Cartography M.Sc.

Master thesis

Sentiment-based spatial-temporal event detection in social media data

Lin Che

Technical University of Munich TUM

TU WIEN TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

TECHNISCHE UNIVERSITÄT DRESDEN

ITC UNIVERSITY OF TWENTE.

2019

Sentiment-based spatial-temporal event detection in social media data

submitted for the academic degree of Master of Science (M.Sc.)
conducted at the Department of Civil, Geo and Environmental Engineering
Technical University of Munich

Author:	Lin, Che
Study course:	Cartography M.Sc.
Supervisor:	Juliane Cron, Ruoxin Zhu
Reviewer:	Dr.-Ing. Eva Hauthal

Chair of the Thesis
Assessment Board: Prof. Dr. Liqiu Meng

Date of submission: 11.09.2019

Statement of Authorship

Herewith I declare that I am the sole author of the submitted Master's thesis entitled:

"Sentiment-based spatial-temporal event detection in social media data"

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Munich, 11.09.2019 Lin, Che

Acknowledgements

Throughout the writing of this thesis, I have received a great deal of support and assistance. I would first like to express my sincere gratitude to my supervisor, Juliane Cron, for her patience, motivation, enthusiasm and guidance, for her insightful comments, continuous assistance during the whole research and thesis writing process. Without her endless spiritual and practical support, this thesis would not have been possible. I owe my deepest gratitude to my supervisor, Ruoxin Zhu, who helped me choose the topic, gave me technical guidance and valuable feedback.

I would like to offer my special thanks to Chenyu Zuo, who gave me advice and guidance on the visualization part of this thesis and gave me unconditional help in life and career all the time. My thanks also go to Diao Lin for his meaningful suggestions on the thesis. I would also like to show great thankfulness to Dr.-Ing. Eva Hauthal and Dr. Corné van Elzakker for providing me enlightening advice during my proposal and mid-term presentation.

My heartfelt appreciation goes to all my friends and colleagues who helped me through the last half-year, kept me company, and gave me constructive suggestions and warm encouragement. I thank all the lecturers, the classmates from Cartography M.Sc, and lovely friends from all over the world I have met in the past two years. Last but not least, nobody has been more important to me than the members of my family. My thanks go to all of them for being caring and supporting all the time. I am truly indebted and deeply grateful to my mother, Song Lin, for her unconditional support, help and encouragement. Without her support, I could not study abroad in the first place and complete my master’s degree.

Abstract

The emergence of social media provides a new source of sensing society. In this thesis, a novel event detection method for detecting real-world events based on population sentiment orientation (PSO) from social media check-in data is proposed. The method is mainly composed of sentiment analysis, spatial-temporal analysis, and event extraction.

The hypothesis guiding this research is that social events change PSO in the dimension of time and space. The ratio of the number of positive and negative records is chosen to indicate the PSO within a specified period and geographical area. First, different sentiment classification methods are compared and a sentiment classification model is trained. Second, spatial-temporal analysis empowers the method to detect multi-scale events in terms of time and space dimension. Specifically, the method can detect events from nationwide festivals to local activities and from annual-scale to day-scale. Furthermore, time series analysis and spatial clustering can interpret the social phenomenon of the event by mining spatial-temporal patterns. At last, A Word Cloud is used to visualize the extracted high-frequency event keywords for visually identifying the event.

For testing the method, a case study is conducted using Sina Weibo data in Shanghai, China, 2014. Events such as Chinese New Year, Mid-Autumn Festival and concerts have been successfully detected. Besides, the interpretation ability of this method is also tested. Main reasons for negative microblogs on the New Year’s Eve are successfully extracted. Traffic as one of the biggest reasons from the result, the spatial pattern of it has been discovered. Related microblogs are mainly distributed at sites of public transportation like Shanghai Pudong international airport or railway station.

Keywords: Spatial-temporal analysis, Event detection, Social media, Sentiment analysis, Machine learning, Data mining.

Contents

List of figures

List of tables

1 Introduction

1.1 Background

Social media has become an indispensable part of many people's daily life. People share their thoughts, opinions, photos, etc. in social media with their friends, family or any other user, all the time. Kaplan and Haenlein gave a definition and discussed the challenges and opportunities of social media (Kaplan & Haenlein, 2010). They said social media is a group of internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of user generated content. By extensive use of social media, like Facebook, Twitter, Instagram, Flickr, networking and content sharing are unprecedented fast and massive. It has not only changed people's way of living but on the other hand, also provided a very valuable data source and a way of understanding this society. Gundecha and Liu introduced the basics of social media mining and illustrated representative research problems and some applications like community analysis, sentiment analysis and opinion mining, social recommendation, Influence modeling, information diffusion and provenance and privacy issues (Gundecha & Liu, 2012).

In this big data era, every individual becomes one social sensing sensor and in total, from that we can understand our socioeconomic environments better (Y. Liu et al., 2015). To understand society, many researchers have made a lot of efforts. Asur and Huberman successfully forecasted box-office revenues for movies using Twitter data (Asur & Huberman, 2010). Another interesting and popular topic in this field is sentiment analysis. Sentiment analysis and opinion mining are not new at all, it has been studied and applied in a wide range in almost every business and social domain (B. Liu, 2012). However, due to the new characteristics of social media, sentiment analysis faces new challenges and opportunities in terms of short text, informal language and rapid spread.

Most of social media websites provide a geo-tag option which allows the user to attach the spatial-temporal attribute – the location – as add-on to additional to the post. Goodchild invented the term "Volunteered geographical information (VGI)" for this kind of additional information (Goodchild, 2007). Liu used the term "check-in record" to denote this kind of geo-tagged content. VGI and check-in records have been used for several studies (B. Liu, 2012). Zheng and Zha discovered travel patterns from geotagged photos (Zheng, Zha, & Chua, 2012). Malik and Lamba investigated population bias in geotagged tweets (Malik, Lamba, Nakos, & Pfeffer, 2015).

VGI play an important role as social sensors. Many researchers trying to detect events from them. Sakaki and Okzaki for example did an event detection on earthquake from Twitter (Sakaki, Okazaki, & Matsuo, 2010). Chae and Thom detected an abnormal event using spatiotemporal social media (Chae et al., 2012). In this thesis, author tries to explore a method that combines sentiment analysis and spatial-temporal analysis to detect social event from social media data.

1.2 Problem statement and research identification

As mentioned before, researchers have explored a variety of ways in social event detection based on social media. However, only a few researchers have used the sentiment as an important feature and indicator of society condition and tried to detect real life events based on them. In this thesis, the author proposed a new indicator called *Population Sentiment Orientation (PSO)* and designed a new event detection framework based on this indicator and conducted many corresponding experiments to verify this method. In this study, we explore how social events like festival, concert or football match can be detected from social media. Combining sentiment analysis and spatial-temporal analysis, we propose a new approach of social event sensing.

The hypothesis guiding this research is that social events change population sentiment orientation in the dimension of time and space.

1.3 Research objectives and questions

The overall objective of this thesis is to develop a method for event detection based on population sentiment from geotagged social media data leading to an interpretation of the population sentiment in the dimension of time and space. The main objective consists of three sub-objectives and related research questions:

1) Find a suitable sentiment indicator to represent the population sentiment orientation.

2) Develop a methodology for spatial-temporal analysis of the population sentiment to detect sentiment fluctuation and abnormalities.

3) Find suitable methods to identify and interpret the event in the spatial-temporal dimensions.

RQ 1) Which sentiment classification method for text of social media posts is most suitable to find the appropriate indicator for representing the population sentiment orientation?

RQ 2) How should the spatial-temporal analysis be designed for population sentiments? Which specific methods or algorithms can be applied? What are the analysis dimensions?

RQ 3) How can events be detected after the spatial-temporal attributes have been determined? What method(s) is (are) suitable to interpret the event?

1.4 Structure of the thesis

This thesis is structured into 6 chapters. The background of this research including problem, objectives and questions is described in chapter 1. In chapter 2, the literature review concerning sentiment analysis and classification methods, Natural Language Processing methods, spatial-temporal analysis methods, event detection methods from social media, data clustering and aggregation methods is provided. Chapter 3 illustrates the methodology of this thesis and the principle of the whole approach. The implemented case study based on microblogs of Shanghai, China is described in chapter 4. Chapter 5 is a discussion of the method. Conclusions are drawn and future work is stated in chapter 6.

2 Literature review

This thesis mainly covers three major research topics: (1) Sentiment analysis of social media, based explicitly on text information; (2) Spatial-temporal analysis of location-based sentiment-tagged social media data to discover hidden patterns in the dimension of time and space; (3) Extraction and interpretation of social events from spatial-temporal patterns.

The following three parts in this chapter are providing the scientific background of these topics relevant for this thesis. The order follows the workflow of this thesis. The subchapter 2.1 is about sentiment analysis. The result of the sentiment analysis should be polarized social media data. The subchapter 2.2 is about spatial-temporal analysis, which is the analyzing process of polarized social media data. The subchapter 2.3 is about event detection which explains how an event can be detected in the spatial-temporal analysis.

2.1 Sentiment analysis

In sentiment analysis, computational and natural language processing (NLP) techniques are mainly combined to extract, identify, characterize, and categorize text information for determining the attitude and opinion of people. (Beigi, Hu, Maciejewski, & Liu, 2016). Usually, sentiment can be categorized as positive and negative polarity. It used mainly in online review posts, for instance, movie, book, and consumer product and has excellent use and potential in business (Taboada, 2016). Sentiment analysis can be applied in four levels: (1) sentence; (2) aspect, (3) document and (4) user level (Ahmed, El Tazi, & Hossny, 2015). Feature selection, data integration, data cleaning, and crowdsourcing are essential and indispensable procedures to improve the sentiment analysis result.

Figure 2.1 shows the sentiment analysis process. First step is sentiment identification and then select the features, after that is sentiment classification and finally output the sentiment polarity.

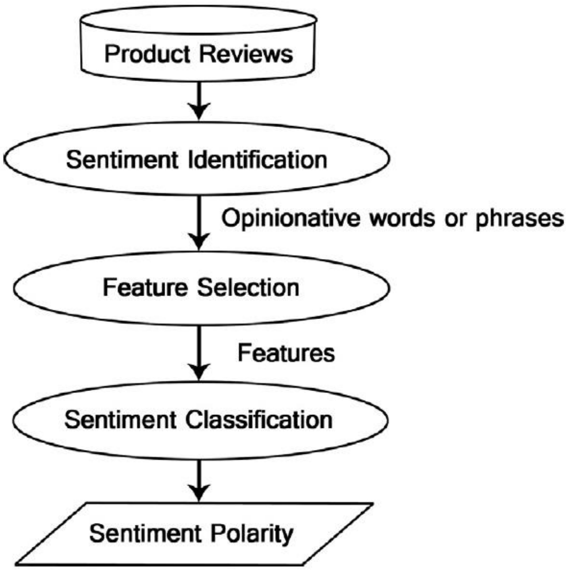


Figure 2.1: Sentiment analysis process (Medhat, Hassan, & Korashy, 2014)

Among all the sentiment analysis methods, it can be mainly divided into two categories, the lexicon-based method and machine learning-based method. Medhat summarizes the different techniques of sentiment classifications as shown in figure 2.2. The main two approach are machine learning approach and lexicon-based approach. Machine learning approach includes supervised learning and unsupervised learning. Some algorithms are shown in the figure. Dictionary-based and corpus-based approaches are the two mian approach of lexicon based method.

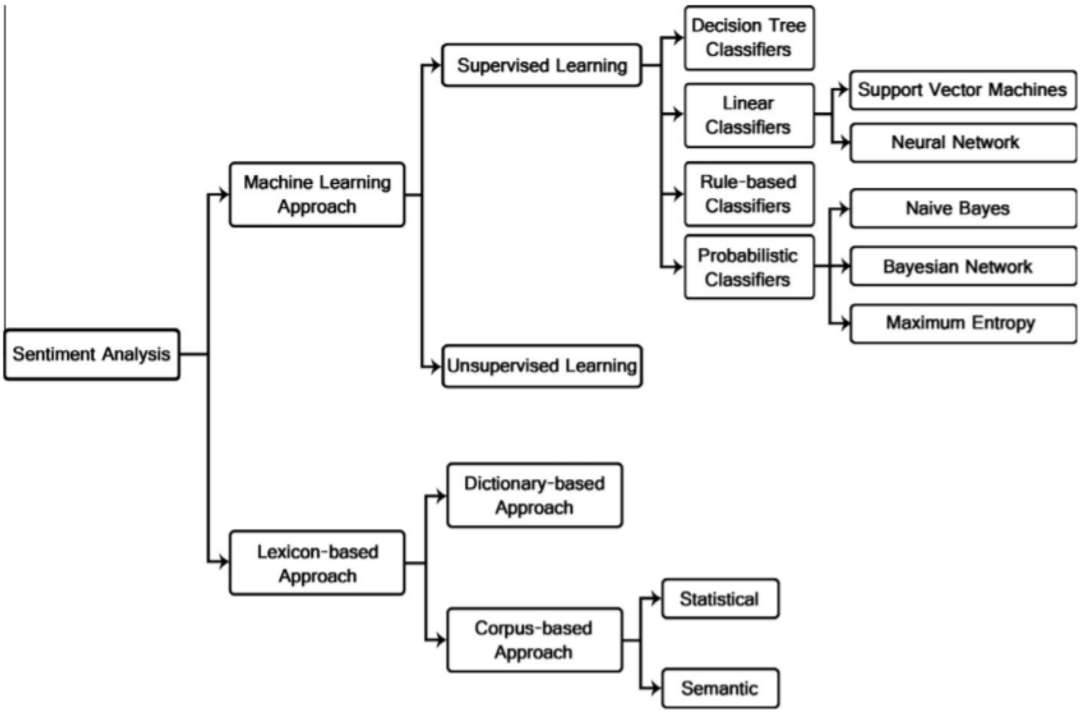
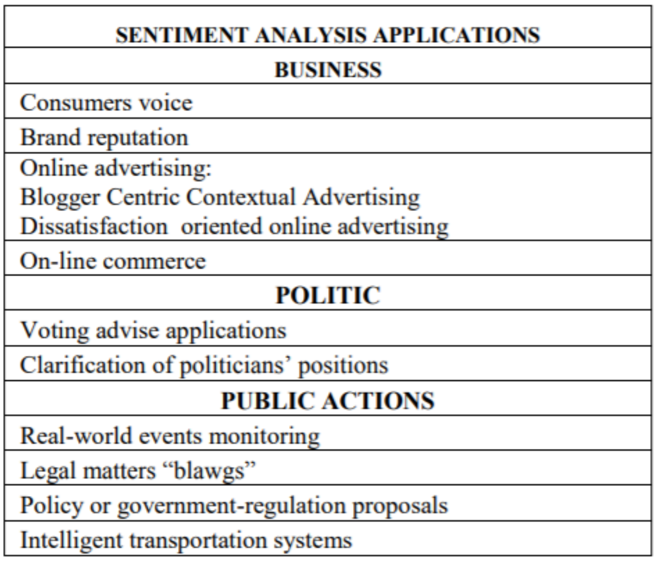


Figure 2.2: Sentiment classification techniques (Medhat et al., 2014)

Sentiment analysis has been applied in a variety of fields, for instance, business, politics, public actions, or finance. Figure 2.3 shows the main areas of applications.



SENTIMENT ANALYSIS APPLICATIONS
BUSINESS
Consumers voice
Brand reputation
Online advertising: Blogger Centric Contextual Advertising Dissatisfaction oriented online advertising
On-line commerce
POLITIC
Voting advise applications
Clarification of politicians' positions
PUBLIC ACTIONS
Real-world events monitoring
Legal matters "blawgs"
Policy or government-regulation proposals
Intelligent transportation systems

Figure 2.3: Sentiment analysis applications (Alessia, Ferri, Grifoni, & Guzzo, 2015)

Sentiment analysis jointly with NLP has also achieved great success in many fields such as, social networking, the ranking of best products, trend detection systems, public opinion monitoring, emotional management and sociology study (Montoyo, MartíNez-Barco, & Balahur, 2012).

Sentiment analysis on social media, especially on microblog text information is challenging, due to the short length status message, flexible words, spelling variation and emoticons, mixture of a different form of text (link, emoji). Pak and Paroubek experimented sentiment analysis on Twitter ((Pak & Paroubek, 2010). Ortigosa presented a new method for sentiment analysis on Facebook using an adaptive e-learning system and the accuracy reached to 83.27% ((Ortigosa, Martín, & Carro, 2014). The following paragraphs give detailed background knowledge of the two main methods: Lexicon-based and Machine learning-based method.

2.1.1 Lexicon-based method

Sentiment analysis is a study of text sentiment polarity classification in word, sentence, paragraph, and document level. The lexicon-based approach is based on the assumption that the sentiment orientation of text information is the sum of sentiment orientation of each word or phrase (Yadav & Elchuri, 2013). There are mainly two methods of creating a lexicon (Taboada, 2016), one is manually creating, and the other one is based on a seed of words automatically expanding (Qiu, Liu, Bu, & Chen, 2009). The types of lexicons are listed below.

Table 2-1: : Types of lexicons (Yadav & Elchuri, 2013)

Types of lexicons	
Sentiment Words	Positive and Negative sentiment words have a sentiment score of +1 or -1 to indicate the respective polarity
Negation Words	Negation words are the words which reverse the polarity of sentiment, normally appear before the sentiment word.
Blind Negation Words	Blind negation words operate at a sentence level and points out the absence or presence of some sense that is not desired in a product feature.
Split Words	Split words are the words used for splitting sentences into clauses. The split words list consists of conjunctions and punctuation marks.

Due to the characteristic of different languages, different types of lexicons are needed. Table 2-1 shows the lexicons in English. In the Chinese language instead, the most commonly used lexicons are sentimental words, stop words, degree words, negation words (Xianghua, Guo, Yanyan, & Zhiqiang, 2013). Each sentiment word in a sentence can form a sentiment unit, and the sentiment polarity of the whole sentence can be scored by the sum score of every sentiment unit in the sentence.

