Social Network Analysis and Mining (2022) 12:59
https://doi.org/10.1007/s13278-022-00881-0

ORIGINAL ARTICLE

Identifying lexical change in negative word-of-mouth on social media

Wienke Strathern[1] · Raji Ghawi[1] · Mirco Schönfeld[2] · Jürgen Pfeffer[1]

Received: 25 March 2021 / Revised: 13 April 2022 / Accepted: 15 April 2022 / Published online: 7 June 2022

Abstract

Negative word-of-mouth is a strong consumer and user response to dissatisfaction. Moral outrages can create an excessive collective aggressiveness against one single argument, one single word, or one action of a person resulting in hateful speech. In this work, we examine the change of vocabulary to explore the outbreak of online firestorms on Twitter. The sudden change of an emotional state can be captured in language. It reveals how people connect with each other to form outrage. We find that when users turn their outrage against somebody, the occurrence of self-referencing pronouns like 'I' and 'me' reduces significantly. Using data from Twitter, we derive such linguistic features together with features based on retweets and mention networks to use them as indicators for negative word-of-mouth dynamics in social media networks. Based on these features, we build three classification models that can predict the outbreak of a firestorm with high accuracy.

1 Introduction

As social media platforms with hundreds of millions of users interacting in real time on topics and events all over the world, social media networks are social sensors for online discussions and are known for quick and often emotional disputes (Chadwick 2017). Online firestorms can be defined as the sudden discharge of large quantities of messages containing negative word of mouth and complaint behavior against a person, company or group in social media networks (Pfeffer et al. 2014). The negative dynamics often start with a collective "against the others" (Strathern et al. 2020).

In social media, negative opinions about products or companies are formed by and propagated via thousands or millions of people within hours. Furthermore, massive negative online dynamics are not only limited to the business domain, but they also affect organizations and individuals in politics. Even though online firestorms are a new phenomenon, their dynamics are similar to the way in which rumors are circulated. In 1947, Gordon Allport and Leo Postman defined a rumor as a "proposition for belief, passed along from person to person, usually by word of mouth, without secure standards of evidence being presented" (Allport and Postman 1947).

When people are active on social media, they act in a socio-technical system that is mediated and driven by algorithms. The goal of social media platforms is to keep users engaged and to maximize their time spent on the platform. Highly engaged users who spend a lot of time on platforms are the core of a social media business model that is based on selling more and better targeted ads. But the question is always which content will be interesting for a particular user? To answer this, recommendation systems are developed to increase the chance that a user will click on a suggested link and read its content. These recommendation algorithms incorporate socio-demographic information, but also data of a user's previous activity (Leskovec et al. 2014; Anderson 2006).

Furthermore, behavioral data of alters (friends) of a user are also used to suggest new content (Appel et al. 2020). Social scientists have studied the driving forces of social relationships for decades, i.e., why do people connect with each other. Homophily and transitivity are the most important factors for network formation. Homophily means that your friends are similar to yourself (McPherson et al. 2001). They like similar things and are interested in similar topics.

✉ Wienke Strathern
wienke.strathern@tum.de

Raji Ghawi
raji.ghawi@tum.de

Mirco Schönfeld
mirco.schoenfeld@uni-bayreuth.de

Jürgen Pfeffer
juergen.pfeffer@tum.de

[1] School of Social Science and Technology, Technical University of Munich, Munich, Germany

[2] University of Bayreuth, Bayreuth, Germany

Transitivity describes the fact that a person's friends are often connected among each other (Heider 1946; Cartwright and Harary 1956). Combining these two aspects results in the fact that most people are embedded in personal networks with people that are similar to themselves and who are to a high degree connected among each other.

The above-described forces of how humans create networks combined with recommendation systems have problematic implications. Recommendation systems filter the content that is presented on social media and suggest new "friends" to us. As a result, filter bubbles (Pariser 2011) are formed around individuals on social media, i.e., they are connected to like-minded people and familiar content. The lack of diversity in access to people and content can easily lead to polarization (Dandekar et al. 2013). If we now add another key characteristic of social media, abbreviated communication with little space for elaborate exchange, a perfect breeding ground for online firestorms emerges. Consider a couple of people disliking a statement or action of a politician, celebrity or any private individual and these people voicing their dislike aggressively on social media. Their online peers, who most likely have similar views (see above), will easily and quickly agree by sharing or retweeting the discontent. Within hours, these negative dynamics can reach tens of thousands of users (Newman et al. 2006). A major problem, however, is to capture first signals of online outrage at an early stage. Knowing about these signals would help to intervene in a proper way to avoid escalations and negative dynamics.

In previous work, Strathern et al. (2020) tackled the question of anomaly detection in a network by exploring major features that indicate the outbreak of a firestorm; hence, the goal was to early detect change and extract linguistic features. Detection of outrage (e.g., hate speech) is based on identification of predefined keywords, while the context in which certain topics and words are being used has to be almost disregarded. To name just one extreme example, hate groups have managed to escape keyword-based machine detection through clever combinations of words, misspellings, satire and coded language (Udupa 2020). The focus of the analysis of Strathern et al. was on more complex lexical characteristics, which they applied as a basis for automated detection.

Our research question is the following: On Twitter, there is constant fluctuation of content and tweets and the question arises if, in these fluctuations, we can detect early that a negative event starts solely based on linguistic features. We assume that the start of a firestorm is a process, and because of a sudden change of emotions it can be early detected in sentiments and lexical items. With this work, we aim at answering the following question: Once we identify the linguistic changes as indicators of a firestorm, can we also predict a firestorm? In an abstract view on a firestorm as depicted in Fig. 1, the indicators show at time point 1), whereas the firestorm takes place starting during the phase marked by 2) in the figure. Hence, in this paper, we build upon and extend the work presented by Strathern et al. (2020).

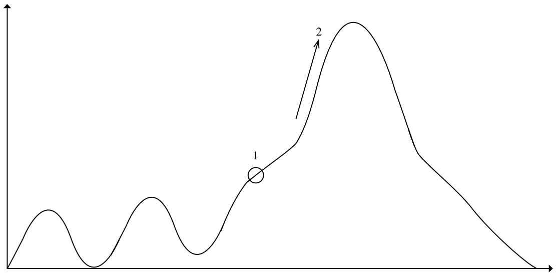


Fig. 1 Early detection of linguistic indicators (1) and prediction of firestorm (2)

Our choice of methods to answer our research question regarding the prediction of the beginning of online firestorms is based on text statistics and social network analysis for longitudinal network data. We assume that anomalies in behavior can be detected by statistical analysis applied to processes over time. Hence, in this work, we extract lexical and network-based properties, measure their occurrence for different tweet periods and use these features to predict the outbreak of a firestorm. For the scope of this work, we are mainly interested in textual data from tweets and in mention and retweet networks. We use quantitative linguistics to study lexical properties. For our linguistic analysis, we apply the Linguistic Inquiry Word Count Tool by Pennebaker et al. (2015). To contrast this linguistic perspective, we also investigate mention and retweet networks. Mentions and hashtags represent speech acts in linguistic pragmatics and are interesting in that they represent behavioral properties in addition to the lexical properties (Scott 2015). For predictive analysis, we define models based on linguistic features as well as models based on features derived from mention and retweet networks and compare them with each other.

