes seven first-authored publications, four of them peer-reviewed papers relevant to the examination, as well as three additional papers and an additional second-author paper.

Publications Relevant to the Examination		
Title	**Authors**	**Venue**
Talking About Torture: A Novel Approach to the Mixed Methods Analysis of Genocide-Related Witness Statements in the Khmer Rouge Tribunal	Schirmer, Miriam; Pfeffer, Jürgen; Hilbert, Sven	Journal of Mixed Methods Research (`https://doi.org/10.1177/15586898231218463`). Published 11-2023.
A New Dataset for Topic-Based Paragraph Classification in Genocide-Related Court Transcripts	Schirmer, Miriam; Kruschwitz, Udo; Donabauer, Gregor	Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022) (`http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.479.pdf`). Published 06-2022.
Uncovering Trauma in Genocide Tribunals: An NLP Approach Using the Genocide Transcript Corpus	Schirmer, Miriam; Olguín Nolasco, Isaac Misael; Mosca, Edoardo; Xu, Shanshan; Pfeffer, Jürgen	Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023) (`https://doi.org/10.1145/3594536.3595147`). Published 06-2023.
Investigating the Increase of Violent Speech in Incel Communities with Human-Guided GPT-4 Prompt Iteration	Matter, Daniel*; Schirmer, Miriam*; Grinber, Nir; Pfeffer, Jürgen	Frontiers in Social Psychology – Section Computational Social Psychology (`https://doi.org/10.3389/frsps.2024.1383152`). Published 07-2024.
Additional Publications in the Dissertation		
GENTRAC: A Tool for Tracing Trauma in Genocide and Mass Atrocity Court Transcripts	Schirmer, Miriam; Brechenmacher, Christian; Jashari, Endrit; Pfeffer, Jürgen	Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (`https://aclanthology.org/2024.lrec-main.677.pdf`). Published 05-2024.
The Language of Trauma: Modeling Traumatic Event Descriptions Across Domains with Explainable AI	Schirmer, Miriam; Leemann, Tobias; Kasneci, Gjergji; Pfeffer, Jürgen; Jurgens, David	Arxiv Preprint (`https://doi.org/10.48550/arXiv.2408.05977`) Published 08-2024.
More Skin, More Likes! Measuring Child Exposure and User Engagement on TikTok	Schirmer, Miriam; Voggenreiter, Angelina; Pfeffer, Jürgen	Arxiv Preprint (`https://doi.org/10.48550/arXiv.2408.05622`). Published 08-2024
Large Language Models and Thematic Analysis: Human-AI Synergy in Researching Hate Speech on Social Media	Breazu, Petre; Schirmer, Miriam; Hu, Songbo; Katsos, Napoleon	Arxiv Preprint (`https://doi.org/10.48550/arXiv.2408.05126`). Published 08-2024.

Table 1: Publications included in this dissertation. Asterisks mark equal author contributions.

Contents

1 Introduction

To effectively address and prevent violence, it is necessary to understand its various facets and impacts from multiple perspectives. This dissertation explores the use of Natural Language Processing (NLP) methods to analyze and comprehend different forms and stages of violence, incorporating diverse viewpoints. Violence, as defined in this dissertation, is the "intentional use of force or power, against individuals or groups, resulting in physical or psychological harm." (Krug et al., 2002) It includes physical, verbal, and psychological forms and can occur in various contexts, including digital environments. How these different aspects of violence can be examined through the lens of state-of-the-art language models is one of the main contributions of this dissertation. Further, it bridges the gap between computational NLP techniques and theory-driven approaches from the social sciences by providing a comprehensive framework for analysis. Finally, the dissertation outlines how these methods can be used to engage with society, disseminate findings to a broader audience, and offer actionable recommendations for real-world interventions to prevent harm. I apply diverse research methods throughout this dissertation, including NLP techniques, mixed-methods analysis, and expert interviews. The studies cover different forms of violence, such as physical violence in genocides, verbal abuse online, and psychological trauma, while considering the impacts on diverse demographic groups, including genocide survivors, online community members, and children.

Recent advancements in NLP have enabled significant applications across various domains in Computational Social Science (CSS) (Lazer et al., 2009; O'Connor et al., 2011), including the detection and analysis of violence-related content (Burley et al., 2020; Macanovic, 2022; Ziems et al., 2024). Analyzing perspectives from those affected by and those involved in violence, this research investigates its impact by examining how individuals discuss their experiences across various contexts, including witness statements in court, popular social media platforms, and online forums.

This kind of research has been conducted extensively for hate speech within the last decades (i.e., speech that promotes hate or violence against people based on their identity, Waseem et al. (2017), c.f., Davidson et al. (2017); Jahan and Oussalah (2023)). Other forms of violence and their nuanced impacts, particularly from the victim's perspective, remain underexplored through NLP methods. For example, the effects of physical violence in genocides and psychological trauma in abuse survivors have not been examined in as much detail. Similarly, verbal violence in online forums and social media has received less comprehensive attention regarding its long-term effects on victims. These gaps highlight the need for a broader understanding of various forms of violence and their impacts on different demographic groups. Building on these gaps, this dissertation answers the following research questions:

Question 1: How can NLP techniques effectively detect and analyze various forms of violence and trauma in different contexts, and what challenges arise?

Question 2: How can NLP effectively consider both victim and perpetrator perspectives while identifying common characteristics of violence across different domains and contexts?

Question 3: How can NLP contribute to making a real-world impact in violence prevention and research?

These research questions capture different levels of analysis in the study of violence and trauma using NLP techniques. **Question 1** addresses the **methodological level**, focusing on the technical challenges and effectiveness of NLP in various contexts. **Question 2** explores the **psychological and perspective level**, investigating how text data can represent both victim and perpetrator[1] viewpoints. **Question 3** looks at the **cross-domain impact and societal level**, identifying common characteristics of violence that NLP can detect across different domains, highlighting broader societal patterns and effects.

1.1 Violence & Trauma

The World Health Organization (WHO) defines violence as the

> intentional use of physical force or power, threatened or actual, against oneself, another person, or a group or community, resulting in or having a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation (Krug et al., 2002, p. 5).

Violence can be broadly categorized into physical, verbal, and psychological forms. Physical violence involves actions such as assault, homicide, and genocide, which cause bodily harm. Verbal violence includes abusive language, threats, and harassment commonly encountered in personal interactions and on online platforms. Psychological violence involves emotional abuse or manipulation, which can affect mental well-being (Hamby, 2017).

In recent academic discourse, the categorization of violence has evolved to encompass the nature of the acts and the contexts in which they occur. This includes distinctions between violence occurring within families, specific communities, or society at large (Ray and Ray, 2018), and extends to violence arising from broader conflicts, such as war (De Jong, 2002). This categorization helps understand the multifaceted nature of violence and its varying impacts on different groups (Cuevas

[1] The term ”perpetrator” in this context refers to individuals who engage in offensive, hateful, or violent language. The term describes their role in perpetuating negative behaviors without implying criminal intent or legal guilt.

and Rennison, 2016; Lindert and Levav, 2016). With the rise of digital technology, the focus has increasingly shifted to online environments where violence manifests in new forms, including online misogyny (Cuklanz, 2022) or the spread of extremism on online platforms (Scrivens et al., 2020). Understanding the various forms and contexts of violence is important for developing appropriate prevention and intervention strategies. As violence evolves and extends into digital spaces, our approaches must adapt to address both traditional and emerging forms of harm, including how violence can affect individuals.

Psychological harm is one of the significant effects of violence beyond physical injury. This harm can lead to trauma that affects a person's mental health over time. Trauma is an emotional response to a distressing event, and while it can result from various sources – such as accidents or natural disasters – violence is a major contributor (Van der Kolk, 2003; Breslau and Kessler, 2001). The connection between violence and trauma lies in the potential for all forms of violence to cause lasting psychological harm. Trauma encompasses the emotional and psychological aftermath experienced by individuals who have been exposed to distressing or life-threatening situations (American Psychiatric Association, 2013; Friedman and Davidson, 2007). In the study of violence, understanding trauma is essential as it offers insights into the long-term psychological effects on those affected, including victims and witnesses. Trauma can manifest in various ways, such as anxiety, depression, post-traumatic stress disorder (PTSD), and other mental health conditions (Gold, 2017; Yehuda, 1998; Van der Kolk, 2003). These outcomes highlight the importance of developing comprehensive prevention and intervention strategies that address both the physical and psychological aspects of violence.

Scope of Violence in this Dissertation. In this dissertation, I focus on those forms of violence that become visible through text in the online world, such as accounts of experienced violence that have been captured in digitized documents or verbal language that has been expressed in online forms or social media platforms. This includes the impact violence can have on individuals, particularly psychological trauma that may develop after having experienced violence.

1.2 Analyzing Violence with NLP

Measuring violence outside a controlled setting (such as in clinical contexts) poses significant challenges due to its subjective and varied nature. With the increasing availability of textual material online, automated processes, particularly NLP, are becoming vital for violence research (Botelle et al., 2022; Ebner et al., 2023; Ni et al., 2020). NLP methods, defined as "a theory-motivated range of computational techniques for the automatic analysis and representation of human language" (Cambria and White, 2014, p. 48), allow researchers to analyze violence through text data, identifying various forms and impacts, such as verbal abuse, threats, and psychological manipulation. By processing large volumes of text from social media, online forums, and historical documents, NLP can uncover insights into the prevalence, nature, and effects of violence, enhancing our understanding of its impact on individuals and communities.

NLP offers numerous advantages in violence research by processing extensive text data from diverse sources. It can identify patterns and trends in violence across large datasets that are otherwise unmanageable manually. For example, NLP tools can monitor online platforms in real-time, detecting instances of verbal violence, hate speech, and threats, which are crucial for early intervention (Batrinca and Treleaven, 2015; Gongane et al., 2022). Furthermore, by analyzing sentiment and emotional content, NLP helps understand the psychological effects of violent language, aiding in mental health and victim support studies (Calvo et al., 2017).

However, it is important to acknowledge the limitations of NLP, as its analysis is inherently constrained to textual data, primarily detecting verbal and written expressions of violence. This limitation may overlook the nuances of violence conveyed through non-textual means, subtle contexts, or intentional obfuscation by the author. In this dissertation, I focus on language-based violence as directly expressed or reported in text. I aim to highlight the subtle yet pervasive forms of violence that, though often unnoticed, can have significant psychological and social consequences and may lead to or support physical violence in real-world scenarios.

1.3 Dissertation Structure & Overview

This dissertation starts with an outline of opportunities and risks that lie at the intersection of NLP and violence studies (Section 2), followed by an overview of selected methods used in this context (Section 3). It then introduces the studies incorporated (Section 4) and concludes by discussing their contributions and potential for future work in the context of other relevant literature (Section 5). The last section includes interviews with subject matter experts to discuss how research conducted on trauma detection and mass atrocities can be used for real-world impact.

The investigation of trauma as a psychological impact within the context of violence lies at the core of this dissertation, with a primary focus on genocide and mass atrocities. This research includes eight key papers (see Figure 1 for an overview). It begins with an example of how to combine NLP techniques and qualitative analyses to maximize insights from the data (Paper 1, Section 4.1). Papers 2 (Section 4.2) and 3 (Section 4.3) focus on creating datasets for NLP analyses related to genocide, aiming to identify when witnesses at international criminal tribunals discuss potentially traumatizing experiences and to conduct benchmark experiments. Paper 4 (Section 4.4) outlines how to develop models for detecting traumatic events and make these models accessible to the public. Additionally, this dissertation features interviews with subject matter experts, including a genocide survivor, a legal activist, and a United Nations interpreter for the International Criminal Tribunal for the former Yugoslavia (ICTY) (Section 5.1.3). These interviews provide valuable firsthand accounts and professional insights into the practical applications of NLP-based trauma detection. In addition to exploring trauma in the context of genocide and mass atrocities, Paper 5 (Section 4.5) examines common linguistic features associated with traumatic events across various contexts, including online mental health forums. Beyond the study of trauma, this dissertation

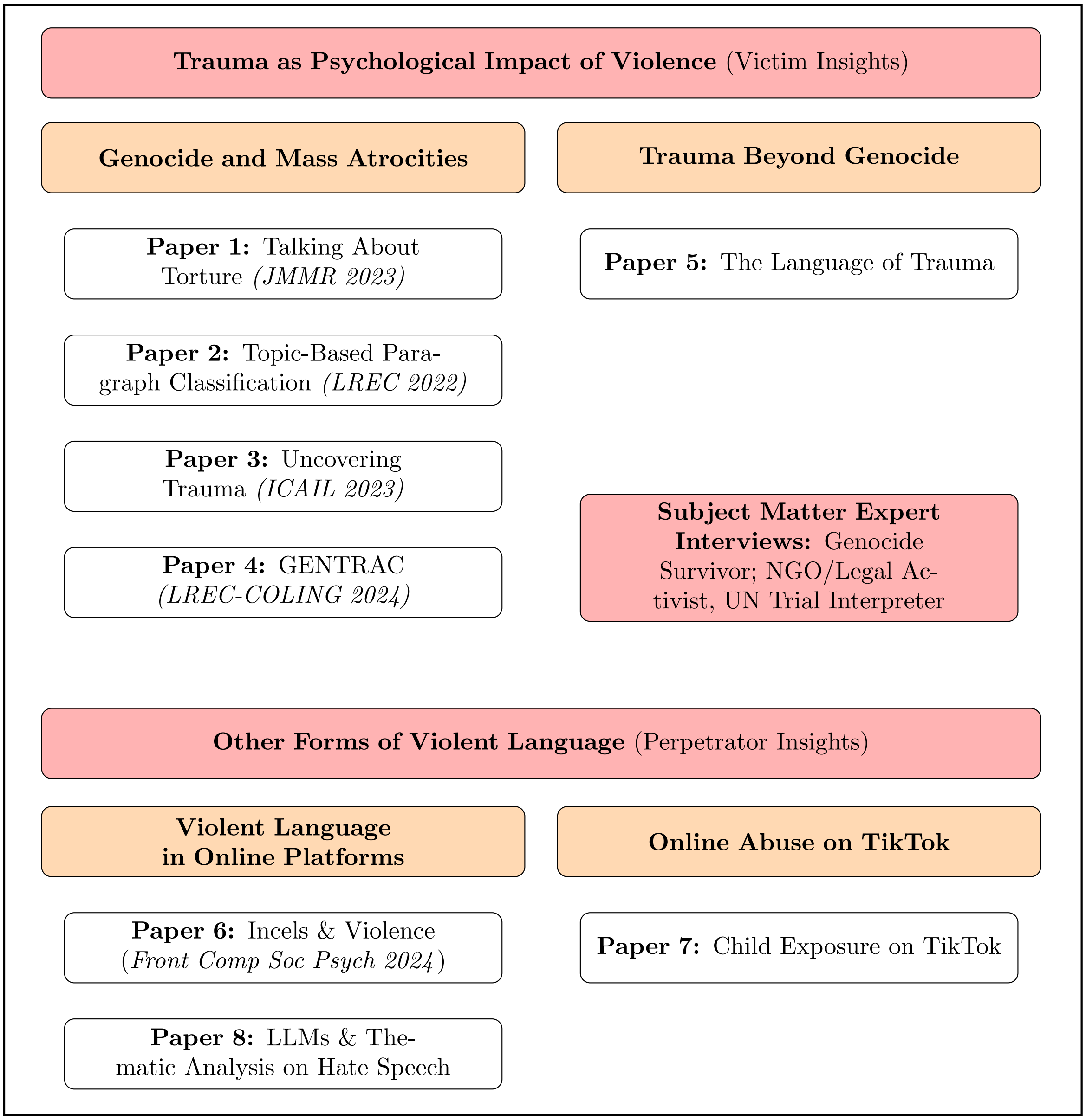

Trauma as Psychological Impact of Violence (Victim Insights)
Genocide and Mass Atrocities
Trauma Beyond Genocide
Paper 1: Talking About Torture *(JMMR 2023)*
Paper 5: The Language of Trauma
Paper 2: Topic-Based Paragraph Classification *(LREC 2022)*
Paper 3: Uncovering Trauma *(ICAIL 2023)*
Subject Matter Expert Interviews: Genocide Survivor; NGO/Legal Activist, UN Trial Interpreter
Paper 4: GENTRAC *(LREC-COLING 2024)*
Other Forms of Violent Language (Perpetrator Insights)
Violent Language in Online Platforms
Online Abuse on TikTok
Paper 6: Incels & Violence (*Front Comp Soc Psych 2024*)
Paper 7: Child Exposure on TikTok
Paper 8: LLMs & Thematic Analysis on Hate Speech

Figure 1: Paper Overview.

also investigates other forms of violent language. It explores violent and misogynistic language in online forums, mainly focusing on the discourse of Involuntary Celibates (Incels) (Paper 6, Section 4.6). Furthermore, it analyzes abusive language and child-endangering parental behaviors on social media platforms like TikTok, highlighting the broader implications of harmful language in digital spaces (Paper 7, Section 4.7). Finally, Paper 8 (Section 4.8) examines how humans and AI work together to categorize and annotate a YouTube dataset of hate speech comments using LLM support.

The structure of this dissertation provides a comprehensive examination of various facets of violence, analyzing language from both victim and perpetrator perspectives. While Paper 1 includes both perspectives, Papers 2 through 5 are centered on the experiences of victims, focusing on trauma detection in the context of genocide and beyond, including the creation of datasets and models for identifying traumatic language. These papers aim to capture the voices of those affected by such events. In contrast, Papers 6 through 8 shift focus to perpetrator perspectives, examining violent and misogynistic language in online forums and exploring abusive language and harmful behaviors on social media platforms like TikTok, thus addressing the linguistic expressions of violence from those who perpetrate it.

1.4 Contributions

This dissertation explores the application of NLP to understand violence in text from multiple angles, including its detection, classification, and thematic analysis. It covers a range of NLP approaches, from simple classification tasks to the use of Large Language Models (LLMs), and offers a framework for researchers to understand the capabilities and limitations of NLP in this context. This dissertation comprehensively analyzes violence detection across different contexts, utilizing diverse research methods, including NLP techniques, mixed-methods analysis, and expert interviews. It examines both victims and perpetrators and analyzes text sources from court transcripts, social media forums, and platforms like TikTok, offering a nuanced understanding of how language reflects and perpetuates violence.

From a practical perspective, the use of NLP in analyzing violence provides valuable contributions by enabling large-scale detection and real-time monitoring of verbal abuse, hate speech, and threats across various platforms. This capability is important for early intervention, helping to prevent the escalation of violence and providing timely support to those affected. Additionally, NLP's ability to identify patterns and trends in violent language enhances understanding of the factors and contexts that contribute to violence, informing policy-making and educational initiatives. By evaluating the emotional and psychological impacts of violent language, NLP also supports mental health efforts, making it a useful tool for addressing and mitigating the effects of violence in modern society.

2 The Need for Nuanced Language Modeling in Violence Research

Detecting violence in text is a challenging task. First, the definition of violence can be ambiguous and broad, making it difficult to determine what constitutes violent content. This includes the challenge of deciding whether to include implications of violence or focus solely on explicit acts. Furthermore, there are various categories of violence, such as physical, verbal, and psychological, which complicates the classification process (Saltzman, 2004). While sociological and criminological research often distinguishes between different forms of violence–e.g., physical vs. psychological violence–NLP-based studies have not always emphasized these distinctions. This is partly due to the predominant focus on social media text data within the field, with less attention given to other forms of text data, such as police reports, news articles, or clinical patient records.

To outline the potential of NLP for violence research, this section first maps out the development of text-based computational violence research, highlighting hate speech and violence in mental health contexts as central areas of focus (Figure 2). It provides the foundation for answering Research Questions 1 and 2 by exploring how NLP techniques have been applied to detect and analyze various forms of violence and addressing the complexities of representing both victim and aggressor perspectives in text data. The section concludes with an overview of the opportunities and challenges in NLP-based violence research.

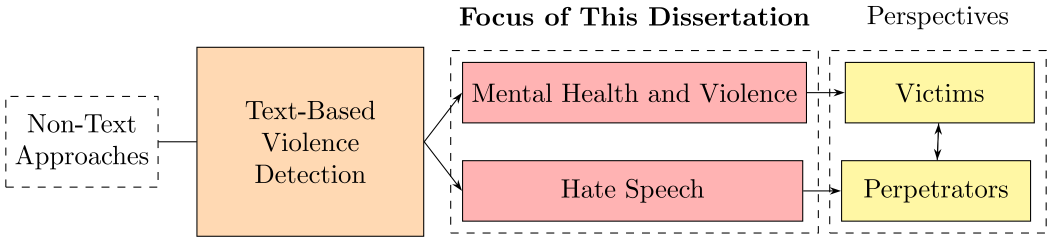

Focus of This Dissertation
Perspectives
Non-Text Approaches
Text-Based Violence Detection
Mental Health and Violence
Hate Speech
Victims
Perpetrators

Figure 2: llustration of the Dissertation's Thematic Focus in Computational Violence Research.

2.1 The Evolution of Computational Violence Research with Text

Understanding the evolution of NLP in violence research, along with key areas and the factors that shape this field, is essential to exploring its opportunities. Computational violence research has seen significant advancements in recent years. Key developments include modeling and predicting events such as war, mass violence, and civic unrest (Chen et al., 2023; Verdeja, 2016), as well as forecasting violent behavior in clinical settings (Parmigiani et al., 2022; Steinert, 2002). These studies often rely on a complex interplay of factors, including individual demographic details and broader political contexts, extending beyond textual data.

2.1.1 The Increase of Text Data as a Catalyst

As computational violence research has evolved, there has been an increasing focus on the use of text data to better understand and predict violent behaviors. Text data offers a rich source of information, capturing nuances and context that other data types might miss (Liu et al., 2018). This shift towards text-based analysis is partly driven by the growing availability of large-scale text datasets from various sources, including social media (Ruths and Pfeffer, 2014), news articles (Zamith and Lewis, 2015), and institutional records (Guiliano and Ridge, 2016). The ability to analyze this text data with NLP techniques has opened new avenues for detecting and understanding violence, allowing researchers to capture subtler forms of aggression and threats that may not be evident through traditional methods.

Social Media. One major influencing factor has been the rise of social media. In recent years, social media platforms have become ubiquitous channels for communication, expression, and information sharing. Users express feelings, opinions, and emotions through public or private conversations, resulting in enormous textual content. These platforms host diverse text data, including posts, comments, tweets, and messages, which researchers have been using to develop NLP models for detecting violent language, hate speech, and threats (Schmidt and Wiegand, 2017; Mishra et al., 2019). By analyzing user-generated content, these models learn patterns associated with aggressive behavior, enabling them to automatically detect violent expressions (Fortuna and Nunes, 2018). The availability of large-scale, labeled datasets from social media platforms has facilitated the training of robust NLP models, enhancing our ability to detect violence in online discourse (Zampieri et al., 2019).

Digitization. Beyond social media, digitization efforts have transformed traditional documents into machine-readable formats. Historical records, legal documents, news articles, and scholarly publications are now available in digital repositories. These digitized sources serve as training data for NLP models focused on violence detection. For instance, legal documents related to domestic violence cases provide insights into patterns of abusive language, victim narratives, and contextual cues (Hachey and Grover, 2006). By applying NLP techniques, researchers can extract relevant features, such as sentiment, aggression, and intent, from legal texts (Zadgaonkar and Agrawal, 2021). Additionally, clinical patient records offer a rich source of data for identifying instances of violence or abuse, facilitating the development of models to detect and analyze violent behavior in healthcare settings (Campbell, 2002; Steinert, 2002). Furthermore, digitized news archives enable the analysis of media coverage of violent incidents, allowing the development of models that recognize violence-related terms and events (de Gibert et al., 2018a).

2.1.2 Hate Speech & Mental Health

NLP research on violence detection spans fields like hate speech and mental health. In hate speech detection, researchers develop algorithms to identify and filter harmful language on social media, tackling the challenges of varying definitions and contexts (MacAvaney et al., 2019; Gongane et al., 2022). Another significant research area focuses on the applications of NLP in mental health, specifically analyzing the impacts of violence in clinical settings or online forums. This research aims to identify signs of trauma, distress, or violent tendencies in textual communications, thereby facilitating early intervention and support for affected individuals (Poletto et al., 2021). Both fields use NLP to improve safety and well-being, even though they have different focuses and applications.

Hate-Speech-Detection. Researchers have increasingly focused on hate speech due to its rise on social media, societal impacts, and advancements in detection technology. The spread of hate speech on these platforms has made it a key area of study (Tontodimamma et al., 2020). Studies have shown that exposure to hate speech can lead to desensitization and increased prejudice (Soral et al., 2018), while technological advances have enabled the development of automated detection systems (Auti et al., 2022). Additionally, the debate on balancing freedom of expression with protecting individuals from hate speech continues to drive research in this area (Siegel, 2020). Effective countermeasures, such as education and promoting positive counter-narratives, have also been a focal point for researchers (Lopez-Sanchez and Müller, 2021).

Effectively addressing harmful language online necessitates a nuanced understanding of its diverse manifestations, including ”abusive language,” ”hate speech,” and ”toxic language” (Nobata et al., 2016; Schmidt and Wiegand, 2017). The overlapping characteristics and varying degrees of subtlety and intensity in these types of content present a significant challenge in distinguishing among them. Davidson et al. (2017) define hate speech as

> [..] language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. In extreme cases, this may also be language that threatens or incites violence (p. 521).

This definition is extended within the research community to include direct attacks against individuals or groups based on race, ethnicity, or sex, often manifesting as offensive and toxic language (Salminen et al., 2020). Hate speech, as a broad category of harmful online language, includes a wide range of hateful behaviors. Research often concentrates on specific areas like toxic language, leading to a fragmented landscape with varied definitions (Caselli et al., 2020; Kansok-Dusche et al., 2023; Nghiem et al., 2024; Waseem et al., 2017). These definitions converge on verbal violence as a fundamental characteristic of harmful language. Researchers have come up with different taxonomies for capturing hate speech with NLP methods, including the target (Nghiem and Morstatter, 2021; Waseem et al., 2017) or the level of aggression (Nghiem and Morstatter, 2021).

Research on content moderation for hate speech detection focuses on developing NLP algorithms to identify and filter harmful content. One challenge is the subjective nature of hate speech, which varies across cultural and linguistic contexts, complicating the creation of universally accepted definitions and datasets (MacAvaney et al., 2019; Gongane et al., 2022). Advanced models, including transformer-based ones, have improved detection accuracy but struggle with nuanced language and multimedia content like emojis and GIFs (Gongane et al., 2022). Explainable AI is emphasized to ensure transparency in automated moderation systems. Benchmark corpora and annotated datasets are crucial for training these models, but ongoing research is needed to address biases and handle code-mixed languages (Poletto et al., 2021).

Violence & Mental Health. Major areas in this field include promoting better health and early disorder identification for intervention (Calvo et al., 2017; Swaminathan et al., 2023). For example, Levis et al. (2021) associated linguistic markers from psychotherapist notes with treatment duration. Analyzing mental health chat conversations, Hornstein et al. (2024) found that words indicating younger age and female gender were associated with a higher chance of re-contacting. More generally, Althoff et al. (2016) developed a framework for text-message-based counseling to correlate various linguistic aspects with conversation outcomes. Recently, the use of Large Language Models (LLMs)[2] has led to the development of specific models for mental health applications (Xu et al., 2024; Yang et al., 2024). While LLMs effectively detect mental health issues and provide eHealth services, their clinical use poses risks, such as the lack of expert-annotated multilingual datasets, interpretability challenges, and issues regarding data privacy and over-reliance (Guo et al., 2024).

For social media data, there has been research on using sentiment analysis and semantic structures to detect anxiety (Low et al., 2020) or depression (Tejaswini et al., 2024) on Reddit posts. In suicide prevention on social media, Sawhney et al. (2020) developed a superior model for suicidal risk screening that identifies emotional and temporal cues, outperforming competitive methods (c.f., Ji (2022) on suicidal risk detection).

Specifically in trauma research, progress is being made in analyzing patient narratives (He et al., 2017) and identifying cases of post-traumatic stress disorder (PTSD) through speech (Marmar et al., 2019). Miranda et al. (2024) developed an NLP workflow using a pre-trained transformer-based model to analyze clinical notes of PTSD patients, revealing consistent reductions in trauma criteria post-psychotherapy. Disruptions in lexical characteristics and emotional valence have been found to contribute to identifying PTSD (Quillivic et al., 2024). Using Twitter data, Ul Alam and Kapadia (2020) investigated whether posts can complete clinical PTSD assessments, achieving promising accuracy in PTSD classification and intensity estimation validated with veteran Twitter users (c.f., Coppersmith et al. (2014); Reece et al. (2017)).

[2] The definition of what constitutes an LLM often varies; for this dissertation, the term refers specifically to generative language models, such as GPT-4.

Identifying Research Gaps in NLP for Hate Speech Detection and Mental Health Applications. Despite significant advancements, several research gaps persist in both areas. In hate speech detection, the subjective nature of what constitutes hate speech presents a major challenge. Definitions and perceptions of hate speech vary widely across cultural and linguistic contexts, complicating the development of universally accepted definitions and datasets (Soni et al., 2024; MacAvaney et al., 2019). Advanced models, particularly transformer-based ones, have improved detection accuracy but still struggle with nuanced language and multimedia content such as emojis and GIFs (Hermida and Santos, 2023). This underscores the need for further research to enhance the detection capabilities for more subtle and varied forms of harmful content. Additionally, existing benchmark corpora and annotated datasets often contain biases, necessitating ongoing efforts to create more balanced and representative datasets (Kovács et al., 2021; Nghiem et al., 2024; Yin and Zubiaga, 2021).

When applying NLP techniques in the field of mental health and violence, such as trauma research, there is a noticeable lack of expert-annotated datasets. This limitation hinders the effective clinical use of NLP tools, pointing to a need to create and validate comprehensive datasets. Recent research highlights that while NLP models, including LLMs like ChatGPT, show promise in mental health analysis, there remains a significant gap due to inadequate expert-annotated data, which affects their reliability and accuracy (Yang et al., 2023). Furthermore, annotation is often costly and time-consuming when relying solely on human experts, suggesting more efficient methods to generate labeled data (Goel et al., 2023; Ji et al., 2023). Additionally, the potential for model hallucination and the production of inaccurate outputs underscores the necessity for rigorous evaluations and the development of inherently interpretable methods (Chung et al., 2023). Research into PTSD detection has progressed, but more scalable NLP workflows are needed to better analyze patient narratives and clinical notes. While NLP techniques show promise in mental health prediction, further refinement is required to improve their performance in clinical settings (Xu et al., 2023).

2.2 How Do People Talk About Violence? - Insights from Social Science Research

Understanding the nuances of communication about violence is essential for grasping the broader social dynamics at play. There is a crucial distinction between talking about violence and talking violently, which can significantly affect how individuals communicate and understand these topics. Talking about violence involves discussing experiences, events, and the impacts of violent acts, often with a focus on understanding, processing, and finding ways to cope with or address such experiences. On the other hand, talking violently refers to using language that is aggressive, hateful, or intended to incite violence. This type of speech is more prevalent in anonymous and unmoderated forums, where individuals feel emboldened to express harmful views without fear of repercussions (Siegel, 2020).

Talking About Violence vs. Speaking Violently. An example of how different individuals speak about violence can be seen in the contrasting accounts of torture between detainees and interrogators. Research indicates that these accounts differ significantly, highlighting the complex ways in which trauma is processed and expressed. Survivors of torture, particularly in the context of genocide, may experience speechlessness due to the severe impact on memory, making it difficult to verbalize their experiences (Sandick, 2012; Lehrner and Yehuda, 2018). Shame, especially in cases involving sexual violence, can further hinder survivors from disclosing details in court (Sharratt, 2016). While some witnesses find the tribunal setting therapeutic, offering a sense of relief and justice, others may be retraumatized by the process (Ciorciari and Heindel, 2016; Brounéus, 2008). In contrast, former interrogators, fearing legal repercussions, may alter or deny their testimonies to avoid responsibility (Holness and Ramji-Nogales, 2016; Kanavou and Path, 2017).

Although this is a specific use case, it illustrates how accounts of violence can differ depending on the individuals involved. Similar effects may be observed in mental health online forums, where victims of sexual abuse discuss their experiences. In these contexts, speechlessness may appear as a challenge in articulating traumatic events, often reflected in the use of metaphors or less direct language to convey their experiences (Bogen et al., 2024). Shame and speechlessness also play a significant role in how individuals discuss violent experiences online. In more anonymous forums, the lack of personal identification can reduce feelings of shame, making it easier for people to discuss their experiences openly. However, the same anonymity can lead to a lack of empathy and an increase in hostile language, as users feel detached from the consequences of their words. In contrast, in more supportive online communities, such as mental health forums, the structured and empathetic environment can help individuals overcome speechlessness and shame associated with their trauma. These forums often have guidelines and moderators that ensure respectful and supportive interactions, which can encourage individuals to share their experiences more openly and honestly (Prescott et al., 2020; Strand et al., 2020).

When looking at hate speech, however, the picture looks completely different. Due to the anonymity provided by online forums, individuals usually do not hesitate to use derogatory or hateful language. The lack of accountability in these spaces encourages people to express violent and offensive views that they might otherwise suppress in more regulated or identifiable environments. This phenomenon is particularly prevalent in free and anonymous forums where the absence of moderation allows hate speech to proliferate unchecked (Windisch et al., 2022).

Discussing Experienced Violence. When individuals talk about their experiences of violence online, several fundamental theories provide insights into this behavior. The Uses and Gratifications Theory (Valkenburg et al., 2006) emphasizes that individuals actively seek media that meets their psychological needs. For some, talking about their experiences of violence online provides a sense of catharsis, allowing them to express pent-up frustrations in a socially acceptable manner. Additionally, the feedback and engagement received from such discussions, including likes, shares, and comments, can fulfill the need for recognition and validation, reinforcing the behavior.

Furthermore, discussing experienced violence online can be a means of seeking validation and support from like-minded individuals, creating a sense of community and belonging. This can appeal to polarized or marginalized groups, where sharing personal stories can reinforce group identity and solidarity (Papacharissi, 2002). The anonymity provided by online platforms can also encourage individuals to share their experiences without fear of judgment or retribution, which is critical for those dealing with shame or trauma.

Engaging in Violent Discourse. Engaging in violent or abusive language online can also be influenced by anonymity, though in a notably different way. One contributing factor is the Online Disinhibition Effect (Suler, 2004), which suggests that anonymity, invisibility, and the absence of authority in online spaces lower social inhibitions, leading individuals to express thoughts and emotions more freely, including aggressive and abusive behavior. This phenomenon is supported by studies linking bystanders and perpetrators of online hate (Wachs and Wright, 2018). Deindividuation Theory (Diener, 1980) further reinforces this idea, proposing that when individuals see themselves as anonymous members of a group, they are more likely to engage in behaviors they would typically suppress, such as aggression and abusive language (c.f., Bilewicz and Soral (2020) on deindividuation and political radicalization).

Online environments can create situations where individuals' aggressive tendencies are triggered and expressed through violent or abusive language. Social media platforms often amplify this behavior through their algorithms, prioritizing content that generates strong emotional responses, including outrage and anger. Posts containing abusive language or violent rhetoric are more likely to be shared and commented on, increasing their visibility and reinforcing the poster's behavior (Craker and March, 2016; Chen, 2017).

2.3 Opportunities and Risks

Having identified key areas where NLP is applied for violence detection and examined the factors influencing how individuals discuss violence, the following subsection explores the current opportunities and risks associated with language modeling in this domain. This section provides an overview of recent advances in computational violence research, highlighting both the potential and challenges in this evolving field.

2.3.1 Measuring Violence

Researchers apply a wide range of methods to measure violence. Surveys and questionnaires collect data on individuals' experiences and attitudes, such as those related to intimate partner violence (Ureña et al., 2015) or political violence (Westwood et al., 2022). Official statistics and records from law enforcement and healthcare agencies provide a more formal perspective on the prevalence and types of violence (Basile et al., 2011; Dumont et al., 2012). Self-report measures offer personal narratives, while observational studies provide direct insights by examining violence in specific settings. Content analysis focuses on media representations of violence, exploring how events like sexual violence cases are portrayed and the potential impact on public policy (Aroustamian, 2020). These methods vary significantly in how they measure violence, each with its operational definitions. Some research takes a broad view, including psychological and emotional abuse, while others focus solely on physical violence. The context–whether domestic, public, or online–also influences the methods and definitions used. This diversity reflects the complexity of violence as a social phenomenon and the need for tailored measurement approaches to address specific research questions and contexts.

Limitations of What NLP Can Measure. In online discourse, key aspects of how people discuss violence—like body language and silence—are not captured in text, limiting NLP's effectiveness in this research. Body language, such as facial expressions and gestures, adds context and depth to communication, particularly in sensitive topics like violence. These nonverbal cues convey emotions and attitudes that text alone cannot fully capture, leading to a less nuanced understanding of an individual's experiences (Kumari and Ganagwar, 2018). Silence, too, plays a significant role in communication, especially in discussions of traumatic events. Pauses and moments of silence can indicate emotional states such as hesitation or difficulty in verbalizing painful memories, but these are invisible in text-based analysis (Sandick, 2012). The online environment also affects how violence is discussed. Anonymity can lead to either more candid or hostile expressions, which NLP struggles to interpret fully, as it cannot grasp the situational and emotional nuances behind the words. Given these limitations, researchers must be cautious in interpreting text-based data on violence. While NLP provides insights into language patterns and sentiment, it cannot fully capture the multifaceted nature of human communication, including body language and silence.

Opportunities for NLP. Despite these challenges, NLP offers significant advantages. It can analyze large volumes of text data from diverse sources like social media, news articles, and legal documents, improving the representativeness of studies. By automating data processing, NLP reduces the resource burden of longitudinal studies and helps protect participant confidentiality through the analysis of anonymized data.

By focusing on a clear and thoughtful operationalization of violence, some challenges in capturing it through NLP can be mitigated. While NLP cannot encompass body language, advancements in detecting non-verbal cues on social media, such as the use of emojis, are offering new insights into how nuances in text can be interpreted (Park et al., 2014). For instance, robust operationalization in hate speech detection might involve using an established categorization framework (Waseem et al., 2017). In violence and mental health research, it could mean identifying concepts that approximate psychological disorders to better capture them through text data (Schirmer et al., 2024b).

2.3.2 Large-Scale Assessments

Methods from computational sciences, particularly NLP, enable the collection and analysis of text data on an unprecedented scale (Lazer et al., 2020). Large-scale text datasets from sources such as web content and social media significantly broaden the scope of social science studies. NLP techniques allow researchers to include more subjects than traditional methodologies permit. For instance, evaluating public opinion on a topic can be done by analyzing thousands of topic-related tweets (or any other messages posted on similar platforms such as Twitter/X) within hours, whereas achieving the same sample size with traditional surveys could take years (Ji et al., 2015; Van Lent et al., 2017).

The efficiency of NLP methods means researchers often aim to capture all available text data on a topic rather than relying on random sampling. This can lead to more extensive, diverse, and potentially more representative samples (Pfeffer et al., 2023). Additionally, large-scale text data collection allows for studying phenomena over broader time intervals and at finer temporal resolutions than traditional methods. For example, researchers can analyze the evolution of public sentiment on social media over time, providing detailed insights into changing opinions and trends.

Information Overload & Professional Search. A notable example is the vast amount of court transcripts from genocide tribunals, such as the *International Criminal Tribunal of the Former Yugoslavia (ICTY)*, which provides approximately 2.5 million pages of transcripts online (ICTY, 2016). Searching for specific content in a text corpus of this magnitude typically requires extensive manual research capacity (Hoang and Schneider, 2018). In another relevant context, researchers studying domestic violence often rely on large datasets of police reports, social services records, and medical reports to identify patterns and risk factors associated with violent incidents. Tools and approaches have been developed to augment this type of search and help limit manual efforts, such as automating search strategies or text extraction from documents (MacFarlane et al., 2021; Russell-Rose et al., 2021). However, searching for specific text passages in large corpora remains challenging even with suitable tools, particularly when the search is recall-oriented (Bache, 2011; Kaptein et al., 2013; Noor and Bashir, 2015).

NLP can efficiently process vast amounts of data, making it manageable in ways that manual analysis cannot. This applies to court documents, pre-existing corpora, and typical data sources in computational social science, like social media and online forums. For instance, in studying online radicalization, NLP techniques can analyze social media posts to identify extremist language and behaviors (Torregrosa et al., 2023). Specifically, researchers can use NLP to detect hate speech, propaganda, and recruitment language by employing sentiment analysis to gauge the tone of posts, topic modeling to uncover underlying themes, and named entity recognition to identify key figures and organizations (Cambria and White, 2014). Furthermore, NLP can track changes in language use over time, allowing researchers to pinpoint the emergence and spread of radical ideologies (El Barachi et al., 2022).

Cost & Speed. Another opportunity that the application of NLP in violence research–and in various fields more generally–holds is the significant saving of cost and time. A clear example is professional search, which involves searching for information in a work context. Professional search is domain-specific and requires expertise. It differs from web searches, often taking significantly longer to meet specific information needs. For instance, librarians spend an average of 26.9 hours on systematic reviews, highlighting the time-consuming nature of this task (Bullers et al., 2018). Professional search also involves limited time and budget constraints, making tools that classify text passages valuable for reducing search efforts (Russell-Rose et al., 2021).

Reducing costs and speeding up processes is crucial across industries, and leveraging NLP is an effective way to achieve this. NLP algorithms can automate the classification and extraction of relevant information from large text corpora, reducing manual effort and mitigating issues like human fatigue and subjectivity. This automation ensures more consistent and reliable results (Li et al., 2020). Various tools and algorithms have been developed to save time when searching through text, though their effectiveness varies by context(MacFarlane et al., 2021). However, these tools are not widely adopted, and not all are suitable for content-based search in text documents, where enhanced keyword search may be more helpful. Further, human factors also play a crucial role in extensive searches. Lengthy, time-consuming searches can lead to fatigue, reducing search quality, and manual searches are more prone to subjectivity. This highlights the need for automated search algorithms.

In automated processes beyond professional search, NLP can be an effective first step. For instance, when working with an annotated dataset for binary classification, the labels can significantly narrow down the text material for deeper analysis. Many NLP models, like BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), are freely available and open-source, offering powerful capabilities at no cost. However, some LLMs are not free, making it essential to assess whether a smaller model suffices for the task. Strong open-source LLMs like LLaMA (Touvron et al., 2023) provide robust performance without costs. In some cases, investing in LLMs is justified if it reduces other expenses, such as annotation (see Section 4.6).

2.3.3 Implementation & Impact

NLP in violence research can facilitate the development of tools that mitigate online violence more effectively, thus amplifying the impact of this research. By automating the detection of violent language, hate speech, and radicalization efforts in real-time, NLP enables the creation of monitoring systems that can swiftly identify and flag harmful content. For example, social media platforms like Twitter and Facebook can integrate NLP algorithms to automatically filter and report posts that promote violence, enabling quicker responses and content removal. For example, Fortuna and Nunes (2018) have demonstrated the effectiveness of NLP in detecting hate speech with high accuracy. NLP has successfully been used to screen for suicide risk on social media. By examining posts for expressions of hopelessness, despair, or suicidal ideation, NLP tools can alert mental health professionals to individuals in crisis, potentially saving lives through timely intervention (Coppersmith et al., 2018; Fernandes et al., 2018). This approach allows for continuous, large-scale monitoring beyond what could be achieved manually. NLP also plays an important role in the detection of sexual violence, a subject often surrounded by stigma and silence. By analyzing narratives shared on social media or support forums, NLP can identify reports of sexual violence, thus helping to "break the silence" and bring attention to these cases (Khatua et al., 2018). This not only aids in supporting survivors but also helps gather data for research and policy-making aimed at preventing such violence. Applying NLP in these sensitive areas requires balancing privacy with the need for intervention. Clear guidelines on when and how to act are important to protect rights and ensure safety. Using NLP tools ethically is key to maintaining public trust and achieving effective violence prevention (Karabacak and Margetis, 2023).

2.3.4 Data Reliability

The reliability of social media data is influenced by its origin. Platforms like Twitter, Facebook, and Instagram provide real-time information, but user-generated content is often informal, abbreviated, and context-specific, making it difficult to interpret accurately (Xiang et al., 2018). Additionally, the presence of bots and fake accounts can skew research findings by spreading misleading content, as seen in Burnap and Williams (2016), which found that automated accounts significantly impacted Twitter data used to study hate speech. Misinformation further complicates the reliability of social media data, distorting the understanding of social behaviors and trends (Wu et al., 2019).

Data annotation is important for ensuring reliability in NLP tasks, especially when dealing with sensitive topics like violence. Expert annotators with backgrounds in psychology, linguistics, or criminology tend to provide more reliable labels than crowd-sourced annotators, who may lack the necessary expertise. For instance, Waseem et al. (2017) found that expert annotations were more consistent and accurate in categorizing abusive language compared to crowd-sourced annotations. To improve reliability, researchers can use detailed guidelines, training sessions, and inter-rater reliability checks.

Interpreting results from NLP analyses of social media data requires caution, as the reliability of interpretations hinges on the quality of the data and the robustness of analytical methods. For instance, sentiment analysis can reveal general trends in public sentiment toward violent events, but it may miss nuanced emotions or contextual subtleties (Balahur et al., 2009). Cultural and linguistic diversity among social media users also means that results can be highly context-dependent, risking misinterpretation if regional dialects, idiomatic expressions, or cultural references are not considered. Olteanu et al. (2015), for example, highlighted the importance of contextual understanding in accurately interpreting Twitter data during crisis events.

2.3.5 Ethical Perspectives

Given the sensitive nature of violence research, ethical considerations must be integrated throughout the study design. This includes responsibly anonymizing data, ethically managing the annotation process, and publishing results in a respectful manner. NLP in violence research presents both ethical advantages and challenges, which are detailed in the following sections.

Pre-existing Data. The text data used, such as clinical reports or social media posts, is typically created independently of the research, avoiding the need for additional participant involvement that might trigger distressing experiences. However, using pre-existing data, especially from public sources like social media, raises ethical concerns about privacy and consent. While social media data is publicly accessible, users may not expect their posts to be used in research. Balancing data use with respect for individual rights and consent is a complex ethical issue. Researchers must implement stringent consent procedures and clearly communicate the research's purpose and scope (Taylor and Pagliari, 2018; Webb et al., 2017).

Participant Burden and Psychological Impact. At the same time, using pre-existing data eliminates the need to recruit participants specifically for the study, thereby reducing potential psychological distress or retraumatization from recalling violent experiences. However, the annotation process can expose annotators to distressing content, posing ethical challenges. Research designs should include provisions for annotators' well-being, such as training, psychological support, and strategies for managing exposure to traumatic material. Providing adequate support and guidelines for annotators is essential but remains a challenging aspect of the research process (Costello et al., 2023; Thomas et al., 2017).

Anonymization and Privacy. Advanced NLP techniques can anonymize data, protecting individuals' identities and privacy. This involves removing personally identifiable information to prevent re-identification, thereby reducing the risk of harm from data breaches or unintended disclosures. However, re-identification remains a potential risk, especially with large, detailed datasets. Researchers must apply robust anonymization methods and continuously assess their effectiveness. Ethical oversight and strong data governance frameworks are also essential to manage and mitigate privacy risks (Ienca et al., 2018; McLachlan and McHarg, 2005).

Responding to Violence Detection. When NLP identifies sensitive information, a key ethical challenge is determining when and how to act. If ongoing risks of harm or violence are detected, researchers must decide whether to intervene, report to authorities, or take other actions. This requires balancing the urgency of preventing harm with the ethical obligations to protect privacy, avoid further distress, and adhere to legal guidelines. Establishing clear protocols for action is essential to ensure the ethical handling of sensitive findings. Researchers should work closely with ethical review boards and legal advisors to develop strategies that prioritize safety while respecting individual rights and confidentiality (Golder et al., 2017; Webb et al., 2017).

2.4 Applying NLP in Violence Research: Summary

This section has shown that despite advancements in hate speech detection and NLP applications in mental health and violence research, significant challenges remain, including the subjective nature of hate speech, biases in existing datasets, and a lack of expert-annotated data. NLP techniques offer opportunities for large-scale assessments and real-time analysis, which are essential for timely intervention and support. However, these techniques must be applied ethically, ensuring data anonymization, privacy, and the responsible handling of sensitive findings. High-quality, carefully curated datasets and meticulous annotation processes are vital for the reliability of NLP research in this domain. Recent research highlights the necessity of expert-annotated datasets, particularly in the field of mental health and violence, such as trauma research, where the lack of such data hinders the effective clinical use of NLP tools.

Despite these advancements, significant research gaps remain:

1. **Inclusive Datasets**: There is a need for more inclusive datasets that go beyond social media to include diverse sources such as legal documents, clinical records, and historical archives. Existing benchmark corpora and annotated datasets often contain biases, necessitating ongoing efforts to create more balanced and representative datasets (Kovács et al., 2021; Yin and Zubiaga, 2021).

2. **Sophisticated Models & Contextual Understanding**: The development of more advanced models is needed to accurately differentiate types of violence and understand the subtleties of implied versus explicit violent content. Although transformer-based models have improved detection accuracy, they still struggle with nuanced language and multimedia content like emojis and GIFs (Hermida and Santos, 2023). Enhancing the models' ability to capture context, considering cultural, linguistic, and regional differences, is crucial, especially given the subjective nature of hate speech and violence definitions across different contexts (MacAvaney et al., 2019).

3. **Ethical Considerations**: Exploring the ethical implications of using NLP in sensitive contexts is necessary to balance the benefits of automated detection with the need to protect privacy and dignity, such as in analyzing data related to intimate partner violence (Tang et al., 2023). Ethical challenges include the need for robust anonymization to prevent reidentification, managing the psychological impact on annotators exposed to distressing content, and deciding when and how to act on sensitive findings detected by NLP (Taylor and Pagliari, 2018; Webb et al., 2017; Costello et al., 2023; Thomas et al., 2017).

4. **Violence Perspectives**: Capturing perspectives of both victims and aggressors enhances understanding of violence's complex dynamics. Research should focus on these perspectives to enable more nuanced prevention efforts tailored to the specific needs of individuals and groups (Grych and Hamby, 2014).

5. **Interdisciplinary Collaboration**: Encouraging more interdisciplinary collaboration between NLP experts and social scientists to develop comprehensive approaches to violence detection and understanding. Creating and validating comprehensive datasets through interdisciplinary efforts are crucial for advancing NLP applications in this field (Goel et al., 2023; Ji et al., 2023).

6. **Impact of NLP-Based Violence Research**: It often remains unclear how research findings can translate into real-world impact. Current research is still falling short in bridging the gap between academic studies and practical applications to effectively support online interventions or policy implementations (Windisch et al., 2022).

These research gaps are addressed through the research questions outlined in this dissertation. Specifically, items 1 to 3 explore how NLP methods can be effectively applied to violence studies (RQ1), item 4 examines the integration of multiple perspectives (RQ2), and items 5 and 6 focus on assessing the dimension of impact within this research area.

3 Modeling Language for Violence Research

As described in Section 2.1.1, the proliferation of potentially violence-related text due to the rise of social media and increased digitization offers a significant opportunity for the application of NLP methods. Social media platforms produce large volumes of unstructured text data, making manual analysis difficult. Traditional methods, such as keyword searches and manual content review, are often insufficient due to the sheer volume and complexity of the data. Additionally, the subtle and varied nature of violent content makes it difficult to detect using conventional approaches.

This chapter provides an overview of the primary methodologies in NLP applied in this dissertation. Given the rapid evolution of the NLP field, with models constantly being improved, this section does not attempt to cover all existing NLP methods. Instead, it focuses on the most relevant and widely used techniques in violence research. Alongside a general overview, specific examples from violence research are provided to illustrate the application of these methods. The chapter is structured to follow the typical NLP pipeline, starting with a discussion on the types of data required, methods for data collection, and the necessary steps for data preparation. It then provides an overview of standard NLP methods and models used in violence research. Finally, the chapter offers insights into how NLP results can be integrated into further statistical modeling, demonstrating the practical applications of these techniques in advancing our understanding of violence. The overall process is illustrated in Figure 3.

3.1 Text as Data

The basic idea of NLP is to use text as data, transforming written language into analyzable information (Cambria and White, 2014; Jurafsky and Martin, 2021). This method allows researchers to harness vast amounts of unstructured text generated daily, converting it into actionable insights. By treating text as data, NLP enables the extraction of specific patterns and trends from sources like social media posts, news articles, and court transcripts.

In violence research, using text-as-data offers several concrete benefits. Firstly, it enhances descriptive analysis, which is crucial for capturing the complexity of violent incidents. For instance, analyzing social media posts about domestic abuse can reveal context, motives, and consequences that purely quantitative data might miss. Secondly, text-as-data promotes discovery. Unlike traditional methods that rely on predefined hypotheses, text-based analysis is more exploratory. This allows researchers to uncover unexpected trends, such as new forms of online harassment or shifts in violent rhetoric on forums. Moreover, text-based automated analyses bridge the gap between quantitative and qualitative research. Quantitative researchers can use NLP to process and analyze large text datasets, converting them into structured formats for statistical analysis. At the same time, qualitative researchers can interpret these analyses to understand rich, contextual details. For example, sentiment analysis can quantify the emotional tone of gang-related tweets, while thematic analysis can uncover underlying narratives in extremist propaganda.

Integrating text-as-data builds on decades of research traditions and theories in social sciences and computational fields. It provides a robust toolkit that enhances established methods, facilitating the cumulative advancement of knowledge. Computational Social Science (CSS) evolves by incorporating new data sources and analytical techniques, allowing researchers to continuously develop and refine traditional methodologies. By employing text-as-data, researchers ensure their methods remain relevant and effective in addressing contemporary issues in violence research.

The shift towards text-as-data signifies a move towards more inductive, discovery-oriented research, enriching quantitative and qualitative approaches. This paradigm shift offers a comprehensive framework for advancing the study of violence, enabling the identification of patterns and trends that might otherwise remain hidden.

3.2 Data Collection & Preparation

Datasets provide empirical evidence to validate theories and hypotheses, ensuring that research findings are credible and impactful. Well-documented datasets enable the replicability of studies, which is fundamental to scientific integrity, allowing other researchers to falsify and build upon existing work. Additionally, access to diverse and comprehensive datasets fosters innovation by enabling the exploration of new questions and the development of novel methodologies.

The data collection and preparation process varies significantly depending on the data source. For example, accessing a social media platform via an API allows researchers to collect large volumes of real-time or historical data. In contrast, archived documents, such as historical newspapers or legal records, provide rich contextual insights but require careful digitization and text extraction processes (Piotrowski, 2012). Researchers often use preexisting datasets to save time and resources, benefiting from their predefined structures and metadata for consistency and easier integration. However, when these datasets are unavailable or insufficient, they may need to scrape data from online sources, requiring custom scripts for extraction and extensive cleaning to ensure accuracy and usability (Luscombe et al., 2022). Creating a dataset involves defining research objectives, selecting data collection methods, and cleaning and preprocessing the data to ensure accuracy and consistency. Key considerations include deciding which information to include, such as specific text fields (e.g., comments or replies), handling multiple languages, and determining the relevance of each data point (Tabassum and Patil, 2020).

3.2.1 Preprocessing & Annotation

Data preprocessing is a crucial step in creating datasets for computational social science research, as it ensures that the raw data collected is transformed into a format suitable for analysis. This stage involves several steps, starting with data cleaning, which includes removing duplicates, correcting errors, and handling missing values. Ensuring the consistency and reliability of the data is paramount, as even minor inaccuracies can significantly impact the results of the analysis.

Preprocessing. To adapt the data structure to further NLP modeling, techniques like tokenization, stopword removal, and text normalization are required as a standard part of the NLP pipeline (Jurafsky and Martin, 2021; Tabassum and Patil, 2020). When working with subtle concepts, such as online violence, the correct data format is important. For TikTok comments, data cleaning may involve decisions around handling emojis, depending on the study's focus. For instance, retaining emojis can benefit sentiment analysis since they often convey strong emotions. Conversely, if the analysis is centered solely on text, removing emojis might help streamline the data. In contrast, preprocessing court data can be more complex due to the structured nature of legal documents. Standardizing formatting inconsistencies, such as varying fonts and spacing, is crucial. It is also important to remove irrelevant elements like page numbers, headers, and footers to keep the focus on content. Additionally, splitting lengthy documents into smaller, manageable sections can facilitate more efficient analysis, but might lead to loss of context.

The extent and nature of these steps also depend on the model used. For instance, minimal preprocessing may be required when working with raw text data, particularly in transformer-based architectures or other large language models (LLMs, such as GPT-4). LLMs can process raw text effectively, even when it is unstructured. However, it is worth noting that these models sometimes fail to capture nuances such as capitalization, which can be significant in specific analyses (Dalal and Singh, 2024). Therefore, while minimal preprocessing might suffice for LLMs, certain refinements may still be necessary to enhance the accuracy of the analysis.

Annotation Annotating text data involves assigning labels or tags to specific text elements based on predefined categories relevant to the research questions. It should be based on clear guidelines to ensure consistency and accuracy (Röttger et al., 2021). There are several methods of annotating text data, each suited to different types of research and data sets. Manual annotation involves human annotators reading the text and applying the appropriate labels, as applied throughout the majority of papers presented in this dissertation (Schirmer et al., 2022, 2023a). This method is highly accurate but can be time-consuming and expensive. Semi-automated annotation combines human effort with machine assistance, where algorithms provide initial annotations that human annotators refine (see Study 6, Section 4.6, Matter et al. (2024)). Fully automated annotation relies on advanced NLP techniques, such as LLMs, to label the text, which can be efficient for large data sets but may lack the nuanced understanding of human annotators (Nasution and Onan, 2024). Each method has advantages and trade-offs, and the choice depends on factors such as the size of the data set, the annotation task's complexity, and available resources.

When focusing on annotating text data related to violence, the process becomes even more complex and sensitive. Violence can be described in numerous ways, ranging from physical acts to psychological threats, and can be explicit or implicit. Accurate annotation in this context requires a deep understanding of violent language and context nuances (Li et al., 2023; Waseem et al., 2017).

The selection of annotators is crucial for this task. Ideally, the team should consist of multiple researchers with expertise in the subject matter to ensure reliability and depth of understanding. Including student assistants and crowdworkers can increase the volume of annotated data, but it is essential that they receive thorough training and clear guidelines to maintain consistency. Having multiple annotators for each text segment to cross-validate annotations and resolve discrepancies through discussion or adjudication by a more experienced researcher is also beneficial. Annotator agreement, which refers to consistency among different annotators, is critical to the annotation process. It is typically measured using statistical metrics such as Cohen's Kappa (Cohen, 1968) or Krippendorff's Alpha (Krippendorff, 2004). High annotator agreement indicates that the guidelines are clear and the annotators apply them consistently, enhancing the annotated data's reliability. Discrepancies can arise from individual biases, varying interpretations of the guidelines, or the inherent ambiguity in the text itself. Regular training and calibration sessions can help mitigate these issues, but it is important to recognize and address the limitations of annotator agreement in the research design (Teruel et al., 2018).

Ethical considerations are paramount when annotating text data on violence. Annotators may be exposed to disturbing content that can have psychological impacts. Therefore, providing support resources and establishing protocols to mitigate these risks is important. Additionally, maintaining the anonymity and confidentiality of sensitive information is critical to protect individuals' privacy and comply with ethical standards. Annotators should be trained to recognize and respect these ethical boundaries, ensuring that the data is handled with the utmost care. Ethical concerns in annotations include potential biases that can impact data accuracy. Annotators must be aware of how their biases, such as subconscious hostility or gender bias, may influence labeling decisions. For instance, these biases could lead to inconsistent or skewed labeling of hate speech or misogynistic content, compromising the reliability of the analysis (Geva et al., 2019).

In conclusion, annotating text data is vital in computational social science research, particularly when dealing with sensitive topics such as violence. It requires careful methodological rigor, ethical sensitivity, and practical considerations to produce high-quality, reliable data that can drive meaningful analysis and insights.

3.2.2 Example: Court Transcripts

To illustrate the application of this pipeline, this subsection provides an overview of the typical preprocessing of court transcripts as conducted in this dissertation, specifically during the creation of the Genocide Transcript Corpus (GTC) (Studies 1-4, with slightly diverging approaches, see Sections 4.1-4.4). This dataset includes texts from genocide tribunals, offering insights into severe human rights violations and the profound trauma experienced by victims and witnesses. The GTC encompasses 90 cases from the International Criminal Tribunal for Rwanda (ICTR), the International Criminal Tribunal for the former Yugoslavia (ICTY), and the Extraordinary Chambers in

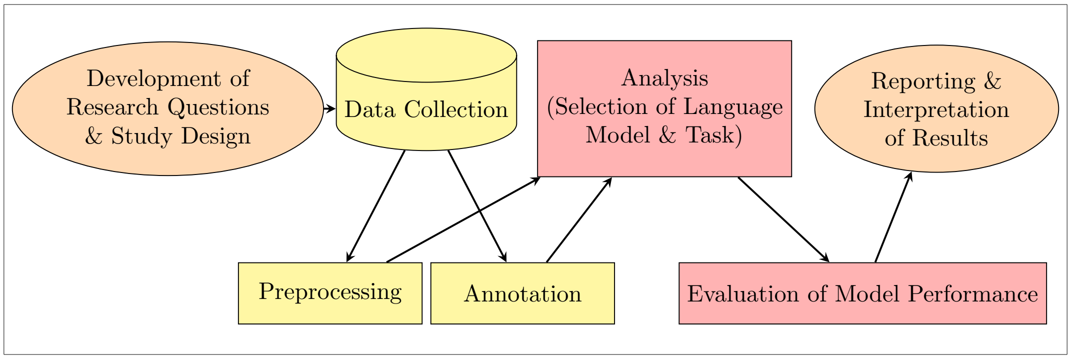

Development of Research Questions & Study Design
Data Collection
Analysis (Selection of Language Model & Task)
Reporting & Interpretation of Results
Preprocessing
Annotation
Evaluation of Model Performance

Figure 3: NLP Analysis Process Flowchart.

the Courts of Cambodia (ECCC).

The Genocide Transcript Corpus (GTC) (Study 4.2, Schirmer et al. (2022)) was initially developed as the first annotated data set for NLP in the field of genocide research, providing important benchmark values for text classification and has been extended with further benchmarking to its final version (Study 4.3, Schirmer et al. (2023a)).

For the ICTR and ICTY, we selected five cases based on final judgments where the accused received life imprisonment. We randomly picked six additional cases featuring witness testimonies for the prosecution, excluding expert witnesses. The ECCC provided only two accessible cases online, with 15 transcripts each. While this case selection might not be fully representative, it captures essential crimes and high-profile cases, which are crucial for understanding the critical aspects of these tribunals. This focus ensures that the most significant and impactful crimes are analyzed, highlighting the core issues addressed by these legal proceedings, thus not diminishing the validity of the research.

The GTC dataset spans a wide range of witness statements from diverse backgrounds, such as former soldiers, tortured prisoners, and guards who committed torture, resulting in transcripts that vary in focus from political and administrative details to vivid descriptions of violence and trauma. This dataset also includes prosecution witnesses who participated in the genocide. To structure the transcripts for NLP analysis, we annotated them based on the speaker's role in legal proceedings. Initially obtained by scraping HTML links, the transcripts were lightly preprocessed to remove line numbers, URLs, HTML tags, and technical document information. We tagged statements by judges, lawyers, witnesses, and the accused, distinguishing between witness questioning (JudgeQA or LawyerQA) and legal proceedings discussions (JudgeProc or LawyerProc), while editorial comments and formal parts were marked as court proceedings. Text segments varied from short responses to multi-paragraph replies, with those exceeding 500 tokens split for NLP efficiency. The updated GTC contains 52,845 text segments from 90 transcripts, categorized by speaker role or court proceedings. This facilitates accurate and consistent NLP analysis for future research (see Table 2 for an overview of GTC variables).

Category	**Information**
Case	Tribunal, case number, accused
Transcript	Document ID, URL-link to the original transcript, date
Witness	Witness name or pseudonym, number of witnesses per transcript
Text	Speaker (e.g., Witness, LawyerQA), text, trauma label
Annotation	Annotation ID, start ID, and document ID

Table 2: Variable Overview Genocide Transcript Corpus (GTC)

Determining whether a text snippet contains trauma-related content is a complex task. For example, a witness might describe observing the deportation of their neighbors, yet this alone does not clarify whether the witness themselves experienced a direct threat to their own life. Consequently, it is important to note that diagnosing whether a witness is traumatized cannot be done through their court testimony alone. Such diagnoses require an in-depth assessment of the witness's personal history, far beyond the scope of witness transcripts. In the context of this dissertation, we used the APA trauma definition outlined in Section 2 to guide our labeling of text snippets that could potentially describe traumatic events. We manually labeled all text segments containing witness statements, including accounts of military attacks, bombings, killings, physical violence, threats, humiliation, and the destruction or looting of property, provided these events were directly observed by the witness and evident from the text segment alone. To ensure consistency, we calculated inter-rater reliability using Fleiss' kappa, an adaptation of Cohen's kappa for more than two annotators Fleiss (1971), yielding high agreement among the annotators and validating the labeling process (for further details, please refer to Study 4.3, Schirmer et al. (2023a)). This resulted in a cleaned, structured dataset with expert-annotated labels well-suited for NLP modeling and the supervised training of a language model.

3.3 From N-Grams to Large Language Models

Model selection is the next critical step once the data is available in a processable format. Which language model is best-suited will heavily depend on the specific task it is required to perform. This subsection first provides an overview of the most common NLP models, with the concrete tasks being described in the next section(Section 3.4).

Recently, large language models (LLMs) have led to a significant push in the development and application of NLP, both in research and public interest. This surge is due to their easy applicability and versatile AI assistant approaches, such as OpenAI's GPT-family (e.g., GPT-3, GPT-4) (Brown et al., 2020; OpenAI, 2023), or open source models, such as Llama3 (Touvron et al., 2023). NLP has evolved through various models, each contributing uniquely to the field. Early models like Bag of Words (BoW) and TF-IDF (Term Frequency-Inverse Document Frequency) laid the groundwork by representing text based on word frequency and importance. For instance, BoW could classify emails as spam by counting specific keywords, while TF-IDF improved search engines by identifying the most relevant documents based on term uniqueness. These models, though simple, were essential in early text classification and sentiment analysis tasks, such as determining whether movie reviews were positive or negative. Later advancements like Word2Vec and GloVe introduced word embeddings, which capture the semantic meaning of words through neural networks. These embeddings enable tasks like word similarity, where Word2Vec might identify that ”king” relates to ”queen” as ”man” does to ”woman,” or analogy generation, supporting more nuanced language understanding in applications such as recommendation systems and question-answering systems (Jurafsky and Martin, 2021).

Deep learning brought about Recurrent Neural Networks (RNNs), which process data sequences, making them ideal for language modeling and text generation tasks. RNNs can predict the next word in a sentence, which is helpful for applications like autocomplete in text editors. Enhancements such as Long Short-Term Memory (LSTM) networks enabled these models to capture long-term dependencies in text. This makes LSTMs effective for tasks like translating long sentences or generating coherent paragraphs. However, while they perform well with sequential data, these models struggle to maintain context over extended passages, such as understanding the plot of a novel from start to finish.

The introduction of Transformer models marked a significant shift in NLP by using self-attention mechanisms to improve performance on tasks requiring context understanding. The BERT (Bidirectional Encoder Representations from Transformers) family (Devlin et al., 2019), including variants like RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019), enables advanced tasks such as question answering and named entity recognition by understanding word context in both directions. For example, BERT can accurately identify entities in a sentence, like recognizing ”Paris” as a location and ”Einstein” as a person.

Recent models like T5 (Text-to-Text Transfer Transformer) and XLNet unify NLP tasks into a text-to-text format, enhancing versatility across applications such as translation and summarization. T5 can translate English to French or summarize long articles, while XLNet improves on BERT by better capturing bidirectional context, enhancing tasks like text classification. The release of LLMs like GPT-3 and GPT-4 further advanced NLP by using extensive data and computational power to generate human-like text.

Notably, BERT-like models and LLMs, like GPT, differ significantly in their architecture. BERT, an encoder-only model, is designed primarily for understanding and generating context-aware embeddings from text, making it effective for tasks like question answering and sentiment analysis. It processes text bi-directionally, analyzing the entire sentence from both directions to grasp context (Devlin et al., 2019). In contrast, LLMs like GPT generate coherent and contextually relevant text by predicting the following word sequence, making them ideal for text completion, creative writing, and conversational agents (Brown et al., 2020).

Summary. There is no definitive model for NLP tasks, and the most sophisticated model is not always necessary. For violence research, traditional models like Bag of Words (BoW) and TF-IDF are useful for basic text classification, such as identifying violent content in social media posts. Word embeddings like Word2Vec and GloVe help understand the semantic context of violence-related terms. RNNs, LSTMs, and GRUs are ideal for analyzing sequential data, such as patterns in longitudinal reports of violence. Transformer models like BERT and XLNet are useful for tasks requiring deep contextual understanding, like detecting nuanced sentiments in discussions about violence. Ultimately, the choice of model should align with the research goals, data nature, and available computational resources.

