# Introduction

## Introduction

With the advancement of Internet technology and the popularity of smart mobile devices, social media has become an indispensable part of people's daily lives. The short text content such as tweets, microblogs, comments, replies, etc. posted by users on social media platforms not only reflects their immediate emotions and attitudes, but also has great research value. Whether it is for the business activities of enterprises (such as public relations) or the social governance of the government, the discovery and tracking of public opinion is very important. The core algorithm in public opinion analysis is the sentiment analysis algorithm.

The main creators of data on social networks are ordinary people. When they create content spontaneously, the content is diverse, including but not limited to short videos, short texts, pictures, emoticons, long texts or long videos. These are different from the traditional language environment. Therefore, this study will focus on the special language environment of social networks, collect and analyze data on online short texts on multiple social media in China, build a new word discovery system based on the language environment, track and collect data on multiple platforms for the same event at the same time, and use natural language processing technology and deep learning models for sentiment analysis.

## Background of the Problem

With the popularity of social media, online public opinion plays an increasingly important role in shaping public opinion and influencing social events. For policymakers, social media is an important channel to understand public opinion and needs. Through deep learning analysis of large-scale text data from online platforms such as social media and news comments, the public's reactions and emotional tendencies to specific policies or events can be quickly captured. This provides policymakers with real-time and comprehensive public opinion feedback, which helps to formulate policies that are more in line with public opinion. For companies, brand crises may break out at any time and have a serious impact on their reputation. By analyzing emotions on social media through deep learning technology, companies can promptly detect the spread of negative emotions and take corresponding measures for crisis management. This analysis can help companies understand consumer emotions, develop effective communication strategies, and mitigate the damage caused by crises.

In recent years, there has been an endless stream of research on online public opinion analysis. Li systematically shared the development and research status of Chinese public opinion research over the past few years in the "Current Status and Research Findings of Internet Public Opinion in China." in 2023. The core algorithm of public opinion analysis, sentiment analysis technology, has developed and iterated very quickly. From the earliest natural language models based on statistics, to deep learning models based on recurrent neural networks, to deep learning methods based on pre-trained models, and the research focus in the past two years, multimodal and multi-task sentiment analysis models. New models are increasingly focusing on sentiment analysis in complex language environments such as online texts, and the requirements for high-quality data sets are also constantly decreasing. For example, in 2023, Chen proposed a high-performance semi-supervised sentiment analysis method.

## Statement of the Problem

At present, China's social media is mainly divided into three categories. The first is short text social media represented by Weibo, Tieba, and Zhihu. The main content is spontaneously created by platform users, including text and replies. The actual content includes short text and emoticons. The second category is vertical social media represented by Xiaohongshu. The content is highly aggregated by field characteristics. The specific content is mainly pictures and texts, and comments and replies include short texts and emoticons. The third category is video websites represented by Bilibili, Kuaishou, and Douyin. The publishers are more formal and mainly create videos. Comments and replies are mainly short texts, including emoticons.

Existing social media research is limited to the first type of social media. There is no good research on the cross-domain issues of the second type of social media, and there is no special research on the third type. When the same event appears on different platforms at the same time, users on different platforms react differently. Traditional research has failed to effectively understand the differences and conduct unified analysis and processing. This study aims to build a cross-platform sentiment analysis algorithm, intending to achieve unified multi-platform data analysis from data collection and core algorithms.

## Research Questions

1. How to collect and organize data on the same event on different types of platforms with different structures?
2. What are the differences in the content of network data from different types of media sources for the same event?
3. Can a deep learning model be built to perform sentiment analysis on short online texts on multiple platforms at the same time?

## Objectives of the Research

In response to the above problems, this study proposed the following research objectives:

1. Build a program that can collect multi-platform data for a specific event.
2. Analyze the data content of the same event on different platforms to find out the similarities and differences.
3. Study the actual effect of different types of sentiment analysis models on the collected data? Is there room for improvement?
4. Build a deep learning model that can perform efficient data analysis on short online texts on multiple platforms.

## Scope of the Study

The research will use publicly available online datasets, including but not limited to SMP2020 and cnsent.

1. The analysis will include at least one of the three types of social media.
2. The analysis will include the impact of introducing traditional text sentiment analysis datasets on online short text analysis.
3. The analysis will limit the content to Chinese social media.
4. The analysis will include the impact of cross-language models, cross-language datasets, and sentiment analysis.
5. The analysis will be limited to online short texts and will not include videos, pictures, or long texts.

## Significance of the Research

This research can more effectively track and discover social public opinion, and solve the problem that the same event has different responses on different platforms, while traditional methods are ineffective. Ultimately, it can effectively track and discover the real impact of the same event on the Internet, which is beneficial to the government's social governance and corporate brand crisis management.

# **Chapter 2**

# **literature review**

## **2.1 Introduction**

In recent years, with the rapid rise of social media in China, cross-platform public opinion tracking and analysis has become a hot topic in the research field. Researchers pay wide attention to the platform characteristics, user behavior, public opinion communication path and the application of key technologies. At the same time, the far-reaching influence of public opinion on enterprise business activities and government decision-making is also gradually emerging. Especially in the spread of network rumors and the management of social expectations, its intervention effect on economic and social operation has attracted more and more attention. This section provides a systematic review of relevant literature, covering the communication mechanism of public opinion, short essays and emotional analysis, application of key articles, technologies, social impact of Internet rumors, and public opinion response in government decision-making.

## **2.2 Public opinion communication mechanism**

The basic mode of public opinion communication is divided into two categories: user-generated content communication and platform algorithm recommendation. The propagation of user-generated content is profoundly influenced by the structure of social networks and user interaction patterns. Liu et al. (2020) pointed out that information transmission has a significant homogeneous aggregation effect, and the intensity of transmission between different groups is regulated by social capital and interest labels.

On the other hand, the platforms recommendation algorithm further shapes the information transmission path by optimizing the users stay time and participation rate (Zhang & Sun, 2022). The heterogeneity of communication path aggravates the complexity of public opinion between platforms, and puts forward higher technical requirements for public opinion tracking and analysis.

## 2.3 Short text and emotion analysis

### ****2.3.1 Short text processing****

Online short texts (such as microblogs, comments) are often incomplete semantically due to their informality and simplicity, which brings significant challenges to natural language processing (NLP). Sun et al. (2021) pointed out that there are many ellipsis, emojis and non-standard expressions in short texts, which are difficult for traditional syntactic analysis methods. The pre-trained model based on Transformer (e. g., BERT, GPT) makes up for the lack of information in the short text through context modeling and semantic extraction at the subword level.