Our contributions are:

- Extracting linguistic and sentimental features from textual data as indicators of firestorms.
- Defining a prediction model that accounts for linguistic features.

The remainder of the paper is organized as follows: Sect. 2 highlights important related works. In Sect. 3, we introduce the dataset used for this analysis together with a few descriptive statistics. What follows in Sects. 4 and 5 is a description of the linguistic and network-based features that our prediction is based upon. The prediction task is described in detail in Sect. 6. Section 7 concludes the paper.

2 Related work

While online firestorms are similar to rumors to some extent, e.g. they often rely on hearsay and uncertainty, online firestorms pose new challenges due to the speed and potential global reach of social media dynamics (Pfeffer et al. 2014). With respect to firestorms on social media, the analysis of social dynamics, their early detection and prediction often involves research from the field of sentiment analysis, network analysis as well as change detection. There is work asking why do people join online firestorms (Delgado-Ballester et al. 2021). Based on the concept of moral panics, the authors argue that participation behavior is driven by a moral compass and a desire for social recognition (Johnen et al. 2018). Social norm theory refers to understanding online aggression in a social–political online setting, challenging the popular assumption that online anonymity is one of the principle factors that promote aggression (Rost et al. 2016).

2.1 Sentiment analysis

Approaches to the analysis of firestorms focusing on the mood of the users and their expressed sentiments unveil, for example, that in the context of online firestorms, non-anonymous individuals are more aggressive compared to anonymous individuals (Rost et al. 2016). Online firestorms are used as a topic of news coverage by journalists and explore journalists' contribution to attempts of online scandalization. By covering the outcry, journalists elevate it onto a mainstream communication platform and support the process of scandalization. Based on a typology of online firestorms, the authors have found that the majority of cases address events of perceived discrimination and moral misconduct aiming at societal change (Stich et al. 2014). Online firestorms on social media have been studied to design an Online Firestorm Detector that includes an algorithm inspired by epidemiological surveillance systems using real-world data from a firestorm (Drasch et al. 2015).

Sentiment analysis was applied to analyze the emotional shape of moral discussions in social networks (Brady et al. 2017). It has been argued that moral–emotional language increased diffusion more strongly. Highlighting the importance of emotion in the social transmission of moral ideas, the authors demonstrate the utility of social network methods for studying morality. A different approach is to measure emotional contagion in social media and networks by evaluating the emotional valence of content the users are exposed to before posting their own tweets (Ferrara and Yang 2015). Modeling collective sentiment on Twitter gave helpful insights about the mathematical approach to sentiment dynamics (Charlton et al. 2016).

Arguing that rational and emotional styles of communication have strong influence on conversational dynamics, sentiments were the basis to measure the frequency of cognitive and emotional language on Facebook. Bail et al. (2017).

Instead, the analysis of linguistic patterns was used to understand affective arousal and linguist output (Sharp and Hargrove 2004). Extracting the patterns of word choice in an online social platform reflecting on pronouns is one way to characterize how a community forms in response to adverse events such as a terrorist attack (Shaikh et al. 2017). Synchronized verbal behavior can reveal important information about social dynamics. The effectiveness of using language to predict change in social psychological factors of interest can be demonstrated nicely (Gonzales et al. 2010). In Lamba et al. (2015), the authors detected and described 21 online firestorms discussing their impact on the network. To advance knowledge about firestorms and the spread of rumors, we use the extracted data as a starting point to follow up on the research findings.

2.2 Network analysis

Social media dynamics can be described with models and methods of social networks (Wasserman and Faust 1994; Newman 2010; Hennig et al. 2012). Approaches mainly evaluating network dynamics are, for example, proposed by Snijders et al. Here, network dynamics were modeled as network panel data (Snijders et al. 2010). The assumption is that the observed data are discrete observations of a continuous-time Markov process on the space of all directed graphs on a given node set, in which changes in tie variables are independent conditional on the current graph. The model for tie changes is parametric and designed for applications to social network analysis, where the network dynamics can be interpreted as being generated by choices made by the social actors represented by the nodes of the graph. This study demonstrated ways in which network structure reacts to users posting and sharing content. While examining the complete dynamics of the Twitter information network, the authors showed where users post and reshare information while creating and destroying connections. Dynamics of network structure can be characterized by steady rates of change, interrupted by sudden bursts (Myers et al. 2012). Network dynamics were modeled as a class of statistical models for longitudinal network data (Snijders 2001). Dynamics of online firestorms were analyzed using an agent-based computer simulation (ABS) (Hauser et al. 2017)—information diffusion and opinion adoption are triggered by negative conflict messages.

2.3 Classification in machine learning

In order to efficiently analyze big data, machine learning methods are used, with the goal of learning from experience in certain tasks. In particular, in *supervised learning*, the goal is to predict some output variable that is associated with each input item. This task is called *classification* when the output variable is a category. Many standard classification algorithms have been developed over the last decades, such as logistic regression, random forests, k nearest neighbors, support vector machines and many more (Friedman et al. 2001; James et al. 2014).

Machine learning methods have been used widely for studying users' behavior on social media (Ruths and Pfeffer 2014), predicting the behavior of techno-social systems (Vespignani 2009) and predicting consumer behavior with Web search (Goel et al. 2010). Moreover, such methods are also used in identifying relevant electronic word of mouth in social media (Vermeer et al. 2019; Strathern et al. 2021).

2.4 Mixed approaches

More recent approaches analyze online firestorms by analyzing both content and structural information. A text-mining study on online firestorms evaluates negative eWOM that demonstrates distinct impacts of high- and low-arousal emotions, structural tie strength, and linguistic style match (between sender and brand community) on firestorm potential (Herhausen et al. 2019). Online Firestorms were studied to develop optimized forms of counteraction, which engage individuals to act as supporters and initiate the spread of positive word of mouth, helping to constrain the firestorm as much as possible (Mochalova and Nanopoulos 2014). By monitoring psychological and linguistic features in the tweets and network features, we combine methods from text analysis, social network analysis and change detection to early detect and predict the start of a firestorm.

3 Data

To address our research question, we examined 20 different firestorms. Some are directed against individuals and a single statement; some are against companies, campaigns and marketing actions. They have all received widespread public attention in social media as well as mainstream media. As shown in Table 1, there are hashtags and also @mentions that name the target.