3.4 Text Classification, Sentiment Analysis, and Topic Modeling as Commonly Used Tasks

Identifying and categorizing violent content in social media posts, news articles, online forums, and other documents is a significant focus in contemporary violence research, addressed by scholars across various disciplines, such as digital humanities (Keydar, 2022), psychology (Botelle et al., 2022), and computer science (Ribeiro et al., 2021). While text classification is a central method for this task, sentiment analysis and topic modeling are also commonly employed. This section offers an overview of these methods, highlighting examples from violence research.

Text Classification. Text classification, a core NLP task, involves tasks such as category labeling or sentiment analysis (Jurafsky and Martin, 2021). Fine-tuning BERT-based models has become a popular strategy for text classification (Devlin et al., 2019), often outperforming traditional machine learning methods and other neural networks (Li et al., 2020; Schirmer et al., 2023a). BERT has been successfully applied across a range of sentiment analysis tasks, from aspect-based sentiment analysis (Sun et al., 2019) to assessing the impact of COVID-19 on social life (Singh et al., 2021). A notable example is hate speech detection, where BERT has been applied to classify tweets for content such as racism, sexism, or hate speech (Mozafari et al., 2020).

BERT has also been adapted to specialized domains. Examples include COVID-Twitter-BERT (Müller et al., 2020) and BioBERT (Lee et al., 2020) for biomedical text. In legal NLP tasks, LegalBERT (Chalkidis et al., 2020) is widely used for its high performance in legal contexts. For topics related to trauma and genocide, ConfliBERT (Hu et al., 2022), trained on text data from international conflicts, is relevant (Schirmer et al., 2023a). Another promising variant is HateBERT (Caselli et al., 2020), trained on over 1 million posts from banned Reddit communities. It is particularly suited for detecting hate speech and potentially traumatic content due to its focus on harmful language.

In violence research, classification tasks can be binary or multi-class, depending on the research question. Binary classification involves categorizing text into two distinct classes, such as identifying whether a piece of content is violent or non-violent. This approach is straightforward and useful for clear-cut distinctions, such as filtering violent content from social media platforms. On the other hand, multi-class classification involves assigning text to one of several categories, such as distinguishing between different types of violence (e.g., physical, psychological, or sexual violence) or identifying various sources of violent content (e.g., hate speech, extremist propaganda, or domestic abuse). Multi-class classification provides a more nuanced understanding of violence, allowing researchers to identify specific patterns and trends within different types of violent behavior, which can inform targeted interventions and policy decisions.

Sentiment Analysis. Sentiment analysis is an NLP task that determines the sentiment or emotional tone behind a body of text (Medhat et al., 2014). It classifies the text into positive, negative, or neutral categories and can further identify more specific emotions like anger, joy, sadness, or fear. Sentiment analysis is widely used to gauge public opinion, monitor brand reputation, and understand customer feedback (Liu, 2020). Sentiment analysis approaches can be broadly classified into lexicon-based and machine learning-based methods. Lexicon-based approaches rely on predefined lists of words (lexicons) associated with specific sentiments, e.g., lexicons such as AFINN (Nielsen, 2011) or VADER (Hutto and Gilbert, 2014). These methods count the occurrences of sentiment-laden words in a text to determine the overall sentiment. While lexicon-based approaches are relatively simple and interpretable, they may struggle with context and sarcasm, leading to less accurate results. Machine learning-based approaches, on the other hand, use algorithms to learn from labeled data and predict sentiment based on patterns in the text. These methods, including supervised learning techniques like Support Vector Machines (SVMs) and neural networks, can capture complex linguistic nuances and context, making them more accurate and robust than lexicon-based methods. However, they require large amounts of annotated data for training and can be more computationally intensive (Medhat et al., 2014; Wankhade et al., 2022).

In violence research, sentiment analysis can help understand the emotional undertones of discussions related to violence. For instance, it can be used to analyze social media posts or online forums to detect rising anger or hostility that might precede violent events (Bermingham et al., 2009). Researchers have employed sentiment analysis to study the public's reaction to violent incidents, such as terrorist attacks or mass shootings, by examining the sentiment expressed in social media posts and news articles (Mansour, 2018). This helps identify emotional patterns and the spread of fear or panic in the aftermath of such events.

Topic Modeling. Topic modeling is an unsupervised machine learning technique used to identify themes or topics within a collection of documents. By analyzing the co-occurrence patterns of words, topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), can discover hidden structures in text data, grouping related terms into topics. This method is particularly useful for summarizing large datasets, uncovering underlying themes, and aiding in content organization (Mohr and Bogdanov, 2013). However, assessing the quality of the topics in terms of human interpretability can sometimes be challenging (Morstatter et al., 2015).

Traditional topic modeling methods like LDA rely on the co-occurrence of words within documents to identify topics. While effective, LDA can struggle with capturing the semantic nuances of language, particularly in complex or large datasets. BERT-based topic modeling (BERTopic) (Grootendorst, 2022) is a more advanced approach that uses the capabilities of transformer models like BERT to generate dense embeddings for each document. BERTopic uses these embeddings to cluster documents into topics, capturing more nuanced relationships between words and improving topic coherence. This approach allows BERTopic to better understand the context and semantic meaning of words than LDA, making it particularly effective for analyzing complex and varied text data.

In the context of violence research, topic modeling helps to identify and categorize different forms and contexts of violence. For example, it can be used to analyze social media posts, news articles, or victim reports to uncover prevalent themes related to various types of violence, such as domestic abuse, hate crimes, or political violence (Xue et al., 2019). Concrete examples from violence studies include using topic modeling to analyze news coverage of violent events to identify how media frames these incidents and the most common narratives (Tourni et al., 2024). Another application is examining online forums and social media platforms to detect emerging trends and hotspots of violent discussions, which can inform early warning systems and intervention strategies (Lee and Jang, 2023; Smoliarova et al., 2018). Moreover, topic modeling has been applied to court documents to offer insights into historical and legal contexts of violence. For example, Keydar (2022) employed topic modeling on the Eichmann trial, one of the most influential trials regarding the Holocaust, uncovering topics related to the deportation and ghettoization of Jews. This demonstrates how topic modeling can be effectively utilized in violence research beyond social media, providing a broader perspective on historical and judicial aspects of violence.

Model Evaluation & Benchmarking. NLP models are evaluated and benchmarked using various metrics and datasets to ensure their effectiveness and reliability. Benchmarking compares a model's performance against standardized datasets, pre-established baselines, or other models. Standard evaluation metrics include precision, recall, F1-score, and accuracy, which measure the model's ability to correctly predict and classify data (Jurafsky and Martin, 2021). Cross-validation techniques and confusion matrices are also employed to assess model performance and identify areas for improvement. Benchmarking helps compare different models and approaches, providing a clear understanding of their strengths and weaknesses. However, benchmarking in the context of violence research can be challenging. While hate speech detection is common and has well-established datasets, other areas, like analyzing genocide court transcripts, lack comparable benchmarks, making evaluation more difficult.

In the context of violence research, NLP models are evaluated and benchmarked using specific datasets related to violent content. For example, models analyzing social media posts for violent themes might be evaluated based on their accuracy in detecting hate speech or threats, using labeled datasets across domains (Guimaraes et al., 2023) or comparing them to an established dataset within the domain, such as the Stormfront dataset based on posts from a white supremacist forum (de Gibert et al., 2018b). Furthermore, topic modeling techniques applied to news articles on violent events are evaluated by their coherence and ability to correctly identify and categorize different forms of violence. By benchmarking these models on relevant datasets, researchers can refine their tools for detecting and understanding violence, leading to better early warning systems and intervention strategies. However, for more specialized tasks, such as analyzing genocide court transcripts, finding appropriate benchmarks is more challenging due to the lack of standardized datasets and the unique nature of the content.

3.5 NLP as the Basis for Further Modeling

After training and evaluating an NLP model or applying techniques such as topic modeling, further statistical methods can be employed to extract deeper insights from the results. Depending on the research question, various approaches can be taken. For instance, comparing sentiment values across different datasets or periods can reveal shifts in public opinion or the emotional tone surrounding specific events. Investigating the increase of violence over time after performing text classification can help identify trends and patterns in violent incidents, potentially uncovering seasonal spikes or responses to particular triggers. Temporal analysis can also examine the evolution of topics or sentiments, allowing researchers to track how discussions about violence change over time.

Statistical Modeling. After training and evaluating an NLP model, further insights can be gained through statistical methods such as regression analysis, time-series analysis, and clustering, each offering specific applications for understanding violence-related data. Regression analysis can help identify relationships between variables, such as the correlation between the frequency of violent language and socioeconomic factors. This can be particularly useful in exploring how different socioeconomic conditions may be associated with the prevalence of violent content in social media posts or news reports. Time-series analysis helps examine trends and patterns over time. For example, it can identify periods of increased violence in social media posts following major political events (Florio et al., 2020). This method can track the rise and fall of hate speech or violent content in response to specific events, such as elections or terrorist attacks. Clustering can group similar documents or posts, aiding in identifying different subcategories of violence (Ni et al., 2020). This can reveal patterns that might not be obvious through manual analysis, such as various forms of violence mentioned in victim reports.

Explainable Artificial Intelligence. To provide interpretable and transparent explanations of NLP model predictions, researchers leverage eXplainable Artificial Intelligence (XAI) to uncover the mechanisms behind specific tasks (Arrieta et al., 2020). XAI is a rapidly evolving field within NLP, particularly in relation to state-of-the-art models like BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), which often function as black boxes (Belinkov et al., 2020; Mosca et al., 2022; Schirmer et al., 2023a). A central method in XAI are local feature attribution explanations, which help us understand why a model made a specific prediction by quantifying the importance of each input feature. Different techniques are used to achieve this. Some methods, like those based on output gradients (Sundararajan et al., 2017), assess how changes in input features affect the prediction. Others, such as DeepLIFT (Shrikumar et al., 2017), analyze neural network structures to trace the contribution of each feature. LIME (Ribeiro et al., 2016) simplifies the model's behavior in a specific instance by creating an easily interpretable approximation, while SHAP (Lundberg and Lee, 2017) uses Shapley values from game theory to fairly distribute importance scores among input features, making the model's decisions more transparent.

XAI can be particularly helpful for violence detection using NLP. For example, in social media analysis, SHAP can identify which words or phrases contribute most to a classification of violent content, aiding in the understanding of hate speech patterns. Concept-based explanations, such as Completeness-Aware Concept-Based Explanations (Yeh et al., 2019), can cluster instances of violent language, revealing underlying themes without needing predefined concept annotations. This interpretability is crucial for developing more effective and transparent early warning systems and intervention strategies in violence research.

3.6 Methods Overview

This section has demonstrated how NLP methodologies provide diverse tools for violence research. These techniques open up significant opportunities for investigating complex social issues related to violence, but they also come with their own set of challenges.

The case studies presented in the next chapter will provide more detailed perspectives on these opportunities and challenges, showcasing how NLP can shed light on aspects of violence that might otherwise remain hidden. I will present seven case studies where NLP methods have been applied to various research questions, including violence detection, media framing of violent events, and the analysis of court transcripts from genocide trials. Table 3 illustrates the mapping between these studies and the methods discussed earlier. Through each study, I will highlight the value of the interdisciplinary application of NLP methods in advancing computational social science research within the context of violence.

Study	**Data Source**	**Methods**
Study 1: Talking about Torture	Court Transcripts (Khmer Rouge Tribunal)	BERT-Based Binary Classification, Sentiment Analysis, Generalized Linear Mixed Regression Model, Qualitative Content Analysis
Study 2: Topic-Based Classification	Court Documents (ICTY, ICTR, ECCC)	Dataset Creation Based on Court Documents, Expert Annotations, BERT-Based Binary Classification, Benchmarking Experiments
Study 3: Uncovering Trauma	Court Documents (ICTY, ICTR, ECCC)	Dataset Creation Based on Court Documents, Expert Annotations, BERT-Based Binary Classification, Active Learning, Benchmarking Experiments, Explainable AI: SHAP Values
Study 4: GENTRAC	Genocide Transcript Corpus as a Basis, Open to Other Court Documents from Selected Courts	Creation of an Open-Source Online Tool, Sophisticated Parsing of Documents, BERT-Based Binary Classification, Web-Based Visualization of Trauma Content
Study 5: Language of Trauma	Court Transcripts and Social Media Forums (Reddit, Counseling Forum, Incel Forum)	Creation of a Cross-Domain Trauma Dataset, Expert & Crowdworker Annotations, Systematic Language Model Comparisons, Explainable AI: SHAP Values, SLALOM Feature Importance, Concept Learning
Study 6: Incels	Full Scraping of Incel Forum incels.is	Web Scraping and Preprocessing, Expert Annotations, LLM-Supported Annotations, Linear Regression and Proportion Analysis
Study 7: Child Exposure on TikTok	TikTok Comments	Topic Modeling (BERTopic), Dictionary-Based Comment Exploration, *t*-Tests for Group Differences
Study 8: LLMs and Thematic Analysis	YouTube Comments	LLM-Supported Annotations, Thematic Analysis

Table 3: Overview of the data collection and analysis methods used for the studies.

4 Studies

This section presents eight studies on NLP for violence detection.

Study 1 lies in the area of **Genocide Studies** and explores the differences in torture-related witness statements during the Khmer Rouge Tribunal. The study employs a **three-phased sequential mixed methods design** combining **NLP**, **sentiment analysis (SA)**, and **qualitative content analysis (QCA)** to identify disparities between testimonies of former detainees and interrogators. It contributes to mixed methods research by integrating digital approaches with NLP and data transformation.

Study 2 is situated at the intersection of **NLP and Genocide Studies** and focuses on developing a new dataset for topic-based paragraph classification in genocide-related court transcripts. The methodology involves creating the **Genocide Transcript Corpus (GTC)** and using **transformer-based approaches** for paragraph identification of violence-related witness statements. The study also explores **transfer learning** within this domain to establish benchmark performances.

Study 3 focuses on **NLP and Trauma Detection** within genocide tribunals. The study extends the Genocide Transcript Corpus (GTC) and applies NLP methods to analyze trauma in witness statements from genocide tribunals. The methodology includes using **binary classification algorithms** with **transformer models (BERTbase and HateBERT)** and applying **Explainable Artificial Intelligence (XAI)** to understand the model's classifications. The study aims to develop trauma-informed legal procedures.

Study 4 lies in the area of **NLP and Legal Studies** and introduces GENTRAC, a tool for detecting and analyzing potentially traumatic content in genocide and mass atrocity court transcripts. The methodology involves developing an **NLP-based tool** to process and analyze court transcripts, visualizing the density of traumatic content, and providing statistical analysis. The tool is designed to handle extensive data from international criminal courts and aims to improve trauma-informed legal procedures.

Study 5 is situated more generally in the fields of **Psychological Trauma and NLP** and models traumatic event descriptions across various domains using explainable AI. The study employs several NLP models, including **RoBERTa and GPT-4**, to predict traumatic events across datasets such as genocide-related court data, PTSD discussions on Reddit, and counseling conversations. The methodology focuses on **training language models**, clustering trauma-related language, and exploring the transferability of findings across different trauma contexts.

Study 6 lies in the area of **Social Psychology and NLP** and investigates the increase of violent speech in Incel communities using human-guided GPT-4 prompt iteration. The study involves scraping a large dataset from incels.is and categorizing the posts into non-violent, explicitly violent, and implicitly violent content. The methodology includes **human coding**, **tuning GPT-3.5 and GPT-4 models**, and evaluating the models' performance in detecting violent speech, with a focus on content moderation and online radicalization.

Study 7 focuses on **Social Media Studies and Child Safety** and examines children's exposure and user engagement on TikTok. The study analyzes comments and content related to children on the platform, categorizing videos by themes such as Family, Fashion, and Sports. The methodology involves **statistical analysis** of comments, focusing on appearance-based comments and the prevalence of revealing clothing in videos, highlighting potential risks and engagement patterns on the platform.

Study 8 is embedded in the field of **Digital Humanities** and explores the synergy between human intelligence and AI in researching hate speech on social media. The study focuses on using **GPT-4** for **thematic analysis (TA)** of a YouTube dataset related to the representation of Roma migrants in Sweden. The methodology involves an experimental study combining human expertise with AI's scalability, analyzing the advantages and limitations of employing large language models in qualitative research.

4.1 Study 1: Talking About Torture: A Novel Approach to the Mixed Methods Analysis of Genocide-Related Witness Statements in the Khmer Rouge Tribunal

This publication is RELEVANT TO THE EXAMINATION.

Authors

Miriam Schirmer, Jürgen Pfeffer, Sven Hilbert

In

Journal of Mixed Methods Research, 0(0), 1-20, peer-reviewed, published 11-2023, https://doi.org/10.1177/15586898231218463.

Abstract

This study investigates differences in torture-related witness statements during the Khmer Rouge Tribunal. It follows a three-phased sequential mixed methods design to identify disparities between testimonies of former detainees and interrogators and to examine how different methods complement each other for a comprehensive perspective on witness accounts. This includes training a natural language processing (NLP) model, sentiment analysis (SA), and qualitative content analysis (QCA). The qualitative and NLP-based analyses showed apparent differences between witness groups; a significant difference in sentiment values could not be detected. This study presents the first mixed methods approach based on court transcripts in genocide research. Its digital approach contributes to mixed methods research (MMR) by showing how NLP and data transformation can contribute to integration.

Contribution of Thesis Author

Theoretical conceptualization, data curation, methodological design, formal analysis, visualization, manuscript writing, revision, and editing.

Journal of Mixed Methods Research
2023, Vol. 0(0) 1–20

DOI: 10.1177/15586898231218463
journals.sagepub.com/home/mmr
Sage

Talking About Torture: A Novel Approach to the Mixed Methods Analysis of Genocide-Related Witness Statements in the Khmer Rouge Tribunal

Miriam Schirmer[1,2], Jürgen Pfeffer[1], and Sven Hilbert[2]

Abstract

This study investigates differences in torture-related witness statements during the Khmer Rouge Tribunal. It follows a three-phased sequential mixed methods design to identify disparities between testimonies of former detainees and interrogators and to examine how different methods complement each other for a comprehensive perspective on witness accounts. This includes training a natural language processing (NLP) model, sentiment analysis (SA), and qualitative content analysis (QCA). The qualitative and NLP-based analyses showed apparent differences between witness groups; a significant difference in sentiment values could not be detected. This study presents the first mixed methods approach based on court transcripts in genocide research. Its digital approach contributes to mixed methods research (MMR) by showing how NLP and data transformation can contribute to integration.

Keywords

genocide, torture, mixed methods, natural language processing, data transformation

Introduction

Witnesses play a critical role in genocide-related trials. Through their testimonies, they provide crucial evidence but also tell personal stories of their survival. However, recounting experiences that might have been traumatic poses a significant challenge for individuals who have experienced extreme violence during a genocide. Consequently, protocols for witness support and psychological assistance are standard practice in most genocide tribunals (e.g., International Criminal Court, n.d.). While researchers have acknowledged the emotional difficulties of testifying about genocide and torture (Ciorciari & Heindel, 2016), there is limited research on how emotionally

[1]School of Social Sciences and Technology, Technical University of Munich, Munich, Germany
[2]Faculty of Human Sciences, University of Regensburg, Regensburg, Germany

Corresponding Author:
Miriam Schirmer, School of Social Sciences and Technology, Technical University of Munich, Germany.
Email: miriam.schirmer@tum.de

difficult experiences manifest in statements of different witness groups and the subsequent impact on individual testimonies.

To close this gap, this study analyzes how individual witnesses recount their experiences with torture in court by analyzing witness statements from Case 001 of the *Extraordinary Chambers in the Courts of Cambodia* (ECCC) against Kaing Guek Eav, who oversaw the torture prison S-21 during the Cambodian genocide between 1975 and 1979. S-21 is particularly relevant in this context since its primary purpose was to extract confessions from perceived enemies of the state (Chandler, 1999).

Generally, post-atrocity trials, such as the ECCC, involve a diverse range of witnesses, including direct victim survivors, family members of deceased victims, civilians, experts, as well as individuals involved in the atrocity, such as guards, soldiers, and political figures. We specifically examine the testimonies of two distinct groups directly involved in the act of torture: survivors who were imprisoned and interrogators who were stationed at the prison. Comparing the statements of these two witness groups potentially identifies differences in their discourse on torture. It potentially informs considerations on whether distinct approaches should be employed when examining different witness groups in court.

This study aims to detect such differences in testimonies of former detainees and interrogators through an exploratory sequential mixed methods design (Fetters et al., 2013; Moseholm & Fetters, 2017) with three phases. The first stage of the study consists of a natural language processing-based (NLP-based) classification task, followed by three different sentiment analyses (SAs) and a qualitative thematic analysis. Natural language processing generally refers to utilizing algorithms to process human language (Jurafsky & Martin, 2021). In the context of a classification task, such an algorithm is trained to assign text segments to predefined categories. While the algorithm gives details about how clear a distinction between different categories can be made, it can be seen as a black box that does not explain which characteristics led to the classification. Therefore, further analysis is needed. Sentiment analysis, an NLP technique that identifies subjective information in text like emotions, shows promise: Diving deeper into group differences, it allows for a more nuanced understanding of the emotional content of the witness testimonies. Finally, the qualitative analysis provides a more profound understanding of the contextual factors that may have influenced the statements. By integrating methods involving the convergence of multiple approaches, we aim to overcome the limitations of individual methods, providing a multi-perspective and more robust view of witness statements (Creswell & Plano Clark, 2018; Tashakkori et al., 2021).

To the best of our knowledge, this type of study design has not found its way into genocide studies so far and thus presents a novel approach to this field of research. Simultaneously, it contributes to mixed methods research (MMR) by addressing the issue of integration (Bryman, 2007; Fetters et al., 2013) on design, methods, and interpretation levels. With digital data transformation being a key aspect, we are presenting an alternative version of a *digital mixed methods design* (O'Halloran et al., 2018) involving transforming qualitative data into quantitative data and comparing patterns across different data dimensions through data mining. Consequently, this study follows recent advances in MMR to combine NLP and machine learning with qualitative analysis (Guetterman et al., 2018; Sripathi et al., 2023). More specifically, both quantitative text mining methods and qualitative content analysis (QCA) are applied to transcripts from ECCC Case 001 to address differences in witness statements between former detainees and former interrogators of the S-21 prison (see Figure 1 for an overview of the study design). First, we follow up on whether speech- and content-based differences between the testimonies of former detainees and interrogators can be identified, and if so, which methods are appropriate for detecting such differences (Research Question (RQ) 1). This is done by answering three subordinate questions of how accurately an NLP-based model can classify text segments as either belonging to a former detainee or interrogator (RQ 1.1, quantitative), whether sentiment values in witness testimonies of former detainees and interrogators exhibit significant differences (RQ 1.2, quantitative), and how thematic patterns in statements of former detainees and interrogators differ (RQ 1.3, qualitative). Finally, this paper synthesizes previous findings by analyzing how the results of the different methodological approaches of RQ 1 complement or challenge each other regarding the differences in witness statements made by former detainees and interrogators (RQ 2).

Background

Trauma and Torture as Part of the Testimony

The American Psychological Association (APA) describes trauma as "exposure to actual or threatened death, serious injury, or sexual violence" that is either experienced directly or witnessed (American Psychological Association, 2013, p. 271). Torture and imprisonment in the context of genocide can clearly be assigned to this trauma concept, especially since former detainees and interrogators were consistently confronted with death, torture, and physical violence in S-21.

How torture was applied in the S-21 prison in Cambodia has been studied by numerous authors (Chandler, 1999; Hinton, 2016). When it comes to testifying in Case 001 of the ECCC, however, only a few works have been published. Among them are studies that mainly focused on the prison leader's role and the situation of individual witnesses in court (Hinton, 2016) or applied a more quantitative approach by applying a software-based text analysis of testimonies (Brönnimann et al., 2013). Addressing the psychological impact of testifying before the ECCC, Ciorciari and Heindel (2016) concluded that the participation of traumatized persons in the court proceedings represented an "emotionally difficult process" (p. 184) and consequently called for sensitivity training of court professionals to provide support for traumatized testifiers and avoid re-traumatization. Apart from the mentioned examples, there is a lack of studies specifically examining how the content of accounts of torture differs between former detainees and interrogators, leaving a gap for further research in this area.

Detainees' Versus Interrogators' Accounts of Torture

Still, there is evidence suggesting that accounts of torture could differ significantly between detainees and interrogators. First, speechlessness can present in cases of experienced trauma, especially in the context of genocide (Sandick, 2012). Experiences of torture and other forms of extreme violence are processed in a complex way, making it especially difficult for victim survivors to verbally express such events from their past. Closely connected to the phenomenon of speechlessness is the impact of genocide and torture on memory: Individuals who have undergone traumatic experiences may either remember them vividly or experience memory repression, where the memories of torture are not consolidated, resulting in an inability to recall the event in detail (Lehrner & Yehuda, 2018). Similar psychological effects were found in perpetrators of genocide, who experienced high levels of post-traumatic stress symptoms in the aftermath of the genocide (Barnes-Ceeney et al., 2019). Especially torture methods involving sexual violence or degrading techniques can create feelings of shame for victim survivors. This shame may affect the witnesses' statements in court, potentially resulting in the omission of uncomfortable details (Sharratt, 2016). At the same time, however, the tribunal might serve as a catalyst in the processing of traumatic experiences: several witnesses stressed their relief and happiness about contributing to justice through their testimony, making it a "cathartic courtroom experience" (Ciorciari & Heindel, 2016, p. 124). While some witnesses might experience this positive outcome, the testimony can have opposite effects on victim survivors and might even lead to re-traumatization (Brounéus, 2008). Lastly, legal implications could impact the testimony of former interrogators. Although it was improbable that former interrogators would face repercussions for testifying, the prospect of being charged for participating in torture may still have impacted their testimony and could have led to denying responsibility (Holness & Ramji-Nogales, 2016; Kanavou & Path, 2017). Altogether, these findings on speechlessness, memory, shame, emotional processing, re-traumatization, and legal implications as potential factors influencing witness testimony imply that former detainees and interrogators talk about torture differently in court.

Combining Mixed Methods, NLP, and Trauma Research

To shed more light on these potential differences, this study applies a mixed methods design to trauma and genocide research, incorporating NLP techniques. Natural language processing-based approaches are used more and more frequently in both mixed methods and trauma research, with authors from different backgrounds repeatedly emphasizing the great potential that NLP brings to MMR (Chang et al., 2021; Guetterman et al., 2018; Reinhold et al., 2022) and discussing the incorporation of big data into mixed methods designs (Bazeley, 2018; O'Halloran et al., 2018). Within the broad range of NLP tools, topic modeling (Ho et al., 2021) and SA (Colditz et al., 2019) have been popular methods applied in the context of mixed methods. Applying these methods in the context of trauma and violence research seems promising, especially with violence being a "too complex and pressing social problem to be subjected to methodological puritanism" (Thaler, 2017, p. 70). Similarly, Creswell and Zhang (2009) highlight the suitability of mixed methods in trauma research, as it allows for the inclusion of qualitative data in a traditionally more quantitative field, such as patient interviews, bridging the gap between research and practice. Several studies have further explored the intersection of violence and trauma using mixed methods, such as investigations into domestic violence and abuse (Bacchus et al., 2018) or childhood trauma (Boeije et al., 2013). Specifically for genocide research, however, mixed methods designs are scarce. For instance, focusing on World War II, Békés et al. (2021) analyzed survivors' narratives through a sequential mixed methods design, combining an interpretative phenomenological approach and quantitative comparison of codes drawn from interview material. Addressing more recent cases of mass atrocities, other studies discussed gender and genocide in Darfur (Kaiser & Hagan, 2015) or sexual violence in the Democratic Republic of the Congo (Kelly et al., 2011) with mixed methods designs. However, these studies did not discuss their contribution to MMR or address MMR-related challenges, such as integration. Additionally, NLP techniques have not yet found their way fully into genocide research except for individual papers that focused on topic-based classification or topic modeling in genocide-related court transcripts (Keydar, 2020; Schirmer et al., 2022; Schirmer et al., 2023). Therefore, the study's value lies in synthesizing state-of-the-art methods in NLP and MMR-specific questions, such as integration, applied to the analysis of transcripts of genocide tribunals. It tackles the challenge of preserving witness accounts of genocide survivors in historical documents (Keydar, 2020) while providing a concrete example of combining computational, statistical, and qualitative methods.

Methods

We follow an exploratory sequential design (Moseholm & Fetters, 2017) to identify differences between former interrogators and detainees when recounting experiences of torture during the ECCC. This is done in three stages, using NLP techniques and QCA (Mayring, 2015) to complement each other. Starting with a broad analysis, an NLP-based language model is applied for binary classification to distinguish between statements made by the two witness groups (Phase I).

This serves as a starting point for obtaining an overview of the differences in statements between the two witness groups. Given that the language model does not explain its classification decision, this step can only provide a general sense of how well an algorithm can distinguish between the two groups. In the second step, three different SAs are conducted to look for differences in witness accounts on a sentiment-based level (Phase II). Sentiment analysis is a valuable tool for analyzing the emotional content of witness statements, especially in the sensitive context of trauma and torture. By examining the sentiment expressed by each witness, we can gain a deeper understanding of the emotional aspects of their experiences. Concretely, SA could reveal that one witness group uses more negative language in their statements, offering valuable insights into the psychological impact of torture on both groups. Although current research suggests that trauma can successfully be detected through SA (Sawalha et al., 2022), it is essential to acknowledge that SA primarily provides insights into the overall sentiment expressed, which does not necessarily equate to trauma or an actual emotional state experienced by the witness. However, despite its limitations in capturing the full extent and complexity of traumatic experiences, analyzing the emotional tone and general sentiment expressed in the testimonies still provides a valuable understanding of the affective impact of traumatic experiences on individuals. Lastly, the statements are subjected to QCA, laying out differences between the two witness groups that go beyond automatically detectable patterns (Phase III). Aligned with this research design, the primary focus of this study is to demonstrate the effective utilization of the three methods within an MMR framework. We further provide a concrete application example from the field of genocide studies, illustrating how the MMR framework can be effectively employed in practice.

Material and Data Transformation

The data basis for this study consists of transcripts of the ECCC's Case 001 that are analyzed throughout all three phases using different methods. In the context of this study, Case 001 is highly relevant as it concerns the S-21 prison, where approximately 18,000 detainees were systematically tortured and killed in the process of obtaining information about perceived threats and conspiracies against the Khmer Rouge regime (Chandler, 1999). Throughout the proceedings of Case 001 from February to November 2009, a total of 55 witnesses were heard, among them nine expert witnesses, 17 fact witnesses, seven character witnesses, and 22 civil parties. In its final verdict for the case, the court concludes that—from a legal perspective—imprisonment in S-21 can be proven for four witnesses (Extraordinary Chambers in the Courts of Cambodia, Case 001, 2010). Since the present study deals explicitly with the conditions of imprisonment in S-21, only statements of these former detainees were included. Additionally, former workers stationed at S-21 in various functions, such as guards, medics, and interrogators, were heard among the 17 fact witnesses. Out of those, four witnesses were identified as former interrogators who admitted to interrogating and torturing detainees—their testimonies were also included in the study. This comprises statements of 8 witnesses, with all four detainee witnesses and four interrogator witnesses as per final judgment (Table 1). The analyzed sample consists of 986 pages of transcripts.

The transcripts used for this study contain organizational details about the case, the date, and the names of the witnesses to be heard. They provide a verbatim record of all spoken proceedings, covering witness testimonies, lawyer arguments, and judge rulings. Non-verbal information, such as physical reactions, is not included. All transcripts of the court proceedings are available to the public on the ECCC Web site in both English and French,[1] with Khmer also being used as an official language during the court proceedings. This study relied on the English versions of the transcripts for the analysis. Despite the possibility of translation biases and inaccuracies, NLP techniques can provide insight into the emotional content and language patterns by detecting patterns between words and phrases. Considering that professional court translators did the

Table 1. Overview of Witnesses Who Were Either Imprisoned in S-21 or Were Part of the S-21 Staff.

Date of Testimony	Witness	Role in S-21
06/29/2009	VN	Detainee
06/30/2009	CM	Detainee
07/01/2009	BM	Detainee
07/02/2009	NC	Detainee
07/15/2009	MN	Interrogator
07/15/2009	HH	Interrogator
07/16/2009		
07/20/2009		
07/21/2009	PK	Guard, interrogator
07/22/2009		
08/03/2009	LM	Guard, interrogator

Note. Due to the topic's sensitivity, witnesses are referred to by their initials. However, the witnesses' full names can be found in the openly published court documents for a comprehensive historical perspective, including additional individual details. Only testimonies that were transcribed in English are included.

translation, a certain quality of the translation can be assumed.[2] Biases are thus likely to affect both groups of former detainees and interrogators equally. Hence, the relative differences between the two groups should remain valid, especially when including the context of the witness statements during the qualitative part of the analysis.

The original transcripts were subjected to an elaborate transformation process to make the data suitable for NLP analysis. In this case, transformation refers to changing qualitative data into quantitative data, also referred to as *quantitizing* (Sandelowski et al., 2009). While quantitizing commonly includes the numerical representation of qualitative data, this study aims to show how quantitizing goes beyond merely assigning numerical values to qualitative data, for example, by including word frequencies. Instead, we created a new data format from the original transcripts, segmenting the transcript documents and assigning meta-variables, such as the witness role (detainee vs. interrogator), witness name, and the sentiment value of the respective statement. Due to its pre-structured form, this leads to a new dataset suitable for NLP and statistical analysis.

Diverging from the originally proposed design, our data transformation approach can be seen as a variation of the *digital mixed methods design* (O'Halloran et al., 2018) that expands data integration to encompass the conversion of qualitative data into quantitative data. However, while O'Halloran et al. (2018) start their approach with a qualitative discourse analysis and build the data transformation on that, this study starts with the data transformation as a first step. By incorporating transformation and data mining as critical components in our study to analyze patterns and trends across diverse data dimensions, we argue that our approach can be classified as a digital mixed methods design.

Transformation from qualitative to quantitative data → NLP exploration → Sentiment analysis and statistical models → Qualitative analysis

Assuming that transcripts provide appropriate insights into how various witness groups discuss torture, they served as data for *all* three steps of analysis. Therefore, the original documents had to be transformed into a form suitable for computational analysis. This was done by dividing each of the transcripts into text chunks of approximately 250 words each, leading to 439 text chunks in total. These text fragments were subsequently labeled according to whether the statement was made by either a former detainee (label "0") or a former interrogator (label "1") ($n_{\text{detainees}} = 204$; $n_{\text{interrogators}} = 235$), resulting in a new dataset for training an NLP model.[3] For the SA, the transcripts were also analyzed in this paragraph format. However, the text was split into individual words to include a word-based SA, which is required for assigning sentiment values through the word-based lexicon. Before finalizing the datasets, preprocessing was conducted. As a regular technique in NLP, preprocessing involves cleaning and transforming the raw text to make it suitable for further analysis. This step is crucial because it helps remove irrelevant text information that can affect the accuracy of the model's predictions. In our case, this involved the removal of punctuation and stop words (i.e., words that are irrelevant to the analysis, such as pronouns). For the qualitative phase, transcripts were analyzed in their original format.

Phase 1: Explorative NLP Classification

One of the state-of-the-art language models used in NLP is called *Bidirectional Encoder Representations from Transformers* (BERT; Devlin et al., 2019). Bidirectional Encoder Representations from Transformer is a pre-trained neural network that can predict outcomes based on the relationships between words and their surrounding context. In the first step of this study, BERT was trained on a labeled dataset to predict whether text passages could be automatically classified as either interrogator or detainee testimonies, aiming to identify potential differences between these groups on a general level. In line with standard practice, accuracy and F1 scores are reported. The accuracy score indicates the total number of correctly identified text segments, while the F1 scores depict the weighted average of precision and recall values.[4] It is important to note that BERT can classify text paragraphs into different categories. However, it does not provide further information on specific text characteristics to help distinguish between the two categories of interrogators and detainees. Nonetheless, the classification is still useful for distinguishing between interrogator and detainee testimonies, as it relies on the inherent differences in language use between these two groups.

Phase 2: Sentiment Analysis

Due to the emotional sensitivity of the analyzed material, SA was selected as a tool to measure specific sentiment values for each witness. This technique categorizes semantic structures according to their underlying emotional content (Liu, 2020). Generally, two types of SA can be distinguished: On the one hand, SA is performed by supervised machine learning, where a text corpus already labeled with different sentiments is used to train an algorithm. On the other hand, lexicon-based SA is based on specific lexicons whose individual words have already been assigned sentiment values. In this study, lexicon-based SA was chosen since a suitable training corpus for this type of text material is not available, and the amount of text used in this study is comparatively small for machine learning approaches.

Three different SAs based on numerical sentiment values were conducted to yield more comparable and generalizable results. The first lexicon used is AFINN (Nielsen, 2011), which assigns a numerical value between −5 and +5 to each word. Negative values indicate a negative emotion, whereas positive values indicate a positive emotion. SentimentR (Rinker, 2019) was used as the second lexicon to detect sentiment values for complete sentences. Accounting for valence shifters and similar modifiers, such as negators or amplifiers, SentimentR makes it easier to calculate sentiment values according to neighbor-word context. The third lexicon applied, VADER (Hutto & Gilbert, 2014), also identifies sentiment values on a sentence level by calculating positive, negative, and neutral sentence components and combining them in a compound value. After assigning sentiment values using all three SA tools, linear hierarchical models were estimated to see if the witness group significantly influenced the sentiment value. Despite the limited number of witnesses, a statistical analysis of this nature remains reliable as the model examines individual words (AFINN) or paragraphs (SentimentR and VADER) to compare sentiment values, thereby utilizing an adequate number of data points.

Phase 3: Qualitative Content Analysis

To prevent a "loss of depth and flexibility" (Driscoll et al., 2007, p. 25) as a disadvantage of quantitizing, witness statements underwent QCA to ensure contextual considerations. As computational emotion analysis has been criticized for not detecting multiple or implicit emotions (Poria et al., 2019), the qualitative phase is crucial to obtaining a comprehensive understanding of the witness testimonies. A qualitative approach also allows us to analyze individual testimonies in detail, identify frequently occurring themes, and avoid overlooking important nuances (Creswell & Plano Clark, 2018; Tashakkori et al., 2021). Since no framework for categories regarding the reconstruction of torture experiences in court exists, categories were inductively developed (Mayring, 2015), closely examining the data and deriving categories from the content itself.

The transcripts were coded using MAXQDA Analytics Pro 2020, analyzing them for each witness individually by reading through each page and establishing categories of recurring topics. This process involved identifying significant words and phrases in the statements that gave insights into how witnesses talked about their experiences of torture, including content and form. For instance, special attention was given to noticeable signs of emotional distress, such as pauses, interruptions, or any coping strategies mentioned by the witnesses during their testimony in court. With this focus in mind, roughly one third of the material was examined in a first step. Categories were developed through an open coding process, which involved two rounds of coding to allow for adjustments. After a meaningful set of codes was established from transcript samples, a second researcher independently went through the same third of the material using coding instructions. The resulting inter-rater reliability of κ = .82 indicates a high degree of agreement between the two researchers. After establishing these categories on a sample of the transcripts, we used them to code the remaining transcripts. Coded categories and themes were then analyzed and interpreted to provide a deeper understanding of the contextual factors that may have influenced the emotional content of the testimonies.

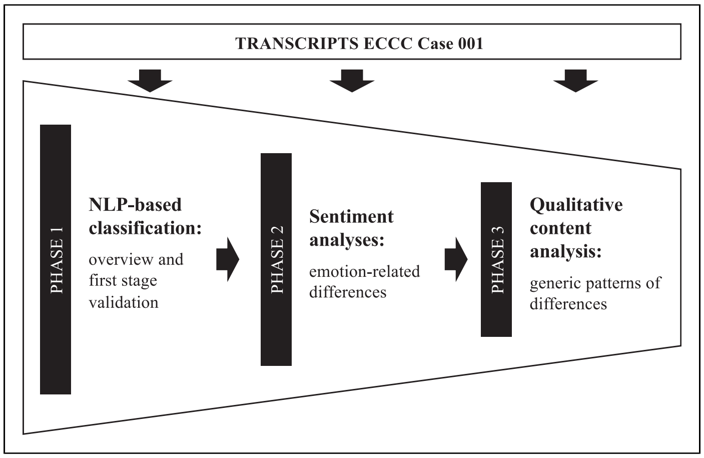

TRANSCRIPTS ECCC Case 001
PHASE 1
NLP-based classification:
overview and first stage validation
PHASE 2
Sentiment analyses:
emotion-related differences
PHASE 3
Qualitative content analysis:
generic patterns of differences

Figure 1. Overall research design. *Note.* This study consists of three sequential phases, with each methodological approach informing the others.

Results

BERT-Based Binary Classification

Applying BERT to predict the witness group of individual testimony passages yielded an accuracy and an F1 score of .95 each. Accordingly, the model correctly classified 95% of the text segments. Considering that the amount of text segments used to train BERT in this study was comparatively low, it is even more remarkable that a high percentage of correct classifications were reached, especially compared to benchmark studies in this field (Zhang et al., 2021). Differences in the accounts of former detainees and former interrogators regarding their testimony in court appear to exist on a speech-based level. *Research question 1.1* about whether an NLP-based model can classify text segments according to the respective witness group can thus be positively answered, emphasizing the model's high accuracy.

Sentiment Analysis

None of the three conducted SAs showed statistically significant differences in sentiment values between former interrogators and former detainees. Nonetheless, all three analyses yielded descriptively lower mean sentiment values for the group of interrogators (see Figure 2): The word-based SA with AFINN revealed a mean sentiment value of −1.64 for former interrogators (*SD* = 1.46) and −1.28 for former detainees (*SD* = 1.74). In the SA conducted with SentimentR, former interrogators had a sentiment value mean of −.21 (*SD* = .31), while former detainees' values were at a mean of −.10 (*SD* = .26). The VADER SA led to similar results: While the interrogators' mean sentiment value was −.58 (*SD* = .61), the detainees' mean sentiment value of −.33 (*SD* = .69) was close (possible sentiment value range for SentimentR and VADER: −1 to +1). Notably, all mean sentiment values were negative.

Results of the general linear models estimated for each SA confirm the descriptive results. Due to the dummy-coded group variable, the resulting parameter estimate equals the mean difference between the two groups controlled for the word or sentence count, respectively. For the AFINN SA, for example, the mean sentiment values for interrogators were lower by –.24 compared to the mean sentiment values for the detainees (see Table 2). Despite the lack of statistically significant differences, descriptively, sentiment values were lower for interrogators throughout all SAs.

To illustrate the descriptive differences more clearly, we explored samples of sentiment values in more detail. As the AFINN analysis is based solely on individual words and lacks contextual information, we focused on SentimentR and VADER. The highest positive and negative sentiment values obtained for both groups are shown in Figures 3 and 4, respectively.

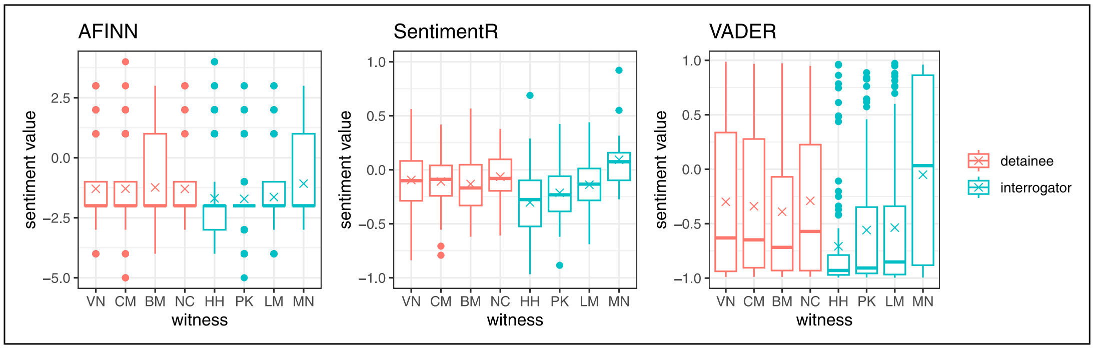


Figure 2. Boxplots of all three sentiment analyses conducted.

These examples confirm the descriptive trend of interrogators having more negative sentiment values assigned to their statements. In Examples 1 and 2, interrogators describe torture and executions in detail, leading to high negative sentiment values. Example 3 contains a detainee description of the living conditions in S-21 that involved the permanent threat of physical violence and death. Interestingly, the highest positive sentiment values were assigned to detainee statements: One statement describes positive emotions about testifying in court (Example 6), while the other draws positivity from painting tasks (Example 7), with the latter not necessarily being connected to actual positive emotions. The statements with the highest positive sentiment values for interrogators both describe political developments (Examples 4 and 5). Again, they are not directly linked to positive emotions experienced by the witnesses and refer to experiences made by the witnesses prior to their work at S-21.

In response to *research question 1.2*, which investigates whether differences in sentiment values between former detainees and interrogators exist, the study found no statistically significant differences. Nevertheless, the provided examples shed light on how such distinctions are established, descriptively illustrating that interrogators exhibited slightly more negative sentiment values than detainees.

Qualitative Analysis

Even though the SAs showed only limited differences between both witness groups, differences in how they talked about experienced violence become clearer by analyzing relevant text passages qualitatively. Through an inductive approach that involved identifying recurring thematic patterns in an open coding process, we identified three main patterns: expressions of emotional distress, technical expertise, and motivation to testify. Expressions of emotional distress referred to statements in the testimonies that conveyed feelings of anxiety, sadness, or trauma. Typical subcodes of this category include the witness needing a break or talking about lasting health problems. Technical expertise, on the other hand, encompassed statements that demonstrated the witness's knowledge and understanding of technical details related to the events they described, such as locations or procedures. A witness describing what kind of torture methods they were

Table 2. Overview of the Three Sentiment Analysis Models.

	Estimate	*SE*	*df*	*T*	*p*
AFINN: R^2_{marginal} = .014; $R^2_{\text{conditional}}$ = .017					
Intercept	−1.06	.12	4.23	−8.75	<.001
Witness group	−.24	.10	3.15	−2.43	.09
Number of words	0	0	3.98	−2.08	.11
SentimentR: R^2_{marginal} = .054; $R^2_{\text{conditional}}$ = .100					
Intercept	.07	.08	7.65	.79	.46
Witness group	−.06	.05	5.67	−1.23	.27
Number of sentences	0	0	7.84	−2.18	.06
Vader: R^2_{marginal} = .047; $R^2_{\text{conditional}}$ = .058					
Intercept	.01	.14	9.48	.05	.96
Witness group	−.15	.08	5.25	−1.90	.11
Number of sentences	−.01	0	9.36	−2.65	.03

Note. The witness group label was '0' for former detainees and '1' for former interrogators. Estimate = estimated parameter value; *SE* = standard error of the parameter estimate; *df* = degrees of freedom; *t* = *t* value; *p* = probability of committing a Type I error; R^2_{Marginal} = variance explained by fixed effects; $R^2_{\text{conditional}}$ = variance explained by both fixed and random effects.

	SentimentR	VADER
interrogator	**Example 1** […] regarding the techniques of torture we were taught how to torture the prisoners and to avoid that the prisoners died otherwise the confessions would be broken and we would be punished and we were trained on how to whip the prisoners with the stick on how to electrocute on how to use the plastic bag to suffocate them […] (witness PK; sentiment value **-1.13**)	**Example 2** […] all prisoners only waited until the day they would be interrogated and executed the prisoners would be interrogated and tortures would have been inflicted on and then they sustain the wounds then they could die of the wounds in the prison cell some people would be taken away to be executed after such interrogations […] (witness HH; sentiment value **-1.00**)
detainee	**Example 3** […] if they found out that we were eating insects we would be beaten also so we could do that only if we avoid being seen by the guards so the death is imminent and people died one after another the corpse would be removed and we ate our meal next to the dead body and we did not care anyway because we were like animals […] (witness VN; sentiment value **-0.84**)	(witness VN; sentiment value **-0.99**)

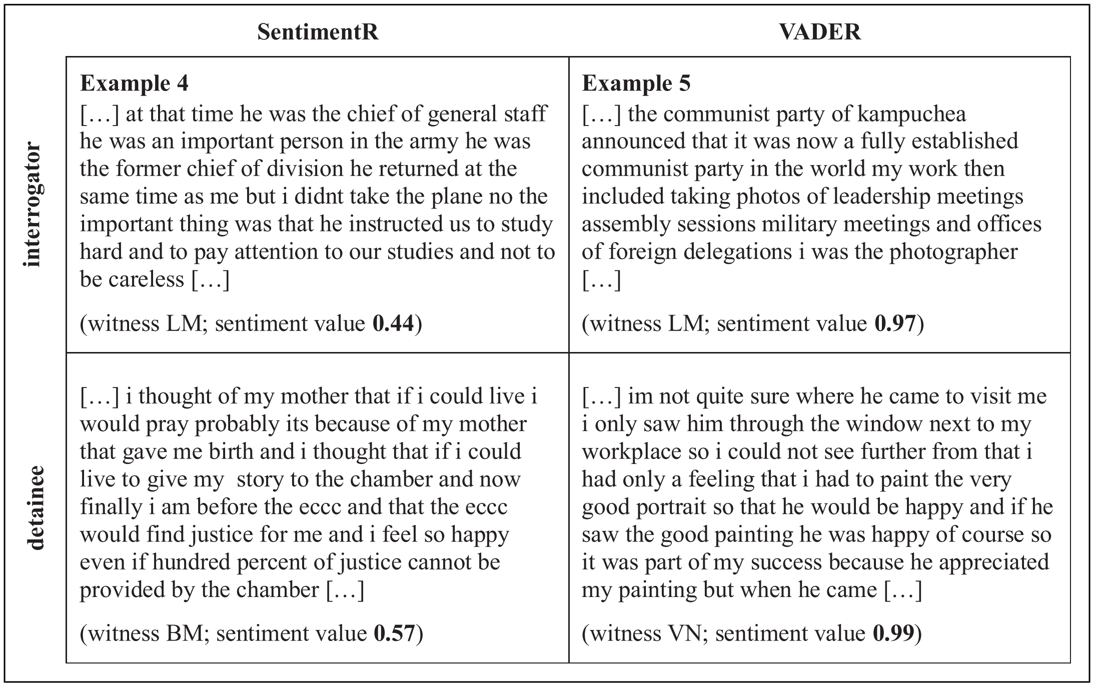
Figure 3. Highest negative sentiment values. *Note.* Highest negative sentiment values for SentimentR and VADER analyses for both groups of interrogators (first row) and detainees (second row). Each text is an excerpt of a paragraph that has been assigned a sentiment value (bold). Witness initials and the sentiment value for the statement are depicted below each text block. The most negative detainee text segment was identical for both the SentimentR and VADER analysis (Example 3). The text has been preprocessed, that is, punctuation and capitalization have been removed.

taught is an example of this category. Finally, motivation to testify refers to statements that reveal why the witnesses decided to testify in court, such as seeking justice or wanting to contribute to the historical record. Subcodes include references to justice and personal feelings toward the tribunal itself.

Emotional Distress. The first important factor deals with the emotional stress during the testimony, which was visible during the former detainees' accounts. Each of the former detainees described their own experience of being tortured, from being beaten by arrival or specific torture during interrogation. They all describe being tortured with their hands tied and by being beaten with bamboo or rattan sticks (BM, 41.1, p. 29; CM, 40.1, p. 13)[5] or being electrocuted regularly (BM, 41.1, p. 30; CM, 40.1, p. 73). Testifying about past experiences of torture in court can be emotionally challenging (Ciorciari & Heindel, 2016), as seen in transcripts where the presiding judge reprimands BM to "control [his] emotion" and "please recompose [him]self" (NN, 41.1, p. 94) while discussing the psychological effects of torture. Similar text passages can be found during the other detainees' testimonies. Also, accounts of experienced torture were described vividly, including descriptions of the emotions felt at that time:

> He asked me to count the lashes, and when I counted up to 10 lashes, he said, "How come you count to 10 lashes? I only beat you for one lash." I felt so painful at the time. There were wounds many wounds on my back and the blood was on the floor flowing from my back. Whips were also used to torture me. (BM, 41.1, p. 13)

Such experiences belonged to the everyday life of detainees at S-21, making both violence and the threat of violence omnipresent and inescapable for them. The detainees further report having



	SentimentR	VADER
interrogator	**Example 4** […] at that time he was the chief of general staff he was an important person in the army he was the former chief of division he returned at the same time as me but i didnt take the plane no the important thing was that he instructed us to study hard and to pay attention to our studies and not to be careless […] (witness LM; sentiment value **0.44**)	**Example 5** […] the communist party of kampuchea announced that it was now a fully established communist party in the world my work then included taking photos of leadership meetings assembly sessions military meetings and offices of foreign delegations i was the photographer […] (witness LM; sentiment value **0.97**)
detainee	[…] i thought of my mother that if i could live i would pray probably its because of my mother that gave me birth and i thought that if i could live to give my story to the chamber and now finally i am before the eccc and that the eccc would find justice for me and i feel so happy even if hundred percent of justice cannot be provided by the chamber […] (witness BM; sentiment value **0.57**)	[…] im not quite sure where he came to visit me i only saw him through the window next to my workplace so i could not see further from that i had only a feeling that i had to paint the very good portrait so that he would be happy and if he saw the good painting he was happy of course so it was part of my success because he appreciated my painting but when he came […] (witness VN; sentiment value **0.99**)

Figure 4. Highest positive sentiment values. *Note*. Highest positive sentiment values for SentimentR and VADER analyses for both groups of interrogators (first row) and detainees (second row). Each text is an excerpt of a paragraph that has been assigned a sentiment value (bold). Witness initials and the sentiment value for the statement are depicted below each text block. The text has been preprocessed, that is, punctuation and capitalization have been removed.

scars that constantly remind them of their suffering. No similar examples were found in the statements of former interrogators.

Technical Expertise. Former interrogators described the torture rather technically, focusing on specific methods and training. In that context, the witnesses explained how interrogations were accompanied by torture if the detainee did not confess (PK, 53.1, p. 21). For that purpose, the interrogators received special training:

> Regarding the techniques of torture, we were taught how to torture the prisoners and to avoid that the prisoners died, otherwise, the confessions would be broken and we would be punished. And we were trained on how to whip the prisoners with the stick, on how to electrocute, on how to use the plastic bag to suffocate them. (PK, 52.1, p. 17)

According to the testimonies, interrogators were trained by more experienced prison staff while watching them interrogate and apply torture techniques, such as electrocution and beating with sticks. Only those who proved to be successful in training were later allowed to use torture on detainees themselves (LM, 57.1, p. 51, p. 88). Notably, former perpetrators' use of collective pronouns like "we" when discussing their training and torture techniques creates a technical, impersonal tone. In contrast, former detainees speak for themselves individually, sharing personal experiences. The witnesses also describe the killing of detainees. One of the former interrogators explains that "the executioners were instructed to kill the prisoners by asking the[m] to kneel down near the rim of the pits" (HH, 50.1, p. 68). Subsequently, the prison staff "would use an oxcart axle to strike the back of the necks and later on they would use a knife to slash the throat, […] then they would untie or remove the cuff and remove the clothes" (HH, 50.1, p. 68). The description of how detainees were killed at S-21 demonstrates clearly that violence is stated in a more factual manner by former interrogators, contrasting the more emotional way in which former detainees recounted their experiences with torture and violence.

Motivation for Testimony. Another factor differentiating testimonies of former detainees and interrogators is closely connected with this observation: the motivation behind the testimony. While former detainees expressed their wish to contribute to justice through their testimony, interrogators might fear being charged for having applied torture. When asked about details on specific torture techniques, the former interrogator HH frequently answered that he "prefer[s] not to answer that question" (e.g., HH, 50.1, p. 68). Former detainees, however, replied openly and in detail, hoping their accounts could contribute to finding truth and justice, as one witness explains:

> What I want is something that is intangible, that is, justice for those that already died. Whatever way the justice could be done is my only hope that can be achieved by this Chamber. And I hope by the end of the Tribunal that justice can be tangible, can be seen by everybody, and that it is something that I expect as a result (VN, 39.1, pp. 55–56).

Following this perspective, former prisoners report to be "happy" testifying in court, "even if 100% of justice cannot be provided by the Chamber" (BM, 41.1, p. 14). Another witness reported being relieved to finally speak about the suffering in front of a court since he "wanted to get it out of [his] chest" (CM, 40.1, pp. 66–67).

The study's qualitative section provides more content-based insights into witness accounts and uncovers differences between both witness groups. Looking at differences in thematic patterns in statements of former detainees and interrogators (*research question 1.3*), our findings indicate that former interrogators provide more detailed information about procedures. In contrast, former detainees primarily focus on recounting their personal experiences, highlighting disparities in emotional distress, technical expertise, and motivation to testify.

Combining the Results: Differences in Testimonies of Torture

In Phase I of the study, we used the language model BERT to classify witness statements as either belonging to former detainees or interrogators. The findings demonstrate the model's ability to effectively differentiate between these two groups based on language. However, the exact differences remain unclear in this phase. Therefore, drawing on the NLP results as a first confirmation of existing differences, we employed SA, focusing on emotional intricacies in the context of torture-related witness statements. The absence of significant differences in sentiment values between both witness groups can be attributed to several factors: First, detainees used strongly negatively weighted words when reporting their torture (e.g., "torture" and "painful"), while interrogators focused more on general procedures associated with less negatively connotated words. Second, the different roles and knowledge of the witnesses in S-21 can be cited as an explanation for differences in their choice of words and their function in the trial. The fact that the court context might play an important role is substantiated by the similarity in sentiment values within witness groups. Although both groups are questioned for reliable information in different areas, the strictly structured legal framework might limit the variety of statements, potentially neutralizing differences in sentiment values (Chlevickaitė et al., 2020). Examples showed that the most negative statements were made by interrogators describing torture, while detainees also expressed hope for justice, further explaining why, descriptively, interrogators' sentiment values were more negative. Phase II of this study thus dives one step deeper into identifying differences in witness statements through SA, complementing the results of Phase I. Finally, building on the previous stages, Phase III uncovers emotional and motivational factors that distinguish the witnesses' statements through its qualitative approach.

This study demonstrates that using a portfolio of NLP and qualitative methods enhances the comprehensive understanding of differences in witness testimonies between former detainees and interrogators, as opposed to using these methods individually: Applying an NLP classification task merely confirms that there is an algorithm-based way to differentiate witness statements, which serves as a justification for investigating those differences further. If we had solely relied on SA, we would not have identified significant differences between the two groups, thereby missing out on non-sentiment-based distinctions. The SA on its own provides us with information on which text segments were the most positive and negative regarding their sentiment value—shedding more light on the relation of sentiment values and witness group. On the other hand, we identify differences by exclusively using QCA without quantitative support but cannot substantiate them on a broader scale. The results challenge each other by providing different insights, particularly when the SA does not show statistically significant differences. This illustrates that NLP classification and QCA prove to be suitable methods to detect differences in testimonies of former detainees and interrogators, while SA might not be an ideal fit (*research question 1*). Still, the study highlights how the three applied methods can complement each other by compensating for their limitations (*research question 2*; see further Section *Contributions to the field of mixed methods*).

Discussion

In a mixed methods exploratory sequential design, this study demonstrated how mixed methods can contribute to analyzing witness statements to find out if and how accounts of torture differ between former detainees and former interrogators in the S-21. We show how using an NLP classification task can serve as a valuable tool to support further analyzing steps and provide an exemplary 3-stage framework for comprehensively analyzing text data from court documents.

Contributions to the Field of Mixed Methods

Addressing the Integration Challenge. This study approaches the integration challenge (Bryman, 2007; Fetters et al., 2013) on different levels. Regarding the research design, this study establishes connections between all three phases. For that purpose, the NLP classification task (Phase I) was set up to approach differentiating between the two witness groups on a speech-based level. The high accuracy of the algorithm in classifying text segments into either former detainees or interrogators served as a baseline for the subsequent phases (Phases II and III), laying the ground for subsequent analyses. The high accuracy score further suggests that the statements from both witness groups exhibit distinct content and narratives, indicating that they can be treated as separate text corpora. This validates the significance of individually analyzing the differences in testimonies between the two groups. Since BERT alone could not provide insights into the content of the testimonies, the study incorporated both SA and QCA to illuminate aspects that could not have been identified by a single approach, allowing for a more thorough understanding of the differences between the witness groups and how contextual factors may have influenced their testimonies. This goes hand in hand with integration on a methodological level, as each method served to inform and justify the concrete selection of the subsequent method in the analysis. By leveraging the insights gained from each method, we were able to make informed decisions regarding the subsequent stages of our analysis. Finally, on an interpretation level, results were integrated following a contiguous narrative approach, reporting findings of each phase separately at first and bringing everything together in the end (Fetters et al., 2013). In this case, the NLP-based classification of text paragraphs could not be replicated through SA, leading to discordance as one of the possible outcomes of MMR (Fetters et al., 2013; Moseholm & Fetters, 2017). The insights provided by the examples in Figures 3 and 4, combined with three different SA techniques, support the validation of the study's results. The qualitative analysis further elaborates on these examples and provides a broader illustration of the study's findings—in this case, complementing the other phases by bringing personal motivations and attitudes of the witnesses into light that could not have been discovered by quantitative approaches alone. That the qualitative findings are supported by the NLP classification and descriptive trends in the SA further underscores the importance of this methodological combination.

Data Transformation and NLP for MMR. This study further highlights opportunities from data transformation, where one type of data is converted into another (Fetters et al., 2013). *Quantitizing* text (Sandelowski et al., 2009) enabled the use of witness transcripts across all three phases of the study. By creating a dataset out of court transcripts amenable for further NLP-based processing, we went beyond classic content analysis (Krippendorff, 2004), where codes of qualitative analyses are mainly counted. Instead, we created a data source that kept most of the original context by separating the original documents into smaller text segments.

Applying NLP techniques to the transformed data confirms some of the advantages of using NLP to "accelerate" MMR (Chang et al., 2021). The NLP classification conducted in this study highlights the usefulness of establishing that differences between the testimonies of former interrogators and detainees exist. This can serve as a potential validation strategy (Chang et al., 2021; Crowston et al., 2012) for the use of further models, enabling researchers to focus on more specific characteristics that set the groups apart, such as distinct emotions or language patterns. Furthermore, our results show the potential to automatically classify witnesses into different groups with similar narratives, allowing for a more accurate individual analysis that uncovers characteristics that are unique to that group and would otherwise have remained hidden. The absence of differences revealed by the SA does not necessarily diminish the value of NLP. In fact, the sentiment values obtained from the analysis offer insights into which paragraphs were associated with more negative or positive sentiment values, allowing for comparison on an individual level and adding a quantitative perspective (Guetterman et al., 2018). Further, results indicate that any differences that may exist are not reflected significantly on a sentiment level, leaving space for the qualitative component to explore other possible sources of distinction. In this study, we selected SA due to the emotional nature of torture-related statements. However, it is important to note that SA is just one of many NLP methods that can be used to identify these differences. For future analyses, techniques such as topic modeling or named entity recognition could also be employed.

Limitations and Future Research

Interpreting the results, some limitations should be acknowledged. First, it must be addressed that the trial transcripts were translated from the original Khmer language into English for analysis. Since the SAs are based on comparing two groups whose statements have been subjected to translation in the same manner, the translation should not have a relevant impact on the analysis. It can be assumed that the translation corresponds to official standards and was performed by professional court translators, ensuring a certain level of translation quality. Hence, comparing the SA results between the two witness groups of former detainees and interrogators can still be meaningful, as any potential translation biases or inaccuracies would likely affect both groups similarly. However, considering that the language itself underwent analysis, it cannot be ruled out that disparities in the original languages were leveled out or amplified during translation, potentially explaining why sentiment values did not differ significantly.

Second, with its highly structured Q&A procedure, the court context may have imposed limits on the variety of statements and subjects, which could have neutralized potential differences in sentiment values. Therefore, exploring the relationship between specific questions and answers, for example, on the emotional impact of the torture for both witness groups, might provide further insights into the differences in language use between the two witness groups.

Third, the SAs have certain methodological limitations. Since AFINN attributes sentiment values per word, contextual information is not included. Further, words that are not categorizable as positive or negative are excluded. This can introduce biases, as certain words like “beat” may be excluded while variations like “beaten” and “beating” have different ratings. While SentimenR incorporates context by assigning values to amplifiers and negations, its scope is limited to direct word neighbors and may not capture relevant context. In contrast, VADER includes positive, negative, and neutral sentence components, making it the model that best incorporates overall context. Finally, for all SAs, it should be noted that while the number of words and sentences was relatively high, the sample size was restricted to 8 witnesses, thus limiting the statistical power of the SAs. On a broader level, it is important to distinguish between trauma and sentiment as separate concepts. While SA can provide an understanding of the general sentiment expressed in witness testimonies, it is crucial to recognize that trauma encompasses a broader range of experiences and has a distinct psychological impact. Therefore, while SA can be informative, it may not capture the full depth and complexity of traumatic experiences and their associated psychological aspects.

Regarding the qualitative analysis, the subjective categorization and selection of text segments should be mentioned as a possible point of critique. However, considering that the SAs include context only in a limited and numerical way, an additional qualitative framework is even more critical. Especially when analyzing sensitive material, such as court testimonies of detainees and interrogators of a torture prison, maintaining the original context cannot be given too much weight.

That differences in the statements of former detainees and interrogators could not be replicated through SA suggests that lexicon-based SAs might not be the best-fitting approach to address our research questions. However, within this 3-stage process, SA still offered valuable quantitative insights, and the absence of significant differences in sentiment values itself is a noteworthy finding. Ultimately, the methodological framework applied in this study is meant to serve as an example of how to combine NLP, statistical, and qualitative methods—its application in practice should be adapted based on the specific research question and the nature of the data.

Conclusion

This paper presents a mixed methods approach for analyzing court transcripts of genocide tribunals, with a specific focus on accounts of experienced torture in different witness groups. By integrating computational and qualitative methods, this study offers new perspectives on analyzing torture-related content in court transcripts. It is worth noting that this design is not exclusively tailored to genocide research but can be applicable to other fields that also seek to examine variations in emotional content within textual data.

The differences found in the last phase of this study provide a starting point for further analysis and the application of a different set of methods for finding out more about emotionally distressing accounts of witnesses in court. Providing a novel approach by bringing together QCA and NLP techniques in an MMR framework, this study leads the way to further research, encouraging the use of mixed methods in genocide research and building bridges between historical analyses and computational methods.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Miriam Schirmer https://orcid.org/0000-0002-6593-3974

Notes

1. https://www.eccc.gov.kh/en/case/topic/90
2. In a job posting for an ECCC interpreter, the court asks for a "minimum of 5 years of experience in interpretation and/or translation, preferably including 3 years of interpretation and/or translation in an international organization or an international body dealing with legal matters" (Extraordinary Chambers in the Courts of Cambodia, 2015).
3. All code used is accessible online at https://osf.io/fnjvt/?view_only=b63c4c1c01364b2285faf045b6f75f77
4. For training, a batch-size of 8 with 4 epochs was used with a train/test/evaluate split of 60/20/20% on the BERT-base-uncased model. Precision refers to the ratio of correctly classified positive samples to all positively predicted samples, whereas recall describes the ratio of correctly classified positive samples to all samples with the specific label.
5. Citations include the witness's initials, document, and page number of the respective statement in the transcript.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Author. https://doi.org/10.1176/appi.books.9780890425596

Bacchus, L. J., Buller, A. M., Ferrari, G., Brzank, P., & Feder, G. (2018). "It's always good to ask": A mixed methods study on the perceived role of sexual health practitioners asking gay and bisexual men about experiences of domestic violence and abuse. *Journal of Mixed Methods Research*, *12*(2), 221–243. https://doi.org/10.1177/1558689816651808

Barnes-Ceeney, K., Gideon, L., Leitch, L., & Yasuhara, K. (2019). Recovery after genocide: Understanding the dimensions of recovery capital among incarcerated genocide perpetrators in Rwanda. *Frontiers in Psychology*, *10*, 637. https://doi:10.3389/fpsyg.2019.00637

Bazeley, P. (2018). *Integrating analyses in mixed methods research*. Sage.

Békés, V., Perry, J. C., & Starrs, C. J. (2021). Coping action patterns in trauma and other autobiographic narratives in Holocaust survivors: A mixed–methods study. *Journal of Aggression, Maltreatment & Trauma*, *30*(10), 1307–1326. https://doi.org/10.1080/10926771.2020.1853296

Boeije, H., Slagt, M., & van Wesel, F. (2013). The contribution of mixed methods research to the field of childhood trauma: A narrative review focused on data integration. *Journal of Mixed Methods Research*, *7*(4), 347–369. https://doi.org/10.1177/1558689813482756

Brönnimann, R., Herlihy, J., Müller, J., & Ehlert, U. (2013). Do testimonies of traumatic events differ depending on the interviewer? *The European Journal of Psychology Applied to Legal Context*, *5*(1), 97–121. https://journals.copmadrid.org/ejpalc/art/8db9264228dc48fbf47535e888c02ae0#resumen

Brounéus, K. (2008). Truth-telling as talking cure? Insecurity and retraumatization in the Rwandan Gacaca courts. *Security Dialogue*, *39*(1), 55–76. https://doi.org/10.1177%2F0967010607086823

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research*, *1*(1), 8–22. https://doi.org/10.1177/2345678906290531

Chandler, D. (1999). *Voices from S-21: Terror and history in Pol Pot's secret prison*. University of California Press.

