



Sentiment analysis in social media

by

Xiaoqi Quan

Supervisor: Dr Nik Whitehead

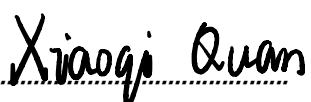
Project submitted as part of the requirements for the award of
MSc Computer Networks and Cyber Security

November 2021

DECLARATION

I, ...Xiaoqi Quan..... declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed :.....

Date : November 12, 2021.....

FORM OF CONSENT

I Xiaoqi Quan..... hereby consent that my Project, submitted in candidature for the M.Sc. Applied Computing degree, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed :.....

Date : ... November 12, 2021.....

Abstract

With the rapid development of the Internet, people all over the country are keen to express their views and opinions on something on social media. In the UK, people are accustomed to using Twitter to express their opinions, while in China, people are more inclined to express their opinions on Weibo. At the same time, with the popularization of mobile phones and the smoothness of the Internet, we can discover new things anytime and anywhere, and we can also express our opinions anytime, anywhere. Therefore, the information on social media has a strong sharing and real-time nature. At the same time, because it is information released by people's subjective wishes, people's views are more subjective and authentic. From social media, we can get almost any information we want. For example, the government wants to listen to people's opinions on the newly promulgated policies, buyers want to get people's comments on products, and movie publishers want to get public comments on movies. Almost all of them can be found on social media. Therefore, after obtaining social media information, it is very important to perform sentiment analysis on the text information of social media.

This project mainly conducts sentiment analysis of social media. Through the sentiment analysis of social media, the keywords of microblogs with different sentiments can be obtained. Analyze the key technical issues of social media, model the emotional analysis of social media, and predict the emotional tendency of the text information of social media. At the same time, through social media sentiment analysis in different languages, analysis methods and model choices suitable for different languages are obtained.

This project uses the Twitter and Weibo data sets for analysis and uses jieba to perform part-of-speech tagging and word segmentation on the Chinese data set. Use statistical methods to analyze the data set, use Tf-IDF and Bag of Words methods to extract features of the data, and use logistic regression and naive Bayes algorithm to classify text. Finally, the Bert pre-training model was used for social media sentiment analysis and achieved the best results.

Table of Contents

CHAPTER 1: INTRODUCTION	7
1.1 INTRODUCTION TO THE PROBLEM	7
1.2 AIM AND OBJECTIVES.....	8
CHAPTER 2 LITERATURE REVIEW	10
2.1 CONCEPT	10
2.1.1 <i>Emotion and sentiment</i>	10
2.1.2 <i>Sentiment analysis</i>	11
2.2 DATASET	11
2.2.1 <i>Twitter dataset</i>	12
2.2.2 <i>IMDB dataset</i>	12
2.2.3 <i>Weibo dataset</i>	12
2.3 DATA PREPROCESSING	13
2.3.1 <i>Format mark removal in data set</i>	13
2.3.2 <i>Remove stop words and stemming</i>	13
2.3.3 <i>Word segmentation</i>	13
2.3.4 <i>Part-of-speech tagging</i>	14
2.4 TECHNIQUES FOR SENTIMENT PROCESSING.....	14
2.4.1 <i>Method based on sentiment dictionary</i>	14
2.4.2 <i>Supervised learning approach</i>	15
2.4.3 <i>Deep learning</i>	16
2.5 FEATURE SELECTION.....	20
2.5.1 <i>Word</i>	21
2.5.2 <i>N-gram</i>	21
2.5.3 <i>TF-IDF</i>	21
2.5.4 <i>Bag of words model</i>	22
2.5.5 <i>Word2Vec</i>	22
2.6 EVALUATION INDICATORS.....	23
2.7 CONCLUSIONS	24
CHAPTER 3 RESEARCH AND DEVELOPMENT METHODOLOGY.....	25
3.1 RESEARCH METHODOLOGY	25
3.2 EXPERIMENTAL DESIGN.....	25
CHAPTER 4: DESIGN AND IMPLEMENTATION.....	28
4.1 UNDERSTANDING OF BUSINESS GOALS	28
4.2 DATA UNDERSTANDING	29
4.2.1 <i>Preliminary analysis of the data set</i>	29
4.2.2 <i>Data quality</i>	31
4.3 DATA PREPROCESSING	32
4.3.1 <i>Twitter data set</i>	32
4.3.2 <i>WeiBo data set</i>	34