### ****2.3.1 sentiment analysis****

As an important tool for public opinion research, emotion analysis has developed from a single emotion classification to multi-dimensional emotion modeling, including emotion intensity, emotional transition pattern, and implicit affective expression (Liu et al., 2022). The emotion analysis model based on LSTM and Transformer can capture the dynamic changes of users emotions in the timeline, and provide theoretical support for public opinion prediction and response (Yu et al., 2023).

## 2.4 Application of key technologies

### 2.4.1 Long-and Short-term Memory Network (LSTM)

LSTM performs well in processing time series tasks, so it is widely used in the field of public opinion analysis. Phaladisailoed and Naruetharadhol (2019) showed that LSTM is better than traditional statistical models in capturing dynamic changes in public opinion. In addition, the LSTM model incorporating multimodal data (including text and pictures) significantly improves the accuracy of emotion prediction (Chen et al., 2021).

### 2.4.2 Transformer architecture and pre-training model

Transformer Effectively capture the context information through the multi-head self-attention mechanism, and perform well in short text processing and emotion analysis. Xu et al. (2023) showed that emotion classification models based on BERT showed significant advantages in dealing with short texts on heterogeneous platforms.

### 2.4.3 Large Language Model (LLM)

Large language model (such as GPT-4) can be applied to complex public opinion analysis tasks due to its strong generation and understanding ability. Kristjanpoller et al. (2021) used LLM to achieve emotional trend prediction in social media, and assisted in policy formulation and the construction of public opinion simulation scenarios.

## 2.5 Social and commercial impact of Internet rumors

### **2.5.1 Rumor communication mechanism**

Research has shown that online rumors attract attention through emotional language and show a multilevel diffusion path (Wang et al., 2022). The platform algorithm has further amplified the influence of these diffusion paths, resulting in the rapid diffusion of public opinion with unreal content.

### **2.5.2 Influence on society**

Online rumors may lead to social panic, policy misleading and declining public trust in authorities. For example, during the COVID-19 outbreak, the spread of misinformation exacerbated social unrest and interfered with the implementation of government public policy policies (Kim & Park, 2021).

### **2.5.3 Impact on the business activities of enterprises**

The impact of Internet rumors on enterprise business activities is reflected in the following aspects:

1. Brand reputation: Negative rumors may lead to a decline in consumer brand trust in the company, which will directly affect sales performance. Chae et al. (2022) found that if enterprises fail to respond to online rumors in time, the negative impact of the brand may last for more than half a year.
2. Supply chain and operational disruptions: Some rumors may mislead public awareness of the quality or safety of products, leading to product recalls or sales stagnation. A typical case is large-scale returns due to rumors (Li et al., 2023).
3. Stock market volatility: Financial rumors usually directly affect the stock price volatility of listed companies. Rahman et al. (2021) found through empirical research that the stock market turmoil caused by negative rumors will aggravate the capital flow pressure of enterprises, and may affect the investment strategy of enterprises.

### ****2.5.4 Governance strategy****

The rumor monitoring system combining LSTM and Transformer model has been widely used in business environments to quickly identify rumor spreading trends and generate coping strategies. In addition, using the large language model to generate "rumor refuting" content, it can effectively control the spread range of rumors in the early stage of communication (Chen et al., 2021).

#### 2.6 Contradiction between government decision-making and public expectations

The governments decisions in response to public events are inconsistent with public expectations, which often leads to secondary public opinion and aggravates social antagonism.

### 2.6.1 Decision lag and information asymmetry

In a public crisis, the government releases a delayed response due to the lag in the information collection and processing process, which contradicts the publics high expectation for instant information transparency (Zhou & Zhang, 2022). When there is a significant gap between public opinion demands and government actions, the publics trust in the government may decrease significantly (Wang et al., 2021).

### 2.6.2 Information release and public credibility construction

Public opinion research shows that the transparency and consistency of information release strategy is the key factor to enhance the credibility of the government. Kim et al. (2022) found that the real-time release of official statements and positive interaction with the public can significantly alleviate the publics emotional response.

### 2.6.3 Technology-assisted decision optimization

The government public opinion monitoring system, which combines big data analysis and machine learning technology, can effectively identify public demands and help optimize policy design. Intelligent response systems based on large language models have been suggested to be used in multi-channel government communication to improve response efficiency and accuracy (Huang et al., 2023).

## 2.7 Study blank

Despite the important progress in related research, there remain the following shortcomings:

1. Short text and multimodal semantic understanding: the current semantic fusion analysis of cross-platform short text and multimodal data is still imperfect.
2. The dynamic relationship between government decision-making and public opinion response: insufficient research on how the government balances its own goals with public expectations, especially the practice in a complex social environment.
3. ptimization strategies for enterprises to deal with network rumors: The current research pays little attention to how to balance the relationship between brand reputation protection and the continuity of business activities, which needs to be discussed in depth.
4. Fairness and ethics of public opinion governance: the application of existing technologies against public opinion control, and the long-term observation of social and commercial impact lacks further research.

# **CHAPTER 3**

# Research Methodology & Design

## 3.1 Introduction

This section introduces the methods and applications that will be used in the study. The overall framework is formulated according to the ACM/IEEE data science life cycle model. First, the project's goals and key issues are clarified. After that, data is collected. This project will write a crawler to obtain data from three different social media and perform data preprocessing. In the exploratory data analysis stage, Chinese word segmentation and new word discovery models will be used in combination with statistical methods and visualization technology to conduct comprehensive text statistical analysis. Finally, this study will use the transform model to perform sentiment analysis on multi-platform network social media text data.

## 3.2 The framework

### 3.2.1 Data Science Life Cycle

This method adopts a structured data science project life cycle mechanism to promote the reasonable and efficient completion of data collection, analysis and model creation. There are many definitions of the life cycle of a data science project. This study adopts the ACM/IEEE data science life cycle model, in which the seven key stages are: problem identification, data collection, data preparation, data analysis, modeling, and evaluation and deployment.

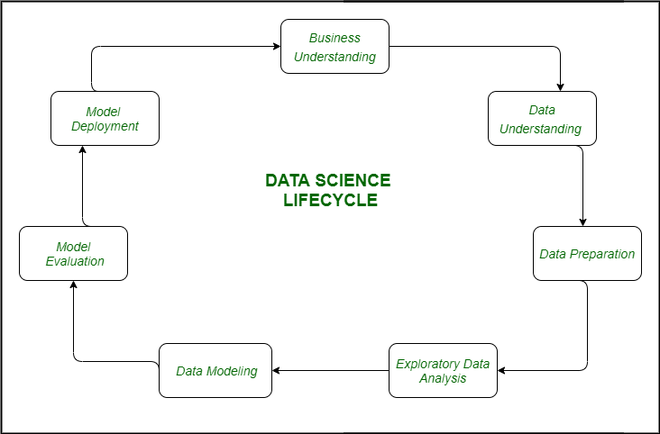


Figure 3.1 Data Science Life Cycle

1. Problem Identification

The definition actually refers to clarifying the goals and key issues of the project. This phase involves communication with stakeholders to understand needs and expectations. The core goal of this study is to perform sentiment analysis on text data of the same event on multiple social media. This goal can be broken down into two parts: data acquisition of the same event on different social media, and sentiment analysis of short online text data.