Chang, T., DeJonckheere, M., Vydiswaran, V. G. V., Li, J., Buis, L. R., & Guetterman, T. C. (2021). Accelerating mixed methods research with natural language processing of big text data. *Journal of Mixed Methods Research*, *15*(3), 398–412. https://doi.org/10.1177/15586898211021196

Chlevickaitė, G., Holá, B., & Bijleveld, C. (2020). Judicial witness assessments at the ICTY, ICTR and ICC: Is there 'standard practice' in International criminal justice? *Journal of International Criminal Justice*, *18*(1), 185–210. https://doi.org/10.1093/jicj/mqaa002

Ciorciari, J. D., & Heindel, A. (2016). Trauma in the courtroom. In B. van Schaack & D. Reicherter (Eds.), *Cambodia's hidden scars: Trauma psychology and the Extraordinary Chambers in the Courts of Cambodia* (2nd ed., pp. 159–189). Documentation Center of Cambodia.

Colditz, J. B., Welling, J., Smith, N. A., James, A. E., & Primack, B. A. (2019). World Vaping Day: Contextualizing vaping culture in online social media using a mixed methods approach. *Journal of Mixed Methods Research*, *13*(2), 196–215. https://doi.org/10.1177/1558689817702753

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Creswell, J. W., & Zhang, W. (2009). The application of mixed methods designs to trauma research. *Journal of Traumatic Stress*, *22*(6), 612–621. https://doi.org/10.1002/jts.20479

Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, *15*(6), 523–543. https://doi.org/10.1080/13645579.2011.625764

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Association for Computational Linguistics (Ed.), *Proceedings of the 2019 Conference of the North American Chapter of the association for computational linguistics* (pp. 4171–4186). ACL. http://doi.org/10.18653/v1/N19-1423

Driscoll, D. L., Appiah-Yeboah, A., Salib, P., & Rupert, D. J. (2007). Merging qualitative and quantitative data in mixed methods research: How to and why not. *Ecological and Environmental Anthropology*, *3*(1), 19–28. https://digitalcommons.unl.edu/icwdmeea/18

Extraordinary Chambers in the Courts of Cambodia. (2010). *Judgment (Kaing Guek Eav alias Duch)*. Case File/Dossier No. 001/18-07-2007/ECCC/TC, 26 July 2010 [Court document]. https://www.eccc.gov.kh/en/document/court/judge-ment-case-001

Extraordinary Chambers of the Courts of Cambodia. (2015, November). *Interpreter* [Job posting]. https://www.eccc.gov.kh/en/jobs/interpreter

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs. Principles and practice. *Health Services Research*, *48*(6), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research*, *20*(6), Article e231. https://doi.org/10.2196/jmir.9702

Hinton, A. L. (2016). *Man or monster? The trial of a Khmer Rouge torturer.* Duke University Press.

Ho, P., Chen, K., Shao, A., Bao, L., Ai, A., Tarfa, A., Brossard, D., Brown, L., & Brauer, M. (2021). A mixed methods study of public perception of social distancing: Integrating qualitative and computational analyses for text data. *Journal of Mixed Methods Research*, *15*(3), 374–397. https://doi.org/10.1177/15586898211020862

Holness, T., & Ramji-Nogales, J. (2016). Participation as reparations: The ECCC and healing in Cambodia. In B. van Schaack & D. Reicherter (Eds.), *Cambodia's hidden scars: Trauma psychology and the Extraordinary Chambers in the Courts of Cambodia* (2nd ed., pp. 213–234). Documentation Center of Cambodia.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

International Criminal Court. (n.d.). *Witnesses*. https://www.icc-cpi.int/about/witnesses

Jurafsky, D., & Martin, J. (2021). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/

Kaiser, J., & Hagan, J. (2015). Gendered genocide: The socially destructive process of genocidal rape, killing, and displacement in Darfur. *Law & Society Review*, *49*(1), 69–107. http://doi.org/10.2139/ssrn.2250867

Kanavou, A. A., & Path, K. (2017). The lingering effects of thought reform: The Khmer Rouge S-21 prison personnel. *The Journal of Asian Studies*, *76*(1), 87–105. https://doi.org/10.1017/S0021911816001625

Kelly, J. T., Betancourt, T. S., Mukwege, D., Lipton, R., & Vanrooyen, M. J. (2011). Experiences of female survivors of sexual violence in Eastern Democratic Republic of the Congo: A mixed-methods study. *Conflict and Health*, *5*(1), 1–8. https://doi.org/10.1186/1752-1505-5-25

Keydar, R. (2020). Listening from Afar: An algorithmic analysis of testimonies from the International Criminal Courts. *Illinois Journal of Law, Technology & Policy*, *1*, 55–83.

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology.* Sage.

Lehrner, A., & Yehuda, R. (2018). Trauma across generations and paths to adaptation and resilience. *Psychological Trauma: Theory, Research, Practice, and Policy*, *10*(1), 22–29. https://doi.org/10.1037/tra0000302

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Mayring, P. (2015). *Qualitative inhaltsanalyse. Grundlagen und techniken [Qualitative content analysis. Foundations and techniques]* (12th ed.). Beltz.

Moseholm, E., & Fetters, M. D. (2017). Conceptual models to guide integration during analysis in convergent mixed methods studies. *Methodological Innovations*, *10*(2), 1–11. https://doi.org/10.1177/2059799117703118

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv Preprint, 1103.2903*, 1–6. https://doi.org/10.48550/arXiv.1103.2903

O'Halloran, K. L., Tan, S., Pham, D.-S., Bateman, J., & Vande Moere, A. (2018). A digital mixed methods research design: Integrating multimodal analysis with data mining and information visualization for big data analytics. *Journal of Mixed Methods Research*, *12*(1), 11–30. https://doi.org/10.1177/1558689816651015

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, *7*, 100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050

Reinhold, A. M., Raile, E. D., Izurieta, C., McEvoy, J., King, H. W., Poole, G. C., Ready, R. C., Bergmann, N. T., & Shanahan, E. A. (2022). Persuasion with precision: Using natural language processing to improve instrument fidelity for risk communication experimental treatments. *Journal of Mixed Methods Research*, *17*(4), 373–395. https://doi.org/10.1177/15586898221096934

Rinker, T. W. (2019). *sentimentr: Calculate text polarity sentiment*. http://github.com/trinker/sentimentr

Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research*, *3*(3), 208–222. https://doi.org/10.1177/1558689809334210

Sandick, P. A. (2012). Speechlessness and trauma: Why the International Criminal Court needs a public interviewing guide. *Northwestern Journal of International Human Rights*, *11*(1), 104–125.

Sawalha, J., Yousefnezhad, M., Shah, Z., Brown, M. R. G., Greenshaw, A. J., & Greiner, R. (2022). Detecting presence of PTSD using sentiment sentiment analysis from text data. *Frontiers in Psychiatry*, *12*, 811392. https://doi.org/10.3389/fpsyt.2021.811392

Schirmer, M., Kruschwitz, U., & Donabauer, G. (2022). A new dataset for topic-based paragraph classification in genocide-related court transcripts. In Proceedings of the Language Resources and Evaluation

Conference (LREC), Marseille, France, 2022, pp. 4504–4512. European Language Resources Association. https://aclanthology.org/2022.lrec-1.479

Schirmer, M., Nolasco, I. M. O., Mosca, E., Xu, S., & Pfeffer, J. (2023). Uncovering trauma in genocide tribunals: An NLP approach using the Genocide Transcript Corpus. In Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL), Braga, Portugal, 19-23 June, 2023, pp. 257–266. https://doi.org/10.1145/3594536.3595147

Sharratt, S. (2016). *Gender, shame and sexual violence: The voices of witnesses and court members at war crimes tribunals*. Routledge.

Sripathi, K. N., Moscarella, R. A., Steele, M., Yoho, R., You, H., Prevost, L. B., Urban-Lurain, M., Merrill, J., & Haudek, K. C. (2023). Machine learning mixed methods text analysis: An illustration from automated scoring models of student writing in biology education. *Journal of Mixed Methods Research*. https://doi.org/10.1177/15586898231153946

Tashakkori, A., Johnson, R. B., & Teddlie, C. (2021). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioural sciences* (2nd ed.). Sage.

Thaler, K. M. (2017). Mixed methods research in the study of political and social violence and conflict. *Journal of Mixed Methods Research*, *11*(1), 59–76. https://doi.org/10.1177/1558689815585196

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2021). Revisiting few-sample BERT fine-tuning. In ICLR 2021 – Ninth International Conference on Learning Representations. Vienna, Austria, 4–8 May 2021. https://doi.org/10.48550/arXiv.2006.05987

4.2 Study 2: A New Dataset for Topic-Based Paragraph Classification in Genocide- Related Court Transcripts

This publication is RELEVANT TO THE EXAMINATION.

Authors

Miriam Schirmer, Udo Kruschwitz, Gregor Donabauer

In

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), 4504–4512, peer-reviewed, `http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.479.pdf`), published 06-2022

Abstract

Recent progress in natural language processing has been impressive in many different areas with transformer-based approaches setting new benchmarks for a wide range of applications. This development has also lowered the barriers for people outside the NLP community to tap into the tools and resources applied to a variety of domain-specific applications. The bottleneck however still remains the lack of annotated gold-standard collections as soon as one's research or professional interest falls outside the scope of what is readily available. One such area is genocide-related research (also including the work of experts who have a professional interest in accessing, exploring and searching large-scale document collections on the topic, such as lawyers). We present GTC (Genocide Transcript Corpus), the first annotated corpus of genocide-related court transcripts which serves three purposes: (1) to provide a first reference corpus for the community, (2) to establish benchmark performances (using state-of-the-art transformer-based approaches) for the new classification task of paragraph identification of violence-related witness statements, (3) to explore first steps towards transfer learning within the domain. We consider our contribution to be addressing in particular this year's hot topic on Language Technology for All.

Contribution of Thesis Author

Theoretical conceptualization, data curation, methodological design, formal analysis, visualization, manuscript writing, revision, and editing.

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4504–4512
Marseille, 20-25 June 2022

A New Dataset for Topic-Based Paragraph Classification in Genocide-Related Court Transcripts

Miriam Schirmer[†,‡], Udo Kruschwitz[†], Gregor Donabauer[†]
[†]University of Regensburg, Germany
[‡]TUM School of Social Science and Technology, Technical University of Munich, Germany
{miriam.schirmer, udo.kruschwitz, gregor.donabauer}@ur.de

Abstract

Recent progress in natural language processing has been impressive in many different areas with transformer-based approaches setting new benchmarks for a wide range of applications. This development has also lowered the barriers for people outside the NLP community to tap into the tools and resources applied to a variety of domain-specific applications. The bottleneck however still remains the lack of annotated gold-standard collections as soon as one's research or professional interest falls outside the scope of what is readily available. One such area is genocide-related research (also including the work of experts who have a professional interest in accessing, exploring and searching large-scale document collections on the topic, such as lawyers). We present GTC (Genocide Transcript Corpus), the first annotated corpus of genocide-related court transcripts which serves three purposes: (1) to provide a first reference corpus for the community, (2) to establish benchmark performances (using state-of-the-art transformer-based approaches) for the new classification task of paragraph identification of violence-related witness statements, (3) to explore first steps towards transfer learning within the domain. We consider our contribution to be addressing in particular this year's hot topic on Language Technology for All.

Keywords: Text classification, BERT, Professional Search, Genocide Studies, Language Technology for All

1. Introduction

Information overload has led to a multitude of search applications of which Web search is just one out of many. Unlike search for leisure or personal interest there is a vast area of search contexts which are found in a work environment. Professional search falls into that scope, i.e. search over domain-specific document collections and often with search tasks that are recall-oriented rather than precision-focused (Kruschwitz and Hull, 2017; Verberne et al., 2019). Beyond applications where such search effort can directly be measured in financial terms (e.g. in patent search, e-discovery or the compilation of systematic reviews) there are many other fields where these costs are more implicit, e.g. in the area of genocide studies that rely on the analysis of vast quantities of different resources (Bachman, 2020; Hinton, 2012).

Looking at the wider picture, searching large text corpora for specific thematic patterns can be very time-consuming and non-trivial, in particular for searchers who do not have a solid foundation in NLP or search technology. The huge amount of court transcripts of genocide tribunals presents a perfect example: the *International Criminal Tribunal of the Former Yugoslavia (ICTY)* alone provides official transcripts for each of its cases online, leading up to approximately 2.5 million pages of transcripts in total (ICTY, 2016). Searching for specific content in a text corpus like this usually requires vast amounts of manual research capacity (Hoang and Schneider, 2018). Tools and approaches to augment this type of search and help limit manual efforts have been developed for a broad range of use cases, e.g. for automating search strategies or text extraction from documents (MacFarlane et al., 2021; Russell-Rose et al., 2021). However, even with the help of suitable tools, searching for specific text passages in large text corpora generally remains a difficult task, in particular when the search is recall-oriented (Bache, 2011; Kaptein et al., 2013; Noor and Bashir, 2015). It should also be pointed out that in many use cases in which experts have to sift through large amounts of textual data a *fully automated* analysis might neither be achievable nor desirable and the provision of support tools that *assist* the expert are the preferred option. The area of fact-checking is one such application context (Nakov et al., 2021).

Turning to our own use case, the search for specific content in transcripts of genocide tribunals further proves difficult because transcripts are only accessible individually (usually one court day per transcript) and in different formats, depending on the tribunal. So far, no datasets of any kind containing genocide court transcripts have been published. Similarly, no other forms of pre-structured or annotated text data in this field of research exist.

This paper addresses this gap by providing a systematically annotated dataset containing text material from three different genocide tribunals: the *Extraordinary Chambers in the Courts of Cambodia (ECCC)*, the *International Criminal Tribunal for Rwanda (ICTR)*, and the *ICTY*. In addition to compiling the sampled corpus, we provide annotations within the text. More specifically, text passages in which witnesses talk about experienced violence have been annotated, focusing on a core part of each testimony. Given that respective passages on violence often cover crimes that are relevant for the indictment of the accused, such as murder or rape, they are essential for judgement. At the same time, they are not easily classifiable due to their potential ambiguity.

Since this dataset of textual documents is the first in this area, we hope that it provides a valuable resource for NLP-based genocide research. To foster generalisability and assess transferability of approaches, the corpus contains a sample from different tribunals.

We consider the provision of the new resource our main contribution, but we also provide experimental work that will serve as a benchmark and allow the contextualisation within the broader field. We use fine-tuned BERT models for that purpose. Utilizing the heterogeneous nature of the corpus we also explore transfer learning and report results.

Beyond the contribution to the NLP community, it is our hope that the results of this paper will be useful for both scholars and practitioners at international criminal tribunals who need to work through large quantities of transcript material as part of their everyday job.

We summarize our contributions as follows:

1. We present GTC, a new reference corpus sampled from different international criminal courts in the context of genocide tribunals. The corpus contains annotations of statements by witnesses about experienced violence.

2. We built state-of-the-art transformer-based classifiers to provide benchmarks for the new classification task of paragraph identification of violence-related witness statements.

3. We provide experimental results for transfer-learning by varying the training and testing data across documents from different tribunals.

4. We make all data as well as code available to the community.[1]

2. Related Work

We touch on the three key areas of interest our work falls into, namely resources, professional search, and text classification. The discussion of each of these should simply serve as both a motivation and basic context.

2.1. Resources

The importance of publicly available language resources to help develop NLP applications has long been recognized, e.g. Calzolari et al. (2010), and the domain-specific nature of many problems is what makes respositories such as the *LRE Map*[2] a valuable starting point for many researchers and practitioners.

For the specific use case of assisting searchers to identify relevant information in genocide-related court transcripts resources are very limited to non-existent. Of course, transcripts of each tribunal are available through the respective courts' websites – however, their quality in terms of digitisation (e.g., object character recognition) varies greatly. Specifically for the ICTY, Fidahić (2021) further criticises that transcripts are only available in certain languages, thus limiting access mainly to English-speaking readers. Considering that the field of genocide research and studies (Totten, 2017) is multi-faceted enough to warrant the provision of suitable resources, we see our own contribution as a starting point to fill this gap, even though we limit our work to English transcriptions.

2.2. Professional Search

Searching through court transcripts can often be framed as an instance of professional search (Koster et al., 2009; Russell-Rose et al., 2021). Professional search describes the process of searching for information in a work context which is commonly domain-specific and requires expertise in a specific area. Key features of professional search are limited time and budget resources, making it desirable to provide support that helps classify specific text passages which ultimately could drastically reduce search efforts (Russell-Rose et al., 2021). It should be noted that professional search is very different from other types of search such as Web search. A common observation is that searches take a lot longer to satisfy a specific information need. For example, Bullers et al. (2018) found that librarians spend 26.9 hours on average on systematic reviews that involve searching for specific content, indicating that this task is highly time-consuming. Similarly, Greene-Colozzi et al. (2021) discuss the time-consuming process of researching court transcripts and other relevant sources related to cybercrime. Professional search in court transcripts in general, however, has not been analysed so far.

Another important aspect dealing with extensive search in large text corpora are human factors. Especially when dealing with time-consuming search in text documents that lasts for hours, fatigue might be an issue that reduces the quality of the search. Additionally, manual search is also more vulnerable to subjectivity, motivating the use of automated search algorithms (Li et al., 2020).

In the context of professional and augmented search, different supported search scenarios could also be helpful as a first step. Especially when working with an annotated dataset that is built around a binary classification task, the classification labels provided can help to significantly narrow down the text material for further in-depth search.

Different tools and algorithms to save time in searching through text have been discussed – varying strongly depending on the specific search context. For example,

[1] https://github.com/MiriamSchirmer/genocide-transcript-corpus.

[2] https://lremap.elra.info

MacFarlane et al. (2021) give a broad overview of different tools for systematic literature reviews, such as tools for text and data extraction or automatic query expansion. While these tools might help review literature more efficiently, the authors note that their use is not widespread. Furthermore, not all of the abovementioned tools are helpful when it comes to content-based search in text documents. In this context – when looking for specific content in court transcripts – tools for enhanced keyword search might prove more useful.

2.3. Text Classification

Topic-based paragraph classification specifically for court transcripts has so far not been discussed in the literature. Nevertheless, extensive research in this area has been done in other fields. As a traditional and fundamental NLP task, text classification covers a wide range of tasks ranging from category labeling over sentiment analysis to authorship attribution (Jurafsky and Martin, 2021). Traditionally effective approaches to supervised machine learning, such as Support Vector Machines (Al Amrani et al., 2018; Tong and Koller, 2001; Zhang et al., 2007) or K-nearest neighbour (Bijalwan et al., 2014), have now largely been replaced by transformer-based approaches (Dhar et al., 2021; Jurafsky and Martin, 2021; Minaee et al., 2021). Fine-tuning BERT has become the standard baseline in text classification (Devlin et al., 2019), not just beating traditional machine learning paradigms but also recurrent neural networks (RNNs), convolutional neural networks (CNNs) or other deep neural networks (DNNs) (Li et al., 2020).

Example topic areas in which BERT has been utilized effectively in text classification include various forms of sentiment analysis ranging from aspect-based sentiment analysis (Sun et al., 2019) to sentiment analysis on the impact of coronavirus in social life (Singh et al., 2021), as well as reading comprehension tasks, e.g., Xu et al. (2019).

Of specific concern to our underlying use case is text classification that requires *text segment classification*, commonly found when applied to social media data, such as tweets and comments. For this type of analysis, splitting larger text data into paragraphs limited to a certain number of words has been established as a regular step in the NLP pipeline (Li et al., 2020). A very prominent example of using BERT sequence classification is hate speech detection (e.g., Mozafari et al. (2020a), Mozafari et al. (2020b), Sohn and Lee (2019)). By applying BERT to Twitter data, tweets can easily be classified according to whether they contained racism, sexism, or hate, among others (Mozafari et al., 2020a).

New BERT models and applications are being reported at rapid speed as the model is continuously applied in new fields. Examples are the recently developed ClimateBert, a pre-trained language model for climate-related text (Webersinke et al., 2021) or COVID-Twitter-BERT (CT-BERT) (Müller et al., 2020).

2.4. Concluding Remarks

The new corpus we provide aims to bridge a (domain-specific) gap that exists in the landscape of annotated text collections. In order to assess the utility of the corpus and the difficulty of the underlying classification task we will adopt the commonly applied baseline approach of fine-tuning BERT. One of the goals is to show whether or not BERT also serves as an efficient tool for this type of text data and whether it can help simplify classification of paragraphs in court data.

This can only be a first step at filling the identified gap – there will be scope for many future directions, not least to replicate the approach to other languages.

3. Genocide Transcript Corpus (GTC)

We introduce *Genocide Transcript Corpus (GTC)*, a corpus of transcripts drawn from the court proceedings of international tribunals dealing with cases of genocide. Following sampling of the original data we also apply an annotation step that assigns binary labels to individual paragraphs. The paragraph labeling is aimed at identifying those parts of the text that refer to violence experienced by witnesses – relevant are only those text segments which are actually part of witness statements.

The dataset used in this study consists of 1475 text passages from three different genocide tribunals. Transcripts from the three biggest ad-hoc genocide tribunals, the ECCC, the ICTR, and the ICTY were selected. In a first step, the courts' databases were searched for witnesses who have actually experienced some form of violence. This pre-selection ensured having a substantial amount of relevant text passages in the dataset and thereby excluding technical or expert witnesses. Three different tribunals were selected to provide a diverse dataset and explore transfer learning, i.e. to show possible differences in the results after training and testing with data from different tribunals. Thus, results are more generalisable and differences in individual tribunals are controlled for.

Between 4 and 7 transcripts were selected per tribunal and were divided into equally large text chunks of 250 words each. Numbers and punctuation were removed in a first preprocessing step. In the final dataset, the number of samples is roughly equally distributed across tribunals (ECCC: 465, ICTY: 530, ICTR: 480). Differences occur since only complete transcripts with varying length (about 40 to 120 pages) were included.

The current version of the GTC contains the following data:

- For the ECCC, transcripts with a total of 438 pages from two different trials (Case 001 against Kaing Guev Eav, Case 002 against Nuon Chea and Khieu Samphan) were selected. This includes the proceedings of 4 full court days and the hearing of 7 witnesses.

- Transcripts of the ICTY were taken from the cases against Slobodan Milošević (IT-02-54) and Duško Tadić (IT-94-1). The material consists of 416 pages of transcripts from 5 trial days, with 15 witnesses testifying in court.
- For the ICTR, 566 pages of transcript material from the cases against Jean-Paul Akayesu (ICTR-96-04) and Pauline Nyiramasuhuko et al. (ICTR-98-42) were included in the dataset. The ICTR data includes 5 witnesses and 7 court days.

In total, 1420 pages of transcripts were incorporated into the dataset. Differences in the number of pages and witnesses are firstly due to different transcript formats regarding digitisation and text density per page. Secondly, legal proceedings vary between the different tribunals and thus lead to slightly different content. For example, in the selected ECCC and ICTR transcripts, witnesses are questioned for approximately one court day, whereas in the selected ICTY transcripts, 2 to 3 witnesses were questioned per day.

4. Methodology

4.1. Label Annotation

All samples were labeled according to whether they contain a witness's description of experienced violence (0 = no violence, 1 = violence). Violence in this context is interpreted broadly and includes accounts of experienced or directly witnessed torture, interrogation, death, beating, psychological violence, experienced military attacks, destruction of villages, looting, and forced displacement. We restrict our interest to a binary classification, i.e. different acts of violence were not categorized further into subcategories. Figure 1 provides an example of a rather clear distinction between the two labels.

Label 0

Q. […] As we discussed before, I will ask you some questions concerning your experiences in Rwanda back in 1994. Back in April of 1994 where did you live? And please you can just specify by commune.
A. We were living in Taba commune.
Q. Is that in Rwanda?
A. It's a commune in Rwanda, in Gitarama prefecture.
Q. Around the beginning of April did you ever receive news of the crash of the president's plane?
A. Yes, I heard this. […]

ICTR-96-4-I, October 23rd 1997, p. 17-18.

Label 1

Q. What happened next?
A. He took me and he had a very long knife that he was wearing in his belt and also a small ax in his hand. We arrived near the primary school. The classrooms are very close to the bureau communal, very close to the place where we were before and it's very close to the road, as well, and when we arrived at that location **this child put down this ax, he also put down the long knife, near me, and you see these things are not very easy to see, a young child like that rape me.** I hope you understand that this is something that is very, very painful. […]

ICTR-96-4-I, October 23rd 1997, p. 60.

Figure 1: Sample abstracts from the corpus demonstrating two clear-cut examples for a text passage that does *not* contain accounts of violence in a witness statement (top example – Label 0) and one that does (bottom example – Label 1). The examples were shortened, and both format and punctuation were adapted for readability.

An important requirement for labeling text passages as containing accounts of violence was whether experienced violence was described by the witness orally in court. Questions by lawyers and judges containing violence-related words were thus labeled '0'. However, since the words used in both cases are the same for the most part, the differentiation between violence-related statements of witnesses vs. lawyers, judges, or the accused makes an automated classification more difficult. Having written statements (e.g., statements recorded previously by court staff, police, or human rights organisations) read out loud during the trial increases this difficulty further: even though reports contain accounts of experienced violence, they are not labeled '1' because they were not expressed orally by the witnesses during the trial, but by a lawyer or another representative of the court (see Figure 2 for an example).

It should have become clear that the task of correct classification in an automated fashion is non-trivial; simple 'bag of words'-based approaches are likely to underperform. Apart from the context that makes it clear how to classify a paragraph, looking at the vocabulary alone will not be sufficient. A similar observation was made when classifying a corpus of tweets which were classed as falling into a number of different classes all to do with violence such as crises, violence, accidents, and crime (Alhelbawy et al., 2016). It was found that the inter-rater agreement varied significantly across the different violence classes.

Since this dataset does not differentiate between subcategories, classification was limited to a binary task. However, to make sure that the categorization is reliable, a random selection of approximately 200 samples were independently labeled by a second researcher (with an inter-rater reliability $\kappa = 0.86$) according to the above-mentioned facets of experienced violence. Even though only a sample of the dataset is labelled by two annotators, the high inter-rater agreement suggests that

> **Label 0 (introduction of witness by lawyer)**
>
> […] The witness is a journalist working for a newspaper and he has reported several materials during the conflict in 1998. […] He describes the situation in Suva Reka on the 25th of March, 1999, including the **killings and burning of houses**. […] The witness also describes that on April 1st, Belanica was shelled, and police, military, and paramilitary forces, numbering about 1.500, subsequently entered the village. **The Serb forces forced people from their houses, looted their homes, loaded the goods on the trucks, and set the houses on fire.**
>
> ICTY, 020424IT, April 24th 2002, p.3361-3362.

Figure 2: Example of a text passage that contains violence-related vocabulary, but is not labeled 1. As in Fig. 1, this example was shortened and adapted.

the labeling process yielded sufficiently plausible results.

Table 1 provides an overview of the number of each label per tribunal. Differences in the label balance are due to the random selection of transcripts.

	n_0	n_1	n_{total}
ALL	946	529	1475
ECCC	286	179	465
ICTY	401	129	530
ICTR	259	221	480

Table 1: Overview of label balance for the complete dataset ("ALL") and the three individual tribunal datasets.

4.2. Experimental Setup

For all experiments, the 12-layer BERT$_{base}$ architecture for sequence classification (Devlin et al., 2019) was used to classify text passages of genocide tribunal transcripts.

As described in Section 3, the dataset consists of 3 subsets with data from different tribunals. 5-fold cross-validation (80:10:10) was applied to each subset and to the full version of the dataset (concatenated subsets).

Overall, BERT$_{base}$ was trained on all possible train, validate and test constellations, leading to a total of 16 different combinations. In those cases, in which training, validation and test data originate from the same subset, the respective splits led exactly to an 80:10:10 distributed number of samples. When training on one (or more) class(es) and testing on samples of a single remaining class, we held out all samples of the target class for testing. Consequently, for some of the combinations the number of test samples equals or even exceeds the number of samples in the train and validation data (for details see Table 2).

In a first step, BERT$_{base}$ was trained on the full dataset to classify samples of all three tribunals, but also to classify tribunal-specific text chunks.

Secondly, we apply the same setup to all three subsets. More specifically, training was performed using tribunal-specific samples to see if BERT is still able to predict class labels of both, the mixed dataset (excluding training class), as well as the remaining tribunal-specific subsets.

To test for the detection of undersampled violence-related paragraphs, additional experiments on this data were set up. All of the subset-specific negative class samples were used and a random proportion of 20% of positive class samples was added.

For training and validation a batch-size of 16 samples and an epoch-number of 3 (compare Devlin et al. (2019)) was used. The training was executed using 4 Nvidia RTX 2080Ti GPUs with an overall memory size of 44GB.

Precision, recall, micro and macro F1 scores for each train/validate/test constellation are provided – in line with common practice, macro F1 scores will be the reference score when comparing results (Jurafsky and Martin, 2021).

5. Results

Our results show that a binary classification based on BERT yields very reliable results across text data from different tribunals. A macro F1 score of 0.81 when training, testing and validating with the complete, mixed dataset that includes all three tribunals shows that BERT can be applied to this type of data and provides reasonably good predictions across the different subsets.

Considering the individual tribunals, using a tribunal-specific dataset for training and validating provided varying test results (ECCC-ECCC macro F1=0.70; ICTY-ICTY macro F1=0.68; ICTR-ICTR macro F1=0.80). Overall, using the mixed dataset for training and validating resulted in the highest F1 scores throughout the tribunal variations (min macro F1=0.78, max macro F1=0.85), independently of the dataset that was used for testing. The highest individual F1 score in our experiments was obtained when predicting data from ICTR transcripts with trained and validated data from the mixed dataset (”ALL") (macro F1=0.85).

Looking at the tribunal-specific outcomes for the respective training/validating/sets also yielded solid results overall: Interestingly, using the ECCC data for training and validating has the highest true prediction rates when testing is conducted with ICTR data (macro F1=0.79), whereas using ECCC data for training, validating *and* testing only led to a comparatively low macro F1 score of 0.70. When training with ICTY data, performance was also best when predicting ICTR data (macro F1=0.81). Results are similar for training and validating with ICTR data: The highest macro F1 score (0.80) was obtained when using ICTR data for testing.

Train/val data	Test data			
Mixed dataset	Mixed dataset (n_{train}=1180, n_{val}=147, n_{test}=148)	ECCC dataset (n_{train}=808, n_{val}=202, n_{test}=465)	ICTY dataset (n_{train}=756, n_{val}=189, n_{test}=530)	ICTR dataset (n_{train}=796, n_{val}=199, n_{test}=480)
ECCC dataset	Mixed dataset (n_{train}=372, n_{val}=93, n_{test}=1010)	ECCC dataset (n_{train}=372, n_{val}=46, n_{test}=47)	ICTY dataset (n_{train}=372, n_{val}=93, n_{test}=530)	ICTR dataset (n_{train}=372, n_{val}=93, n_{test}=480)
ICTY dataset	Mixed dataset (n_{train}=424, n_{val}=106, n_{test}=945)	ECCC dataset (n_{train}=424, n_{val}=106, n_{test}=465)	ICTY dataset (n_{train}=424, n_{val}=53, n_{test}=53)	ICTR dataset (n_{train}=424, n_{val}=106, n_{test}=480)
ICTR dataset	Mixed dataset (n_{train}=484, n_{val}=96, n_{test}=995)	ECCC dataset (n_{train}=384, n_{val}=96, n_{test}=465)	ICTY dataset (n_{train}=484, n_{val}=96, n_{test}=530)	ICTR dataset (n_{train}=484, n_{val}=48, n_{test}=48)

Table 2: Overview of sample balance for the complete, mixed dataset ("ALL") and the three individual tribunal datasets for each train/validate/test constellation.

Overall, precision and recall turned out to be fairly balanced throughout the different training and testing processes. See Table 3 for a detailed overview of the results. When conducting the experiments with undersampled violence-related data, results turn out to be different. Despite using class weights for training (due to the underrepresented positive label), the results obtained are much lower than those reported for the full dataset. For each subset (ECCC: macro F1=0.51, micro F1=0.81; ICTY: macro F1=0.45, micro F1=0.74; ICTR: macro F1=0.45, micro F1=0.75) as well as for the mixed dataset (macro F1=0.47, micro F1=0.77) macro F1 scores are about half of the values reported so far. Since positive samples are heavily underrepresented (e.g. 1 out of 31 samples in the test set) precision, recall and binary F1 for this class amount to 0.0 for a range of data splits. This leads to the overall poor results for this setup. It also offers directions for future experiments.

6. Discussion

General Discussion: This study presented a new type of dataset for NLP-based research in the field of genocide and violence studies. BERT$_{base}$ was further used to predict if text passages from court transcripts of three different genocide tribunals contain accounts of experienced violence by the respective witnesses.

The results, in line with expectations, indicate that the mixed dataset is most successful when predicting if a certain text passage from one of three genocide tribunals contains accounts of experienced violence by a witness. Even when classifying paragraphs of one specific tribunal (e.g., the ECCC) with the model that was trained with data from the same tribunal (ECCC in this case), the model trained on the complete dataset provides better results. Including additional data from other tribunals thus improves the quality of the classification.

Contextualisation: Looking at the wider picture, binary classification scores vary widely across NLP applications (Arase and Tsujii, 2019; Wang et al., 2019) – direct comparisons with other studies must therefore be interpreted with caution. Nonetheless, the ballpark figures we obtained are comparable to state-of-the-art (BERT-based) performance on some other commonly used binary classifications such as MRPC (Zhang et al., 2021), but fall short of performance levels expected for other settings (d'Sa et al., 2020). On the one hand, this confirms once more that BERT can be successfully applied to our corpus and perfectly presents how well this language model has been developed in recent years. On the other hand, further fine-tuning will be necessary to solve performance-related shortcomings.

Precision vs. Recall: The overall similarity of precision and recall rates in our dataset implies that this type of classification might be useful for a broad range of applications. In some cases, recall rates might be more important than precision rates: for example, similar to patent search (Bache, 2011; Bashir and Rauber, 2010), a high recall is especially important when avoiding missed positive classifications is crucial. In a genocide-transcript-related context, this could apply to staff members who have to work through court transcripts as part of their daily work routine, e.g. for preparing a case. For this option specifically, applying the classification algorithm reduces the time spent on manual search drastically, making sure that no sample is missing and leaving time for manual adaptions. On the other hand, in the context of fast and efficient search with less time for manual adaptions, high precision rates would be more useful, e.g., when only some examples of relevant text segments are required and correctness is more important than completeness (Kong and Allan, 2016).

	ALL				ECCC				ICTY				ICTR			
	P	R	mac. F1	mic. F1	P	R	mac. F1	mic. F1	P	R	mac. F1	mic. F1	P	R	mac. F1	mic. F1
ALL	0.81	0.83	0.81	0.83	0.81	0.82	0.82	0.82	0.78	0.78	0.78	0.83	0.85	0.85	0.85	0.85
ECCC	0.77	0.77	0.77	0.79	0.77	0.72	0.70	0.75	0.73	0.71	0.71	0.78	0.81	0.79	0.79	0.80
ICTY	0.77	0.78	0.77	0.78	0.77	0.78	0.77	0.78	0.70	0.73	0.68	0.74	0.81	0.81	0.81	0.81
ICTR	0.74	0.74	0.74	0.78	0.79	0.77	0.78	0.79	0.69	0.74	0.70	0.75	0.83	0.78	0.80	0.85

Table 3: Results for macro precision (P), macro recall (R) and macro/micro F1 scores on test data (columns) with respect to different training/evaluation set (rows) combinations.

Number of Text Chunks and Labeling Balance: When looking at the results of this study, imbalances in the number of text chunks, labels and in the train-validate-test-ratio must be kept in mind. Still, in spite of the ICTY data containing fewer violence-related text segments, results did only differ slightly, indicating that this label imbalance does not impact the results significantly.

However, extending the dataset further would be a first step in making the results more stable. More text data could also help to improve the label balance: by selecting more transcripts per tribunal, the chances of choosing transcripts that contain no/few or above-average accounts of violence can be reduced.

An extended dataset would also make it easier to experiment with undersampled violence-related paragraphs. As already mentioned in Section 5, this setup currently lacks a sufficient number of positive labels in the test sets (when undersampling this class in an adequate ratio to keep the overall number of samples stable). Thus, adding more (non-violent) text chunks would make it easier to generate representative training/validation/test splits with a sufficient number of paragraphs for both classes regarding this setup. However, the dataset as it is offers directions for a range of possible experiments including the identification of violence related text chunks when heavily underrepresented.

Future Research: This dataset has the potential of serving as a basis for a variety of research approaches in the field of genocide research in the future. For example, more in-depth comparisons between linguistic or content-based characteristics between the three tribunals could be made, building bridges between the interdisciplinary field of genocide research and NLP-approaches. Since the provided dataset is violence-based, further research could, for example, build on psycho-linguistic aspects of violence-related trauma in witness statements of genocide tribunals.

From an NLP perspective, next steps could include further fine-tuning of BERT and establishing a model version that is pre-trained specifically on court transcripts of genocide tribunals. Conducting the experiments with more recent transformer architectures or machine learning techniques could also yield interesting results and would therefore be a good starting point for future research. Given that the full annotated dataset is publicly available online, further studies could also include a detailed error analysis of the misclassified paragraphs. Not least we see our work as a first step towards a downstream practical search system.

7. Conclusion

This paper introduces a new dataset of genocide transcript data as a basis for further NLP research and applications. In addition, a baseline for classifying the transcript samples into violent or non-violent text chunks respectively is provided. The results, based on the well-established BERT architecture, demonstrate that such models can successfully be applied to this new domain and its related classification task. Although the number of text segments used in this study could be further extended (as it especially was observable during experiments with undersampled violence related paragraphs), classification with BERT proved to be successful, emphasizing once more the potential this language model holds even for research areas that have not been in the focus of NLP applications.

8. Ethical Considerations

All of the transcripts used in this paper are published online on the respective courts' websites and are publicly accessible. Since this type of text material contains personal and highly sensitive information about witnesses before international criminal courts, special care was taken to ensure that text fragments were not taken out of context. The use of witness names (or their anonymisation) in the dataset was adopted according to the original court document.

9. Acknowledgements

This project was supported by the Foundation of German Business (Stiftung der deutschen Wirtschaft e.V., sdw). We would further like to thank the three anonymous reviewers for their helpful comments.

10. Bibliographical References

Al Amrani, Y., Lazaar, M., and El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, 127:511–520.

Alhelbawy, A., Poesio, M., and Kruschwitz, U. (2016). Towards a corpus of violence acts in arabic social media. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1627–1631, Portorož, Slovenia, May. European Language Resource Association (ELRA).

Arase, Y. and Tsujii, J. (2019). Transfer fine-tuning: A BERT case study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5393–5404, Hong Kong, China, November. Association for Computational Linguistics.

Bache, R. (2011). Measuring and improving access to the corpus. In Mihai Lupu, et al., editors, *Current Challenges in Patent Information Retrieval*, pages 147–165. Springer, Berlin, Heidelberg.

Bachman, J. (2020). Cases studied in genocide studies and prevention and journal of genocide research and implications for the field of genocide studies. *Genocide Studies and Prevention: An International Journal*, 14(1):2–20.

Bashir, S. and Rauber, A. (2010). Improving retrievability of patents in prior-art search. In Cathal Gurrin, et al., editors, *ECIR 2010: Advances in Information Retrieval*, pages 457–470, Milton Keynes, UK, March. Springer.

Bijalwan, V., Kumar, V., Kumari, P., and Pascual, J. (2014). KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, 7(1):61–70.

Bullers, K., Howard, A. M., Hanson, A., Kearns, W. D., Orriola, J. J., Polo, R. L., and Sakmar, K. A. (2018). It takes longer than you think: Librarian time spent on systematic review tasks. *Journal of the Medical Library Association: JMLA*, 106(2):198–207.

Calzolari, N., Soria, C., Gratta, R. D., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J., and Piperidis, S. (2010). The LREC map of language resources and technologies. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dhar, A., Mukherjee, H., Dash, N. S., and Roy, K. (2021). Text categorization: Past and present. *Artificial Intelligence Review*, 54(4):3007–3054.

d'Sa, A. G., Illina, I., and Fohr, D. (2020). BERT and fastText embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, pages 1–5. IEEE.

Fidahić, B. (2021). Case study: The International Criminal Tribunal for the former Yugoslavia's court transcripts in Bosnian/Croatian/Serbian—part 1: Needs, feasibility, and output assessment. *Genocide Studies and Prevention: An International Journal*, 15(2):37–48.

Greene-Colozzi, E. A., Freilich, J. D., and Chermak, S. M. (2021). Developing open-source databases from online sources to study online and offline phenomena. In Anita Lavorgna et al., editors, *Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches*, pages 169–190. Springer International Publishing, Cham.

Hinton, A. L. (2012). Critical genocide studies. *Genocide Studies and Prevention: An International Journal*, 7(1):4–15.

Hoang, L. and Schneider, J. (2018). Opportunities for computer support for systematic reviewing - a gap analysis. In Gobinda Chowdhury, et al., editors, *Transforming Digital Worlds – 13th International Conference, iConference 2018, Proceedings*, pages 367–377, Germany, March. Springer International Publishing.

ICTY. (2016). *Infographic: ICTY Facts & Figures*. International Criminal Tribunal for the former Yugoslavia. https://www.icty.org/node/9590.

Jurafsky, D. and Martin, J. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd (draft) edition. https://web.stanford.edu/˜jurafsky/slp3/.

Kaptein, R., Van den Broek, E. L., Koot, G., and Huis in 't Veld, M. A. (2013). Recall oriented search on the web using semantic annotations. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR '13)*, page 45–48, San Francisco, California, USA, October. Association for Computing Machinery.

Kong, W. and Allan, J. (2016). Precision-oriented query facet extraction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 1433–1442, New York, NY, USA. Association for Computing Machinery.

Koster, C. H., Oostdijk, N. H., Verberne, S., and D'hondt, E. K. (2009). Challenges in professional search with PHASAR. In *Proceedings of the Dutch-Belgian Information Retrieval Workshop*, pages 101–102, Netherlands.

Kruschwitz, U. and Hull, C. (2017). Searching the enterprise. *Foundations and Trends in Information Retrieval*, 11(1):1–142.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2020). A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

MacFarlane, A., Russell-Rose, T., and Shokraneh, F. (2021). Search strategy formulation for systematic reviews: Issues, challenges and opportunities. *arXiv preprint arXiv:2112.09424*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys*, 54(3):1–40.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2020a). A BERT-based transfer learning approach for hate speech detection in online social media. In Hocine Cherifi, et al., editors, *Complex Networks and Their Applications VIII*, pages 928–940, Cham. Springer International Publishing.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2020b). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8):e0237861.

Müller, M., Salathé, M., and Kummervold, P. E. (2020). COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Nakov, P., Corney, D. P. A., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S., and Martino, G. D. S. (2021). Automated fact-checking for assisting human fact-checkers. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4551–4558. ijcai.org.

Noor, S. and Bashir, S. (2015). Evaluating bias in retrieval systems for recall oriented documents retrieval. *International Arab Journal of Information Technology (IAJIT)*, 12(1):53–59.

Russell-Rose, T., Gooch, P., and Kruschwitz, U. (2021). Interactive query expansion for professional search applications. *Business Information Review*, 38(3):127–137.

Singh, M., Jakhar, A. K., and Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining*, 11:33.

Sohn, H. and Lee, H. (2019). Mc-bert4hate: Hate speech detection using multi-channel BERT for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559, Beijing, China, November. IEEE.

Sun, C., Huang, L., and Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tong, S. and Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(Nov):45–66.

Totten, S. (2017). *Advancing Genocide Studies: Personal Accounts and Insights from Scholars in the Field*. Routledge.

Verberne, S., He, J., Wiggers, G., Russell-Rose, T., Kruschwitz, U., and de Vries, A. P. (2019). Information search in a professional context - exploring a collection of professional search tasks. *CoRR*, abs/1905.04577.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, Louisiana, May.

Webersinke, N., Kraus, M., Bingler, J. A., and Leippold, M. (2021). ClimateBert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010v1*.

Xu, H., Liu, B., Shu, L., and Yu, P. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zhang, W., Yoshida, T., and Tang, X. (2007). Text classification based on multi-word with support vector machine. In *2007 IEEE International Conference on Systems, Man and Cybernetics*, pages 3519–3524, Montreal, Canada. IEEE.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. (2021). Revisiting few-sample BERT fine-tuning. In *International Conference on Learning Representations*, Vienna, Austria.

4.3 Study 3: Uncovering Trauma in Genocide Tribunals: An NLP Approach Using the Genocide Transcript Corpus

This publication is RELEVANT TO THE EXAMINATION.

Authors

Miriam Schirmer, Isaac Misael Olguín Nolasco, Edoardo Mosca, Shanshan Xu, Jürgen Pfeffer

In

Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), 257-266, peer-reviewed, published 06-2023, `https://doi.org/10.1145/3594536.3595147`

Abstract

This paper applies Natural Language Processing (NLP) methods to analyze the exposure to trauma experienced by witnesses in international criminal tribunals when testifying in court. One major contribution of this study is the creation of a substantially extended version of the Genocide Transcript Corpus (GTC) that includes 15,845 text segments of transcripts from three different genocide tribunals. Based on this data, we first examine the prevalence of trauma-related content in witness statements. Second, we are implementing a binary classification algorithm to automatically detect potentially traumatic content. Therefore, in a preparatory step, an Active Learning (AL) approach is applied to establish the ideal size for the training data set. Subsequently, this data is used to train a transformer model. In this case, the two models BERTbase and HateBERT are used for both steps, allowing for a comparison of a base-level model with a model that has already been pre-trained on data more relevant in the context of harmful vocabulary. In a third step, the study employs an Explainable Artificial Intelligence (XAI) model to gain a deeper understanding of the reasoning behind the model's classifications. Our results suggest that both BERTbase and HateBERT perform comparatively well on this classification task, with no model clearly outperforming the other. The classification outcomes further suggest that a reduced data set size can achieve equally high performance metrics and might be a preferable choice in certain use cases. The results can be used to establish more trauma-informed legal procedures in genocide-related tribunals, including the identification of potentially re-traumatizing examination approaches at an early stage.

Contribution of Thesis Author

Theoretical conceptualization, data curation, methodological design, formal analysis, visualization, manuscript writing, revision, and editing.

Uncovering Trauma in Genocide Tribunals: An NLP Approach Using the Genocide Transcript Corpus

Paper ID 182

ABSTRACT

Warning: Due to the overall purpose of the study, this paper contains descriptions of violent events in Section 4.1 (Examples 1 and 2) and in Figure 3 that may be distressing for some readers.

This paper applies Natural Language Processing (NLP) methods to uncover and analyze exposure to trauma to which witnesses in international criminal tribunals are subjected when recounting their experiences in court. One major contribution of this study is the creation of a substantially extended version of the Genocide Transcript Corpus (GTC) that includes 15,845 text segments of transcripts from three different genocide tribunals. Based on this data, we first examine the prevalence of trauma-related content in witness statements. Second, we are implementing binary classification algorithm to automatically detect potentially traumatic content. Therefore, in a preparatory step, an Active Learning (AL) approach is applied to establish the ideal size for the training data set. Subsequently, this data is used to train a transformer model. In this case, the two transformer models BERTbase and HateBERT are used for both steps, allowing for a comparison of a base-level model with a model that has already been pre-trained on data more similar to the concept of harmful vocabulary. In a third step, the study employs an Explainable Artificial Intelligence (XAI) model to gain a deeper understanding of the reasoning behind the model's classifications. Our results suggest that both BERTbase and HateBERT perform comparatively well on this classification task, with no model clearly outperforming the other. The classification outcomes further suggest that a reduced data set size can achieve equally high performance metrics and might be a preferable choice in certain use cases. The results can be used to establish more trauma-informed legal procedures in genocide-related tribunals.

CCS CONCEPTS

• **Computing methodologies** → *Information extraction*; **Language resources**; **Supervised learning by classification**; Active learning settings; • **Applied computing** → **Law**.

KEYWORDS

trauma, genocide, classification, BERT, XAI

ACM Reference Format:
Paper ID 182. 2023. Uncovering Trauma in Genocide Tribunals: An NLP Approach Using the Genocide Transcript Corpus. In *Proceedings of the 19th International Conference for Artificial Intelligence and Law (ICAIL '23).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

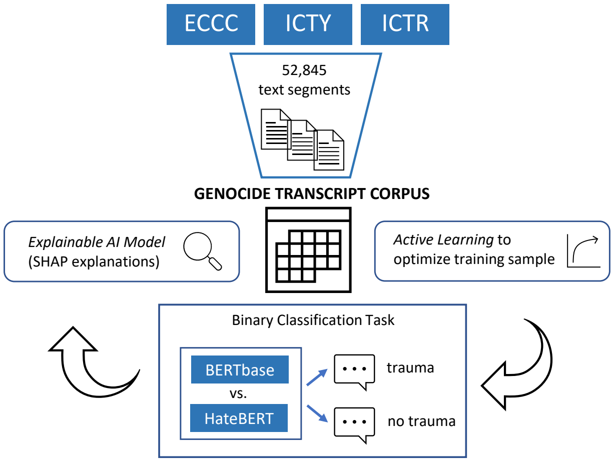


Figure 1: This work presents a significantly updated version of the Genocide Transcript Corpus that is used for an Active-Learning- and Explainable-AI-supported binary classification task to detect trauma in witness statements.

1 INTRODUCTION

Psychological trauma is defined as “exposure to actual or threatened death, serious injury, or sexual violence” that is either experienced directly or witnessed. It further includes “learning that the traumatic event(s) occurred to a close family member or close friend” and “experiencing repeated or extreme exposure to aversive details" of such events [3]. In the context of international criminal courts that deal with the legal processing of mass atrocities and genocide, trauma is a relevant and frequent issue in testimonies, for example, when witnesses recount severe cases of violence, such as having experienced torture or witnessed mass killings. Given that re-accounting such events can be emotionally challenging for witnesses and negatively impact their testimony, it is crucial to identify potentially traumatizing content as early as possible to provide witness support in international genocide tribunals and improve the quality of their testimony at the same time. The early identification of traumatic content thereby further reduces the emotional toll on witnesses.

Legal Natural Language Processing (Legal NLP) provides useful tools to address this challenge through the application of machine

ICAIL '23, June 19–23, 2023, Braga, Portugal
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

learning algorithms to analyze and classify text. While typical research topics in Legal NLP usually center around knowledge modeling, legal reasoning, and interpretability [51], witness perspectives are rarely focused on. Therefore, taking witness statements extracted from court transcripts more closely into account presents another important challenge to Legal NLP.

Our paper addresses this issue by identifying potentially traumatizing content in witness statements before international criminal tribunals. Through the implementation of an algorithm that automatically classifies trauma-related text segments, this paper seeks to uncover the prevalence of trauma and to enable its early detection during trial. In this work, we introduce an updated and refined version of the *Genocide Transcript Corpus* (GTC, Version 2) [37] for NLP with a focus on detecting potential traumatic content in witness statements. The data set includes transcripts of the three biggest genocide tribunals: the *Extraordinary Chambers in the Courts of Cambodia* (ECCC), the *International Criminal Tribunal for Rwanda* (ICTR), and the ICTY. From an NLP perspective, we define potential traumatic content detection as a binary classification task: given a snippet of court transcripts that include witness testimonies, the model should classify whether or not it contains trauma-related content. We evaluate the performance of the two large pre-trained transformer models BERT base [13] and HateBERT [6] the trauma detection task and compare them regarding their accuracy, precision, recall, and F1 score. To find the ideal size of the training data set and thus improve data efficiency in learning, we apply Active Learning (AL) as a preparatory step for the trauma detection task. Lastly, this study utilizes an explainable AI model to gain a deeper insight into why the model makes certain classifications.

In the context of uncovering trauma in international criminal tribunals, this paper makes several significant contributions to the field of legal AI by:

1. creating an extensive data set of court transcripts for NLP tasks in genocide research that consists of 52,845 text segments with 18,854 witness statements that were manually labeled regarding traumatic content,
2. providing benchmark values on BERT-based binary classification tasks for two different models and different data set sizes,
3. proving the effectiveness of the model with an Explainable AI approach that gives further insights in the reasoning behind the models' predictions,
4. making all data and code available to the community.[1]

2 RELATED WORK

2.1 Transcripts as Resources for Legal NLP

The rapid growth of Legal NLP is demonstrated by the introduction of numerous data sets containing legal documents, for instance Legal Judgement Prediction [7, 48], Competition on Legal Information Extraction/Entailment [20, 21], Legal Document Summarization [40, 52], etc. To the best of our knowledge, most of the data sets in Legal NLP source from written text, such as course facts. Meanwhile, our data set is built out of transcripts of court hearings – since this type of documents contain loosely-structured dialogue, its processing presents particular challenging. Similar work using court transcripts as a language resource for NLP has been done by Hong et al. [18], where the authors extracted factual information from parole hearings. Still, studies in Legal NLP using court transcripts as a resource remain scare. Usually, case facts are drafted by the court register, focusing on the background information and factual events. Given that transcripts of the court hearing also include personal accounts, their analysis helps to uncover the actual situation that witnesses have to face during the proceedings and provide advantages for research questions that center around the witness perspective.

2.2 Trauma and Genocide in Court

Since many witness testimonies given in genocide tribunals include detailed descriptions of "exposure to actual or threatened death, serious injury, or sexual violence", they clearly contain content that can be classified as traumatic according to the APA definition provided in Section 1 [3]. When it comes to defining trauma for real-life events, defining and measuring trauma is a complex and challenging task with many of the fundamental issues remaining unresolved [47]. However, according to this definition, torture, political persecution, and imprisonment in the context of genocide can be categorically assigned to the trauma concept – particularly given that survivors of genocide were routinely subjected to violence and faced an ongoing and imminent threat to their lives.

A comprehensive analysis of trauma in the aftermath of the genocide and its impact on genocide tribunals has been conducted by various authors. For example, referring to the ECCC, Ciorciari and Heindel [9] conclude that the participation of traumatized individuals in the court proceedings represents an "emotionally-difficult process", in which a certain degree of re-traumatization is unavoidable. The authors therefore urge judges and attorneys to receive appropriate sensitivity training, as well as to provide traumatized witnesses with professional support. Similar observations have been made by Viebach [45] and Soueid et al. [42] who discuss the difficulties in addressing traumatic experiences of genocide during the ICTR and the ICTY. Accounts of traumatic experiences in genocide tribunal can vary widely depending on the type of testimony and the witness's individual experience. Therefore, witness accounts might differ regarding the emotional involvement and the risk of re-traumatization [11].

2.3 Trauma Detection with NLP

Considering the variety of traumatic experiences with their subjective nature and the difficulty of a clear conceptual definition of trauma, its detection in text material is a complex task. Despite these challenges, recent research has shown how NLP methods can improve the detection of psychological disorders or adapt therapeutic treatment [24, 49].

Specifically in trauma research, advances are being made to detect and evaluate traumatic experiences, such as in analyzing patient narratives [17] or in identifying cases of post-traumatic stress disorder (PTSD) based on speech samples from veterans [30]. Even though there is a growing interest in using NLP techniques to identify mental illnesses, it remains difficult to identify trauma from text.

[1] https://anonymous.4open.science/r/GTC-V2-F312/ (anonymous for review process)

Most of the studies trying to detect trauma are set in a clinical context. They aim at detecting specific psychological disorders, such as PTSD, and rely on diagnoses that have already been made to evaluate the performance of the NLP model. For the detection of trauma-related statements in witness testimonies, however, the detection of psychological disorders is not feasible: Without any further information on the psychological situation of the witnesses or prior diagnoses, it is impossible to draw conclusions about their the mental health from court transcripts alone. This is why this study focuses on witness accounts that describe events that can be categorized as traumatic, but do not necessarily have led to a traumatic response.

Except for one paper that established a new data set for genocide-related court transcripts [37], none of the above-mentioned techniques have been applied to transcripts of genocide tribunals, including the detection of trauma-related content. Therefore, applying NLP techniques to transcripts of genocide tribunals can provide new insights and advance the field of Legal NLP.

2.4 Optimized Text Classification Through Active Learning Support

2.4.1 Binary Text Classification. A promising approach to detect trauma-related content in witness statements is binary classification through a supervised learning algorithm using a pre-trained NLP model. Among different machine learning models for text classification, transformer-based architectures have been established as state-of-the-art models that outperform other algorithms, such as convolutional (CNNs) or recurrent neural networks (RNNs) [26]. One of the most extensively researched transformer-based models for NLP tasks is *Bidirectional Encoder Representations from Transformers* (BERT) [13]. Designed to understand the context and relationships between words in a text, BERT's pre-trained 12-layer-architecture allows the model to learn general representations of language very efficiently and thus make it a powerful model for text classification.

So far, BERT has been adapted for a wide variety of different domains, yielding highly efficient pre-trained transformer-models, such as COVID-Twitter-BERT [33] or BioBERT [25]. For NLP tasks specifically in the field of legal AI, LegalBERT [8] has been widely used as a state-of-the-art model with high performance metrics for legal use cases. Beside these domain specific transformer models, several BERT variations for topics more connected to trauma and genocide exist. One example is ConfliBERT [19] that has been pre-trained on text data related to international conflicts, such as news data and government reports. Other BERT variations that seem promising in the context of this study are models pre-trained on harmful language.

For the context of this study, HateBERT [6] – as an English pre-trained BERT model that has been further trained with more than 1 million posts from banned Reddit communities – seems especially promising: Being closely linked to violence, hate speech can be intertwined with trauma-related speech in the context of genocide in that it can be part of violent actions that cause trauma or be traumatizing by itself. Given the thematic focus of trauma in this study, HateBERT might thus prove to be more relevant for detecting potentially traumatic content than, for example, LegalBERT, that would be better suited to identify legal language and concepts. For that reason, in addition to using BERTbase as a baseline model, we will be conducting the classification task with HateBERT as well.

2.4.2 Active Learning Approach. One way of enhancing binary classification is the application of an *Active Learning* (AL) approach to optimize the size of the training data set. This technique allows the model to actively select samples from the data set that are accompanied by a higher uncertainty for the classification process [14]. Thus, instead of randomly selecting data samples, the AL algorithm focuses on samples that are more informative. Concentrating on the quality of the data rather than its quantity, AL helps to select which data are required to enhance model performance. Starting with a small training data set and gradually increasing the amount of training data, AL can help determine the saturation of the model or the amount of samples needed to reach optimal performance for a binary text classification task [38].

In this study, we will be employing an AL approach to determine the optimal size of our training data set. The optimized data set can subsequently be used for the classification task.

2.5 Explainable Artificial Intelligence

To provide interpretable and transparent explanations of our resulting model's predictions, we will use *eXplainable Artificial Intelligence* (XAI) to analyze the mechanisms behind the classification task [2]. XAI is a relatively novel research field that has recently gained popularity in NLP as state-of-the-art models—such as BERT [13], Bloom [36], and GPT-3 [5]—behave like black boxes [4, 32]. In particular, *post-hoc explainability* approaches enable us to explain the model's reasoning for a certain prediction even when the architecture is not inherently interpretable. Thus, they can be applied without sacrificing predictive performance [29].

Most relevant to this work are *local feature attribution explanations*, seeking to quantify the relevance of each input feature for the current prediction instance. Hence the adjective *local*, which indicates the explanation refers to a specific input-output pair [27, 29].

Methods vary greatly in terms of how they compute the relevance of input features. For instance, some rely on computing output gradients w.r.t. each feature [12, 43]. Others are instead tailored to explain neural networks and take advantage of their layered architecture to propagate importance in a backward fashion: DeepLIFT [41]. LIME [34] approximates complex models around a single instance via a local surrogate that allows for direct interpretation. Finally, Lundberg and Lee [28] propose SHAP, an explainability framework inspired by the game-theoretic concept of Shapley values [28]. For our purpose, we choose SHAP because of its solid theoretical background and its availability in well-maintained interpretability libraries (see 4.4 for more details).

2.6 Concluding Remarks

The purpose of this paper is to shed more light on witnesses' subjection to possible trauma during international criminal tribunals. With the application of AL, transformer-based classification, and XAI models, it brings together three key concepts of NLP and provides benchmark values based on a new corpus. So far, NLP-based research in the context of genocide is very scarce – only a few studies exist that have shown how transcript material can be analyzed through an NLP lens [23, 37]. The new version of the GTC serves as a representative data source of three big genocide tribunals and fills the gap in available language resources in the field of NLP-based genocide research – thus contributing to the development of legal AI research in the context of international criminal courts.

3 THE GENOCIDE TRANSCRIPT CORPUS

3.1 Data Selection

The Genocide Transcript Corpus (GTC) [37] was initially developed as the first annotated data set for NLP in the field of genocide research, providing important benchmark values for text classification. However, the data set is limited in size and only consisted of 20 transcripts, making the results based on analyses less reliable and impactful. To address this issue, our version of the GTC (Version 2) has been significantly extended to 90 transcripts from the three largest genocide tribunals: the ECCC, the ICTY, and the ICTR.

To select the transcripts for the ICTR and ICTY, cases were chosen based on the final judgment. We chose five cases both of these tribunals where the accused was sentenced to life imprisonment, and randomly picked six cases from the transcripts that included witness testimonies for the prosecution (excluding expert witnesses). Regarding the ECCC, only two cases are available online. Therefore, we included those two cases with 15 transcripts each in our data set. The GTC includes a wide range of witness statements with various backgrounds, such as soldiers, prisoners who were subjected to torture, and guards who carried out torture. As a result, some of the transcripts have a more political or administrative focus whereas others provide more detailed descriptions of violence, and thus including more statements related to trauma. This data set also includes witnesses for the prosecution who have committed crimes during the genocide. See Table 1 for an overview of the transcripts and cases included in the GTC.

3.2 Data Annotation and Final Data Set

To structure the transcript text and to make it suitable for NLP research, we annotated all transcripts based on the speaker's role in the legal proceedings. Prior to the annotation process, transcripts were obtained by scraping them from their HTML links. They were lightly pre-processed, removing line numbers, url-links, html tags, and some parts of the technical document information. In the adapted transcripts, we identified and tagged statements made by judges, lawyers, witnesses, and the accused. Further, we differentiated between statements made during questioning of witnesses (JudgeQA or LawyerQA) and discussions about legal proceedings (JudgeProc or LawyerProc). Individual persons were not distinguished. We also marked formal parts of the transcripts, such as editorial comments, as court proceedings. Text segments can range from short one-word-sentences (e.g., The witness answers with "A. Yes.") to replies that span multiple paragraphs. To optimize the efficiency of NLP tasks, text snippets exceeding the length of about 500 token were split.

The updated version the GTC contains 52,845 text segments of a total of 90 transcripts that can be attributed to an individual person or court proceedings. The final data set includes the following variables:

- **Case** information: Tribunal, case number, accused
- **Transcript** information: Document ID, url-link to the original transcript, date
- **Witness** information: Witness name or pseudonym, number of witnesses per transcript
- **Text** information: Speaker (e.g., Witness, LawyerQA), text, trauma label
- **Annotation** information: Annotation ID, start ID, and document ID

4 METHODS

4.1 Labeling Trauma

Determining whether a text snippet contains trauma-related content is a complex task. It is very important to note that it is not possible to determine if a witness is actually traumatized either before their statement in court or through the testimony itself. Respective diagnoses require a profound assessment of the witness's personal history that exceeds the contents of the witness transcripts by far. In this study, we rely on the APA trauma definition outlined in Section 2.2 as a guideline to label text snippets that could potentially be describing a traumatic event and manually labeled all text segments containing witness statements (n = 18,854) accordingly. Our trauma label includes accounts of witnessed military attacks and bombings, killings, physical violence, threats, and humiliation directed to oneself or close people. Destruction and looting of one's own property is also labeled as potentially traumatic. It should be pointed out that text snippets were only labeled as potentially traumatic when the traumatic event became evident solely through the respective text segment and was directly observed by the witness. For illustration, Examples 1 and 2 provide witness statements that were labeled as describing a potentially traumatizing event. Both examples depict answers made by witnesses during a Question and Answering part of the trial session ("A." indicating "Answer).

To ensure consistency in labeling, inter-rater reliability was calculated. For this purpose, we used Fleiss' kappa as an adaption of Cohen's kappa for more than two annotators [16]. In our case, a sub-sample of the data (n = 2021) was labeled by three different annotators to see if the labeling itself is valid. All annotators were graduate-level psychologists to ensure a sufficient level of annotator expertise regarding trauma. The resulting inter-rater reliability of κ = .84 suggests a high agreement between the individual annotators and thus validates the consistency and accuracy of the labeling process used in our study.

Example 1 – Trauma Label

A. I saw two or three other bodies that jumped off the bridge, or were thrown off the bridge, after I had jumped. Not everyone jumped off the bridge. After the two or three bodies that jumped, the two or three people that jumped into the water, you could hear them opening bursts of fire for quite some time, and then there was silence for a while. You could no longer hear anything.

(ICTY, Case IT-09-92, Transcript 120904IT)

Table 1: Overview of the number of transcripts, the number of text segments (in total and for witnesses only), the number of witnesses whose testimonies are captured by the included transcripts, and the year of the hearings in the GTC (each per case and per accused)

Cases	*n* transcripts	*n* segments (all / witness)	*n* witnesses	Year of hearing
ECCC	**30**	**15876 / 6120**	**49**	
Kaing Guev Eav	15	8189 / 3199	27	2009
Nuon Chea & Khieu Samphan	15	7687 / 2921	22	2015, 2016
ICTY	**30**	**19217 / 6637**	**44**	
Ratko Mladić	6	4222 / 1246	11	2012
Vujadin Popović et al.	6	4011 / 1339	9	2006, 2007
Milan & Sredoje Lukić	6	5115 / 1946	10	2008
Zdravko Tolimir	6	3229 / 1048	8	2010
Milomir Stakić	6	2640 / 1058	6	2002
ICTR	**30**	**17752 / 6097**	**45**	
Callixte Nzabonimana	6	3089 / 869	6	2004, 2005, 2006
Édouard Karemera et al.	6	5041 / 1859	8	2005
Sylvestre Gacumbitsi	6	3402 / 1262	11	2003
Tharcisse Renzaho	6	2546 / 893	10	2007
Athanase Seromba	6	3674 / 1214	10	2004, 2005
TOTAL	**90**	**52845 / 18854**	**138**	

Example 2 – Trauma Label
A. I don't think I remember it. I think I had been terrified ever after having seen my mother being beaten.
(ECCC, Case 001, Transcript E1/42.1)

4.2 Active Learning with BERT and HateBERT

We applied an active learning approach to both BERT and HateBERT to detect the ideal amount of training data for an optimal performance of the binary classification model. For both the AL part and the final binary classification, we used the 12-layer BERT base model (uncased) [13] and GroNLP/hateBERT [6]. Training was implemented on a subsample of the GTC that only contains witness statements (*n* = 18,854).

In accordance with the AL setup, training was done repeatedly with an incrementally increasing size of data. Each time, new samples were selected to detect when the model converges or reaches maximum performance. Given the overall size of the data set, we opted for a total of 8 iterations for the AL process to achieve a satisfactory level of variability while avoiding an overly dense configuration. For training, the same distribution was constrained for the train, val and test sets. For the latter one, 5% of the dataset was extracted randomly, and from the remaining 95%, a train-val-split of 90:10 was used, with a batch size of 32 and 3 epochs. To evaluate the performance of the AL algorithm, we tested the model against a holdout test set and provide macro F1 as well as recall scores and report both metrics for the epoch in which the model performed best. We decided to report recall scores separately since, in our use case, it is more important to identify as many relevant instances as possible, even if this results in some false positives. This approach makes it more likely that no significant cases of trauma go unnoticed. This prioritization of recall over precision just serves as a reference for the selection of the ideal data set size – it was not technically implemented in the model. See Table 2 for an overview of all relevant model parameters.

Table 2: Model Setup

Parameter	Value
General parameters	
learning model	1e-6
epsilon	1e-08
epochs	3
β_1	0.9
β_2	0.99
batch size	32
GPU	NVIDIA A100-SXM4-40GB
AL-specific parameters	
iterations	8
epochs	4

4.3 Binary Classification with BERT and HateBERT

After the completion of the active learning process, the final models are used for a binary classification task to first see how both models perform generally with this type of data and to compare the model performance of both transformers in a second step. To optimize the classification results, we re-train both models with the amount of data that provides most satisfactory results in the AL classification. To ensure validity of the results, we implemented a 5-fold cross validation for the final classification tasks.