2.1.2 Machine learning-based method

With the rapid development of machine learning techniques, sentiment analysis is one of the popular topics in this field. Supervised learning method and the semi-supervised method are the two main method of machine learning-based method.

a) Supervised learning
Training classification model by standard sentiment labelled documents and supervised learning can be applied in sentiment analysis. Pang and Lee introduced the supervised classification into sentiment analysis for the first time (Pang, Lee, & Vaithyanathan, 2002). They used movie reviews as data and performed three machine learning methods, Naïve Bayes, Maximum entropy classification and support vector machine. The result is much better than human-produced baselines. Based on this research, Dong and Wei proposed adaptive recursive Neural Network (AdaRNN) for sentiment classification using Twitter data as an example in 2014 (L. Dong et al., 2014). With the development of deep learning, Tang and Wei developed a deep learning system (called Coooolll) to do sentiment analysis on Twitter data (Tang, Wei, Qin, Liu, & Zhou, 2014).

b) Semi-supervised learning
The increase in the amount of data is significant rapid and more researchers started combining the advantages of the lexicon-rules method and supervised learning. Training the classification model with a small amount of labelled data and a large amount of unlabeled data can decrease labor and time cost. Tan and Cheng proposed a method that mad the maximum use of both the old-domain data and the unlabeled new-domain data to address the problem that when transferring the supervised sentiment classifier to another domain the classifier often performs not so well (Tan, Cheng, Wang, & Xu, 2009). They proposed the Adapted Naïve Bayes Transfer Classifier (NTBC) which improved the result a lot. Zhou and Chen proposed a two-step semi-supervised learning method called fuzzy deep belief networks (DBN) for sentiment classification and tested on

movie reviews and DVD reviews. The accuracy reached to 79.4% (S. Zhou, Chen, & Wang, 2014).

2.2 Spatial-temporal analysis

After sentiment analysis, spatial-temporal analysis is needed for further processing and analyzing polarized social media data. Looking back into Cartography and Geographic Information Science (GIS) history, spatial-temporal analysis is nothing new. However, spatial-temporal data analysis is fast developed and widely applied due to fast developing novel computational techniques and fast growing data volume and increasing data complexity (Meliker & Sloan, 2011). Data with spatial and temporal property is normally referred to as spatial-temporal data. The most commonly used temporal property is the timestamp, and spatial property is often geographic coordinates. The spatial-temporal analysis is widely used in discovering hidden patterns inside large chaotic data. Turner, for example, presented a grid cell based spatial analysis program (SPAN) to observe landscape patterns (Turner, 1990). Day and Britsch evaluated the relationship between direct wetland loss because of canal dredging (Day et al., 2000). Liu did spatial-temporal analysis on Landsat satellite imagery to improve the estimation of cropland area (J. Liu et al., 2005). Meliker and Soloan stated that the two primary goals of spatial-temporal analysis are prediction and description. They gave a spatiotemporal data analysis workflow which is shown in figure 2.4. The main components of the workflow are: (1) Collect and prepare data; (2) Map and Examine; (3) Pre-process; (4) Define and model spatial structure; (5) Evaluate model; (6) Utilize results (Meliker & Sloan, 2011).

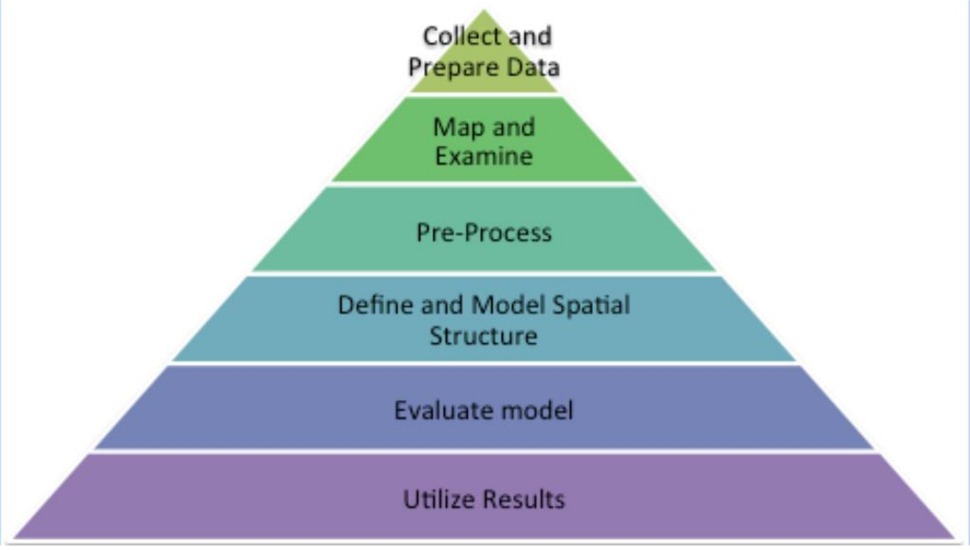


Figure 2.4: Spatial-temporal analysis workflow (Meliker & Sloan, 2011)

In spatial-temporal analysis, time series analysis is the commonly used method for temporal analysis and clustering is the most common geographic spatial analysis method in terms of pattern discovery. These methods are respectively explained in the following subchapters.

2.2.1 Time series analysis

Time series analysis is using statistical methods to discover hidden meaningful knowledge inside a series of time-indexed data. According to Box and Jenkins, the most critical research topics and applications are (Box, Jenkins, Reinsel, & Ljung, 2015):

.

- Based on current and past time series time to predict the values in the future.
- The use of indicator input to determine and assess the effects of unusual intervention events on the behavior of a time series.
- The examination of interrelationships among the time series and determination of appropriate models to represent these relations.

The essential research problems are:

- Forecasting time series
- Estimation of transfer functions
- Analysis of effects of unusual intervention events to a system
- Analysis of multivariate time series
- Discrete control systems.

Figure 2.5 shows a time series example of yearly average global temperature (Wei, 2006).

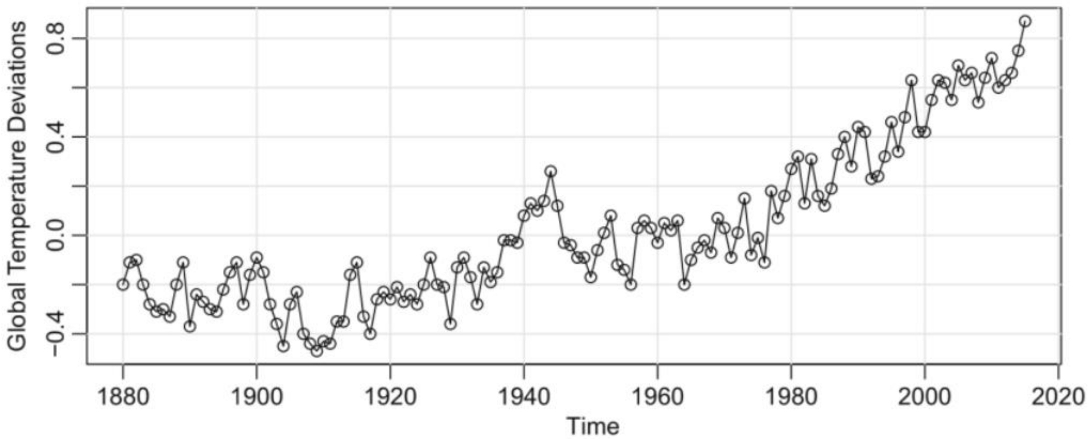


Figure 2.5: Yearly average global temperature (Wei, 2006)

In this thesis, the time series analysis is needed for determining the unusual intervention in order to detect an unusual event.

2.2.2 Cluster analysis

Cluster analysis can categorized as one of the pattern recognition techniques and may be characterized by the use of resemblance or dissemblance measures between the objects to be identified (Diday & Simon, 1976). It is also a task of data mining, and widely used in machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. Main algorithms categories are hierarchical clustering, centroid-based clustering, distribution-based clustering and density-based clustering (Bhui, Everitt, & Jones, 2014). After spatial clustering, the objects within a cluster have high similarity in comparison to one another but are dissimilar to objects in other clusters. Many sophisticated cluster algorithms have been built in a business statistics software system such as S-Plus, SPSS, and SAS. In the field of machine learning, clustering can be seen as a type of unsupervised learning because it does not requires predefined class and no labelled training dataset is needed (Han, Kamber, & Tung, 2001).

Figure 2.6 shows the illustration of one of the most well-known clustering algorithm, the K-means algorithm. Given a representation of n objects, find K groups based on a measure of similarity. The similarities between objects in the same group are higher than the similarities between objects in different groups.

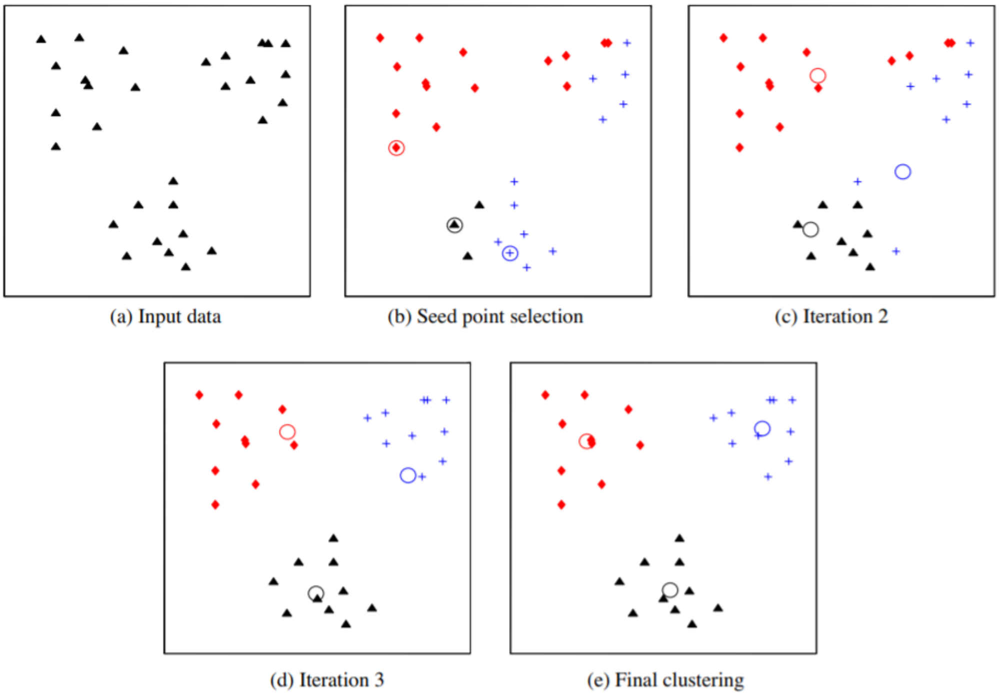


Figure 2.6: Illustration of K-means algorithm (Jain, 2010)

Another clustering algorithm called DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is much used in spatial analysis. DBSCAN first proposed by Ester and Kriegel and is a non-parametric, density-based clustering technique and performs significantly useful in discovering clusters of arbitrary shape (Ester, Kriegel, Sander, & Xu, 1996). Figure 2.7 shows a graphic representation of the DBSCAN algorithm. Wu and Leahy proposed an optimal graph theoretic clustering approach to accomplish image segmentation (Wu & Leahy, 1993). Cuzick and Edwards proposed a method for detecting spatial clustering of an event in populations with non-uniform density (Cuzick & Edwards, 1990). DBSCAN is also used in many applications, such as astronomy, biology, earth science and geography (Sander, Ester, Kriegel, & Xu, 1998).

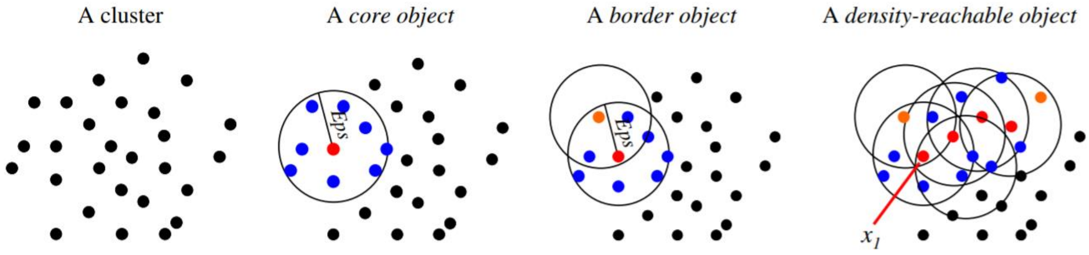


Figure 2.7 Graphical presentation of DBSCAN (Tran, Drab, & Daszykowski, 2013)

2.3 Event extraction

In this subchapter, the event extraction is specially referred to as event extraction in social media. With the rapid development of social media, such as Twitter, Facebook, Instagram, Sina Weibo, etc. people can conveniently express their opinion, and attitude on events and the rapid spread characteristic of social media is incomparable with traditional media. According to a report on Twitter, they have over 300 million users and generating over 300 million tweets and handling over 1.6 billion queries per day (X. Zhou & Chen, 2014). Statistically, it has great potential in monitoring population opinion trend and detects a social event. The study of it has many possible applications, such as crisis management, decision making, product sales, an opinion survey, etc.

The most commonly used model in discovering topic is Latent Dirichlet Allocation (LDA) model. In natural language processing, LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar (Girolami & Kabán, 2003).Figure 2.8 shows the graphic of its mechanism.

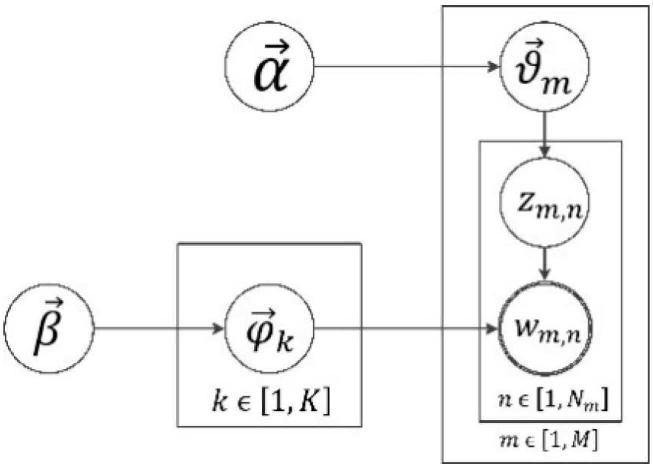


Figure 2.7: Graphical model of Latent Dirichlet Allocation (LDA)

Generally, based on specific context or task, the definition of the *event* might be different, and the workflow and method are different as well. For example, Zhao and Mitra defined the event as a set of relations between social users on a specific topic over a time period, so the events are extracted by three steps, the text-based clustering, the temporal segmentation, and the graph cuts of social networks (Zhao & Mitra, 2007). In another case, Yao and Cui proposed to extract bursts from multiple social media sources using a state-based model (Yao, Cui, Huang, & Jin, 2010).

Unlike these, this thesis is proposing a method how to sense and monitor the society by population sentiment orientation and detect social events by spatial-temporal analysis.

3 Methodology

This chapter elaborates details of methods that adopted in this thesis based on the background, literature review and state of the art from the former two chapters. To be more specific, the first part of the workflow is sentiment analysis using *Natural Language Processing* techniques. The result of the sentiment analysis should be polarized data set by positive and negative sentiment. Preprocessing methods and the comparison of different classification methods and results are discussed in this part. The following part is the event detection based on the spatial-temporal analysis. First, the structure and mechanism design of the spatial-temporal analysis is described. Second, the pipeline, algorithms and details of the spatial-temporal analysis are illustrated as well as the technique selection and example presentation. Figure 3.1 shows the workflow chart of this thesis.

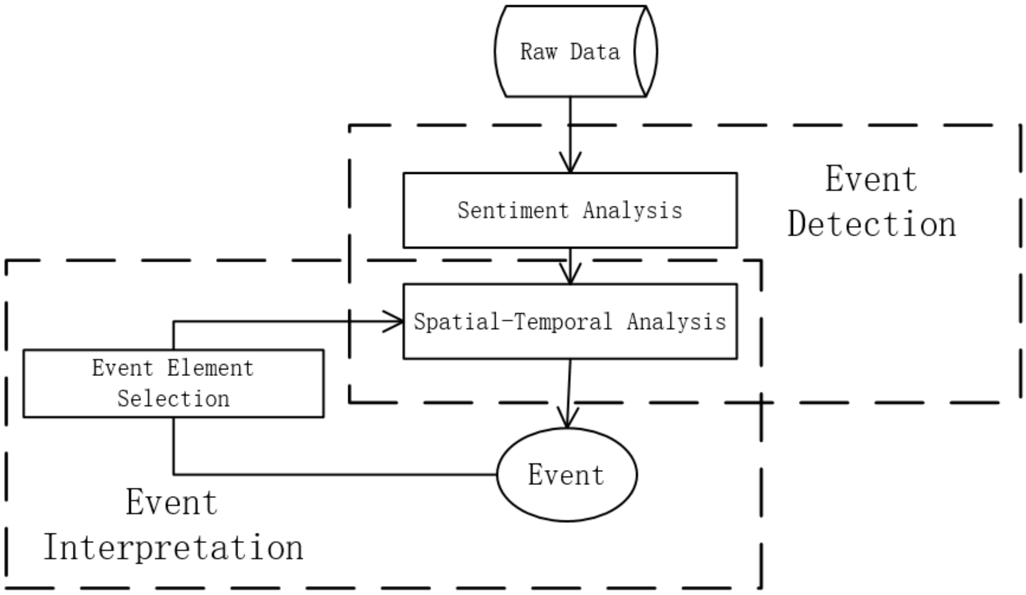


Figure 3.1: Thesis workflow chart

3.1 Data preprocessing

In most cases, data selection is always the first step because raw data regularly cannot meet the need. Data preprocessing is the first and vitally important step. Well performed data preprocessing can lead to a good result and vice versa (Rahm & Do, 2000).