3.1 Dataset

We used the same set of firestorms as in Lamba et al. (2015), whose data source is an archive of the Twitter decahose, a random 10% sample of all tweets. This is a scaled up version of Twitter's Sample API, which gives a stream of a random 1% sample of all tweets.

Table 1 Firestorm events sorted by number of tweets

Firestorm hashtag/mention	Tweets	Users	First day
#whyimvotingukip	39,969	32,382	2014-05-21
#muslimrage	15,721	11,952	2012-09-17
#CancelColbert	13,277	10,353	2014-03-28
#myNYPD	12,762	10,362	2014-04-23
@TheOnion	9959	8803	2013-02-25
@KLM	8716	8050	2014-06-29
#qantas	8649	5405	2011-10-29
@David_Cameron	7096	6447	2014-03-06
suey_park	6919	3854	2014-03-28
@celebboutique	6679	6189	2012-07-20
@GaelGarciaB	6646	6234	2014-06-29
#NotIntendedtobeaFactualStat.	6261	4389	2011-04-13
#AskJPM	4321	3418	2013-11-14
@SpaghettiOs	2890	2704	2013-12-07
#McDStories	2374	1993	2012-01-24
#AskBG	2221	1933	2013-10-17
#QantasLuxury	2098	1658	2011-11-22
#VogueArticles	1894	1819	2014-09-14
@fafsa	1828	1693	2014-06-25
@UKinUSA	142	140	2014-08-27

Mention and retweet networks based on these samples can be considered as *random edge sampled* networks (Wagner et al. 2017) since sampling and network construction is based on Tweets that constitute the links in the network. As found by Morstatter et al. (2013), the Sample API (unlike the Streaming API) indeed gives an accurate representation of the relative frequencies of hashtags over time. We assume that the decahose has this property as well, with the significant benefit that it gives us more statistical power to estimate the true size of smaller events.

The dataset consists of 20 firestorms with the highest volume of tweets as identified in Lamba et al. (2015). Table 1 shows those events along with the number of tweets, number of users, and the date of the first day of the event. The set of tweets of each firestorm covers the first week of the event. We also augmented this dataset via including additional tweets, of the same group of users, during the same week of the event (7 days) and the week before (8 days), such that the volume of tweets is balanced between the 2 weeks (about 50% each). The fraction of firestorm-related tweets is between 2 and 8% of the tweets of each event (Table 1)—it is important to realize at this point that even for users engaging in online firestorms, this activity is a minor part of their overall activity on the platform.

Thus, for each of the 20 firestorms, we have three types of tweets: (1) tweets related to the firestorm, (2) tweets posted 1 week before the firestorm and (3) tweets posted during the firestorm (same week) but not related to it. Let us denote these three sets of tweets T_1, T_2 and T_3, respectively.

For each event, we also extracted tweets metadata including timestamp, hashtags, mentions and retweet information (user and tweet ID).[1]

4 Linguistic features

Negative word-of-mouth sometimes contains strong emotional expressions and even highly aggressive words against a person or a company. Hence, the start of a firestorm might be indicated by a sudden change of vocabulary and emotions. Do people become emotionally thrilled and can we find changes in tweets? Can we capture a change of perspective in the text against a target? Emotionality is reflected in words, the first analysis is based on the smallest structural unit in language: words (Bybee and Hopper 2001).

4.1 Extraction of features

To extract linguistic features and sentiment scores we use the Linguistic Inquiry Word Count classification scheme, short LIWCTool (Pennebaker et al. 2015). In this way, first textual differences and similarities can be quantified by simple word frequency distribution (Baayen 1993). Furthermore, to understand emotions in tweets we use the sentiment analysis provided by the LIWCTool. Essentially, sentiment analysis is the automatic determination of the valence or polarity of a text part, i.e., the classification of whether a text part has a positive, negative or neutral valence. Basically, automatic methods of sentiment analysis work either lexicon based or on the basis of machine learning. Lexicon-based methods use extensive lexicons in which individual words are assigned positive or negative numerical values to determine the valence of a text section (usually at the sentence level) of a text part (mostly on sentence level) (Tausczik and Pennebaker 2009).

LIWC contains a dictionary with about 90 output variables, so each tweet is matched with about 90 different categories. The classification scheme is based on psychological and linguistic research. Particularly, we were interested in sentiments to see if users show ways of aggressiveness during firestorms compared to non-firestorm periods. Furthermore, we would like to know which lexical items differ in different phases. We extracted 90 lexical features for each tweet of each of the 20 firestorms. We used variables that give standard linguistic dimensions (percentage of words in the text that are pronouns, articles, auxiliary verbs) and informal language markers (percentage of words that refer to the category assents, fillers, swear words, netspeak). To discover sentiments, we also used the variables affective processes, cognitive processes, perceptual processes. The categories provide a sentiment score of positivity and negativity to every single tweet. We also considered the category posemo and negemo to see if a tweet is considered positive or negative. We also constructed our own category ‘emo’ by calculating the difference between positive and negative sentiments in tweets. Thus, weights of this category can be negative and should describe the overall sentiment of a tweet.

These categories each contain several subcategories that can be subsumed under the category names. The category of personal pronouns, for example, contains several subcategories referring to personal pronouns in numerous forms. One of these subcategories ‘I,’ for example, includes—besides the pronoun ‘I’—‘me,’ ‘mine,’ ‘my,’ and special netspeak forms such as ‘idk’ (which means “I don’t know”).

Netspeak is a written and oral language, an internet chat, which has developed mainly from the technical circumstances: the keyboard and the screen. The combination of technology and language makes it possible to write the way you speak (Crystal 2002).

Finally, for each individual subcategory, we obtain the mean value of the respective LIWC values for the firestorm tweets and the non-firestorm tweets. Comparing these values gives first insights about lexical differences and similarities.