For a comprehensive comparison of the models, we report macro F1, recall, and precision scores, as well as the accuracy scores for both transformer models. Except for the AL-specific characteristics, the model parameters and technical details are identical to the previous AL setup (see Table 2).

4.4 Explainable AI

Complementary to performance metrics, we apply a post-hoc explainability pipeline to extract further insights about single predictions and overall model behavior. More specifically, we are interested in discovering which lexicon features are most relevant for the model to detect trauma in text. To this end, we apply SHAP [28], a feature attribution explanation method. This choice is based on the solid theoretical foundation of the method and its wide usage for analyzing models in NLP [31, 32].

SHAP is based on Shapley values [39], a concept from classical game theory originally defined as a fair measure to reward players contributing to a specific outcome. In a machine learning setting, the input text tokens represent the players whereas the outcome is the model's prediction. In other words, Shapley values fairly attribute an importance score to each part of the input based on their impact on the final classification result [28].

Computing Shapley values for a text instance of length N is unfeasible as it involves perturbing the input text and rerunning the model $O(2^N)$ times. Thus, we utilize *DeepSHAP* from the official SHAP library[2] to approximate Shapley values for our classifiers. We choose to use DeepSHAP as it is specifically tailored for deep architectures such as transformers and is thus more efficient and accurate than model-agnostic alternatives [28]. Concrete examples of SHAP explanations are presented in 5.3.

5 RESULTS

5.1 Prevalence of Trauma in the GTC

Our analysis revealed that 13.54% of all witness statements contain trauma-related content, indicating that a significant proportion of witness accounts discuss potentially distressing experiences. With a proportion of 19.64%, the number of possibly traumatic witness statements was highest for the ECCC, followed by the ICTY with 11.41% and the ICTR with 9.73%. The high number for the ECCC could be explained by the limitation of included cases for this tribunals, especially considering that one of the ECCC Cases (Case001 against Kaing Guev Eav) specifically dealt with one of the biggest prisons during the Cambodian Genocide, where detainees were subjected to interrogation and torture on a regular basis.

It is important to note that this number only includes cases that could be classified as traumatic based solely on specific text segments, and does not consider the broader context of the witness' accounts or physical reactions. As such, the actual amount of trauma-related content in these statements is likely to be higher, which further stresses the need for legal practices that rely on a more trauma-informed and witness-centered approach.

Table 3: Prevalence of trauma-related witness statements in the GTC

Tribunal	*n* witness segments	Trauma label (*n*)	Trauma label (%)
ECCC	6120	1202	19.64
ICTY	6637	757	11.41
ICTR	6097	593	9.73
TOTAL	18854	2552	13.54

5.2 Optimized Text Classification Through an Active-Learning-Driven Data Set

5.2.1 Active Learning for an Optimal Data Set Size. We applied an active learning approach to show how the number of labeled data required to achieve state-of-the-art-results can be reduced. However, having implemented an AL algorithm with 8 iterations on our data set, no point was reached where the model yielded stable performance metrics that could clearly not be improved. Figure **??** shows the progress of the AL iterations with an increasing data set size. The algorithm was tested against a holdout test set. For evaluation, F1 and recall scores are reported for the best performing epoch respectively.

While, for both models, the F1 score seemed relatively stable with a test data size from $n = 478$ to $n = 597$, there was a slight drop for an increased test data set of 836 samples. The best overall performance was reached when the models were evaluated with a data set size of $n = 478$. The same is true for the recall score. It is interesting to note that from the fourth AL iteration with a test sample size of $n = 479$, the model performance slightly decreased on average throughout the remaining AL iterations (e.g., from an macro F1 score of 0.889 to 0.870 for the BERTbase model).

Our results did not show an unambiguous saturation point for the model performance, which can be explained by various factors, such as insufficient data diversity or unrepresentative data samples (see Section 6 for further discussion). However, the AL approach showed that the model performed best with approximately half of the training data (test data size = 478). On this basis, we opted for conducting the binary classification task for both the reduced sample size and the full witness data set to compare performance values.

5.2.2 Binary Classification Task. Since the AL approach did not unambiguously suggest a higher model performance with a reduced number of training samples, both the full GTC data set (only including witness statements; $n = 18{,}854$) and a reduced data set ($n = 9{,}552$) were used to re-train both BERT and HateBERT and perform a binary classification task. The results are summarized in Table 4.

Examining the classification outcomes, the best overall model performance was achieved when training BERTbase with the reduced training data set (*F1* = 0.856; *Accuracy* = 0.947). Looking closer at the results after training with the *reduced data set size*, macro F1 scores of 0.886 for BERTbase and 0.836 for HateBERT are

[2] github.com/slundberg/shap

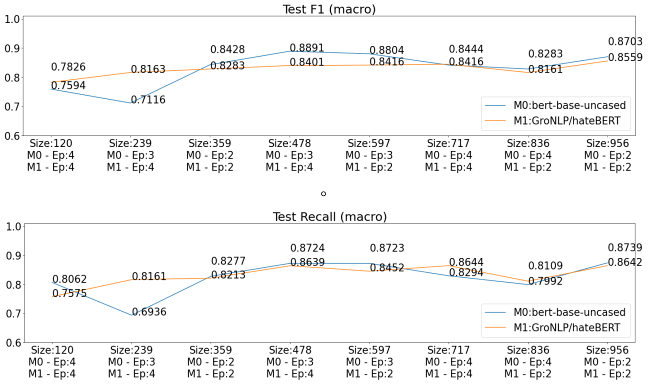


Figure 2: Active learning for data size optimization with F1 and recall scores for both BERTbase and HateBERT.

similar, which suggests that their performance is relatively equal. The same is true for the accuracy score, where BERTbase performs slightly better with a score of 0.947 while the accuracy score for BERTbase is 0.926. For the *full data set size*, differences in F1 and Accuracy scores where slightly higher between both models, with macro F1 scores of 0.812 for BERTbase and 0.858 for HateBERT. Interesting to note is that while BERTbase outperforms HateBERT with regard to the overall classification results, HateBERT performs slightly better when using the full witness data set for training.

Although on a very small scale only, more variation between the model performance becomes visible when taking a closer look at precision and recall values. As described in Section 4.2 recall scores are particular relevant for our use case, because we want to ensure that we detect as many cases of trauma as possible, even though this could go in hand with a higher score of false positives. Consequently, with a score of 0.881 BERTbase trained on the reduced data set provides the best overall recall performance.

Table 4: Result metrics of the binary classification (macro*)

Model	F1*	Recall*	Precision*	Acc
Reduced Witness Data (n = 9,552)				
bert-base-uncased	0.8855	0.8807	0.8909	0.9470
GroNLP/hateBERT	0.8360	0.8588	0.8186	0.9260
Full Witness Data (n = 18,854)				
bert-base-uncased	0.8162	0.8429	0.7956	0.9163
GroNLP/hateBERT	0.8581	0.8560	0.8632	0.9355

5.3 Explainable AI

Figure 3 and figure 4 show explanation examples for a traumatic and a non-traumatic instance respectively. The *base value* indicates the average model's prediction score across the whole dataset and $f(inputs)$ represents the predicted score for the selected instance. Tokens in red drive the predictions toward the *trauma* class while blue ones towards the *no trauma* one. Concretely, in the case of figure 3, starting from the base value (∼ −9.9) and adding up all token contributions all the way ∼ 11.3.

For the traumatic instance (Figure 3), the two-token segment *dead bodies* is the main contributor to the *trauma* output. We can observe that our BERT architecture successfully uses context to understand that the word *body* (referring to the truck's size) is not playing a strong role while *bodies* does as it comes together *dead*. This is possible thanks to the transformer-based nature of the classifier—able to learn contextualized text features.

Figure 4, instead, refers to a non-traumatic instance and displays most tokens not playing a strong role in either direction. In other words, in the absence of trauma, most tokens do not have a substantial impact with reference to the output, and the sample is predicted as non-traumatic. This is confirmed by the very negative base value, indicating that the default behavior of the classifier is to predict as *no trauma*.

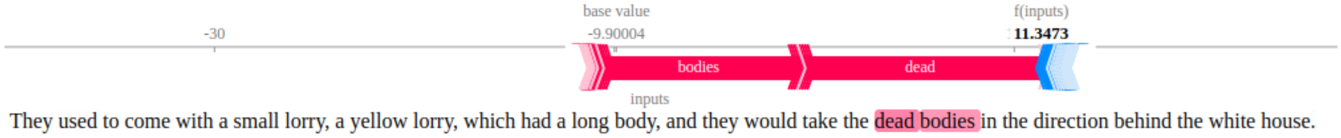
6 DISCUSSION

6.1 Contextualization

6.1.1 Prevalence of Trauma in the GTC. This paper is the first to assess the prevalence of trauma-related content in international criminal tribunals through NLP methods. While other authors have approached this topic through a rather qualitative lens and restricted to mostly one particular tribunal (e.g., Ciorciari and Heindel [9] for the ECCC or Dembour and Haslam [11] for the ICTR), our analysis reveals a first quantitative estimation of possibly traumatizing content in witness statements in the context of genocide tribunals. Given that we found a substantial amount of almost 14% of witness testimonies were potentially distressing, our paper stresses the claim for improved witness support in international criminal trials and supports it with empirical evidence.

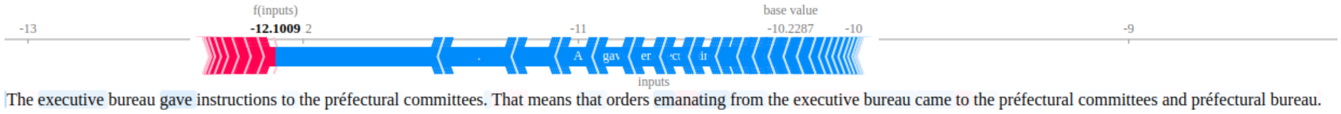
6.1.2 Detecting Trauma Through Binary Classification. Taking a broader view at binary classification tasks, scores vary significantly across NLP applications [1, 46]; consequently, direct comparisons with other studies should be treated with caution. However, the performance metrics obtained in our study are comparable to state-of-the-art (BERT-based) models on some other commonly used binary classification tasks such as the Microsoft Research Paraphrase Corpus (MRPC) [50]. Looking at HateBERT, the model performs similar to classical tasks evaluated by the original authors of the model [6]. Also compared to other models used for classification tasks in the context of hate speech detection results are similar – for the GTC data, even a slight improvement is detectable [35, 44]. However, when comparing the results of HateBERT in a different context, such as the detection of harmful speech against LGBTQIA+ individuals, HateBERT had lower evaluation metrics in our application case [10].

Examining the results in regard to a binary classification task specifically performed with genocide transcript data, our models seemed to be performing significantly better [37]. This could be due to the substantially larger data set than the one used for performing binary classification with the first version of the GTC.



They used to come with a small lorry, a yellow lorry, which had a long body, and they would take the dead bodies in the direction behind the white house.

Figure 3: SHAP explanation generated for an instance classified as *trauma* from our BERT uncased model.



The executive bureau gave instructions to the préfectural committees. That means that orders emanating from the executive bureau came to the préfectural committees and préfectural bureau.

Figure 4: SHAP explanation generated for an instance classified as *no trauma* from our BERT uncased model.

Regarding the model comparison, it is interesting to note that the application of HateBERT as a model specifically pre-trained on harmful language did not clearly outperform BERTbase. This could, for example, be due to the differences in the text material used for training (Reddit posts vs. text segments from court transcripts). Nonetheless, HateBERT led to slightly better results when training with the full data set than BERTbase, which might make its use more advantageous for that use case.

6.1.3 Active Learning Implementation. Contrary to our original goal, implementing the AL algorithm did not lead to a clear point of saturation where the model performance could not be improved. This could be explained by the way in which the samples were selected: The actively selected text examples might not have been sufficiently diverse to cover the entire range of possible inputs and could have let to an overfitted model. This effect would even increase if the algorithm did not effectively select the most informative samples. Nonetheless, the improvements in model performance in the fourth AL iteration with a test sample size of $n = 479$ does only slightly increase for the final model trained with $n = 956$ samples in the last iteration.

Consistent with the results from our AL approach, our classification outcome cannot clearly be interpreted in regard to whether a reduced data set size leads to the best model results. However, considering that the best overall performance was reached with the reduced data set, could indicate that a smaller amount of labeled data is sufficient when performing this type of classification task with data from court transcripts. Since overall differences between the model performance were only little, this could be especially useful in application cases where the amount of work needed for manually labeling data is extremely high or data resources are limited.

6.2 Limitations

6.2.1 Data Selection and Annotation. While the GTC in its current form provides a representative sample of genocide-related trials, 90 transcripts cannot capture the full breadth of witness statements in such tribunals. It could therefore be interesting to look at testimonies in other than the 12 selected court cases to validate the representativeness of the GTC.

Important to discuss is also the reduction of text segments to a maximum of about 500 tokens. This was done to ensure an uncomplicated processing of our data through the most common transformer models. However, this means that text paragraphs in which witnesses described their experiences across longer text segments were split (splitting was also applied to statements made by judges and lawyers). While creating an NLP-suitable data set, some context information could therefore be lost.

6.2.2 Label Balance. With an overall proportion of almost 14% out of all witness text segments, the trauma label is clearly imbalanced – which is a common phenomena for data in the context of violent speech [22]. One way of addressing the issue of class imbalance is data augmentation that helps increase the number of unrepresented data in the data set to improve the NLP learning algorithm [15]. Considering that our models performed well compared to classification tasks in similar domains, we did not implement further steps to augment the imbalanced data.

6.2.3 Error Analysis. Taking a deeper look into why certain samples were misclassified by our models, the SHAP values depicted in Figure 5 provide valuable insights. Words shown in the graph contributed significantly to classifying the text snippet *falsely* as traumatic. Especially for the first five words "stole", "kill", "breasts", "bullets", or "killed", it seems logical that the words might be used more often in text segments labeled as traumatic. Concretely, these words were used regularly in the context of violent lootings ("stole"), killings ("kill(ed)", "bullets"), or physical violence directed against women ("breasts"). These words may have been learned as trauma-related by the algorithm as a result.

Some of the misclassifications may further be explained by the fact that only witness statements that were directly expressed by the witness met the criteria for the trauma label. Asking witnesses about trauma-related events or having written statements read out during the trial increases this difficulty further: even though those reports include accounts of trauma, they are not classified as traumatic because the witnesses did not express them verbally.

7 CONCLUSION AND FUTURE RESEARCH

With the Genocide Transcript Corpus, this paper provides a new, extensive language resource for NLP tasks in the field of genocide

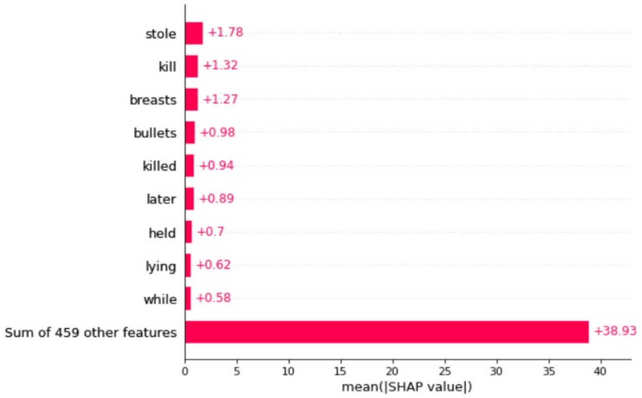


Figure 5: Bar plot of words with the highest SHAP values (i.e. the input tokens' impact) that led to misclassified samples (*trauma* instead of *no trauma*) when classifying with BERT-base.

research that contains 52,845 text segments of which 18,854 (= statements by witnesses) were manually labeled regarding traumatic content. To promote further research with the GTC, we make both data and code openly available (see Section 1). Our results show that trauma can successfully be detected through a binary classification task, indicating that even smaller data set sizes can lead to meaningful results. The XAI approach not only validates the effectiveness of our model in identifying traumatic content, but also demonstrates its capability to distinguish between traumatic and non-traumatic text.

This data set has the potential to serve as a language resource for a variety of research questions in the intersection of genocide and NLP research, building links between both disciplines. Regarding traumatic content, a more qualitatively analysis into the specific linguistic characteristics of trauma-related text could be conducted, allowing for a comparison across the three different tribunals. Further research might also include current cases before the International Criminal Court (ICC) to apply the insights gained from past cases to more recent ones.

From an NLP perspective, next steps could involve further fine-tuning of the transformer models to see how the performance metrics can be improved. The GTC is further suitable for a multi-class classification, focusing on the different speaker roles (e.g., lawyer, witness, judge) and their narratives. Given that the GTC contains information on several meta variables, it could also be very interesting to explore connections between, e.g., the proportion of trauma-related content in witness statements and the final judgment.

By providing an algorithm that automatically detects potentially traumatizing content, this study contributes to the development of techniques for automatic witness support and the improvement of trauma-informed practices in international criminal courts. Our results stress the importance of attaching more relevance to the witness perspective and to adapt legal strategies to reduce their risk of re-traumatization in court.

8 ETHICAL CONSIDERATIONS

Due to the sensitivity of the data, both this paper and the GTC only use information that is publicly available on the respective courts' websites. This applies to the transcripts (and excerpts) referred to in this article and personal information on the accused and witnesses. Witnesses' names and personal information are not disclosed beyond the information that has been published by the courts.

9 ACKNOWLEDGMENTS

Personalized thanks for usage of an annotation tool and to the annotators upon final revision for camera-ready version of the paper.

REFERENCES

[1] Yuki Arase and Jun'ichi Tsujii. 2019. Transfer Fine-Tuning: A BERT Case Study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5393–5404. https://doi.org/10.18653/v1/D19-1542

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[3] American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition.* American Psychiatric Publishing.

[4] Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*. 1–5.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472* (2020).

[7] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4317–4323. https://doi.org/10.18653/v1/P19-1424

[8] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559* (2020).

[9] John D Ciorciari and Anne Heindel. 2016. Victim testimony in international and hybrid criminal courts: Narrative opportunities, challenges, and fair trial demands. *Virginia Journal of International Law* 56 (2016), 265–338.

[10] Jamell Dacon, Harry Shomer, Shaylynn Crum-Dacon, and Jiliang Tang. 2022. Detecting Harmful Online Conversational Content towards LGBTQIA+ Individuals. *arXiv preprint arXiv:2207.10032* (2022).

[11] Marie-Bénédicte Dembour and Emily Haslam. 2004. Silencing hearings? Victim-witnesses at war crimes trials. *European Journal of International Law* 15, 1 (2004), 151–177.

[12] Misha Denil, Alban Demiraj, and Nando De Freitas. 2014. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815* (2014).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7949–7962.

[15] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075* (2021).

[16] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[17] Qiwei He, Bernard P Veldkamp, Cees AW Glas, and Theo de Vries. 2017. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment* 24, 2 (2017), 157–172.

[18] Jenny Hong, Catalin Voss, and Christopher Manning. 2021. Challenges for Information Extraction from Dialogue in Criminal Law. In *Proceedings of the 1st Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Online, 71–81. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.8

[19] Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. 2022. ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5469–5482.

[20] Yoshinobu Kano, Mi-Young Kim, Randy Goebel, and Ken Satoh. 2017. Overview of COLIEE 2017.. In *COLIEE@ ICAIL*. 1–8.

[21] Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2019. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2018 Workshops, JURISIN, AI-Biz, SKL, LENLS, IDAA, Yokohama, Japan, November 12–14, 2018, Revised Selected Papers*. Springer, 177–192.

[22] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. The gab hate corpus: A collection of 27k posts annotated for hate speech. *Language Resources & Evaluation* 56 (2022), 79–108. https://doi.org/10.1007/s10579-021-09569-x

[23] Renana Keydar. 2022. Changing the Lens on Survivor Testimony: Topic Modeling the Eichmann Trial. *BJewish Studies Quarterly* 29, 4 (2022), 412–435.

[24] Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research* 23, 5 (2021), e15708.

[25] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[26] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2020. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364* (2020).

[27] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3 (2018), 31–57.

[28] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[29] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-Hoc Interpretability for Neural NLP: A Survey. *ACM Comput. Surv.* 55, 8, Article 155 (dec 2022), 42 pages. https://doi.org/10.1145/3546577

[30] Charles R Marmar, Adam D Brown, Meng Qian, Eugene Laska, Carole Siegel, Meng Li, Duna Abu-Amara, Andreas Tsiartas, Colleen Richey, Jennifer Smith, et al. 2019. Speech-based markers for posttraumatic stress disorder in US veterans. *Depression and anxiety* 36, 7 (2019), 607–616.

[31] Edoardo Mosca, Katharina Harmann, Tobias Eder, and Georg Groh. 2022. Explaining Neural NLP Models for the Joint Analysis of Open-and-Closed-Ended Survey Answers. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*. Association for Computational Linguistics, Seattle, U.S.A., 49–63. https://doi.org/10.18653/v1/2022.trustnlp-1.5

[32] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4593–4603. https://aclanthology.org/2022.coling-1.406

[33] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv preprint arXiv:2005.07503* (2020).

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[35] Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. Fbert: A neural transformer for identifying offensive content. *arXiv preprint arXiv:2109.05074* (2021).

[36] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR* abs/2211.05100 (2022). https://doi.org/10.48550/arXiv.2211.05100 arXiv:2211.05100

[37] Miriam Schirmer, Udo Kruschwitz, and Gregor Donabauer. 2022. A New Dataset for Topic-Based Paragraph Classification in Genocide-Related Court Transcripts. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. European Language Resources Association, Marseille, France, 4504–4512. https://aclanthology.org/2022.lrec-1.479

[38] Burr Settles. 2009. Active learning literature survey. (2009).

[39] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[40] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. [n. d.]. Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[41] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 3145–3153.

[42] Marie Soueid, Ann Marie Willhoite, and Annie E Sovcik. 2017. The survivor-centered approach to transitional justice: Why a trauma-informed handling of witness testimony is a necessary component. *The George Washington International Law Review* 50 (2017), 125–179.

[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3319–3328.

[44] Kanishk Verma, Tijana Milosevic, Keith Cortis, and Brian Davis. 2022. Benchmarking Language Models for Cyberbullying Identification and Classification from Social-media texts. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*. 26–31.

[45] Julia Viebach. 2018. *Trauma on trial: Survival and witnessing at the International Criminal Tribunal for Rwanda*. Springer, 1011–1030.

[46] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*. New Orleans, Louisiana.

[47] Frank W Weathers and Terence M Keane. 2007. The Criterion A problem revisited: Controversies and challenges in defining and measuring psychological trauma. *Journal of traumatic stress* 20, 2 (2007), 107–121.

[48] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478* (2018).

[49] Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine* 5, 1 (2022), 46.

[50] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting Few-sample BERT Fine-tuning. In *International Conference on Learning Representations*. Vienna, Austria.

[51] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158* (2020).

[52] Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D Ashley, and Matthias Grabmair. 2019. Automatic summarization of legal decisions using iterative masking of predictive sentences. In *Proceedings of the seventeenth international conference on artificial intelligence and law*. 163–172.

Received 12 February 2023; revised tbc; accepted tbc

4.4 Study 4: GENTRAC: A Tool for Tracing Trauma in Genocide and Mass Atrocity Court Transcripts

Authors

Miriam Schirmer, Christian Brechenmacher, Endrit Jashari, Jürgen Pfeffer

In

Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 7666–7671, peer-reviewed, published 05-2024, `https://aclanthology.org/2024.lrec-main.677.pdf`

Abstract

This paper introduces GENTRAC, an open-access web-based tool built to interactively detect and analyze potentially traumatic content in witness statements of genocide and mass atrocity trials. Harnessing recent developments in natural language processing (NLP) to detect trauma, GENTRAC processes and formats court transcripts for NLP analysis through a sophisticated parsing algorithm and detects the likelihood of traumatic content for each speaker segment. The tool visualizes the density of such content throughout a trial day and provides statistics on the overall amount of traumatic content and speaker distribution. Capable of processing transcripts from four prominent international criminal courts, including the International Criminal Court (ICC), GENTRAC's reach is vast, tailored to handle millions of pages of documents from past and future trials. Detecting potentially re-traumatizing examination methods can enhance the development of trauma-informed legal procedures. GENTRAC also serves as a reliable resource for legal, human rights, and other professionals, aiding their comprehension of mass atrocities' emotional toll on survivors.

Contribution of Thesis Author

Theoretical conceptualization, data curation, methodological design, formal analysis, visualization, manuscript writing, revision, and editing.

GENTRAC: A Tool for Tracing Trauma in Genocide and Mass Atrocity Court Transcripts

Miriam Schirmer, Christian Brechenmacher, Endrit Jashari, and Jürgen Pfeffer
School of Social Sciences and Technology
Technical University of Munich, Germany
{miriam.schirmer, christian.brechenmacher, endrit.jashari, juergen.pfeffer }@tum.de

Abstract

This paper introduces GENTRAC, an open-access web-based tool built to interactively detect and analyze potentially traumatic content in witness statements of genocide and mass atrocity trials. Harnessing recent developments in natural language processing (NLP) to detect trauma, GENTRAC processes and formats court transcripts for NLP analysis through a sophisticated parsing algorithm and detects the likelihood of traumatic content for each speaker segment. The tool visualizes the density of such content throughout a trial day and provides statistics on the overall amount of traumatic content and speaker distribution. Capable of processing transcripts from four prominent international criminal courts, including the International Criminal Court (ICC), GENTRAC's reach is vast, tailored to handle millions of pages of documents from past *and future* trials. Detecting potentially re-traumatizing examination methods can enhance the development of trauma-informed legal procedures. GENTRAC also serves as a reliable resource for legal, human rights, and other professionals, aiding their comprehension of mass atrocities' emotional toll on survivors.

Keywords: trauma, genocide, topic-based classification, digital history

1. Introduction

In March 2022, the International Criminal Court (ICC) initiated an investigation into potential war crimes and crimes against humanity in Ukraine. With 17 ongoing investigations and 31 cases, the ICC is responsible for persecuting genocide, war crimes, and crimes against humanity as the world's first permanent international criminal court (International Criminal Court, 2023). Witness accounts play a crucial role in such investigation, often encapsulating traumatic experiences that are revisited later in court settings. Given that re-accounting such events can be emotionally challenging for witnesses and may negatively impact their testimony, it is essential to identify potentially traumatizing content to provide adequate witness support and improve the quality of the testimony at the same time (Soueid et al., 2017).

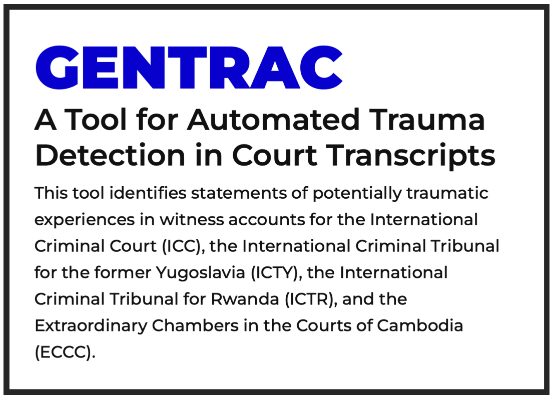
Recognizing the importance of accurately identifying such trauma and harnessing the advancements in NLP, we introduce the Genocide Trauma Tracing Tool "GENTRAC" – a tool designed to automatically detect potentially traumatic content in witness statements of international criminal courts. Utilizing a publicly available, BERT-based model for trauma detection in the context of genocide (Schirmer et al., 2023a), our tool identifies potential trauma in transcripts from a total of 187 cases before the ICC, the International Criminal Tribunal for the former Yugoslavia (ICTY), the International Criminal Tribunal for Rwanda (ICTR), and the Extraordinary Chambers in the Courts of Cambodia (ECCC). This includes cases that have been concluded and ongoing cases (as of 2023). Notably, *future cases* that will be heard before the ICC can also be analyzed by this tool.

The sheer volume of about 2.5 million pages of transcripts originating solely from the ICTY underscores GENTRAC's expansive applicability. While the ICTY, the ICTR, and the ECCC were limited to addressing regionally specific atrocities, the ICC was established as a perpetual legal institution to handle any future mass atrocities worldwide. As of 2023, the ICC is adjudicating 31 cases. Being the sole international court tasked with addressing genocide and mass atrocities on a global scale, it is probable that this caseload will keep expanding, consequently increasing the volume of documents that GENTRAC can process.

Our web-based interactive tool serves a dual purpose:

- First, it employs an advanced parsing algorithm to process and structure court transcripts of the ICC, the ICTY, the ICTR, and the ECCC, enabling subsequent NLP analysis.
- Second, using the parsed transcript output, GENTRAC runs a BERT (Bidirectional Encoder Representations from Transformers; (Devlin et al., 2019))-based binary classifier to detect traumatic content in each segment of the witness' statement and provides information on trauma density and speech proportions of individual speakers in the transcript.

Through this process, the tool streamlines access to historical cases of mass atrocities, like those addressed by the ICTY, for future research. Simultaneously, it gives insights into contemporary



GENTRAC

A Tool for Automated Trauma Detection in Court Transcripts

This tool identifies statements of potentially traumatic experiences in witness accounts for the International Criminal Court (ICC), the International Criminal Tribunal for the former Yugoslavia (ICTY), the International Criminal Tribunal for Rwanda (ICTR), and the Extraordinary Chambers in the Courts of Cambodia (ECCC).

Figure 1: Section of the landing page: general information.

and future cases before the ICC, extending its relevance to situations in, for instance, Ukraine, Darfur, and Afghanistan. GENTRAC is a publicly available[1], web-based, interactive tool, making it accessible to researchers, legal and human rights professionals, and anyone interested in examining trauma in the context of mass atrocities.

2. Detecting Trauma

Psychological trauma, as defined by the American Psychological Association (APA), encompasses experiences of "exposure to actual or threatened death, serious injury, or sexual violence," whether directly encountered or witnessed. It also includes instances where individuals "learn that the traumatic event(s) occurred to a close family member or close friend" (American Psychiatric Association, 2013). Within the context of international criminal courts tasked with the legal proceedings of mass atrocities and genocide, trauma becomes a recurring phenomenon in testimonies. This is notably the case when witnesses recount harrowing episodes of violence, such as personal experiences of torture or witnessing large-scale massacres.

2.1. NLP for Trauma Detection

Considering the variety of traumatic experiences with their subjective nature, their detection in text material is complex. Despite these challenges, recent research has shown how NLP methods can improve the detection of psychological disorders or adapt treatment (Ahmed et al., 2022; Le Glaz et al., 2021; Zhang et al., 2022). In trauma research, progress is being made in analyzing patient narratives (He et al., 2017) and identifying cases of post-traumatic stress disorder (PTSD) through speech (Marmar et al., 2019). Specifically within genocide research, researchers have developed a model to identify potentially traumatic content within witness statements from international criminal courts, including a manually labeled dataset from three genocide tribunals (Schirmer et al., 2023a, 2022) or an in-depth mixed method analysis of witnesses talking about torture in court (Schirmer et al., 2023b).

While NLP methods have become prevalent in psychological diagnosis and applications, they have not yet been integrated into a tool specifically designed for processing court transcripts. So far, tools in this context have primarily focused on transcription challenges (e.g., Downey, 2006; Saadany et al., 2022). A tool that brings together insights from trauma research and legal AI in international criminal courts does not exist so far.

2.2. Defining the Scope of Trauma Detection

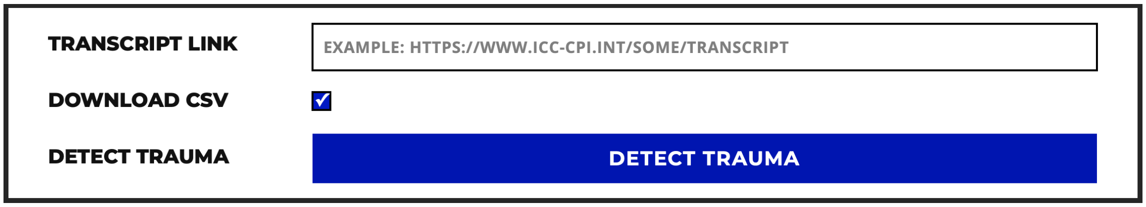
GENTRAC does not intend to provide psychological diagnoses or comprehensively grasp intricate psychological phenomena like trauma. Drawing conclusions about the mental health of witnesses from court transcripts alone is impossible without additional information about their psychological well-being or previous diagnoses. Therefore, this study concentrates on witness accounts describing events categorized as traumatic but doesn't assume that these events have necessarily resulted in a traumatic response.

The tool specifically aims to pinpoint instances that meet the APA's definition of trauma. This encompasses not just events that are merely frightening, but those with a substantial likelihood of leading to psychological distress or trauma. By closely following the APA's criteria, we aim to minimize subjectivity in determining what constitutes trauma. Consequently, statements that GENTRAC classifies as positive include those where witnesses mention events that could be traumatizing, focusing on the exposure to potentially traumatic events rather than the psychological trauma that could arise from recounting such distressing experiences.

3. General Page Setup and Functionality

GENTRAC offers an intuitive and interactive approach to trauma detection in court transcripts through a publicly accessible web page. Upon accessing the tool's landing page, users can insert a link to any transcript from one of the four international courts (ICC, ICTR, ICTY, and ECCC), including any future trials before the ICC. Clicking the "Detect Trauma" button initiates the parsing of the document. In response, GENTRAC generates a CSV file, available for optional download. This

[1] https://gentrac.tox.report/



TRANSCRIPT LINK EXAMPLE: HTTPS://WWW.ICC-CPI.INT/SOME/TRANSCRIPT

DOWNLOAD CSV ☑

DETECT TRAUMA

Figure 2: Section of the landing page: URL input and processing.

Speaker	Role	Statement	Exam	Trauma
Mr. President	Presiding Judge	Regarding your personal matte..	True	-
Mr. Vann Nath	Witness	Mr. President, on the day I was..	True	0.95
Mr. President	Presiding Judge	Thank you for your information ..	True	-
Mr. President	Presiding Judge	Next, I would like to inquire if an.	False	-
Judge Lavergne	Judge	Could you explain to us the diffi..	True	-

Table 1: Preview of the segmented transcript available for download as a CSV file (see Section 4).

file includes segmented transcripts, accompanied by annotations that signal the potential presence of trauma. This format ensures users have a clear and structured output, facilitating further research, analysis, or legal examination.

On the bottom of the page, users find further information about the mechanisms behind the tool. This includes information on how we define trauma, technical details on the classification model and links to further resources.

3.1. Document Preprocessing and Parsing

Initiating the process, the tool first engages in a segmentation phase. We have developed a multi-level parser architecture designed explicitly for ICC, ICTY, ICTR, and ECCC transcripts. Those transcripts are fetched from URLs or document uploads and converted into machine-readable formats. The primary objective during this phase is to dissect the entire transcript, transforming a collection of character sequences into distinct statements and assigning these to speakers, thereby differentiating statements based on the speaker's role.

Given that all courts rely on different transcript templates and file formats, each parser is tailored to a specific court, ensuring precise information extraction. For instance, if a witness is labeled simply as "The Witness" later on in the transcript, our tool can match this information with names provided in previous sections of the transcript, remembering the witness's name for further occurrences. Similarly, GENTRAC can distinguish between lawyers and judges, appending their names to the appropriate speaker segment.

Users can download this processed transcript version as a machine-readable CSV file. Besides details on the speaker, their role and the actual statement, the file includes context information, like marking ongoing witness examinations. Finally, the last column presents the likelihood for a positive trauma classification, e.g. 0.95 in the example provided in Table 1 (see Section 4 for more detail).

3.2. BERT-Based Trauma Detection

After the segmentation phase, GENTRAC proceeds to the task of trauma detection. For this process, we utilize a publicly available model for trauma detection in the context of genocide (Schirmer et al., 2023a). Trained using BERT-base-uncased (Devlin et al., 2019) on a dataset of over 18,000 witness statement segments, the model was benchmarked against a human baseline. Manual labeling was performed by three trained psychologists. These professionals adhered strictly to APA definitions, aiming to minimize subjectivity.

The model demonstrated robust performance, achieving an F1 score of 0.89 and an accuracy of 0.95. These metrics suggest that the model offers a dependable approach for trauma detection in court transcripts. The model demonstrates efficient inference capabilities when deployed on a CPU, allowing us to minimize latency within our system.

3.3. Visualizations and Statistics

Upon processing the input document, GENTRAC's interface provides users with a multifaceted view of the results:

1. **Trauma Quantification in Witness Statements:** GENTRAC showcases the cumulative count of segments within witness statements that are identified as potentially traumatic as its primary purpose (Figure 4).

2. **Trauma Evolution Overview:** The tool offers a visual representation of the unfolding of traumatic content throughout the document, enabling users to track the ups and downs of such content during a hearing day (Figure 3).

3. **CSV Preview:** A snapshot of the CSV file is presented, allowing users to get an overview of the structure and content of the segmented transcript (Table 1).

4. **Speaker Analysis:** The tool offers insights into transcript contributors, enabling users to identify each speaker's role (e.g., lawyer or judge) and their relative discourse contribution (Figure 4).

This comprehensive display ensures that users can quickly grasp the document's essence, speaker dynamics, and, most critically, the prevalence of traumatic content within it.

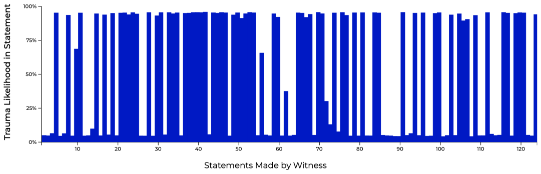


Figure 3: Progression of trauma-related witness statements as displayed on the website.

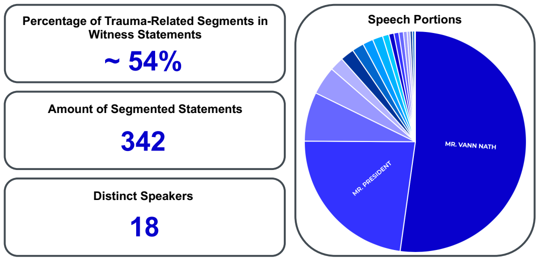
4. Example Case

To illustrate the functionality of GENTRAC, we demonstrate results of a transcript from the ECCC, featuring the testimony of Vann Nath.[2] He is one of the few survivors of the S-21 torture prison during the Khmer Rouge Regime in Cambodia and has released an autobiography and spoken publicly to educate others about his experiences and raise awareness (Chandler, 2023; Nath and Nariddh, 1998).

Upon processing the document, a preview of the segmented transcripts is presented, including the individual speakers, their roles, and the actual statement. The preview further includes whether the statement is part of a witness examination and the likelihood that a witness statement contains traumatic content. Table 1 displays five speech contributions during the examination of Vann Nath. The tool extracts speaker names, as referenced in the transcript, and assigns respective roles (e.g., "Presiding Judge"). Table 1 further illustrates that the trauma classification is only applied to witness segments and thus does not apply to statements made by judges and lawyers.

The user is further presented a visualization depicting the progression of traumatic content experienced by the respective witness throughout the document. Typically, this content corresponds to a single day of court hearings. Figure 3 shows the density of traumatic content in our sample case. In this analysis, we observe a distinct pattern: a small amount of detected trauma at the outset, followed by a significant increase in the first half, a subsequent intermission, and finally, a minor resurgence. This pattern aligns closely with the typical structure of a day in court, where proceedings commence with the exchange of personal information, progress to probing questions regarding the individual's experiences during imprisonment, and culminate in a cross-examination phase towards the end.

Transcript statistics reveal further details about the processed document. In the example case, trauma-related statements make up approximately 54% of the statement segments of the witness. We also receive information on the total amount of unique speaker segments (342) and the number of distinct speakers (18) (see Figure 4).


Percentage of Trauma-Related Segments in Witness Statements

~ 54%

Amount of Segmented Statements

342

Distinct Speakers

18

Speech Portions

Figure 4: Examples of transcript statistics displayed by the tool.

In this instance, nearly half of the witness statements cover potentially traumatic accounts. This is notably higher than previous findings, which indicated that roughly 14% of witness statements from various genocide tribunals contained trauma-related content (Schirmer et al., 2023a). The heightened trauma percentage in Vann Nath's testimony could stem from his prolonged imprisonment and consistent exposure to torture and death at the S-21 prison. An example of a text segment classified as potentially traumatic can be seen in Table 2. The tool further shows a list of these speakers and their share of statements made; e.g., in this case, the main proportion of speaker segments is made by the witness (~53%), followed by the presiding judge (~23%).

"And when we were allowed to do exercise, our legs were still shackled to the metal bars and we could like hop to do exercise. If we didn't hop then they would beat us also."

Table 2: Example of a witness statement that was classified as potentially traumatic (excerpt).

[2] The transcript is available at `https://www.eccc.gov.kh/en/witness-expert-civil-party/mr-vann-nath`

5. Discussion

5.1. Contribution

GENTRAC provides a web-based, publicly accessible tool for detecting and analyzing potentially traumatic content in witness statements from genocide and mass atrocity trials. This is significant for various reasons. Firstly, it employs state-of-the-art NLP techniques to identify trauma, which is essential for comprehending mass atrocities' emotional and psychological impact on survivors and victims. Secondly, GENTRAC aids in the identification of potentially re-traumatizing examination methods, contributing to the development of more trauma-informed legal procedures that ensure sensitive witness treatment (Soueid et al., 2017). This is enhanced through GENTRAC's capability to visualize traumatic content throughout a hearing day and to provide corresponding transcript statistics. Thus, GENTRAC aids human rights and legal professionals, including judges, lawyers, and prosecutors, in gaining insights into witness statements' emotional context and potentially adjusting their approaches accordingly. Additionally, GENTRAC's compatibility with transcripts from prominent international criminal courts, such as the ICC, makes it a valuable resource for analyzing a wide range of cases. This relevance extends to past, ongoing, *and future* ICC proceedings, ensuring its enduring utility.

5.2. Future Work

GENTRAC's functionality is currently tailored to the existing structure of the respective court transcript formats. Should there be any changes in the transcript structure, the parsers will require updates and adjustments to maintain their effectiveness. This is particularly relevant for the ICC, which stands out as the only court expected to conduct future trials and thereby continuously generate new transcripts for analysis with GENTRAC. This contrasts with other tribunals, where proceedings have largely concluded. This ongoing process ensures GENTRAC remains fully operational.

Next steps will encompass user studies to improve the user interface and to add statistical features and contextual insights about the trial to give users a more comprehensive understanding of the text material. While the tool is interactive in that users can choose which transcript to analyze, it may increase the user experience to allow the selection of specific transcript statistics.

Considering the complex nature of trauma, relying solely on a binary classifier to identify relevant text segments presents a notable limitation. This approach may oversimplify the nuanced expressions of trauma, potentially overlooking critical subtleties in the text. Enhancing the model to recognize a broader spectrum of trauma-related expressions could significantly improve its accuracy and sensitivity, thereby offering a more comprehensive analysis of traumatic content.

Lastly, we are exploring the adaptation of GENTRAC to include a broader range of text sources that contain traumatic content. This expansion aims not only to incorporate transcripts and documents from various legal forums, enhancing computational analyses with results from other scholars in that area (Hawes et al., 2009; Keydar et al., 2022; Keydar, 2020), but also to extend beyond legal texts. The inclusion of materials unrelated to court proceedings, such as personal narratives, social media posts, and journalistic accounts, could significantly enrich our understanding and detection of trauma across different contexts.

6. Ethical Considerations

Given the sensitivity of the data, both GENTRAC and this paper exclusively utilize publicly available information from the respective courts' websites. This encompasses the transcripts mentioned in this article and personal information about witnesses. It's important to note that the names and personal details of witnesses are handled with utmost discretion and are not disclosed beyond what has been officially published by the courts.

7. Bibliographical References

Arfan Ahmed, Sarah Aziz, Carla T Toro, Mahmood Alzubaidi, Sara Irshaidat, Hashem Abu Serhan, Alaa A Abd-Alrazaq, and Mowafa Househ. 2022. Machine learning models to detect anxiety and depression through social media: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2:100066.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition*. American Psychiatric Publishing.

David Chandler. 2023. *Voices from S-21: Terror and history in Pol Pot's secret prison*. University of California Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Greg Downey. 2006. Constructing" computer-compatible" stenographers: The transition to real-time transcription in courtroom reporting. *Technology and Culture*, 47(1):1–26.

Timothy Hawes, Jimmy Lin, and Philip Resnik. 2009. Elements of a computational model for multi-party discourse: The turn-taking behavior of supreme court justices. *Journal of the American Society for Information Science and Technology*, 60(8):1607–1615.

Qiwei He, Bernard P Veldkamp, Cees AW Glas, and Theo de Vries. 2017. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment*, 24(2):157–172.

International Criminal Court. 2023. About the court.

Renana Keydar. 2020. Changing the lens on survivor testimony: Topic modeling the eichmann trial. *Journal of Law, Technology & Policy*, (1):55–84.

Renana Keydar, Yael Litmanovitz, Badi Hasisi, and Yoav Kan-Tor. 2022. Modeling repressive policing: Computational analysis of protocols from the israeli state commission of inquiry into the october 2000 events. *Law & Social Inquiry*, 47(4):1075–1105.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

Charles R Marmar, Adam D Brown, Meng Qian, Eugene Laska, Carole Siegel, Meng Li, Duna Abu-Amara, Andreas Tsiartas, Colleen Richey, Jennifer Smith, et al. 2019. Speech-based markers for posttraumatic stress disorder in us veterans. *Depression and Anxiety*, 36(7):607–616.

Vann Nath and Moeun Chhean Nariddh. 1998. *A Cambodian Prison Portrait: One Year in the Khmer Rouge's S-21*. White Lotus.

Hadeel Saadany, Constantin Orăsan, and Catherine Breslin. 2022. Better transcription of uk supreme court hearings. *arXiv preprint arXiv:2211.17094*.

Miriam Schirmer, Udo Kruschwitz, and Gregor Donabauer. 2022. A new dataset for topic-based paragraph classification in genocide-related court transcripts. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4504–4512, Marseille, France. European Language Resources Association.

Miriam Schirmer, Isaac Misael Olguín Nolasco, Edoardo Mosca, Shanshan Xu, and Jürgen Pfeffer. 2023a. Uncovering trauma in genocide tribunals: An nlp approach using the genocide transcript corpus. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 257–266.

Miriam Schirmer, Jürgen Pfeffer, and Sven Hilbert. 2023b. Talking about torture: A novel approach to the mixed methods analysis of genocide-related witness statements in the khmer rouge tribunal. *Journal of Mixed Methods Research*, page 15586898231218463.

Marie Soueid, Ann Marie Willhoite, and Annie E Sovcik. 2017. The survivor-centered approach to transitional justice: Why a trauma-informed handling of witness testimony is a necessary component. *The George Washington International Law Review*, 50:125–179.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digital Medicine*, 5(1):46.

4.5 Study 5: The Language of Trauma: Modeling Traumatic Event Descriptions Across Domains with Explainable AI

Authors

Miriam Schirmer, Tobias Leemann, Gjergji Kasneci, Jürgen Pfeffer, David Jurgens

In

ArXiv, `https://doi.org/10.48550/arXiv.2408.05977`, published 08-2024

Abstract

Psychological trauma can manifest following various distressing events and is captured in diverse online contexts. However, studies traditionally focus on a single aspect of trauma, often neglecting the transferability of findings across different scenarios. We address this gap by training various language models with progressing complexity on trauma-related datasets, including genocide-related court data, a Reddit dataset on post-traumatic stress disorder (PTSD), counseling conversations, and Incel forum posts. Our results show that the fine-tuned RoBERTa model excels in predicting traumatic events across domains, slightly outperforming large language models like GPT-4. Additionally, SLALOM-feature scores and conceptual explanations effectively differentiate and cluster trauma-related language, highlighting different trauma aspects and identifying sexual abuse and experiences related to death as a common traumatic event across all datasets. This transferability is crucial as it allows for the development of tools to enhance trauma detection and intervention in diverse populations and settings.

Contribution of Thesis Author

Theoretical conceptualization, data curation, methodological design, formal analysis, visualization, manuscript writing, revision, and editing.

The Language of Trauma: Modeling Traumatic Event Descriptions Across Domains with Explainable AI

Miriam Schirmer[1], Tobias Leemann[1], Gjergji Kasneci[1], Jürgen Pfeffer[1], David Jurgens[2]
[1]Technical University of Munich
[2]University of Michigan

Abstract

Psychological trauma can manifest following various distressing events and is captured in diverse online contexts. However, studies traditionally focus on a single aspect of trauma, often neglecting the transferability of findings across different scenarios. We address this gap by training language models with progressing complexity on trauma-related datasets, including genocide-related court data, a Reddit dataset on post-traumatic stress disorder (PTSD), counseling conversations, and Incel forum posts. Our results show that the fine-tuned RoBERTa model excels in predicting traumatic events across domains, slightly outperforming large language models like GPT-4. Additionally, SLALOM-feature scores and conceptual explanations effectively differentiate and cluster trauma-related language, highlighting different trauma aspects and identifying sexual abuse and experiences related to death as a common traumatic event across all datasets. This transferability is crucial as it allows for the development of tools to enhance trauma detection and intervention in diverse populations and settings.

1 Introduction

Post-Traumatic Stress Disorder (PTSD) is a significant mental health condition that can develop after experiencing a traumatic event. For an event to potentially lead to PTSD, it must involve actual or threatened death, serious injury, or a threat to one's physical integrity, causing intense fear, helplessness, or horror (Friedman et al., 2007; Gold, 2017). Although about 70% of Americans will encounter such traumatic events in their lifetime, only about 5-7% develop PTSD, highlighting that PTSD is relatively rare despite high trauma exposure. However, this figure could be higher, as many cases may go undiagnosed (Bonn-Miller et al., 2022; Atwoli et al., 2015).

This discrepancy suggests that various factors, including psychological resilience, the nature of the trauma, and access to mental health support, influence the development of PTSD. Definitions of trauma and responses to it can vary widely across cultures and social contexts, affecting the prevalence and expression of PTSD.

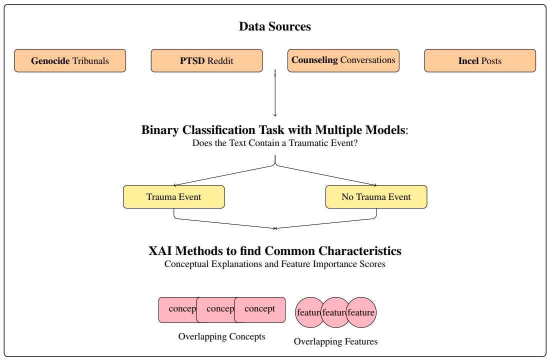

Data Sources
Genocide Tribunals
PTSD Reddit
Counseling Conversations
Incel Posts
Binary Classification Task with Multiple Models:
Does the Text Contain a Traumatic Event?
Trauma Event
No Trauma Event
XAI Methods to find Common Characteristics
Conceptual Explanations and Feature Importance Scores
concep
concep
concept
featur
featur
feature
Overlapping Concepts
Overlapping Features

Figure 1: We (1) create a cross-domain trauma dataset, (2) classify traumatic events with models of different complexity, and (3) use XAI methods to identify overlapping characteristics of traumatic events.

To investigate the interplay of these factors, we are proposing a Natural Language Processing (NLP) approach to identify traumatic events across different domains. Understanding the cross-cutting mechanisms of trauma is crucial for developing comprehensive support systems and interventions that are adaptable to various contexts. We are following up on these research questions:

RQ1: Given the diverse forms of trauma, what are the most effective methods for modeling and predicting its manifestations?

RQ2: How transferable is the detection of multifaceted traumatic events across domains?

RQ3: What are the cross-cutting mechanisms related to trauma that can be identified across different types and contexts of traumatic events?

Our work advances trauma detection by applying NLP and XAI methods to offer detailed insights not yet explored in the literature. We contribute by: (1)

identifying key trauma concepts from psychological literature and replicating them using NLP methods, (2) modeling traumatic event detection with various language models and creating a dataset that includes genocide court transcripts, PTSD-related Reddit posts, counseling conversations, and "Involuntary Celibates" Incel forum posts, (3) developing a three-stage XAI framework that approximates Shapley values, assesses feature importance, and identifies task-relevant concepts, providing a comprehensive understanding of trauma at both the instance and dataset levels, and (4) automating trauma detection to enhance online psychological support by displaying hotline information and resources in forums where trauma is frequently discussed.

2 Traumatic Events & Language

2.1 Definition & Scope

Psychological trauma, as defined by the American Psychological Association (APA), encompasses experiences of "exposure to actual or threatened death, serious injury, or sexual violence," whether directly encountered or witnessed. This includes instances where individuals "learn that the traumatic event(s) occurred to a close family member or close friend" (American Psychiatric Association, 2013).

While psychological trauma and PTSD are frequently discussed in the context of childhood abuse and the military, trauma can manifest in a variety of situations (Van der Kolk, 2003; Yehuda, 1998). It can arise in interpersonal violence like domestic abuse and sexual assault; and accidents or natural disasters. Trauma can also result from medical issues, bereavement and loss, emotional and psychological abuse, and its manifestation can vary depending on cultural beliefs and values (Smelser et al., 2004).

2.2 Trauma Contexts & Categorization

Within the psychological literature, key events have been identified that are typical for specific trauma contexts. In armed conflict and mass atrocities, exposure to severe violence and death is prevalent. This often includes the death of close family members, forced displacement, and sexual abuse (Powell et al., 2003). For instance, Dyregrov et al. (2000) found that most child survivors of the Rwandan genocide had witnessed severe injuries and deaths, with more than half witnessing massacres.

In domestic trauma, the most common forms are physical abuse (e.g., intimate partner violence), emotional abuse, and neglect (McCloskey and Walker, 2000). Emotional abuse is particularly hard to detect due to its subtle nature, including consistent belittling, criticizing, or bullying (Dye, 2020; Idsoe et al., 2021). Sexual violence, whether in war or domestic contexts, is an especially devastating form of trauma (Kiser et al., 1991). This includes childhood sexual abuse, rape, and exploitation.

The range of traumatic events makes conceptualizations of trauma complex. Researchers have categorized trauma in line with diagnostic manuals like the Diagnostic and Statistical Manual of Mental Disorders (DSM) into types such as assaultive violence (e.g., military combat, rape, threats with weapons), other injuries or shocking events (e.g., serious car accidents and life-threatening illnesses) (Breslau et al., 2004). Identifying these events is crucial, as most subsequent issues are linked to the initial trauma due to the development of trauma-specific fears in PTSD (Terr, 2003).

2.3 NLP for Trauma Detection

Given the variety and subjective nature of traumatic experiences, detecting them in text is complex. Despite these challenges, recent research has shown that NLP methods can improve the detection of psychological disorders and aid in treatment adaptation (Ahmed et al., 2022; De Choudhury and De, 2014; Le Glaz et al., 2021; Malgaroli et al., 2023; Zhang et al., 2022).

NLP and Mental Health. Major areas in this field include promoting better health and early disorder identification for intervention (Calvo et al., 2017; Swaminathan et al., 2023). For example, Levis et al. (2021) associated linguistic markers from psychotherapist notes with treatment duration. Analyzing mental health chat conversations, Hornstein et al. (2024) found that words indicating younger age and female gender were associated with a higher chance of re-contacting.

Recently, the use of Large Language Models (LLMs) has led to the development of specific models for mental health applications (Xu et al., 2024; Yang et al., 2024). While LLMs effectively detect mental health issues and provide eHealth services, their clinical use poses risks, such as the lack of expert-annotated multilingual datasets, interpretability challenges, and issues regarding data privacy and over-reliance (Guo et al., 2024).

Specifically for social media data, there has been research on using sentiment analysis and semantic structures to detect anxiety (Low et al., 2020) or depression (Tejaswini et al., 2024) on Reddit posts. In suicide prevention on social media, Sawhney et al. (2020) developed a superior model for suicidal risk screening that identifies emotional and temporal cues, outperforming competitive methods (c.f., Ji (2022) on suicidal risk detection).

Trauma Detection. In trauma research, progress is being made in analyzing patient narratives (He et al., 2017) and identifying cases of post-traumatic stress disorder (PTSD) through speech (Marmar et al., 2019). Miranda et al. (2024) developed an NLP workflow using a pre-trained transformer-based model to analyze clinical notes of PTSD patients, revealing consistent reductions in trauma criteria post-psychotherapy. Disruptions in lexical characteristics and emotional valence have been found to contribute to identifying PTSD (Quillivic et al., 2024). Using Twitter data, Ul Alam and Kapadia (2020) investigated whether posts can complete clinical PTSD assessments, achieving promising accuracy in PTSD classification and intensity estimation validated with veteran Twitter users (cf. Coppersmith et al. (2014); Reece et al. (2017)).

2.4 Trauma Event Detection in this Study

Previous work has identified language markers of PTSD, such as overuse of first-person singular pronouns, increased use of words related to depression, anxiety, and death, and more negative emotions. However, these markers are not specific to trauma and can also be associated with other psychological disorders, complicating accurate identification. Additionally, the transferability of detection methods is often lacking (Coppersmith et al., 2014; Quillivic et al., 2024).

Trauma detection in NLP is distinct in that it involves identifying a specific traumatic event that precedes a PTSD diagnosis, unlike the detection of depression or anxiety, which do not require a concrete event in their definitions. This study focuses on detecting such events in online resources, avoiding symptom or diagnosis analysis. Drawing conclusions about mental health from public text data alone is impossible without additional psychological information. We aim to identify instances meeting the APA's definition of trauma, minimizing subjectivity by closely following their criteria.

3 Data & Labeling

3.1 Data Sources

Our final dataset is built from four datasets, each offering unique perspectives on traumatic experiences (Table 1) to identify common characteristics of trauma that extend beyond specific events, such as those related to war: The Genocide Court Transcripts (GTC; Schirmer et al., 2023a) dataset comprises text from genocide tribunals, providing insights into severe human rights violations and the profound trauma experienced by victims and witnesses. This encompasses 90 cases across the International Criminal Tribunal for Rwanda, the International Criminal Tribunal for the former Yugoslavia, and the Extraordinary Chambers in the Courts of Cambodia. The Reddit PTSD Dataset includes posts from the PTSD subreddit of the Reddit Mental Health Dataset (Low et al., 2020), where individuals discuss their experiences with post-traumatic stress disorder, sharing personal stories and support. The Mental Health Counseling Conversations Dataset (Amod, 2024) features questions and answers sourced from online counseling and therapy platforms. The questions cover a wide range of mental health topics, and qualified psychologists provide the answers.

The Incel Posts Dataset (Matter et al., 2024) contains posts from Incel community forums and reflects extreme misogynistic viewpoints. This dataset serves as a control in our study: Though not explicitly trauma-related, it includes posts on depression, bullying, and violence directed towards women. The violent and aggressive language in this dataset helps quantify our models' ability to distinguish explicit trauma from related emotional distress.

3.2 The Trauma Event Dataset TRACE

We present the final trauma event dataset TRACE (Trauma Event Recognition Across Contextual Environments). To that end, all source datasets were pre-processed to ensure comparability for the detection task, including the removal of URLs and standardization of formatting. Due to their varied origins, the samples from each dataset differ in size, with instances ranging from single-word sentences to more elaborate descriptions of events and personal thoughts across all datasets. For compatibility with the BERT-architecture, we split instances exceeding the 512-token limit into smaller segments. Our approach treats each segment as independent,

Dataset	Description	Size & Balance	AA
Genocide Transcript Corpus (GTC)	Witness statements from 90 different cases across three different genocide tribunals.	15,845 samples (trauma: 13.54%)	n/a
PTSD Subreddit (PTSD)	Post-Traumatic Stress Disorder (PTSD) subset of the Reddit Mental Health Dataset.	1,200 samples (trauma: 47.19%)	(1) $\alpha = .63$ (2) $F1 = .77$
Counseling Dataset	Queries submitted by users seeking advice, with answers provided by professionals.	1,200 samples (trauma: 8.16%)	(1) $\alpha = .69$ (2) $F1 = .95$
Incel Dataset	Posts from the Incel online forum *incels.is*.	300 samples (trauma: 2.67%)	(1) $\alpha = .43$ (2) $F1 = .78$

Table 1: Dataset Overview. Note: Annotator agreement (AA) was calculated (1) among crowd workers (Krippendorff's α) and (2) for the crowd worker majority vote vs. the expert vote (Binary F1).

with trauma classification based solely on its content. While some segments from the same text may appear in both training and test sets, we consider label leakage minimal, since the model must rely on the segment's content for accurate prediction. 7-20% (depending on the dataset) of segments were split overall.

Our study aims to demonstrate cross-domain transferability on realistic data, making it crucial to use datasets with their expected class distribution, even if they differ in context and trauma event rates. We matched the size of all datasets to the Counseling Dataset, which had the fewest samples and the most significant class imbalance. Despite these constraints, the Counseling Dataset remains highly valuable for its unique perspective on online mental health conversations, particularly in seeking expert advice.

Annotation Process. The GTC already contains a binary trauma variable that psychologists have annotated according to the APA definition of trauma. For the PTSD and Counseling datasets, 1,200 instances each were annotated by crowdworkers. We used the Portable Text Annotation Tool (Potato; Pei et al., 2022) to set up an annotation interface for crowdworkers using Prolific as a recruitment platform for annotators. Each instance was labeled by three annotators, and all annotators received an hourly reimbursement of approximately 12 US$. The crowdworkers were provided detailed instructions, the APA definition of a traumatic event, and three examples. Both the Prolific pre-screening and the instructions contained a trigger warning, ensuring that participants were free to pause or stop the study at any time (Appendix A, Figure 6). Annotators were based in either the US or the UK and fulfilled English language requirements.

We conducted a pilot study comparing single-choice and span annotation setups, where participants highlighted traumatic events in the text. The final annotation task used the span setup to ensure accurate detection (Appendix A, Figure 7). Annotations were quality-checked, resulting in the removal of two annotator entries who labeled an unlikely number of samples as trauma, without affecting the total sample count (e.g., 1,200). For the Incel dataset, we only labeled 300 instances since it serves as a control test set. To ensure quality, two researchers with psychology degrees annotated a subset of 200 instances from each dataset and resolved disagreements through discussion (Cohen's $\kappa = .82$).

Annotator Agreement. To assess annotator consistency, we report Krippendorff's α for agreement among crowdworkers and provide Binary F1 scores to measure agreement between the crowdworker majority vote and the expert vote, with the latter serving as the 'true' reference (Table 1). Both agreements were best for the Counseling Dataset. All agreement scores indicate at least moderate agreement (Krippendorff, 2018). Despite variability, our primary focus is on the accuracy of labels from majority voting. The moderate F1 scores indicate that majority votes are reliable labels, supporting the robustness of our annotation process. Given the subjective nature of interpreting trauma-related constructs, some disagreement is expected, similar to lower agreement seen in tasks like hate speech detection (Li et al., 2024). This level of agreement, while not perfect, provides a solid foundation for the study.

4 Methods

4.1 Models and Hyperparameters

In this work, we implement five sequence classification models for natural language inputs. The suitability of the these models for trauma detection

Model	Complexity	Interpretability	Hyperparameters	Scalability	Prediction
BoW-Naive-Bayes	Low	High	binary, smoothing param. α	High	After training
N-Gram Logistic Regression	Low	Medium	TF-IDF, n-grams	High	After training
TF-IDF Fully-Connected NN	Medium	Medium	Hidden layers, layer width	Medium	After training
BERT-based Models	High	Low	Learning rate, layers, heads	Low	One-shot or after fine-tuning
Black-box API (GPT-3.5/4)	High	Low	Prompt template, API settings	Low	One-shot or after fine-tuning

Table 2: Model Categorization According to General Suitability Criteria

in different contexts is defined by criteria such as complexity, interpretability, hyperparameter optimization, and scalability. To help in understanding the trade-offs and strengths of each approach, we provide an overview of the models considered in Table 2. The hyperparameters given are optimized with a hyperparameter optimization framework.

BoW-Naive-Bayes Model. The simplest model is obtained by fitting a Naive-Bayes model on the word counts in both classes. Let $\mathbf{t} = [t_1, t_2, \ldots, t_N]$ be an input sequence. We model the log-odds by combining two key components. First, we calculate the prior odds, which is the log of the initial ratio of the probabilities of the two categories. Second, we add the word-specific weights, which are summed over all elements in the input sequence. Each weight represents the log of the ratio of the probabilities of that element occurring in each category.

We obtain the weight of a term by counting its occurrences in documents from both classes and applying Laplace smoothing with a specified hyperparameter α. The main advantage of this linear model is its interpretability due to the individual weights of each token that are explicitly computed.

N-Gram Logistic Regression Model. We compute n-grams for the datasets and fit a logistic regression model on the TF-IDF represenation of the n-Grams, where n is $[1, 2, 3]$.

TF-IDF Fully-Connected Model. Furthermore, we compute TF-IDF vectors for the samples and train a fully connected neural network using this representation as an input. We use either one or two hidden layers, with the number of hidden layers and their width as a hyperparameter.

BERT-based Models. We train the popular encoder-only transformer models BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). We experiment with both pretrained and non-pretrained versions of these models. We find that the pretrained models yield superior performance, which is why we restrict our analysis to these models for the main paper. We use the learning rate, number of layers, and number of heads as hyperparameters.

Black-box API models (GPT-3.5/GPT-4). We use a prompt template to access publicly available foundation model APIs for GPT-3.5 and GPT-4 (Achiam et al., 2023). We rephrase the classification tasks as a sequence completion tasks by using prompt template, which instructs the model to either output "0" or "1", and apply basic prompt engineering, including a task definition, the trauma definition, and labeling instructions (see Appendix A.2). We use the top token log-probabilities returned by the API to compute class log-odds, which can be used to compute calibration measures and ROC curves.

4.2 Explainable AI Methods

We use explainable AI approaches to gather insights on how trauma is described and recognized across different domains. Feature-based explanations allow us to gain insights into the importance of individual input features, i.e., tokens. We chose model-agnostic approaches that treat the predictive model as a black-box function and can be applied to any model (SHAP values) and model-specific, mechanistic approaches that are only applicable to specific models but can more faithfully describe the output of certain model classes. Additionally, concept-based explanations allow us to move beyond individual feature attributions to a higher level of abstraction, and help us identify interpretable concepts that are crucial for trauma detection without requiring extensive supervision. These methods collectively enhance our ability to interpret model predictions and validate their reliability.

SHAP Explanations. Shapley values originate from game theory and have been proposed to compute the contribution of individual features to the output of a non-linear function. They are a form of feature attribution explanation that assigns each input token a numerical score. The score corresponds to the average contribution to the output obtained when this feature is added. We compute SHAP values using an efficient sampling-based algorithm with the implementation of Lundberg and Lee (2017).

SLALOM Explanations. Leemann et al. (2024) have shown that single attribution scores cannot fully describe the inner workings of modern transformer language models. The authors propose SLALOM, a model to assess the role of input tokens along two dimensions: A *token value* score, describes the effect each token has on its own, while the *token importance* describes how much weight is placed on each token when tokens are concatenated to sequences. While SLALOM can be used to approximate any model's behavior in principle, it is particularly suited for transformer models, like the BERT and RoBERTa models used in this work.

Concept-based Explanations. Concept-based explanations have been proposed as an alternative to feature-wise explanations. They do not reason over individual input features (tokens, pixels, etc.) but instead use a higher level of abstraction (Kim et al., 2018; Koh et al., 2020). However, it is difficult to discover meaningful concepts from the data without supervision (Leemann et al., 2023). In case no concept annotations are present in the data, they identify clusters in a model's latent space that best describe a model's decision. In this work, we turn to Completeness-Aware Concept-Based Explanations (Yeh et al., 2019), which are one of the few conceptual explanation techniques that are applicable to textual inputs and do not require supervision in terms of the data. The concepts are represented as a set of salient examples, i.e., sample snippets that most strongly exhibit the discovered concept.

In this study, we focus on the RoBERTa architectures for concept-based text classification, which proved reliable across all datasets. We use the logit outputs of this model to obtain SHAP and SLALOM explanations and use the latent representation before the classification head as the latent space where the concept vectors are identified. Details on explanation approaches and their hyperparameters are provided in Appendix A.1.

5 Model Performance Results

Classification Performance We fit all the models to the respective datasets after performing hyperparameter optimization (cf. Appendix A.2) and report their performance metrics in Table 3. The evaluation across GTC, PTSD, and Counseling datasets shows clear trends. Transformer-based models, especially fine-tuned BERT and RoBERTa, significantly outperform traditional models and feedforward neural networks. The Naive-Bayes-BoW and NGram Logistic Regression models show moderate performance but lag behind due to their simpler architectures. The feedforward model performs reasonably well but is outclassed by transformer models. Fine-tuned BERT and RoBERTa exhibit substantial improvements in all metrics, with RoBERTa achieving the highest F1 scores in the GTC dataset ($F1 = .74$) and the PTSD dataset ($F1 = .71$), highlighting its effective language comprehension capabilities. To control for dataset size effects, we ran an additional experiment using 1,000 randomly selected GTC samples in the training set to match the size of other datasets. The performance remained consistent, indicating that our findings on smaller datasets likely extend to larger ones (Appendix A, Table 7).