1. Data Collection

The definition actually means collecting the required data, including internal data (such as company databases) and external data (such as public data sets). At this stage, pay attention to the source of the data and how to obtain it. The data collected in this study are the content posted by users of Weibo, Xiaohongshu, and Bilibili, as well as the comments and replies to the related content. All of the above data are public data from social media, but there is no data source directly used for research, so crawler technology is needed to obtain data. The main implementation of this study is implemented in Python, and the main technical tools are the requests library and the Scrapy crawler framework. And use CSV files and MongoDB database for storage.

1. Data preparation

The definition actually means cleaning and transforming the collected data, including processing missing values, removing outliers, and converting data formats. This is an important step to ensure data quality. The data obtained in this study is network text data. The original data contains special tags and hyperlinks, which need to be processed. At the same time, since the data comes from multiple platforms, the data format needs to be unified. The specific tool is Pandas in Python.

1. Exploratory Data Analysis (EDA)

Conduct exploratory data analysis (EDA) and use statistical methods and visualization tools to understand the basic characteristics and structure of the data. This helps identify potential patterns and relationships. Chinese text analysis projects need to use models for new word discovery, Chinese word segmentation, and stop removal at this stage. Then perform data analysis based on specific analysis goals. The visualization tools used during this period are Echarts and Matplotlib.

1. Data Modelling

Modeling means selecting and applying appropriate algorithms to build models, such as regression, classification, and clustering. At this stage, model evaluation and adjustments are also performed to optimize performance. The model finally established in this study is a sentiment analysis model. For supervised deep learning, it is necessary to select a suitable training data set. There are many specific implementation methods, such as time series models such as LSTM and its variants, Transformers series models, or LLM models. Sentiment analysis belongs to a specific field of text classification. Generally speaking, depending on the application scenario, the Transformers series model and the LLM model have the best performance respectively. This model will be trained based on one of the Transformers series models to obtain an optimal model for sentiment classification in the field of Chinese online short texts.

1. Model Evaluation

Different evaluation indicators (such as accuracy, precision, recall, F1 score, etc.) are used to evaluate the effect of the model. Cross-validation and other methods are also used to ensure the robustness of the model. The main evaluation criteria for this study are accuracy, precision and recall. Specific evaluation criteria are selected based on the different task types of word segmentation, new word discovery and sentiment analysis, as well as the extreme performance of the model.

1. Deployment

Apply modeling results to actual business or projects. This includes model launch, monitoring, and integration with other systems. The final model deployed this time needs to be expanded with a long-term training data set after deployment, and the model needs to be continuously adjusted based on the prediction results. The data collection program needs to continuously update the crawler program according to the changes of social media websites.

Figure 3.2 Research Framework(Complete framework will be added after the project is completed)

### 3.2.2 Problems to be solved

The primary task of this study is to build a sentiment analysis model for Chinese online short texts. The figure below is a simple flow chart for data acquisition, processing and modeling. To obtain truly effective conclusions, there are still some issues that need to be addressed in each specific link.

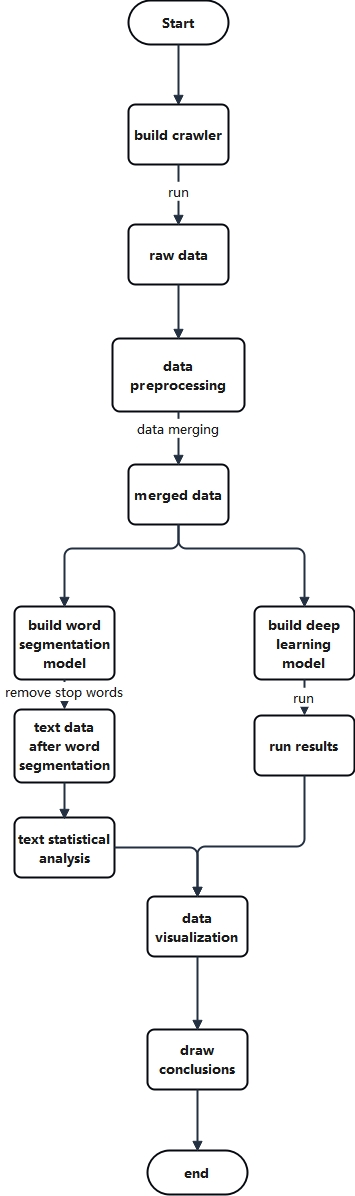


Figure 3.3 Simple data flow chart

1. Data Collection and Quality

* Web crawler to directly crawl web text data will result in some unnecessary data, such as various icons, emoticons and other special texts, as well as special words such as replies and topics that belong to the main text. These contents are collectively referred to as markers. In addition, there are various network links that do not belong to the main text, so these contents need to be identified and processed.
* The data comes from multiple platforms, so the data format needs to be unified and the final data needs to be merged into the same table.

1. Chinese word segmentation and stop word removal

* Chinese text has no spaces between characters, so a word segmentation model is needed for word segmentation. However, the word segmentation models in different fields vary greatly, so a reasonable word segmentation model needs to be built. Here we choose the COARSE\_ELECTRA\_SMALL\_ZH model in HanLP as the word segmentation model. The original algorithm of this model is Electra (Clark et al. 2020) small model trained on coarse-grained CWS corpora. Its performance is P: 98.34% R: 98.38% F1: 98.36% which is much higher than that of MTL model. The corpus comes from the Chinese corpus constructed by HanLP, which mainly comes from Chinese published text data, with a total corpus of more than 100 million.The figure below is a flowchart of sentiment analysis, where the input is "这里是测试文本", It means "Here is the test text".