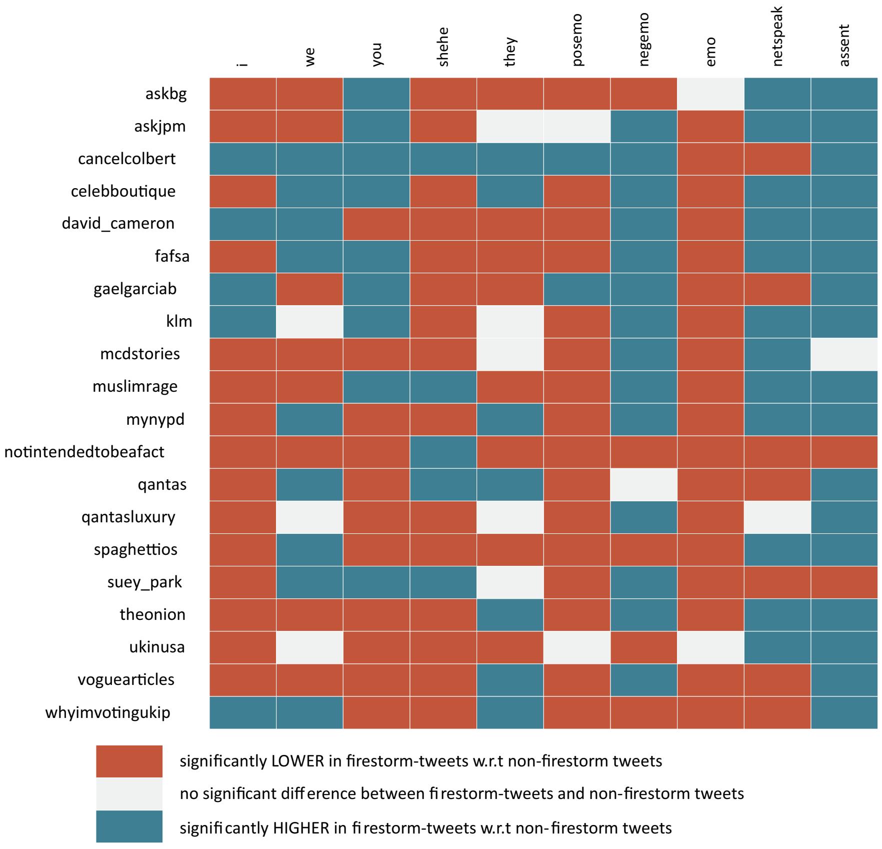
4.2 Comparing firestorm and non-firestorm tweets

In order to explore how the linguistic and sentiment features of tweets change during firestorms, we perform comparisons between firestorm tweets and non-firestorm tweets with regard to the individual LIWC subcategories. The firestorm tweets (T_1) were compared with tweets from the same user accounts from the week immediately before the firestorm (T_2) and the same week of the firestorm (T_3). We used t-tests to compare the mean value of the respective LIWC values for the firestorm tweets and the non-firestorm tweets, where the level of statistical significance of those tests is expressed using p-values (we used $p < 0.01$).

Figure 2a depicts the comparisons between firestorm tweets and non-firestorm tweets with regard to the individual subcategories. Every subcategory was examined separately for all 20 firestorms.

[1] Comparing with (Lamba et al. 2015), we have excluded ‘AskThicke’ firestorm, because it has a gap of 24 h between T_2 and T_1; hence, we added ‘suey_park’ firestorm instead.

Fig. 2 Comparison between firestorm-related tweets ($T1$) and non-firestorm tweets ($T2$ and $T3$) w.r.t various linguistic features using T-tests with p value < 0.01



	i	we	you	shehe	they	posemo	negemo	emo	netspeak	assent
askbg	🟥	🟥	🟦	🟥	🟥	🟥	🟥	⬜	🟦	🟦
askjpm	🟥	🟥	🟦	🟥	⬜	⬜	🟦	🟥	🟦	🟦
cancelcolbert	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟥	🟥	🟦
celebboutique	🟥	🟦	🟦	🟥	🟦	🟥	🟦	🟥	🟦	🟦
david_cameron	🟦	🟦	🟥	🟥	🟥	🟥	🟦	🟥	🟦	🟦
fafsa	🟥	🟦	🟦	🟥	🟥	🟥	🟦	🟥	🟦	🟦
gaelgarciab	🟦	🟥	🟦	🟥	🟥	🟦	🟦	🟥	🟥	🟦
klm	🟦	⬜	🟦	🟥	⬜	🟥	🟦	🟥	🟦	🟦
mcdstories	🟥	🟥	🟥	🟥	⬜	🟥	🟦	🟥	🟦	⬜
muslimrage	🟥	🟥	🟦	🟦	🟥	🟥	🟦	🟥	🟦	🟦
mynypd	🟥	🟦	🟥	🟥	🟦	🟥	🟦	🟥	🟦	🟦
notintendedtobeafact	🟥	🟥	🟥	🟦	🟥	🟥	🟥	🟥	🟥	🟥
qantas	🟥	🟦	🟥	🟦	🟦	🟥	⬜	🟥	🟥	🟦
qantasluxury	🟥	⬜	🟥	🟥	⬜	🟥	🟦	🟥	⬜	🟦
spaghettios	🟥	🟦	🟥	🟥	🟥	🟥	🟥	🟥	🟦	🟦
suey_park	🟥	🟦	🟦	🟦	⬜	🟥	🟦	🟥	🟥	🟥
theonion	🟥	🟥	🟥	🟥	🟦	🟥	🟦	🟥	🟦	🟦
ukinusa	🟥	⬜	🟥	🟥	🟥	⬜	🟥	⬜	🟦	🟦
voguearticles	🟥	🟥	🟥	🟥	🟦	🟥	🟦	🟥	🟥	🟦
whyimvotingukip	🟦	🟦	🟥	🟥	🟦	🟥	🟥	🟥	🟥	🟦

🟥 significantly LOWER in firestorm-tweets w.r.t non-firestorm tweets

⬜ no significant difference between firestorm-tweets and non-firestorm tweets

🟦 significantly HIGHER in firestorm-tweets w.r.t non-firestorm tweets

(a) Results of comparison

	I	we	you	shehe	they	posemo	negemo	emo	netspeak	assent
lower	15	8	11	15	8	16	5	18	7	2
same	0	3	0	0	5	2	1	2	1	1
higher	5	9	9	5	7	2	14	0	12	17

(b) Number of firestorms

The blue (turquoise) cells represent the firestorms in which terms from the respective category occurred more frequently during the firestorms. The red (brick) cells represent the firestorms in which the same words occurred less frequently during the firestorms. The light gray cells represent the firestorms in which there is no significant difference between firestorm tweets and non-firestorm tweets.

The results of comparison are aggregated in the table in Fig. 2b, which shows, for each feature, the number of firestorms according to the three cases of comparison: lower, higher and same (no significant difference).