OpenAI's GPT-4 also performs particularly well on the PTSD and Counseling datasets and even outperforms BERT in the F1 metric on Counseling, showcasing its strong generalization abilities despite not being further fine-tuned and relying on a single prompt for these tasks. Interestingly, all models perform reasonably well, which may be attributed to the specific task of trauma event detection. However, the Counseling dataset proved more challenging due to its very imbalanced class distribution and the presence of very few trauma event samples. This is reflected GPT-4 F1 score of .36, which was the highest for this dataset but still indicates the difficulty of the task. RoBERTa achieves strong performance metrics overall, highlighting the impact of architectural improvements and extensive training on larger datasets, though it does not outperform BERT on the Counseling dataset.

Cross-Domain Performance Figure 2 presents the cross-domain results of RoBERTa models fine-tuned on one dataset and evaluated on other datasets, using the AUC-ROC metric (cf., Appendix A, Table 5). Models trained on the GTC dataset showed the highest generalizability, per-

Dataset	GTC		PTSD		Counseling	
LM	F1 (bin.)	AU-ROC	F1 (bin.)	AU-ROC	F1 (bin.)	AU-ROC
NaiveBayes-BoW	0.53 ± 0.09	0.82 ± 0.09	0.56 ± 0.04	0.70 ± 0.02	0.17 ± 0.01	0.70 ± 0.02
NGramLogisticRegression	0.51 ± 0.10	0.83 ± 0.09	0.58 ± 0.02	0.70 ± 0.02	0.15 ± 0.05	0.79 ± 0.01
FeedForwardModel	0.52 ± 0.10	0.84 ± 0.09	0.52 ± 0.05	0.74 ± 0.01	0.03 ± 0.03	0.78 ± 0.01
BERT (finetuned)	0.71 ± 0.01	0.96 ± 0.00	0.66 ± 0.02	0.80 ± 0.01	**0.35** ± 0.05	**0.91** ± 0.01
RoBERTa (finetuned)	**0.74** ± 0.01	**0.97** ± 0.00	**0.71** ± 0.01	**0.83** ± 0.01	0.18 ± 0.09	0.88 ± 0.02
OpenAI GPT-4	0.64	0.94	0.69	0.82	**0.36**	0.85

Table 3: Classification performance of the language models used in this work. We report Binary F1-Scores, and Area under the Receiver-Operator Curve ("AU-ROC"). We report standard errors over cross-validation with 5 runs for all models but the Black-box API models, where computation costs are prohibitive.

forming well across all test sets. Those trained on the PTSD dataset excelled on their own test set and performed strongly on others. Models trained on the Counseling dataset achieved top performance on their own set but did less well on others. The model trained on all combined datasets showed robust and consistent performance across all test sets, maintaining high accuracy and reliability. Despite differences in trauma types across datasets, significant overlaps contribute to strong cross-testing results. For example, both the GTC and PTSD datasets include trauma related to death, acute stress reactions, and physical violence, aiding models' cross-dataset performance. However, the GTC dataset's unique military component may cause some performance differences. Overall, high cross-domain performance suggests that shared trauma themes enable effective generalization across different contexts.

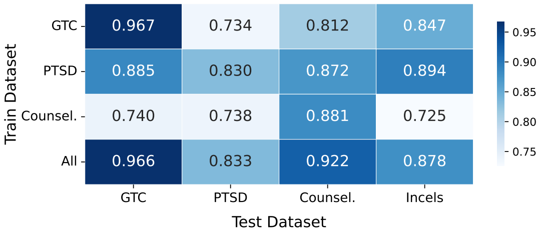


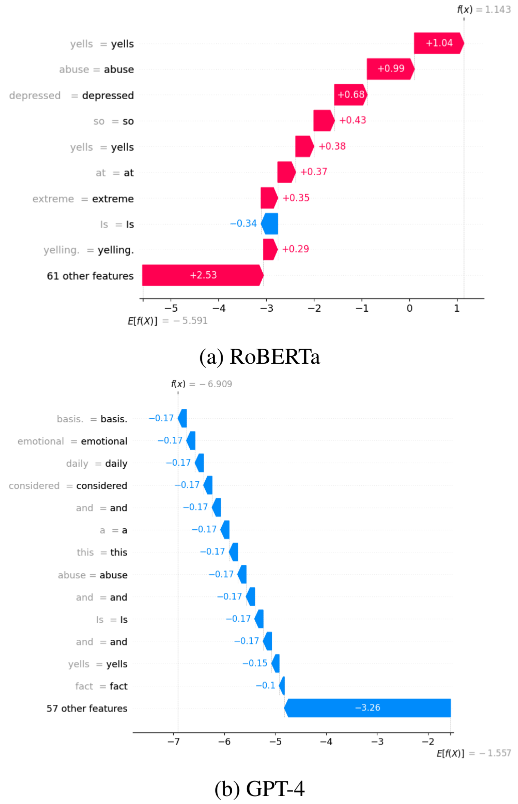
Figure 2: Cross-domain performance (AUC-ROC) when a RoBERTa model is trained on one dataset and tested on other datasets.

The results show that the RoBERTa model fine-tuned on the PTSD dataset has the best generalizability across different datasets, with models trained on the full data also performing well. Given the diversity of traumatic events across datasets, this result suggests the trauma features in the PTSD dataset are broadly applicable for learning a general event type, rather than causing models to pick up on only keywords. Counseling-trained models perform well on their own dataset but do not generalize as effectively. Performance on the Incel dataset indicates all models effectively differentiate trauma-related vocabulary from control data.

SHAP Explanations To understand how the models attribute feature importance to the trauma label, we calculated SHAP values for some samples from all datasets, focusing on comparing RoBERTa and GPT-4 due to their high performances and the interesting differences in how these language models classify trauma. While most classifications aligned (see Figure 8 in Appendix A), we found that, in several instances, GPT-4 provided more non-trauma attributions for certain features compared to RoBERTa.

Figure 3 shows a counseling dataset example where RoBERTa and GPT-4 disagree. RoBERTa assigns high relevance to words like *yells*, *abuse*, and *depressed*, while GPT-4 does not, possibly due to the forum user's uncertainty about defining abuse. This discrepancy may stem from GPT-4's closer adherence to the APA definition of trauma, with less variation and personal bias than human annotators, who may classify events based on their own experiences and interpretations.

These findings, though based on exemplary instances, highlight the challenge of detecting mental abuse. RoBERTa may rely more on specific keywords related to abuse, whereas GPT-4 seems to consider contextual nuances. Human annotators might interpret such incidents as traumatic based on subjective judgment and empathy, while GPT-4, adhering strictly to the APA definition of trauma, did not classify these incidents as trauma.



(a) RoBERTa

(b) GPT-4

Figure 3: SHAP values for an instance from the **Counseling Dataset**: "My dad doesn't like the fact that I'm a boy. He yells at me daily because of it and he tells me I'm extreme and over dramatic. I get so depressed because of my dad's yelling. He keeps asking me why I can't just be happy the way I am and yells at me on a daily basis. Is this considered emotional abuse?"

6 Characteristics of Trauma Across Domains

Feature Characteristics with SLALOM The SLALOM feature importance scores from all datasets focus on the highest value features for trauma classification. Features like *dream* and *shattered*, in the top right corner, contribute most to the trauma classification. For clarity, overlapping features were excluded (blue dots remain in the figure) (Figure 4).

Notable feature variability includes war-related vocabulary (e.g., *bombardment*, *bullets*) likely from genocide-related data, and more generalizable words (e.g., *dreams*, *accident*, *dead*) applicable across domains. Amplifying words like *intense*, *suddenly*, and *gloomy* also appear, fitting traumatic contexts without specific events.

Groups of thematically related words are evident: *dead* and *assassinated* represent death, *wounded*, *choking*, and *slapped* indicate physical injury and violence, and *dreams*, *shattered*, and *replay* are associated with trauma's psychological impact.

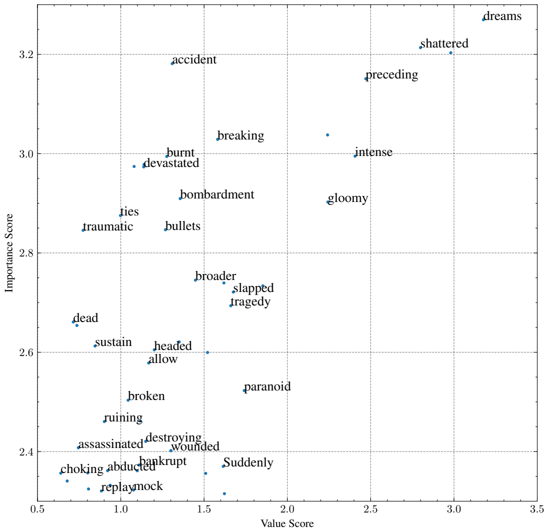


Figure 4: SLALOM feature importance scores based on the full dataset and the RoBERTa model.

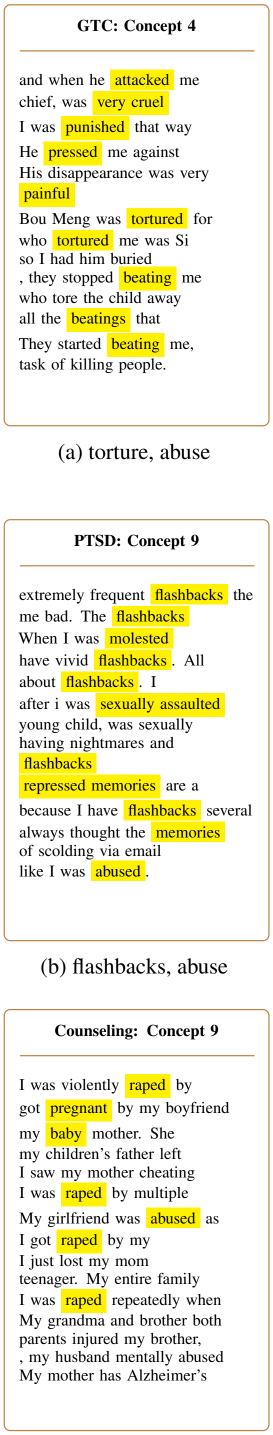
Conceptual Explanations For each dataset, we assessed conceptual explanations to detect context-specific trauma concepts. We select the concepts that have the highest number of traumatic instances in the neighborhood closely associated with the corresponding concept (Figure 5).

In the genocide dataset, concepts related to killings, death, and severe injuries were prominent, reflecting the extreme nature of the content. In contrast, the PTSD and counseling datasets, which address more everyday trauma, contained more references to domestic violence and abuse. The smaller size of the counseling dataset made it challenging to identify unique concepts without overlap.

Across all contexts, death and sexual violence were prevalent. In the genocide dataset, these were depicted through killings and executions, whereas in other datasets, they were associated with grief, loss, and suicide. Sexual violence, particularly rape, consistently appeared as a common source of PTSD, which is consistent with the psychological literature (Atwoli et al., 2015).

7 Conclusion

Traumatic events shape millions of lives. Computational tools to recognize these events can help third parties provide support. However, their diversity makes classification challenging. This paper introduces a new dataset for recognizing traumatic events and analyzes (i) NLP models' performance, (ii) their generalizability across domains, and (iii) if they learn general trauma features using XAI techniques. We show that transformer-based models offer strong performance and generalization, though simpler models still perform well in-domain. However, zero-shot performance by GPT-4 lags behind fine-tuned models. Our analysis shows that while certain features of trauma are context-specific, there are also universal elements across different experiences. However, certain types of traumatic events—notably mental abuse—are particularly challenging to classify due to their less defined nature and greater variability, highlighting the need for clear definitions and enhanced model performance.



GTC: Concept 4

and when he attacked me
chief, was very cruel
I was punished that way
He pressed me against
His disappearance was very painful
Bou Meng was tortured for
who tortured me was Si
so I had him buried
, they stopped beating me
who tore the child away
all the beatings that
They started beating me,
task of killing people.

(a) torture, abuse

PTSD: Concept 9

extremely frequent flashbacks the
me bad. The flashbacks
When I was molested
have vivid flashbacks. All
about flashbacks. I
after i was sexually assaulted
young child, was sexually
having nightmares and flashbacks
repressed memories are a
because I have flashbacks several
always thought the memories
of scolding via email
like I was abused.

(b) flashbacks, abuse

Counseling: Concept 9

I was violently raped by
got pregnant by my boyfriend
my baby mother. She
my children's father left
I saw my mother cheating
I was raped by multiple
My girlfriend was abused as
I got raped by my
I just lost my mom
teenager. My entire family
I was raped repeatedly when
My grandma and brother both
parents injured my brother,
, my husband mentally abused
My mother has Alzheimer's

(c) rape, pregnancy

Figure 5: Trauma-related concepts found in the three datasets (Most salient examples, RoBERTa Model). For more examples see Appendix A.

8 Limitations

The different contexts of the datasets and label imbalance, especially in the Counseling dataset, affect the cross-testing results and overall model performance in trauma detection. Label imbalance is particularly challenging because models may become biased towards the more frequent non-trauma events, leading to poorer performance in detecting the less common trauma events. It is normal to have a smaller number of trauma event samples, making it harder for models to learn and accurately identify these underrepresented cases. However, given that the primary goal of this study is to demonstrate cross-domain transferability on realistic data, it is essential to use datasets with an expected and realistic class distribution.

Technical limitations include the summative nature of the explanations, which only provide high-level insights into the different natures of trauma across domains. Additionally, sampling-based explanations such as SLALOM and SHAP are only approximations of the true model behavior, and their fidelity can be increased with more samples, though this incurs higher computational costs.

Another limitation is that people discuss traumatic events differently depending on the context, which might limit the comparability of the datasets used in this study. Conversations with mental health professionals often use clinical terms, focusing on symptoms, triggers, and coping mechanisms (Tong et al., 2019), while online forums blend informal and semi-formal language where anonymity allows for candid sharing, but responses may vary in depth and understanding (Lahnala et al., 2021; Stana et al., 2017). This contrasts with court testimonies, which require precise, factual language focused on specific events and details for legal documentation (Ciorciari and Heindel, 2011; Schirmer et al., 2023b).

We chose the span annotation method, where annotators select the text indicating a traumatic event, because pilot experiments showed it improved performance by focusing attention on specific events rather than a simple "yes" or "no" decision. Although this was a design choice and not a central research question, analyzing these spans could offer insights into annotation quality and inform future training. Investigating the detection of specific traumatic event spans rather than general segments is a promising direction for future research.

Finally, our analysis partially relies on social media data. This type of data provides vast, real-time insights into public mental health trends but can be noisy and less reliable. It would be important for future studies to replicate our results with clinical data to ensure the findings' robustness and applicability in medical settings.

Ethics Statement

Our data processing procedures did not involve any handling of private information. No user names were obtained at any point of the data collection process. The human annotators were informed of and aware of the potentially violent content before the annotation process, with the ability to decline annotation at any time. The same is true for crowdworkers, who were presented several trigger warnings throughout the process. Both human coders were given the chance to discuss any distressing material encountered during annotation. As discussions on the potential trauma or adverse effects experienced by annotators while dealing with distressing material become more prevalent (Kennedy et al., 2022), we have proactively provided annotators with a recommended written guide designed to aid in identifying changes in cognition and minimizing emotional risks associated with the annotation process.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arfan Ahmed, Sarah Aziz, Carla T Toro, Mahmood Alzubaidi, Sara Irshaidat, Hashem Abu Serhan, Alaa A Abd-Alrazaq, and Mowafa Househ. 2022. Machine learning models to detect anxiety and depression through social media: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2:100066.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition*. American Psychiatric Publishing.

Amod. 2024. mental_health_counseling_conversations (revision 9015341).

Lukoye Atwoli, Dan J Stein, Karestan C Koenen, and Katie A McLaughlin. 2015. Epidemiology of posttraumatic stress disorder: prevalence, correlates and consequences. *Current opinion in psychiatry*, 28(4):307–311.

Marcel O Bonn-Miller, Megan Brunstetter, Alex Simonian, Mallory J Loflin, Ryan Vandrey, Kimberly A Babson, and Hal Wortzel. 2022. The long-term, prospective, therapeutic impact of cannabis on post-traumatic stress disorder. *Cannabis and cannabinoid research*, 7(2):214–223.

Naomi Breslau, EL Peterson, LM Poisson, LR Schultz, and VC Lucia. 2004. Estimating post-traumatic stress disorder in the community: lifetime perspective and the impact of typical traumatic events. *Psychological medicine*, 34(5):889–898.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

John D Ciorciari and Anne Heindel. 2011. Trauma in the courtroom. *Cambodia's hidden scars: Trauma psychology in the wake of the Khmer Rouge. Phnom Penh: Documentation Center of Cambodia (DC-Cam)*.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 579–582.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Heather L Dye. 2020. Is emotional abuse as harmful as physical and/or sexual abuse? *Journal of Child & Adolescent Trauma*, 13(4):399–407.

Atle Dyregrov, Leila Gupta, Rolf Gjestad, and Eugenie Mukanoheli. 2000. Trauma exposure and psychological reactions to genocide among rwandan children. *Journal of traumatic stress*, 13:3–21.

Matthew J Friedman, Terence M Keane, and Patricia A Resick. 2007. *Handbook of PTSD: Science and practice*. Guilford press.

Steven N Gold. 2017. *APA handbook of trauma psychology: foundations in knowledge, Vol. 1*. American Psychological Association.

Zhijun Guo, Alvina Lai, Johan Hilge Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language model for mental health: A systematic review. *arXiv preprint arXiv:2403.15401*.

Qiwei He, Bernard P Veldkamp, Cees AW Glas, and Theo de Vries. 2017. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment*, 24(2):157–172.

S. Hornstein, J. Scharfenberger, U. Lueken, et al. 2024. Predicting recurrent chat contact in a psychological intervention for the youth using natural language processing. *npj Digital Medicine*, 7:132.

Thormod Idsoe, Tracy Vaillancourt, Atle Dyregrov, Kristine Amlund Hagen, Terje Ogden, and Ane Nærde. 2021. Bullying victimization and trauma. *Frontiers in psychiatry*, 11:480353.

Shaoxiong Ji. 2022. Towards intention understanding in suicidal risk assessment with natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4028–4038.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677. PMLR.

Laurel J Kiser, Jerry Heston, Pamela A Millsap, and David B Pruitt. 1991. Physical and sexual abuse in childhood: Relationship with post-traumatic stress disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 30(5):776–783.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K Kummerfeld, Lawrence An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. *arXiv preprint arXiv:2106.12976*.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

Tobias Leemann, Alina Fastowski, Felix Pfeiffer, and Gjergji Kasneci. 2024. Attention mechanisms don't learn additive models: Rethinking feature importance for transformers. *arXiv preprint arXiv:2405.13536*.

Tobias Leemann, Michael Kirchhof, Yao Rong, Enkelejda Kasneci, and Gjergji Kasneci. 2023. When are post-hoc conceptual explanations identifiable? In *Uncertainty in Artificial Intelligence*, pages 1207–1218. PMLR.

Maxwell Levis, Christine Leonard Westgate, Jiang Gui, Bradley V Watts, and Brian Shiner. 2021. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychological medicine*, 51(8):1382–1391.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Daniel M Low, Laurie Rumker, John Torous, Guillermo Cecchi, Satrajit S Ghosh, and Tanya Talkar. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Matteo Malgaroli, Thomas D Hull, James M Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Charles R Marmar, Adam D Brown, Meng Qian, Eugene Laska, Carole Siegel, Meng Li, Duna Abu-Amara, Andreas Tsiartas, Colleen Richey, Jennifer Smith, et al. 2019. Speech-based markers for post-traumatic stress disorder in us veterans. *Depression and Anxiety*, 36(7):607–616.

Daniel Matter, Miriam Schirmer, Nir Grinberg, and Jürgen Pfeffer. 2024. Investigating the increase of violent speech in incel communities with human-guided gpt-4 prompt iteration. *Frontiers in Social Psychology*, 2:1383152.

Laura Ann McCloskey and Marla Walker. 2000. Post-traumatic stress in children exposed to family violence and single-event trauma. *Journal of the American Academy of Child & Adolescent Psychiatry*, 39(1):108–115.

Oshin Miranda, Sophie Marie Kiehl, Xiguang Qi, M Daniel Brannock, Thomas Kosten, Neal David Ryan, Levent Kirisci, Yanshan Wang, and LiRong Wang. 2024. Enhancing post-traumatic stress disorder patient assessment: leveraging natural language processing for research of domain criteria identification using electronic medical records. *BMC medical informatics and decision making*, 24(1):1–14.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Steve Powell, Rita Rosner, Willi Butollo, Richard G Tedeschi, and Lawrence G Calhoun. 2003. Posttraumatic growth after war: A study with former refugees and displaced people in sarajevo. *Journal of clinical psychology*, 59(1):71–83.

Robin Quillivic, Frédérique Gayraud, Yann Auxéméry, Laurent Vanni, Denis Peschanski, Francis Eustache, Jacques Dayan, and Salma Mesmoudi. 2024. Interdisciplinary approach to identify language markers for post-traumatic stress disorder using machine learning and deep learning. *Scientific reports*, 14(1):12468.

Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. 2017. Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, 7(1):13006.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697.

Miriam Schirmer, Isaac Misael Olguín Nolasco, Edoardo Mosca, Shanshan Xu, and Jürgen Pfeffer. 2023a. Uncovering trauma in genocide tribunals: An nlp approach using the genocide transcript corpus. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 257–266.

Miriam Schirmer, Jürgen Pfeffer, and Sven Hilbert. 2023b. Talking about torture: A novel approach to the mixed methods analysis of genocide-related witness statements in the khmer rouge tribunal. *Journal of Mixed Methods Research*, page 15586898231218463.

Neil J Smelser et al. 2004. Psychological trauma and cultural trauma. *Cultural trauma and collective identity*, 4:31–59.

Alexandru Stana, Mark A Flynn, and Eugenie Almeida. 2017. Battling the stigma: Combat veterans' use of social support in an online ptsd forum. *International Journal of Men's Health*, 16(1).

Akshay Swaminathan, Iván López, Rafael Antonio Garcia Mar, Tyler Heist, Tom McClintock, Kaitlin Caoili, Madeline Grace, Matthew Rubashkin, Michael N Boggs, Jonathan H Chen, et al. 2023. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. *NPJ Digital Medicine*, 6(1):213.

Vankayala Tejaswini, Korra Sathya Babu, and Bibhudatta Sahoo. 2024. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–20.

Lenore C Terr. 2003. Childhood traumas: An outline and overview. *Focus*, 1(3):322–334.

Janet Tong, Katrina Simpson, Mario Alvarez-Jimenez, and Sarah Bendall. 2019. Talking about trauma in therapy: Perspectives from young people with post-traumatic stress symptoms and first episode psychosis. *Early intervention in psychiatry*, 13(5):1236–1244.

Mohammad Arif Ul Alam and Dhawal Kapadia. 2020. Laxary: A trustworthy explainable twitter analysis model for post-traumatic stress disorder assessment. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 308–313.

Bessel A Van der Kolk. 2003. *Psychological trauma*. American Psychiatric Pub.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2019. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32.

Rachel Yehuda. 1998. *Psychological trauma*. American Psychiatric Pub.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digital Medicine*, 5(1):46.

A Appendix

A.1 Implementation Details: Explanation Methods

In this section, we give more details on how we computed the explanations shown in this paper.

SHAP Values. To obtain SHAP values, we use the official `shap`[1] package. We use the TextExplainer class.

SLALOM. We use the SGD algorithm proposed in Leemann et al. (2024) to estimate the SLALOM model on 100k background samples of length 2. We use all the tokens that appear in the samples from the datasets used and fit one global SLALOM model.

Conceptual Explanations. We use the completeness-aware loss proposed by Yeh et al. (2019) with snippets of length of 5 token as snippets for the algorithm. We trained with concept discovery module to discover $K = 10$ concepts using the Adam optimizer at an initial learning rate of 1×10^{-3}, decaying to 5×10^{-4} and 1×10^{-4} in subsequent epochs. Training lasted 3 epochs with a batch size of 12. The model weights used were obtained from the best-performing model. We identified the 25 closest activations per concept. Evaluation on a separate test set involved dot products between latent representations and concept vectors, selecting the top activations.

A.2 Implementation Details: Models

We use the `optuna`[2] framework for hyperparameter optimization with 50 steps for each model/dataset. We then train the models using different seeds and on five random data splits using the discovered hyperparameters. Through the optimization we obtain the parameters given in Table 4.

Prompt Template. We use the following prompt template to prompt the GPT models as the system prompt.

"You are tasked with detecting trauma in text segments of transcripts of genocide tribunals. Specifically, detect instances that meet the APA's definition of trauma. Psychological trauma, as defined by the APA, includes experiences of exposure to actual or threatened death, serious injury, or sexual violence, either directly encountered or witnessed. It also includes instances where individuals learn that the traumatic event(s) occurred to a close family member or friend. Label the text with '1' if there are indicators of trauma based on this definition, and '0' if there are no indicators of trauma. Note that trauma is rare and occurs in less than 20% of the cases. Only answer with either '0' or '1'."

The samples are then passed as a user prompt.

A.3 Annotation Details

Participants were prescreened using Prolific based on self-reported English-language proficiency. We did not collect demographic data from the annotators as such data was not central to the questions our study is focused on and Prolific does not normally include this metadata.

A.4 Metrics

For completeness, we additionally report accuracy, recall, and precision for the trained models in Table 6.

[1] https://github.com/shap/shap

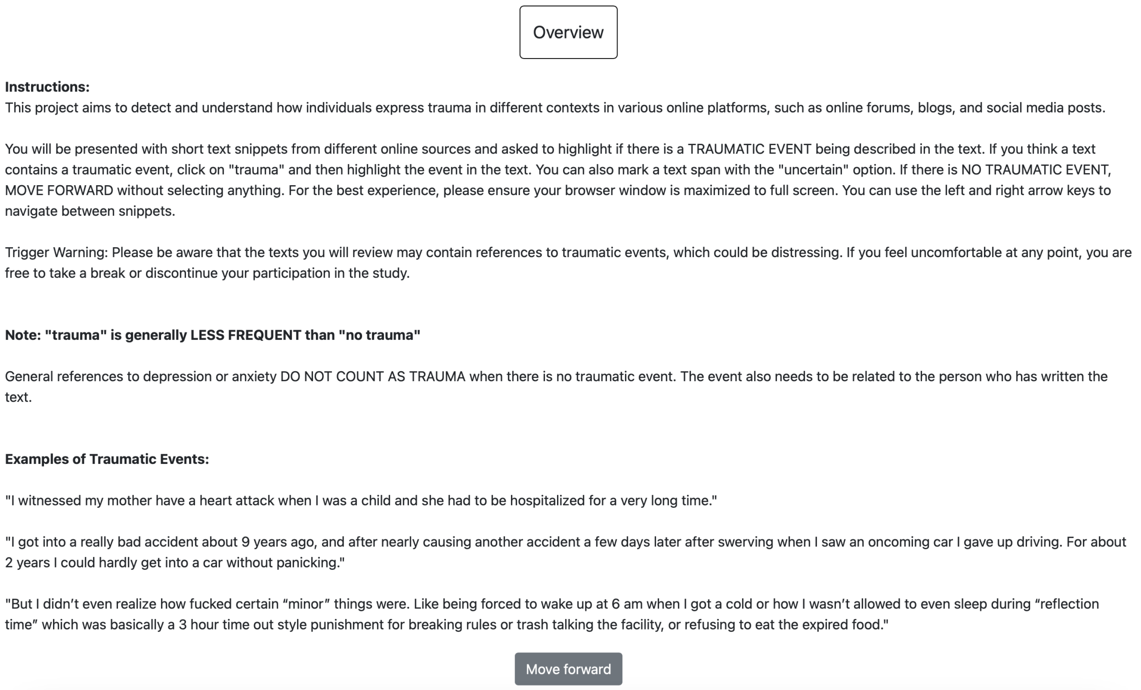
[2] https://optuna.org/

	Parameters		
Model	**GTC**	**PTSD**	**Counseling**
NaiveBayes-BoW	multiplicities: true alpha: 1.01	multiplicities: true alpha: 5.97	multiplicities: false alpha: 1.01
NGramLogisticRegression	n_gram_range: [1, 2] C: 0.92 penalty: l2	n_gram_range: [2, 3] C: 0.0 penalty: none	n_gram_range: [1, 2] C: 9.36 penalty: l2
FeedForwardModel	hidden_dim1: 50 hidden_dim2: 80 lr: 5.72e-05	hidden_dim1: 50 hidden_dim2: none lr: 1.79e-04	hidden_dim1: 200 hidden_dim2: 50 lr: 5.72e-05
BERT (finetuned)	n_layers: 5 lr: 2.32e-05	n_layers: 12 lr: 1.10e-05	n_layers: 6 lr: 1.41e-05
RoBERTa (finetuned)	n_layers: 12 lr: 2.04e-06	n_layers: 7 lr: 6.43e-06	n_layers: 4 lr: 9.54e-05
OpenAI	target_model: gpt-4-turbo	target_model: gpt-4-turbo	target_model: gpt-4-turbo

Table 4: Automatically selected hyperparameters for the different datasets

Dataset	Test Dataset			
Train	GTC	PTSD	Counsel.	Incels
GTC	**0.967** ± 0.000	0.734 ± 0.005	0.812 ± 0.020	0.847 ± 0.003
PTSD	0.885 ± 0.010	0.830 ± 0.006	0.872 ± 0.014	**0.894** ± 0.010
Counsel.	0.740 ± 0.017	0.738 ± 0.018	0.881 ± 0.016	0.725 ± 0.027
All	0.966 ± 0.001	**0.833** ± 0.013	**0.922** ± 0.012	0.878 ± 0.005

Table 5: Cross-Testing models trained on one dataset on other datasets. Model: RoBERTa finetuned with AU-ROC metric



Instructions:

This project aims to detect and understand how individuals express trauma in different contexts in various online platforms, such as online forums, blogs, and social media posts.

You will be presented with short text snippets from different online sources and asked to highlight if there is a TRAUMATIC EVENT being described in the text. If you think a text contains a traumatic event, click on "trauma" and then highlight the event in the text. You can also mark a text span with the "uncertain" option. If there is NO TRAUMATIC EVENT, MOVE FORWARD without selecting anything. For the best experience, please ensure your browser window is maximized to full screen. You can use the left and right arrow keys to navigate between snippets.

Trigger Warning: Please be aware that the texts you will review may contain references to traumatic events, which could be distressing. If you feel uncomfortable at any point, you are free to take a break or discontinue your participation in the study.

Note: "trauma" is generally LESS FREQUENT than "no trauma"

General references to depression or anxiety DO NOT COUNT AS TRAUMA when there is no traumatic event. The event also needs to be related to the person who has written the text.

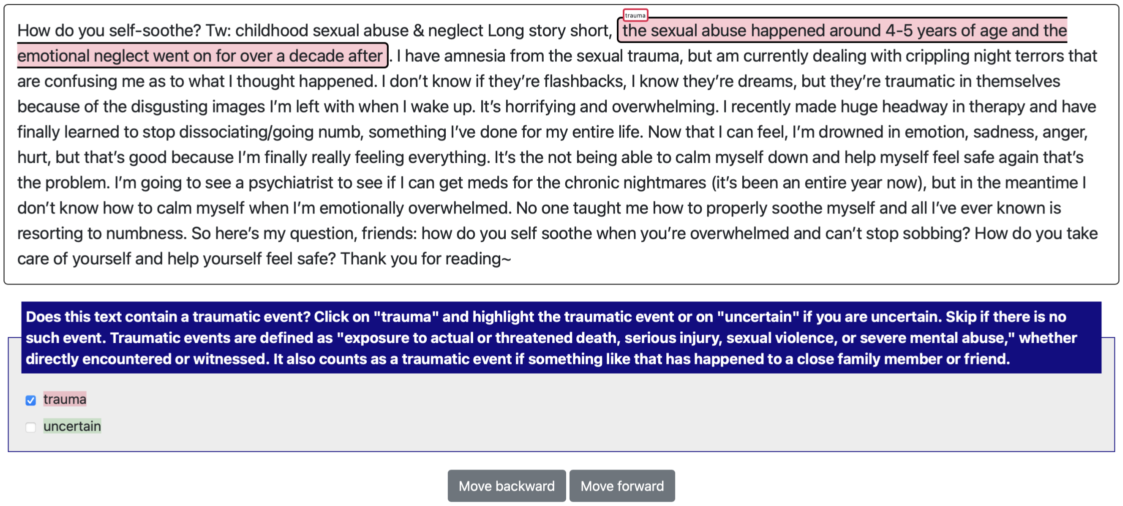
Examples of Traumatic Events:

"I witnessed my mother have a heart attack when I was a child and she had to be hospitalized for a very long time."

"I got into a really bad accident about 9 years ago, and after nearly causing another accident a few days later after swerving when I saw an oncoming car I gave up driving. For about 2 years I could hardly get into a car without panicking."

"But I didn't even realize how fucked certain "minor" things were. Like being forced to wake up at 6 am when I got a cold or how I wasn't allowed to even sleep during "reflection time" which was basically a 3 hour time out style punishment for breaking rules or trash talking the facility, or refusing to eat the expired food."

Figure 6: Instructions for Annotators. Note: We selected these examples because they were the most frequently mislabeled in the pilot, making them particularly relevant. Additionally, we kept the instruction page concise to avoid overwhelming the annotators, as excessive detail could deter them or lead to less careful reading.



How do you self-soothe? Tw: childhood sexual abuse & neglect Long story short, the sexual abuse happened around 4-5 years of age and the emotional neglect went on for over a decade after. I have amnesia from the sexual trauma, but am currently dealing with crippling night terrors that are confusing me as to what I thought happened. I don't know if they're flashbacks, I know they're dreams, but they're traumatic in themselves because of the disgusting images I'm left with when I wake up. It's horrifying and overwhelming. I recently made huge headway in therapy and have finally learned to stop dissociating/going numb, something I've done for my entire life. Now that I can feel, I'm drowned in emotion, sadness, anger, hurt, but that's good because I'm finally really feeling everything. It's the not being able to calm myself down and help myself feel safe again that's the problem. I'm going to see a psychiatrist to see if I can get meds for the chronic nightmares (it's been an entire year now), but in the meantime I don't know how to calm myself when I'm emotionally overwhelmed. No one taught me how to properly soothe myself and all I've ever known is resorting to numbness. So here's my question, friends: how do you self soothe when you're overwhelmed and can't stop sobbing? How do you take care of yourself and help yourself feel safe? Thank you for reading~

Does this text contain a traumatic event? Click on "trauma" and highlight the traumatic event or on "uncertain" if you are uncertain. Skip if there is no such event. Traumatic events are defined as "exposure to actual or threatened death, serious injury, sexual violence, or severe mental abuse," whether directly encountered or witnessed. It also counts as a traumatic event if something like that has happened to a close family member or friend.

☑ trauma

☐ uncertain

Figure 7: Interface of the Span Annotation Task.

Model	Accuracy	Precision	Recall
	GTC		
NaiveBayesBOWmodel	0.84 ± 0.03	0.44 ± 0.08	0.69 ± 0.12
NGramLogisticRegression	0.88 ± 0.02	0.60 ± 0.12	0.44 ± 0.09
FeedForwardModel	0.88 ± 0.02	0.60 ± 0.12	0.46 ± 0.09
BERTmodel	0.88 ± 0.03	0.58 ± 0.12	0.46 ± 0.10
RoBERTamodel	0.91 ± 0.00	0.70 ± 0.02	0.59 ± 0.05
BERTPretrainedmodel	0.92 ± 0.00	0.74 ± 0.03	0.70 ± 0.04
RoBERTaPretrainedmodel	0.93 ± 0.00	0.75 ± 0.03	0.74 ± 0.04
OpenAI GPT-4	0.91	0.68	0.61
	PTSD		
NaiveBayesBOWmodel	0.69 ± 0.01	0.63 ± 0.02	0.52 ± 0.06
NGramLogisticRegression	0.68 ± 0.01	0.62 ± 0.03	0.54 ± 0.03
FeedForwardModel	0.70 ± 0.01	0.71 ± 0.04	0.42 ± 0.05
BERTPretrainedmodel	0.72 ± 0.01	0.64 ± 0.01	0.69 ± 0.06
RoBERTaPretrainedmodel	0.75 ± 0.01	0.66 ± 0.02	0.78 ± 0.04
OpenAI GPT-4	0.69	0.58	0.84
	Counseling		
NaiveBayesBOWmodel	0.26 ± 0.01	0.09 ± 0.01	0.99 ± 0.01
NGramLogisticRegression	0.92 ± 0.01	0.55 ± 0.17	0.09 ± 0.03
eedForwardModel	0.92 ± 0.01	0.10 ± 0.10	0.02 ± 0.02
BERTPretrainedmodel	0.93 ± 0.01	0.54 ± 0.04	0.27 ± 0.05
RoBERTaPretrainedmodel	0.91 ± 0.01	0.36 ± 0.19	0.20 ± 0.12
OpenAI GPT-4	0.91	0.42	0.31

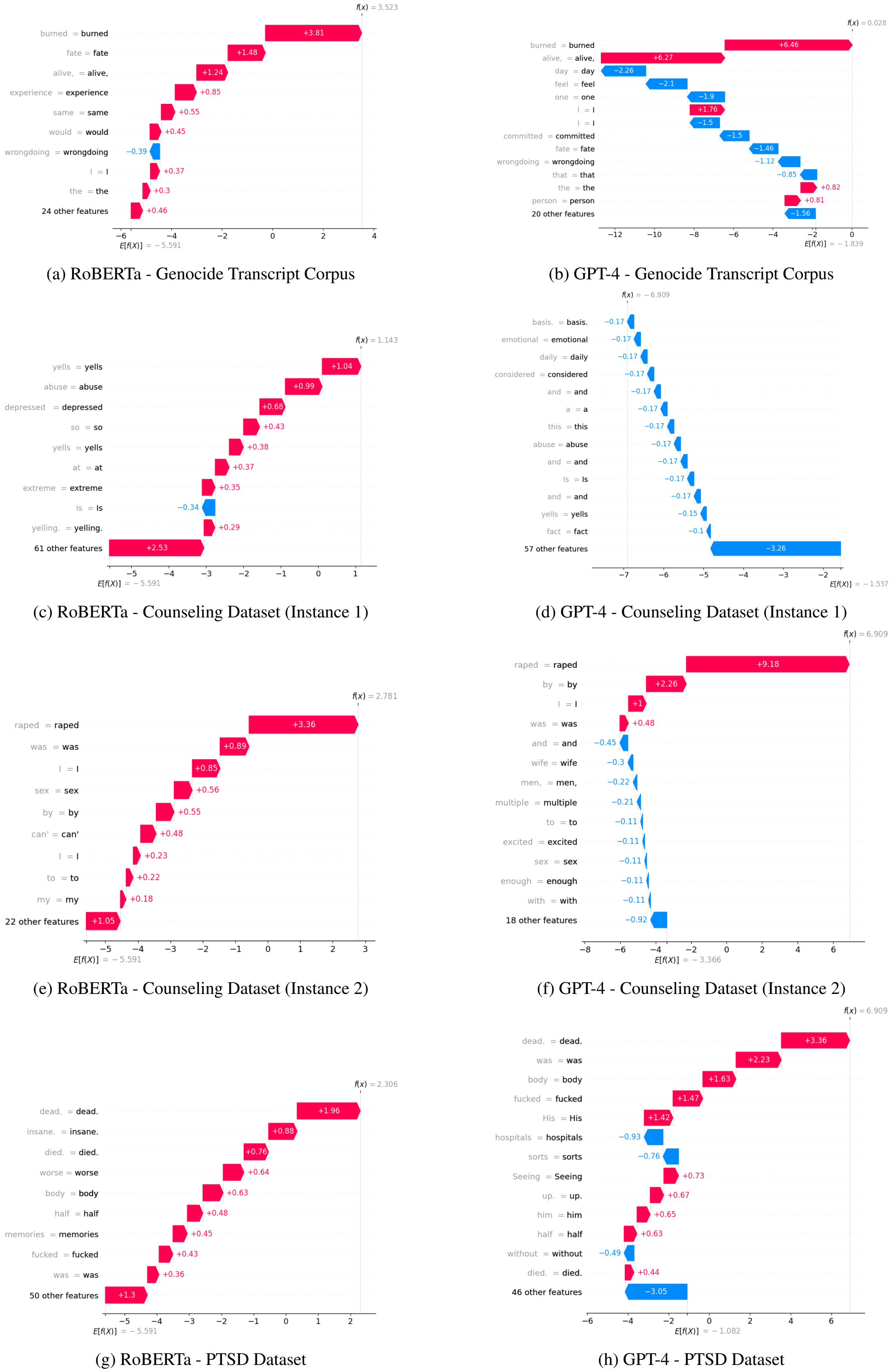
Table 6: Additional model performance metrics. We see that the non-pretrained versions of BERT/RoBERTa do not perform on par with the pretrained ones on GTC. Therefore, we consider only the pretrained versions for the rest of the paper.

	GTC-1000		GTC-All	
LM	F1 (bin.)	AU-ROC	F1 (bin.)	AU-ROC
FeedForwardModel	0.38 ± 0.01	0.86 ± 0.00	0.52 ± 0.10	0.84 ± 0.09
BERTPretrainedmodel	0.61 ± 0.03	0.93 ± 0.00	0.71 ± 0.01	0.96 ± 0.00
RoBERTaPretrainedmodel	0.66 ± 0.03	0.95 ± 0.00	0.74 ± 0.01	0.97 ± 0.00

Table 7: Additional experiments with a smaller GTC (Genocide Transcript Corpus) sample size to control for dataset size effects.

Dataset	Instance
Genocide Transcript Corpus	I can feel that the person committed any wrongdoing would be burned alive, and I would also see that one day if I committed any wrongdoing I would experience the same fate.
Counseling Dataset (Instance 1)	My dad doesn't like the fact that I'm a boy. He yells at me daily because of it and he tells me I'm extreme and over dramatic. I get so depressed because of my dad's yelling. He keeps asking me why I can't just be happy the way I am and yells at me on a daily basis. Is this considered emotional abuse?
Counseling Dataset (Instance 2)	I was raped by multiple men, and now I can't stand the sight of myself. I wear lingerie to get my self excited enough to have sex with my wife.
PTSD Dataset	It's nearly been 4 years (trigger warning) It's almost been 4 years since he died. I can't look at hospitals without the memories coming back. Seeing him half dead. His body was all sorts of fucked up. I can't deal with this any longer. I'm going to go insane. Every day it gets worse.

Table 8: Instances from various datasets used for SHAP value analysis (see Figure 8).



(a) RoBERTa - Genocide Transcript Corpus

(b) GPT-4 - Genocide Transcript Corpus

(c) RoBERTa - Counseling Dataset (Instance 1)

(d) GPT-4 - Counseling Dataset (Instance 1)

(e) RoBERTa - Counseling Dataset (Instance 2)

(f) GPT-4 - Counseling Dataset (Instance 2)

(g) RoBERTa - PTSD Dataset

(h) GPT-4 - PTSD Dataset

Figure 8: SHAP Values for various instances from different datasets. See Table 8 for the full text of each instance.

Concept 2

some cows were wounded.
or diarrhoea.
supply themselves with food.
so itchy everywhere.
bullet in the head;
those various gunshots.</s>
once I was wounded.
, one hit me in
three bursts of gunfire.
themselves from the bullets.
to work I kept crying
and killed animals for wedding
the gunshots and we were
I even saw dead bodies
I cry every night.

(a) firearms, physical injuries

Concept 4

and when he attacked me
chief, was very cruel
I was punished that way
He pressed me against
His disappearance was very
painful
Bou Meng was tortured for
who tortured me was Si
and then my eyes would
olded me for being blinded
so I had him buried
, they stopped beating me
who tore the child away
all the beatings that
They started beating me,
task of killing people.

(b) torture, abuse

Concept 6

long you fell unconscious.
falling into the ground.
another place two bodies,
them were lying on the
weapon in my mouth.
I started freezing, and
up the dead bodies and
us drowned in the river
a lot of dead and
people who perished there.
children die of hunger,
then we all got stuck
people and throwing their bodies
her sister burned to death
who died of hunger,

(c) death, motionless bodies

Concept 7

December he called me and
my husband called cadres
sometimes he called me to
his phone call because he
or called me and then
was called Lucia, Ruk
know what happened to Ph
man called Rukara went
since I was busy looking
school was called Hasan Ve
really is still in my
one called me up and
my wife asked them what
tell you exactly when this
It arrived in Kosovo Pol

(d) communication, asking questions (non-traumatic)

(a) Concepts Found in the **GTC Dataset** (Examples, RoBERTa Model): (a)-(c) Trauma-Related, (d) Non-Trauma Related

Concept 0

in the dream with the
In this dream,
awake in the dream,
I survived veteran suicide.
Or in a dream rel
toxic relationships with men.
,500 military sexual assaults
shot and killed himself in
ago I had a dream
only dated for a few
after having a dream that
harmed while in a
was a dream. This
shot herself in the head
, on a medical discharge

(e) dreams, military, suicide

Concept 5

. Our car clipped the
his car multiple times.
my skull. It wasn
raped at least three times
hit another car, causing
22 year old man (
were later found and her
a lot of death,
SA and death
of sexual abuse, two
murder-suicides each
tried to apprehend the two
/SuicideWatch and
after service. Fires are
experiences with SSRIs

(f) accidents, death, sexual abuse

Concept 6

my parents were kids,
my childhood over to my
my mom divorced my violent
I was a child,
had an abusive ex and
my mother died 2 years
memories of my childhood,
abusive and cheating gay ex
first serious boyfriend repeatedly d
realize TW: childhood sexual
my trauma from childhood,
my now ex of 4
My parents were divorced when
my parents raised me and
throughout my childhood by my

(g) childhood abuse, family trauma

Concept 9

extremely frequent flashbacks the
me bad. The flashbacks
When I was molested
have vivid flashbacks. All
aphobia in the last
about flashbacks. I
phobia related to the
after i was sexually assaulted
young child, was sexually
having nightmares and
flashbacks
repressed memories are a
because I have flashbacks several
always thought the memories
of scolding via email
like I was abused.

(h) flashbacks, memories, sexual abuse, phobia

(b) Concepts Found in the **PTSD Dataset** (Examples, RoBERTa Model): (a)-(d) Trauma-Related

Concept 4

My boyfriend lost his dad
my mom. My dad
shot and killed my rapist
His mom, my grandma
My dad cheated on my
I lost my mother recently
his wife. My uncle
, my mother and father
I was kidnapped at fourteen
. My father cheated on
My ex-wife married
niece whom my sister abandoned
My daughter was overly tired
my mom. Years later
misses mom and dad in

(i) family members: cheating, loss

Concept 5

My boyfriend lost his dad
my mom. My dad
shot and killed my rapist
his wife. My uncle
His mom, my grandma
My dad cheated on my
I lost my mother recently
, my mother and father
. My father cheated on
My ex-wife married
I was kidnapped at fourteen
Post Traumatic Stress Disorder
Post Traumatic Stress Disorder
niece whom my sister abandoned
posttraumatic stress disorder

(j) PTSD (overlap Concept 4)

Concept 6

My boyfriend lost his dad
my mom. My dad
abusive father and his wife
his wife. My uncle
His mom, my grandma
My dad cheated on my
I lost my mother recently
, my mother and father
I was kidnapped at fourteen
My sister never defended me
, my doctor gave me
. He bought me a
all of my family left
broken apart after she got
rehabilitation program and got
kicked

(k) clinical context, dependencies (overlap Concept 4)

Concept 9

I was violently raped by
got pregnant by my boyfriend
my baby mother. She
my children's father left
I saw my mother cheating
I was raped by multiple
My girlfriend was abused as
I got raped by my
I just lost my mom
teenager. My entire family
I was raped repeatedly when
My grandma and brother both
parents injured my brother,
, my husband mentally abused
My mother has Alzheimer's

(l) rape, abuse, pregnancy

(c) Concepts Found in the **Counseling Dataset** (Examples, RoBERTa Model): (a)-(d) Trauma-Related

Figure 9: Examples of Concepts Discovered on Various Datasets for the RoBERTa Model: (a) GTC dataset, (b) PTSD dataset, (c) Counseling dataset.

4.6 Study 6: Investigating the Increase of Violent Speech in Incel Communities with Human-Guided GPT-4 Prompt Iteration

This publication is RELEVANT TO THE EXAMINATION.

Authors

Daniel Matter*, Miriam Schirmer*, Nir Grinberg, Jürgen Pfeffer

*The authors contributed equally.

In

Frontiers in Social Psychology 2:1383152, peer-reviewed, published 07-2024, `https://doi.org/10.3389/frsps.2024.1383152`.

Abstract

This study investigates the prevalence of violent language on incels.is. It evaluates GPT models (GPT-3.5 andGPT-4) for content analysis in social sciences, focusing on the impact of varying prompts and batch sizes on coding quality for the detection of violent speech. We scraped over 6.9M posts from incels.is and categorized a random sample into non-violent, explicitly violent, and implicitly violent content. Two human coders annotated 3,028 posts, which we used to tune and evaluate GPT-3.5 and GPT-4 models across different prompts and batch sizes regarding coding reliability. The best-performing GPT-4 model annotated an additional 45,611 posts for further analysis. We find that 21.91% of the posts on the forum contain some form of violent language. Within the overall forum, 18.12% of posts include explicit violence, while 3.79% feature implicit violence. Our results show a significant rise in violent speech on incels.is, both at the community and individual level. This trend is particularly pronounced among users with an active posting behavior that lasts for several hours up to one month. While the use of targeted violent language decreases, general violent language increases. Additionally, mentions of self-harm decline, especially for users who have been active on the site for over 2.5 years. We find substantial agreement between both human coders (Cohen's kappa = 0.65), while the best GPT-4 model yields good agreement with both human coders (Cohen's kappa = 0.54 for Human A and 0.62 for Human B). Overall, this research offers effective ways to pinpoint violent language on a large scale, helping with content moderation and facilitating further research into causal mechanisms and potential mitigations of violent expression and online radicalization in communities like incels.is.

Contribution of Thesis Author

Theoretical conceptualization, methodological design, manuscript writing, revision, and editing.

Frontiers in Social Psychology

TYPE Original Research
PUBLISHED 03 July 2024
DOI 10.3389/frsps.2024.1383152

OPEN ACCESS

EDITED BY
Clemens Stachl,
University of St. Gallen, Switzerland

REVIEWED BY
Rudolf Debelak,
University of Zurich, Switzerland
Jan Digutsch,
University of St. Gallen, Switzerland
Suhaib Abdurahman,
University of Southern California, United States

*CORRESPONDENCE
Miriam Schirmer
✉ miriam.schirmer@tum.de

†These authors have contributed equally to this work and share first authorship

RECEIVED 06 February 2024
ACCEPTED 29 May 2024
PUBLISHED 03 July 2024

CITATION
Matter D, Schirmer M, Grinberg N and Pfeffer J (2024) Investigating the increase of violent speech in Incel communities with human-guided GPT-4 prompt iteration. *Front. Soc. Psychol.* 2:1383152.
doi: 10.3389/frsps.2024.1383152

Investigating the increase of violent speech in Incel communities with human-guided GPT-4 prompt iteration

Daniel Matter[1†], Miriam Schirmer[1*†], Nir Grinberg[2] and Jürgen Pfeffer[1]

[1]Department of Governance, School of Social Sciences and Technology, Technical University of Munich, Munich, Germany, [2]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beersheba, Israel

This study investigates the prevalence of violent language on *incels.is*. It evaluates GPT models (GPT-3.5 and GPT-4) for content analysis in social sciences, focusing on the impact of varying prompts and batch sizes on coding quality for the detection of violent speech. We scraped over 6.9*M* posts from *incels.is* and categorized a random sample into non-violent, explicitly violent, and implicitly violent content. Two human coders annotated 3,028 posts, which we used to tune and evaluate GPT-3.5 and GPT-4 models across different prompts and batch sizes regarding coding reliability. The best-performing GPT-4 model annotated an additional 45,611 posts for further analysis. We find that 21.91% of the posts on the forum contain some form of violent language. Within the overall forum, 18.12% of posts include explicit violence, while 3.79% feature implicit violence. Our results show a significant rise in violent speech on *incels.is*, both at the community and individual level. This trend is particularly pronounced among users with an active posting behavior that lasts for several hours up to one month. While the use of targeted violent language decreases, general violent language increases. Additionally, mentions of self-harm decline, especially for users who have been active on the site for over 2.5 years. We find substantial agreement between both human coders ($\kappa = 0.65$), while the best GPT-4 model yields good agreement with both human coders ($\kappa = 0.54$ for Human A and $\kappa = 0.62$ for Human B). Overall, this research offers effective ways to pinpoint violent language on a large scale, helping with content moderation and facilitating further research into causal mechanisms and potential mitigations of violent expression and online radicalization in communities like *incels.is*.

KEYWORDS

GPT-4, violent language, Incels, annotator agreement, time analysis

1 Introduction

The term "Incels" ("Involuntary Celibates") refers to heterosexual men who, despite yearning for sexual and intimate relationships, find themselves unable to engage in such interactions. The online community of Incels has been subject to increasing attention from both media and academic research, mainly due to its connections to real-world violence (Hoffman et al., 2020). Scrutiny intensified after over 50 deaths have been linked to Incel-related incidents since 2014 (Lindsay, 2022). The rising trend of Incel-related violence underscores societal risks posed by the views propagated within the community, especially those regarding women. In response, various strategic and administrative measures have been implemented. Notably, the social media platform Reddit officially banned the largest Incel subreddit *r/incel* for inciting violence against women (Hauser, 2017). The Center for Research and Evidence on Security Threats has emphasized the community's violent, misogynistic tendencies, classifying its ideology as extremist (Brace, 2021). Similarly, the Texas Department of Public Safety has labeled Incels as an "emerging domestic terrorism threat" (Texas Department of Public Safety, 2020).

Incels mainly congregate on online platforms. Within these forums, discussions frequently revolve around their feelings of inferiority compared to male individuals known as "Chads," who are portrayed as highly attractive and socially successful men who seemingly effortlessly attract romantic partners. Consequently, these forums often serve as outlets for expressing frustration and resentment, usually related to physical attractiveness, societal norms, and women's perceived preferences in partner selection. These discussions serve as an outlet for toxic ideologies and can reinforce patterns of blame and victimization that potentially contribute to a volatile atmosphere (Hoffman et al., 2020; O'Malley et al., 2022).

As public attention on Incels has grown, researchers have also begun to study the community more comprehensively, focusing on abusive language within Incel online communities (Farrell et al., 2019; Jaki et al., 2019), Incels as a political movement (O'Donnell and Shor, 2022), or mental health aspects of Incel community members (Broyd et al., 2023). Despite the widespread public perception that links Incels predominantly with violence, several studies found that topics discussed in Incel online communities cover a broad range of subjects that are not necessarily violence-related, e.g., discussions on high school and college courses and online gaming (Mountford, 2018). Nevertheless, the prevalence of abusive and discriminatory language in Incel forums remains a significant concern as it perpetuates a hostile environment that can both isolate members further and potentially escalate into real-world actions.

This paper follows up on how violent content is presented and evolves on *incels.is*, the largest Incel forum. We examine the prevalence and changes in violent content, analyzing specific forms of violence in individual posts and their progression over time at the user level. Our study classifies various types of violent content—explicit vs. implicit, and directed vs. undirected—using both manual labeling and Large Language Models (LLMs). We also assess the effectiveness of OpenAI's GPT-3.5 and GPT-4 models in annotating this content, exploring the challenges associated with these models.

While previous studies have explored the dynamics of violence in Incel forums broadly (cf., Farrell et al., 2019 with a focus on misogyny), there exists a significant research gap in understanding the specific forms of violence articulated in individual posts and the progression of such content at the user level (see the following paragraphs for a more detailed literature review). This distinction is critical as it allows us to determine the extent of violent content on the overall forum level and analyze users' trajectories of posting violent content in their posts, offering insights beyond the collective forum atmosphere.

We initially perform manual labeling on a subset of the data to establish a human baseline and ensure precise categorization for our violence typology, e.g., explicit vs. implicit violence; see Section 5.1. We then employ OpenAI's GPT-3.5 and GPT-4 APIs to classify a greater number of posts, enabling a comprehensive annotation of our dataset. We use the human baseline to assess the performance and ensure the accuracy of the categorization process, and discuss different experimental setups and challenges associated with annotating Incel posts. We then examine how the prevalence of violent content within the forum evolves for each category on the individual and forum levels.

2 Violent language in Incel communities

Within computational social science (Lazer et al., 2009), a diverse body of research has explored the multifaceted landscape of incel posts and forums. Natural language processing techniques have been employed to analyze the linguistic characteristics of Incel discourse, uncovering patterns of extreme negativity, misogyny, and self-victimization. Sentiment analysis, for instance, has illuminated the prevalence of hostile sentiments in these online spaces (Jaki et al., 2019; Pelzer et al., 2021), while topic modeling has unveiled recurrent themes and narratives driving discussions (Mountford, 2018; Baele et al., 2021; Jelodar and Frank, 2021). Other studies have focused on broader communities of misogynistic movements, tracking their evolution over time (Ribeiro et al., 2021a). These studies offer invaluable insights into the dynamics of Incel online communication and serve as a valuable foundation for more comprehensive research to fully understand the complexities of these communities.

Due to misogynistic and discriminating attitudes represented in Incel forums, research focusing on violent content constitutes the majority of academic studies related to this community. Pelzer et al. (2021), for instance, conducted an analysis of toxic language across three major Incel forums, employing a fine-tuned BERT model trained on ∼20,000 samples from various hate speech and toxic language datasets. Their research identified seven primary targets of toxicity: women, society, incels, self-hatred, ethnicities, forum users, and others. According to their analysis, expressions of hatred toward women emerged as the most prevalent form of toxic language (see Jaki et al., 2019 for a similar approach). On a broader level, Baele et al. (2021) employed a mix of qualitative and quantitative content analysis to explore the Incel ideology prevalent in an online community linked to recent acts of politically motivated violence. The authors emphasize that this particular community occupies a unique and extreme position within the broader misogynistic movement, featuring elements that not only encourage self-destructive behaviors but also have the potential to incite some members to commit targeted acts of violence against women, romantically successful men, or other societal symbols that represent perceived inequities.

The rise of research on the Incel community has also shifted the spotlight on users within the "Incelverse," driven by both qualitative and computational approaches. Scholars have embarked on demographic analyses, identifying prevalent characteristics, such as social isolation and prevailing beliefs within the Incelverse. A recent study on user characteristics in Incel forums analyzed users from three major Incel platforms using network analysis and community detection to determine their primary concerns and participation patterns. The findings suggest that users frequently interact with content related to mental health and relationships and show activity in other forums with hateful content (Stijelja and Mishara, 2023). Similarly, Pelzer et al. (2021) investigated the spread of toxic language across different incel platforms, revealing that the engagement with toxic language is associated with different subgroups or ideologies within the Incel communities. However, these studies have generally focused on smaller subsets of users and have not examined user behavior across the entirety of the *incels.is* forum. This gap in research is noteworthy, especially when broader studies indicate that content from hateful users tends to spread more quickly and reach a larger audience than non-hateful users (Mathew et al., 2019).

3 Categorizing violent language with language models

Effectively approaching harmful language requires a nuanced understanding of the diverse forms it takes online, encompassing elements such as "abusive language," "hate speech," and "toxic language," (Nobata et al., 2016; Schmidt and Wiegand, 2017). Due to their overlapping characteristics and varying degrees of subtlety and intensity, distinguishing between these types of content poses a significant challenge. In addressing this complexity, Davidson et al. (2017) define hate speech as "language that is used to express hatred toward a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group." Within the research community, this definition is further extended to include direct attacks against individuals or groups based on their race, ethnicity, or sex, which may manifest as offensive and toxic language (Salminen et al., 2020).

While hate speech has established itself as a comprehensive category to describe harmful language online, the landscape of hateful language phenomena spans a broad spectrum. Current research frequently focuses on specific subfields, e.g., toxic language, resulting in a fragmented picture marked by a diversity of definitions (Waseem et al., 2017; Caselli et al., 2020a). What unites these definitions is their reliance on verbal violence as a fundamental element in characterizing various forms of harmful language. Verbal violence, in this context, encompasses language that is inherently aggressive, demeaning, or derogatory, with the intent to inflict harm or perpetuate discrimination (Waseem et al., 2017; Soral et al., 2018; Kansok-Dusche et al., 2023). Building on this foundation, we adopt the terminology of "violent language" as it aptly encapsulates the intrinsic aggressive and harmful nature inherent in such expressions. To operationalize violent language, Waseem et al. (2017) have developed an elaborate categorization of violent language online. This categorization distinguishes between explicit and implicit violence, as well as directed and undirected forms of violence in online contexts. It will serve as the fundamental concept guiding the operationalization of violent speech in this paper (see Section 5.1). By addressing various degrees of violence, this concept encompasses language employed to offend, threaten, or explicitly indicate an intention to inflict emotional or physical harm upon an individual or group.

Supervised classification algorithms have proven successful in detecting hateful language in online posts. Transformer-based models like HateBERT, designed to find such language, have outperformed general BERT versions in English (Caselli et al., 2020a). While HateBERT has proven effective in recognizing hateful language, its adaptability to diverse datasets depends on the compatibility of annotated phenomena. Additionally, although these models exhibit proficiency in discovering broad patterns of hateful language, they are limited in discerning specific layers or categories, such as explicit or implicit forms of violence. Ultimately, the capability of BERT-based models to identify nuanced patterns of hateful language, including explicit and implicit forms, depends on the dataset used for fine-tuning.

Large Language Models (LLMs) present a promising alternative in scenarios where an evaluated, labeled dataset is unavailable. Recent research has found that using LLMs, particularly OpenAI's GPT variants, to augment small labeled datasets with synthetic data is effective in low-resource settings and for identifying rare classes (Møller et al., 2023). Further, Gilardi et al. (2023) found that GPT-3.5 outperforms crowd workers over a range of annotation tasks, demonstrating the potential of LLMs to drastically increase the efficiency of text classification. The efficacy of employing GPT-3.5 for text annotation, particularly in violent language, has been substantiated, revealing a robust accuracy of 80% compared to crowd workers in identifying harmful language online (Li et al., 2023b). Even in more challenging annotation tasks, like detecting implicit hate, GPT-3.5 demonstrated a commendable accuracy by correctly classifying up to 80% of the provided samples (Huang et al., 2023). Specifically for identifying misogynistic language, Morbidoni and Sarra (2023) found that GPT-3.5 outperformed supervised baselines. While these results showcase the effectiveness of GPT-3.5 in-text annotation, there remains room for improvement, particularly in evaluating prompts and addressing the inherent challenges associated with establishing a definitive ground truth in complex classification tasks like violent language classification (Li et al., 2023b).

Although smaller, fine-tuned, discriminative language models have shown superior performance in many cases (Abdurahman et al., 2023; Kocoń et al., 2023; Mu et al., 2023; Rathje et al., 2023), LLMs stand out for their adaptability across varied tasks and their capacity to incorporate context-specific information without additional training. Their ability to generate relevant insights without requiring highly specialized datasets offers a distinct advantage, bridging the gap in research contexts with limited data resources (Huang et al., 2023; Kocoń et al., 2023; Liu et al., 2023). Given the reduced technical complexity of making API calls compared to training a BERT model, LLMs may further provide enhanced accessibility for researchers across various disciplines, making data annotation more efficient and accessible (Li et al., 2023b).

4 Summary and study outline

The Incel community has become a subject of growing academic interest due to its complex interplay of extreme views and connections to real-world violence over the last few years. While previous research has illuminated linguistic and ideological dimensions of violent language in online forums, a forum-wide analysis based on different violence categories remains lacking. By further including the user level, this study makes it possible to distinguish between the overall evolution of violent speech prevalence within the forum and observe how the prevalence of violent content shifts for individual users over their active periods in the forum. Using manual annotation in conjunction with GPT-4 for this task offers a cost-effective and flexible approach, given its pre-trained capabilities for understanding a wide range of textual nuances. By classifying different categories of violent speech, we aim to determine whether various forms of violence exhibit differing levels of prevalence within the forum and if they evolve differently over time. Results can be used to assess the threat of violence in Incel forums and help tailor intervention strategies and content moderation to the specific nature of the content, enhancing the effectiveness of efforts to mitigate harm and promote safety within online communities.

5 Materials and methods

Besides *incels.is*, platforms like *looksmax.org* and Incel-focused subreddits are key communication channels for the Incel community. After Reddit officially banned the biggest Incel subreddit *r/incel* for inciting violence against women (Hauser, 2017; Ribeiro et al., 2021b), many users migrated to alternative platforms. With a self-proclaimed 22,000 members and over 10 million posts,[1] *incels.is* has become the leading Incel forum, making it an essential resource for understanding the community.

We scraped all publically available threads from *incels.is*, yielding over 400*k* threads with more than 6.9*M* posts. These were generated by 11,774 distinct users.[2] The web scraping was performed in May 2023. We collected the raw HTML responses from the website, focusing solely on text-based content and disregarding all non-text forms of media, primarily images, which were present in ∼6.3% of posts. Most of the media content was consistent with the posts, serving as supporting references. These included memes and short clips that reinforced the points made within the posts. Given the complexity of conducting a multimodal analysis, especially regarding the assessment of violence within memes, and our specific focus on directly expressed violent language in the text, we opted not to include such media content.

Next, we employed a three-step approach, leveraging the GPT-3.5[3] and GPT-4[4] APIs. A low temperature of 0.1 for both GPT-3.5 and GPT-4, which controls the randomness of the model's output, was chosen to ensure consistent and reliable responses while maintaining the model's creativity and flexibility (Jin et al., 2023). Note that a temperature above zero does not equate to non-deterministic behavior, as OpenAI now allows for seeded randomness in their models. Following a round of manual annotation of a random sample of 3,028 posts, we iterated prompts and the number of posts per query (batch size) for both models to align their classification of violent language with the human baseline. See Section 5.2 for more detail on the content of each prompt and their iterations. Finally, we used the best-performing prompt to classify an additional 45,611 posts, which we then analyzed for temporal patterns.[5]

5.1 Categories of violence

For categorizing different types of violent language, we used a slightly adapted version of Waseem et al. (2017)'s typology of abusive language. To bridge the challenges of navigating through the variety of definitions of hate speech, Waseem et al. (2017) have identified mutual characteristics that combine previous classifications of harmful content. This makes their typology a valid reference point when classifying violent language in online forums. This concept encompasses expressions that offend, threaten, or insult specific individuals or groups based on attributes such as race, ethnicity, or gender. It extends to language indicating potential physical or emotional harm directed at these individuals or groups. Additionally, differentiating between different types of violence (explicit vs. implicit and general vs. directed) helps gain a more nuanced picture of how violence manifests online. Following this classification scheme, we distinguish violent posts between explicitly and implicitly violent, as well as between directed, undirected/general, and self-directed violence. Each post is assigned an explicit/implicit and a directed/undirected/self-directed label. Table 1 provides examples for each category.

In the context of this classification framework, explicit violent language is a very straightforward and usually directly recognizable form of violence, e.g., racist or homophobic threats. While such language can vary in context, its harmful intent is generally unambiguous. Implicit violent language is subtler and more challenging to detect. It may involve ambiguous terms or sarcasm and lacks prominent hateful words, making it difficult for human annotators and machine learning algorithms to identify (cf., Caselli et al., 2020b for a similar distinction between explicit and implicit hate speech). On the second dimension, directed violent language refers to posts that target a specific individual, either within the forum or outside. General violent language, on the other hand, addresses a group of individuals. In the Incel context, for example, this type of language is often addressed toward women or a specific ethnic group. In our analysis, we focused solely on analyzing the textual content of posts without further differentiating between violent language targeted at particular genders or forum members.

[1] These numbers are extracted from the landing page and could not be reproduced in our attempts. Out of the 22,000 users, only 11,774 appear to have engaged by posting content.

[2] This includes 890 delete users. Once a user deletes their profile, the forum replaces all occurrences of their username with *Deleted User [XXX]* but retains the now anonymous posts.

[3] `gpt-3.5-turbo-0301` at temperature 0.1.

[4] `gpt-4-1106-preview` at temperature 0.1.

[5] Our final dataset, which includes 48,639 annotated Incel posts, as well as the code to replicate our findings, can be accessed at https://osf.io/ugtz8/.

TABLE 1 Classification examples for each category.