3.1.1 Data selection

Raw data can be a mess because of many possible reasons. For instance, loss of some attributes contains invalid values or duplicate records. Data selection can significantly reduce the data redundancy in the first place.

Data attributes selection

One record of social media contains many types of information, not only the text but also many people like to attach a photo, a video, location or a link. According to Twitter official developer document (Twitter, 2019), a tweet object contains attributes such as “*created_*at”, “*id*”, “*id_str*”, “*text*”, “*source*”, “*truncated*”, “*user*”, “*coordinates*”, “*place*”, “*entities*”, “*retweeted*”, “*lang*” and etc. For the further use of this thesis, the valuable attributes are “*user*”, “*id*”, “*created_*at”, “*text*”, “*coordinates*”. All the other attributes can be dropped in order to reduce the redundancy of the dataset. The attributes table after selection should look like summarized in table 3-1.

Table 3-1: Data attributes after selection

Attribute	Type	Description
Created_at	String	UTC time when this Tweet was created.
id	Int64	The integer representation of the unique identifier for this Tweet.
user	User Object	The user who posted this Tweet.
text	String	The actual UTF-8 text of the status update.
Coordinates	Coordinates	Nullable. Represents the geographic location of this Tweet as reported by the user or client application. The inner coordinates array is formatted as geoJSON (longitude first, then latitude).

Data range selection

After data attributes selection, data can be further selected by the attributes left based on the study purpose. For example, specific region selection based on the coordinates, specified period selection based on the time attribute.

3.1.2 Data cleaning

Integrity check

After data selection, the data is still not “clean” and not ready to process. Integrity check means deleting all the invalid data. First, a check if the data type is correct and consistent is performed. Second, a check if there are empty records is done. Finally, a check if there are outliers is needed. For example, check if the id column is int64 type, check if there are records that the text attribute of is empty. Delete the invalid data or modify them into correct.

Deduplication

Due to some technical reason or human-made disoperation, in social media dataset duplicate values often occur. Duplication increase the data redundancy and much of this can cause statistical bias and affect experiment results.

3.1.3 Text cleaning

The text attribute is the essential information for further process (sentiment analysis) in this thesis. According to Unicode standard (Allen et al., 2012), the Unicode standard contains a set of unified Han ideographic characters used in the written Chinese, Japanese, and Korean languages. This thesis is based on Chinese Sina Weibo microblog, so the text language should be in Chinese. All other languages should be filtered. Chinese is in UTF-8 range 4E00-9FA5. By using the regular expression, all the link, emoji, tag, punctuation, white space, digits should be removed. The data types need to be removed are shown in Table 3-2.

Table 3-2: The data types need to be removed

Data type	Example
Link	http://abcde/efg
Emoji	😊 (U+1F60A)
Punctuation	! , . ? ~
White space	“ ”

Digits	12345
Tag	#News#

3.2 Sentiment analysis

The goal of the sentiment analysis is to get a sentiment polarized dataset by sentiment classification. As summarized in chapter 2, there are mainly two classification categories: lexicon-based method and machine learning-based method. In this thesis, both of the methods are conducted and compared in the case study (see chapter 4.3). The best-performed algorithm and model will be finally chosen for the sentiment classification task, and the result will be the input for the spatial-temporal analysis performed later on. The workflow of the sentiment analysis is shown in Figure 3.2. The chapter 3.2.1 explains the left part of the figure and the chapter 3.2.2 explains the right part of the figure. This figure guides the workflow of the sentient analysis in this thesis.

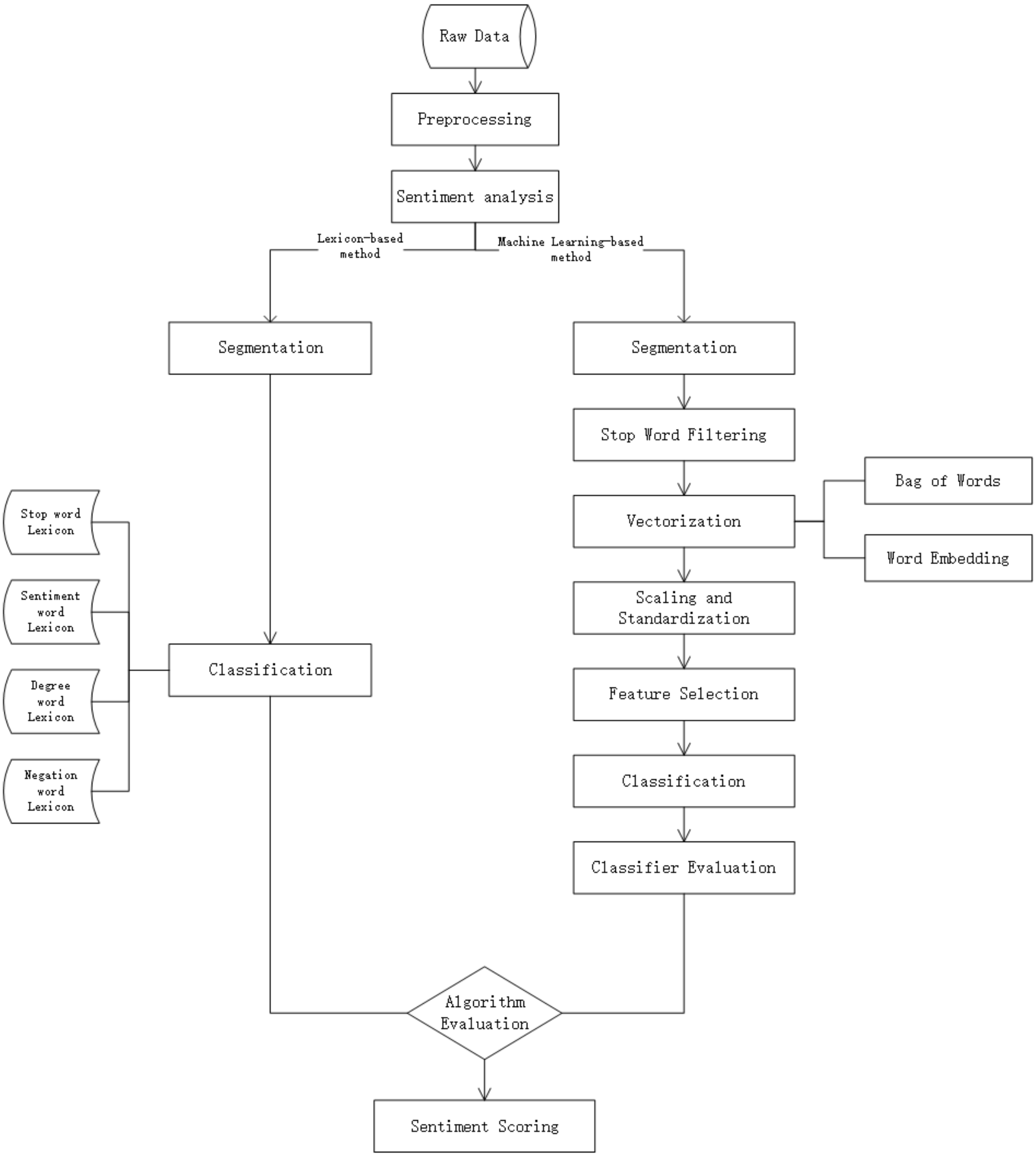


Figure 3.2: Workflow chart of sentiment analysis

3.2.1 Lexicon-based method

As mentioned in the second chapter, the lexicon-based sentiment analysis method is building lexicons based on the different components of a sentence in order to classify sentences. This study focuses on Chinese language, so the lexicons are categorized as sentiment *word lexicon*, *degree word lexicon*, *stop word lexicon*, and *negation words lexicon*. Each sentiment word can locate and form a sentiment unit and each sentiment unit must have one sentiment word, may or may not have degree word or negation word which can change the sentiment level and polarity. The final score of the sentence is the sum of the score of every unit, and it mathematically shows in Equation 1:

$$\mathrm{P(T)} = \sum_{i=1}^{n} \mathrm{P}(\mathrm{U}_i) \qquad (1)$$

$\mathrm{P}(\mathrm{U}_i)$ means the sentiment score of the sentiment unit. P(T) is the sum score of every sentiment unit.

The degree word lexicon is from HowNet, a Chinese knowledge base (Z. Dong, Dong, & Hao, 2010). It contains 219 degree words. The 6 degree level weights are rearranged to (0.5, 0.7, 1.3, 1.5, 1.7, 2) to fit this thesis because the original weight range is too broad. The degree factor γ for the *i*th sentimental word ω_i is defined in Equation (2):

$$\Upsilon(\omega_i) = \prod_{k=1}^{M} d_{ki} \qquad (2)$$

The d_{ki} is the weight of the *k*th degree word for *i*th sentimental word.

Negation word reverses the sentiment polarity of the sentiment unit entirely. There are no official, or standard negation word lexicon available. In this thesis, the most frequently used 28 negation words are used as the negation lexicon. The impact factor of negation words τ on the sentiment words ω_i is defined as Equation 3:

$$\tau(\omega_i) = (-1)^n \qquad (3)$$

The *n* is the number of negation words in the *i*th sentiment unit. All the factors above form a sentiment unit and the total score of it ($\mathrm{P}(\mathrm{U}_i)$) can be calculated as Equation 4:

$$\mathrm{P}(\mathrm{U}_i) = P(\omega_i) \times \Upsilon(\omega_i) \times \tau(\omega_i) \qquad (4)$$

The P(ω_i) is the sentiment value of the *i*th sentiment word. The sentiment word lexicon used in this thesis is provided by Dalian University of Technology (Yu & Jianmei, 2008). After polar categorization, the lexicon contains 11267 positive words and 10797 negative words. Each positive word has a score of 1, and each negative word has a score of -1.

3.2.2 Machine Learning-based method

Machine learning–based method is the other commonly used method for sentiment analysis. This thesis is Chinese language oriented, and the workflow of the method is shown in figure 3.2. The main components of the workflow: (1) Segmentation; (2) stop word filtering; (3) Vectorization; (4) Scaling and standardization; (5) Features selection; (6) Classification; (7) Classifiers evaluation.

Segmentation

In Chinese NLP Word segmentation is always the crucial first step because Chinese words can be composed of multiple characters but with no space appearing between words, which is different from languages like English has natural spacing. In this thesis, a Python Chinese word segmentation module called “jieba” (Chinese for “to stutter”) is used for word segmentation (jieba, 2012). The key words extraction of this algorithm is based on Term Frequency-Inverse Document Frequency (TF-IDF) or Text Rank. It also provides part of speech tagging function.

Stop word filtering

This step is optional and depends on the purpose and the following processing method of the task. The aim of stop word filtering are: (1) Reduce data redundancy; (2) Prevent the central meaning of the sentence from meaningless stop words. However, this is not always the best choice. For instance, Word2Vec is based on the contextual syntactic relationship, so filtering stop words will ruin the contextual syntactic relationship and leads to an unsatisfactory result. This will be discussed more in this following subchapter. The stop word dictionary used in this thesis is published by the Information Retrieval Laboratory of Harbin Institute of Technology. It contains 767 stop words, such as “the”, “a”, “an”, “in”.

Vectorization

In machine learning algorithms, input should be numeric feature vectors so that the computer can understand. Thus when working with text documents, it needs to be converted into a numeric vector. This process is known as text vectorization. In this thesis, Word2vec model is chosen and used for vectorization. Word2vec model is a group of related models that are used to produce word embedding, and it was created and published in 2013 by a team of researchers led by Tomas Mikolov at Google (Mikolov, Chen, Corrado, & Dean, 2013). There are two models in Word2vec, CBOW and skip grams. In general, the models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.

Scaling and standardization

In machine learning, the aim of scaling and standardization is to transfer data range to [0, 1]. The meaning and advantages of this process are: (1) it speeds up the convergence of gradient descent; (2) it avoids the different dimension of features issue and makes the features comparable.

Feature selection

Not all the features are equally important and not every feature can represent the data very well, so the feature selection step may be needed. Filter method, wrapper method and embedded method are commonly used in the feature selection step. Dimensionality reduction can also be seen as a type of feature selection. Because in many cases, the dimension of the features can be significantly large, and that causes the calculation to be too expensive so dimensionality reduction can help. Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Pearson, 1901). PCA is one of the most well-known dimensionality reduction methods.

Classification

In machine learning algorithms, classification is the task that identifying which class a new observation belongs to, based on a training dataset (Alpaydin, 2009). Classification algorithms mainly composed of two types, supervised classification and unsupervised classification. Unsupervised classification is usually known as clustering. Some classic classification algorithms that perform in this study are shown below.

(1) Logistic Regression

Logistic regression also called sigmoid function which takes any real input t, and outputs a value between zero and one. Logistic regression can predicts the probability of an outcome that can only have two values. So it mostly solves binary classification problem. The equation can be written as Equation 5:

$$\mathrm{P}(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{5}$$

P(x) is the probability that the dependent variable equals a case, given some linear combination of the predictors. β_0 is the intercept from the linear regression equation. $\beta_1 x$ is the regression coefficient multiplied by the coefficient.

(2) K Nearest Neighbors Classifier

K nearest neighbors (KNN) is a algorithm that stores all available cases and classifies new cases based on a similarity measure (Cover & Hart, 1967). KNN has been used in statistical estimation and pattern recognition already at the beginning of 1970s as a non-parametric technique. There are some different distance functions for different usage such as Euclidean distance, Manhattan distance, Minkowski distance, etc.

(3) Support Vector Machine

Support Vector Machine (SVM) is an outstanding classifier formally defined by a separating hyperplane. In other words, given labelled training data, the algorithm outputs an optimal hyperplane which categorizes new examples (Scholkopf & Smola, 2001). There are linear SVM and non-linear classification depends on different types of kernels. In this study, SVM with a linear kernel is used. In linear SVM, the hyperplane can be written as the set of point $\vec{x}$ satisfying Equation 6:

$$\vec{w} \cdot \vec{x} - b = 0 \tag{6}$$

Where $\vec{w}$ is the normal vector to the hyperplane. The parameter $\frac{b}{||\vec{w}||}$ determines the offset of the hyperplane from the origin along the normal vector $\vec{w}$.

(4) *Naive Bayes classifier*

Naïve Bayes classifier is based on Bayes theorem with strong independence assumption between the features. Naïve Bayes is a conditional probability model (Hand & Yu, 2001). When facing continuous data, in this theory the assumption is that the continuous values associated with each class are distributed according to the normal distribution also known as Gaussian distribution. In this study, Gaussian naïve Bayes is used for classification. The probability distribution can be written as Equation 7:

$$\mathrm{p(x = v}|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \tag{7}$$

where x is the continuous attribute, μ_k is the mean of the values in x associated with class C_k. σ_k^2 is the Bessel corrected variance of the values in x. v is the observation value.

(4) *Random Forest Classifier*

Random forest consists of a large number of individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction, and the class with the most votes becomes the prediction of the model and it correct for decision tree's habit of overfitting to their training set (Ho, 1995). The prediction for the unseen sample x' can be either made by averaging of the predictions from all the individual decision tree or by taking the majority vote. The Equation of the former can be written as Equation 8:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{8}$$

By applying the general technique of bootstrap aggregating, B is the times of bagging. b=1 … B. f_b is the classification tree on the bth training dataset.

Classifiers evaluation

After performing the classification methods, an evaluation is needed to select the best classifier of the given data set. There are several evaluation metrics exist in the field of machine learning. Accuracy and F1 score are used for evaluation in this thesis. Precision, recall, and F1 score are written in Equation 9-11:

$$\text{Precision} = \frac{t_p}{t_p + f_p} \tag{9}$$

$$\text{Recall} = \frac{t_p}{t_p + f_n} \tag{10}$$

$$\text{F1 Score} = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

Where t_p is the number of correctly classified positive samples, f_p is the number of incorrectly classified positive samples, and f_n is the number of incorrectly classified negative samples (Goutte & Gaussier, 2005). F1 score is the harmonic average and trade-off of the precision and recall. The higher score of F1 means the better the result.

Once the best classifier is chosen, all the labelled dataset will be used for training the classification model and apply the model on the whole experimental data. The entire experimental data should be classified by the model into positive and negative sentiment dataset.

3.2.3 Sentiment analysis indicator

In order to further analysis sentiment changes and perform spatial-temporal analysis, an indicator as an entry point is necessary. In this thesis, the ratio of the number of positive and negative records to indicate the population sentiment orientation within a certain period of time and specified geographical area is used. The author names this indicator *Population Sentiment Orientation (PSO)*. The equation is written as Equation 12:

$$\mathrm{PSO} = \frac{P}{N} \tag{12}$$

Where p is number of positive records and N is the number of negative records.

3.3 Spatial-temporal analysis

In this thesis, sentiment information can be seen as thematic information, and the goal of the thesis is to detect a social event by abnormal spatial and temporal distribution pattern based on the hypothesis that social events change population sentiment orientation (PSO). By given the sentiment indicator, spatial-temporal analysis is the next step of sentiment analysis. It can be inferred that, in the time dimension, social events can be determined by the global or local maxima or minima of the time series of PSO. For verifying the hypothesis and implementing the proposed method, a time series analysis is conducted in this study.

3.3.1 Time series analysis

A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time (Hamilton, 1994). In time series plot of this study, there is no doubt that Y-axis of the time series is the population sentiment orientation value. The X-axis unit can vary depending on the characteristics of the social event to be detected. The X-axis of the default regular time series is every single day from 00:00 to 24:00 but it can be different, for instance, if analyzing PSO variation in different periods of a day, the X-axis should be 00:00 to 24:00. Figure 3.3 shows a wind speed time series plot example (Huang et al., 1998). In this thesis, the maxima and minima help to detect the event, and from that, the time of the event can also be seen and determined. One step further, run the time series analysis on the date of this event, the specific time period when the event happened, and the trend of PSO can also be determined.