Results For category 'I' this means that in five firestorms people used words of this category significantly more often, while in 15 firestorms these words were used significantly less. Similar results are observed for category 'she/he' In addition to the category 'I', the categories 'posemo' and 'negemo' should also be highlighted. Words representing positive emotions like 'love,' 'nice,' 'sweet'—the 'posemo' category—are used significantly less in almost all firestorms: positive emotions were less present in 16 out of 20 firestorms. For the category 'negemo,' which contains words representing negative emotions, this effect is reversed for all tweets—words in this category are used significantly more often during most of the firestorms (14 out of 20). There are 18 firestorms in which the 'emo' values were significantly lower during a firestorm. At the same time, there

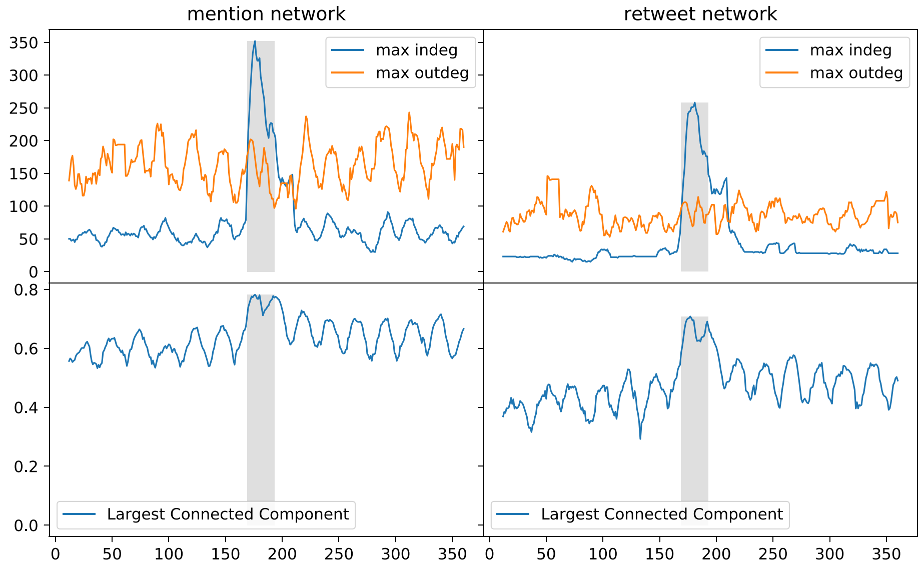


Fig. 3 Evolution of network features over time (#myNYPD firestorm). Highlighted area indicates the start of the firestorm (first 24 h)

were only two firestorms where the differences in the values of 'emo' were not significant. Another remarkable category is 'assent,' which contains words like 'agree,' 'OK,' 'yes.' In this category, the effect is also reversed—words in this category are used significantly more often during almost all firestorms (17 out of 20). *Interpretation.* We can state that during firestorms, the I vanishes and users talk significantly less about themselves compared to non-firestorm periods. Simultaneously, the positivity in firestorms tweets vanishes and negativity rises.

5 Mention and retweet networks

Besides linguistic features and sentiments expressed in tweets, online firestorms have also impact on the structure of user's social networks, such as mention and retweet networks.

To get insight on the evolution of each firestorm over time, we first split the time-line of each of the firestorm datasets into buckets of *one hour* and assign tweets to buckets based on their timestamp. The result of this splitting is a series of about 360 time slices (since the studied time-span of an event is 15 days). This allows us to perform analysis at fine granularity.

First, at each time slice, we extract several *basic features* of the corresponding hourly buckets of tweets, including:

- Number of tweets N_t
- Number of mention tweets N_{mt}
- Number of mentions N_m
- Ratio of mention tweets to all tweets N_{mt}/N_t.
- Mention per tweet ratio: N_m/N_t.

Moreover, at each time point we construct *mention networks*, and *retweet networks* taking into account all the tweets during the last 12 h. This way, we obtain a moving window of tweets: with a window size of 12 slices at steps of 1 h. The *mention network* of each moving window contains an edge ($user_1$, $user_2$) if a tweet (among tweets under consideration) posted by $user_1$ contains a mention to $user_2$. The *retweet network* of each moving window contains an edge ($user_1$, $user_2$) if a tweet (among tweets under consideration) posted by $user_1$ is a retweet of another (original) tweet posted by $user_2$.

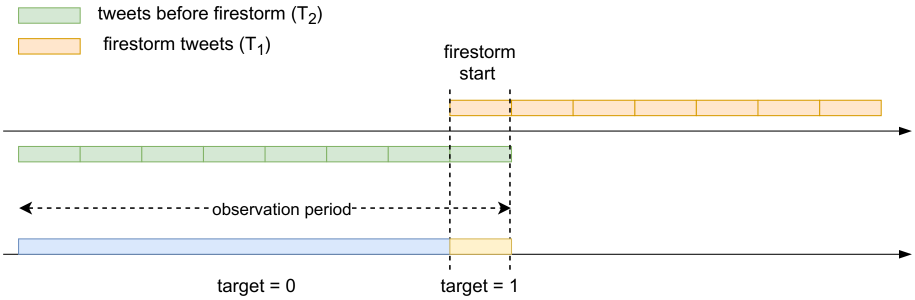
For each event, the mention networks constructed at different time points are directed, unweighted networks. We performed several types of social network analysis and extracted a set of metrics, including:

- Number of nodes N and edges E,
- Average out-degree (which equals avg. in-degree).
- Maximum out-degree and maximum in-degree.
- Relative size of the largest connected component.

Each of the aforementioned features leads to a time-series when taken over the entire time-span of the event. For example, Fig. 3 depicts some of those time-series for the features of the mention and retweet networks of #myNYPD firestorm, showing how those features evolve over time. While network metrics are affected by sampled datasets, we still believe that these metrics are meaningful since the sampling process was consistent over all firestorms.

Results One can clearly observe the oscillating behavior of those features. This oscillation is due to the alternation of tweeting activity between daytime and night. More interesting observation is the manifest change of behavior that occurs near the middle of the time span, which evidently

Fig. 4 Timeline of a firestorm



signals the beginning of the firestorm event. This apparent change can be observed in most of the features for the event at hand. However, not all the features are useful to detect the trigger of the firestorm in all events. In particular, we find the maximum in-degree feature is one of the best features to detect this change. This feature can clearly detect the start of the firestorm (in all events). The maximum in-degree in mention networks means the highest number of mentions received by a particular user.

Interpretation Thus, the ability of this feature to detect a firestorm can be interpreted by considering that, generally speaking, a firestorm occurs when one user is being mentioned unusually high. This result is intuitive since Tweets related to a certain Firestorm normally mention the victim's Twitter account.

Monitoring this feature in real-time would be certainly handy at detecting firestorms as early as possible, by signaling abnormal changes (increase) in this feature. However, the change of focus to a particular user can be the result of different (including positive) events.

From a network perspective, an online firestorm occurs when one user is mentioned unusually high, focusing on a Twitter handle or a hashtag. The maximum in-degree in @ mention networks is significantly deviating from comparable time periods.

6 Predicting the start of a firestorm

In the previous section we identified slight changes in lexical and sentimental cues as indicators of a firestorm. From a network perspective, we identified the maximum in-degree to be a very good indicator for a firestorm to occur. Based on these findings we want to test and compare our extracted features for a classification task in order to build models for predicting the start of a firestorm.

6.1 Prediction models (predictor variables)

As mentioned earlier, we split the time-line of each firestorm into buckets of *one hour* and assign tweets to buckets based on their timestamp.