Category	Example
Non-violent	*Pleasure has become my main purpose of getting new hobbies, music mainly is maintaining me with life*
Explicit, directed	*I hope the whore gets raped then she can press actual sexual assault charges*
Explicit, general	*Cliquey, superficial western women deserve the rope, along with the Jews that made them this way*
Explicit, self-directed	*I'm so ugly I should be killed*
Implicit, directed	*He looks like he just got back from Auschwitz*
Implicit, general	*If only women weren't like this. But females love brutality, power, and domination, so in the end they get what they deserve*
Implicit, self-directed	*The world would be better off without men like me*

5.2 Augmented classification

Based on this classification scheme, two human annotators independently labeled a subsample of 3,028 posts. Annotation was performed by one of the authors of this study (Human A) and a research assistant familiar with the field of research (Human B), both being female. They were supported by an annotation manual providing definitions and examples for each violence category, as they are presented in Section 5.1 (general description) and Table 1 (classification examples). The annotators were tasked with reviewing each comment and categorizing it accordingly. They had the option to label comments as unclear. Those comments were subsequently excluded from the baseline sample. Additionally, the research assistant could discuss any open questions or ambiguous comments with the rest of the research team for clarification. By involving multiple annotators to establish a human baseline, we ensure a robust assessment of inter-coder consistency, enabling reliable comparisons with the models' annotations. We report Cohen's Kappa (κ) (Cohen, 1968) for intercoder reliability, as it accounts for chance agreement and adjusts for imbalanced data distributions. We also report weighted and macro F1 scores to assess the performance of the classification against the human baseline. The weighted F1 score differentiates between ground truth and predicted labels, making it a suitable metric for comparing the performance of the models against the human annotators. The macro F1 score, on the other hand, is an appropriate metric for inspecting the performance regarding underrepresented classes, as it computes the F1 score for each class individually and then takes the average of those scores. We used the manually annotated sample of 3,028 posts to evaluate the performance of different query prompts and batch sizes for both GPT-3.5 and GPT-4.

We started with a basic prompt employing role-prompting, a fundamental method in prompt engineering. Assigning the model a specific role, such as an expert, has been proven to be particularly effective in guiding the model's responses (Chen et al., 2023). In our prompts, we assigned the model the role of a "moderator of an online forum, aiming to moderate abusive and hateful language." The initial prompt only included information on our classification scheme, i.e., the categories of violence. Following best practices in prompt engineering (Chen et al., 2023; Liu et al., 2023; Hu et al., 2024), we successively added additional information and instructions to the prompt. Mu et al. (2023) demonstrated that enhancing GPT-3.5 prompts with task and label descriptions notably boosts its performance. In our case, including contextual information, specifically about the posts originating from an Incel forum, significantly improved the model's performance. To further improve the prompt, we kept looking at posts where the model's classification differed from the manual annotation and tried to find patterns in the misclassifications. Further, we used a form of self-instruction, presenting those misclassifications to the model itself and asking it for advice on improving the prompt. Finally, we included instructions to explain the reasoning behind the decision in the prompt, usually in the form of the most important words. The model must produce these hints before generating the label to ensure the model focuses on the right parts of the text and avoids *post-hoc* rationalization. The instruction to provide reasons is part of all final queries, which we provide in our OSF repository created for this study (see above).

GPT-3.5 allows for a maximum of 4*k* tokens for input and output, which can contain multiple messages with different roles, such as system and user messages. The LLM treats the system message as the central reference point for its behavior, while the user message is part of the ongoing conversation. Hence, we provide the task description and classification scheme in the system message and post them in the user message. GPT-4 has a context window of 128*k* tokens. Batching multiple posts into a single classification request made the speed and cost of the classification process manageable. Otherwise, reiterating the same system prompt for each post would substantially inflate the required number of tokens. We experimented with different batch sizes, ranging from 10 to 200 posts per batch.

In practice, each classification batch looked like

```
[System Message]
<Prompt>
The posts are:
```

followed by the batch of posts

```
[User Message]
Post 1: <Post 1>
Post 2: <Post 2>
...
```

GPT-4[6] introduces a novel JSON output mode, enabling the model to generate outputs in a JSON object format instead of

[6] Since we conducted this study, OpenAI has also released a version of GPT-3.5, which supports guaranteed JSON outputs.

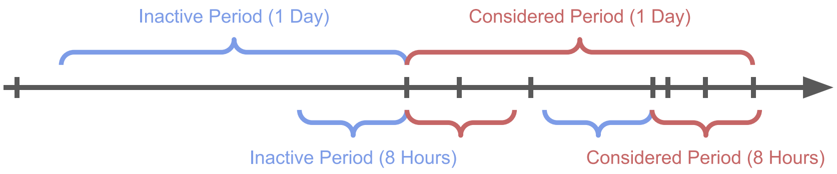

Inactive Period (1 Day)
Considered Period (1 Day)
Inactive Period (8 Hours)
Considered Period (8 Hours)

FIGURE 1
An exemplary timeline of user activity, ticks indicate posts. The user is 1-day-inactive once, while they are 8-h-inactive twice (blue). After being inactive for a given period, we observe the user's behavior for the same length of time (red).

plain text. The prompt must specify JSON schema. Our findings indicate that this mode does not alter the model's performance but significantly simplifies parsing its outputs. We used this mode for all our final classifications. Regarding data preprocessing, we limited our intervention to consolidating multiple new lines into one line. We found the model could handle the posts' raw text very well. Notably, it did not miss or confuse any post at any time. After iterating over the queries, we chose the one that performed best against the human baseline to annotate another 45,611 posts.

5.3 Time-based patterns of violent user posts

Figure 1 illustrates our method for distinguishing between active and inactive periods for individual users. We classify users as inactive if they have not made a post for at least T (e.g., 1 h, 1 day). Upon their return, we observe their behavior for the same duration, T, which we term a *session*. Posts can belong to multiple sessions since being inactive for 1 day inherently includes being inactive for 1 h, but not vice versa. This approach enables us to analyze the impact of inactivity on the prevalence of violent language in posts. To detect activity, we consider all posts, including unlabeled ones. As we cannot access viewing behavior, we need to limit our analysis of user activity to posting behavior. We repeat the following procedure for session lengths T of 1 h, 6 h, 12 h, 1 day, 1 week, 2 weeks, 30 days, and 180 days. The choice of these timespans allows us to capture the short-term, medium-term, and long-term effects of inactivity on the prevalence of violent language in posts. Due to small sample sizes, we do not report results for $T \geq 365$ days.

For each session length, we aggregate all annotated posts by their relative time since the user's first post of the session. We then divide the data into 12 equally sized bins and calculate the share of each category in each bin. To identify statistically significant trends in the prevalence of violent language, we conduct a χ^2 trend test on the resulting multinomial distribution over time for each timespan. The null hypothesis assumes no variation in the usage of violent language over time, and significance is evaluated against this assumption. To account for multiple testing, we apply Bonferroni correction, dividing the significance level by the number of tests performed (10 in our case). Significance levels are reported as $\hat{p} < 0.05$, ** indicating $\hat{p} < 0.01$, and *** indicating $\hat{p} < 0.001$ for the corrected significance levels $\hat{p}$ of the χ^2 trend tests.

To describe the trend direction, if any, we perform an ordinary least squares linear regression for each timespan, using the share of violent posts as the dependent variable and the time since the user's first post of the session as the independent variable. If inactivity reduces the prevalence of violent language, we would expect a statistically significant trend and a positive coefficient, indicating an increase in violent posts following a period of inactivity. Although the data suggests a multi-level model with random effects for users, with an average of four annotated posts per user, it is too sparse to estimate such a model reliably. Therefore, we rely on linear regression results instead.

6 Results

6.1 Performance of automated classification

Table 2 shows the pairwise Cohen's Kappa and weighted/macro F1 scores of all relevant annotation methods. Human A and B indicate the two human annotators, while GPT-3.5 presents the best-performing GPT-3.5 query and batch-size combination. GPT-4/X showcases the performance of GPT-4 with batch-size X for the best-performing query, each. Since the instruction to provide reasons for the models' decisions improved the results, it is part of all final queries.

GPT-3.5 is outperformed by GPT-4 in all metrics when comparing its labels against both human annotators. The rest of the analysis hence focuses on the performance of the different GPT-4 variants. The inter-annotator agreement between Human A and Human B, as measured by Cohen's Kappa (κ), is 0.69, indicating a substantial level of agreement. Their weighted and macro F1 scores of 0.85 and 0.77, respectively, illustrate apt performance with distinct yet varying levels of precision and recall in their annotations. Overall, Human A is less likely to label a post as violent than Human B, with 66% of posts labeled as violent by Human A, compared to 75% by Human B.

The analysis of different batch sizes reveals notable variations in the performance of GPT-4. Batch size 20 shows the highest agreement with Human A, as evidenced by its superior performance metrics. Conversely, batch size 100 aligns more closely with Human B, particularly regarding κ and weighted F1 scores. For the macro F1 score, batch size 50 exhibits the best alignment with Human B. The achieved Kappa values of 0.54 against Human A and 0.62 against Human B indicate moderate to

TABLE 2 Cohen's Kappa/weighted F1-score/macro F1-score.

	Human A	Human B	GPT3.5	GPT4/10	GPT4/20	GPT4/50	GPT4/100	GPT4/200
Human A	–	0.69/0.85/0.77	0.40/0.70/0.52	0.53/0.74/0.63	**0.54/0.76/0.63**	0.52/0.74/0.62	0.52/0.75/0.60	0.36/0.71/0.49
Human B	0.69/0.87/0.77	–	0.39/0.75/0.54	0.58/0.79/0.67	0.55/0.79/0.65	0.61/0.83/**0.67**	**0.62/0.84**/0.67	0.40/0.77/0.52
GPT3.5	0.40/0.67/0.52	0.39/0.68/0.54	–	**0.54/0.75/0.62**	0.49/0.72/0.59	0.49/0.71/0.59	0.47/0.70/0.56	0.37/0.67/0.48
GPT4/10	0.53/0.73/0.63	0.58/0.76/0.67	0.54/0.74/0.62	–	**0.75/0.86/0.78**	0.60/0.77/0.67	0.58/0.76/0.66	0.46/0.68/0.55
GPT4/20	0.54/0.75/0.63	0.55/0.77/0.65	0.49/0.74/0.59	**0.75/0.87/0.78**	–	0.69/0.83/0.74	0.65/0.81/0.71	0.44/0.71/0.51
GPT4/50	0.52/0.77/0.62	0.61/0.82/0.67	0.49/0.76/0.59	0.60/0.80/0.67	0.69/0.85/**0.74**	–	**0.72/0.87**/0.72	0.47/0.75/0.55
GPT4/100	0.52/0.78/0.60	0.62/0.84/0.67	0.47/0.77/0.56	0.58/0.80/0.66	0.65/0.84/0.71	**0.72/0.88/0.72**	–	0.51/0.80/0.59
GPT4/200	0.36/0.77/0.49	0.40/0.81/0.52	0.37/0.79/0.48	0.46/0.79/0.55	0.44/0.80/0.51	0.47/0.82/0.55	**0.51/0.83/0.59**	–

Bold numbers indicate the best performance per row, excluding humans. For the F1-scores, left indicates the ground truth, while top indicates predictions.

substantial agreement. They are similar to scores observed in other studies with comparable tasks (e.g., Haddad et al., 2019, although the authors achieved a higher agreement in one of three pairs of annotators). Macro and weighted F1 scores of 0.63 and 0.76 against Human A and 0.67 and 0.84 against Human B, respectively, indicate a high level of precision and recall in the classification of all three categories. Our weighted F1 scores of $\sim$ 0.8 align with those reported by other studies on the detection of violent language with GPT-3.5and GPT-4 (Huang et al., 2023; Li et al., 2023b). Very similar results hold for directed, undirected, and self-directed violence, which we do not report here for brevity.

Table 3 elucidates the overall label distribution across varying batch sizes, in which we observe a statistically significant shift. With increasing batch sizes, there is a discernible trend of fewer posts being classified as explicitly or implicitly violent and more as non-violent. This trend is more pronounced in the classification of implicit violence. Using a batch size of 10, 14% of all posts were labeled as implicitly violent. At batch size 200, this drops by 84%–2% of the total posts. The share of posts labeled as explicitly violent only decreases by 43% from 28 to 16%.

The label distribution generated at batch size 50 most closely aligns with the average distribution generated by the human annotators, suggesting an optimal batch size for achieving a human-like understanding of content classification. We further investigated the correlation between a post's position in a batch and its likelihood of being labeled violent. Posts positioned later in the batch were less frequently tagged as violent for larger batch sizes. This trend was consistent across different batch sizes but did not reach statistical significance. Due to the high level of agreement with humans A and B and the match in the overall class distribution, we used the labels generated by GPT-4 with batch size 50 for the remainder of our analysis.

6.2 Time-based patterns of violent user posts

Our results show that posts containing violent language, whether explicit or implicit, constitute 21.91% of all posts. 18.12% of posts contain explicit violent language, while implicit violent language accounts for 3.79% of forum posts. This leaves 78.09% of forum posts non-violent. The user analysis reveals a wide range of engagement levels. While an average of 586 posts per user appears substantial, a median of 24 posts per user indicates a very skewed distribution. About 10% of users maintained forum activity for at least 2.5 years at the time of scraping, highlighting their sustained engagement. Approximately 23.8% of forum users contributed only one post, underscoring the presence of occasional contributors within the platform's user community, while the 10% most active users have posted at least $1,152$ times. These findings underscore the diverse spectrum of user activity within the platform, ranging from highly engaged, long-term participants to sporadic contributors with limited involvement.

Figures 2A–H illustrates the temporal evolution of violent language in posts, with different time intervals as predictors for the prevalence of each violence category. Significance refers to the χ^2 tests for trends in proportions for each time interval. Regression lines are added to illustrate the overall trend. Our results indicate that within the 5 years since the forum's creation and our data collection (Figure 2A), violent language has been slightly increasing overall on a statistically significant level ($\beta = 0.006$ for explicit violence and $\beta = 0.0005$ for implicit violence). Plotting violence against time since the first post (Figure 2B), this trend is not reproduced. We find that the share of violent content remains relatively stable ($\beta = 0.0004$ for implicit violent language), with no significant changes over multiple years.

Figures 2C–J explore the impact of temporary inactivity on the prevalence of violent language. Each figure follows users for a period T, as indicated in the subfigures. The tracking takes place after these specific users have remained inactive for at least the same designated period. While we compute inactivity on the entire dataset, the plots only show annotated posts. From these figures, we observe varying results. We do not observe any statistically significant change in violent language for the 1-day (Figure 2F) and 180-day (Figure 2J) intervals. Within all other intervals, however, we observe a slight but significant increase in violent language overall, accompanied by a decrease in non-violent language. This trend is most prominent for the 2-week interval (Figure 2H) ($\beta = 0.01$ for explicit violence) and least pronounced for the 12-h window (Figure 2E) ($\beta = 0.004$ for explicit violence).

Figure 3 showcases the same analysis for the different categories of directedness. Since they do not contain any statistically relevant results, indicating that no substantial change in directed, general, or self-directed violence can be observed within the examined time

TABLE 3 Class distribution for different batch sizes *s*.

	s = 10	s = 20	s = 50	s = 100	s = 200	H-∅
Non-Violent	0.58 (1.00)	0.62 (1.07)	0.70 (1.20)	0.72 (1.24)	0.82 (1.41)	0.70 (1.21)
Explicit	0.28 (1.00)	0.26 (0.96)	0.21 (0.78)	0.22 (0.80)	0.16 (0.57)	0.22 (0.81)
Implicit	0.14 (1.00)	0.12 (0.81)	0.09 (0.61)	0.06 (0.41)	0.02 (0.16)	0.07 (0.52)

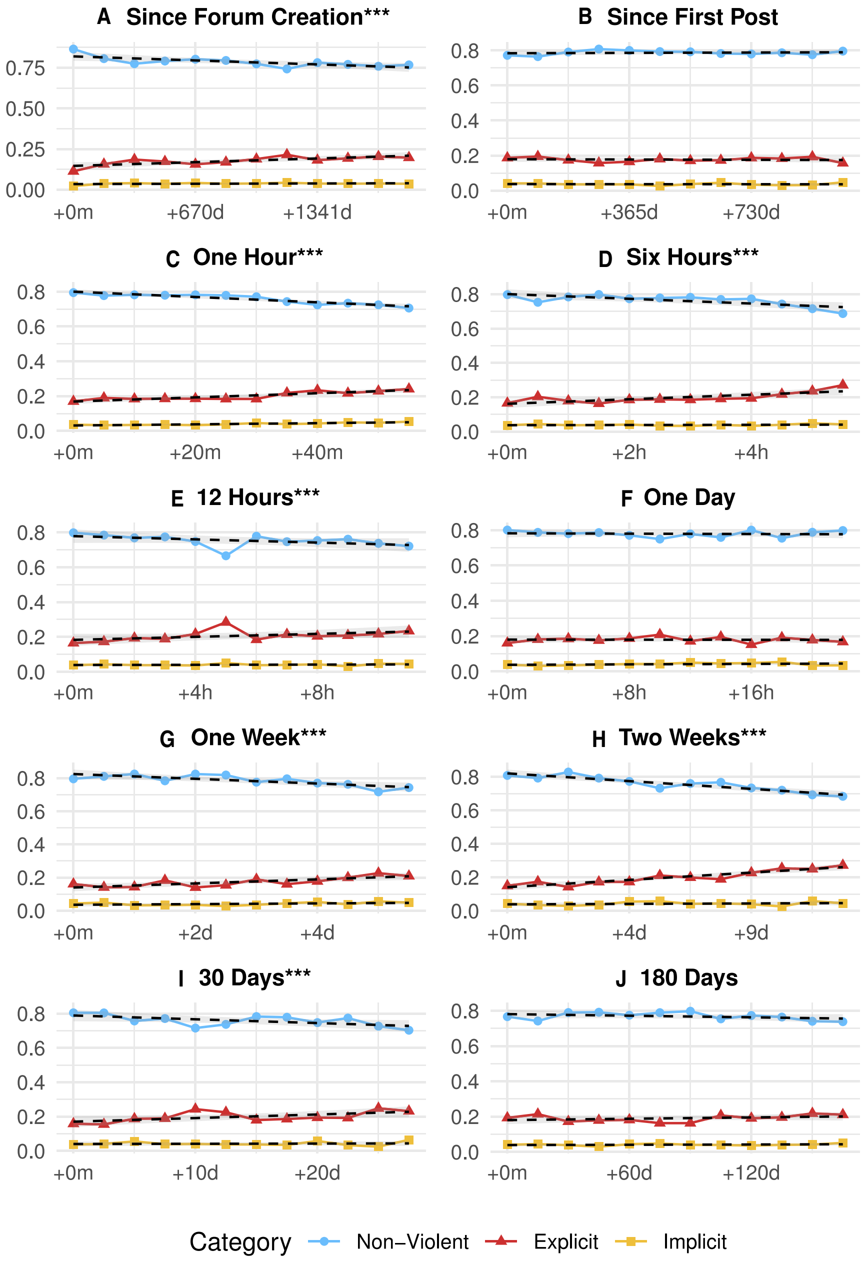


FIGURE 2
Evolution of the share of violent posts within the forum (non-violent vs. explicitly violent vs. implicitly violent). Each subfigure presents the results of separate trend tests for each time span, with the share of posts as the dependent variable and the elapsed time as the independent variable. Asterisks indicate the corrected level of significance (***$p < 0.001$). Linear regressions, represented by dashed lines, illustrate the overall trend. **(A)** Since Forum Creation***. **(B)** Since First Post. **(C)** One hour***. **(D)** Six hours***. **(E)** 12 hours***. **(F)** One day. **(G)** One week***. **(H)** Two week***. **(I)** 30 days***. **(J)** 180 days.

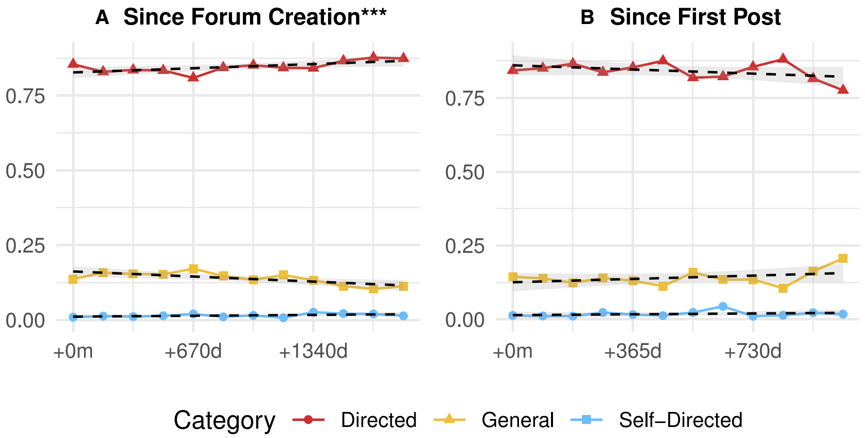


FIGURE 3
Evolution of the share of Violent Posts within the Forum (Directedness). Each subfigure presents the results of separate trend tests for each time span, with the share of posts as the dependent variable and the elapsed time as the independent variable. Asterisks indicate the corrected level of significance (***$p < 0.001$). Linear regressions, represented by dashed lines, illustrate the overall trend. **(A)** Since Forum Creation***. **(B)** Since First Post.

frames. Figure 3A reveals that the share of directed (i.e., targeted) violence increases significantly over time within the overall forum ($\beta = 0.004$). This is accompanied by a decrease in non-directed (general) violence ($\beta = -0.004$). Only considering aggregated user behavior for the time since the first post (Figure 3B), this trend appears reversed, with a slight decrease in directed violent language and an increase in general violent language. These changes, however, are not statistically significant. In this particular case, we also observe more variability in the share of violent content over time, making it harder to detect a pronounced trend. The share of self-harm content remains stable over time for both the forum and individual users (both $\beta = 0.001$).

6.3 GPT cost and speed

For the scope of this study, we spend a total of $\sim$ \$66 for OpenAI's APIs, including many iterations over all the human-annotated posts and the additionally annotated posts. Overall, we estimate GPT-3.5 and GPT-4 annotated $\sim$ 120,000 posts, including prompt iteration and batch-size experiments, which amounts to $\sim$ \$0.0005 per annotated post.

A key component of keeping the cost low is proper input batching. Our prompts are around 500 tokens long, whereas the average post is around 50 tokens long. Naively sending each post individually would have cost $550T \times \frac{\$0.01}{1{,}000T} = \$0.0055$ per post, or $\sim$ \$260 for the final set of 45,611 annotated posts. Increasing the batch-size to 50 yields a cost per batch of $3,000T \times \frac{\$0.01}{1{,}000T} = \$0.03$, or \$20 for the final set of 45,611 annotated posts. GPT-3.5 is significantly cheaper.

The average time for GPT-4 to annotate a single post was 1 s at batch size 50. The total time for GPT-4 to annotate 120,000 posts was $\sim$ 33 h. At the time of writing, OpenAI employs strong rate limiting on their APIs, preventing us from speeding up the process by running multiple instances in parallel, rendering time constraints the more limiting factor than cost. On multiple occasions, we experienced significant slow-downs in the APIs' response time, which are confirmed by OpenAI.[7] Moving our long-running jobs to the early European morning significantly improved the experience of working with the API.

7 Discussion

Our findings reveal that 21.91% of all posts feature violent language, either explicit or implicit. We detect a subtle but statistically significant increase in overall violence on *incels.is* within the forum. The same trend is found to be more pronounced in user activity for particular time intervals, particularly in user engagement within the 2-week period. Additionally, directed violence increases over time, while self-harm consistently remains very low within the forum. This shift implies a change in the type of aggression within the community, where users resort to more targeted hostility. While these trends are very subtle, they could be explained by evolving community norms, which become more tolerant toward specific forms of violent content over time, user familiarity, or moderation effects (Gibson, 2019). Our observations align with findings from other research indicating an increase of misogynistic content and violent attitudes within Incel communities (Farrell et al., 2019) and a general rise in hate speech across various online spaces (Laub, 2019; Zannettou et al., 2020; Peters, 2022). With 21.91% of the posts exhibiting violent language, it is crucial to recognize the substantial presence of violence within these forums, emphasizing the imperative to closely monitor such platforms and contemplate legislative actions, such as implementing stricter regulations on online hate speech and harassment.

[7] https://status.openai.com

7.1 Classifying violent language with GPT

Our study indicates that LLMs can produce a sensible starting point for the zero- and few-shot classification of violent content, providing a solid foundation for further analyses. Instructing the model to identify keywords that underpin its decisions has been particularly helpful, improving its accuracy and providing a valuable reference point for a more informed comparison with human evaluators. This strategy offered a transparent framework for comprehending the model's logic, serving as a neutral benchmark for evaluating its decision-making process. However, its performance was not assessed against a standardized corpus. Other models, such as HateBERT (Caselli et al., 2020a), may perform better on datasets they are fine-tuned on. Despite this, it's important to recognize that models specialized in hate speech, including HateBERT, face difficulties in accurately classifying varied forms of violent content (Poletto et al., 2021; Yin and Zubiaga, 2021). Additionally, these models may not be explicitly designed to differentiate within distinct categories of violent language, introducing an additional layer of complexity to the classification process. Given the subtle increase in the context of a wider rise in online violent language and the large size of our dataset, which might lead to artificial effects, we must interpret these trends with caution.

The difficulty in detecting certain kinds of violent language differs significantly between categories. While explicit acts of violence, such as physical assault or overt verbal abuse, may be easier to detect through keywords or contextual cues, implicit violence often manifests in more nuanced ways that are hard even for humans to identify (Strathern and Pfeffer, 2023). These include coded language that carries a threatening subtext. For instance, users often refer to Elliot Rodger, who committed an Incel-related attack in 2014, stating posts like "Just go ER." Also, Incel-specific language is frequently inherently derogative toward women, calling them *foids*, short for feminine humanoids, and uses racist slang, e.g., *Currycel* for an Indian Incel. Herein lies an apparent strength of LLMs, which proved to be very effective at finding and classifying these Incel-specific terms. Having been trained on large parts of the internet, it is very probable that the model has encountered these terms before and learned to associate them with violence. Although misclassifications may have occurred, particularly given the challenges inherent in detecting violence of this nature, their potential impact on our work is expected to be minimal. This is because our primary emphasis is on analyzing broad trends within the platform, which means that occasional inaccuracies in classification do not impact our analysis substantially.

While the change in sensitivity for different batch sizes might seem discerning at first, it also serves as a tuneable hyperparameter. We found that manipulating the model's overall sensitivity by altering the query instead of sensitivity toward a specific class is challenging during query optimization. The batch size allows us to adjust the sensitivity to match the overall label distribution of the human annotators. It is worth noting that this adjustment substantially impacts the model's speed and cost, as discussed in Section 6.3. While other authors find similar behavior, e.g., Li et al. (2023a), we did not find research primarily focusing on this particular aspect of prompt engineering and believe a more thorough investigation could be beneficial.

The substantial agreement between GPT-4 and human annotators, alongside its accessibility and cost-effectiveness, make GPT-4 a viable alternative to traditional embedding-based classification models. Our human annotator agreement scores are comparable to those reported in prior research (Haddad et al., 2019), underscoring the challenge of attaining a Cohen's Kappa score above 0.8. Still, our agreement might be influenced by methodological limitations within the annotation process. The study relied on just two annotators, potentially skewing the analysis due to the subjective nature of detecting violent content, especially regarding more complex categories. This limitation, though resulting from practical constraints, points to an opportunity for improvement. Expanding to a broader and more diverse pool of annotators could mitigate interpretation variances and enhance classification reliability, possibly employing majority voting to achieve more balanced and unbiased results.

This study emphasizes the effectiveness of leveraging LLMs, specifically GPT-4, as annotators in intricate classification tasks, especially in identifying different types of violent content in online communities—an inherently challenging task for human annotators. By providing reasons for its classification, GPT-4 can drastically streamline situations where human annotators are uncertain. While our results provide a baseline, further research is needed to evaluate the performance of GPT-4 compared to other hate-speech-focused models. Moreover, employing LLMs, such as GPT-4, to augment the annotated sample offers distinct advantages, as it spares human annotators from the potential emotional distress of reading content containing violence against specific individuals or groups.

7.2 Violence trends within the Incel community

The results of our study align with previous research focused on radicalization within the Incel community. As noted by Habib et al. (2022), users who become part of online Incel communities exhibit a 24% increase in submitting toxic content online and a 19% increase in the use of angry language. The authors conclude that Incel communities have evolved into platforms that emphasize expressing anger and hatred, particularly toward women. In the context of online discussions on conspiracy theories, Phadke et al. (2022) modeled various radicalization phases for Reddit users, identifying different stages in radicalization that could also be applied to the Incel context in future studies.

The analyses for the 1-day (Figure 2F) and the 180-day interval (Figure 2J), as well as the period that captures the overall time since the first post on an aggregated user level (Figure 2B), do not show any statistically significant changes over time. Particularly for the longer time intervals capturing more than a month, the forum's overall increase in violent language can thus not be reproduced. However, for shorter time periods of less than a month (e.g., 1 h, 6 h, 12 h, 1 week, and 2 weeks), the increase is significant, indicating that violent language tends to spike over shorter intervals. While the 1-day interval might initially appear as an anomaly, the deviation could result from chance or other factors not accounted for in the current analysis. Therefore, it might be valuable to validate these findings with additional data to determine the reliability of this particular observation. Future research could also benefit from advanced time-series analyses to uncover deeper insights into specific trends or events within the forum.

Our findings highlight the complex relationship between user engagement duration and violent content generation. Further research may be needed to explore the underlying motivations and dynamics driving these temporal patterns in online Incel discussions. Exploring broader time-related factors, including the potential impact of COVID-19-related dynamics on online behavior—especially relevant as the pandemic overlaps with our analysis of posts from the past 5 years—holds significant importance. This consideration stems from previous studies suggesting that the pandemic contributed to shifts in behavioral patterns, leading to increased radicalization across various online forums, including those associated with Incel communities (Davies et al., 2021). Additional (computational) studies and in-person surveys with community members could provide deeper insights and guide interventions to foster more positive interactions within the forum.

Additionally, individual beliefs and attitudes of users, including their affiliation with specific subgroups within the Incel community that vary in extremism, could correlate with observed trends. It is plausible that belonging to a particular ideological subgroup may influence how members express violent content. These ideologies may affect the time spent online, the duration of active online engagement, and the posting frequency, making them relevant factors to consider in this context. It might be fruitful to examine whether the observed trends are more pronounced among specific subgroups within the community or whether they are evenly distributed over the user population. Although our results are too subtle to account for an actual pattern of radicalization, it might also be interesting to build upon these results and dive more deeply into the content of violent posts within specific time windows to see if phases of escalation can be identified.

Understanding the driving factors behind the increase in violent speech is essential to address and mitigate overall aggression levels within the forum. Investigating whether this generalized violence specifically targets certain groups, such as women or non-Incel men, could provide valuable insights into the dynamics of hostility within the community (Pelzer et al., 2021). In light of these findings, refining our analytical framework could enhance the precision of our results. Although Waseem et al. (2017)'s typology offers a solid starting point, an Incel-specific framework, such as the one proposed by Pelzer et al. (2021), which categorizes posts based on their targets—ranging from women and society to Incels themselves and ethnic groups—might yield more nuanced insights. Future research should consider these distinctions to better understand the variability in the direction of violent content. This is particularly pertinent given the observation that a significant portion of violent posts targets not only women but also "Chads," "normies," and society at large, suggesting a broad spectrum of animosity that extends beyond a single focal group.

In summary, our investigation into the evolution of violent speech within Incels forums and the intricate dynamics of ideology-driven aggression underscores the complexity of online radicalization. While we offer an overview of the evolution of specific subcategories of violence, the significance of temporal factors, ideological underpinnings, and community-specific behaviors in the online violence landscape necessitates further research. Our analysis has been limited to textual data, yet incorporating other forms of data, such as memes and short videos, through a multimodal analysis could enhance our insights (Gomez et al., 2020; Kiela et al., 2020; Bhandari et al., 2023; Chhabra and Vishwakarma, 2023). Despite the technical challenges associated with image recognition and determining the level of violence in these media, a multimodal approach in future research promises a more comprehensive understanding of the factors driving violent speech in digital communities.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Our data processing procedures did not involve any handling of private information. The user names obtained during the scraping process do not contain sufficient and valid information to make conclusions about online users' personal information. The same is true for posts directly cited in this paper. Despite offensive language in these posts, we have included them to enhance the clarity and understanding of our categorization for our readers. Both human annotators were informed of and aware of the potentially violent content in Incel posts before the annotation process, with the ability to decline annotation at any time. Both coders were given the chance to discuss any distressing material encountered during annotation. As discussions on the potential trauma or adverse effects experienced by annotators while dealing with hate speech become more prevalent (Kennedy et al., 2022), we have proactively provided annotators with a recommended written guide designed to aid in identifying changes in cognition and minimizing emotional risks associated with the annotation process.

Author contributions

DM: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. MS: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. NG: Writing – original draft, Writing – review & editing, Supervision. JP: Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., et al. (2023). Perils and opportunities in using large language models in psychological research. *PsyArXiv*. [Preprint]. doi: 10.31234/osf.io/d695y

Baele, S. J., Brace, L., and Coan, T. G. (2021). From "Incel" to "Saint": analyzing the violent worldview behind the 2018 Toronto attack. *Terror. Political Violence* 33, 1667–1691. doi: 10.1080/09546553.2019.1638256

Bhandari, A., Shah, S. B., Thapa, S., Naseem, U., and Nasim, M. (2023). "Crisishatemm: multimodal analysis of directed and undirected hate speech in text-embedded images from Russia-Ukraine conflict," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC: IEEE), 1993–2002. doi: 10.1109/CVPRW59228.2023.00193

Brace, L. (2021). *A short introduction to the involuntary celibate sub-culture*. Centre for Research and Evidence on Security Threats. Available online at: https://crestresearch.ac.uk/resources/a-short-introduction-to-the-involuntary-celibate-sub-culture/ (accessed March 10, 2024).

Broyd, J., Boniface, L., Parsons, D., Murphy, D., and Hafferty, J. D. (2023). Incels, violence and mental disorder: a narrative review with recommendations for best practice in risk assessment and clinical intervention. *BJPsych Adv*. 29, 254–264. doi: 10.1192/bja.2022.15

Caselli, T., Basile, V. Mitrović, J., and Granitzer, M. (2020a). Hatebert: retraining bert for abusive language detection in English. *arXiv* [Preprint]. arXiv:2010.12472. doi: 10.48550/arXiv.2010.12472

Caselli, T., Basile, V. Mitrović, J., Kartoziya, I., and Granitzer, M. (2020b). "I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 6193–6202.

Chen, B., Zhang, Z. Langrené, N., and Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv* [preprint]. arXiv:2310.14735. doi: 10.48550/arXiv.2310.14735

Chhabra, A., and Vishwakarma, D. K. (2023). A literature survey on multimodal and multilingual automatic hate speech identification. *Multimed. Syst*. 29, 1203–1230. doi: 10.1007/s00530-023-01051-8

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull*. 70, 213. doi: 10.1037/h0026256

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proce. Int. AAAI Conf. Web Soc. Media* 11, 512–515. doi: 10.1609/icwsm.v11i1.14955

Davies, G., Wu, E., and Frank, R. (2021). A witch's brew of grievances: the potential effects of COVID-19 on radicalization to violent extremism. *Stud. Confl. Terror*. 46, 1–24. doi: 10.1080/1057610X.2021.1923188

Farrell, T., Fernandez, M., Novotny, J., and Alani, H. (2019). "Exploring misogyny across the manosphere in reddit," in *Proceedings of the 10th ACM Conference on Web Science* (New York, NY: ACM), 87–96. doi: 10.1145/3292522.3326045

Gibson, A. (2019). Free speech and safe spaces: how moderation policies shape online discussion spaces. *Soc. Media Soc*. 5:2056305119832588. doi: 10.1177/2056305119832588

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv* [Preprint]. arXiv:2303.15056. doi: 10.48550/arXiv.2303.15056

Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. (2020). "Exploring hate speech detection in multimodal publications," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Snowmass, CO: IEEE), 1470–1478. doi: 10.1109/WACV45572.2020.9093414

Habib, H., Srinivasan, P., and Nithyanand, R. (2022). Making a radical misogynist: how online social engagement with the manosphere influences traits of radicalization. *Proc. ACM Hum. Comput. Interact*. 6(CSCW2), 1–28. doi: 10.1145/3555551

Haddad, H., Mulki, H., and Oueslati, A. (2019). "T-hsab: a Tunisian hate speech and abusive dataset," in *International Conference on Arabic Language Processing* (Cham: Springer), 251–263. doi: 10.1007/978-3-030-32959-4_18

Hauser, C. (2017). Reddit bans 'Incel' group for inciting violence against women. *The New York Times*. Available online at: https://www.nytimes.com/2017/11/09/technology/incels-reddit-banned.html (accessed March 10, 2024).

Hoffman, B., Ware, J., and Shapiro, E. (2020). Assessing the threat of incel violence. *Stud. Confl. Terror*. 43, 565–587. doi: 10.1080/1057610X.2020.1751459

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., et al. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *J. Am. Med. Inform. Assoc*. ocad259. doi: 10.1093/jamia/ocad259

Huang, F., Kwak, H., and An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *arXiv preprint arXiv*:2302, 07736. doi: 10.1145/3543873.3587368

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., and De Pauw, G. (2019). Online hatred of women in the Incels.me forum: linguistic analysis and automatic detection. *J. Lang. Aggress. Conf*. 7, 240–268. doi: 10.1075/jlac.00026.jak

Jelodar, H., and Frank, R. (2021). Semantic knowledge discovery and discussion mining of Incel online community: topic modeling. [*arXiv*] [Preprint]. arXiv:2104.09586. doi: 10.48550/arXiv.2104.09586

Jin, Y., Li, D., Yong, A., Shi, J., Hao, P., Sun, F., et al. (2023). RobotGPT: robot manipulation learning from ChatGPT. *arXiv* [Preprint]. arXiv:2312.01421. doi: 10.48550/arXiv.2312.01421

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., et al. (2023). A systematic review on hate speech among children and adolescents: definitions, prevalence, and overlap with related phenomena. *Trauma Violence Abuse* 24, 2598–2615. doi: 10.1177/15248380221108070

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., et al. (2022). Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Lang. Resour. Eval*. 56, 1–30. doi: 10.1007/s10579-021-09569-x

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., et al. (2020). "The hateful memes challenge: detecting hate speech in multimodal memes," in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC: Curran Associates Inc), 14.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., et al. (2023). ChatGPT: jack of all trades, master of none. *Inform. Fusion* 99:101861. doi: 10.1016/j.inffus.2023.101861

Laub, Z. (2019). *Hate Speech on Social Media: Global Comparisons*. Washington, DC: Council on Foreign Relations, 7.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabsi, A.-L., Brewer, D., et al. (2009). Computational social science. *Science* 323, 721–723. doi: 10.1126/science.1167742

Li, J., Zhao, R., Yang, Y., He, Y., and Gui, L. (2023a). "OverPrompt: enhancing ChatGPT through efficient in-context learning," in *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Li, L., Fan, L., Atreja, S., and Hemphill, L. (2023b). "HOT" ChatGPT: the promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv* [Preprint]. arXiv:2304.10619. doi: 10.48550/arXiv.2304.10619

Lindsay, A. (2022). Swallowing the black pill: involuntary celibates'(Incels) anti-feminism within digital society. *Int. J. Crime Justice Soc. Democr*. 11, 210–224. doi: 10.5204/ijcjsd.2138

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G., et al. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv*. 55, 1–35. doi: 10.1145/3560815

Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. (2019). "Spread of hate speech in online social media," in *Proceedings of the 10th ACM Conference on Web Science* (New York, NY: ACM), 173–182. doi: 10.1145/3292522.3326034

Møller, A. G., Dalsgaard, J. A., Pera, A., and Aiello, L. M. (2023). Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv*:2304, 13861.

Morbidoni, C., and Sarra, A. (2023). "Can LLMs assist humans in assessing online misogyny? Experiments with GPT-3.5," in *CEUR Workshop Proceedings, Vol. 3571* (CEUR-WS), 31–43.

Mountford, J. (2018). Topic modeling the red pill. *Soc. Sci.* 7:42. doi: 10.3390/socsci7030042

Mu, Y., Wu, B. P., Thorne, W., Robinson, A., Aletras, N., Scarton, C., et al. (2023). Navigating prompt complexity for zero-shot classification: a study of large language models in computational social science. *arXiv* [preprint]. arXiv:2305.14310. doi: 10.48550/arXiv.2305.14310

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web* (Geneva), 145–153. doi: 10.1145/2872427.2883062

O'Donnell, C., and Shor, E. (2022). "This is a political movement, friend": WHY "incels" support violence. *Br. J. Sociol.* 73, 336–351. doi: 10.1111/1468-4446.12923

O'Malley, R. L., Holt, K., and Holt, T. J. (2022). An exploration of the involuntary celibate (Incel) subculture online. *J. Interpers. Violence* 37, NP4981–NP5008. doi: 10.1177/0886260520959625

Pelzer, B., Kaati, L., Cohen, K., and Fernquist, J. (2021). Toxic language in online incel communities. *SN Soc. Sci.* 1, 1–22. doi: 10.1007/s43545-021-00220-8

Peters, M. A. (2022). Limiting the capacity for hate: hate speech, hate groups and the philosophy of hate. *Educ. Philos. Theory* 54, 2325–2330. doi: 10.1080/00131857.2020.1802818

Phadke, S., Samory, M., and Mitra, T. (2022). Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. *Proc. Int. AAAI Conf. Web Soc. Media* 16, 770–781. doi: 10.1609/icwsm.v16i1.19333

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Eval.* 55, 477–523. doi: 10.1007/s10579-020-09502-8

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., Van Bavel, J. J., et al. (2023). GPT is an effective tool for multilingual psychological text analysis. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/sekf5

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., and Long, S. (2021a). The evolution of the manosphere across the web. *Proc. Int. AAAI Conf. Web Soc. Media* 15, 196–207. doi: 10.1609/icwsm.v15i1.18053

Ribeiro, M. H., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., and De Cristofaro, R. (2021b). Do platform migrations compromise content moderation? Evidence from r/the_donald and r/incels. *Proc. ACM Hum. Comput. Interact.* 5(CSCW2), 1–24. doi: 10.1145/3476057

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-g., Almerekhi, H., and Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Hum.-Centric Comput. Inf. Sci.* 10, 1–34. doi: 10.1186/s13673-019-0205-6

Schmidt, A., and Wiegand, M. (2017). "A survey on hate speech detection using natural language processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (Valencia), 1–10. doi: 10.18653/v1/W17-1101

Soral, W., Bilewicz, M., and Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggress. Behav*. 44, 136–146. doi: 10.1002/ab.21737

Stijelja, S., and Mishara, B. L. (2023). Characteristics of Incel forum users: social network analysis and chronological posting patterns. *Stud. Conf. Terror*. 1–21. doi: 10.1080/1057610X.2023.2208892

Strathern, W., and Pfeffer, J. (2023). *Identifying Different Layers of Online Misogyny*. doi: 10.48550/arXiv.2212.00480

Texas Department of Public Safety (2020). *Texas Domestic Terrorism Threat Assessment*. Austin, TX.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: a typology of abusive language detection subtasks. *arXiv* [Preprint]. arXiv:1705.09899. doi: 10.48550/arXiv.1705.09899

Yin, W., and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Comput. Sci.* 7:e598. doi: 10.7717/peerj-cs.598

Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S., and Stringhini, G. (2020). "Measuring and characterizing hate speech on news websites," in *Proceedings of the 12th ACM conference on web science* (New York, NY: ACM), 125–134. doi: 10.1145/3394231.3397902

4.7 Study 7: More Skin, More Likes! Measuring Child Exposure and User Engagement on TikTok

Authors

Miriam Schirmer, Angelina Voggenreiter, Jürgen Pfeffer

In

ArXiv, https://doi.org/10.48550/arXiv.2408.05622, Published 08-2024

Abstract

Sharenting, the practice of parents sharing content about their children on social media platforms, has become increasingly common, raising concerns about children's privacy and safety online. This study investigates children's exposure on TikTok, offering a detailed examination of the platform's content and associated comments. Analyzing 432,178 comments across 5,896 videos from 115 user accounts featuring children, we categorize content into Family, Fashion, and Sports. Our analysis highlights potential risks, such as inappropriate comments or contact offers, with a focus on appearance-based comments. Notably, 21% of comments relate to visual appearance. Additionally, 19.57% of videos depict children in revealing clothing, such as swimwear or bare midriffs, attracting significantly more appearance-based comments and likes than videos featuring fully clothed children, although this trend does not extend to downloads. These findings underscore the need for heightened awareness and protective measures to safeguard children's privacy and well-being in the digital age.

Contribution of Thesis Author

Theoretical conceptualization, data curation, methodological design, formal analysis, visualization, manuscript writing, revision, and editing.

More Skin, More Likes! Measuring Child Exposure and User Engagement on TikTok

Miriam Schirmer[a], Angelina Voggenreiter[a] and Jürgen Pfeffer[a]

[a]School of Social Sciences and Technology, Technical University of Munich, Germany

Author Note

Correspondence concerning this article should be addressed to Miriam Schirmer, School of Social Sciences and Technology, Technical University of Munich, Germany. Email: miriam.schirmer@tum.de

Abstract

Sharenting, the practice of parents sharing content about their children on social media platforms, has become increasingly common, raising concerns about children's privacy and safety online. This study investigates children's exposure on TikTok, offering a detailed examination of the platform's content and associated comments. Analyzing 432,178 comments across 5,896 videos from 115 user accounts featuring children, we categorize content into Family, Fashion, and Sports. Our analysis highlights potential risks, such as inappropriate comments or contact offers, with a focus on appearance-based comments. Notably, 21% of comments relate to visual appearance. Additionally, 19.57% of videos depict children in revealing clothing, such as swimwear or bare midriffs, attracting significantly more appearance-based comments and likes than videos featuring fully clothed children, although this trend does not extend to downloads. These findings underscore the need for heightened awareness and protective measures to safeguard children's privacy and well-being in the digital age.

Keywords: TikTok, children, sharenting, exposure, parents

More Skin, More Likes! Measuring Child Exposure and User Engagement on TikTok

Introduction

In a recent investigation of more than 2.1 million Instagram posts associated with content that shows children, New York Times journalists have discovered a "marketplace of girl influencers" typically managed by the girls' mothers. Often portraying young girls in exposing attire, reporters found that these posts draw the attention of men sexually attracted to children and have shed light on the complex and potentially exploitative dynamics behind online content of minors (Valentino-DeVries & Keller, 2024). This investigation also sheds light on the broader digital landscape, including platforms like TikTok. With its short-form video format and widespread popularity, TikTok has become a significant platform for self-expression, creativity, and social interaction. As of the latest available data, TikTok boasts a staggering user base, with an estimated 900 million in 2024 (Statista, 2024).

Despite TikTok's explicit age restrictions, prohibiting children under 14 years old from creating accounts, the platform remains a magnet for younger users. Consequently, children actively generate content that reflects their interests, talents, and daily lives, shaping the platform's dynamic and content ecosystem (Pedrouzo & Krynski, 2023). In the United States of America, for example, the largest proportion of TikTok Users (25%) are between 10 and 19 years old (Howarth, 2024). Similarly, in the UK, almost one-third of 5-7-years-olds, half of 8-11-years-olds and more than two-thirds of 12-15-years-olds use TikTok (Ofcom, 2022).

This study explores the complex interactions between children and TikTok, focusing on how the platform's environment can impact young users. Given the limited research on children's exposure on TikTok, our goal is to provide an overview of content featuring minors on the platform, identify the most common types of content involving children, and categorize these types accordingly. We also analyze the comments on these videos, qualitatively assessing the nature of viewer feedback and any notable aspects. Additionally, we examine specific risks for children on social media, particularly the potential for sexual exploitation on TikTok (Are, 2023). Considering previous findings that skin exposure increases user engagement (Kernen et al., 2021; Ramsey & Horan, 2018), we investigate how attire and skin exposure influence likes and comments, highlighting the vulnerability of young users to exploitative behaviors.

Background

Sharenting on Social Media

"Sharenting", that is "parents sharing" information about their children on social media, has become a frequent phenomenon in our digitalized society (Amon et al., 2022; Cataldo et al., 2022; Verswijvel et al., 2019; Yegen & Mondal, 2021). This practice encompasses a wide range of activities, from publishing photographs and videos to posting anecdotes and milestones. While sharenting enables families to keep in touch and share joyous moments with a broader community, it also raises questions about privacy, consent, and the implications of establishing a digital footprint for children early in their lives (Stephenson et al., 2024; Walrave et al., 2022). Although the overall amount of sharenting is unknown; Amon et al. (2022), for example, surveyed almost 500 parents living in the United States who regularly use social media and found that almost 90% of them have distributed content of their children on social media platforms. In a survey involving 2,900 Spanish schoolchildren, nearly 20% of children reported that their parents had posted information about them (Garmendia et al., 2022). Exploring the reasons behind parents' decisions to post content featuring their children online, Latipah et al. (2020) interviewed 10 parents about their activity on social media platforms. The authors identified four primary reasons for sharing: seeking affirmation and social support, demonstrating caregiving capabilities, engaging in social participation, and documenting family experiences (cf., Amon et al. (2022) for similar results). Campana et al. (2020) found that fathers on Instagram share photos with their children to both connect with others and showcase their family lives.

Sharenting activities largely imperil children's right to privacy, especially as few parents seem to ask their children for permission to disclose information about them. In a survey with 1,460 Czech and Spanish parents, of whom around 80% published pictures of their child, only 20% obtained their child's consent (Kopecky et al., 2020). Some parents even deliberately ignored the will of their child, as reported in interviews with 12-14-years olds (Ouvrein & Verswijvel, 2019). In another study, around 4% children mentioned experiencing negative outcomes out of sharenting, such as receiving negative or hurtful comments from others, and 12% of children requested their parents to remove the information shared about them (Garmendia et al., 2022). Such information often included sensitive details such as th child's first name or the date of birth (Brosch, 2016). In addition, the child-related content shared by parents often violates the child's dignity. Stormer et al. (2023) identified 184 videos of 35 TikTok accounts containing psychological maltreatment towards children through caregivers, such as yelling at, ignoring, and pranking them. They found that these videos received higher engagement in form of likes, views, and comments than those without maltreatment of children. Even more, in the investigation of Brosch (2016), about 45% of the parents posted photos that could be considered inappropriate, such as images of their child in the nude or semi-nude, typically taken during baths or beach visits, involving children under 3 years old. Similarly, in the study of Kopecky et al. (2020), 20% of the parents admitted having posted photos in which their children were partially exposed, and 3.5% of the Czech sample (n=1,093) had shared photographs of their naked child at a neonatal or infant stage online. These finding are in line with Stephenson et al. (2024)'s conclusion that parents frequently share intimate posts aiming for viral content.

Sexualization of Children on TikTok

Previous research has linked skin exposure on social media to increased user engagement. On Instagram, for example, more revealing photos tend to attract more likes, as shown by Park and Lee (2017). Additionally, a study on young women found that although self-sexualization rates in photos were relatively low, sexualized images garnered more likes and followers (Ramsey & Horan, 2018). Non-government organizations have also warned that algorithms may prioritize images showing more skin (Kayser-Bril et al., 2020). However, this trend has not been specifically validated for TikTok or for content involving children. While inappropriate content of children seems to be prevalent on many social media platforms, the video-sharing platform TikTok has been consistently criticized in the past for enabling the sexual exploitation of children and adolescents. Whereas the general level of suggestive and sexualized behavior is high amongst the TikTok community, this behavior is imitated by minor users, who perform sensual or provocative dances or show themselves in swimsuits or underwear (Suárez-Álvarez et al., 2023). Users on TikTok react to minors' videos by sending sexually explicit comments and requests, as both interviews with children and adolescents (Soriano-Ayala et al., 2023), as well as a BBC investigation of TikTok videos, have shown (Silva, 2019). Comments of this type often focus on the physical appearance of children, complimenting their looks, and sometimes extend to inappropriate interactions, such as invitations for further personal engagement or offers to meet up, highlighting a dangerous aspect of online behavior towards minors (Silva, 2019). Even more, a Forbes investigation has revealed child sexual abuse material being shared within private TikTok accounts (Levine, 2022).

Most investigations into online exploitation of children have been carried out by investigative journalists from major newspapers (Barry et al., 2021; Levine, 2022; Silva, 2019; Valentino-DeVries & Keller, 2024), with scientific research on the subject being scarce. Existing academic work primarily consists of qualitative reports and case studies (Khan & Bhattacharjee, 2022; Soriano-Ayala et al., 2023), while quantitative analysis—essential for understanding the scale and patterns of such issues—remains notably insufficient.

Child Protection Mechanisms on Social Media Platforms

Investigating child protection mechanisms on social media is vital. TikTok's guidelines prohibit harassment of minors through public or private interactions and commit to reporting content that endangers children to law enforcement (TikTok, 2024). However, despite removing most sexually explicit comments within 24 hours of reporting, TikTok has not consistently eliminated messages that are inappropriate for children (Silva, 2019). At the same time, TikTok has increased efforts in content moderation of sexually explicit language on their platform, e.g., deleting videos that contain captions such as "sex" or "lesbian" (Steen et al., 2023). Research has shown how these automated detection algorithms can be circumvented by using alternative words and negative implications of the automated deletion of sexual content that might not be harmful but is aimed at educating young users (Steen et al., 2023). TikTok has faced criticism for inadequately protecting children's privacy, often enhancing protections only in response to public outrage and regulatory pressure, rather than proactively as recommended (Polito et al., 2022). Additionally, TikTok has been accused of prioritizing profit over public interest, with calls for a better balance between online freedoms and responsibilities, emphasizing children's rights and social justice over financial gains (Salter & Hanson, 2021). General education and education tailored explicitly to sexual content have been suggested as a countermeasure by many scholars

(Bozzola et al., 2022), however, mostly in the context of children being exposed to such content instead of them being exposed directly. The same is true for protection algorithms that are being used to restrict children's content (Badillo-Urquiola et al., 2019; Taylor & Brisini, 2024).

We address the critical issue of children's sexual exploitation and exposure on TikTok, emphasizing the need for targeted strategies and policies to protect young users from harmful content and interactions. While children's encounters with sexual content on platforms like TikTok are well-documented (Barry et al., 2021), this study specifically assesses the risks associated with such exposures, rather than focusing on the broader issue of exposure to inappropriate content during online activities.

Summary and Research Questions

Previous research has shown the risks of sexual exploitation of children's images and videos on social media. While TikTok restricts explicitly exposing content, its handling of concerns about child exposure remains criticized. Due to limited mechanisms and research in this area, we conducted a thorough study to assess the extent of children's exposure on TikTok, guided by the following questions: How are children portrayed on TikTok (RQ1)? How do users react to videos of children (RQ2)? Can TikTok content featuring children be further traced to private devices and other websites (RQ3)? Finally, is there a relationship between the nature of the video content—such as the degree of exposure—and user reactions (RQ4)?

Methods

Data

Since TikTok's guidelines prohibit users under 13 from holding accounts (TikTok, 2024), this study focuses on accounts managed by adults, typically parents, that feature children under 13. We created our dataset by searching for accounts using keywords like "child" or "kid" and additional terms suggested by TikTok's search console (e.g., "family," "child model"). We collected the IDs of the first 100 videos from each matching account and excluded those not featuring children under 13. We determined age using visual cues or age information provided in videos or profiles. For each video, we gathered the first 500 comments and relevant metadata, focusing on English-language comments to ensure consistency and minimize misinterpretation. Our final dataset includes 432,178 comments from 5,896 videos across 115 TikTok accounts. Table 1 provides an overview of the dataset.

Table 1

Number of Accounts, Videos, and Comments per Account Category

Category	n_accounts		n_videos		n_comments	
	count	%	count	%	count	%
Family	78	67.83%	4,073	69.08%	340,921	78.88%
Fashion	21	18.26%	1,336	22.66%	83,708	19.37%
Sports	16	13.91%	487	8.26%	7,549	1.75%
Total	115	100.00%	5,896	100.00%	432,178	100.00%

Recurrent Themes

Literature on concurrent TikTok topics and themes indicates that a majority of content revolves around comedy, sports and fitness, beauty, and popular TikTok dances and challenges (Pryde & Prichard, 2022; Vaterlaus & Winter, 2021). To answer RQ1 ("How are children portrayed on Social Media"), we first used an explorative approach, carefully going through the videos and collecting frequently occurring themes, as well as noteworthy trends and TikTok challenges. This way, we identified recurring themes consistent with previous research, such as beauty and sports (Pryde & Prichard, 2022). Videos featuring children were less about storytelling and more focused on playful, daily routines shared by parents. Therefore, we introduced a third category, "Family," encompassing videos depicting day-to-day family life and accounts dedicated to this purpose. This led to the final categorization of accounts into the following groups (see Figure 1 for an illustration of typical content):

- Family: content predominantly revolving around family-oriented themes on TikTok, such as parents showcasing their daily routines with their children or playing together.
- Sports: content of children engaged in sports, predominantly gymnastics, and dancing.
- Fashion: fashion-related clips typically featuring a single child showcasing different outfits and modeling poses in front of the camera.

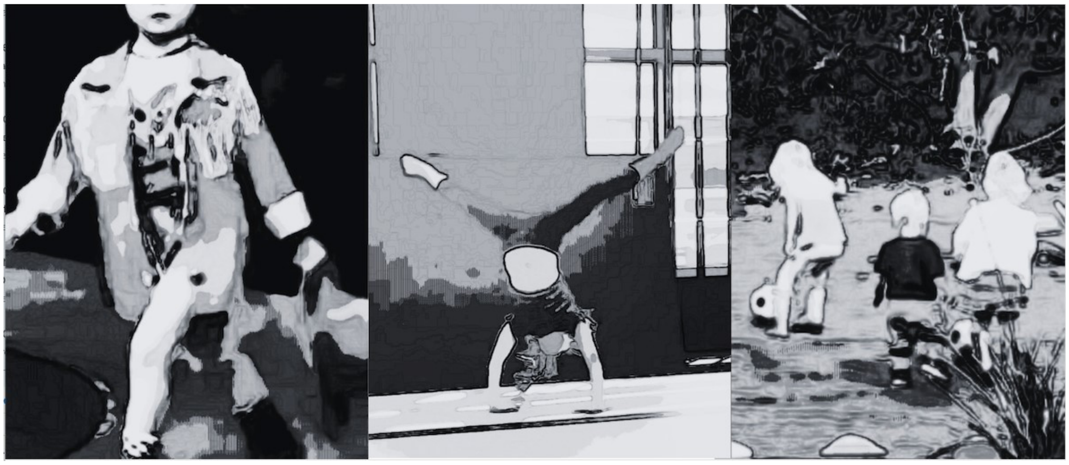
Video analysis

Topic Modeling

To objectively explore and analyze the content of the videos, we applied BERTopic (Grootendorst, 2022) to identify recurrent themes within the dataset. This advanced topic modeling technique leverages BERT embeddings and Term Frequency-Inverse Document Frequency (TF-IDF) to cluster semantically similar comments, providing a more nuanced understanding compared to traditional methods like Latent Dirichlet Allocation (LDA). Each preprocessed comment was transformed into a vector representation using pre-trained BERT embeddings. These vectors are then reduced in dimensionality using UMAP (Uniform Manifold Approximation and Projection) and clustered using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). We specified the number of topics (k =50) to strike a balance between capturing detailed nuances and maintaining broader thematic coherence.

Figure 1

Preview of Child-Related Video Content, With Examples From Each Category (From Left to Right: Fashion, Sports, Family)



Video Annotation

After identifying frequently occurring themes within our dataset, we manually annotated each video (n = 5,896) with the following attributes, referring to the child(ren) in the focus of the video: First, we annotated the child's perceived gender[1] to account for potential differences, categorizing each instance as female, male, or both. When we could not detect a gender based on the video, the gender was categorized as 'unknown'. Second, we analyzed the level of skin exposure in each video, consistent with our research question. We marked a video as showing skin exposure if the child appeared in revealing clothing, such as being naked, in swimwear, or wearing outfits that expose the belly, under, or upper body. Third, we annotated whether a child was shown wearing makeup to account for additional appearance-based factors, as makeup can significantly alter how a child is perceived.

Two researchers conducted the labeling of the 5,896 videos included in our dataset. Both researchers labeled the full dataset individually. We calculated Cohen's Kappa for inter-annotator agreement (Cohen, 1968) for skin exposure, yielding a score of κ = .67, indicating substantial agreement, and κ = .41 for the makeup category, indicating moderate agreement. Cases of ambiguity were discussed within the research team. The moderate agreement in the makeup category was due to common visual filters, making it difficult to distinguish between actual makeup and image enhancements.

[1] We recognize that gender identity is diverse and extends beyond male and female categories. Our use of visual cues for categorization is based on conventional perceptions and is not meant to exclude or invalidate non-binary or other gender identities.

Comment Classification

We then studied user reactions towards videos depicting children (RQ2) by evaluating the video comments. We used a quantitative approach to get a first overview of frequent reactions, analyzing the most frequent words, bigrams, trigrams, and emojis in user comments. Next, we specifically investigated whether children were being targeted by inappropriate comments (i.e., referring to the child as a sex object), contact offers, and whether other users expressed concerns about the child's exposure on the platform. To achieve this, we developed a carefully curated dictionary-based approach, selecting keywords, bi- and trigrams, and appearance-based terms (e.g., *"cute," "beautiful,"* and *"hot"*) that were most likely to indicate these categories.[2] This approach resulted in a set of 100,043 appearance-based comments. Each comment was then manually inspected and labeled with one of the codes: inappropriate, contact, concern, or none.

While contact offers and concerns about a child's exposure were easy to detect, inappropriate comments posed challenges. Many videos featured both the parent and child, making it unclear whether sexually explicit comments (e.g., *"sexy"*) referred to the child or parent, so we did not classify them as inappropriate. The cultural diversity of TikTok users also required a nuanced approach, recognizing that appropriateness varies across cultures. To account for this, we adopted a conservative stance, labeling comments as inappropriate only when they were clearly so in a broad cultural context, focusing on the most overt instances. During the analysis, it became evident that two other categories were exceedingly prevalent within the dataset: comments on the visual appearance of the child (*"She is so pretty," "He has beautiful eyes"*) and comments showing a very strong affection towards the child (e.g., *"I love you," "You*

[2] We make all code available at: https://osf.io/huf76/?view_only=4dbfb7991f3e47b0af2cb07b2cad6c45

are my girl"). We thus created two other dictionaries for extracting these comments and analyzed them using quantitative methods, given the high number of results.

Regarding RQ3, we studied whether and to which extent the content shown on TikTok was distributed a) on private devices and b) on websites other than TikTok. To answer the first part of the question, we inspected the number of times users downloaded a video on TikTok. For the second part, which was mainly motivated by the concern expressed in one comment in our dataset, indicating that the TikTok video was used on a child pornography website, we employed two strategies: First, for each account, we used the Bing reverse image search utilizing a screenshot of the child of one of the videos. Second, we searched for the username of each account using Bing image search. We then evaluated the search results and collected all websites that used a copy of an image or video of the child from TikTok.

Finally, to answer RQ4, we conducted a quantitative analysis to compare videos displaying children wearing exposing clothing and those that do not. We employed statistical tests to examine differences in various metrics, including attachment, appearance-based comments, offers of contact, concerns raised by other users, and engagement metrics, such as the number of likes and downloads. Means and standard deviations were calculated for each variable, and *t*-tests were applied to determine statistical significance between groups. We applied Bonferroni correction to account risk of Type I errors due to multiple comparisons.

Ethical Considerations

While society is concerned about risks to children on social media, these concerns may not align with the interests of account creators, guardians, or TikTok. We, therefore, had to carefully balance minimizing harm with ensuring informed consent. Given that the children's privacy had already been compromised due to widespread viewing, commenting, and downloading of their content, we aimed to avoid bringing additional attention to these children. To adhere to ethical standards, we did not share any identifying details such as account names, pictures, video links, or non-aggregated metadata. Anonymized comments were included only after ensuring they could not be used to identify individuals or accounts through search engines or TikTok. Although we were prepared to report any content classified as child pornography under German law, we did not encounter such material. All data analyzed was from publicly accessible sources, and our study did not involve direct research with human subjects.

Results

In the 5,896 videos analyzed (see Table 2), 19.57% show children in exposed clothing, with girls more frequently appearing in such attire across all categories. Similarly, 3.73% of the total videos feature children wearing makeup, with the Fashion category showing the highest prevalence at 14.52%. The highest share of exposed clothing is in the Sports category, where 46.41% of videos depict children in revealing outfits, likely due to the nature of sports content, particularly gymnastics. In the Family category, which focuses on familial themes, only 0.39% of videos show children wearing makeup, and 14.35% depict skin exposure. In contrast, the Fashion category, centered on style and beauty, has 25.75% of videos showing exposed clothing, reflecting its connection to fashion-related content. Meanwhile, in the Sports category, makeup appears in only 2.05% of videos, indicating its minor role in this context.

Looking at gender differences, we find that videos featuring female children are the most prevalent, comprising 66.79% of the total dataset. Among these, 14.96% depict the children in exposed clothing, and 3.36% show them wearing makeup. In contrast, videos featuring male children account for 18.37% of the dataset, with only 2.07% showing exposed clothing and a minimal 0.15% featuring makeup. Videos with children of both genders constitute 13.52% of the total, with 2.39% showing exposure and 0.22% depicting makeup use. Notably, videos with unknown gender representation are rare, comprising only 1.32% of the dataset, with a very small fraction showing exposure (0.15%) and none depicting makeup.

Comment Analysis

Comments on Visual Appearance

In 88,627 comments (79.0% to Family accounts, 19.32% Fashion, 1.78% Sports), words related to visual appearance were used. Among the most prominent ones were cute, beautiful, adorable, amazing, sweet, pretty, and gorgeous, as well as hair, face, eyes, and dress. Although most of these comments were written in a very positive tone, there were also some negative comments (e.g., *"y does China want me watching ugly middle eastern children"* or *"I hope these two [children] improve with time. Looks are pretty disappointing watching they have really attractive parents."*) When looking at the most frequent words, there is a very big overlap between appearance-based words and overall most frequent words (Figure 3B). This is particularly true when it comes to general terms such as *"cute," "beautiful,"* etc. However, when looking only at appearance-based words, we can see that body parts, such as *"face," "hair,"* and *"eyes,"* belong to the most frequent words in all categories. For the Fashion and Family category, we also observe *"dress"* and *"shoes"* as the most frequent terms relating to what the children are wearing. On the contrary, the 15 most frequent words in the Sports category do not feature any clothing, but mention more body parts, such as *"feet," "neck,"* and *"toe"*. This shift suggests a focus on physical performance and anatomical aspects rather than attire. In the Fashion category, further words such as *"stunning"* and *"slay"* appear but are not present in the other categories, indicating a distinct emphasis on style and presentation not observed in other categories.

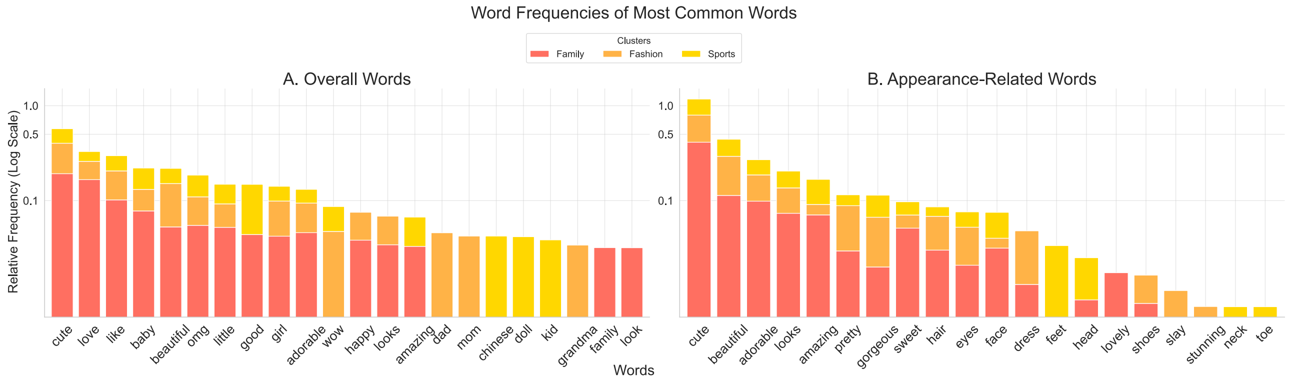
Table 2

Overview of Video Distribution per Label and Category

Label	Gender	Exposed (% within Label and Category)	Makeup (% within Label and Category)
Family	Female	372 (9.13%)	13 (0.32%)
(*n* = 4,073)	Male	97 (2.38%)	-
	Both	106 (2.60%)	3 (0.07%)
Total		584 (14.34%)	16 (0.39%)
Fashion	Female	327 (24.48%)	176 (13.17%)
(*n* = 1,336)	Male	5 (0.37%)	9 (0.67%)
	Both	12 (0.90%)	9 (0.67%)
Total		344 (25.75%)	194 (14.52%)
Sports	Female	183 (37.58%)	9 (1.85%)
(*n* = 487)	Male	20 (4.11%)	-
	Both	23 (4.72%)	1 (0.21%)
Total		226 (46.41%)	10 (2.05%)
Overall	Female	882 (14.96%)	198 (3.36%)
(*n* = 5,896)	Male	122 (2.07%)	9 (0.15%)
	Both	141 (2.39%)	13 (0.22%)
Total		1,154 (19.57%)	220 (3.73%)

Figure 3

Most Frequent Appearance-Based Words per Category



Inappropriate Comments, Attachment & Contact Offers

We found 12 clearly inappropriate comments directed toward children in the dataset. While 12 may seem like a relatively small number, it is still concerning given the nature of these comments. These disturbing comments were all directed at videos from eight unique accounts (four fashion and four family accounts), with six of the videos featuring toddlers and two showing around 6-year-olds. Even though the number might seem low, the presence of any such comments is significant and troubling, especially given the vulnerable age of the children involved. They included, for example, comments such as *"sexy bi**h," "I like the way you suck on your glasses,"* or *"Hot babies"* referring to toddlers performing model poses, *"Save her for me when she's 18+"* referring to a toddler in pajamas, or *"That laugh at nothing makes me want to kiss you with a lot of passion and marry you [...]."* referring to a six-year-old playing dolls with her dad. Five of these comments referred to videos showing children in exposing clothing.