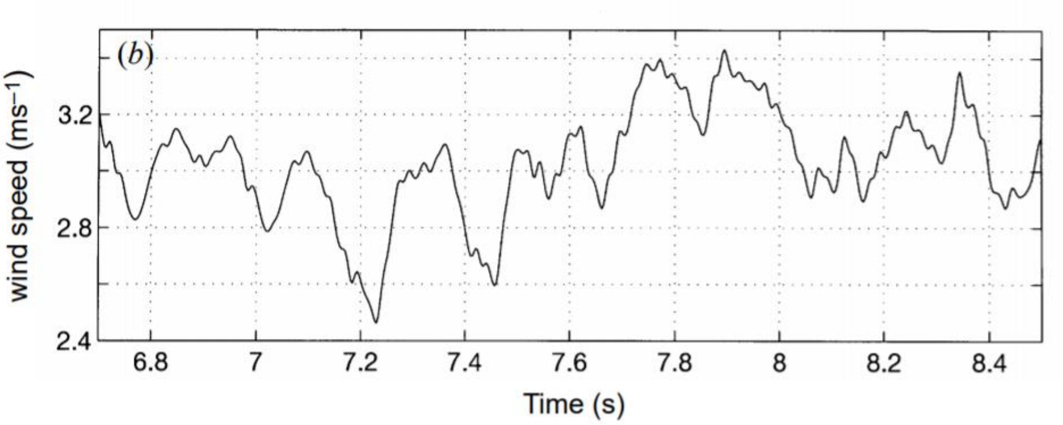


Figure 3.3: Wind speed time series plot example

3.3.2 Cluster analysis

In geography, cluster analysis is a very powerful method and tool to discover geographical patterns. In this thesis, cluster analysis is used for detecting the location and location pattern of the social event and to allow further interpretation based on the pattern. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is chosen in this thesis.

DBSCAN

DBSCAN groups points that are close to each other based on a distance measurement and a minimum number of points. The haversine formula is chosen to be the distance measurement here because the experimental data tagged with geographic coordinates.

The haversine formula

The haversine formula calculates the great-circle distance between two points on a sphere, in this case, the Earth, given their longitudes and latitudes. The formula was published by (Van Brummelen, 2012). It is a particular case of a more general formula in spherical trigonometry. It can be written as Equation 13:

$$d = 2r\ \arcsin(\sqrt{\sin\left(\frac{\varphi_2-\varphi_1}{2}\right)^2 + \cos(\varphi_1)\cos(\varphi_2)\sin\left(\frac{\lambda_2-\lambda_1}{2}\right)^2}) \quad (13)$$

where d is the distance between the two points along a great circle of the sphere. r is the radius of the sphere. $\varphi_1,\ \varphi_2$ are the latitude of the point 1 and point 2. λ_1, λ_2 are the longitude of point 1 and point 2 (all in radians).

DBSCAN clustering algorithm requires two parameters, *eps* and *minPoints*. *Eps* specifies how close points should be to each other to be considered a part of a cluster. *Minpoints* is the minimum number of points to form a dense region. There are three types of points in this algorithm, core points, density reachable points and outliers. Core points are that at least *Minpoints* points are within distance *eps* of it. Density reachable points are that within the *eps* distance of core points. Outliers are points that are not reachable. Figure 3.4 is an illustration of DBSCAN (Chire, 2011), minPt is 4, and red points are core points, yellow points are density reachable points and blue point is outlier.

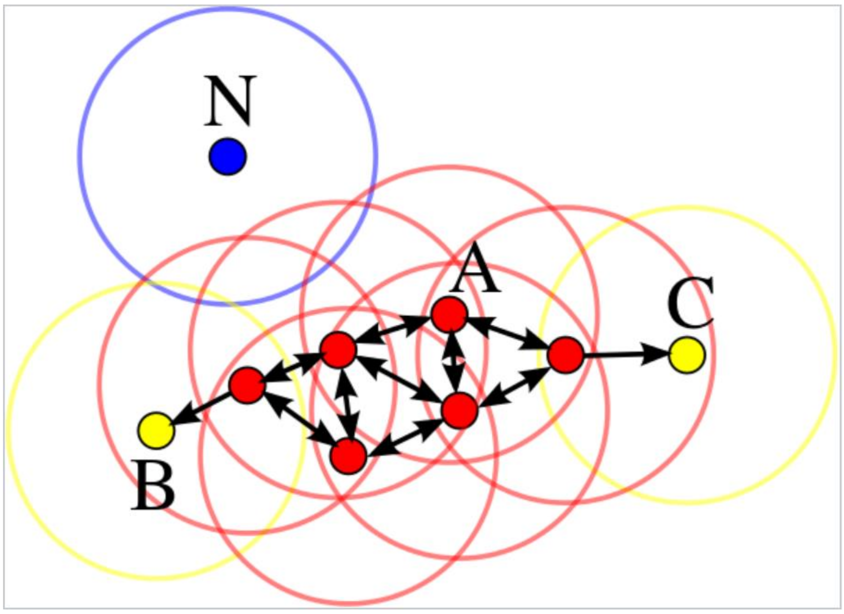


Figure 3.4: DBSCAN illustration

Applying DBSCAN in the spatial-temporal analysis is on the one hand in order to discover the geographical distribution pattern of the social event, on the other hand, is to interpret the event from the pattern.

Visualization

Geographical information visualization is needed in spatial analysis. Precise base map, rich Point of Interest (POI) and harmonious symbolization and visualization conduce to visual analysis. In this thesis, Carto (formerly CartoDB) is used for visualization. Carto is a software as a service could computing platform that provides Geographical Information System (GIS), web mapping, and spatial data science tools (Carto, 2011,). Figure 3.5 shows an example of Carto map.

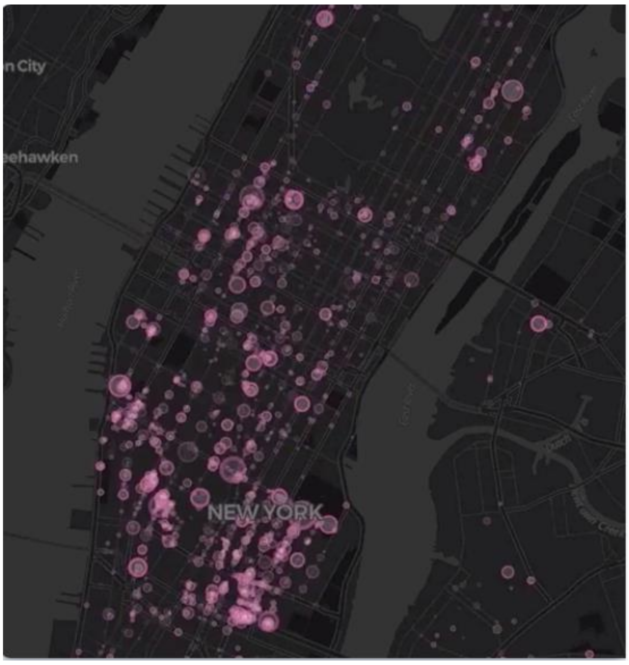


Figure 3.5: Carto map example

3.4 Event extraction

The event detection of this thesis is based on the hypothesis that if there is an event happening, the name of the event and related things shall appear in high frequency in social media text information. So the event extraction part is last but essential. So the corresponding event extraction method applied in this thesis is structured as follow:

1. Extract keywords from each sentence based on the TF-IDF algorithm.
2. Count and store the keywords and the frequency of them.
3. Generate a *Word Could* of the keywords based on the frequency for visualization.

Term frequency-inverse document frequency (TF-IDF) is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Hand, 2006). It weighs a keyword in any content and assigns the importance to that keyword based on the number of times it appears in the document. Jieba Chinese text segmentation Python library (jieba, 2012) provides TF-IDF keywords extraction while segmenting sentences. Besides, due to the fact that event name and related words most likely are verbs and nouns, so filtering of the part of speech is beneficial here. Only the verbs and nouns are shown on the *Word Cloud*.

Generating *Word Cloud* is achieved by a Python library called word cloud created by Andreas Mueller (Andreas Mueller, 2019). This library provides different visualization styles and different implementation methods. In this thesis, the word cloud is created from the keywords frequency dictionary. Figure 3.6 is an example of *Word Cloud* visualization. The size of the word represents its frequency and different colors have no meaning.

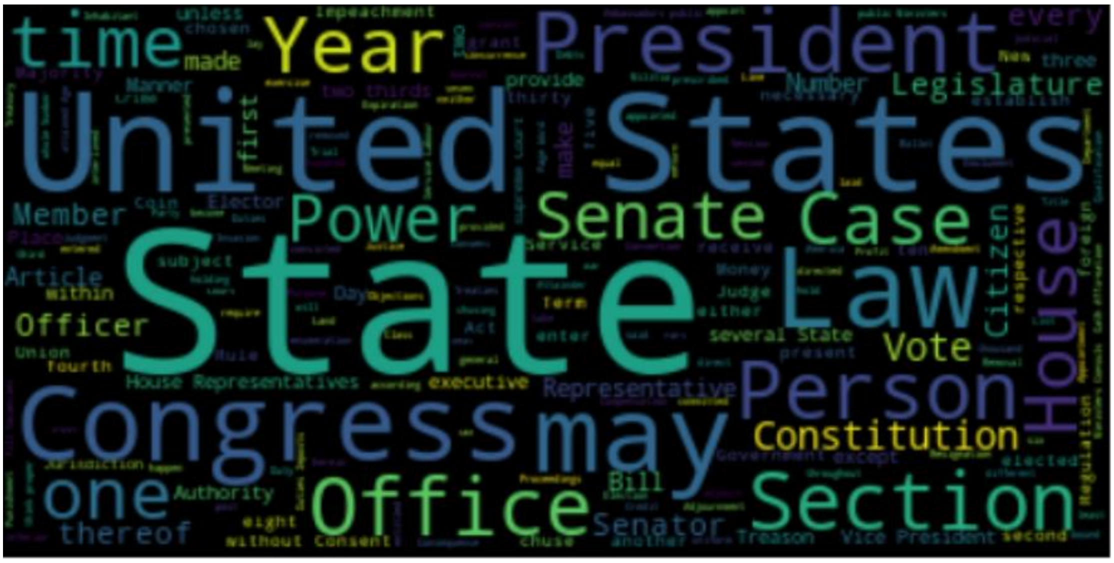


Figure 3.6: Word Cloud visualization example

 In summary, this chapter explains the method and the workflow of this thesis in details. Specifically, the process of the method is broken down into data preprocessing, sentiment analysis, spatial-temporal analysis and event detection method. In the next chapter, the hypothesis of this thesis will be tested. Verifying this method will be done by demonstrating the workflow through a case study in Shanghai, China.

4 Social events detection for the city of Shanghai

In this chapter, the experiment is conducted to test the hypotheses and the workflow of the social event detection method proposed in chapter 3. Specifically, the results will be presented for the case study of social event detection in Shanghai, China, of 2014. This chapter goes into details of how the methods, including sentiment analysis, spatial-temporal analysis, and event detection method is implemented in practice.

4.1 Case study data description

The data source is Sina Weibo (Chinese microblogging website), and the Weibo data used in this experiment is provided by Michael Jendryke (2015), Wuhan University (China). The raw data is about 4GB and stored in a Comma-separated values (CSV) file. Each row in the CSV file represents one record of the data, and the attributes columns are separated by commas. The dataset is stored in Unicode Transformation Format (UTF-8) encoding system format, and it has 28 columns and 11794009 rows. Table 4-1 shows an example of a record.

Table 4-1: Records example

Column name	Meaning	Value
"_id"	Entry ID	ObjectId("57f8fe82e040a8da7a33a836")
"idNearByTimeLine"	ID near by timeline	8
"createdAT"	Create time of this ID	"2014-07-17 04:06:34"
"createdATUnixTime"	Create Unix time of this ID	1405569994
"msgID" :	Message ID	NumberLong("3154515945889755")
"msgmid" :	Message ID	NumberLong("3154515945889755")
"msgtext"	The text content	"今天好热啊，都晒黑了呢！http://t.cn/RPzdvNZ"
"msgin_reply_to_status_id"	Reply number of this message	0
"msgin_reply_to_user_id"	The ID of the reply user	0
"msgin_reply_to_screen_name"	The name of the reply user	""

"msgfavorited"	The number of like	0
"msgsource"	Message source	""
"geoTYPE"	Geographical type	"Point"
"distance"	distance	6800
"userID"	User ID	NumberLong("5217853530")
"userscreen_name"	User Name	"XIA 小妍"
"userprovince"	User province ID	44
"usercity"	User city ID	3
"userlocation"	User location	"广"
"userdescription"	User description	""
"userfollowers_count"	Follower number of the user	44
"userfriends_count"	Friends number of the user	86
"userstatuses_count"	User statuses number	52
"userfavourites_count"	User favorites number	2
"usercreated_at"	Create time of the user	"2014-07-15 05:51:57"
"usergeo_enabled"	If user location enabled	1
"userverified"	If the user verified	0
"userbi_followers_count"	Follower number	16

4.2 Data preprocessing

As stated in chapter 3, only “created_at”, “id”, “user”, “text”, “coordinates” attributes will be used in this study. After integrity and duplication check, 11376875 records are valid and can be used in the experiment later. Specified about integrity check, for instance, this dataset is from Shanghai city, so the “WGS84Latitude” should be roughly between 29 and 34, and the “WGS84Longitude” should be roughly between 118 and 124. Besides, the text information needs to be cleaned. All the links, emoji, and punctuations, white spaces, digits, and tags need to be filtered by the regular expression. The data after preprocessing is described in Table 4-2.

Table 4-2: Data description after cleaning

Item	Value
Attributes	“created_at”, “id”, “user”, “text”, “coordinates”
Number of rows	11376875
Data size	2.2GB

4.3 Sentiment analysis

Following the methods proposed in chapter 3.2: the sentiment analysis consists of two parts: The lexicon-based method and the machine learning-based method. In the case study, both methods will be applied and evaluated. The best-performed model will be used in the later analysis (chapter 4.4). The results and a discussion about the result are also stated in this chapter.

4.3.1 Lexicon-based method

The type and source of the lexicons used in this experiment are listed in Table 4-3.

Table 4-3: Lexicons description

Lexicon	Source
Sentiment lexicon	Dalian University of Technology
Degree word lexicon	HowNet
Negation word lexicon	Unofficial source (28 words)
Stop word lexicon	Harbin Institute of Technology

In this thesis, a labelled ground truth test dataset from the same source Sina Weibo is used for evaluating the classification algorithms. This test dataset contains 119990 rows of records. 59994 of them are positive records, and 59996 are negative records. After running the lexicon-based classification method, the result and the evaluation score as shown in Table 4-4 have been derived.

Table 4-4: Evaluation of lexicon-based method

	Positive	Negative

Precision	0.63	0.74
Recall	0.29	0.11
F_1	0.40	0.19

From the result, in general, the F_1 score of both positive and negative records are not satisfactory, and some conclusions can be drawn here.

1. The algorithm performs better on positive than negative records.
2. The recall values generally are significantly lower than the precision value.
3. The precision is acceptable, but the recall is too low.

Discussion of some possible reasons:

1. The positive words are a little more (11267: 10797) than the negative words in the sentiment lexicon from Dalian University of Technology, and this may makes the result of positive records better than the negative records.
2. Microblog contains many informal words and slang that are not in the lexicon, so many sentiment sentences cannot be targeted. This might be the biggest reason for low recall.
3. High precision means that once the sentiment word can be found in the lexicon, the classification works not bad.

4.3.2 Machine learning-based method

The machine learning-based method is applied to compare it with the lexicon-based method.

Segmentation

The Jieba Python library (jieba, 2012) is used for segmentation. It offers Full Mode, Default Mode and Search Engine Mode. Illustrations from Jieba website are shown below:

```
#encoding=utf-8
import jieba
```

```
seg_list = jieba.cut("我来到北京清华大学", cut_all=True)
print("Full Mode: " + "/ ".join(seg_list))  # Full Mode

seg_list = jieba.cut("我来到北京清华大学", cut_all=False)
print("Default Mode: " + "/ ".join(seg_list))  # Default Mode (Accurate
Mode)

seg_list = jieba.cut_for_search("小明硕士毕业于中国科学院计算所，后在日本京都大
学深造")  # Search Engine Mode
print(", ".join(seg_list))
```

Output:

```
[Full Mode]: 我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学

[Accurate Mode]: 我/ 来到/ 北京/ 清华大学

[Search Engine Mode]： 小明，硕士，毕业，于，中国，科学，学院，科学院，中国科学
院，计算，计算所，后，在，日本，京都，大学，日本京都大学，深造
```

The algorithm segments the sentence into words. Search engine mode returns the most redundant version, the words have overlapping but the advantage is that it does not miss any possible segmentation. The accurate mode returns the most reduced version, all the words appeared only once and no overlapping. Full mode is the intermediate redundancy version. The Accurate Mode is chosen in this experiment because redundant segmentation result will affect the results of subsequently keywords frequency statistics.

Vectorization

By now, all the sentences can be accurately segmented. The next step is converting all words into vectors. In this thesis, the Word2vec model is applied. Gensim, A Natural Language Processing Python library (Řehůřek, 2009) offers word2vec algorithms. The whole dataset to train the vectorization model has been tested, but the result seems not so good. The mechanism of word2vec is based on the syntactic relationship of the context, so the quality of the model can be tested manually by comparing the input word and the top similar words of the output. An already well trained Word2vec model can return the most similar words and their similarity value. However, in the training process of Word2vec model using the whole Weibo dataset, the result is not very ideal, the best similarity is generally below 0.6. After fine-tuning all the parameters such as the size (dimensionality of the word vectors), window (Maximum distance between the current and predicted word within a sentence), min_count (ignores all the words with a total frequency lower than this), etc. the improvement of the quality is still very limited. Some possible reasons are:

1. The language used in the microblog text does not follow strict grammar and syntactic rules which the Word2vec algorithm is totally based on.
2. The quality of the preprocessing step influences the quality of the input sentences. For instance, after the text cleaning the sentence may not make sense anymore. Another example, the segmented words may not be the most accurate ones, and that also influence the syntactic structure.