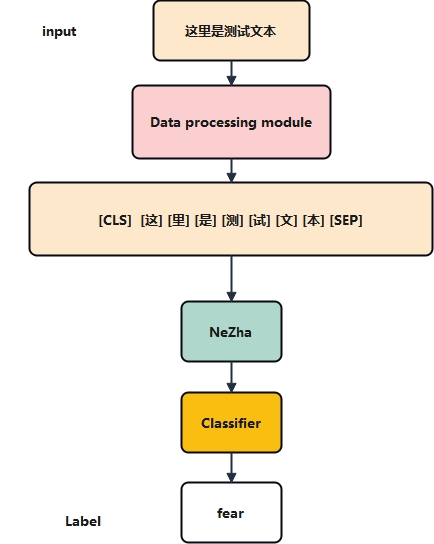


Figure 3.4 Flowchart of sentiment analysis

* Sources of stop words: Different languages have some special words such as textual expressions or modal particles that have no practical meaning. Stop words come from a wide range of sources, and there are certain differences in stop words in different corpora. This stop word list is constructed by me (Tian Fengshou) based on the stop words of Baidu, Google and Sogou, combined with the actual situation of online short texts. It is mainly aimed at online texts, and some commonly used symbols and meaningless words are added.

1. Building a Sentiment Analysis Model

* Construction and selection of training data set: The actual prediction ability of supervised deep learning models is highly related to their training data, and the data we are going to process this time is short text data on the Internet. This is a sentiment analysis in a special scenario, which is different from the traditional sentiment analysis scenario. The main scenario of traditional sentiment analysis is long text. Therefore, we need to build our own training data set according to the actual situation. Here we choose SMP2020 as our training data set. The annotated data set used in this technical evaluation is provided by the Social Computing and Information Retrieval Research Center of Harbin Institute of Technology. The original data comes from Sina Weibo and is provided by the Micro Hotspot Big Data Research Institute. The data set is divided into two parts. The first part is the general Weibo data set. The Weibo content in this data set is randomly obtained from Weibo content, not targeting specific topics, and covers a wide range. The second part is the epidemic Weibo data set, which is not applicable. Each Weibo is labeled as one of the following six categories: neutral, happy, angry, sad, fear, surprise. The general Weibo training data set includes 27,768 Weibo posts, the validation set contains 2,000 Weibo posts, and the test data set contains 5,000 Weibo posts.
* Selection of pre-trained models: In the field of natural language, if you want to obtain a better model on a small amount of data, the best way is to pre-train on a basic model. This involves how to choose a basic pre-trained model. After multiple rounds of testing, we finally chose the NAZHA model provided by Huawei as the final pre-trained model.

### 3.2.2 Future Development

More training data. Currently, there is still little training data, and the performance of the algorithm has not been fully explored due to the lack of training data. Therefore, it is necessary to further increase the training data in the future to improve the performance of the final model.

Correlation tracking of public opinion events: The spread of an event on multiple platforms has a time relationship. This study did not conduct in-depth research on this aspect.

## 3.3 Data sources and collection methods

Since the data comes from multiple platforms: Weibo, Bilibili, and Xiaohongshu. Therefore, the crawler here is built according to the actual situation. The Weibo data platform is built using the Requests and Scrapy frameworks. Since the platform has two network entrances cn and com, com has login restrictions, so cookies need to be added manually. cn has no login restrictions, and the cookie-free version can be used directly to obtain the required data. The data acquisition of the Xiaohongshu platform also requires cookies to obtain complete data, but the difference between the cookie-free version is not large, so the cookie-free version can also be used directly. The bullet screen and comments of Bilibili do not need to be logged in, so the cookie-free version is used directly.

All downloaded data is stored in csv files and saved directly locally. The crawled content includes text, comments, relationships, and users.

The text data is the content officially released by the user. The official content of Bilibili is mainly video, and only links are saved. Weibo videos and Xiaohongshu videos are processed in the same way. Comment data refers to all comments and replies under the official content. User data refers to the personal information of different users. Considering that the data is not directly related to this study, the three platforms use different tables for storage. No data merging is performed. Relationship data refers to the relationship between followers and followed persons. This study does not use this data directly, so it is only stored and not processed.

## 3.4 Data preprocessing

### 3.4.1 Data merging

User data and relationship data are only stored, so they are not merged. The text and comment content of the three platforms are not much different, so they are directly stored in a unified format.

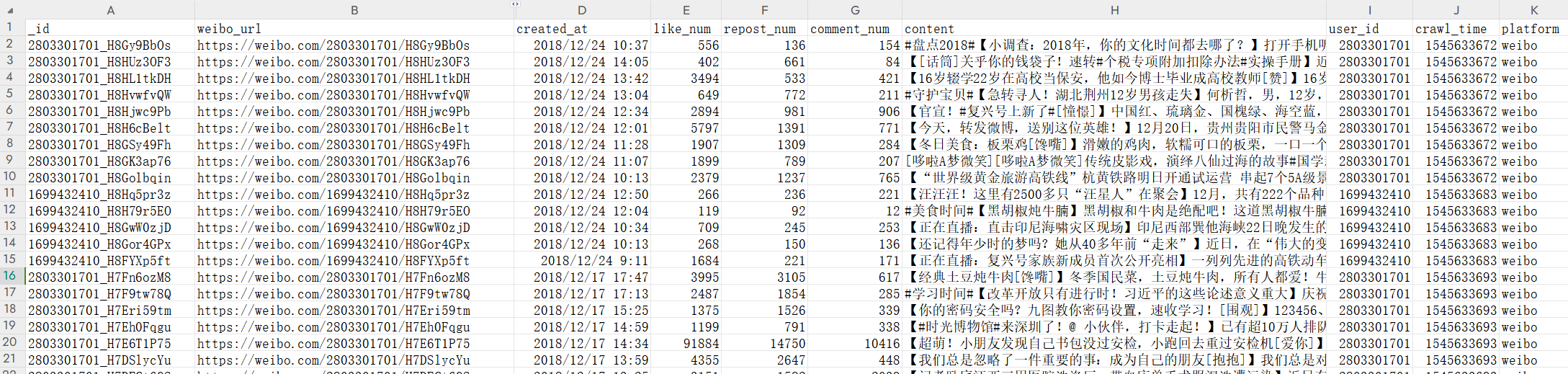


Figure 3.5 Screenshot of the text data

The fields in the body are: \_id, url, created\_at, like\_num, repost\_num, comment\_num, content, user\_id, crawl\_time, platform. Among them, \_id is the unique identifier, url is the body link, created\_at is the creation time, like is the number of likes, repost\_num is the number of reposts, comment\_num is the number of comments, content is the body content, user\_id is the publisher id, crawl\_time is the crawl time, platform is the publishing platform.

The content of the comment data is similar, but only contains \_id, comment\_user\_id, content, url, created\_at, crawl\_time, platform. Among them, comment\_user\_id is the id of the commenter, and url is the id of the content being commented on.

### 3.4.2 Data cleaning

Data cleaning of online text mainly removes text data that does not belong to the main content. For hyperlinks, we can directly use regular expressions to remove them, or use the hyperlink recognition built into the word segmentation model to remove them. In addition, Weibo and Xiaohong also contain five special markers that need to be processed separately.