In addition to these comments, a significant number of strangers on TikTok expressed affection for the depicted children. In 4,206 comments, users expressed their love (love you/him/her, love this/your baby/girl/boy), and in 122 comments users referred to the child as my girl/boy/baby (e.g., *"my baby girlfriend," "[...] I love you my baby doll you are so sweet."*, *"[...] dance for me my girls [heart-emojis]"*). In various comments, users asked whether they could adopt the child, with some expressing more than joking intentions (e.g., *"she's just so cute, I love [heart-emoji] her. can I adopt her & I'm serious."*). In addition, multiple comments showed a strong protection motive towards the child, even though there was no need for protection expressed in the respective video (*"I will protect this child with my life. she too precious and seems so sweet"*, *"I'm ready to donate myself to protect that angel [...]"*). While short expressions of love and affection are common on TikTok, the strength of bonding some users seemed to establish towards the child, should not be underestimated. In our dataset alone (which only included comments of at most 100 videos per account), we found 150 users who sent more than 30 comments to a single account, and it was unclear whether these users were strangers or acquaintances to the child. For example, one user with a private account sent 82 comments to a toddler, such as *"Hello little love!!," "Sweet [child's name]!!,"* or *"Hi cutie pie [child's name]!!"*, whereas another one sent 79 comments to a toddler's model account ranging from *"She so beautiful," "Awwww you are a pretty girl,"* and *"I can watch she over n over [...]."*

In an additional 114 comments, users tried to directly contact the account, e.g., by asking for an exchange of direct messages or postal addresses. While many of these offers included collaboration requests, possibly due to the influencer role of the account, it was not possible for us to determine whether real companies/other content creators or private interests stood behind these offers. Also, multiple users asked for an address to send gifts to the child, but it was unclear whether this was due to company interests in content creators, admiration of fans, or darker intent. In addition, there were several comments ranging from nice and joking messages to disturbing ones, depending on the interpretation, e.g., *"I would pay you to babysit her," "mami plz* *can u send me pic in may tik tok acount [...] am in huge luv wz ur baby boyyy [...]"* or *"[...] is it possible for you to send me a box of all her old dresses [....]"*.

Comments Expressing Concerns

In 560 comments, users expressed concerns about the way the child was depicted in the video, including worries about the way in which the child was dressed *"You need to learn how to dress your child because that is so inappropriate"*, the context in which the child-related content was used (e.g., suggestive poses, inappropriate background music, or performance of 18+ related TikTok trends), the number of saves (e.g., *"It's terrifying how many saves this has."*), the form of comments (e.g., *"There are some disgusting comments on this video. Please if this is really your daughter, protect her from grown men who are watching these"*) or the future of the child (*"I can't even imagine how she will feel in 10 years knowing millions of people saw this. It's so sad you choose money over her well being and privacy"*). Some users even criticized TikTok's regulations concerning such content, e.g., *"I really wish TikTok would ban minors from being in videos [...]"* or *"how is this not inappropriate??? [...]"*. Interestingly, there seemed to be a strong common ground in which videos/accounts were regarded as inappropriate by the TikTok community: 206 (36%) of concerned comments were directed to a fashion account of a six-year-old girl and 153 (27%) to a fashion account of a three-year-old girl. Further four fashion and three family accounts received 10-40 comments (together 28%), and 88 accounts in our dataset received no such comment at all.

Exposure and Its Influence on Viewer Interaction

The results from the *t*-tests offer comparisons between videos with exposure and those without, across several key metrics (Table 3). We found that the prevalence of appearance-based comments notably differs between the two groups. Videos with exposure had a higher mean (*M* =

.2360, SD = .42) compared to those without (*M* = .1847, *SD* = .39), with the differences being statistically significant (*t* = -29.63, *p* < .001). This suggests that videos with exposure were more likely to have comments related to appearance. Looking at comments featuring attachment, slight differences were observed between videos with exposure, without being statistically significant. Contact offers also showed no significant difference between videos with and those without. Raised concerns revealed a significant discrepancy, with videos with exposure (*M* = .0037, *SD* = .06) exhibiting more raised concerns in comments than those without exposure (*M* = .0009, *SD* = .03) (*p* < .001).

Table 3

Group Comparisons for Videos with Content of Exposed Children and Without

	Videos with Exposure (*n* = 1,154)	Videos without Exposure (*n* = 4,742)		
	M (SD)	*M (SD)*	*t*	*p*
Attachment	.0094 (.10)	.0101 (.10)	1.71	0.09
Appearance	.2360 (.42)	.1847 (.39)	-29.63	< .001***
Contact Offers	.0002 (.02)	.0003 (0.02)	.47	.64
Expressed Concerns	.0037 (.06)	.0009 (.03)	-17.86	< .001***
N Likes	413,704.20 (931,821.27)	387,034.25 (903,999.81)	-6.66	< .001***
N Downloads	3,203.55 (9,511.63)	5,857.85 (3,9301.72)	16.42	< .001***

Note: The numbers indicate the share of comments featuring attachment, appearance, contact offers, and expressed concerns, alongside absolute numbers for likes and downloads. *** indicating a *p*-value below .001.

The number of likes and downloads showed significant differences between videos with and without exposing content. Videos with exposure received more likes on average (*M* = 413,704.20, *SD* = 931,821.27) compared to those without exposure (*M* = 387,034.25, *SD* = 903,999.81), a difference that was statistically significant (*p* < .001). However, the pattern was different for downloads. Videos with exposure had fewer downloads on average compared to those without exposure (*M* = 5,857.85, *SD* = 39,301.72), with this difference also being statistically significant (*t* = 16.42, *p* < .001). In summary, while exposure appears to increase the likelihood of receiving likes, it inversely correlates with the number of downloads, highlighting the complex dynamics of audience engagement in digital environments.

Secondary Distribution of TikTok Child-Related Content

On TikTok, users can download videos, and the video metadata reveals the number of times a video has been downloaded. First, we examined how frequently users saved videos of children to their own devices, which complicates efforts by platforms and parents to remove such content later. Each video was downloaded on average 137 times, while this number heavily varied per video. While 35% of videos were not downloaded at all, 43% were saved between 1 and 100 times, 13% between 100 and 1,000 times, and 8% more than 1,000 times, with a video of a sneezing baby receiving the maximum number of 1,069,362 downloads. Download numbers were significantly positively correlated to the popularity of the video, measured in the number of views, likes, comments, and shares (all Pearson correlations with p < .001).

In a second step, we investigated whether users would share or repurpose these videos or parts of them on other platforms, potentially without the permission of the parental guardians. For each account, we analyzed the distribution of content on other websites by performing a Google search with the account name as well as a reverse image search on Bing using a screenshot of the child of one of the videos. For 23 accounts, copies of the content were found on other platforms and web pages. Besides social media platforms, like Instagram, these websites included a platform containing duplicates of all TikTok videos with the defined goal to let users watch TikTok videos in an anonymous way; numerous Pinterest collections with pictures of children, some of them tagged with titles such as *"cute babies"* or *"[...] dancing like a stripper"*; and multiple websites on ’social media celebrities’, containing profiles of children on TikTok, including information on the child’s full name, birth date, height, waist size, and medical details.

General Video Content

The child-related videos in our dataset span a wide range of content, from everyday life and sports to family performances and children’s modeling. These videos feature children across various age groups, with a significant focus on pregnancy, childbirth, and early child-rearing. Content often features intimate family moments, including prenatal appointments, childbirth, and newborn care, with creators sharing advice on routines, breastfeeding, and sleep schedules. "Routine videos" are common, showing day-to-day activities with toddlers, from morning prep to bedtime. While meant to connect with other parents, these videos raise privacy concerns by exposing personal details to a large audience, potentially attracting unwanted attention. Figure 3A showcases the top 15 most common words in the three distinct categories Fashion, Family, and Sports, highlighting both overlaps and differences in word usage across these categories. Notably, words such as *"cute," "like," "love," "baby," "omg," "beautiful,"* and *"girl"* appear across all three categories, suggesting universal themes of affection, admiration, and personal interest that transcend specific contexts. Comments on videos in the fashion category are characterized by a blend of aesthetic appreciation (*"beautiful," "adorable"*) and social/family roles (*"dad," "mom"*). The consistent presence of family-related terms alongside fashion-centric vocabulary suggests a notable connection between familial themes and modeling. The family category shows a stronger emphasis on personal and relational expressions (*"love," "baby," "little,"* and *"adorable"*) alongside a higher frequency of words, indicating more intense discussions or more content volume around family topics.

Certain trends raise further concerns, especially those emphasizing children's physical appearance. Some videos feature parents criticizing their child's appearance or making comparisons between themselves and their children. A notable trend involves parents posting a photo with the caption *"When you think I'm pretty..."* followed by an image of their child with *"... you should see my daughter(s)."* Additionally, there are instances where children appear inappropriately, dancing to mature songs or participating in age-inappropriate TikTok challenges.

Topic Modeling

The results from BERTopic aligned with our initial observations about the video content. The topics and their most significant words are detailed in Figure 4. After excluding topics mainly composed of names (e.g., TikTokers or children) and those lacking coherent themes, we focused on 20 meaningful topics.[3] Key categories included descriptions of physical appearance, expressions of cuteness, family dynamics, and lifestyle interests such as food, fashion, and dance.

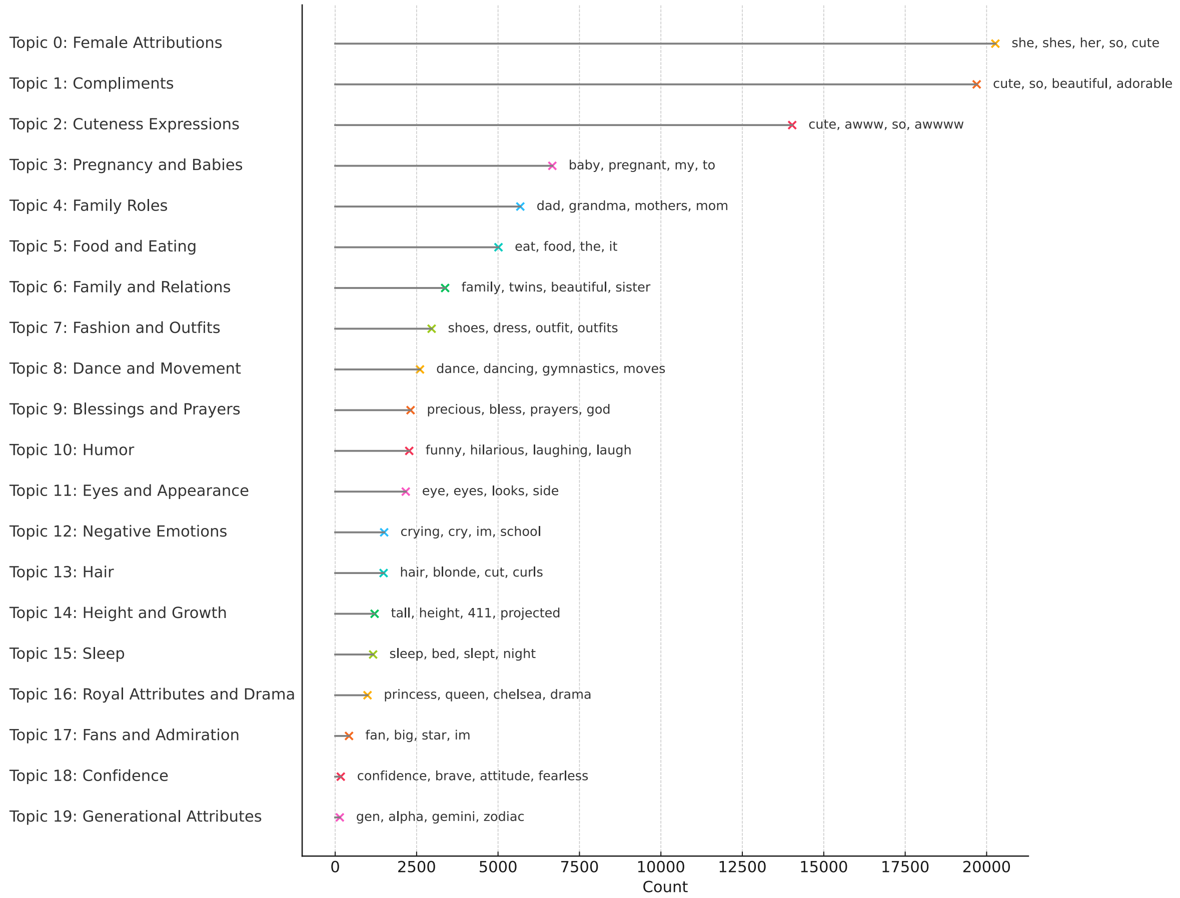
The dataset shows a significant variation in topic prevalence, with "Compliments" leading with 20,269 comments, while topics like "Generational Attributes" have as few as 143

[3] Full topic overview at: https://osf.io/huf76/?view_only=4dbfb7991f3e47b0af2cb07b2cad6c45

comments. This distribution underscores the dominance of themes related to compliments, cuteness, and family, with a strong emphasis on positive and affectionate language, as seen in "Compliments," "Cuteness Expressions," and "Humor," highlighting admiration and endearment throughout the dataset. Family and relational dynamics also feature prominently, as seen in topics like "Family Roles" (Topic 8), "Pregnancy and Babies" (Topic 6), and "Family and Relations" (Topic 17), suggesting a strong focus on familial relationships and life events. Additionally, topics such as "Fashion and Outfits" (Topic 19) and "Hair" (Topic 30) point to an interest in personal appearance and style.

Figure 4

Overview of Topics and Their Most Salient Words (Selected Topics Based on Their Coherence)



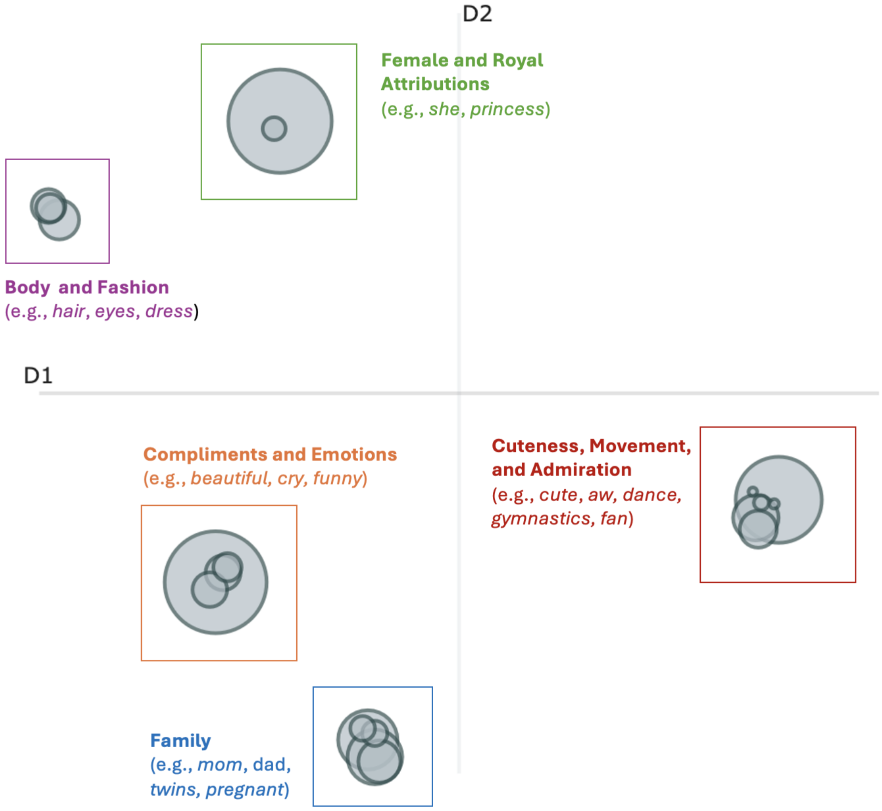
The intertopic distance map in Figure 5 provides a visual overview of the relationships between topics identified by the BERTopic model. The map delineates several thematic clusters within the dataset. The "Body and Fashion" cluster in the upper-left quadrant centers on physical appearance and style, featuring terms like *"hair," "eyes,"* and *"dress."* The "Family" cluster positioned in the lower-middle emphasizes familial roles and relationships, highlighted by words such as "*mom," "dad," "twins,"* and *"pregnant."* Conversely, the "Cuteness, Movement, and Admiration" cluster in the lower-right underscores endearing qualities and activities with terms like *"cute," "dance,"* and *"gymnastics."* The close proximity of circles within each cluster indicates strong thematic connections, while the distinct separation between clusters, such as between "Female and Royal Attributions" and "Body and Fashion," suggests diverse areas of discourse within the dataset.

Discussion

This paper is among the first to comprehensively examine the impact of social media exposure of children on TikTok, focusing on user engagement and interaction. Nearly 20% of the analyzed videos feature children in revealing clothing, a significant proportion. Our analysis shows that users often react strongly to such content, not only through inappropriate comments or contact offers but also by expressing intense attachment. This aligns with previous research on the widespread presence of revealing images of children on social media (Kopecky et al., 2020). We found significant differences in engagement between videos with and without exposing attire—those with exposing content received more appearance-related comments, concerns, and likes but fewer downloads. This pattern suggests that while viewers may like such content in the moment, they may hesitate to download it due to privacy and child protection concerns, reflecting a discrepancy in engagement behavior. Similar findings in Stormer et al., 2023's study on child maltreatment further indicate that certain inappropriate behaviors might drive increased engagement.

Figure 5

Intertopic Distance Map



Note. The x-axis (D1) and y-axis (D2) represent the principal components derived from word embeddings. Each topic is represented as a circle, with the size of the circle indicating the topic's prevalence—larger circles correspond to more dominant topics within the dataset.

Our results reveal that, alongside the high prevalence of sharenting, parents—especially mothers—often face criticism and harassment for sharing videos of their children online. This backlash highlights growing concerns about the risks of exposing children on social media, such as cyberbullying, exploitation, and unwanted attention. The criticism reflects broader societal fears about children's safety and privacy in the digital age. Although not all parents engage in sharenting, those who do often face scrutiny. These findings align with previous research, emphasizing the complex social dynamics and challenges parents encounter with public perceptions of sharenting (Stephenson et al., 2024; Valentino-DeVries & Keller, 2024).

Limitations and Future Research

Video Selection

We primarily utilized children-related keywords in our search, such as ’child’ or ’kid.’ This way, our keyword search methodology was not fully systematic and likely influenced by automated recommendations. While this approach still yielded a diverse range of TikTok accounts, including those centered around family, sports, and fashion themes, the categories we identified may not fully capture the entirety of content featuring children on TikTok. Although we tried to encompass a broad range, our sample is not fully representative due to the platform’s algorithm and our language restriction. In different languages or cultural settings, the nature of comments may vary significantly. This limitation suggests that our findings may not capture the full spectrum of underage users presented on TikTok across different regions and cultures.

Social Acceptance

The social acceptance of children wearing revealing clothing or being depicted in minimal attire, such as diapers, on social media might vary with context and with age. For babies and very young children, posting images where they are naked or in diapers is often seen as more acceptable, reflecting societal norms that view such depictions as innocuous or adorable representations of early childhood. However, as children grow older, societal expectations and concerns about privacy and appropriateness come into play, leading to a decrease in the acceptance of sharing images that expose too much. For example, the share of 45% of exposing baby pictures on Facebook in Brosch (2016)'s study is substantially higher than what we found when looking at a diverse age group. Another example is sportswear: attire that is often short and reveals the midriff might be more accepted for children of various ages due to the specific context of athletic activities. The discrepancy between parents posting videos of children in revealing attire and other users voicing concerns in the comments highlights a clear divergence in perspectives. Our findings, showing that videos featuring children in revealing clothing tend to attract more comments of concern, are consistent with other research indicating a rising awareness regarding privacy issues in children's videos (Walrave et al., 2022).

Educating Parents & Policy Implementations

Educating parents on safely sharing children's content on TikTok requires understanding the platform's safety features, privacy settings, and potential risks. Resources like the TikTok Safety Center, ConnectSafely (ConnectSafely, 2023) and Internet Matters (Matters, 2023) offer essential guidance. TikTok provides tools like restricted mode, private accounts, and Family Pairing features to control who can view and interact with posts. Parents should also consider the long-term risks of sharenting, as content can be difficult to remove and may be misused. Research shows that many parents are unaware of the privacy risks of sharenting, underscoring the need for targeted education (Barnes & Potter, 2020; Williams-Ceci et al., 2021). Parents who have faced negative experiences often adopt mindful practices, such as blurring faces or avoiding identifiable features (Walrave et al., 2023). Future studies could examine how educational interventions influence parental behavior

(Williams-Ceci et al., 2021). Additionally, policymakers could improve child safety through stricter age verification, enhanced reporting mechanisms, and greater transparency in content moderation.

Conclusion

This study critically examines children's exposure on TikTok, analyzing 463,165 comments across 5,896 videos. We found that 19.57% of these videos featured children in revealing clothing, with such content receiving significantly more appearance-related comments. The research highlights the complex dynamics of sharenting, reflecting societal concerns about online risks like exploitation and unintended use of video content. This study contributes to the broader conversation on child safety in digital spaces, emphasizing the urgent need for strategies to protect young users and laying the groundwork for future research and policy development in this area.

Data Availability Statement

The data analyzed in this study will not be publicly shared to prevent any potential misuse and to safeguard the privacy of the individuals involved. Upon careful consideration, the data may be shared upon reasonable request to the corresponding author. Our code is available at https://osf.io/huf76/?view_only=4dbfb7991f3e47b0af2cb07b2cad6c45.

References

Amon, M. J., Kartvelishvili, N., Bertenthal, B. I., Hugenberg, K., & Kapadia, A. (2022). Sharenting and children's privacy in the united states: Parenting style, practices, and perspectives on sharing young children's photos on social media. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 1–30.

Are, C. (2023). Flagging as a silencing tool: Exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. *New Media & Society*, 14614448241228544.

Badillo-Urquiola, K., Smriti, D., McNally, B., Golub, E., Bonsignore, E., & Wisniewski, P. J. (2019). Stranger danger! social media app features co-designed with children to keep them safe online. *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, 394–406.

Barnes, R., & Potter, A. (2020). Sharenting and parents' digital literacy: An agenda for future research. *Communication Research and Practice*, *7*(1), 6–20. https://doi.org/10.1080/22041451.2020.1847819

Barry, B., Wells, G., West, J., Stern, J., & French, J. (2021). How TikTok serves up sex and drug videos to minors. https://www.wsj.com/articles/tiktok-algorithm-sex-drugs-minors-11631052944

Bozzola, E., Spina, G., Agostiniani, R., Barni, S., Russo, R., Scarpato, E., Di Mauro, A., Di Stefano, A. V., Caruso, C., & Corsello, G. (2022). The use of social media in children and adolescents: Scoping review on the potential risks. *International Journal of Environmental Research and Public Health*, *19*(16), 9960.

Brosch, A. (2016). When the child is born into the internet: Sharenting as a growing trend among parents on facebook. *The New Educational Review*, *43*, 225–235.

Campana, M., Van den Bossche, A., & Miller, B. (2020). # Dadtribe: Performing sharenting labour to commercialise involved fatherhood. *Journal of Macromarketing*, *40*(4), 475–491.

Cataldo, I., Lieu, A. A., Carollo, A., Bornstein, M. H., Gabrieli, G., Lee, A., & Esposito, G. (2022). From the cradle to the web: The growth of “sharenting”—a scientometric perspective. *Human Behavior and Emerging Technologies*, *2022*, 1–12.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

ConnectSafely. (2023). Parent’s guide to TikTok.

https://www.connectsafely.org/parents-guide-to-tiktok/

Garmendia, M., Martínez, G., & Garitaonandia, C. (2022). Sharenting, parental mediation and privacy among spanish children. *European Journal of Communication*, *37*(2), 145–160.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

arXiv preprint arXiv:2203.05794.

Howarth, J. (2024). Tiktok user age, gender, & demographics (2024).

https://explodingtopics.com/blog/tiktok-demographics

Internet Matters (2023). TikTok app safety - what parents need to know.

https://www.internetmatters.org/resources/tiktok-app-safety-what-parents-need-to-know/

Kayser-Bril, N., Richard, É., Duportail, J., & Schacht, K. (2020). Undress or fail: Instagram's algorithm strong-arms users into showing skin. *AlgorithmWatch*. https://algorithmwatch.org/en/instagram-algorithm-nudity/

Kernen, L., Adriaensen, B., & Tokarski, K. O. (2021). Social influencer. In J. Schellinger, K. O. Tokarski, & I. Kissling-Näf (Eds.), *Digital business* (pp. 237–251). Springer Gabler, Wiesbaden. https://doi.org/10.1007/978-3-658-32323-3_15

Khan, M. M. I., & Bhattacharjee, H. (2022). A new avenue of crime in bangladesh: Tiktok as a weapon of violence against women. *1st International Conference of Social Sciences on Bangladesh at*, *50*.

Kopecky, K., Szotkowski, R., Aznar-Díaz, I., & Romero-Rodríguez, J.-M. (2020). The phenomenon of sharenting and its risks in the online environment: Experiences from czech republic and spain. *Children and Youth Services Review*, *110*, 104812.

Latipah, E., Kistoro, H. C. A., Hasanah, F. F., & Putranta, H. (2020). Elaborating motive and psychological impact of sharenting in millennial parents. *Universal Journal of Educational Research*, *8*(10), 4807–4817.

Levine, A. S. (2022). These TikTok accounts are hiding child sexual abuse material in plain sight. https://www.forbes.com/sites/alexandralevine/2022/11/11/tiktok-private-csam-childsexual-abuse-material/

Ní Bhroin, N., Dinh, T., Thiel, K., Lampert, C., Staksrud, E., & Ólafsson, K. (2022). The privacy paradox by proxy: Considering predictors of sharenting. *Media and Communication*, *10*(1), 371–383.

Ofcom. (2022). Children and parents: Media use and attitudes report 2022 – interactive data. https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens/childrenand-parents-media-use-and-attitudes-report-2022/interactive

Ouvrein, G., & Verswijvel, K. (2019). Sharenting: Parental adoration or public humiliation? a focus group study on adolescents' experiences with sharenting against the background of their own impression management. *Children and Youth Services Review*, *99*, 319–327. https://doi.org/10.1016/j.childyouth.2019.02.011

Park, H., & Lee, J. (2017). Do private and sexual pictures receive more likes on instagram? *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*, 1–6. https://doi.org/10.1109/ICRIIS.2017.8002525

Pedrouzo, S. B., & Krynski, L. (2023). Hyperconnected: Children and adolescents on social media. the tiktok phenomenon. *Archivos Argentinos de Pediatria*, e202202674.

Polito, V., Valença, G., Sarinho, M. W., Lins, F., & Santos, R. P. d. (2022). On the compliance of platforms with children's privacy and protection requirements: An analysis of tiktok. *International Conference on Software Business*, 85–100.

Pryde, S., & Prichard, I. (2022). Tiktok on the clock but the# fitspo don't stop: The impact of TikTok fitspiration videos on women's body image concerns. *Body Image*, *43*, 244–252.

Ramsey, L. R., & Horan, A. L. (2018). Picture this: Women's self-sexualization in photos on social media. *Personality and Individual Differences*, *133*, 85–90.

Salter, M., & Hanson, E. (2021). "i need you all to understand how pervasive this issue is": User efforts to regulate child sexual offending on social media. In *The emerald international handbook of technology-facilitated violence and abuse* (pp. 729–748). Emerald Publishing Limited.

Silva, M. (2019). Video app TikTok fails to remove online predators. https://www.bbc.com/news/blogs-trending-47813350

Soriano-Ayala, E., Bonillo Díaz, M., & Cala, V. C. (2023). TikTok and child hypersexualization: Analysis of videos and narratives of minors. *American Journal of Sexuality Education*, *18*(2), 210–230.

Statista. (2024). Number of TikTok users worldwide from 2020 to 2025. https://www.statista.com/statistics/1327116/number-of-global-tiktok-users/

Steen, E., Yurechko, K., & Klug, D. (2023). You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media+ Society*, *9*(3), 20563051231194586.

Stephenson, S., Page, C. N., Wei, M., Kapadia, A., & Roesner, F. (2024). Sharenting on tiktok: Exploring parental sharing behaviors and the discourse around children's online privacy. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–17.

Stormer, B., Chandler-Ofuya, N., Baker, A. J., Balin, T., Brassard, M. R., Kagan, J., & Rosenzweig, J. F. (2023). Caregiver psychological maltreatment behaviors toward children on TikTok. *Child Maltreatment*, 10775595231211616.

Suárez-Álvarez, R., García-Jiménez, A., & Urbina Montana, M. L. (2023). Sexualising characteristics of adolescent on TikTok: Comparative study Great Britain–Spain. *Convergence*, *29*(5), 1262–1282. https://doi.org/10.1177/13548565231187728

Taylor, S. H., & Brisini, K. S. C. (2024). Parenting the TikTok algorithm: An algorithm awareness as process approach to online risks and opportunities. *Computers in Human Behavior*, *150*, 107975.

TikTok. (2024). Community guidelines. https://www.tiktok.com/community-guidelines/de-de/

Valentino-DeVries, J., & Keller, M. H. (2024). A marketplace of girl influencers managed by moms and stalked by men. https://www.nytimes.com/2024/02/22/us/instagram-child-influencers.html

Vaterlaus, J. M., & Winter, M. (2021). TikTok: An exploratory study of young adults' uses and gratifications. *The Social Science Journal*, 1–20. https://doi.org/10.1080/03623319.2021.1969882

Verswijvel, K., Walrave, M., Hardies, K., & Heirman, W. (2019). Sharenting, is it a good or a bad thing? understanding how adolescents think and feel about sharenting on social network sites. *Children and Youth Services Review*, *104*, 104401.

Walrave, M., Robbé, S., Staes, L., & Hallam, L. (2023). Mindful sharenting: How millennial parents balance between sharing and protecting. *Frontiers in Psychology*, *14*, 1171611.

Walrave, M., Verswijvel, K., Ouvrein, G., Staes, L., Hallam, L., & Hardies, K. (2022). The limits of sharenting: Exploring parents' and adolescents' sharenting boundaries through the lens of communication privacy management theory. *Frontiers in Education*, *7*, 803393.

Williams-Ceci, S., Grose, G. E., Pinch, A. C., Kizilcec, R. F., & Lewis Jr, N. A. (2021). Combating sharenting: Interventions to alter parents' attitudes toward posting about their children online. *Computers in Human Behavior*, *125*, 106939.

Yegen, C., & Mondal, S. (2021). Sharenting: A new paradigm of digital entertainment of new age parenting and social media. In S. Das & S. Gochhait (Eds.), *Digital Entertainment* (pp. 229–247). Palgrave Macmillan, Singapore. https://doi.org/10.1007/978-981-15-9724-4_11

4.8 Study 8: Large Language Models and Thematic Analysis: Human-AI Synergy in Researching Hate Speech on Social Media

Authors

Petre Breazu, Miriam Schirmer, Songbo Hu, Napoleon Katsos

In

ArXiv, https://doi.org/10.48550/arXiv.2408.05126, Published 08-2024

Abstract

In the dynamic field of artificial intelligence (AI), the development and application of Large Language Models (LLMs) for text analysis are of significant academic interest. Despite the promising capabilities of various LLMs in conducting qualitative analysis, their use in the humanities and social sciences has not been thoroughly examined. This article contributes to the emerging literature on LLMs in qualitative analysis by documenting an experimental study involving GPT-4. The study focuses on performing thematic analysis (TA) using a YouTube dataset derived from an EU-funded project, which was previously analyzed by other researchers. This dataset is about the representation of Roma migrants in Sweden during 2016—a period marked by the aftermath of the 2015 refugee crisis and preceding the Swedish national elections in 2017. Our study seeks to understand the potential of combining human intelligence with AI's scalability and efficiency, examining the advantages and limitations of employing LLMs in qualitative research within the humanities and social sciences. Additionally, we discuss future directions for applying LLMs in these fields.

Contribution of Thesis Author

Data curation, methodological design, formal analysis, visualization, manuscript writing, revision, and editing.

Large Language Models and Thematic Analysis: Human-AI Synergy in Researching Hate Speech on Social Media

Petre Breazu (University of Cambridge), **Miriam Schirmer** (Technical University of Munich), **Songbo Hu** (University of Cambridge), **Napoleon Katsos** (University of Cambridge)

Abstract

In the dynamic field of artificial intelligence (AI), the development and application of Large Language Models (LLMs) for text analysis are of significant academic interest. Despite the promising capabilities of various LLMs in conducting qualitative analysis, their use in the humanities and social sciences has not been thoroughly examined. This article contributes to the emerging literature on LLMs in qualitative analysis by documenting an experimental study involving GPT-4. The study focuses on performing thematic analysis (TA) using a YouTube dataset derived from an EU-funded project, which was previously analyzed by other researchers. This dataset is about the representation of Roma migrants in Sweden during 2016—a period marked by the aftermath of the 2015 refugee crisis and preceding the Swedish national elections in 2017. Our study seeks to understand the potential of combining human intelligence with AI's scalability and efficiency, examining the advantages and limitations of employing LLMs in qualitative research within the humanities and social sciences. Additionally, we discuss future directions for applying LLMs in these fields.

Keywords: *Large Language Models (LLMs)*, *thematic analysis*, *AI in qualitative research*, *human-AI synergy*

Introduction

In this article, we examine the capabilities of GPT-4 (OpenAI, 2024), the state-of-the-art Large Language Model (LLM) that powers ChatGPT, to perform a thematic analysis (TA) of YouTube comments related to the representation of Roma beggars in Sweden. The aim of this experiment is not to endorse LLMs for undertaking research tasks and making automatic decisions in relation to data analysis but rather to explore the advantages and limitations of a potential human-AI synergy to accelerate the analytical process.

TA is a well-established qualitative research method used across humanities and social sciences which is perfectly suited for innovative experiments with LLMs. Braun and Clarke (2006) outlined a clear methodology for TA which consists of six sequential steps: (i) familiarizing oneself with the data, (ii) generating initial codes, (iii) identifying themes, (iv) refining these themes, (v) defining and naming the themes, and (vi) preparing the final report. Traditionally, researchers have conducted both inductive and deductive TA. Inductive TA is a bottom-up, data-driven approach, according to which themes are derived directly from the data, without being influenced by the researcher's preconceptions or theoretical framework (Nowell, Norris et al. 2017). Conversely, deductive TA is a top-down, theory-driven approach, which starts with a predefined set of themes or theoretical framework that the researcher ‘expects’ to find in the data (Kennedy and Thornberg 2018). Both approaches are well-matched with the capabilities of LLMs. Inductive TA enables us to explore how LLMs independently identify themes directly from the dataset. This approach allows us to examine the models' ability to analyze data without predefined categories and input from researchers. Conversely, deductive TA offers an opportunity to observe how LLMs perform analysis within a structured framework in a more controlled setting. In this scenario, the researchers guide the model by introducing theoretical concepts and definitions of specific themes beforehand. This method facilitates a more directed analysis and shows how LLMs apply and adhere to predefined analytical criteria. In this study, we test the applicability of LLMs in supporting both inductive and deductive approaches to TA, with a view to explore the extent to which LLMs enhance the efficiency and comprehensiveness of qualitative research methodologies.

LLMs, as a category of Generative AI, are developed on an unprecedented scale in terms of model size (number of parameters), training data, and computational resources. For example, Meta AI's recent LLaMA-3 model is equipped with 70 billion parameters and was trained using 15 trillion tokens over 6.4 million Graphics Processing Unit (GPU) hours (Meta AI, 2024).

Such extensive training enables LLMs to efficiently perform a wide range of language-related tasks with zero-shot, one-shot, and few-shot learning[1], which require little or no task-specific data (Kaplan et al., 2020). The capability extends to both generative tasks such as text generation, translation, summarization, question answering, and dialog systems, as well as analytical tasks (often termed 'discriminative' in machine learning contexts) including sentiment analysis, named entity recognition, part of speech tagging, and text classification.

There is an emerging body of academic literature which shows how various LLMs have been used in qualitative research or display potential for performing some forms of TA (Dai, Xiong and Ku 2023, De Paoli 2023, Bano, Hoda et al. 2024). While LLMs have demonstrated remarkable proficiency in processing and understanding complex textual information (Bano, Zowghi and Whittle 2023, De Paoli 2023), their ability to effectively analyze and synthesize different types of data remains less explored. Previous studies on the use of LLMs in thematic analysis have employed a diverse array of data sources, including government reports (Khan et al., 2024), survey responses (Dai et al., 2024), semi-structured interviews (De Paoli, 2024), and legal documents, including criminal court opinions (Drápal et al., 2024). This variety of data types highlights the adaptability of LLMs across different fields and research contexts.

Our study contributes to the existing academic debates by focusing on a specific type of data: comments from social media platforms like YouTube, which are frequently laden with hate speech and inflammatory content (Breazu, 2023). This choice of dataset is significant for several reasons. Firstly, it represents a domain that has been less explored in existing research, particularly in the context of LLMs like GPT-4. Secondly, working with data that contains hate speech presents a unique challenge as it may not be processed by LLMs due to content policy restrictions. Such data, by its nature, often violates the guidelines set forth to ensure respectful and safe interactions within digital environments, and there is a risk that LLMs might not recognize the analysis of this data as a valid research task. This constraint is a critical area of concern in employing LLMs for research purposes. Our experimental study seeks to understand how to navigate these barriers responsibly. In this article, we examine GPT-4's

[1] Zero-shot learning refers to the ability of LLMs to complete tasks they have never been explicitly trained on, relying only on their pre-existing knowledge. One-shot learning refers to an LLM's ability to perform a task after receiving a single example or instruction, while few-shot learning adapts an LLM to a new task by being exposed to only a minimal number of specific examples. These methods, in contrast to fine-tuning, highlight the versatility and generalisation abilities of LLMs. They allow LLMs to complete a wide range of tasks without extensive task-specific training datasets, which typically require thousands of examples, to reach optimal performance.

capabilities to perform TA, identify potential contributions and limitations, and possibly provide a new perspective on future human-AI collaboration in academic research.

LLMs and Qualitative Research

The integration of LLMs in qualitative research presents both promising opportunities and challenges (Dai, et al., 2023; De Paoli, 2023). In what follows we highlight recent academic debates which addressed the potential role of LLMs in enhancing qualitative research methodologies. Most academic literature located so far, particularly focused on TA and the complex interplay between human researchers and AI technologies.

Although the deployment of LLMs in qualitative research is in 'its status nascendi' [in the state of being born] (De Paoli, 2023:3), there are ongoing debates about their role and effectiveness in processing and analyzing data. Some researchers (Byun et al 2023; Rietz and Maedche, 2021) suggest that AI can match human capabilities in processing qualitative data and highlight the potential for LLMs to learn human coding practices which suggest that these models can adapt to the subjective nature of the qualitative analysis. Byun et al. (2023) argue that LLMs could rival human capabilities in generating and analyzing qualitative content, especially because of their ability to process large amounts of data and expedite the analysis. Researchers also highlight the potential of LLMs to overcome typical limitations of qualitative research performed by human researchers, especially in relation to processing large datasets, generalizing results to larger contexts, and avoiding subjectivity.

Other studies (Bano et. al 2024; Rudolph et al. 2023) caution against over-reliance on LLMs, pointing out discrepancies between AI and human reasoning that could affect the interpretation of qualitative data. These authors also point to the limitations of LLMs, especially in relation to fully understanding the context of research or the complex nature of human communication. It remains unclear how LLMs compare to human intelligence when performing various qualitative analytical tasks. Bano et al. (2023) and Rudolph et al. (2023) also draw attention to the risks of 'hallucinations' — instances where LLMs generate inaccurate or fabricated information. These inaccuracies, alongside the issue of 'model drift'[2] and the limitations imposed by LLMs' inability to access or interpret the full breadth of relevant literature, present

[2] Model drift refers to the degradation of model performance over time due to changes in the underlying data distribution or the relationship between input features and the target variable, which can result in reduced accuracy and reliability of predictions.

significant difficulties to the validity and replicability of research findings. Furthermore, the evolving nature of copyright and intellectual property concerns (Balel, 2023; Polonsky and Rotman, 2023) needs a thoughtful approach to the integration of LLMs in academic work [here we refer to LLM as a co-author].

It is undeniable that LLMs offer unparalleled advantages in processing large datasets and significantly streamline the analysis process. The ongoing debates surrounding their application in qualitative research point to a delicate balance between embracing technological advances and exercising prudence. While acknowledging that LLMs augment our research capabilities with their speed and scale, AI should complement rather than substitute the critical insight that only human expertise can provide (De Paoli, 2023; Gao et al. 2023).

Data and Context of Research

This article uses data from an EU-funded research project that explores the representation of Roma in Swedish media and political discourse, focusing specifically on Romaphobia as evidenced in comments on YouTube videos about Roma beggars in Sweden. Following an analysis of Roma beggars' portrayal in four leading Swedish newspapers (Breazu, 2024; Breazu and Machin 2024), this data set aims to understand how such discourses resonate or not with the general public on social media. For this experiment, we selected a set of 474 YouTube comments which were thematically categorized by an early career researcher, using NVivo[3].

The socio-political backdrop of this research is crucial for contextual understanding. In Sweden, begging is legally considered a form of free expression and is protected by the Constitution. The 2007 EU enlargement, which saw Romania and Bulgaria's accession, led to many migrants, including ethnic Roma, into Sweden, drawn by the promise of better economic prospects (Breazu, 2024). However, challenges such as limited education and language barriers left some Roma migrants unable to find work or housing, pushing them towards begging or busking. The consequent visibility of Roma begging in public areas sparked debates on public order, safety, and well-being which led to an increase in anti-Roma sentiments (Hansson, 2023; Wigerfelt and Wigerfelt, 2015).

Throughout the years, discussions on potentially banning begging have surfaced repeatedly. The refugee crisis in 2015 notably shifted public and political discourse, intensifying debates

[3] NVivo is a qualitative data analysis (QDA) software that helps researchers organize, analyze, and find insights in unstructured or qualitative data such as interviews, open-ended survey responses, articles, social media, and web content.

around begging bans. By 2016, amidst rising political attention, especially before the 2017 elections, the dilemma of Roma begging and the prospect of instituting localized bans emerged as significant issues in Swedish politics. The analysis of these data sets seeks to offer insights into how these debates are taken up by social media users in their online engagement.

Experiment Design

For our experimental design, we employed a two-fold approach using OpenAI's GPT-4 architecture via the OpenAI API. First, GPT-4 was given various segments of the dataset containing YouTube comments about Roma migrants in Sweden and was tasked to inductively categorize these comments. We assigned ChatGPT-4 the role of a researcher and tasked it to follow Braun and Clarke's (2006) six steps of conducting a thematic analysis on our YouTube dataset about the representation of Eastern European Roma beggars in Sweden. The steps included initial reading of data, coding the data by highlighting key phrases or sentiments, identifying overarching themes based on the coded data, and providing brief descriptions for each identified theme. We fed the dataset in seven separate batches, and we allowed the model to independently analyze the comments without providing pre-defined categories or theoretical framework to ensure an organic emergence of themes based solely on ChatGPT-s's reading of the comments. This thorough approach allowed us to assess the thematic classification capabilities of GPT-4 and gain insights into its alignment with human evaluators and its overall efficacy in qualitative research tasks. These categories identified by GPT-4 were then compared with those found by a human qualitative researcher. Additionally, four more experts in qualitative thematic analysis who were familiar with the dataset assessed the quality of the categories.

Second, we used the identified categories to instruct GPT-4 to deductively assign each comment to one of the previously established categories. We used the OpenAI GPT-4 API with a temperature setting of 0.1 for all API analyses. Throughout this process, we experimented with multiple variations of prompts to optimize our results. We started with a basic prompt employing role-prompting, a fundamental technique in prompt engineering (Chen, Zhang, Langrené, & Zhu, 2023). Assigning the model a specific role, such as an expert, has been proven to be more effective in guiding the model's responses. In our prompts, we assigned the model the role of a 'qualitative researcher investigating the representation of Roma in YouTube comments.' The initial prompt included a basic task description and a short description of the categories. Following best practices in prompt engineering (Chen et al., 2023; Hu et al., 2024;

Liu et al., 2023), we progressively enriched the prompt with additional information and instructions. Mu et al. (2023) demonstrated that augmenting GPT prompts with detailed task and label descriptions significantly boosts its performance. To further refine the prompt, we manually reviewed randomly selected comments and their assigned labels. During this process, experts agreed that some of the comments, such as the one discussed above, did not fit any of the categories found by GPT-4. For example, the comment *“’facts’” that everyone can see are usually the wrong "facts". Facts are and can be verified, not just something one person makes up as he goes.”* This was labeled as “Ethnic Misunderstanding” although there was no clear hint on who the comment was targeted at or what it was about in detail. Recognizing this limitation, we added a "None" option to the original version of the prompt, allowing for responses that didn't neatly fit the existing categories. This step was important for identifying discrepancies between the model's classification and expert evaluations, as well as for detecting patterns in any misclassifications.

To obtain a more generalizable comparison of GPT's categorization quality, we conducted the deductive analysis twice using different sets of categories. In the first instance, as has been described in this section, the categories were created through the ChatGPT interface. In the second instance, they were assembled by a qualitative researcher familiar with the field.

Findings and Discussions

The initial observation is that, beyond its high level of efficiency and ability to process large data sets in seconds, GPT-4 follows steps similar to those of human researchers.

Initial ChatGPT Categorization Scheme

Leveraging its neural architecture and attention mechanism, GPT-4 can attend to every word in the comments, and the initial analysis resulted in the identification of 152 themes. This mirrors the experience of human researchers, where an initial analysis often produces a multitude of themes. Subsequent steps involve instructing GPT-4 to eliminate redundancy, statistically insignificant themes, and overlapping repetitions and generate a refined set of themes. By applying the same process, GPT-4 ultimately distilled the data into five main categories, as illustrated in Table 1.

In our initial round of categorizing YouTube comments, GPT-4 identified the following categories that best describe the dataset after analyzing 474 posts in 7 batches: Ethnic Misunderstanding and Identity Confusion, Stereotyping and Social Prejudice, Economic

Concerns and Welfare Debates, Cultural Clash and Integration Challenges, and Polarization of Public Opinion (Table 1).

Table 1: ChatGPT categories to describe the dataset.

Category	Description
Ethnic Misunderstandings and Identity Confusion:	Emphasizes the distinction between Roma individuals and the ethnic majorities of Romania and Bulgaria, urging accurate ethnic identification and addressing misconceptions about nationality versus ethnicity. Discusses the Roma's historical migration from India, emphasizing their unique cultural evolution in Europe, the complexity of their identity, and the challenges in categorizing Roma strictly based on their ancient origins.
Stereotyping and Social Prejudice:	Stereotypes and prejudices evident in public discourse about begging, crime, theft; and the impact of stereotypes on national reputations. Concerns about crime, exploitation, and the association of Roma with organized criminal activities, including aggressive begging and scams.
Economic Concerns and Welfare Debates:	Explores the economic implications of migration, public perceptions of immigrants and minorities, and critiques of current policies affecting societal integration, welfare systems, and public services.
Cultural Clash and Integration Challenges:	Highlights the differences in cultural norms and legal adherence between Roma communities and the broader populations, touching on the perception of separate legal systems or tribal laws within Roma communities.
Polarization of Public Opinion:	Engages in a broader debate on what constitutes national identity and citizenship in the context of global migration, including discussions on multiculturalism, societal change, and the preservation of cultural identity amidst demographic shifts.

The initial classification performed by GPT-4 on a dataset of YouTube comments about Roma in Sweden yielded categories that aligned well with those identified in an earlier thematic analysis by a qualitative researcher. These categories were reviewed to ensure they adequately captured the scope and addressed every aspect of the comments. To validate the accuracy and relevance of the categories produced by GPT-4, four domain experts independently compared the results. Upon review, the experts agreed that the categories made sense and were comparable to the established thematic framework and that all relevant aspects of the comments were appropriately addressed.

We will now examine how GPT-4's thematic insights into YouTube comments about Eastern European Roma beggars in Sweden compare to the ones by the human researcher.

Both analyses identified the theme of misconceptions surrounding the ethnic identity of the Roma people. GPT-4 categorizes this under ‘Ethnic Misunderstandings and Identity Confusion,’ and frames it around the historical migration of Roma from India and the complexity of Roma identity. In contrast, the human researcher’s category, ‘(Non)Belonging,’ emphasizes a common discriminatory public discourse about Roma as the ‘other’ European, which although live in Europe are not part of the nation (Marin Thornton, 2014; McGarry, 2014) This complex view reflects the human researcher's reliance on academic literature on the discursive representations of Roma and racism to capture the layers of identity politics.

The prevalence of negative stereotypes and prejudices against Roma is another shared theme. GPT-4's category, ‘Stereotyping and Social Prejudice, discusses the impact of these stereotypes on specific national reputations and the association with criminal activities. Meanwhile, the human researcher identifies ‘Perceptions and Stereotypes,’ which focuses specifically on the depiction of Roma as beggars, thieves, and unproductive citizens. The human researcher’s detailed focus underscores the informed understanding of stereotypes’ roots and impacts, influenced by scholarly insights (Rosenhaft and Sierra, 2000, Tremlett, 2022, van Dijk, 2000)

Economic implications and the strain on welfare systems feature prominently in both analyses. GPT-4's ‘Economic Concerns and Welfare Debates’ captures public perceptions of Roma as economic migrants exploiting welfare policies. In contrast, the human researcher categorizes this discourse under ‘Populism’ and ‘Nativism.’ These terms reflect a more precise academic framing of how economic concerns intersect with political narratives about immigration, national identity, and cultural threats, drawing on established theories in political science and sociology (Betz, 2019, Krzyżanowski et. al, 2021; Newth, 2023).

Both analyses highlight the cultural differences and integration challenges faced by Roma communities. GPT-4 uses the category ‘Cultural Clash and Integration Challenges’ to discuss perceived legal and cultural separations. The human researcher’s ‘Cultural Racism’ points to racism expressed through cultural markers of otherness, rather than biological categories. This distinction is informed by academic discussions on modern forms of racism that focus on cultural incompatibility instead of overt biological racism (Bonilla-Silva, 2013; Breazu and Machin, 2024).

Finally, both analyses address polarized public opinions. GPT-4's ‘Polarization of Public Opinion’ discusses broader debates on national identity and multiculturalism, capturing the societal divide. The human researcher, however, identifies ‘Extreme Hate Speech,’ clearly identifying various forms of abusive and dehumanizing language that endorses violence. This specificity is informed by existing academic research on hate speech (Guiora and Park, 2017; Matamoros-Fernández and Farkas, 2021). Additionally, GPT-4's focus on neutral language could also contribute to its broader categorization, as the model avoids to explicitly label the comments as hate speech.

Potential Causes for Differences

The differences between GPT-4 and human researcher analyses can be attributed to several factors. GPT-4's approach tends to categorize themes in a broad, generalized manner, focusing on overarching social and cultural issues. This is likely due to its design as an AI model trained to process and summarize vast amounts of text without the depth of specialized academic training. Consequently, its analysis offers a wide lens on the topics, suitable for capturing a broad spectrum of public discourse.

On the other hand, the human researcher's analysis is deeply informed by existing academic literature, which provides a more detailed understanding of the issues. The use of specific terms like 'Populism,' 'Nativism,' and 'Extreme Hate Speech' reflects a thorough grounding in scholarly work on the representation of Roma migrants. This specificity is essential in identifying the subtle manifestations of racism and discrimination, which may not be as readily apparent in a more generalized analysis.

Moreover, the human researcher's focus on socio-political narratives and the role of media and political elites demonstrates an understanding of the broader context in which these public opinions are formed. This perspective is crucial for comprehending how public discourse is shaped and the implications it has for societal attitudes and policies.

When evaluating the classification, experts agreed that some of the comments did not fit neatly into the categories identified by GPT-4. For example, comments specifically targeting Roma in a derogatory manner were usually labeled as 'Ethnic Misunderstandings and Identity Confusion' or 'Stereotyping and Social Prejudice.' This is one example:

with the lack of Gypsies in Bulgaria, property prices have been on the rise lately. Keep the Gypsies ... send down the sexy Swedish bikini team to Sunny Beach Bulgaria, where blonds are welcome and Gypsies are not"

The human researcher labeled this comment as hate speech due to its dehumanizing language which reinforces harmful stereotypes (e.g. the use of the pejorative *Gypsies*[4]), and explicitly endorses the exclusion of Roma individuals based on racial prejudice (*where blonds are welcome and Gypsies are no*t).

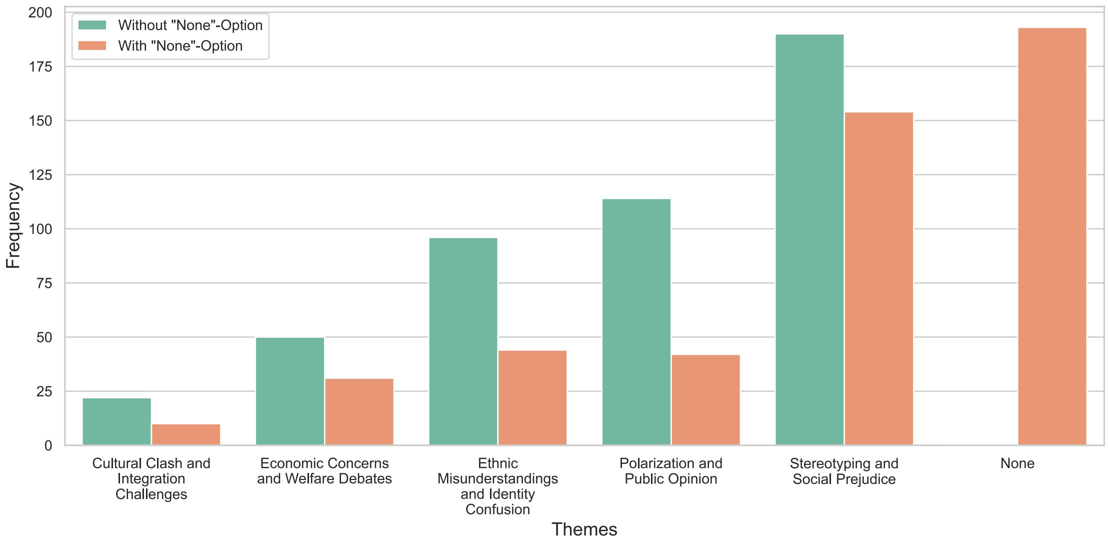
Recognizing this limitation, a ‘None’ option was added to the categorization process to allow for responses that didn't neatly fit the existing categories. The addition of the option not to assign a label to a comment proved to serve multiple purposes: it enhanced the accuracy of categorization by preventing misclassification, revealed potential gaps in the thematic framework by highlighting the proportion of responses that fall outside predefined categories, reduced bias by avoiding forced categorization, and acknowledged the complex nature of immigration discussions that may not be easily captured by broad themes.

Figure 1 illustrates the distribution of themes related to immigration discussions as categorized by GPT-4, comparing scenarios with and without a ‘None’ option. The data reveals that ‘Stereotyping and Social Prejudice’ is the most prevalent theme when specific categories are required, followed by ‘Polarization and Public Opinion’ and ‘Ethnic Misunderstandings and Identity Confusion’ However, when a ‘None’ option is introduced, it becomes the most frequently selected choice (193 samples; 40.72% overall), surpassing all other categories. This shift suggests that many responses do not neatly fit into the predefined themes, highlighting the complexity of immigration discourse. The introduction of the ‘None’ option also leads to a decrease in the frequency of all other themes, indicating that forced categorization may overestimate the prevalence of certain topics.

Notably, ‘Cultural Clash and Integration Challenges’ remains the least common theme in both scenarios. Overall, this visualization illustrates the complex nature of immigration discourse and the importance of flexible categorization in capturing the full range of perspectives.

[4] The term ‘Gypsy’ in the Eastern European context is considered derogatory. It does not have the semantic value to accurately reflect the ethnic identity of the Roma people but rather carries negative connotations such as being unreliable, lazy, dirty, quarrelsome, or deviant. It is recommended to use ‘Roma’ or ‘Romani’ to refer to this ethnic minority respectfully.

Figure 1: GPT-4 Theme Distribution (GPT-Themes)



Human-Supported Categorization

Unlike GPT-4, the human researchers identified more specialized labels, such as belonging, unbelonging, nativism, populism, or cultural racism, due to their knowledge of academic literature in the field. This expertise enables them to associate comments with these specific concepts. This observation suggests that when tasking LLMs with analysis through a specific conceptual framework, it is essential to train the model through in-context learning[5] and provide specific examples of concepts we want to be identified. In the evaluation of the classification results, it was particularly noteworthy that the categories identified by GPT-4 remained neutral and did not introduce or enhance any stereotypes present in the comments. While GPT-4 identified broad themes such as stereotypes or prejudice, it refrained from labeling any content as racist. For example, it would point out references to cultural differences between Roma and non-Roma but maintained a very neutral description and avoided labeling the discourse as (cultural) racism (Table 2). It is noteworthy to mention that even though the model repeatedly reminded researchers that the comments contained inflammatory or discriminatory content, it preserved a high level of neutrality when labeling the content.

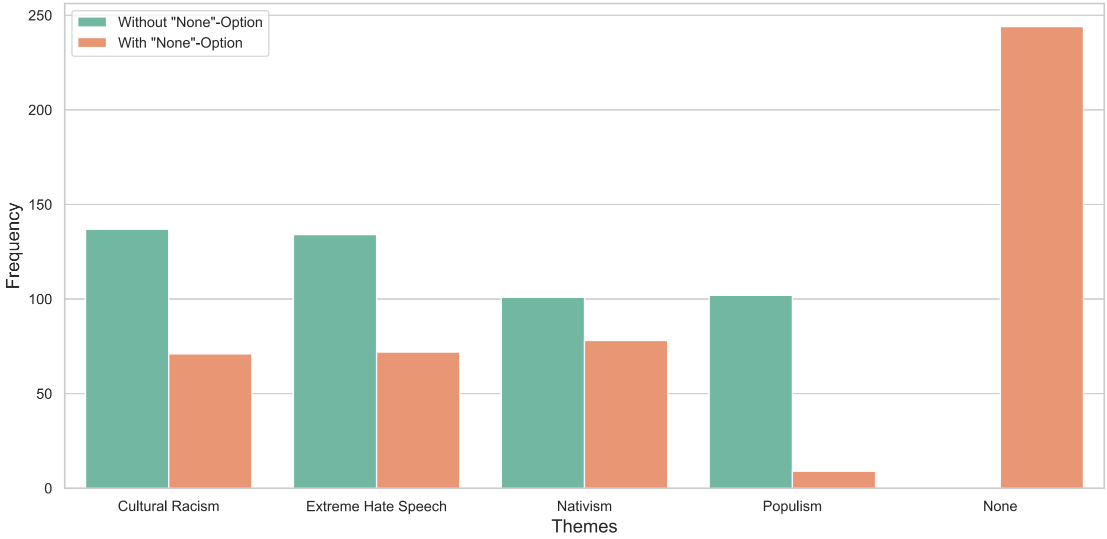
[5] In-context learning is a computational technique used with LLMs where the model learns to perform specific tasks by analyzing examples presented in its immediate input through prompt engineering, rather than through explicit prior fine-tuning on a similar task.

Table 2: Human categories to describe the dataset.

Category	Description
Populism:	Populism is a discourse that is used by elites including mainstream media, politicians, and academics to advance political interests in a manner that reflects the alleged will of the "ordinary people", for example, "fears" of uncontrollable immigration, declining economic prosperity, a decline in moral, cultural and religious values, and a loss of national identity and autonomy.
Nativism:	Nativism is an exclusionary citizenship discourse constructed around adverse narratives about "'us' (the natives) versus 'them' (the non-natives)," with the latter being perceived as dangerous, as social, economic, or cultural threats to the people of the land. It is a mythicized idea about a disenfranchised group of people, the natives of the land, who themselves appear to be forgotten and suffer the consequences wrought by immigration and mainstreaming of multiculturalism such as higher demographics, lower wages, unemployment, increase in crime, decline in safety, cultural changes and altering of immediate surroundings.
Extreme hate speech:	Extreme hate speech encompasses abusive or dehumanizing language invoking well-trodden stereotypes about groups of people, at times endorsing violence, even in playful, humorous ways.
Cultural Racism:	Racism is not expressed about biological categories but alludes to culture as a marker of otherness.
(Non)Belonging	Roma as the 'other' European who should not be confused with Romanians or Bulgarians. References to their Indian origin and physical resemblance.
Perceptions and Stereotypes	Roma as beggars, thieves, scammers and unproductive citizens

As with the previous categories, we instructed GPT-4 to assign one of the human-defined categories to each comment as a first step and only added the 'None' option in the second step. Surprisingly, with this approach, GPT-4 assigned specific labels to less than half of the posts, with 51.48% (244 samples) being classified as not belonging to any of the assigned categories. This highlights the importance of careful selection of categories and prompt design to ensure accurate and meaningful classification.

Figure 2: GPT-4 Theme Distribution (Human Categories)



When evaluating the classification, experts agreed that some of the comments did not fit neatly into the categories identified by GPT-4. For example, comments specifically targeting Roma in a derogatory manner were usually labeled as 'Ethnic Misunderstandings and Identity Confusion' or 'Stereotyping and Social Prejudice.' This is one example:

with the lack of Gypsies in Bulgaria, property prices have been on the rise lately. Keep the Gypsies ... send down the sexy Swedish bikini team to Sunny Beach Bulgaria, where blonds are welcome and Gypsies are not.

While the experts agreed with most of GPT-4's classifications, some discrepancies were noted. For instance, the comment: *They need to be called GYPSIES...any of them is not Romanian or Bulgarian. The largest community of gypsies from the EU is in Hungary*, which was categorized under 'Cultural Racism' by GPT-4. Experts, however, identified this as hate speech due to the use of a derogatory term that lacks semantic value to denote an ethnic group. The term 'GYPSIES,' particularly in the Eastern European context, is laden with negative connotations such as being unreliable, lazy, or dirty (Breazu, 2020). This language is clearly intended to demean and dehumanize the Roma community and can be labeled as hate speech. Additionally, GPT-4's failure to address the capitalization issue further emphasizes the need for context sensitivity, as the capitalized term 'GYPSIES' indicates an aggressive and derogatory emphasis typical of hate speech. The comment also reflects a discourse of unbelonging, suggesting that Roma should not be associated with or referred to as Romanian or Bulgarian.

Another example, *We will holiday in Sweden no longer. It is not Sweden anymore...........*, was classified as ‘Nativism’ by GPT-4. While the borderline between nativist and populist discourses is very thin (Riedel, 2018) and sometimes hard to distinguish, experts contended that this comment is more indicative of populism rather than nativism. Populism often involves rhetoric that appeals to the ‘common people’ against a perceived elite or cultural threat (Stavrakakis, 2017) which fits the sentiment expressed in the comment. The statement reflects a broader sense of discontent and nostalgia for an imagined past (in this case, Sweden as a great place for holidays which has lost its appeal because of the presence of Roma beggars), characteristic of populist discourse. This misalignment in thematic understanding highlights a critical gap; GPT-4's classification missed the broader political and cultural context implied by the comment, focusing instead on a narrower interpretation related to native identity versus foreign influence.

These examples illustrate why the human researchers' expertise provided a more detailed and contextualized classification. Their deeper understanding of the socio-political and cultural background allowed them to correctly identify hate speech and the broader populist sentiment, which GPT-4's broader and more generalized categories failed to capture. This shows the importance of incorporating specific examples and context-driven learning when training AI models to analyze complex and sensitive issues like immigration discourse.

Summary and Conclusion

This study has explored the potential and limitations of using GPT-4, a state-of-the-art Large Language Model (LLM), to perform thematic analysis (TA) on YouTube comments related to Roma beggars in Sweden. Our experimental study highlights the efficiency and scalability of LLMs in processing large datasets and identifying broad themes within qualitative data. However, it also shows the necessity for human oversight to ensure depth, context, and accuracy in qualitative research.

Human-AI Synergy in Qualitative Analysis

The integration of LLMs in qualitative research presents promising opportunities for enhancing research methodologies, but it also introduces significant challenges. As Byun et al. (2023) suggest, AI can match human capabilities in processing qualitative data, but it requires careful guidance to avoid discrepancies between AI and human reasoning. Our study demonstrates that while GPT-4 can generate useful initial categorizations, the depth and specificity provided by human researchers are crucial for accurate and meaningful analysis. The broader, neutral approach of GPT-4 often fails to capture the thorough understanding that human expertise brings to thematic analysis (Bano et al., 2024).

Context Learning

One of the critical insights from this study is the importance of context learning in deploying LLMs for qualitative research. GPT-4's ability to process and categorize data can be significantly improved by training the model with specific contextual information. Providing the model with a detailed background and socio-political context allows it to produce more accurate and relevant classifications. This step is essential for refining the model's capabilities to ensure it can effectively interpret and analyze complex qualitative data (Gao et al., 2023; De Paoli, 2023).

Theory-Driven Prompts

Another key finding is the effectiveness of theory-driven prompts in guiding LLMs. Using predefined theoretical frameworks and detailed task descriptions helps LLMs like GPT-4 align more closely with human categorizations (Mu et al., 2023). This approach leverages the strengths of LLMs in processing large volumes of data while ensuring that the analysis adheres to established academic theories and frameworks (Kennedy & Thornberg, 2018). By incorporating specific examples and theoretical concepts into the prompts, researchers can improve the model's performance and reduce the risk of misclassification (Chen et al., 2023).

Discrepancies and Refinement

The discrepancies observed in the thematic analysis indicate a need for further refinement in GPT-4's understanding of context and subtle language. Teaching the model context and providing it with sufficient background information are crucial steps to improve its accuracy. Currently, we are employing a top-down (deductive) approach, where the model operates with predefined beliefs and normative values, which introduces its own subjectivity. To reduce this subjectivity, it is essential to refine the model's analysis through context-specific training and clearer methodological steps. By instructing the model about the context of the comments and feeding it relevant theoretical frameworks, we can enhance its ability to provide more accurate and meaningful classifications.

The two category schemes reveal distinct approaches to categorizing immigration discussion themes. The first uses broader, neutral terms like 'Cultural Clash and Integration Challenges,' while the second employs more specific, potentially controversial labels such as 'Cultural

Racism' and 'Extreme Hate Speech.' Despite these differences, both graphs show a similar pattern: the introduction of a 'None' option significantly alters response distribution. Without the 'None' option, responses spread across available categories. However, when 'None' is introduced, it becomes the dominant choice, suggesting many responses don't fit neatly into predefined themes. This shift is more pronounced in the second graph, where provocative labels may have pushed more responses toward 'None.' The comparison highlights how category selection and the inclusion of a 'None' option impact data interpretation. The first graph's neutral categories might lead to less charged discussions, while the second's terminology could prompt more contentious debates. In both cases, the high frequency of 'None' responses underscores the complexity of immigration discourse and the limitations of rigid categorization.

Future Directions

Looking ahead, future research should focus on further refining the synergy between human intelligence and LLM capabilities. This involves developing more sophisticated methods for integrating in-context learning and theory-driven prompts into the training and deployment of LLMs. Additionally, it is crucial to address the ethical considerations and limitations associated with using LLMs, particularly in handling sensitive data and ensuring the accuracy and reliability of the analysis (Polonsky & Rotman, 2023; Bano et al., 2024).

Moreover, exploring ways to improve the interpretative abilities of LLMs and incorporating feedback mechanisms where human researchers can interactively refine and guide the model's analysis will be vital. This collaborative approach can use the strengths of both AI and human expertise, which can lead to more thorough and comprehensive qualitative research outcomes (De Paoli, 2023; Rudolph et al., 2023).