In order to improve the quality of the result, a Word2vec model, trained by Kyubyong Park (2016) using 1GB Wikipedia text data, has been applied. The vector size of the vector is 300 and contains 50101 vocabularies. Even the data sources are different, but the model works quite well on the Weibo dataset. So this model is selected for this experiment.

Classification

The most important process of the machine learning method is classification. In this step, several machine learning classifiers on the already labelled dataset have been tried and evaluated using the already explained F1 score evaluation metric (see chapter 3). The best performed one is selected to train the classification model and applied on the whole experiment dataset to complete the sentiment classification task.

As discussed in chapter 3, logistic regression, k nearest neighbours classifier, support vector machine, naïve Bayes classifier, and random forest classifier will be tested and compared in this study. The optimal parameters and evaluation scores are shown in Table 4-5.

Table 4-5: The optimal hyperparameters and the evaluations of classification methods

Methods	Parameters	Accuracy	F_1 score

Logistic Regression	C=0.1, Max_iter=100, Penalty='l2', solver='lbfgs'	0.73	0.73
K Nearest Neighbors	N_neighbors=5, Leaf_size=30, metric='minkowski'	0.68	0.68
Support Vector Machine (linear kernel)	C=1, kernel='linear'	0.69	0.69
Naïve Bayes (GaussianNB)	Var_smoothing= 1e-9	0.68	0.66

Random Forest	n_estimators=200,
	min_samples_split=800, max_depth=10
		0.72	0.71

After fine-tuning of the classification models, the best-performed classifier is *Logistic Regression*. So here take Logistic Regression as an example to illustrate the fine-tuning process. Figure 4.1 shows the learning curve of the logistic regression model.

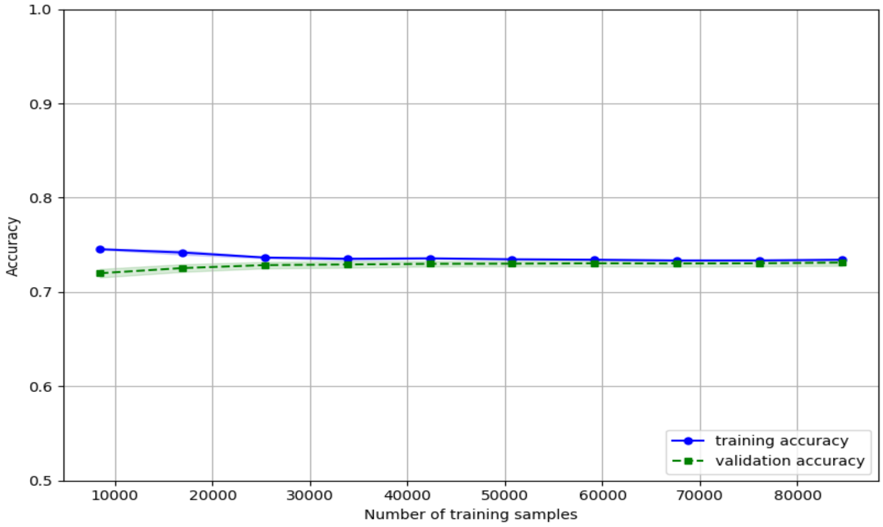


Figure 4.1: The learning curve of the logistic regression estimator

The X-axis is the number of training samples, and the Y-axis is the accuracy. From the learning curve, we can observe that the validation score and the training score converged after around 30000 training records. It means that it does not suffer from a high variance error, in another word, the estimator is not overfitting.

Besides, figure 4.2 shows a validation curve of parameter C of logistic regression estimator.

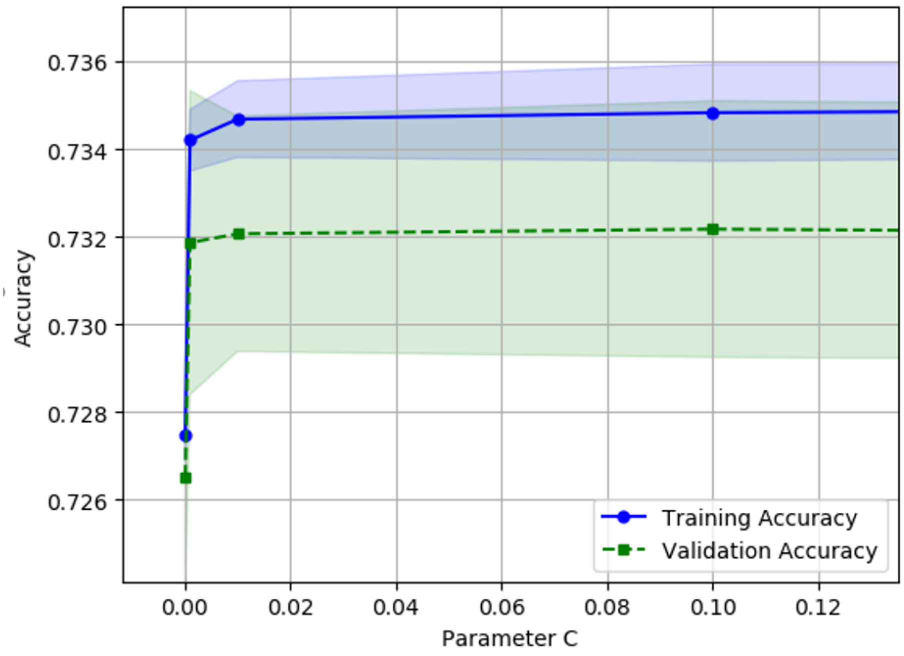


Figure 4.2: The validation curve of parameter C

Form figure 4.2, we can conclude that the parameter C does not influence the estimator so much, and it roughly reaches to a plain after the C equals to 0.1.

At last, the best parameters are determined by the grid search technique with cross-validation. In this experiment, 5-fold cross-validation was used. The final optimal parameter are: C equals 0.1 and solver is “lbfgs”.

After finding the best parameters, the whole labelled set is used to train the best classification model. Afterwards, the already trained model on the experiment dataset has been applied to accomplish the sentiment classification task. Positive records are labelled 1 and negative records are labelled 0. There are some no sentiment records because these sentences are composed of stop words. A sentence like such has no sentiment and is assigned a value of -99. No sentiment records will not be used in the following experiment. The classification result of experiment data is shown in Table 4-6. This classified data set will be used for further analysis.

Table 4-6: Classification result

Classes	Quantity
Positive	3939058
Negative	6811234
No sentiment	196411
Total	10946703

As stated in chapter 3.2.3, the ratio of the number of positive and negative records is to indicate the population sentiment orientation (PSO).

4.4 Spatial-temporal analysis-based event detection

Spatial-temporal analysis of population sentiment orientation in this study helps to determine the time, and geographical distribution of the happened social events and the word cloud technique is used for specifically visual detect what the exact event is. In the experiment part of this thesis, several case studies for testing and demonstrating the principle and application of the proposed method and workflow has been done.

4.4.1 Event detection

The hypothesis of time series analysis of this study is that the global or local maxima or minima of the population sentiment orientation of the time series can indicating social events. For testing this hypothesis, the case study is designed and implemented as follows.

First, the whole year (minimum unit is one single day) and entire Shanghai region data to detect city-scale event happened in a day of the entire year has been used. Figure 4.3 shows the time series plot.

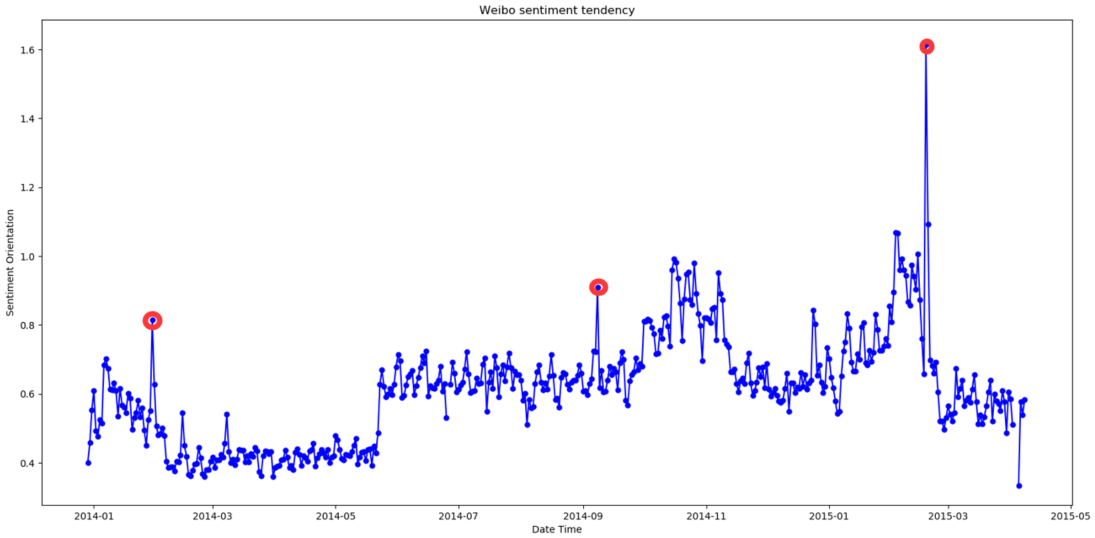


Figure 4.3: The time series plot of population sentiment orientation

In the plot, the time-series display the tendency of PSO, and in general, the PSO score is below 1.0 which is in line with common sense that people complain and express negative opinions more often in Social media. The local minima and maxima are very quickly visual detected from the plot. Therefore, the top 3 maxima points have been picked to verify the experimental hypothesis:

(1) 18.02.2015
(2) 08.09.2014
(3) 30.01.2014

As stated in chapter 3.4, extracting keywords from each sentence is based on the TF-IDF algorithm. Then the final step is to count the frequency of them and generate a Word Cloud of the keywords. Figure 4.4-4.6 show the Word Cloud of the top 3 maxima.

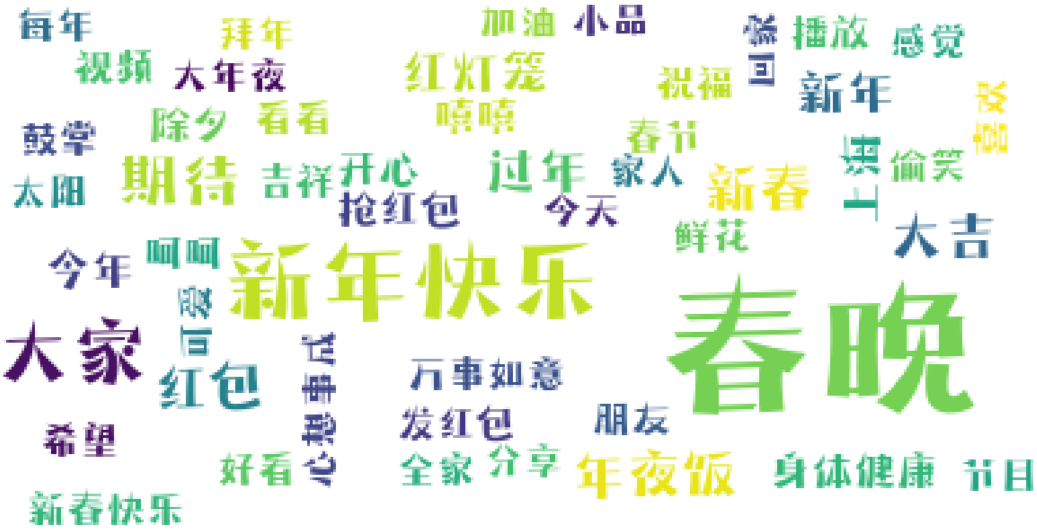


Figure 4.4: The Word Cloud of 18.02.2015, local maximum

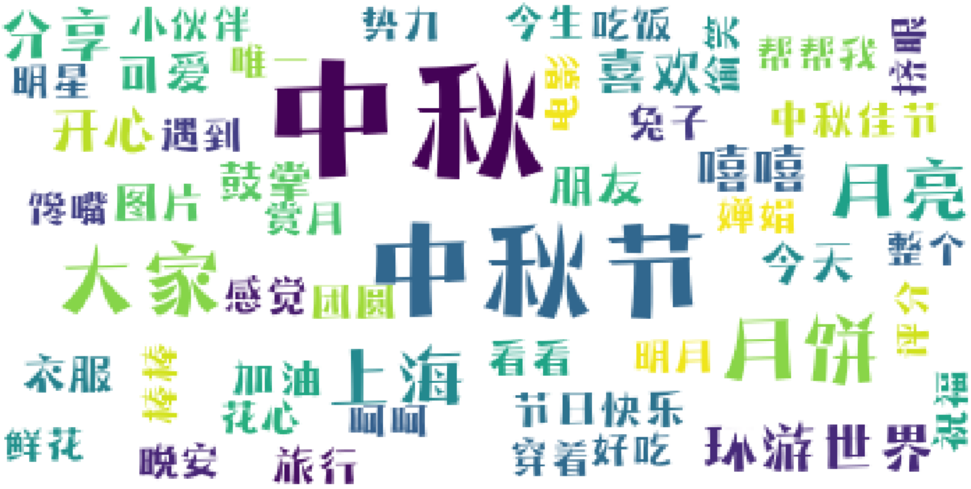


Figure 4.5: The Word Cloud of 08.09.2014, local maximum

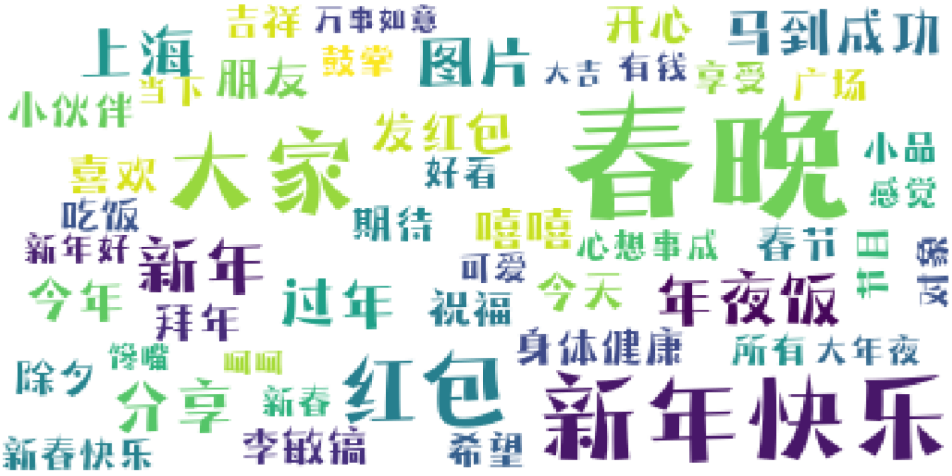


Figure 4.6: The Word Cloud of 30.01.2014, local maximum

The keywords in the Word Cloud are translated from Chinese language into English in Table 4-7.

Table 4-7: the keywords translation in Word Cloud

Date of maximum	Top frequent keywords
18.02.2015	“Spring Festival Gala”, “Happy New Year”, “Red Packets”, “New Year’s Eve”, “Family”, “New Year”, ”Family Dinner”, “Red Lantern”, “Good Wish”, “Friends”, etc.
08.09.2014	“Mid-Autumn”, “Mid-Autumn Festival”, “Moon”, “Moon Cake”, “Happy Holiday”, “Enjoying the Moon”, “Reunion”, “Happy”, “Blessing”, etc.
30.01.2014	“Spring Festival Gala”, “Happy New Year”, “Spring Festival”, “New Year’s Eve”, “Red Packets”, “Expecting”, “Friends”, “Good Health”, “Hope”, etc.

From the keywords in the Word Cloud, people can very intuitively tell what event is or what is happening. The first maximum is the New Year’s Eve of 2015, the second maximum is the New Year’s Eve of 2014, and the third maximum is Mid-Autumn festival of 2014. By the way, the Mid-Autumn Festival is an East Asian harvest festival and is held on the 15th day of the 8th of the lunar calendar. At that time, the moon should be the most round and bright at night which means family reunion (Z. Zhang, 1993). The theme of the festival is about gathering, thanksgiving and praying. Every year family and friends come together on that day. So that is an essential traditional festival of a year and has similar influence as the New Year.

The results verify the correctness of the hypothesis:

1. The local maxima or minima of the POS time series can be used as a social event indicator.
2. Abnormal population sentiment orientation value is very likely to indicate social events.
3. Word Cloud-based visual event detection method is intuitive and effective.

Small temporal scale event detection

As previously stated, this is a multi-scale spatial-temporal analysis event detection method. The time series analysis above is an annual analysis on a daily basis, so only the date of the event can be detected. In order to detect the effectiveness of this method on a fine-grained small scale data, a time series analysis of PSO on the New Year's Eve of 2014 is conducted and the plot is drawn in Figure 4.7.

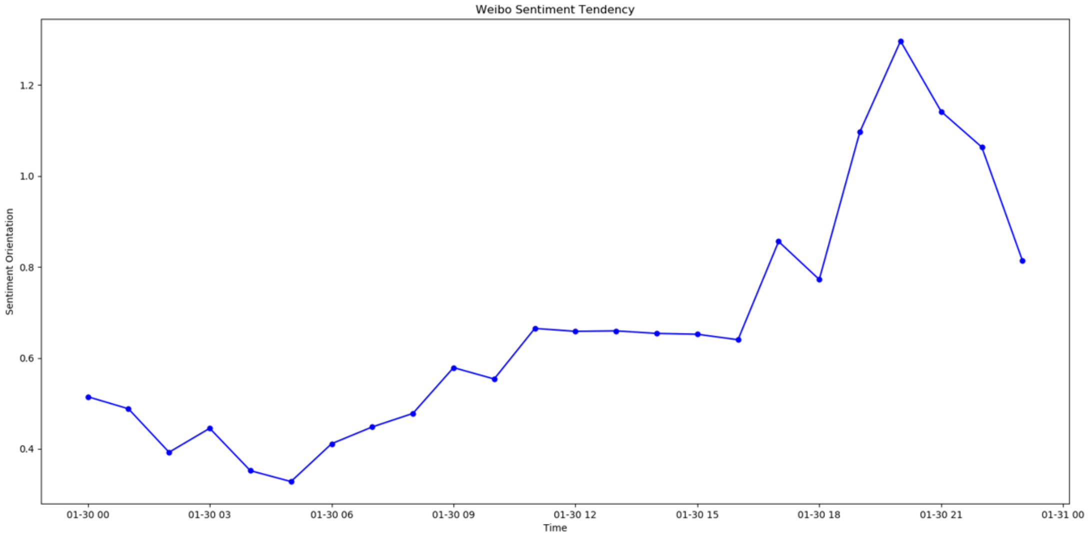


Figure 4.7: Time series analysis of the New Year's Eve of 2014

From the plot, we can see that the PSO value starts going up from around 4 pm and reaches the maximum value around 8 pm. It means that people's celebrations were mainly concentrated during this time. This result of this experiment is in line with common sense that people start preparing the dinner around 4 pm and the Spring

Festival Gala begins at 8 pm every year which makes the PSO reach the peak and gradually falling after that.