1. One is the additional information in the comment, For example, this example comes from actual data obtained, "回复@\*:", similar to the following: "回复@齁甜齁甜的彼得潘:我知道你们这些都是海归或是海归的父母，搞笑国内没有好大学？ \和你们这些把国外的[大便]都奉若珍宝人有什么好说的！！！"
2. One type of content is @user. For example, this example comes from actual data obtained, "nihao @dfugo @jb51 haha"
3. One category is emoticons and icons, For example, this example comes from actual data obtained, "铭记历史，勿忘国耻。老爷爷，您一路走好[蜡烛]。" In the original data of Weibo, emoticons are stored in a format similar to `[text]`.
4. Title, Weibo and Xiaohongshu titles will also be directly obtained. For example, this example comes from actual data obtained, "【汪汪汪！这里有2500多只“汪星人”在聚会】12月，共有222个品种，超过2500只小狗参加第二届波兰卢布林国际犬展。戳视频，一起来看它们的萌样子～[下]️[下]️[下]️新华视点的秒拍视频"
5. Topics, Weibo or Xiaohongshu content is sometimes related to some internal topics. For example, this example comes from actual data obtained, "【官宣！#复兴号上新了#[憧憬]】中国红、琉璃金、国槐绿、海空蓝，复兴号大家族再添高颜值新成员！时速350公里17辆长编组、时速250公里8辆编组、时速160公里动力集中等多款复兴号新型动车组首次公开亮相，TA们既有颜值，更有内涵，最快在明年1月5日上线！你期待吗？@中国铁路"

Among the five markers, 1, 2, 4, and 5 can be removed using regular expressions, as shown in the code diagram.

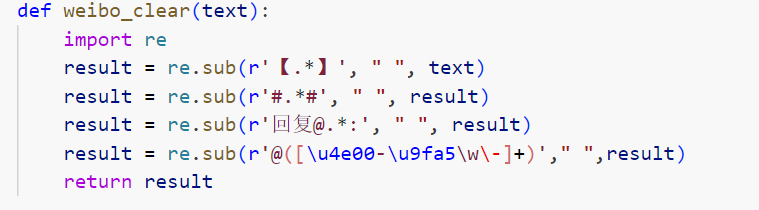


Figure 3.6 Code to remove markers

# However, the formats of emoticons and icons are special and may be mixed with normal symbols. The two tasks of this study are text statistical analysis and sentiment analysis. In sentiment analysis, the original training data contains emoticons, so they do not need to be removed. In the text statistical analysis stage, emoticons need to be specially processed. Therefore, we have two options. One is to directly remove them with regular expressions, and the other is to build an emoticon discovery system for statistical analysis. The second method needs to be combined with a word segmentation model, so it will be explained in detail in the next chapter.**CHAPTER 4**

# Initial Findings

## 4.1 Data Exploration

After the data is merged, if the text needs to be processed, two types of models are needed. Traditional statistical models need to remove meaningless words, so for this type of model, the emojis in the data need to be removed. For deep learning models, the original data is directly retained, so for deep learning models, the original data also has a practical role.

At the same time, due to the large amount of data obtained, only one independent event, the data related to the "Yibin Earthquake", was selected for subsequent operations. After screening according to the subject and keywords, 55,570 pieces of data related to the event were actually obtained.

### 4.1.1 Statistical analysis of emoticons

First, we use the emoji discovery system we built earlier to obtain all the emojis used in the dataset, and then we perform statistical analysis on the emojis we found. We find that the most commonly used emoji is "祈祷" (which means "prayer" in English), and the top 20 usage frequencies are as follows:

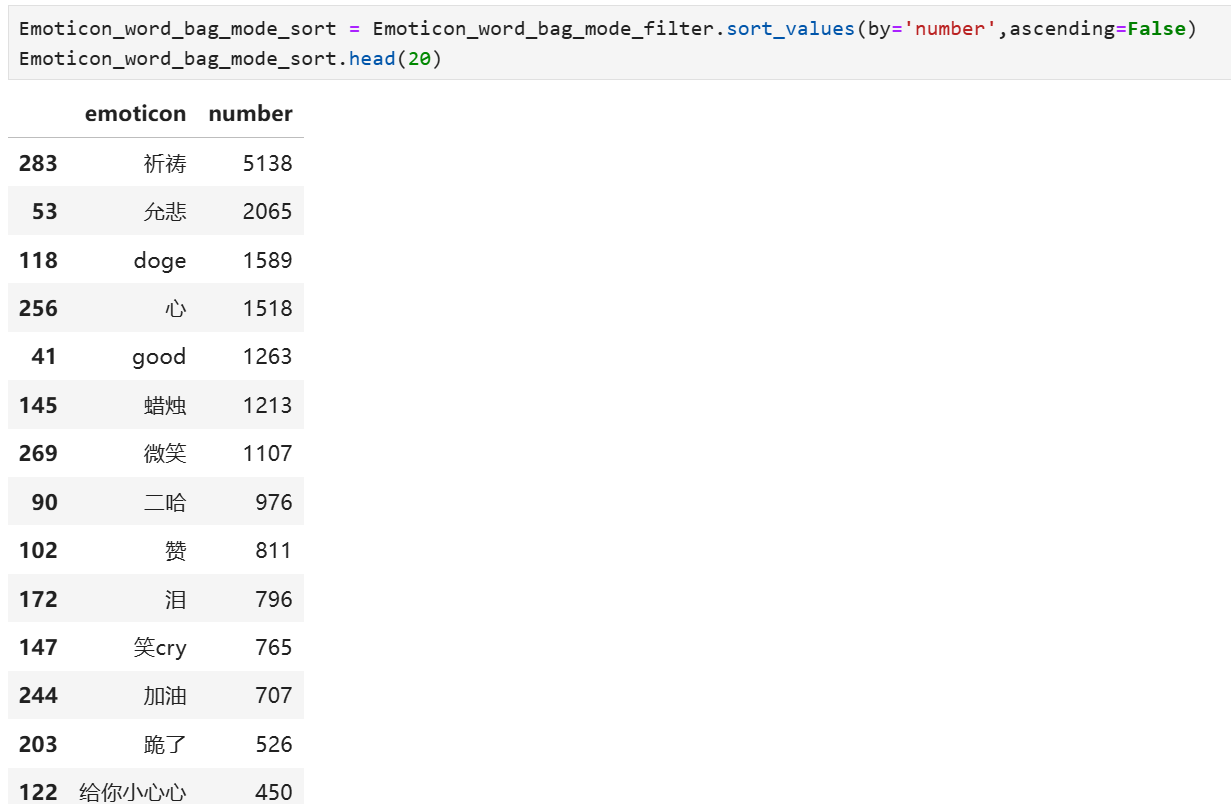


Figure 4.1 Number of emoticons used

Using a bar chart to visualize the data, the results are as follows:

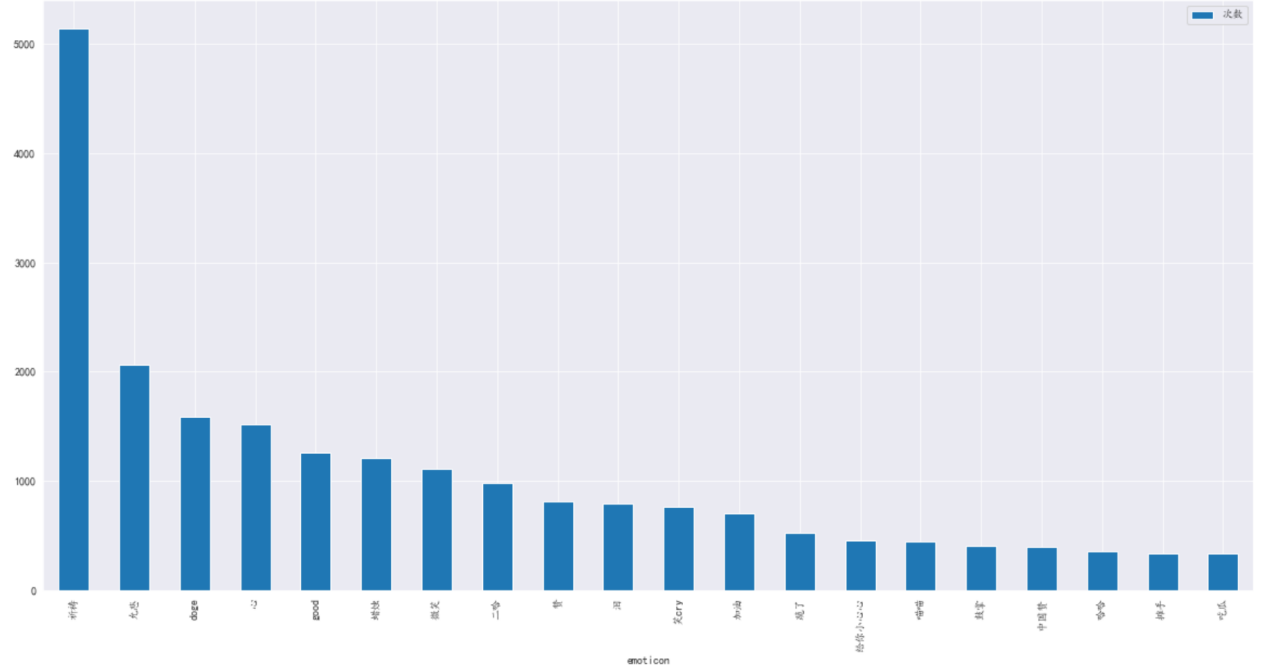


Figure 4.2 Histogram of the number of emoticon packages used

After analyzing the data set, the user ID that used the most emojis was 3853279141, who used the most emojis in this offline analysis, with 52 emojis. A bar chart was used to plot the number of emoji usages, and the results are shown in the figure:

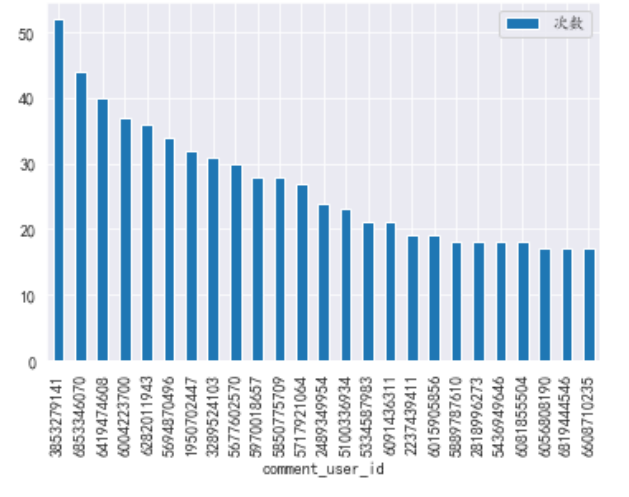


Figure 4.3 Statistics on the total number of emoticons used by users

By counting the number of posts by the id, we can calculate the frequency of each user using the emoji package. We found that the user who used the emoji package the most was user 5694870496.

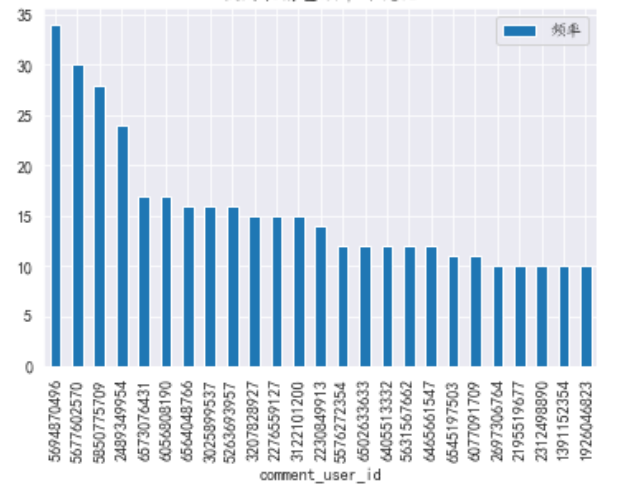


Figure 4.4 Statistics on the frequency of users' emoji usage

### 4.2 Word segmentation model

The word segmentation model directly uses the structured perceptron model in HanLP, which is trained from a large comprehensive corpus of 99.7 million words, covering multiple fields such as news, social media, finance, and law. It is the largest Chinese word segmentation corpus in the world. The size of the corpus determines the actual effect. The corpus for production environments should be in the tens of millions of words. Natural semantic linguistics experts have been continuously annotating the corpus to keep up with the times and maintain the most advanced word segmentation quality. In terms of word segmentation standards, HanLP provides two granularities: fine-grained and coarse-grained. Fine-grained is suitable for search engine business, and coarse-grained is suitable for text mining business. This study uses coarse-grained word segmentation according to actual needs.

### 4.3 Word cloud

The data set after word segmentation is clean content words. By statistically analyzing the data, a word cloud can be obtained. By specifying the shape of the word cloud as a heart shape and specifying the color, a word cloud that fits this earthquake event can be obtained.