While LLMs like GPT-4 hold significant potential for transforming qualitative analysis, their deployment must be carefully managed to ensure they complement rather than replace human expertise. Our findings emphasize this need: although GPT-4 can quickly process large datasets and identify broad themes, it often misses the contextual insights that human expertise provides. For example, GPT-4's broader, neutral approach sometimes led to the misclassification of comments, while human researchers, with their deep understanding of socio-political contexts and academic literature, could identify specific themes more accurately. Furthermore, the AI's tendency to avoid explicitly labeling hate speech highlighted the need of human oversight to interpret and categorize sensitive data correctly. The future of qualitative research could definitively benefit from a synergistic approach that combines the scalability and efficiency of AI with the critical thinking and contextual understanding of human researchers.

References

Balel, Y. (2023). "The Role of Artificial Intelligence in Academic Paper Writing and Its Potential as a Co-Author." European Journal of Therapeutics **29**(4): 984-985.

Bano, M., et al. (2024). "Large language models for qualitative research in software engineering: exploring opportunities and challenges." Automated Software Engineering **31**(1): 8.

Bano, M., D. Zowghi and J. Whittle (2023). "Exploring Qualitative Research Using LLMs." arXiv preprint arXiv:2306.13298.

Betz, H. G. (2019). Facets of nativism: a heuristic exploration. *Patterns of Prejudice*, *53*(2), 111-135.

Bonilla-Silva, E. (2013). “New racism,” color-blind racism, and the future of Whiteness in America. In *White out* (pp. 268-281). Routledge.

Breazu, P. (2022). Decontextualizing a Ban on Begging: A Multimodal Critical Analysis of Media and Political Discourse in Sweden. *Critical Romani Studies*, *5*(2), 28-46.

Breazu, P. (2023). "Entitlement Racism on YouTube: White injury—the licence to Humiliate Roma migrants in the UK." Discourse, Context & Media **55**: 100718.

Breazu, P. and D. Machin (2024). "Humanitarian discourse as racism disclaimer: The representation of Roma in Swedish press." Journal of Language and Politics.

Byun, C., P. Vasicek and K. Seppi (2023). Dispensing with humans in human-computer interaction research. Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv preprint arXiv:2310.14735*.

Dai, S.-C., A. Xiong and L.-W. Ku (2023). "LLM-in-the-loop: Leveraging large language model for thematic analysis." arXiv preprint arXiv:2310.15100.

De Paoli, S. (2023). "Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of the Approach." Social Science Computer Review: 08944393231220483.

Gao, J., et al. (2023). CollabCoder: a GPT-powered workflow for collaborative qualitative analysis. Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing.

Guiora, A., & Park, E. A. (2017). Hate speech on social media. *Philosophia*, *45*, 957-971.

Hansson, E. (2023). *The Begging Question: Sweden's Social Responses to the Roma Destitute*. University of Nebraska Press.

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., & others. (2024). Improving large language models for clinical named entity recognition via prompt engineering. Journal of the American Medical Informatics Association. https://doi.org/ocad259.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Kennedy, B. L. and R. Thornberg (2018). "Deduction, induction, and abduction." The SAGE handbook of qualitative data collection: 49-64.

Krzyżanowski, M., Ekman, M., Nilsson, P. E., Gardell, M., & Christensen, C. (2021). Uncivility, racism, and populism: Discourses and interactive practices in anti-& post-democratic communication. *Nordicom Review*, *42*(s1), 3-15.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys: **55**(9), 1-35.

Marin Thornton, G. (2014). The outsiders: Power differentials between Roma and non-Roma in Europe. *Perspectives on European Politics and Society*, *15*(1), 106-119.

McGarry, A. (2014). Roma as a political identity: Exploring representations of Roma in Europe. *Ethnicities*, *14*(6), 756-774.

Meta. (2024). Introducing LLaMA-3. Meta. Retrieved April 18, 2024, from https://llama.meta.com/llama3/

Mu, Y., Wu, B. P., Thorne, W., Robinson, A., Aletras, N., Scarton, C., Bontcheva, K., & Song, X. (2023). Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. arXiv preprint arXiv:2305.14310.

Newth, G. (2023). Rethinking 'nativism': beyond the ideational approach. *Identities*, *30*(2), 161-180.

Nowell, L. S., et al. (2017). "Thematic analysis: Striving to meet the trustworthiness criteria." International journal of qualitative methods **16**(1): 1609406917733847.

OpenAI (2024). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Polonsky, M. J. and J. D. Rotman (2023). Should artificial intelligent agents be your co-author? Arguments in favour, informed by ChatGPT, SAGE Publications Sage UK: London, England. **31:** 91-96.

Rosenhaft, E., & Sierra, M. (2022). *European Roma: Lives beyond Stereotypes* (p. 352). Liverpool University Press.

Riedel, R. (2018). Nativism versus nationalism and populism–bridging the gap. *Central European Papers*, *6*(2), 18-28.

Rietz, T. and A. Maedche (2021). Cody: An AI-based system to semi-automate coding for qualitative research. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

Rudolph, J., S. Tan and S. Tan (2023). "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" Journal of applied learning and teaching **6**(1): 342-363.

Stavrakakis, Y. (2017). Discourse theory in populism research: Three challenges and a dilemma. *Journal of language and politics*, *16*(4), 523-534.

Tremlett, A. (2017). Visualising everyday ethnicity: moving beyond stereotypes of Roma minorities. *Identities*, *24*(6), 720-740.

Van Dijk, T. A. (2000). New (s) Racism: A Discourse. *Ethnic minorities and the media*, *37*, 33-49.

Wigerfelt, B., & Wigerfelt, A. S. (2015). Anti-gypsyism in Sweden: Roma´ s and Travllers´ Experiences of Bias-motivated Crime. *Internet Journal of Criminology*, (6743), 1-28.

5 Discussion

This dissertation explores the use of Natural Language Processing (NLP) to analyze and understand various forms of violence from multiple perspectives. It makes three central contributions: (1) applying advanced language models to reveal nuanced representations of violence in text, (2) combining computational NLP methods with qualitative analyses, and (3) offering actionable recommendations for societal engagement and real-world interventions. Key themes from this work are summarized below:

Diverse Contexts and Datasets (RQ1). The studies span a wide range of contexts, including genocide tribunals (Studies 1-4) (Schirmer et al., 2022, 2023a), online mental health forums (Study 5) (Schirmer et al., 2024b), Incel communities (Study 6) (Matter et al., 2024), and social media platforms like TikTok and YouTube (Studies 7-8) (Breazu et al., 2024; Schirmer et al., 2024c). Contributions include the creation of the Genocide Transcript Corpus (GTC), which achieved benchmark performance in identifying violence-related witness statements using transformer-based approaches and demonstrated successful transfer learning (Studies 2-3). Additionally, the TRACE dataset (Trauma Event Recognition Across Contextual Environments) was developed, encompassing GTC data as well as posts from a PTSD subreddit and a counseling forum (Study 5). In the Incel community study, analysis of the *incels.is* forum revealed that 21.91% of posts contained violent language (Study 6). The research also involved a dataset of TikTok videos on ”sharenting,” which refers to parents sharing content of their children online (Study 7).

Application of Diverse NLP models (RQ1). All studies applied and fine-tuned advanced NLP models, such as BERT, RoBERTa, and GPT-4, to effectively identify and classify violence-related content, trauma, and hate speech. We further used NLP techniques to support classification, such as active learning to achieve high performance in identifying violence-related witness statements (Study 3), comparing BERT variants, such as BERTbase and HateBERT, who were effective in classifying trauma-related content, even with limited data (Study 3). Extending trauma analysis to online forums, the fine-tuned BERT and RoBERTa models outperformed GPT-4 in predicting traumatic events across diverse datasets (Study 5). By using GPT-enhanced annotation, we found that 21.91% of posts in the *incels.is* community contained violent language. At the same time, the substantial agreement between human and AI annotations demonstrated the potential of AI for large-scale content moderation (Study 6). Using a dictionary-based NLP approach, the TikTok study found that 21% of comments on videos featuring children were appearance-based, with a significantly higher frequency of such comments and likes on videos where children wore revealing clothing. This highlights the risks of children's exposure on the platform and demonstrates how NLP results can be further analyzed through statistical models.

Combining Quantitative and Qualitative Methods (RQ1). These studies demonstrate how integrating quantitative NLP methods with qualitative analyses offers a deeper and more detailed understanding of textual data. In Study 1, while sentiment analysis showed no significant difference between groups, qualitative and NLP analyses revealed distinct differences in emotional involvement and the detail in which torture experiences were described. The TikTok study (Study 7) highlighted risks to children's exposure by combining NLP techniques with qualitative analysis to categorize inappropriate comments, such as contact requests or sexualizing content. Additionally, thematic analysis using GPT-4 on a YouTube dataset (Study 8) demonstrated the effectiveness of combining AI scalability with human expertise to fine-tune categories of hate speech and annotate data.

Common Characteristics & Victim and Aggressor Perspectives (RQ2). Throughout these studies, different perspectives are carefully considered, ranging from victims of mass atrocities to traumatized individuals sharing their experiences in court (Study 1-5) and online (Study 5), children vulnerable to exploitation on social media (Study 7), and individuals who engage in hate speech (Study 6). When analyzing the victim perspective, the study design and results primarily focused on the psychological impact of trauma and how it manifests in language. Despite the context differences, we found common characteristics when talking about trauma that evolves around sexual abuse, death, and language related to the impact of trauma, such as "flashback" (Study 5). In contrast, studies from the aggressor perspective concentrated on identifying the targets of hateful language. For both perspectives, we operationalized concepts of trauma and violence to make them accessible for NLP, categorizing them into specific traumatic events or different hate speech types, such as language targeting a specific group versus general hateful comments.

Practical Applications and Interventions (RQ3). The findings from these studies offer actionable insights for real-world interventions. For example, GENTRAC (Study 4) is an open-access tool designed to aid legal professionals in identifying traumatic content and enhancing trauma-informed legal procedures. Similarly, insights from the TikTok study (Study 7), which found that 21% of comments on videos featuring children were appearance-based, can inform strategies to protect children from inappropriate exposure and interactions online.

5.1 Towards Successful Violence Detection with NLP

Each of the studies presented above addressed a specific problem and made a distinct contribution. The following pages will detail how this dissertation addresses the research questions outlined in Section 1 and the research gaps defined in Section 2.4, while also highlighting the significant scope that remains for future work.

5.1.1 RQ1: Capturing Violence and Trauma with NLP

Researchers can use NLP techniques to detect and analyze various forms of violence and trauma by applying advanced language models and integrating them with social science frameworks. These methods can reveal subtle forms of violent language and trauma that traditional approaches might overlook. In this dissertation, I provide several examples of how this can be achieved.

Reliable Data & Annotation. While creating datasets is not an NLP method per se, it is fundamental to such analysis. Throughout this dissertation, various datasets have been constructed from scratch, including the Genocide Transcript Corpus (GTC) (Studies 2 and 3), the Trauma Dataset (TRACE) (Study 5), the Incel Dataset (Study 6), and a dataset of TikTok comments from videos featuring children (Study 7). When working with court documents, as seen in Studies 2 and 3, the primary challenges often involve dealing with inconsistencies in digitization. Older documents might be scanned images with varying quality, requiring OCR (Optical Character Recognition) for text extraction, which can introduce errors (van Strien et al., 2020). The format of court documents is typically structured and formal, which aids in the extraction process but requires careful handling to maintain the integrity of legal language and nuances. In the studies presented in this dissertation, we addressed these issues by, for example, manually correcting recurring OCR errors and tagging witness names to ensure the sensitive handling of information. In contrast, social media data, such as the TikTok comments (Study 7) and content from online forums (Studies 5 and 6), presents different challenges. Social media data is often unstructured and informal and contains various languages, slang, abbreviations, and emojis. Additionally, this data is subject to platform-specific formatting and character limits, which can affect the analysis (Clark and Araki, 2011). The trustworthiness of social media data can also be an issue due to potential fake or misleading content, requiring robust methods for validation and filtering (Guo et al., 2020; Moturu and Liu, 2011). To address these challenges, we adjusted our approach by incorporating appearance-related emojis into our analysis (Study 7) and ensuring accurate representation of the forum structure in the Incel paper (Study 6). This allowed us to base our main analysis on relevant entries rather than mere references or forwarded comments without substantial content. While this dissertation has concentrated on building datasets from specific social media forums and documents, future research could extend these methods to more diverse data sources, such as large-scale Twitter datasets that capture entire days (Pfeffer et al., 2023).

Moving to reliability in data annotation, recent research highlights the potential of LLMs in annotating violent content (Li et al., 2023). Studies 6 and 8 demonstrate that LLMs can significantly enhance data annotation, particularly in handling large datasets. Human annotations often show high disagreement, suggesting that LLMs may offer a more consistent and objective approach, especially for detecting implicit violence. In Study 6, LLMs proved particularly useful in tasks where human annotators struggled due to high disagreement. Moreover, LLMs can foster productive ”discussions” with researchers in cases of disagreement, especially when the model explains its decisions. Another advantage is cost savings: while acquiring LLMs can be initially expensive, they often prove more cost-effective than ongoing human annotation. Using open-source models can further reduce costs, making advanced annotation techniques more accessible to researchers (Goel et al., 2023).

However, the effectiveness of LLM-based annotations varies by task. Study 8 shows that while LLMs perform well in many contexts, they may lack the nuanced understanding and expert knowledge necessary for complex literary concepts. LLM categorizations can sometimes lack the theoretical grounding that human experts provide, indicating a need to better integrate LLM capabilities with expert insights (Breazu et al., 2024). Accurate annotations provide the foundation for NLP analysis in sensitive areas like violence and hate speech detection, where interpreting complex constructs like trauma and aggression is challenging. Discrepancies in annotators' interpretations, as seen in lower agreement rates, highlight this challenge (Li et al., 2023). To improve annotation quality, future research could explore the interplay between crowdworkers, expert annotators, and LLMs. Integrating these sources could leverage LLM scalability while ensuring quality through expert validation. Fine-tuning LLMs based on feedback from crowdworkers and experts could improve accuracy, balancing scalability with reliability. Incorporating LLM reasoning can also assist in making annotation processes more objective by providing consistent, data-driven insights (as proposed by Matter et al. (2024) and Wang et al. (2024)). This approach could enhance data annotation, especially in complex areas like violence and hate speech detection.

Combining NLP with Qualitative Analysis. By integrating NLP with qualitative methods, as demonstrated in Study 1, researchers can examine trauma within specific contexts, such as genocide, to gain deeper insights into the experiences and narratives of both victims and survivors. Study 1 advocates for a mixed-methods framework where NLP and qualitative analysis complement each other, exemplified by the combination of BERT-based binary classification, sentiment analysis, and qualitative content analysis. Our findings highlight the necessity of these combined methods to fully understand the perspectives of both perpetrators and victims. Researchers have presented frameworks to demonstrate the synergy between NLP and qualitative analysis (Chang et al., 2021), with research applications ranging from Twitter analyses on mass shootings (Criss et al., 2023) to misogyny on online platforms like 4Chan (Phillips et al., 2024).

In Study 8, we extended this approach using large language models (LLMs), specifically GPT-4, to assist with annotations and thematic analysis of YouTube comments. The benefits of this mixed-methods approach became particularly evident in tasks requiring an understanding of the intent behind comments, such as determining whether a YouTube comment was intended to be racist or understanding the personal background of a witness in court documents. These tasks were especially challenging when dealing with short comments. Regarding the inductive creation of categories from the entire dataset of YouTube comments in Study 8, we found that while the categories generated by GPT-4 generally made sense, they lacked the theoretical grounding of those developed by qualitative researchers. Research from other fields has come to similar conclusions, stressing the potential for improvement in human-LLM synergy, particularly in enhancing the theoretical robustness of categories developed through NLP methods (Xiao et al., 2023).

For future directions, multimodal violence detection, integrating text, images, and videos could provide a more comprehensive analysis of violent content (Tan et al., 2018; Xue et al., 2021). For example, Study 7 could be extended to a comprehensive analysis of TikTok videos, including the video content itself, to better capture the nuances of violent behavior depicted visually rather than relying solely on textual data (Peixoto et al., 2020; Wu et al., 2020). The same applies to working with court testimonies, where video recordings are available in some prominent cases, such as those from the ICTY. Integrating video material alongside textual transcripts can provide a richer, more nuanced understanding of the testimonies, capturing not only the spoken words but also the visual cues and emotional expressions that may be critical in interpreting the content accurately.

Text Classification & Exploratory NLP Analyses. The majority of studies in this dissertation applied text classification as an effective NLP technique to detect violent content, both for mental health contexts and hate speech. Text classification proves to be a powerful tool for distinguishing who is discussing violence, as demonstrated in Study 1 (Schirmer et al., 2023b), and for identifying whether a text segment contains references to violence or trauma, as shown in Study 2, providing a transparent, structured approach to handling large datasets. The method can be easily adapted to other topics of interest, making it easy to apply to violence detection in general or more nuanced categories such as traumatic events.

Enhancing text classification to a multiclass framework allows for a more nuanced approach, capturing the complexity of violent experiences. However, effective violence detection in NLP faces challenges due to highly unbalanced datasets. For example, the rarity of traumatic instances complicates multiclass classifications, such as distinguishing different forms of violence, making accurate model training more difficult (Talpur and O'Sullivan, 2020). For other areas of NLP-based violence studies, such as domestic violence, multiclass models have been successfully used to detect violent online posts (Subramani et al., 2019), showing potential for adapting multiclass classifications also for the detection of traumatic events. Strategies to address this include data augmentation, synthetic data generation, and leveraging transfer learning from related tasks to improve model robustness (Endres et al., 2022). However, sometimes, it can be beneficial to work with unbalanced datasets, as demonstrated in Study 5, to better reflect the complexities of real-world scenarios. In specific areas like mental health and violence, obtaining comprehensive datasets is particularly challenging due to the sensitive nature of the content and the scarcity of documented cases (Le Glaz et al., 2021; Montejo-Ráez et al., 2024). These datasets are often small and cannot be easily augmented, making it challenging to train models effectively.

In Study 7, which focuses on analyzing TikTok comments, the primary method used was the application of dictionaries and keyword filtering. This involved creating dictionaries of relevant terms, such as intrusive behavior, appearance-based comments, and child safety. Researchers can filter TikTok comments using predefined dictionaries to isolate relevant content, such as those containing words like ”bullying,” ”abuse,” ”threat,” or ”harassment.” This reduces the dataset to a manageable size for focused analysis. Similar approaches have been used in analyzing electronic mental health records (Van Le et al., 2018), terrorist manifestos (Ebner et al., 2024), and violence against women (Stephanie et al., 2024). These methods also help identify trends or spikes in discussions, which can be linked to real-world events or policy changes. Using dictionaries and keyword filtering in NLP allows efficient analysis of large-scale social media data, providing insights into violence and trauma themes. While advanced models offer precision, simpler methods often suffice for exploratory analysis, delivering meaningful results with less complexity.

Modeling language and culture is challenging when classifying violence, as its expression varies across cultures. For translated texts, for example, methods must be carefully chosen or adapted to ensure that translation issues do not impact the results. In the court context, studies have shown how translated court transcripts can still be used sensitively, ensuring that the overall context is not lost by relying on standardized translations (Fishman, 2006; Gilbert and Heydon, 2021). This aligns with expert opinions, which emphasize the importance of standardizing official translations to maintain the integrity of the original content (see Section 5.1.3). Social media posts present a different challenge. Not only do they involve diverse languages, but they also feature slang, age-specific language, and varying meanings depending on cultural context. For instance, certain phrases or symbols might be interpreted differently across age groups or cultural backgrounds, making it challenging to accurately detect violence or hate speech. Additionally, hate speech on social media is highly culture-sensitive; a term considered offensive in one culture might be harmless or even commonplace in another (Bhattacharya et al., 2020; Paz et al., 2020). Nonetheless, advances are being made in developing culturally aware models capable of understanding diverse linguistic contexts of violent language, involving training on multilingual datasets and incorporating cultural context into the model’s learning process (Aluru et al., 2020; Corazza et al., 2020; Röttger et al., 2022). In contrast, cultural differences play a significant role in dealing with court data, particularly in how sensitive testimonies are handled and how language is used. The way trauma is expressed, the level of detail provided, and how testimonies are delivered can all vary greatly depending on cultural norms and expectations (DeVries, 1996). This cultural sensitivity must be considered when analyzing such data, as it influences both the language used and the overall interpretation of the testimony.

Summary. The framework outlined in this dissertation guides researchers in effectively using NLP for violence detection by emphasizing three key aspects: (1) data preparation, focusing on creating and curating specialized datasets tailored to the specific research context; (2) methodological integration, combining selected NLP techniques with qualitative analysis to capture subtle forms of violence; and (3) human-AI collaboration, leveraging LLMs for annotation while recognizing the essential role of human expertise in providing context and ensuring accurate interpretations. Future work in violence detection with NLP must address challenges in data quality and cultural sensitivity while balancing the scalability of LLMs with expert-driven accuracy. Additionally, integrating multimodal data (text, images, videos) remains a complex yet crucial area for more comprehensive analysis.

5.1.2 RQ2: Common Characteristics and Perspectives on Violence in Text

Contextual analysis is important in understanding the historical, social, and cultural backgrounds that shape the narratives of both victims and perpetrators. Utilizing diverse data sources, such as court documents, personal testimonies, and social media posts, ensures a comprehensive view of the different perspectives. Ensuring balanced representation within datasets is critical to avoiding bias and capturing the full scope of experiences.

Perspectives on Violence. Capturing multiple perspectives in a single study is often challenging. Study 1 serves as a positive example, successfully incorporating both perpetrator and victim perspectives within the same documents, offering a more holistic view of the events. However, this becomes more difficult with data from sources like police reports or social media, where typically only one perspective is represented. For example, in Study 5, we examined victim or survivor descriptions in mental health forums, while in Study 6, we focused on perpetrators engaging in hate speech on Incel forums. Additionally, when filtering abusive content online, victim perspectives are rarely directly available as victims do not typically engage in these forums. To address these gaps, alternative methods such as victim surveys are necessary to bring together the disparate perspectives of victims and perpetrators.

The studies on cross-domain trauma detection (Studies 1, 3, 5) align with the Uses and Gratifications Theory by employing NLP to identify and analyze how individuals express their experiences of violence. These studies reveal that sharing traumatic experiences online can provide catharsis and social validation. For example, Study 5 extends the analysis to online mental health forums, demonstrating how NLP can uncover patterns of support and validation in discussions about trauma. This aligns with theories suggesting online discussions provide a sense of community and belonging, particularly in anonymous forums where users feel safer sharing their experiences (Papacharissi, 2002; Valkenburg et al., 2006). This contrasts with trauma discussed in court (Studies 1-3), where the context is more formal and structured, focusing on legal testimonies and the search for justice rather than community support and personal validation (Ciorciari and Heindel, 2016).

The studies on violent or inappropriate language detection (Studies 6, 7, 8) suggest that group dynamics and anonymization may influence how individuals discuss experienced violence or engage in violent language online, though direct effects have not been clearly demonstrated. Future research could explore these influences in more detail by designing controlled experiments to observe how victims of violence communicate their experiences in different contexts and to examine how exposure to hateful language affects discussions of online hate. Further research on violence could also investigate how constant interaction with social media impacts aggressive behavior or the willingness to share violence-related experiences online, building on existing challenges related to the impact of digital environments on well-being (Montag and Diefenbach, 2018). This might involve analyzing digital traces to identify violence triggers and creating digital environments that help mitigate potential negative effects on mental health and social communication. Such studies would benefit from platforms that replicate real-world social media environments, such as specially designed clones of Facebook (Voggenreiter et al., 2024) or Instagram (Hartl et al., 2024), to facilitate controlled experiments.

Common Characteristics. By examining these types across diverse datasets, the dissertation highlights the importance of creating cross-domain datasets, as exemplified in Study 5, which combined data from different sources (Schirmer et al., 2024b). This cross-domain approach gave a more comprehensive understanding of how violent language manifests across various contexts and platforms, revealing common patterns and unique characteristics. From the victim's perspective, we identified common characteristics in discussions of trauma related to sexual abuse, death, and the language surrounding trauma impacts, such as the terms "impact," "dreams," "flashback," or physical injury with terms such as "wounded," "killed," and "tortured" (Study 5). In contrast, studies from the aggressor's perspective focused on identifying targets of hateful language. Across studies 6-8, racist and misogynistic comments emerged as recurring themes on various social media platforms. However, the intent and nature of these comments varied, influenced by the scope of each study. For instance, Study 7 focused more on inappropriate comments rather than explicitly hateful ones. For both perspectives, we defined and categorized concepts of trauma and violence to make them easier to analyze with NLP. This involved grouping them into specific traumatic events or different types of hate speech, such as language targeting a particular group versus general hateful comments.

Despite these findings, it remains challenging to capture common characteristics of violence while incorporating multiple perspectives in a single study. A comprehensive NLP framework for analyzing violence should include identifying both the sender and the target of violent language, as this aspect is generally detectable across various contexts and data sources. Even in anonymous forums, violent language is typically directed at someone, either in a general or specific manner. By systematically identifying the sender and target, researchers can gain insights into the dynamics of violent interactions and better understand the relationships and power structures involved. This comprehensive approach allows for a more nuanced analysis of violence, considering the content and the participants, as well as their roles in the communication.

5.1.3 RQ3: Impact of Violence Detection

NLP-based research on online violence has demonstrated significant real-world impact across various contexts. For instance, the identification and analysis of violent language on the Incel subreddit contributed to its eventual ban, underscoring NLP's role in moderating harmful online communities (Hauser, 2017). In another case, NLP techniques were instrumental in the sentencing of 24 individuals for hate comments on the Stormfront forum, an online community known for promoting white supremacist ideologies, highlighting the legal implications of detecting online hate speech (ANSA, 2020). Additionally, tools like the Violentometer, developed by Xavier (2023), use NLP to assess and quantify violent content, aiding in prevention and intervention efforts.

These examples demonstrate how NLP can enhance online safety, inform policy decisions, and support legal actions against online violence. To further explore the practical applications and challenges of violence detection, this section examines ways to bridge the gap between academic research and practical implementation based on the studies presented in this dissertation. First, I present insights from interviews with experts in genocide and mass atrocities, highlighting their perspectives on the usefulness of this research. The second part of the section focuses on online tools that play a central role in translating academic violence research into practical solutions.

5.1.3.1 Learning from the Experts: Practical Perspectives on Trauma Detection in the Context of Mass Violence

To gain a comprehensive perspective on the topic of genocide and its implications for my work, I conducted interviews with three subject matter experts from diverse backgrounds and experiences. These interviews were conducted between January and April 2024. To protect their privacy, their identities will remain anonymous.

1. A survivor of the July 1995 Srebrenica massacre during the Bosnian War who is actively engaged in genocide education and prevention efforts.

2. An investigator at *Yazda*, a community-led institution that offers programs to aid and enable survivors of genocide in Iraq and around the world, who conducts interviews with survivors of mass atrocities in Northern Iraq.

3. A United Nations interpreter who participated in the International Criminal Tribunal for the former Yugoslavia.

I provided all interviewees with background information on the Genocide Transcript Corpus, along with a layman's explanation of how I used NLP to capture trauma in court transcripts. I then asked them for their opinions on the usefulness of this approach and any further suggestions they might have. These interviews facilitated an exchange of ideas, highlighting how this research can benefit stakeholders involved in prosecuting mass atrocities.[3]

Genocide Survivor. In the interview with a survivor of the Srebrenica massacre, several key insights were gained regarding the impact of testifying about traumatic experiences and the potential benefits of automated trauma classification systems. The survivor expressed a strong sense of "moral obligation" to testify about their experiences following the loss of family members in the massacre (14). Despite this sense of duty, the survivor described the process of recounting their trauma as "personally very difficult" (26). The initial interviews with prosecutors were extensive, spanning four days and requiring the survivor to provide detailed recollections of their experiences (29). The survivor recounted that the process was profoundly stressful: "It was so stressful for me, you know, and I had to go through all horror again, you know, and all the time when they put me there, [...] all the time I was crying" (59-60). This was the first time they spoke in such detail about their experiences (63-65). The survivor noted that the International Criminal Tribunal for the former Yugoslavia (ICTY) represented a source of hope during this period: "And at such circumstances, the ICTY was for us only hope and only light" (33-36). They reported feeling strengthened after their testimony at the ICTY: "After that, after the ICTY, after my testimony, I felt stronger" (242).

The survivor emphasized the importance of sharing their story: "It is most important for us to be heard our story, you know" (87). However, there were concerns about perceived inequities in the treatment of victims versus perpetrators, as the survivor felt that more attention and care were given to the rights of criminals: "I felt that they gave more care and more attention to the criminals and they wanted to protect their rights more than us victims" (148). This perception underscores the need for mechanisms to ensure equitable treatment of victims' voices.

While the expert specifically criticizes the gap between research and the general public (102), they acknowledge that automated trauma classification systems could offer significant benefits in this context by alleviating the need for survivors to repeatedly recount their traumatic experiences. Such systems could help classify and analyze testimonies, potentially reducing the emotional burden placed on survivors while ensuring their narratives are documented and accessible. Furthermore, the survivor acknowledged the value of their testimony for legal education and training purposes despite the challenges in reaching the public (212). This suggests that automated systems could also support the broader dissemination of testimonies for educational and training contexts, thereby enhancing the effectiveness of legal processes.

Legal Activist. The second interviewee is an investigator who has been working for Yazda in Northern Iraq since 2016. They conduct interviews with survivors of mass atrocities, including those of ISIS attacks, in Kurmanji, a Northern Kurdish language, within refugee camps. Their

[3] The interviews were transcribed using Whisper, a speech recognition system developed by OpenAI (Radford et al., 2022). Numbers in the transcripts indicate line numbers. To protect the experts'privacy, full transcripts are not published but are available upon request from the author.

organization has collected over 2,000 interviews and documents mass graves, focusing extensively on the documentation of genocidal acts and their aftermaths. With five to six years of experience, their work involves interviewing survivors of sexual violence, children, and other genocide-affected individuals. This extensive documentation aims to prevent future genocides and ensure that survivors' experiences are recorded accurately.

They emphasize the importance of documentation, stating, "Documentation is one of the, like, most important projects now, like, in Yazda, because we try to document the genocide to, like, not happen in the future" (73). According to them, the thorough documentation process is important because, historically, many atrocities went undocumented, leaving survivors without evidence to support their claims. The expert notes the shift in approach: "So many times, no one has evidence, no one has documented, but this time, like, we stand and we try to document what happened in order to get a bit of our rights and survivors' rights" (74). This documentation could potentially support future trials or lawsuits to achieve justice for the survivors (75). One of the main goals of her organization is to hold perpetrators accountable (81).

However, the current process is manual and labor-intensive, involving coding everything by hand, which has made creating a searchable database challenging. The expert points out the need for automation in this area, stating it would be helpful if the search could be automated as they have been working on a searchable database for a very long time (93, 104, 136). Interviews typically last between 12 to 50 hours (139), and the interviewees are often heavily traumatized and prepared by psychologists (196).

Confidentiality and trust are critical in their work. They stress the importance of gaining consent from survivors before sharing their information, stating, "This information is really confidential; they cannot share with everybody because they have trust and know how we are dealing with their testimony" (247-249). The expert believes that automated trauma detection could be highly beneficial in preparing witnesses for trials by helping them understand how questions might be asked, thus lessening their trauma: "Automated trauma detection is helpful to prepare witnesses for trials, so they see how questions are asked" (252-272). According to her, this technology could significantly enhance the support provided to survivors, ensuring their experiences are documented and used effectively in seeking justice.

UN Interpreter. Using court transcripts translated into English from various other languages is often subject to critique within NLP research. However, the appropriateness and validity of such translations are contingent upon the specific research objectives. According to the interpreter, international criminal courts and the United Nations implement stringent measures to ensure translation standardization and accuracy. For instance, annual reports provide detailed assessments of transcript accuracy (44). Additionally, native speakers are typically present within each legal team, whether part of the defense, prosecution, or judiciary, to identify and address potential inaccuracies or ambiguities in the translations (101). Moreover, prosecution teams engage with witnesses before testimony to familiarize themselves with the content and request reinterpretation if necessary to ensure clarity (105).

In the context of the International Criminal Tribunal for the former Yugoslavia (ICTY) and the International Criminal Tribunal for Rwanda (ICTR), only a limited number of transcripts are produced in the original languages of the respective countries. Predominantly, transcripts are generated in English (18-31). While deliberate omissions and retelling are considered valid interpreting methods, they can also clarify the intended message (46-55). Certain non-verbal expressions, such as sobbing or variations in tone, are not typically translated (68). The transcripts are designed to capture all audible elements of the testimony (99).

The interpreter emphasized the importance of having English transcripts to uphold international accountability. This standardization facilitates broader access and scrutiny, ensuring that justice processes are transparent and understandable to the global community (190-196).

Summary of Expert Interviews. These expert insights underscore the multifaceted benefits of automated trauma detection systems in the context of mass violence and genocide. Fine-tuned and pre-trained language models, such as the one for trauma detection, can significantly enhance legal and educational processes by reducing the emotional burden on survivors, ensuring the equitable representation of their voices, and facilitating the efficient documentation and analysis of testimonies. The practical perspectives subject matter experts provide highlight the critical need for continued research and development in this area to make these technologies accessible and impactful beyond academic circles. Concretely, collaborating with organizations like Yazda, which conducts and processes witness interviews on the ground, would be invaluable in adapting the tools developed in this dissertation for real-world applications. By working closely with such organizations, these tools can be tailored to specific use cases, making them more accessible and practical for human rights activists.

5.1.3.2 Collaborations and Tools for Enhanced NLP Violence Detection

Collaboration between social scientists, NLP researchers, and non-governmental organizations (NGOs) is vital for advancing violence detection. In the context of this dissertation's focus on survivors of mass atrocities, I collaborated with the NGO *Auschwitz Institute for the Prevention of Mass Atrocities*. Building on my research, I developed comprehensive guidelines for interviewing survivors and interpreting research findings in this field. The Auschwitz Institute now utilizes these guidelines in their internal education seminars, designed for government officials shaping policy in this area and human rights activists conducting witness interviews on the ground.

Similarly, in the context of hate speech, other partnerships with organizations like HateAid, an NGO that focuses on combating online hate and misogyny, can provide real-world data and scenarios for testing and refining NLP tools. In addition, case studies of particularly viral incidents, such as misogyny in media-heavy court cases (Strathern and Pfeffer, 2023), can highlight the strengths and limitations of current methodologies.

Developing specialized tools can further foster interdisciplinary engagement. For instance, GENTRAC (Study 4) is an openly accessible resource for researchers and professionals, facilitating collaborative efforts (Schirmer et al., 2024a). Freely accessible online tools for violence detection, such as web-browser applications that detect harmful language (Modha et al., 2020), can support researchers in detecting and analyzing violent content. To make these tools accessible to non-NLP researchers, they should be designed to handle the complexities of different data sources while providing robust, scalable solutions. Ensuring that these tools are user-friendly, adaptable to various research needs, and require minimal technical expertise is essential for effectively aiding in the comprehensive analysis of violent content. Content moderation is important in this context as it can help manage harmful behavior and enhance online safety (Waltenberger et al., 2023; Wilson and Land, 2020). One direction of content moderation includes AI-supported mechanisms, which have been shown to successfully mitigate online hate against the LGBTQ+ community (Thiago et al., 2021).

Another future direction of computational violence research is to explore how social media can contribute to real-life violence through the spread of misinformation. Social media platforms can amplify false or misleading information, creating misunderstandings and increasing community tensions. Misinformation can spread quickly during social or political unrest, leading to heightened fear and division (Vosoughi et al., 2018). Additionally, the algorithms that drive social media engagement often prioritize sensational content, inadvertently boosting the visibility of inflammatory or misleading posts (Allcott and Gentzkow, 2017). Future research should focus on developing advanced NLP models to detect and mitigate misinformation in real-time. Understanding the social and psychological mechanisms that turn online misinformation into offline actions is also essential. By fostering interdisciplinary collaborations among technologists, social scientists, and policymakers, we can create effective strategies to manage misinformation and enhance online safety. This approach can help prevent digital misinformation from escalating into real-world violence, contributing to more resilient and informed communities (Pennycook and Rand, 2018).

5.2 Conclusion

This dissertation has highlighted the advantages of using NLP techniques to study violence across different contexts. Based on multiple studies, I have demonstrated that computational methods offer significant opportunities for analyzing social science research problems related to violence. These methods enable access to new datasets and allow researchers to analyze large-scale data efficiently, contributing to a deeper understanding of violent behavior with reduced costs, time, and intrusiveness.

Throughout this dissertation, I have evaluated the potential of interdisciplinary and mixed-methods approaches for NLP research on violence and have demonstrated the importance of creating cross-domain datasets by combining data from different sources and drawing on concepts from social science, psychology, and computer science, enhancing the comprehensiveness of the analysis.

Nevertheless, as shown by this dissertation and exemplified within the various case studies, applying NLP methods in violence research also holds many challenges. Successful violence detection with NLP requires a combination of innovative research methodologies, ethical data collection practices, cultural and linguistic sensitivity, and strong interdisciplinary collaboration. By addressing these challenges and exploring new directions, researchers can enhance the effectiveness and impact of NLP in understanding and mitigating violence in various contexts, ensuring studies are robust, ethical, and impactful.

References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Althoff, T., Clark, K., and Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders*. Author, 5th ed. edition.

ANSA (2020). Rome sentences 24 for hate comments on 'stormfront' forum. `https://www.infomigrants.net/en/post/22714/rome-sentences-24-for-hate-comments-on-stormfront-forum`. Published on 2020/02/12.

Aroustamian, C. (2020). Time's up: Recognising sexual violence as a public policy issue: A qualitative content analysis of sexual violence cases and the media. *Aggression and Violent Behavior*, 50:101341.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Auti, N., Ghadge, S., Jadhav, R., Jagtap, P., and Ranaware, S. (2022). Social media based hate speech detection using machine learning. *International Journal for Research in Applied Science and Engineering Technology*, 8(6):443–450.

Bache, R. (2011). Measuring and improving access to the corpus. In Lupu, M., Mayer, K., Tait, J., and Trippe, A. J., editors, *Current Challenges in Patent Information Retrieval*, pages 147–165. Springer, Berlin, Heidelberg.

Balahur, A., Steinberger, R., et al. (2009). Rethinking sentiment analysis in the news: from theory to practice and back. *Proceeding of WOMSA*, 9:1–12.

Basile, K. C., Black, M. C., Breiding, M. J., Chen, J., Merrick, M. T., Smith, S. G., Stevens, M. R., and Walters, M. L. (2011). National intimate partner and sexual violence survey: 2010 summary report.

Batrinca, B. and Treleaven, P. C. (2015). Social media analytics: A survey of techniques, tools and platforms. *AI & Society*, 30:89–116.

Belinkov, Y., Gehrmann, S., and Pavlick, E. (2020). Interpretability and analysis in neural nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5.

Bermingham, A., Conway, M., McInerney, L., O'Hare, N., and Smeaton, A. F. (2009). Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *2009 International Conference on Advances in Social Network Analysis and Mining*, pages 231–236. IEEE.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., and Ojha, A. K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*.

Bilewicz, M. and Soral, W. (2020). Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Bogen, K. W., Orchowski, L. M., and Ullman, S. E. (2024). Online disclosure of sexual victimization and social reactions: What do we know? In *Resistance & Recovery in the# MeToo era, Volume I*, pages 116–131. Routledge.

Botelle, R., Bhavsar, V., Kadra-Scalzo, G., Mascio, A., Williams, M. V., Roberts, A., and Stewart, R. (2022). Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: An applied evaluative study. *BMJ Open*, 12(2):e052911.

Breazu, P., Schirmer, M., Hu, S., and Kastos, N. (2024). Large language models and thematic analysis: Human-ai synergy in researching hate speech on social media. *arXiv preprint arXiv:2408.05126*.

Breslau, N. and Kessler, R. C. (2001). The stressor criterion in dsm-iv posttraumatic stress disorder: An empirical investigation. *Biological Psychiatry*, 50(9):699–704.

Brounéus, K. (2008). Truth-telling as talking cure? Insecurity and retraumatization in the Rwandan Gacaca courts. *Security Dialogue*, 39(1):55–76.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bullers, K., Howard, A. M., Hanson, A., Kearns, W. D., Orriola, J. J., Polo, R. L., and Sakmar, K. A. (2018). It takes longer than you think: Librarian time spent on systematic review tasks. *Journal of the Medical Library Association: JMLA*, 106(2):198–207.

Burley, T., Humble, L., Sleeper, C., Sticha, A., Chesler, A., Regan, P., Verdeja, E., and Brenner, P. (2020). Nlp workflows for computational social science: Understanding triggers of state-led mass killings. In *Practice and Experience in Advanced Research Computing*, pages 152–159. ACM.

Burnap, P. and Williams, M. L. (2016). Us and them: Identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5:1–15.

Calvo, R. A., Milne, D. N., Hussain, M. S., and Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Cambria, E. and White, B. (2014). Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

Campbell, J. C. (2002). Health consequences of intimate partner violence. *The Lancet*, 359(9314):1331–1336.

Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2020). HateBERT: Retraining BERT for abusive language detection in English. *arXiv preprint arXiv:2010.12472*.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chang, T., DeJonckheere, M., Vydiswaran, V. G. V., Li, J., Buis, L. R., and Guetterman, T. C. (2021). Accelerating mixed methods research with natural language processing of big text data. *Journal of Mixed Methods Research*, 15(3):398–412.

Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv preprint arXiv:2310.14735*.

Chen, G. M. (2017). *Online Incivility and Public Debate: Nasty Talk*. Palgrave Macmillan.

Chung, N. C., Dyer, G. C., and Brocki, L. (2023). Challenges of large language models for mental health counseling. *ArXiv*.

Ciorciari, J. D. and Heindel, A. (2016). *Trauma in the courtroom. In B. van Schaack & D. Reicherter (Eds.), Cambodia's hidden scars: Trauma psychology and the Extraordinary Chambers in the Courts of Cambodia*. Documentation Center of Cambodia, 2nd ed. edition.

Clark, E. and Araki, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual english. *Procedia-Social and Behavioral Sciences*, 27:2–11.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213.

Coppersmith, G., Harman, C., and Dredze, M. (2014). Measuring post traumatic stress disorder in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 579–582.

Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Costello, E., Brunton, J., Bolger, R., Soverino, T., and Juillerac, C. (2023). Massive omission of consent (mooc): Ethical research in educational big data studies. *Online Learning*.

Craker, N. and March, E. (2016). The dark side of facebook: The dark tetrad, negative social potency, and trolling behaviors. *Personality and Individual Differences*, 102:79–84.

Criss, S., Nguyen, T. T., Michaels, E. K., Gee, G. C., Kiang, M. V., Nguyen, Q. C., Norton, S., Titherington, E., Nguyen, L., Yardi, I., et al. (2023). Solidarity and strife after the atlanta spa shootings: A mixed methods study characterizing twitter discussions by qualitative analysis and machine learning. *Frontiers in Public Health*, 11:952069.

Cuevas, C. A. and Rennison, C. M., editors (2016). *The Wiley Handbook on the Psychology of Violence*. Wiley.

Cuklanz, L. M., editor (2022). *Gender Violence, Social Media, and Online Environments: When the Virtual Becomes Real*. Taylor & Francis.

Dalal, E. and Singh, P. (2024). Textrefine: A novel approach to improve the accuracy of llm models. *Data and Metadata*, 3:331–331.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.

de Gibert, O., Perez, N., Garcia-Pablos, A., and Cuadros, M. (2018a). Detection of hate speech in online media: Application of nlp tools and techniques. *Proceedings of the Sixth International Conference on Learning Representations (ICLR)*.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018b). Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

De Jong, J. T. V. M., editor (2002). *Trauma, War, and Violence: Public Mental Health in Socio-Cultural Context*. Springer Science & Business Media.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

DeVries, M. W. (1996). *Trauma in cultural perspective.* Guilford Press.

Diener, E. (1980). Deindividuation: The absence of self-awareness and self-regulation in group members. In Paulus, P. B., editor, *Psychology of Group Influence*, pages 209–242. Lawrence Erlbaum Associates, Hillsdale, NJ.

Dumont, C., Meisinger, S., Whitacre, M. J., and Corbin, G. (2012). Nursing2012 horizontal violence survey report. *Nursing2023*, 42(1):44–49.

Ebner, J., Kavanagh, C., and Whitehouse, H. (2023). Assessing violence risk among far-right extremists: A new role for natural language processing. *Terrorism and Political Violence*, pages 1–18.

Ebner, J., Kavanagh, C., and Whitehouse, H. (2024). Measuring socio-psychological drivers of extreme violence in online terrorist manifestos: An alternative linguistic risk assessment model. *Journal of Policing, Intelligence and Counter Terrorism*, 19(2):125–143.

El Barachi, M., Mathew, S. S., Oroumchian, F., Ajala, I., Lutfi, S., and Yasin, R. (2022). Leveraging natural language processing to analyse the temporal behavior of extremists on social media. *Journal of Communications Software and Systems*, 18(2):195–207.

Endres, M., Mannarapotta Venugopal, A., and Tran, T. S. (2022). Synthetic data generation: A comparative study. In *Proceedings of the 26th International Database Engineered Applications Symposium*, pages 94–102.

Fernandes, A. C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., and Chandran, D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports*, 8(1):7426.

Fishman, C. S. (2006). Recordings, transcripts, and translations as evidence. *Washington Law Review*, 81:473.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Florio, K., Basile, V., Polignano, M., Basile, P., and Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Friedman, M. J. and Davidson, J. R. (2007). PTSD. *Handbook of PTSD: Science and Practice.*

Geva, M., Goldberg, Y., and Berant, J. (2019). Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898.*

Gilbert, D. and Heydon, G. (2021). Translated transcripts from covert recordings used for evidence in court: Issues of reliability. *Frontiers in Communication*, 6:779227.

Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I., and Feder, A. (2023). Llms accelerate annotation for medical information extraction. *ArXiv.*

Gold, S. N. (2017). *APA handbook of trauma psychology: Foundations in knowledge, Vol. 1.* American Psychological Association.

Golder, S., Ahmed, S., Norman, G., and Booth, A. (2017). Attitudes toward the ethics of research using social media: A systematic review. *Journal of Medical Internet Research*, 19.

Gongane, V. U., Munot, M. V., and Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining*, 12(1):129.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794.*

Grych, J. and Hamby, S. (2014). Advancing the measurement of violence: Challenges and opportunities. *Psychology of Violence*, 4(4):363.

Guiliano, J. and Ridge, M. (2016). The future of digital methods for complex datasets: an introduction. *International Journal of Humanities and Arts Computing*, 10(1):1–7.

Guimaraes, S., Kakizaki, G., Melo, P., Silva, M., Murai, F., Reis, J. C., and Benevenuto, F. (2023). Anatomy of hate speech datasets: Composition analysis and cross-dataset classification. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, pages 1–11.

Guo, B., Ding, Y., Yao, L., Liang, Y., and Yu, Z. (2020). The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36.

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., and Li, K. (2024). Large language model for mental health: A systematic review. *arXiv preprint arXiv:2403.15401.*

Hachey, B. and Grover, C. (2006). Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14:305–345.

Hamby, S. (2017). On defining violence, and why it matters. *Psychology of Violence*, 7(2):167–180.

Hartl, A., Starke, E., Voggenreiter, A., Holzberger, D., Michaeli, T., and Pfeffer, J. (2024). Empowering digital natives: Instaclone-a novel approach to data literacy education in the age of social media. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 484–490.

Hauser, C. (2017). Reddit bans 'Incel' group for inciting violence against women. *The New York Times*. `https://www.nytimes.com/2017/11/09/technology/incels-reddit-banned.html`.

He, Q., Veldkamp, B. P., Glas, C. A., and de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment*, 24(2):157–172.

Hermida, P. C. D. Q. and Santos, E. M. D. (2023). Detecting hate speech in memes: A review. *Artificial Intelligence Review*, 56(11):12833–12851.

Hoang, L. and Schneider, J. (2018). Opportunities for computer support for systematic reviewing - a gap analysis. In Chowdhury, G., McLeod, J., Gillet, V., and Willett, P., editors, *Transforming Digital Worlds – 13th International Conference, iConference 2018, Proceedings*, pages 367–377, Germany. Springer International Publishing.

Holness, T. and Ramji-Nogales, J. (2016). Participation as reparations: The eccc and healing in cambodia. In van Schaack, B. and Reicherter, D., editors, *Cambodia's Hidden Scars: Trauma Psychology and the Extraordinary Chambers in the Courts of Cambodia*, pages 213–234. Documentation Center of Cambodia, 2nd edition.

Hornstein, S., Scharfenberger, J., Lueken, U., et al. (2024). Predicting recurrent chat contact in a psychological intervention for the youth using natural language processing. *NPJ Digital Medicine*, 7:132.

Hu, Y., Hosseini, M., Parolin, E. S., Osorio, J., Khan, L., Brandt, P., and D'Orazio, V. (2022). Conflibert: A pre-trained language model for political conflict and violence. In *Proceedings of NAACL-HLT 2022*, pages 5469–5482.

Hutto, C. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

ICTY (2016). *Infographic: ICTY Facts & Figures*. International Criminal Tribunal for the Former Yugoslavia. https://www.icty.org/node/9590.

Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., and Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review. *PloS One*, 13.

Jahan, M. S. and Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Ji, S. (2022). Towards intention understanding in suicidal risk assessment with natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4028–4038.

Ji, S., Zhang, T., Yang, K., Ananiadou, S., and Cambria, E. (2023). Rethinking large language models in mental health applications. *ArXiv*.

Ji, X., Chun, S. A., Wei, Z., and Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining*, 5:1–25.

Jurafsky, D. and Martin, J. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd (draft) edition. https://web.stanford.edu/˜jurafsky/slp3/.

Kanavou, A. A. and Path, K. (2017). The lingering effects of thought reform: The Khmer Rouge S-21 prison personnel. *The Journal of Asian Studies*, 76(1):87–105.

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., and Bilz, L. (2023). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, Violence, & Abuse*, 24(4):2598–2615.

Kaptein, R., Van den Broek, E. L., Koot, G., and Huis in 't Veld, M. A. A. (2013). Recall oriented search on the web using semantic annotations. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR '13)*, pages 45–48, San Francisco, California, USA. Association for Computing Machinery.

Karabacak, M. and Margetis, K. (2023). Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5).

Keydar, R. (2022). Changing the lens on survivor testimony: Topic modeling the eichmann trial. *BJewish Studies Quarterly*, 29(4):412–435.

Khatua, A., Cambria, E., and Khatua, A. (2018). Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400. IEEE.

Kovács, G., Alonso, P., and Saini, R. (2021). Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2(2):95.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.

Krug, E. G., Mercy, J. A., Dahlberg, L. L., and Zwi, A. B. (2002). The world report on violence and health. *The Lancet*, 360(9339):1083–1088.

Kumari, R. and Ganagwar, R. (2018). A critical study of digital nonverbal communication in interpersonal and group communication: In context of social media. *International Journal of Communication and Media Studies*, 8(4):1–12.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Alstyne, M. V. (2009). Computational social science. *Science*, 323(5915):721–723.

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., et al. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., et al. (2021). Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.

Lee, C. S. and Jang, A. (2023). Questing for justice on twitter: Topic modeling of# stopasianhate discourses in the wake of atlanta shooting. *Crime & Delinquency*, 69(13-14):2874–2900.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lehrner, A. and Yehuda, R. (2018). Trauma across generations and paths to adaptation and resilience. *Psychological Trauma: Theory, Research, Practice, and Policy*, 10(1):22–29.

Levis, M., Westgate, C. L., Gui, J., Watts, B. V., and Shiner, B. (2021). Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychological Medicine*, 51(8):1382–1391.

Li, L., Fan, L., Atreja, S., and Hemphill, L. (2023). ”hot” chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2020). A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*.

Lindert, J. and Levav, I., editors (2016). *Violence and Mental Health*. Springer.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Wert, G., Greenawald, B., Al Boni, M., and Brown, D. E. (2018). Predicting violent behavior using language agnostic models. *In Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 102–109.

Lopez-Sanchez, M. and Müller, A. (2021). On simulating the propagation and countermeasures of hate speech in social networks. *Applied Sciences*, 11(24).

Low, D. M., Rumker, L., Torous, J., Cecchi, G., Ghosh, S. S., and Talkar, T. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of Medical Internet Research*, 22(10):e22635.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

Luscombe, A., Dick, K., and Walby, K. (2022). Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*, 56(3):1023–1044.

Macanovic, A. (2022). Text mining for social science–the state and the future of computational text analysis in sociology. *Social Science Research*, 108:102784.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

MacFarlane, A., Russell-Rose, T., and Shokraneh, F. (2021). Search strategy formulation for systematic reviews: Issues, challenges and opportunities. *arXiv preprint arXiv:2112.09424*.

Mansour, S. (2018). Social media analysis of user's responses to terrorism using sentiment analysis and text mining. *Procedia Computer Science*, 140:95–103.

Marmar, C. R., Brown, A. D., Qian, M., Laska, E., Siegel, C., Li, M., Abu-Amara, D., Tsiartas, A., Richey, C., Smith, J., et al. (2019). Speech-based markers for posttraumatic stress disorder in us veterans. *Depression and Anxiety*, 36(7):607–616.

Matter, D., Schirmer, M., Grinberg, N., and Pfeffer, J. (2024). Investigating the increase of violent speech in incel communities with human-guided gpt-4 prompt iteration. *Frontiers in Social Psychology*, 2:1383152.

McLachlan, J. and McHarg, J. (2005). Ethical permission for the publication of routinely collected data. *Medical Education*, 39(9).

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Miranda, O., Kiehl, S. M., Qi, X., Brannock, M. D., Kosten, T., Ryan, N. D., Kirisci, L., Wang, Y., and Wang, L. (2024). Enhancing post-traumatic stress disorder patient assessment: Leveraging natural language processing for research of domain criteria identification using electronic medical records. *BMC Medical Informatics and Decision Making*, 24(1):1–14.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2019). Tackling online abuse: A survey of automated abuse detection methods. *Journal of Computational Linguistics*, 45(4):123–145.

Modha, S., Majumder, P., Mandl, T., and Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. *Expert Systems with Applications*, 161:113725.

Mohr, J. W. and Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter. *Poetics*, 41(6):545–569.

Montag, C. and Diefenbach, S. (2018). Towards homo digitalis: Important research issues for psychology and the neurosciences at the dawn of the internet of things and the digital society. *Sustainability*, 10(2):415.

Montejo-Ráez, A., Molina-González, M. D., Jiménez-Zafra, S. M., García-Cumbreras, M. Á., and García-López, L. J. (2024). A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review*, 53:100654.

Morstatter, F., Pfeffer, J., Mayer, K., and Liu, H. (2015). Text, topics, and turkers: A consensus measure for statistical topics. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 123–131.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., and Groh, G. (2022). SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Moturu, S. T. and Liu, H. (2011). Quantifying the trustworthiness of social media content. *Distributed and Parallel Databases*, 29:239–260.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2020). A BERT-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII*, pages 928–940, Cham. Springer International Publishing.

Müller, M., Salathé, M., and Kummervold, P. E. (2020). COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Nasution, A. H. and Onan, A. (2024). Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*.

Nghiem, H., Gupta, U., and Morstatter, F. (2024). "define your terms": Enhancing efficient offensive speech classification with definition. *arXiv preprint arXiv:2402.03221*.

Nghiem, H. and Morstatter, F. (2021). "stop asian hate!": Refining detection of anti-asian hate speech during the covid-19 pandemic. *arXiv preprint arXiv:2112.02265*.

Ni, Y., Barzman, D., Bachtel, A., Griffey, M., Osborn, A., and Sorter, M. (2020). Finding warning markers: Leveraging natural language processing and machine learning technologies to detect risk of school violence. *International Journal of Medical Informatics*, 139:104137.

Nielsen, F. A. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153.

Noor, S. and Bashir, S. (2015). Evaluating bias in retrieval systems for recall oriented documents retrieval. *International Arab Journal of Information Technology (IAJIT)*, 12(1):53–59.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 994–1009.

OpenAI (2023). Gpt-4 technical report. Technical report.

O'Connor, B., Bamman, D., and Smith, N. A. (2011). Computational text analysis for social science: Model assumptions and complexity. In *Second NeurlPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Papacharissi, Z. (2002). The presentation of self in virtual life: Characteristics of personal home pages. *Journal of Broadcasting & Electronic Media*, 46(3):346–367.

Park, J., Baek, Y. M., and Cha, M. (2014). Cross-cultural comparison of nonverbal cues in emoticons on twitter: Evidence from big data analysis. *Journal of Communication*, 64(2):333–354.

Parmigiani, G., Barchielli, B., Casale, S., Mancini, T., and Ferracuti, S. (2022). The impact of machine learning in predicting risk of violence: A systematic review. *Frontiers in Psychiatry*, 13:1015914.

Paz, M. A., Montero-Díaz, J., and Moreno-Delgado, A. (2020). Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022.

Peixoto, B., Lavi, B., Bestagini, P., Dias, Z., and Rocha, A. (2020). Multimodal violence detection in videos. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2957–2961. IEEE.

Pennycook, G. and Rand, D. G. (2018). The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. *Management Science*, 66(11):4944–4957.

Pfeffer, J., Matter, D., Jaidka, K., Varol, O., Mashhadi, A., Lasser, J., Assenmacher, D., Wu, S., Yang, D., Brantner, C., Romero, D. M., Otterbacher, J., Schwemmer, C., Joseph, K., Garcia, D., and Morstatter, F. (2023). Just another day on twitter: A complete 24 hours of twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1073–1081.

Phillips, J. B., Ingram, K. M., and Campion, K. (2024). Gendered extremism in the pacific on 4chan: A mixed-methods exploration of australian and new zealanders' concepts of women, gender, and sexual violence on/pol. *Terrorism and Political Violence*, pages 1–22.

Piotrowski, M. (2012). *Natural language processing for historical texts*, volume 17. Morgan & Claypool Publishers.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Prescott, J., Rathbone, A. L., and Brown, G. (2020). Online peer to peer support: Qualitative analysis of uk and us open mental health facebook groups. *Digital Health*, 6:2055207620979209.

Quillivic, R., Gayraud, F., Auxéméry, Y., Vanni, L., Peschanski, D., Eustache, F., Dayan, J., and Mesmoudi, S. (2024). Interdisciplinary approach to identify language markers for post-traumatic stress disorder using machine learning and deep learning. *Scientific Reports*, 14(1):12468.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Whisper: Openai's speech recognition system. `https://openai.com/research/whisper`. Accessed on 2022/11/11.

Ray, L. and Ray, P. L., editors (2018). *Violence and Society.* SAGE Publications.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., and Langer, E. J. (2017). Forecasting the onset and course of mental illness with twitter data. *Scientific Reports*, 7(1):13006.

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S., and Zannettou, S. (2021). The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 196–207.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Röttger, P., Seelawi, H., Nozza, D., Talat, Z., and Vidgen, B. (2022). Multilingual hatecheck: Functional tests for multilingual hate speech detection models. *arXiv preprint arXiv:2206.09917*.

Röttger, P., Vidgen, B., Hovy, D., and Pierrehumbert, J. B. (2021). Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Russell-Rose, T., Gooch, P., and Kruschwitz, U. (2021). Interactive query expansion for professional search applications. *Business Information Review*, 38(3):127–137.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213):1063–1064.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-G., Almerekhi, H., and Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10:1–34.

Saltzman, L. E. (2004). Issues related to defining and measuring violence against women: Response to kilpatrick. *Journal of Interpersonal Violence*, 19(11):1235–1243.

Sandick, P. A. (2012). Speechlessness and trauma: Why the International Criminal Court needs a public interviewing guide. *Northwestern Journal of International Human Rights*, 11(1):104–125.

Sawhney, R., Joshi, H., Gandhi, S., and Shah, R. (2020). A time-aware transformer-based model for suicide ideation detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697.

Schirmer, M., Brechenmacher, C., and Pfeffer, J. (2024a). Gentrac: A tool for tracing trauma in genocide and mass atrocity court transcripts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7666–7671.

Schirmer, M., Kruschwitz, U., and Donabauer, G. (2022). A new dataset for topic-based paragraph classification in genocide-related court transcripts. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4504–4512, Marseille, France. European Language Resources Association.

Schirmer, M., Leemann, T., Kasneci, G., Pfeffer, J., and Jurgens, D. (2024b). The language of trauma: Modeling traumatic event descriptions across domains with explainable ai. *arXiv preprint arXiv:2408.05977*.

Schirmer, M., Nolasco, I. M. O., Mosca, E., Xu, S., and Pfeffer, J. (2023a). Uncovering trauma in genocide tribunals: An nlp approach using the genocide transcript corpus. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 257–266.

Schirmer, M., Pfeffer, J., and Hilbert, S. (2023b). Talking about torture: A novel approach to the mixed methods analysis of genocide-related witness statements in the khmer rouge tribunal. *Journal of Mixed Methods Research*, page 15586898231218463.

Schirmer, M., Voggenreiter, A., and Pfeffer, J. (2024c). More skin, more likes! measuring child exposure and user engagement on tiktok. *arXiv preprint arXiv:2408.05622*.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Scrivens, R., Gill, P., and Conway, M. (2020). The role of the internet in facilitating violent extremism and terrorism: Suggestions for progressing research. In Holt, T. and Bossler, A., editors, *The Palgrave Handbook of International Cybercrime and Cyberdeviance*. Palgrave Macmillan, Cham.

Sharratt, S. (2016). *Gender, shame and sexual violence: The voices of witnesses and court members at war crimes tribunals*. Routledge.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org.

Siegel, A. A. (2020). Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pages 56–88.

Singh, M., Jakhar, A. K., and Pandey, S. (2021). Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining*, 11:33.

Smoliarova, A. S., Bodrunova, S. S., Yakunin, A. V., Blekanov, I., and Maksimov, A. (2018). Detecting pivotal points in social conflicts via topic modeling of twitter content. In *International Conference on Internet Science*, pages 61–71. Springer.

Soni, H. K., Sharma, S., and Sinha, G., editors (2024). *Text and Social Media Analytics for Fake News and Hate Speech Detection*. Chapman and Hall/CRC, 1st edition.

Soral, W., Bilewicz, M., and Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, 44(2):136–146.

Steinert, T. (2002). Prediction of inpatient violence. *Acta Psychiatrica Scandinavica*, 106:133–141.

Stephanie, E. M. A., Ruiz, L. G. B., Vila, M. A., and Pegalajar, M. C. (2024). Study of violence against women and its characteristics through the application of text mining techniques. *International Journal of Data Science and Analytics*, 18(1):35–48.

Strand, M., Eng, L. S., and Gammon, D. (2020). Combining online and offline peer support groups in community mental health care settings: A qualitative study of service users' experiences. *International Journal of Mental Health Systems*, 14:1–12.

Strathern, W. and Pfeffer, J. (2023). Identifying different layers of online misogyny. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media.*

Subramani, S., Michalska, S., Wang, H., Du, J., Zhang, Y., and Shakeel, H. (2019). Deep learning for multi-class identification from domestic violence online posts. *IEEE Access*, 7:46210–46224.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, 7(3):321–326.

Sun, C., Huang, L., and Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org.

Swaminathan, A., López, I., Mar, R. A. G., Heist, T., McClintock, T., Caoili, K., Grace, M., Rubashkin, M., Boggs, M. N., Chen, J. H., et al. (2023). Natural language processing system for rapid detection and intervention of mental health crisis chat messages. *NPJ Digital Medicine*, 6(1):213.

Tabassum, A. and Patil, R. R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, 7(06):4864–4867.

Talpur, B. A. and O'Sullivan, D. (2020). Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in twitter. In *Informatics*, volume 7, page 52. MDPI.

Tan, S., O'Halloran, K. L., Wignell, P., Chai, K., and Lange, R. (2018). A multimodal mixed methods approach for examining recontextualisation patterns of violent extremist images in online media. *Discourse, Context & Media*, 21:18–35.

Tang, X., Zou, A., Zhang, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. B. (2023). Medagents: Large language models as collaborators for zero-shot medical reasoning. *ArXiv*.

Taylor, J. and Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, 14:1–39.

Tejaswini, V., Sathya Babu, K., and Sahoo, B. (2024). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–20.

Teruel, M., Cardellino, C., Cardellino, F., Alemany, L. A., and Villata, S. (2018). Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Thiago, D. O., Marcelo, A. D., and Gomes, A. (2021). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732.

Thomas, D. R., Pastrana, S., Hutchings, A., Clayton, R., and Beresford, A. (2017). Ethical issues in research using datasets of illicit origin. In *Proceedings of the 2017 Internet Measurement Conference*.

Tontodimamma, A., Nissi, E., Sarra, A., and Fontanella, L. (2020). Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126:157–179.

Torregrosa, J., Bello-Orgaz, G., Martínez-Cámara, E., Ser, J. D., and Camacho, D. (2023). A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):9869–9905.

Tourni, I. C., Guo, L., Hu, H., Halim, E., Ishwar, P., Daryanto, T., Jalal, M., Chen, B., Betke, M., Zhafransyah, F., et al. (2024). Detecting frames in news headlines and lead images in us gun violence coverage. *arXiv preprint arXiv:2406.17213*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ul Alam, M. A. and Kapadia, D. (2020). Laxary: A trustworthy explainable twitter analysis model for post-traumatic stress disorder assessment. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 308–313.

Ureña, J., Romera, E. M., Casas, J. A., Viejo, C., and Ortega-Ruiz, R. (2015). Psychometric properties of psychological dating violence questionnaire: A study with young couples. *International Journal of Clinical and Health Psychology*, 15(1):52–60.

Valkenburg, P. M., Peter, J., and Schouten, A. P. (2006). Friend networking sites and their relationship to adolescents' well-being and social self-esteem. *CyberPsychology & Behavior*, 9(5):584–590.

Van der Kolk, B. A. (2003). *Psychological trauma.* American Psychiatric Publishing.

Van Le, D., Montgomery, J., Kirkby, K. C., and Scanlan, J. (2018). Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. *Journal of Biomedical Informatics*, 86:49–58.

Van Lent, L. G., Sungur, H., Kunneman, F. A., Van De Velde, B., and Das, E. (2017). Too far to care? measuring public attention and fear for ebola using twitter. *Journal of medical Internet research*, 19(6):e193.

van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., Mcgillivray, B., Colavizza, G., et al. (2020). Assessing the impact of ocr quality on downstream nlp tasks. In *ICAART 2020-Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, volume 1, pages 484–496. SciTePress.

Verdeja, E. (2016). Predicting genocide and mass atrocities. *Genocide Studies & Prevention*, 9(3).

Voggenreiter, A., Brandt, S., Putterer, F., Frings, A., and Pfeffer, J. (2024). The role of likes: How online feedback impacts users' mental health. In *Proceedings of the 16th ACM Web Science Conference*, pages 302–310.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wachs, S. and Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. *International Journal of Environmental Research and Public Health*, 15(9):2030.

Waltenberger, F., Höferlin, S., and Froehlich, M. (2023). Reddit insights: Improving online discussion culture by contextualizing user profiles. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Wang, X., Kim, H., Rahman, S., Mitra, K., and Miao, Z. (2024). Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Wankhade, M., Rao, A. C. S., and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Webb, H., Jirotka, M., Stahl, B., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O., and Burnap, P. (2017). The ethical challenges of publishing twitter data for research dissemination. In *Proceedings of the 2017 ACM on Web Science Conference*.

Westwood, S. J., Grimmer, J., Tyler, M., and Nall, C. (2022). Current research overstates american support for political violence. *Proceedings of the National Academy of Sciences*, 119(12):e2116870119.

Wilson, R. A. and Land, M. K. (2020). Hate speech on social media: Content moderation in context. *Conneticut Law Review*, 52:1029–1076.

Windisch, S., Wiedlitzka, S., Olaghere, A., and Jenaway, E. (2022). Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, 18(2):e1243.

Wu, L., Morstatter, F., Carley, K. M., and Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.

Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., and Yang, Z. (2020). Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer.

Xavier, H. S. (2023). Violentometer: Measuring violence on the web in real time. In *Companion Proceedings of the ACM Web Conference 2023*, pages 272–275.

Xiang, Z., Du, Q., Ma, Y., and Fan, W. (2018). Assessing reliability of social media data: lessons from mining tripadvisor hotel reviews. *Information Technology & Tourism*, 18:43–59.

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., and Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 75–78.

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., and Wang, D. (2024). Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Xu, X., Yao, B., Dong, Y., Yu, H., Hendler, J., Dey, A. K., and Wang, D. (2023). Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Xue, J., Chen, J., and Gelles, R. (2019). Using data mining techniques to examine domestic violence topics on twitter. *Violence and Gender*, 6(2):105–114.

Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., and Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610.

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z.-Z., and Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. *ArXiv*, pages 6056–6077.

Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., and Ananiadou, S. (2024). Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Yeh, C.-K., Kim, B., Arik, S. O., Li, C.-L., Pfister, T., and Ravikumar, P. (2019). On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32.

Yehuda, R. (1998). *Psychological trauma.* American Psychiatric Publishing.

Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Zadgaonkar, A. V. and Agrawal, A. J. (2021). An overview of information extraction techniques for legal document analysis and processing. *International Journal of Electrical & Computer Engineering (2088-8708)*, 11(6).

Zamith, R. and Lewis, S. C. (2015). Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*, 659(1):307–318.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1415–1420.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.