This experiment proves that the time series analysis of PSO also works on small temporal scale event detection.

Small spatial scale local event detection

Furthermore, all the events detected in this case are national festivals such as New Year' Eve and Mid-Autumn festival. That probably because the data used in the experiment is from the whole Shanghai city, so the events detected correspondingly are a city-scale event. The spatial-temporal analysis based event detection method of this study is supposed to be spatial multi-scale as well. For testing, if the method can not only work on the large spatial scale but also can detect local events, a small scale local event detection experiment was conducted as follow.

The Mercedes-Benz Arena has been selected, a representative spot in Shanghai, formerly known as the Shanghai World Expo Cultural Center as a case study of local event detection. This indoor arena located on the former grounds of Expo 2010 in Pudong, Shanghai. The facility seats 18000 people and includes some venues, and the arena hosts many ceremonies, events and activities. Figure 4.8 is the arena on the Google satellite map.

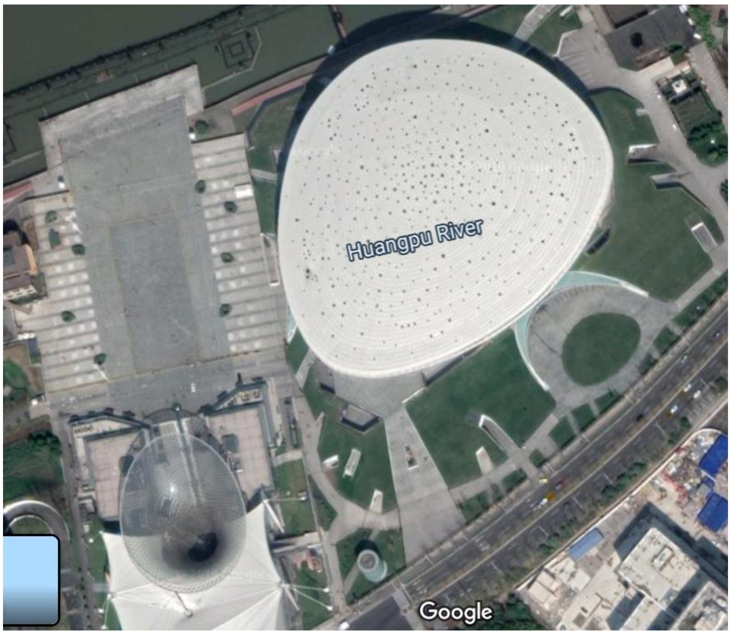


Figure 4.8: Mercedes-Benz Arena on Google map

The first step is selecting data by the location of this arena. In the experiment, the latitude range should be between 31.1900 and 31.1921, and the longitude range should be between 121.4874 and 121.4900.

Next step is the time series analysis. Figure 4.9 shows the PSO time series plot of Mercedes-Benz Arena.

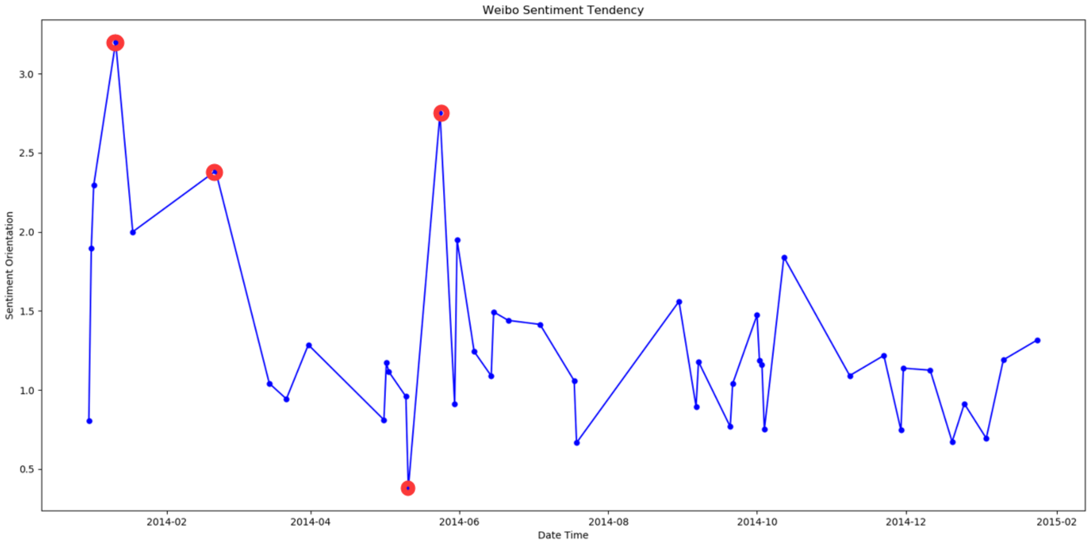


Figure 4.9: PSO time series plot of Mercedes-Benz Arena

In small-scale PSO time series analysis, there is one more thing to note compared with the large-scale analysis. Due to the small scale, the amount of data becomes very small. Thus the POS indicator has a high probability of losing its statistical significance. In this experiment, all the days with less than 40 microblogs have been filtered out to prevent the high probability.

The top 3 maxima point are:

(1) 11.01.2014

(2) 24.05.2014

(3) 21.02.2014

The minimum point is:

(1) 11.05.2014

The corresponding detection Word Clouds are shown in figure 4.10 – 4.13.

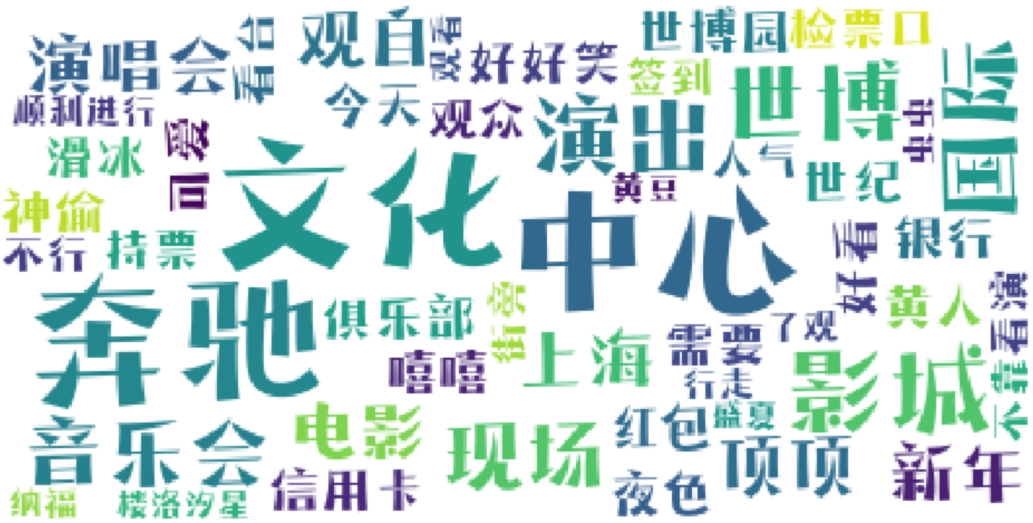


Figure 4.10: The Word Cloud of 11.01.2014, local maximum

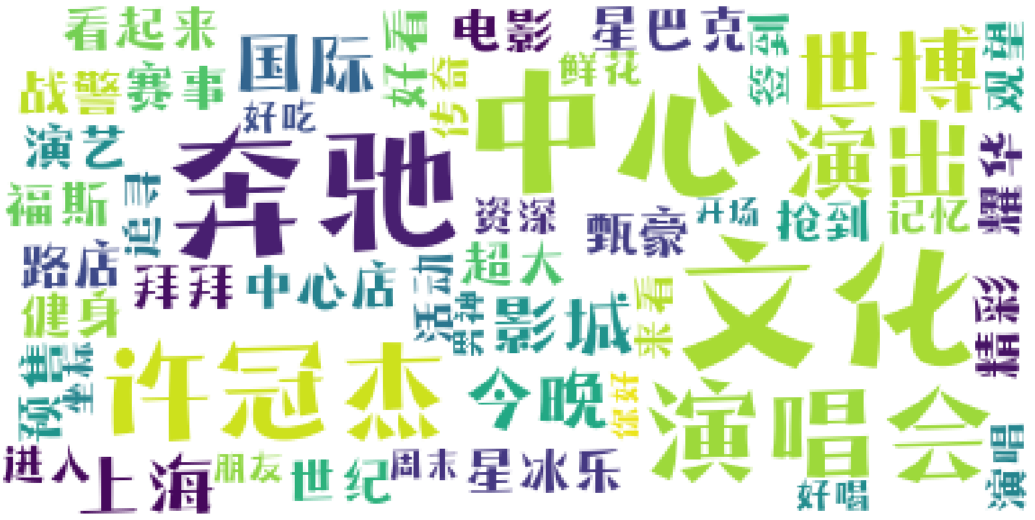


Figure 4.11: The Word Cloud of 24.05.2014, local maximum

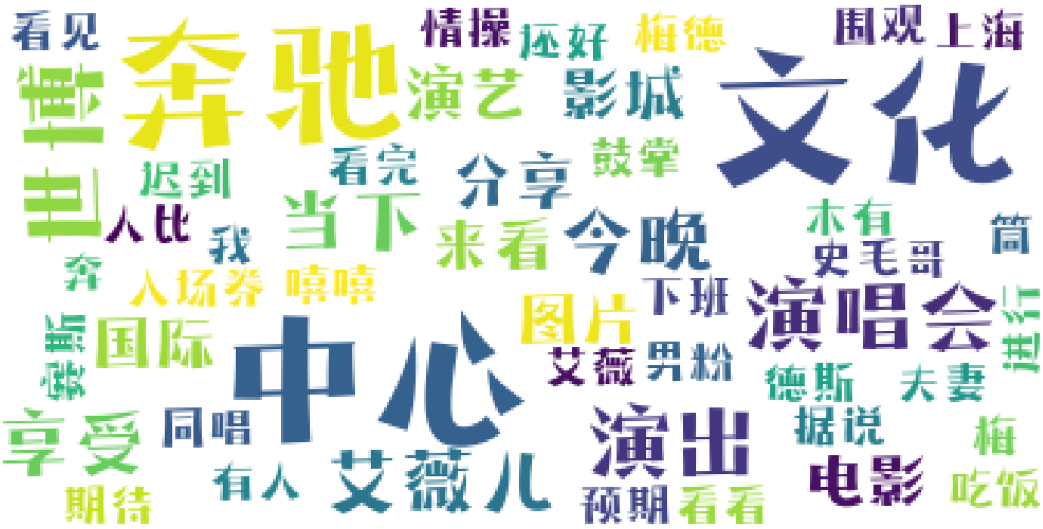


Figure 4.12: The Word Cloud of 21.02.2014, local maximum

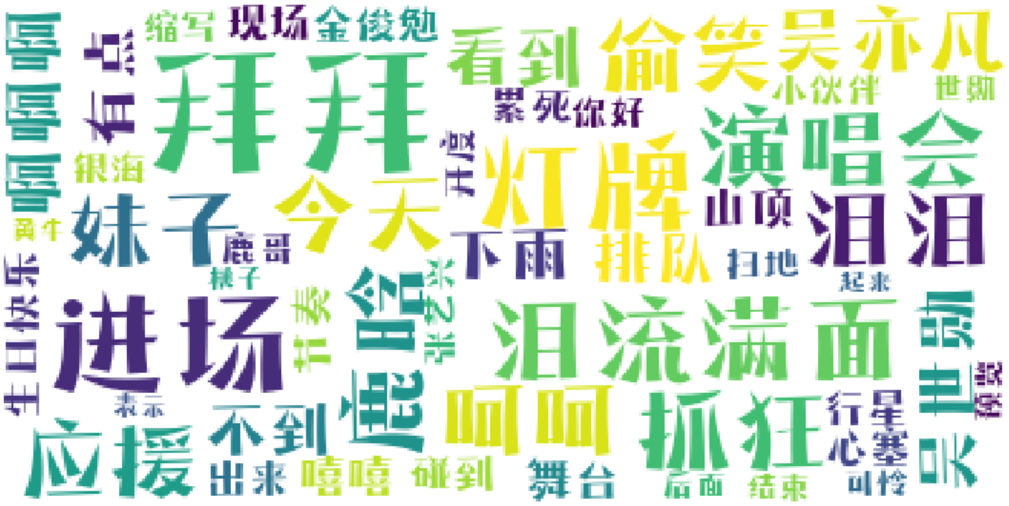


Figure 4.13: The Word Cloud of 11.05.2014, local minimum

The keywords of the Word Clouds are translated in Table 4-8.

Table 4-8: the keywords translation in Word Cloud

Date of extrema	Top frequent keywords

11.01.2014	“Cultural Center”, “Show”, “Concert”, “On Site”, ”Dingding”, “Audience”, ”Bleachers”, “Popular”, “Awesome”, ”Watch”, “International”, etc.
24.05.2014	“Benz”, “Guanjie Xu”, “Concert”, “Show”, “Opening”, “Presale”, “Activity”, “Performance”, “International”, “Sing”, “Idol”, “Bravo”, “Flowers” etc.
21.02.2014	“Benz”, “Center”, “Culture”, “Concert”, “Show”, “Expecting”, “Avril Lavigne”, “Sing”, “Watching”, “Performance”, “Fans”, etc.
11.05.2014	“Concert”, “Crying”, “Han Lu”, “Kris Wu”, “Oh SeHun”, “Suho”, “Lay Zhang”, “Stage”, “Light Board”, “Entering”, “Crazy”, “Queuing”, ”Smirking”, “Tired”, etc.

From the keywords in the Word Cloud, we can know what happened and what the event was. On 11.01.2014, there must be a concert, and in the Word Cloud, there is a singer called “Dingding”. The author looked it up and proved that on 11.01.2014, famous singers Dinging Sa and Yu Qi held a concert at Mercedes-Benz Arena Shanghai. Similarly, Guanjie Xu held a concert on 24.05.2014, and Avril Lavigne held a show on 21.02.2014.

The result of the minimum point is very interesting and need to be further explained and discussed. In the Word Cloud of 11.05.2014, there are many singer names and they are a South Korean-Chinese boy band based in Seoul called Exo. Exo held a concert on 11.05.2014 at Mercedes-Benz Arena Shanghai. Different from other events that have been successfully detected by the PSO time series, this concert event is a PSO minimum point. From the Word Cloud, we can also observe the cause and make an explanation. During the concert, people were posting words such as “Crying”, “Crazy”, “Tired”, “Queuing” and so on. It reflects that the audience and fans were crazy about this concert, and these microblogs were classified as negative class. This case proves that both maxima and minima can be used as event detection indicators.

Other than this, small spatial scale event detection on another case, HongKou Football Stadium Shanghai has been tested additionally. Local events such as football matches and concerts can also be well detected.

To sum up, in this subchapter some experiments and case studies to verify the effectiveness of the sentiment-based spatial-temporal event detection method has been done. Furthermore, this method can work on multi-scale spatial-temporal events.

It will be demonstrated and explained in the next subchapter how this method also be used for interpreting social phenomenon.

4.4.2 Event interpretation

Geographical data mining can help people to better sense and understand social phenomena. The spatial-temporal analysis-based method of this study covers not only social event detection but also support social phenomenon mining and interpretation. Taking the new year's eve of 2014 as an example, the time series plot has shown that the PSO reaches a maximum. Therefore, it should be possible to explore what the unhappy people are complaining about and what the adverse events are.

For detecting negative events, the Word Cloud based on only negative sentiment microblogs has been gathered. Figure 4.14 shows the result.

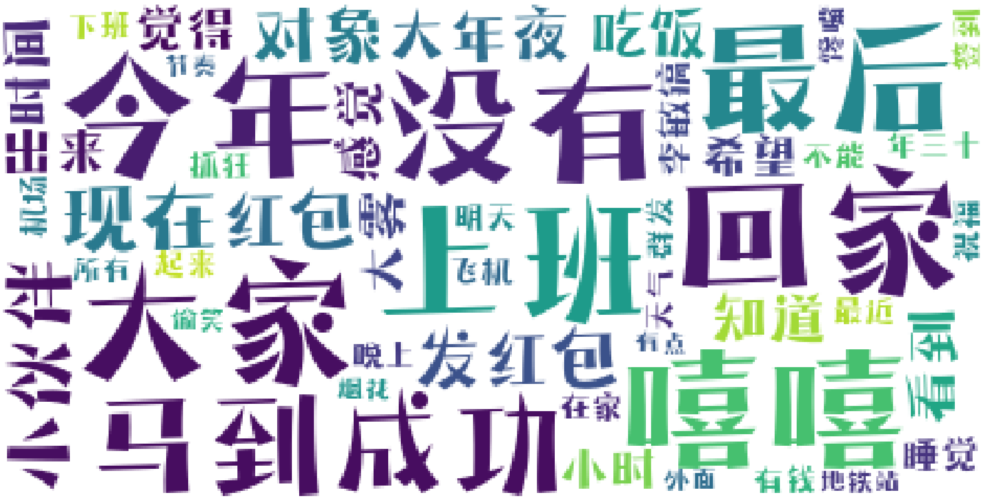


Figure 4.14: The negative Word Cloud of New Year's Eve 2014

In the course of the experiment, detecting negative events in a positive day or detecting positive events in a negative day, the words related to the primary sentiment events still appear very frequent in the opposite sentiment microblogs. So draw a conclusion here, in the process of detecting non-dominant sentiment events, frequency filtering of keywords is necessary. In this case of the new year's eve of 2014, the frequency of the keyword is set below 500.