Figure 4.5 Colorful love word cloud



Figure 4.6 Red love word cloud

From the statistical analysis, we can see that the most used word is "Sichuan", which is the place where the earthquake occurred, followed by "earthquake", which means earthquake. The words that are used more frequently are "bless (保佑)", "hope (希望)", "peace (平安)", etc. It can be basically seen that the content and comment areas of various social media platforms are mainly blessings for this incident.

### 4.4 Sentiment Analysis

Due to the particularity of the selected events, the data inherently contains both negative and positive comments, and different algorithms can be used for sentiment analysis. DistilBERT trained on the SST-2 dataset was first added with synthetic multilingual data generated by advanced LLMs, and then fine-tuned. The final test dataset accuracy was about 93%. The prediction result of the model is label + score. The model was used to perform statistical analysis on the collected dataset, and the final prediction results are as follows:

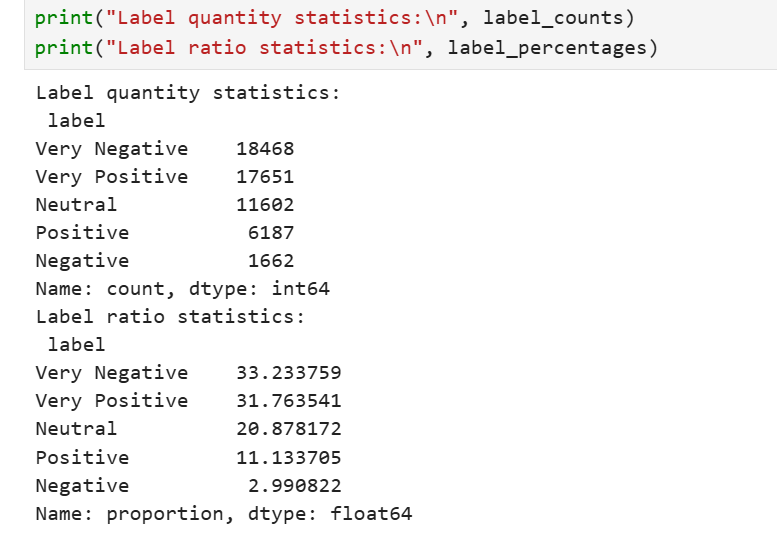


Figure 4.7 Tag statistics results

Data visualization technology can also be used to well display the distribution of sentiment analysis results.