Thus, for each time slice, the corresponding bucket of tweets is described by several features. Mainly, we distinguish between different types of features; each type of them defines a prediction model:

- *Baseline model* includes the basic features, such as number of tweets N_t, number of mentions N_m, etc. (see Sect. 5).
- *Mention-network model* includes network features, such as, number of nodes and edges, density, reciprocity, average and max in-degree and out-degree, etc., extracted from *mention* networks.
- *Retweet-network model* includes the same set of network features extracted from *retweet* networks.
- *Linguistic model* extends the basic model by including linguistic features, i.e., the mean values of extracted LIWC features (over the hourly bucket of tweets). In particular, we are interested in the following features: pronouns: namely: 'i,' 'we,' 'you,' 'shehe,' and 'they'; emotions: 'posemo,' 'negemo' and 'emo'; and 'netspeak' and 'assent.'

By doing so, we create separate time series for each of the features mentioned above.

6.2 Target variable

As shown in Fig. 4, the time span of the two sets of tweets T_2 and T_1 is 8 and 7 days, respectively, with an overlap of 1 day between the two periods. We consider the first day of the firestorm as its start. Hence, we create a *target* variable whose value is 0 for the time points t occurring entirely before the firestorm (the first 7 days of T_2) and 1 for the time points t occurring during the first day of the firestorm. The rest of the firestorm days are omitted. Hence, we obtain about $7 \times 24 = 168$ time points[2] where $target = 0$ (negative instances), as well as 24 points where $target = 1$ (positive instances).

[2] This number slightly varies from one firestorm to another.

Table 2 Pearson correlation of basic features, network features and linguistic features with the *target* variable (#myNYPD firestorm)

Basic features		Network features			Linguistic features	
			Mention	Retweet		
N_t	0.70	N	0.76	0.83	i	−0.16
N_{mn}	0.71	E	0.80	0.86	we	0.12
N_m	0.67	density	−0.61	−0.68	you	−0.15
N_{mt}/N_t	0.40	recip.	0.34	−0.38	she/he	−0.06
N_m/N_t	0.21	lwcc	0.82	0.85	they	0.35
		avg d_{in}	0.82	0.87	posemo	0.01
		max d_{in}	0.96	0.96	negemo	0.27
		max d_{out}	−0.00	0.11	emo	−0.23
					netspeak	0.56
					assent	0.19

Our objective is thus to predict the value of this *target* variable using the aforementioned sets of predictors. Hence, the prediction turns into a binary classification task, where we want to classify whether a time point t belongs to the period of firestorm start (target = 1) or not (belongs to the period before the firestorm: target = 0), using different types of features of the tweets. This classification task needs to be performed for each firestorm separately and independently from other firestorms.

6.3 Comparing features between before and the start of the firestorm

Before we dive deeper into the details of the classification task, it is interesting at this point to look at how different predictor features correlate with our target variable (which indicates the firestorm start). This would help us get insight on the ability of those features to *predict* that target variable. For this purpose, we calculate the Pearson correlation of each feature with the target variable (its numeric value 0 or 1). Table 2 shows the correlation values for the case of #myNYPD firestorm.

We can observe that basic features—in particular, number of tweets N_t, number of mention tweets N_{mt} and number of mentions N_m—have a relatively strong positive correlation with the target variable.

This effect of strong positive correlation can be also observed for most of network features, such as number of nodes N and edges E, relative size of largest (weakly) connected component *lwcc*, avg. and max. in-degree. In contrast, density has a strong negative correlation, which means that this feature is lower at the start of the firestorm compared to before the firestorm. On the other hand, reciprocity has rather a weak correlation with the target variable; this correlation is positive for mention networks (+0.34) and negative for retweet networks (-0.38). Finally, max d_{out}, the maximum out-degree, has no correlation at all.

Regarding linguistic features, most of those features have weak correlation (positive or negative), or no correlation with the target variable. The highest correlations are for 'netspeak' (0.56) and 'they' (0.35).

6.4 Design of the classification task

6.4.1 Split into training and test sets

As in any supervised machine learning task, data instances need to be split into training and test subsets: the first is used to train the classifier while the other is used to test it, i.e., to evaluate its performance. Typically, such splitting of the dataset is performed in a *random* fashion, with, for example, 75% of instances for training and the remaining 25% for testing. Moreover, in order to make a more reliable evaluation, a *cross validation* approach is typically used, such as the k-folds method. In k-folds cross-validation, the dataset is split into k consecutive folds, and each fold is then used once as a validation while the $k-1$ remaining folds form the training set. This method generally results in a less biased model compared to other methods, because it ensures that every observation from the original dataset has the chance of appearing in the training and test set.

However, in our firestorm dataset(s), positive and negative classes are highly *unbalanced*, with a 1:7 ratio, i.e., for each positive instance there are 7 negative instances. To tackle this unbalanced issue, we use *stratified k-folds*, which is a variation in k-folds cross-validation that returns *stratified* folds, that is, the folds are made by preserving the percentage of samples for each class.

In this study, we opt to use $k = 4$, and the dataset is split hence into 4 stratified folds. Thus, when the dataset contains 24 positive samples, and 168 negative ones, then each fold will contain $24/4 = 6$ positive samples, and about $168/4 = 42$ negative ones. The training is also performed 4 times, each time one of the folds is used as a test set while the remaining 3 folds are used as a training set.

This means that, each time, the 24 positive instances will be distributed such that 6 instances will be in the test set and 18 instances in the training set. This approach avoids the undesired situations where the training is performed with very few or with too many positive instances. The overall evaluation score is calculated as the average over the 4 training times.

6.4.2 Feature scaling

In our case, different features have their values on very different scales. For instance, regarding network features, the number of nodes N and edges E are usually $> 10^3$, while density is $< 10^{-3}$ and reciprocity is $< 10^{-2}$. Thus, in order to improve the prediction accuracy, we need to avoid some low-scale features being overwhelmed by other high-scale ones; therefore, we use feature scaling in order to put the features roughly on the same scale.

We use the *standard scaling* approach, where each feature is standardized by centering and scaling to unit variance. The standard score of a sample x is calculated as: $z = (x - \mu(x))/\sigma(x)$ where μ is the mean of the samples, and σ is the standard deviation of the samples.

Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using transform. Standardization of a dataset is a common requirement for many machine learning algorithms, as they might behave badly if the individual features do not roughly look like standard normally distributed data (e.g., Gaussian with 0 mean and unit variance).

6.4.3 Algorithm

As a classification algorithm, we used the logistic regression algorithm. Logistic regression is a well-known and widely used classification algorithm which extends linear regression. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The logistic function is defined as: $\sigma(x) = 1/(1 + \exp(-x))$

6.4.4 Evaluation

As an evaluation measure, we used Accuracy, which is simply the fraction of correctly classified instances (to all instances). For each firestorm, the prediction accuracy is calculated as the average of the accuracy over the 4 folds.