The keywords in the Word Cloud are: “Working”, “Come back home”, “No”, “cannot”, “Airport”, ”This year”, “Outside”, “Off work”, “Airplane”, “Train station”, “Subway station”, etc. We can infer that some people are complaining that they are still working, some people are complaining that they haven’t returned home yet, and some are complaining about the traffic.

Many traffic-related words are in the Word Cloud such as “Air Plane”, “Train station”, “subway station” and ”Airport”. As an example, for people who were complaining about the traffic, a spatial clustering analysis on these microblogs has been performed trying to discover the geographical distribution patterns.

Spatial clustering analysis

First, all the traffic-related microblogs based on the traffic keywords in the Word Cloud have been extracted. Figure 4.15 is the point map of the extracted negative traffic data and Figure 4.16 is the heat map of it.

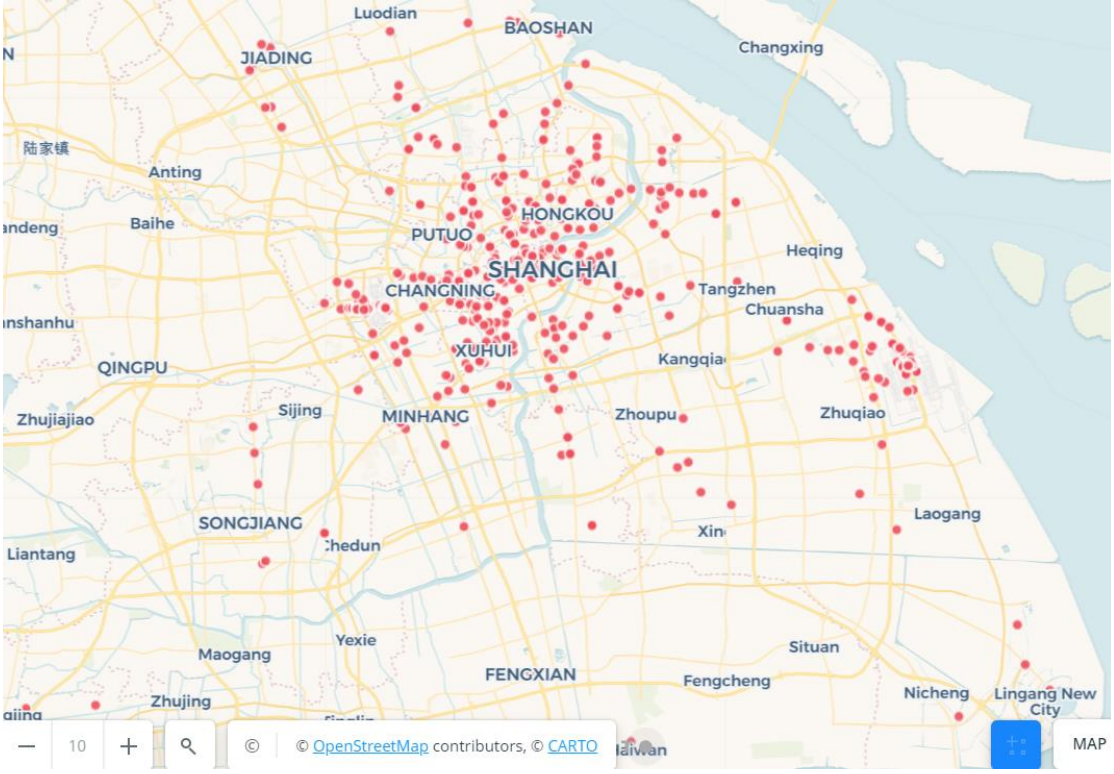


Figure 4.15: The point map of negative traffic data

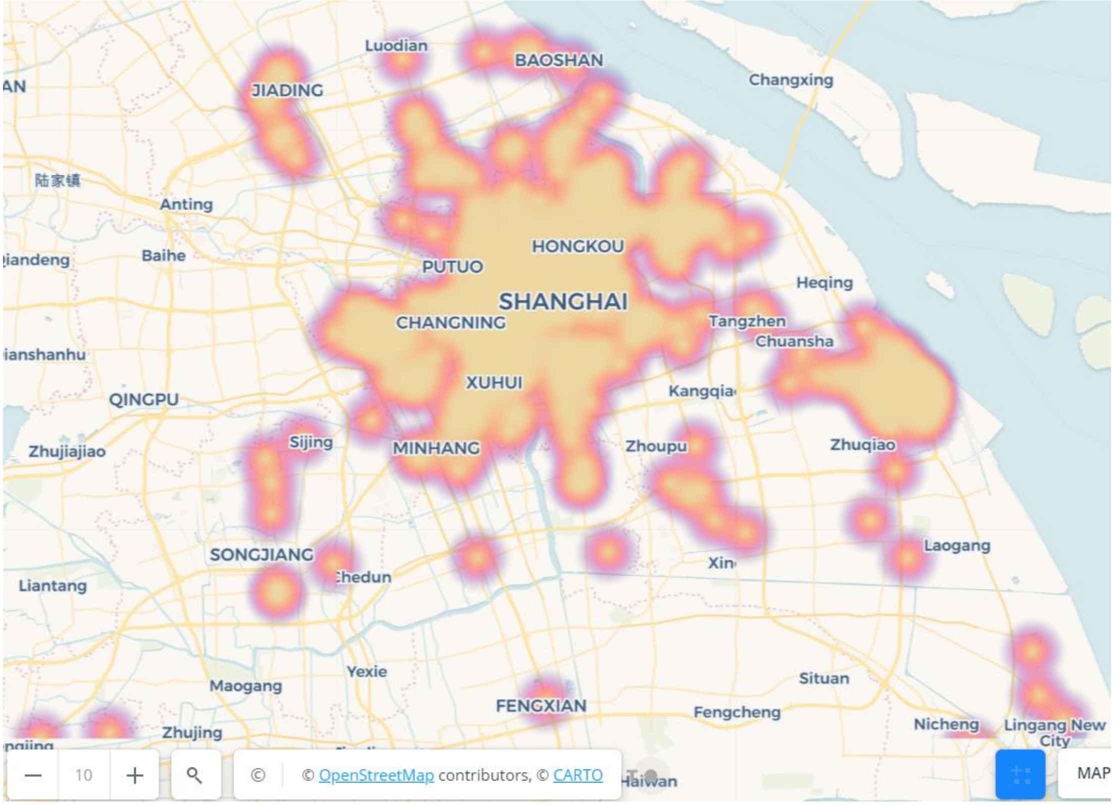


Figure 4.16: The heat map of negative traffic data

From the point map or heat map, we can get an overall impression of the geographical distribution of the negative traffic-related data. In general, the closer to the city centre, the higher the density. To deeply mining the spatial pattern, DBSCAN is applied to the data set based on the haversine formula (distance metric). Figure 4.17 is a small multiple showing the clustering results with different parameter combination. The best clustering result can be visually judged.

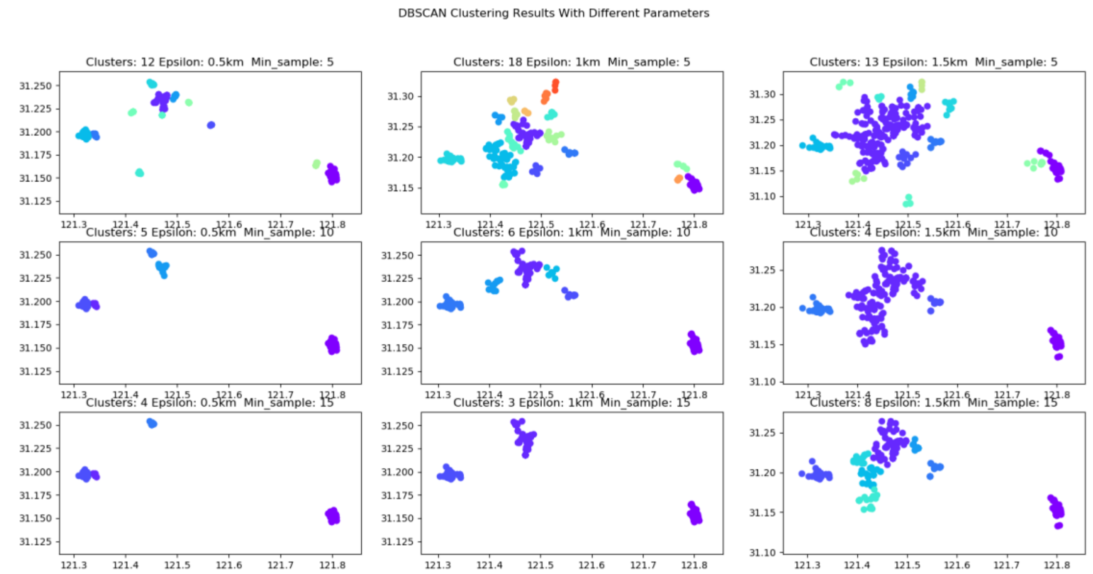


Figure 4.17: Sample multiple of clustering results

The X and Y axes are representing the geographical coordinates system. The title of each subplot is the cluster number, the parameters (Epsilon and Min_sample). The different color of the points in the plot represents a different cluster. The range of the Epsilon is from 0.5 km to 1.5 km, and the scope of the Min_sample is from 5 to 15. From the plot, we can see that there are roughly three clusters. With the 1 km Epsilon and 15 Min_sample, the algorithm returns 3 cluster, and the clustering effect seems relatively good compared with others. Therefore, the three clusters have been placed on a map to see the spatial patterns. Figure 4.18 is the map and Figure 4.19 – 4.21 are showing the zoomed-in maps of each cluster.

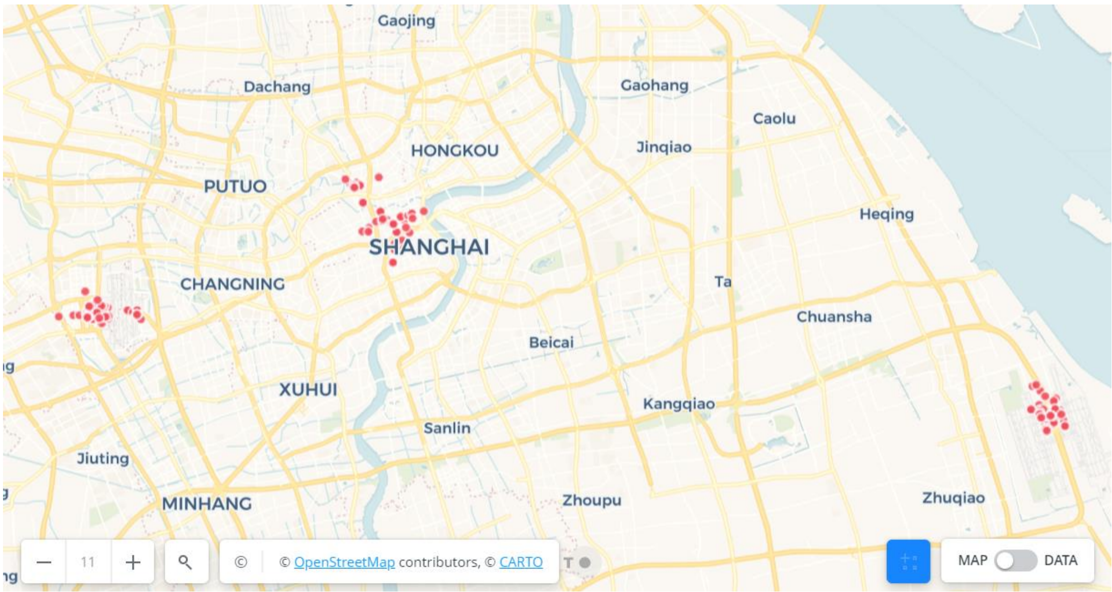


Figure 4.18: Cluster map

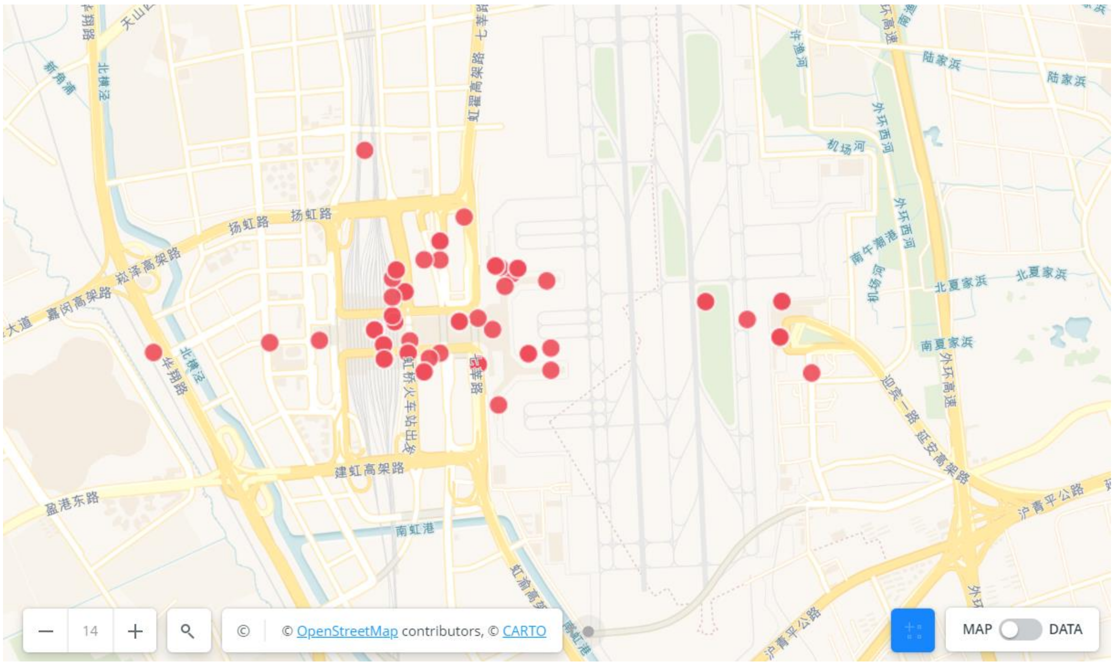


Figure 4.19: Zoom to thehe first cluster

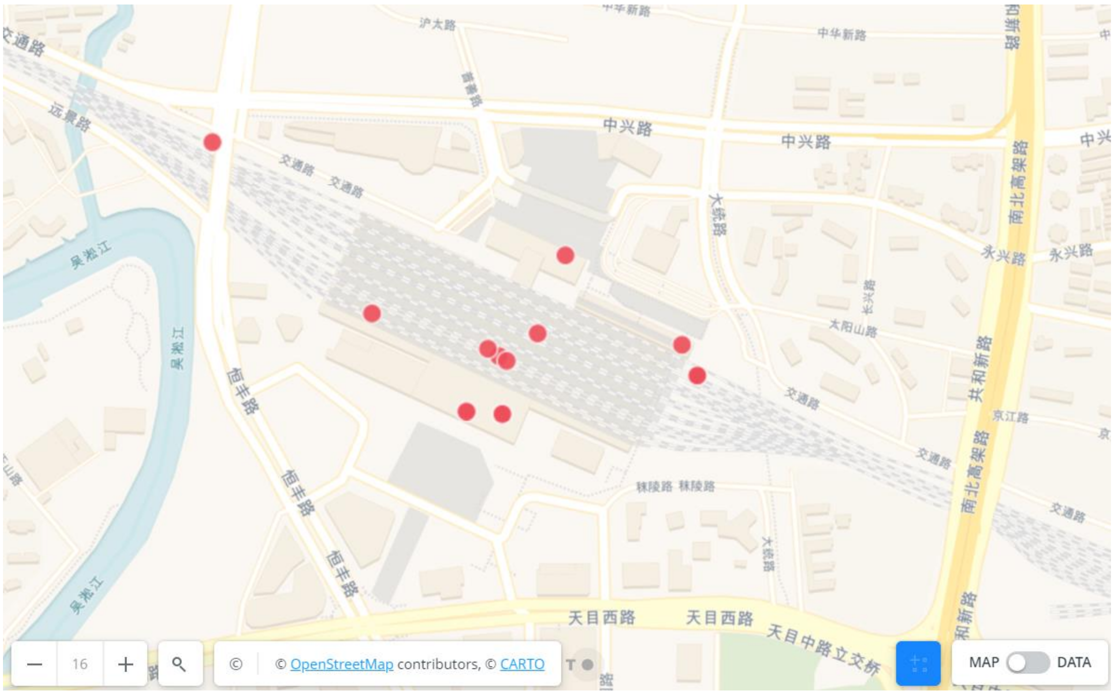


Figure 4.20: Zoom to the second cluster

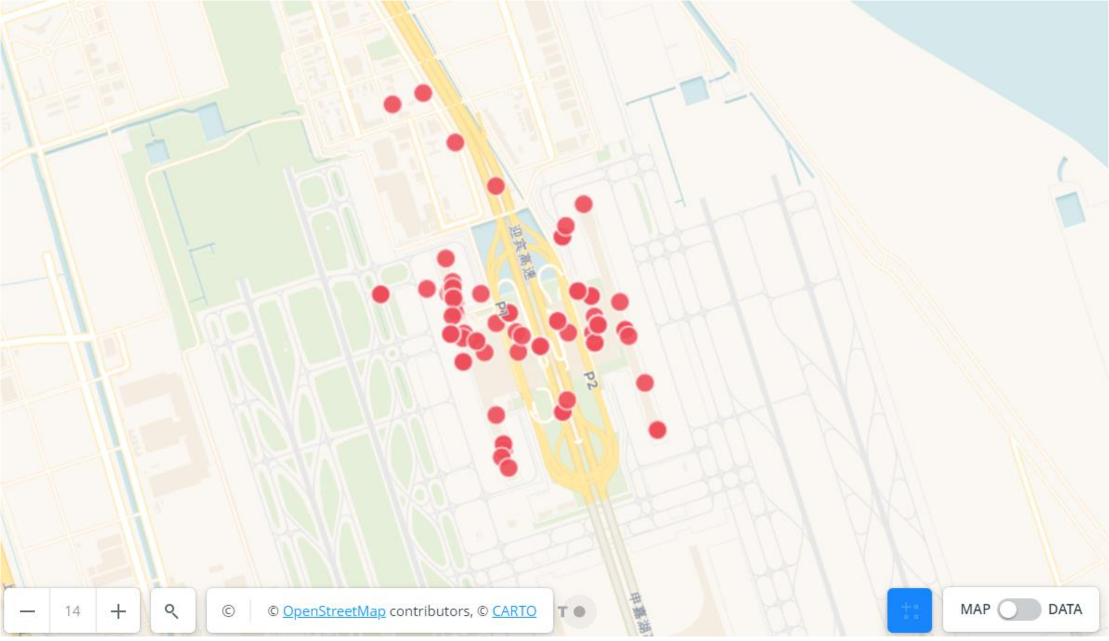


Figure 4.21: Zoom to the third cluster

With additional geographical information on the Open Street Map base map, from figure 4.19 – 4.21 we can see that the first cluster is Shanghai Hongqiao International Airport and railway station region. The second cluster is Shanghai railway station region, and the third cluster is Shanghai Pudong International Airport.