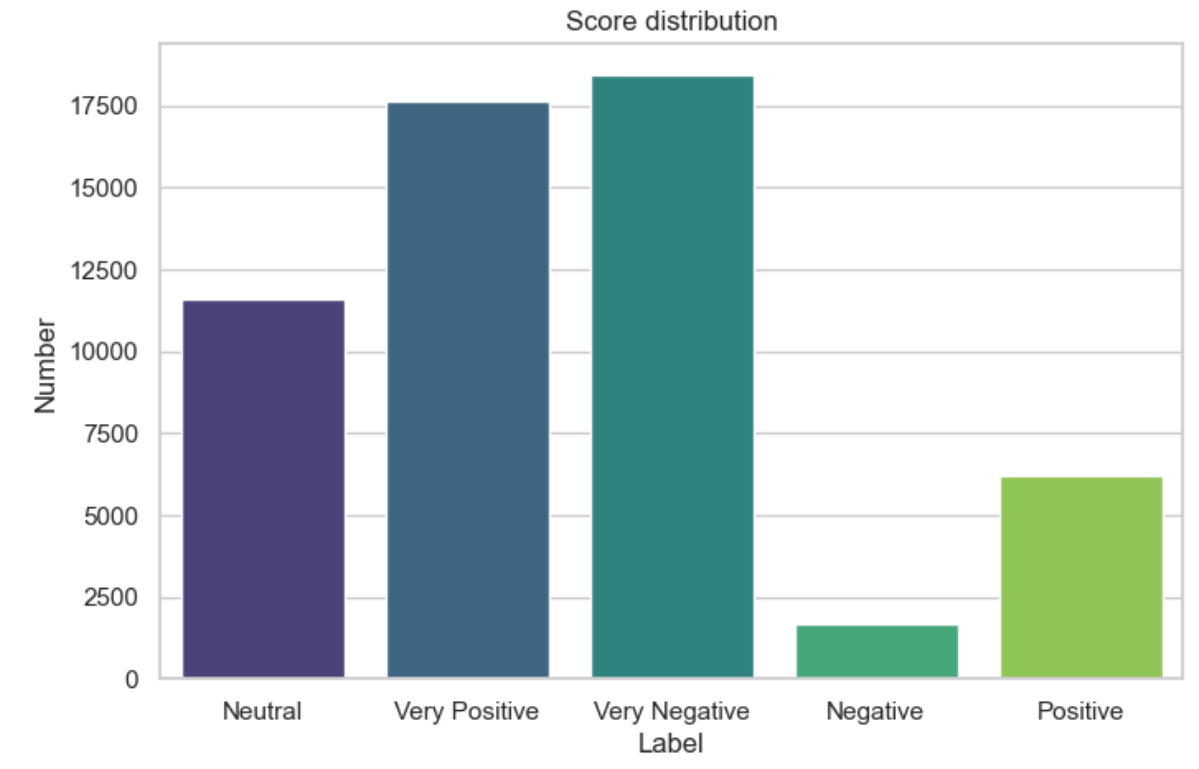


Figure 4.8 Score distribution histogram

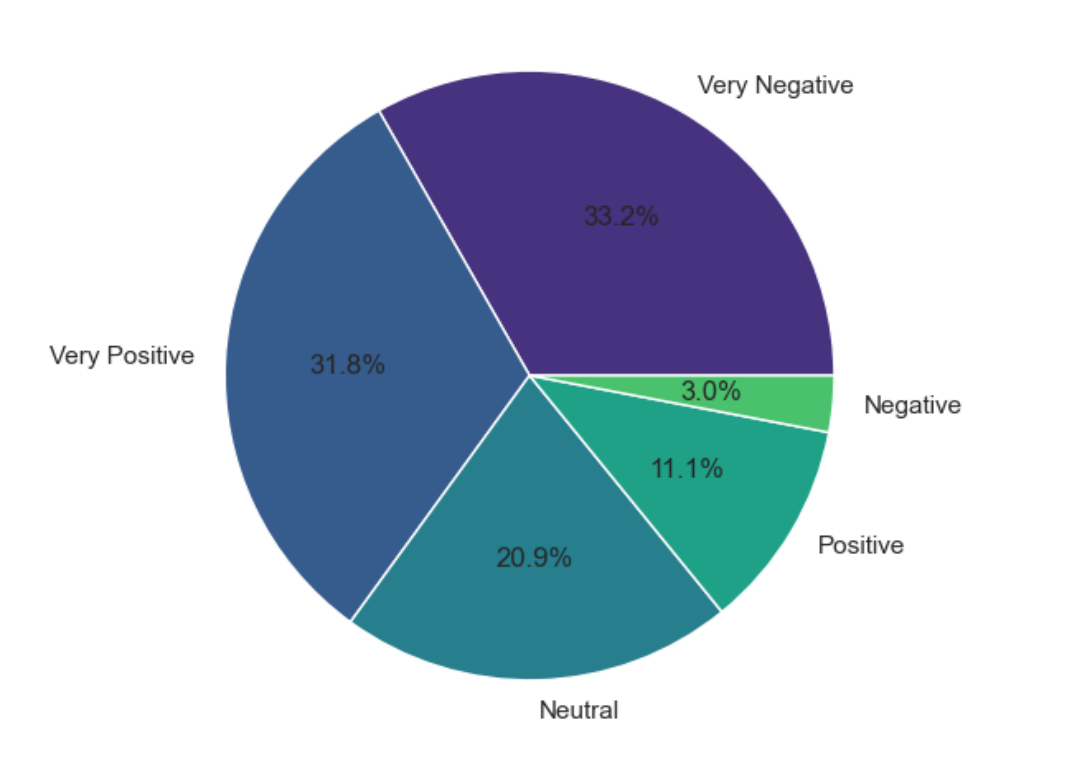


Figure 4.9 Label ratio pie chart

### 4.5 Summarize

Based on the emoji discovery system, we conducted statistical analysis of emojis, analyzed the frequency of users using emojis, and preliminarily implemented sentiment analysis of cross-platform data of a single event based on the sentiment analysis model. The entire process can basically achieve the design goal.

# **CHAPTER 5**

# CONCLUSION AND RECOMMENDATIONS

## 5.1 Summary

This study completed three core tasks: (1) Building a crawler that can crawl Weibo, Xiaohongshu, Zhihu, and Bilibili. (2) Building a data cleaning and conversion system that can merge data from multiple social media. (3) Building a word segmentation and sentiment analysis model to perform word segmentation and sentiment analysis on the acquired data. In this process, an emoji discovery system was also built that can automatically discover emojis based on text format and vocabulary frequency.

In the end, the overall project basically achieved the design goal, and the project can provide effective support for subsequent research and analysis related to Chinese social media.

## 5.2 Future Works

There are three directions for improvement and enhancement in future work: (1) Optimizing data acquisition capabilities. The current concurrent performance of crawlers is average. Compared with the hundreds of millions of active users of Chinese social media, the data acquisition capability is too weak and cannot fully acquire real-time hot data. Therefore, the concurrent performance of the current crawlers needs to be further improved in the future. This requires starting from the account pool and proxy pool to build a large-scale distributed crawler system that can acquire a large amount of social media data.

1. Improving the prediction ability of sentiment analysis models. Currently, mainstream sentiment analysis models are trained on large-scale data sets and then perform transfer learning. However, the data sets used for transfer learning are relatively small. Therefore, in the future, native data sets of Chinese social media should be built to fully explore the learning ability of the model. Secondly, we can start from specific means in the field of sentiment analysis such as sentiment dictionaries to further improve relevant algorithms.
2. For hot events on social media, there is a propagation mechanism between multiple platforms, but this study did not involve research on related content. In the future, we can combine professional knowledge such as communication to study the mechanism of event propagation, so as to build a more complete event tracking and analysis mechanism.

References

Bai, X., Chen, Y., & Zhang, Y. (2022). Graph pre-training for AMR parsing and generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 6001–6015). Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.415

Cen, Z., Zhang, J., & Liu, Y. (2021). A framework for multi-source data fusion in cross-platform sentiment analysis. Journal of Computational Social Science, 4(2), 130-142.

Chae, J., Kim, Y., & Lee, S. (2022). The economic impact of online rumors on brand trust and consumer behavior. Journal of Marketing Research, 59(3), 120–134.

Chen, H., Guo, C., Li, Y., Zhang, P., & Jiang, D. (2023). Semi-supervised multimodal emotion recognition with class-balanced pseudo-labeling. In MM '23: Proceedings of the 31st ACM International Conference on Multimedia (pp. 9556–9560). https://doi.org/10.1145/3581783.3612864

Clark, K., Luong, M., Le, Q., & Manning, C. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. International Conference on Learning Representations (ICLR). https://openreview.net/pdf?id=r1xMH1BtvB

Graham, S., Sweeney, P., & Alas, B. (2018). Data science lifecycle: A review and framework. In 2018 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 360-369). IEEE. https://doi.org/10.1109/DSAA.2018.00057

He, H., & Choi, J. (2021). The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 5555–5577). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.451

Huang, Y., Zhao, X., & Lin, W. (2023). AI-powered public opinion management in governance. Journal of Public Administration, 9(1), 102-125.

Kim, H., & Park, S. (2021). The role of social media in corporate reputation management: A sentiment analysis approach. Journal of Business Research, 131, 352–367.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Departmental Papers (CIS).

Li, X. (2023). Current status and research findings of internet public opinion in China. Office Automation, 28(2), 16–18. https://d.wanfangdata.com.cn/periodical/bgzdh-bgsbygc202302005

Li, X., Liu, J., Li, S., & Zhao, X. (2020). SMP2020-EWECT: A dataset for Weibo sentiment classification. Proceedings of the China National Conference on Social Media Processing (SMP 2020). https://smp2020.ewect.com

Li, X., Wang, Y., & Chen, L. (2023). The ripple effect of food safety rumors on supply chain management. Supply Chain Management Review, 29(2), 34–49.

Rahman, M., Hassan, K., & Sarker, F. (2021). Financial rumors and stock market volatility: Evidence from emerging markets. International Review of Financial Studies, 14(1), 77–92.

Shin, S., Ryu, W., Park, S., Cho, K., & Lee, D. (2021). Effective sentence scoring method using BERT for speech recognition. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 5554–5566). https://doi.org/10.18653/v1/2021.emnlp-main.451

Wang, C., & Xu, B. (2017). Convolutional neural network with word embeddings for Chinese word segmentation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Vol. 1, pp. 163–172). Asian Federation of Natural Language Processing. https://www.aclweb.org/anthology/I17-1017

Wei, J., Wang, X., Tian, H., Bai, Y., Lan, Y., Wu, J., Wang, S., & Li, X. (2019). NEZHA: Neural contextualized representation for Chinese language understanding. arXiv preprint arXiv:1909.00204. https://arxiv.org/abs/1909.00204

Zhou, P., & Zhang, Y. (2022). The challenges of government-public expectation alignment in crisis management. Asian Journal of Governance, 5(2), 87-104.