6.5 Results

We applied the logistic regression algorithm to each firestorm using different prediction models: basic model, mention network model and linguistic model. Table 3 shows the overall accuracy for each firestorm, with respect to each prediction model. We can see that the prediction accuracy is pretty high in general where the accuracy is within the range of 87% to 100%.

Table 3 Accuracy of prediction models

	Basic	Linguistic	Mention	Retweet
askbg	0.926	0.916	0.958	0.953
askjpm	0.953	0.953	0.995	0.995
cancelcolbert	0.948	0.953	0.990	0.984
celebboutique	0.915	0.945	0.937	0.963
david_cameron	1.000	0.995	0.995	0.995
fafsa	0.932	0.943	0.989	0.989
gaelgarciab	0.906	0.885	0.956	0.956
klm	0.891	0.902	0.907	0.907
mcdstories	0.943	0.938	0.923	0.961
muslimrage	0.956	0.990	0.980	0.969
mynypd	0.958	0.984	1.000	0.995
notintendedto.	0.990	0.984	0.989	0.989
qantas	0.922	0.922	0.939	0.956
qantasluxury	0.932	0.943	0.972	0.972
spaghettios	0.944	0.964	0.989	0.989
suey_park	0.943	0.948	0.990	0.995
theonion	0.974	0.974	0.989	0.989
ukinusa	0.870	0.875	0.995	0.995
voguearticles	0.951	0.967	0.971	0.977
whyimvotingukip	0.943	0.969	0.956	0.950
avg.	0.940	0.948	0.971	0.974

For the basic model, the accuracy ranges between 87% (for ‘ukinusa’) and 100% (for ‘david_cameron’), with an average of 94%. For the linguistic model, the accuracy ranges between about 87% (e.g., ‘ukinusa’) and 99.5% (@David_Cameron), with an average of 95%.

Finally, the two network models, mention and retweet, show very similar results in general. The accuracy ranges between about 90% (klm) and 100% (myNYPD), with an average of 97%. Overall we can see that all the prediction models are able to predict the start of the firestorm with very high accuracy.

Interpretation Network models are slightly more accurate than the linguistic model, which is in turn slightly more accurate than the basic model. It is logical that in times of firestorms there are a lot of mentions, hashtags and retweets, i.e., explicit network properties. Even more important and interesting is the result that we can measure early changes already in the language and that these properties are much more important for the early detection of changes. The fact that we make a comparison here should illustrate how well our model works alongside other more explicit models.

7 Conclusion

Our goal was to predict the outbreak of a firestorm using linguistic and network-based features. Therefore, we examined the vocabulary of tweets from a diverse set of firestorms and compared it to non-firestorm tweets posted by the same users. Additionally, we measured features describing the mention and retweet networks also comparing firestorm with non-firestorm tweets. We used the features in a logistic regression model to predict the outbreak of firestorms. The identified linguistic and sentimental changes were good indicators for the outbreak of a firestorm.

Observing linguistic features, we found that during firestorms users talk significantly less about themselves compared to non-firestorm periods which manifested in significantly fewer occurrences of self-referencing pronouns like ‘I,’ ‘me’ and the like. Simultaneously, the positivity in firestorm tweets vanishes and negativity rises. Especially the change in the use of personal pronouns served as a good indicator for the outbreak of online firestorms. This change of subject to a different object of discussion could be observed in an increased mentioning of a user or a hashtag who/that was the target of a firestorm, hence the perspective changes. Users start pointing at others. This expressed itself in a maximum in-degree in mention networks that significantly deviated from comparable time periods giving evidence for the pragmatic action from a network perspective. However, we are aware of the fact that we have only measured cases in which the in-degree change happens in the context of something negative.

Our models were able to predict the outbreak of a firestorm accurately. We were able to classify the outbreak of a firestorm with high accuracy (above 87%) in all scenarios. It showed, however, that classification models using features derived from the mention and retweet networks performed slightly better than models based on linguistic features.

Overall, verbal interaction is a social process and linguistic phenomena are analyzable both within the context of language itself and in the broader context of social behavior (Gumperz 1968). From a linguistic perspective, the results give an idea of how people interact with one another. For this purpose, it was important to understand both the network and the speech acts. Changes in the linguistic and sentimental characteristics of the tweets thus proved to be early indicators of change in the parts of social media networks studied. Besides the fact that users changed their perspective, we could also observe that positivity in words vanished and negativity increased.

Future work could consider clustering firestorms according to their dynamics, i.e., can firestorms be differentiated in the way users ally against a target? This is of interest insofar as we know that negative PR can also mean profit for a company and that this is seen as less bad. Another pathway worth following would be to leverage contextualized word embeddings (Peters et al. 2018) to identify especially harmful words that demand early attention. Generally, the question of what motivates people to ally against a target is of great scientific and social interest.

Our results give insights about how negative word-of-mouth dynamics on social media evolve and how people speak when forming an outrage collectively. Our work contributed to the task of predicting outbreaks of firestorms. Knowing where a firestorm is likely to occur can help, for example, platform moderators to know where an intervention in a calming manner will be required. Ultimately, this can save individuals from being harassed and insulted in online social networks.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The lists of Tweet IDs of the analyzed data can be shared upon request.

Declarations

Conflict of interests The author(s) gratefully acknowledge the financial support from the Technical University of Munich—Institute for Ethics in Artificial Intelligence (IEAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the IEAI or its partners.

References

Allport G, Postman L (1947) The psychology of rumor. J Clin Psychol 3(4):402

Anderson C (2006) The long tail: why the future of business is selling less of more. Hyperion Books, New York

Appel G, Grewal L, Hadi R, Stephen AT (2020) The future of social media in marketing. J Acad Mark Sci 48(1):79–95

Auger IE, Lawrence CE (1989) Algorithms for the optimal identification of segment neighborhoods. Bull Math Biol 51(1):39–54

Baayen H (1993) Statistical models for frequency distributions: a linguistic evaluation. Comput Humanit 26:347–363

Bail CA, Brown TW, Mann M (2017) Channeling hearts and minds: advocacy organizations, cognitive-emotional currents, and public conversation. Am Sociol Rev 82(6):1188–1213

Brady WJ, Wills JA, Jost JT, Tucker JA, Bavel JJV (2017) Emotion shapes the diffusion of moralized content in social networks. PNAS 114(28):7313–7318

Bybee J, Hopper P (2001) Frequency and the emergence of linguistic structure. In: Bybee J, Hopper P (eds) Typological studies in language, vol 45. John Benjamins Publishing Company, Amsterdam, pp 1–24