From the result of spatial analysis, we learned that the traffic-related negative microblogs on the New Year's Eve of 2014 are mainly distributed in the airport and railway area of Shanghai. That is very likely because the people were complaining that they are still on the way home. Besides, the negative data around the airport are much more than the railway station. That may be because people were complaining about delayed flights but by contrast, trains usually are not late.

In summary, the reasons why those people were in a bad mood and wanted to explain it spatially can be found by mapping the clusters. First, the Word Cloud is a helpful visualization to detect negative keywords. Many keywords are related to traffic. Secondly, a spatial clustering on the traffic-related negative data has been done. At last, the result of the clustering shows that the people complaining were mainly distributed in the airport and train station area. Therefore, this sentiment-based spatial-temporal analysis method has proved by the experiment that it can also be used for interpreting social phenomenon.

5 Discussion

After experiment and case study, a comprehensive discussion about research hypotheses, research questions and problems occurred during the experiment is necessary. The discussion mainly consists of parts: data, sentiment analysis and spatial-temporal analysis.

5.1 Data

The data part of this thesis is mainly discussed in terms of data quality and data language. Data quality directly affects the experimental results, and data language needs to be considered when applying this method in different fields and using various data sources.

5.1.1 Data quality

The proposed event detection method is designed and developed based on social media data and it takes the advantage that social media websites provide time and location tagging functions. It should be noted that although today many people like to post a microblog with location. The data used in the case study of this thesis are all geo-tagged. In general, only a tiny percent of total social media are with location. Moreover, open APIs of social media companies are all restricted nowadays. The proposed method has the potential to perform even better than in the case study with more data, especially the ability of small scale event detection. The interpretation of this method also requires the data to be statistically significant. In this regard, a process of filtering the low-frequency research objects is added to ensure the data is statistically significant.

On the other hand, although data quality very much depends on the data source. Consequently, the preprocessing also plays an important role. In this thesis, the preprocessing includes data selection, integrity check, duplication check and text cleaning. All the invalid records are deleted and not used in the experiment because the amount of invalid records only accounts for a tiny proportion of the entire data set. However, if the ratio of invalid records is too large to be ignored, a process of correcting the records is necessary.

5.1.2 Data language

This thesis is the Chinese language-oriented, so the method and the experiment are designed and tested in Chinese. For someone may be interested in porting this method to other languages, only the text cleaning and sentiment analysis parts need to be adjusted according to the language.

5.2 Sentiment analysis

The main task of sentiment analysis of the proposed method is sentiment classification. The two main methods, lexicon-based method and machine learning-based method, have been respectively tested.

5.2.1 Lexicon-based method

The main pros and cons of lexicon-based method are:

1. High precision and low recall
   This is a commonly-known problem of the lexicon-based method (L. Zhang, Ghosh, Dekhil, Hsu, & Liu, 2011), and it does appear in the experiment. This method can only recognize the sentiment words that exist in the lexicon. This will result in a significant percentage of the data that cannot be classified, thus this means a lot of waste of data. In the experiment, there is almost half of the data that cannot be classified, which is the main reason for the low recall.

1. Lexicon dependency
   In the lexicon-based method, the quality of classification results depends on the quality of the lexicons. In practice, it is almost unrealistic to build lexicons only for a classification task. Besides, there are not so many lexicons researchers can choose, especially in Chinese. So under these circumstances, there is not much room for the improvement of lexicon-based classification if the result is not very good. In the experiment, all the well-known Chinese sentiment lexicons are compared and the one from Dalian University of Technology is picked, but the results show that still a vast amount of sentiment words appeared in Weibo are not in the lexicon.

2. No training set required

One advantage of lexicon-based method compared with machine learning-based method is that this method does not require ground truth labelled training data set. Labelling data is also very time-consuming and labor-intensive work.

5.2.2 Machine learning-based method

The Word2vec model is used. The result trained with the Weibo data is not so satisfying, but the result trained with Wikipedia data is relatively good. The possible reasons have been discussed in chapter 4.3.2 already. There is another popular way of vectorization called One Hot Encoding. However, it is not guaranteed that this vectorization can perform better than Word2vec model.

Five supervised classification algorithms are used and compared in the experiment. The Logistic Regression classification got the best score, and the scores of other algorithms are between 0.6 and 0.7. Some possible reasons and explanations are listed below.

1) Logistic Regression has an advantage for binary classification problems and converges very fast.

2) All the results are around 0.7. The possible limiting factors of the classification results are:
   a) Text cleaning might destroy the semantic structure of sentences.
   b) The segmentation is not accurate enough.
   c) The vector of each record is the average of every word vector in the sentence. Averaging is not an accurate way of representing the sentence. Ideally, some words are more representative and therefore should be given a higher weight.

3) The quality of the classification depends more on the preprocessing of the input data than on the selection of different classifiers. The quality of the preprocessing and the input data determine the upper limit of the classification result.

5.3 Spatial-temporal analysis

In the experiment, the proposed event detection method has been proved that it can successfully detect multi-scale spatial-temporal events. Accurately, in the time dimension, the day of activities can be detected from a large time scale like a year and the period of a happening event can be detected from a small time scale like a day. In the spatial dimension, a large spatial scale event can be detected like a national festival and a small spatial scale local event can also be successfully detected like a concert.

The development of an interactive web system based on the proposed method is recommended. This system would be more flexible in the spatial-temporal analysis as an interactive system can provide more analysis dimensions. To be more specific, in the spatial dimension, this method has been proven effective in multi-scale event detection by two case studies. With a system, researchers can choose any spatial scale and region or any temporal dimension. In the experiment, only the time series analysis in days (00:00 to 24:00) has been tested. However, most events are not happening the whole day, and this reduces the accuracy of the results. Ideally, researchers can choose the unit of time series based on what kind of event they want to detect. Moreover, there is other dimension of time. For example, the dimension of 24 hours of a day can help to detect periodic events in days. Morning, afternoon, evening and night can also be a dimension. The season dimension can help to detect seasonally-period events.

6 Conclusion and outlook

6.1 Conclusion

The main innovation of this thesis is a new approach of event detection which combines sentiment analysis and spatial-temporal analysis. Moreover, this method also has the ability of event interpretation.

In the thesis, the hypotheses have been verified and the research questions (see chapter 1) have been answered through experiments. Finally, the conclusions can be summarized as follows:

1. Population sentiment orientation (PSO) can be used to detect events in the dimension of time and space. This further proves the validity of the fundamental hypothesis of this thesis that social events change PSO.

2. The ratio of positive and negative records of the social media can serve as an indicator for PSO.

3. This event detection method can successfully detect multi-scale events. From nationwide festival to local activity and from annual-scale to day-scale detection, the method works pretty well.

4. In general, the extrema of the time series of PSO most likely represent social events. Therefore, social events can be visually detected by the time series plot.

5. This method can also be used to interpret events. Specifically, spatial-temporal analysis of this method can help mining spatial-temporal patterns of the event.

6. From the experimental results, overall, the machine learning-based sentiment classification is better than the lexicon-based sentiment classification in terms of the given Weibo data set.

7. Spatial-temporal analysis in this thesis mainly contains time-space scale transformation, time series analysis and clustering analysis.

8. Unlike event detection for specific events, this sentiment-based event detection method does not require related keywords training set, so it can detect wider types of events.

6.2 Outlook

In general, the proposed event detection method has been successfully verified through experiments and case studies. However, some details can be added or improved in the future. Based on this method and theory, many interesting research can be carried out further in the future.

In the sentiment classification part, many classification algorithms are tried, and the best score is 0.73. More sophisticated algorithms can be tested on the data set like Neural Networks. Besides, other vectorization models can be compared with Word2vec like One Hot Encoding with TF-IDF. Besides, emoji can also be used to help improve the sentiment classification.

An interactive web event analysis system can be developed based on this thesis method. It will provide more flexibility and a better visualization. All the techniques applied in this thesis can serve as prototype functions of this web system. Users can upload data files in the specified format. The sentiment classification algorithm runs on the server based on an already well-trained classification model. In the front end, users can select the analysis region and period interactively. The system will automatically plot the time series and generate the Word Cloud to detect events.

Additionally, there is a potential to develop an Event Monitoring and Early Warning System. To be specific, with enough computing power, it is possible to calculate population sentiment orientation in real-time, in other words, monitoring PSO. With some thresholds, abnormal PSO values can be detected, and with a further Word Cloud inspection, an event can be finally determined.

Multi-source data fusion can be a direction to improve the reliability and accuracy. In this thesis, a concept called social sensing is introduced (see chapter 1.1). This thesis is based on that and sentiment information is one factor of social sensing. The source of sentiment information in this thesis is social media, specifically Sina Weibo. It can be inferred that this method can be improved with more and different data sources. For example, a fusion data of Twitter, Flickr and Instagram will definitely sense the social sentiment more comprehensively. Multi-source data fusion is definitely a trend, and it will increase the credibility and robustness of the proposed method.

References

Ahmed, K., El Tazi, N., & Hossny, A. H. (2015). *Sentiment analysis over social Networks: An overview.* Paper presented at the 2015 IEEE International Conference on Systems, Man, and Cybernetics.

Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications, 125*(3).

Allen, J. D., Anderson, D., Becker, J., Cook, R., Davis, M., Edberg, P., . . . Ishida, R. (2012). *The Unicode Standard* (Vol. 8): Citeseer.

Alpaydin, E. (2009). *Introduction to machine learning*: MIT press.

Andreas Mueller. (2019). Word Cloud. Retrieved from https://amueller.github.io/word_cloud/.

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.* Paper presented at the Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01.

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. In *Sentiment analysis and ontology engineering* (pp. 313-340): Springer.

Bhui, K., Everitt, B., & Jones, E. (2014). Might depression, psychosocial adversity, and limited social assets explain vulnerability to and resistance against violent radicalisation? *PloS one, 9*(9), e105918.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*: John Wiley & Sons.

Carto. (2011,). Carto. Retrieved from https://carto.com.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., & Ertl, T. (2012). *Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition.* Paper presented at the IEEE VAST.

Chire. (2011). DBSCAN Illustration. Retrieved from https://en.wikipedia.org/wiki/DBSCAN.

Cover, T. M., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory, 13*(1), 21-27.

Cuzick, J., & Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society: Series B (Methodological), 52*(1), 73-96.

Day, J. W., Britsch, L. D., Hawes, S. R., Shaffer, G. P., Reed, D. J., & Cahoon, D. (2000). Pattern and process of land loss in the Mississippi Delta: a spatial and temporal analysis of wetland habitat change. *Estuaries, 23*(4), 425-438.

Diday, E., & Simon, J. (1976). Clustering analysis. In *Digital pattern recognition* (pp. 47-94): Springer.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., & Xu, K. (2014). *Adaptive recursive neural network for target-dependent twitter sentiment classification.* Paper presented at the Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers).

Dong, Z., Dong, Q., & Hao, C. (2010). *Hownet and its computation of meaning.* Paper presented at the Proceedings of the 23rd international conference on Computational Linguistics: Demonstrations.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* Paper presented at the Kdd.

Girolami, M., & Kabán, A. (2003). *On an equivalence between PLSI and LDA.* Paper presented at the

SIGIR.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211-221.

Goutte, C., & Gaussier, E. (2005). *A probabilistic interpretation of precision, recall and F-score, with implication for evaluation.* Paper presented at the European Conference on Information Retrieval.

Gundecha, P., & Liu, H. (2012). Mining social media: a brief introduction. In *New Directions in Informatics, Optimization, Logistics, and Production* (pp. 1-17): Informs.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2): Princeton university press Princeton, NJ.

Han, J., Kamber, M., & Tung, A. K. (2001). Spatial clustering methods in data mining. *Geographic data mining and knowledge discovery*, 188-217.

Hand, D. J. (2006). Data Mining. *Encyclopedia of Environmetrics, 2*.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all? *International statistical review, 69*(3), 385-398.

Ho, T. K. (1995). *Random decision forests.* Paper presented at the Proceedings of 3rd international conference on document analysis and recognition.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., . . . Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 454*(1971), 903-995.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.

jieba. (2012). "jieba" Chinese text segmentation. Retrieved from https://github.com/fxsjy/jieba.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53*(1), 59-68.

Kyubyong Park. (2016). Pre-trained Chinese Word2vec model. Retrieved from https://github.com/Kyubyong/wordvectors.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5*(1), 1-167.

Liu, J., Liu, M., Tian, H., Zhuang, D., Zhang, Z., Zhang, W., . . . Deng, X. (2005). Spatial and temporal patterns of China's cropland during 1990–2000: an analysis based on Landsat TM data. *Remote sensing of Environment, 98*(4), 442-456.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., . . . Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers, 105*(3), 512-530.

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). *Population bias in geotagged tweets.* Paper presented at the Ninth international AAAI conference on web and social media.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal, 5*(4), 1093-1113.

Meliker, J. R., & Sloan, C. D. (2011). Spatio-temporal epidemiology: principles and opportunities. *Spatial and spatio-temporal epidemiology, 2*(1), 1-9.

Michael Jendryke, T. B. (2015). Weibo data 2014. Retrieved from http://www.lmars.whu.edu.cn/index.php?l=en.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Montoyo, A., MartíNez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems, 53*(4), 675-679.

Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior, 31*, 527-541.

Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining.* Paper presented at the LREc.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques.* Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2*(11), 559-572.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2009). *Expanding domain sentiment lexicon through double propagation.* Paper presented at the Twenty-First International Joint Conference on Artificial Intelligence.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull., 23*(4), 3-13.

Řehůřek, R. (2009). gensim. Retrieved from https://radimrehurek.com/gensim/.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery, 2*(2), 169-194.

Scholkopf, B., & Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*: MIT press.

Taboada, M. (2016). Sentiment analysis: an overview from linguistics. *Annual Review of Linguistics, 2*, 325-347.

Tan, S., Cheng, X., Wang, Y., & Xu, H. (2009). *Adapting naive bayes to domain adaptation for sentiment analysis.* Paper presented at the European Conference on Information Retrieval.

Tang, D., Wei, F., Qin, B., Liu, T., & Zhou, M. (2014). *Coooolll: A deep learning system for twitter sentiment classification.* Paper presented at the Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014).

Tran, T. N., Drab, K., & Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems, 120*, 92-96.

Turner, M. G. (1990). Spatial and temporal analysis of landscape patterns. *Landscape ecology, 4*(1), 21-30.

Twitter. (2019). Tweet Object. Retrieved from https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html

Van Brummelen, G. (2012). *Heavenly mathematics: The forgotten art of spherical trigonometry*: Princeton University Press.

Wei, W. W. (2006). Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

Wu, Z., & Leahy, R. (1993). An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis & Machine*

Intelligence(11), 1101-1113.

Xianghua, F., Guo, L., Yanyan, G., & Zhiqiang, W. (2013). Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems, 37*, 186-195.

Yadav, V., & Elchuri, H. (2013). *Serendio: Simple and Practical lexicon based approach to Sentiment Analysis.* Paper presented at the Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013).

Yao, J., Cui, B., Huang, Y., & Jin, X. (2010). *Temporal and social context based burst detection from folksonomies.* Paper presented at the Twenty-Fourth AAAI Conference on Artificial Intelligence.

Yu, X. L. L. H. P., & Jianmei, R. H. C. (2008). Constructing the Affective Lexicon Ontology [J]. *Journal of the China Society for Scientific and Technical Information, 2*, 6.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011, 89*.

Zhang, Z. (1993). A brief account of traditional Chinese festival customs. *Journal of popular culture, 27*(2), 13.

Zhao, Q., & Mitra, P. (2007). *Event Detection and Visualization for Social Text Streams.* Paper presented at the ICWSM.

Zheng, Y.-T., Zha, Z.-J., & Chua, T.-S. (2012). Mining travel patterns from geotagged photos. *ACM Transactions on Intelligent Systems and Technology (TIST), 3*(3), 56.

Zhou, S., Chen, Q., & Wang, X. (2014). Fuzzy deep belief networks for semi-supervised sentiment classification. *Neurocomputing, 131*, 312-322.

Zhou, X., & Chen, L. (2014). Event detection over twitter social media streams. *The VLDB Journal—The International Journal on Very Large Data Bases, 23*(3), 381-400.

Appendix

The attached DVD contains the data and code of this thesis. The data includes the case study experiment data from Sina Weibo, all the lexicons used in sentiment analysis and machine learning models traind in the experiment. The code part includes all the data processing script and algorithms mentioned in the thesis.