Cartwright D, Harary F (1956) Structural balance: a generalization of Heider's theory. Psychol Rev 63(5):277–293

Chadwick A (2017) The hybrid media system: politics and power, 2nd edn. Oxford University Press, Oxford

Charlton N, Singleton C, Greetham DV (2016) In the mood: the dynamics of collective sentiments on Twitter. R Soc Open Sci 3(6):160162

Crystal D (2002) Language and the internet. IEEE Trans Prof Commun 45:142–144

Dandekar P, Goel A, Lee DT (2013) Biased assimilation, homophily, and the dynamics of polarization. Proc Natl Acad Sci 110(15):5791–5796

Delgado-Ballester E, López-López I, Bernal-Palazón A (2021) Why do people initiate an online firestorm? the role of sadness, anger, and dislike. Int J Electron Commer 25:313–337

Drasch B, Huber J, Panz S, Probst F (2015) Detecting online firestorms in social media. In: ICIS

Ferrara E, Yang Z (2015) Measuring emotional contagion in social media. PLOS ONE 10(11):e0142390

Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning. Springer series in statistics. Springer, New York

Goel S, Hofman J, Lahaie S, Pennock D, Watts D (2010) Predicting consumer behavior with Web search. In: Proceedings of the National Academy of Sciences. National Academy of Sciences Section: Physical Sciences, pp 17486–17490

Gonzales AL, Hancock JT, Pennebaker JW (2010) Language style matching as a predictor of social dynamics in small groups. Commun Res 37(1):3–19

Gumperz J (1968) The speech community. In: Duranti A (ed) Linguistic anthropology: a reader. Wiley, New York, pp 166–173

Hauser F, Hautz J, Hutter K, Füller J (2017) Firestorms: modeling conflict diffusion and management strategies in online communities. J Strat Inf Syst 26(4):285–321

Heider F (1946) Attitudes and cognitive organization. J Psychol 21:107–112

Hennig M, Brandes U, Pfeffer J, Mergel I (2012) Studying social networks. A guide to empirical research. Campus Verlag, Frankfurt

Herhausen D, Ludwig S, Grewal D, Wulf J, Schoegel M (2019) Detecting, preventing, and mitigating online firestorms in brand communities. J Mark 83(3):1–21

Jackson B, Scargle JD, Barnes D, Arabhi S, Alt A, Gioumousis P, Gwin E, Sangtrakulcharoen P, Tan L, Tsai TT (2005) An algorithm for optimal partitioning of data on an interval. IEEE Signal Process Lett 12(2):105–108

James G, Witten D, Hastie T, Tibshirani R (2014) An introduction to statistical learning. Springer, Cham

Johnen M, Jungblut M, Ziegele M (2018) The digital outcry: what incites participation behavior in an online firestorm? New Media Soc 20(9):3140–3160

Killick R, Fearnhead P, Eckley IA (2012) Optimal detection of changepoints with a linear computational cost. J Am Stat Assoc 107(500):1590–1598

Lamba H, Malik MM, Pfeffer J (2015) A tempest in a teacup? Analyzing firestorms on Twitter. In: 2015 IEEE/ACM ASONAM. New York, NY, USA, pp 17–24

Leskovec J, Rajaraman A, Ullman JD (2014) Mining of massive datasets, 2nd edn. Cambridge University Press, Cambridge

McCulloh I, Carley KM (2011) Detecting change in longitudinal social networks. J Soc Struct 12(3):1–37

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444

Mochalova A, Nanopoulos A (2014) Restricting the spread of firestorms in social networks. In: ECIS 2014 proceedings

Morstatter F, Pfeffer J, Liu H, Carley K (2013) Is the sample good enough? comparing data from Twitter's streaming API with Twitter's firehose. In: Proceedings of the international AAAI conference on web and social media, vol 7, no 1

Myers S, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 33–41

Newman M (2010) Networks: an introduction. Oxford University Press Inc, New York

Newman M, Barabási AL, Watts DJ (2006) The structure and dynamics of networks. Princeton University Press, Princeton

Pariser E (2011) The filter bubble. What the internet is hiding from you. The New York Press, New York

Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015. Technical report, The University of Texas at Austin

Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: NAACL-HLT 2018, pp 2227–2237

Pfeffer J, Zorbach T, Carley KM (2014) Understanding online firestorms: negative word-of-mouth dynamics in social media networks. J Mark Commun 20(1–2):117–128

Rost K, Stahel L, Frey BS (2016) Digital social norm enforcement: online firestorms in social media. PLoS ONE 11(6):e0155923

Ruths D, Pfeffer J (2014) Social media for large studies of behavior. Science 346:1063–1064

Scott K (2015) The pragmatics of hashtags: inference and conversational style on Twitter. J Pragmat 81:8–20

Scott AJ, Knott M (1974) A cluster analysis method for grouping means in the analysis of variance. Biometrics 30:507–512

Sen A, Srivastava MS (1975) On tests for detecting change in mean. Ann Stat 3(1):98–108

Shaikh S, Feldman LB, Barach E, Marzouki Y (2017) Tweet sentiment analysis with pronoun choice reveals online community dynamics in response to crisis events. In: Advances in cross-cultural decision making. Springer, pp 345–356

Sharp WG, Hargrove DS (2004) Emotional expression and modality: an analysis of affective arousal and linguistic output in a computer vs. paper paradigm. Comput Hum Behav 20(4):461–475

Snijders TAB (2001) The statistical evaluation of social network dynamics. Sociol Methodol 31(1):361–395

Snijders TA, Koskinen J, Schweinberger M (2010) Maximum likelihood estimation for social network dynamics. Ann Appl Stat 4(2):567–588

Stich L, Golla G, Nanopoulos A (2014) Modelling the spread of negative word-of-mouth in online social networks. J Decis Syst 23(2):203–221

Strathern W, Schönfeld M, Ghawi R, Pfeffer J (2020) Against the others! Detecting moral outrage in social media networks. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 322–326

Strathern W, Ghawi R, Pfeffer J (2021) Advanced statistical analysis of large-scale web-based data. In: Data science in economics and finance for decision makers. Edited by Per Nymand-Andersen

Tausczik YR, Pennebaker JW (2009) The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 29:24–54

Udupa S (2020) Artificial intelligence and the cultural problem of extreme speech. Social Science Research Council (20 December 2020)

Vermeer S, Araujo T, Bernritter S, Noort G (2019) Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media. Int J Res Mark 36:492–508

Vespignani A (2009) Predicting the behavior of techno-social systems. Science 325:425–428

Wagner C, Singer P, Karimi F, Pfeffer J, Strohmaier M (2017) Sampling from social networks with attributes. In: Proceedings of the WWW conference, pp 1181–1190

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge, MA

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.