4.4 DATA ANALYSIS.....	36
4.4.1 Twitter data set.....	36
4.4.2 Weibo data set.....	40
4.4.3 Conclusions.....	44
4.5 MODEL.....	44
4.5.1 Data set.....	44
4.5.2 Experiment.....	44
CHAPTER 5 RESULTS AND DISCUSSION	46
5.1 RESULTS	46
5.2 DISCUSSION	46
CHAPTER 6 CONCLUSIONS AND RECOMMENDATIONS.....	48
6.1 CONCLUSIONS	48
6.2 RECOMMENDATIONS.....	48
CHAPTER 7 REFLECTION	49
REFERENCES.....	50

List of Figures

Figure 1 The two-dimensional structure of emotion[1]	10
Figure 2 Classification algorithm process [20]	16
Figure 3 RNN model diagram.....	17
Figure 4 RNN model diagram [21]	17
Figure 5 Text-CNN model diagram [22]	18
Figure 6 The attention mechanism model [34]	19
Figure 7 Word2vec CBOW model [39][40].....	23
Figure 8 Word2vec skip-gram model [39][40]	23
Figure 9 The CRISP-DM model [50].....	26
Figure 10 The stages and their tasks	26
Figure 11 Negative emotional tweets	30
Figure 12 Positive tweets.....	30
Figure 13 Original Twitter data set.....	33
Figure 14 Two-attribute Twitter data set.....	33
Figure 15 Twitter data set after cleaning	33
Figure 16 Twitter data set after cleaning	34
Figure 17 Twitter data set after cleaning	34
Figure 18 Twitter data set after cleaning	34
Figure 19 WeiBo data set.....	35
Figure 20 The number of microblogs.....	35
Figure 21 The word segmentation results.....	36
Figure 22 The distribution of the number of emotional polarity labels.....	37
Figure 23 The sentence length distribution.....	37
Figure 24 The 30 most frequent words of negative attitude tweets.....	38

Figure 25 Negative emotion tweets word cloud.....	38
Figure 26 The 30 most frequent words of tweets with positive attitudes	38
Figure 27 Positive emotion tweets word cloud.....	39
Figure 28 Category distribution map.....	40
Figure 29 The sentence length distribution.....	40
Figure 30 Words with the highest frequency of occurrence.....	41
Figure 31 Word Cloud	41
Figure 32 Words with the highest frequency of occurrence.....	42
Figure 33 Word cloud.....	42
Figure 34 Words with the highest frequency of occurrence.....	43
Figure 35 Word cloud.....	43

List of Tables

Table 1 Confusion matrix [27]	24
Table 2 Fields in the Twitter data set	29
Table 3 Fields in the IMDB data set	30
Table 4 Fields in the Weibo data set.....	31
Table 5 The experimental results.....	46

Chapter 1: Introduction

1.1 Introduction to the problem

Tim Berners-Lee created the first website in 1991. With the rapid development of Internet technology, various products led by companies such as Google, Facebook, Amazon, Baidu, and Alibaba have become the mainstream of the current Internet world. It also makes the Internet the third largest media besides television broadcasting, newspapers, and magazines. It also makes people more and more accustomed to publishing information and obtaining information through the Internet, and more accustomed to publishing some information containing their own feelings on the Internet. Therefore, various text messages expressing emotions are widely distributed on the Internet, such as BBS, chat rooms, personal blogs, and social networks such as Weibo and Twitter.

With the rapid development of the Internet, we can obtain more and more information. On the online shopping sites, we can find a lot of experience about the same item that another person purchases before we buy it by ourselves. On the various blog sites, we can see the views of bloggers about one thing. On the social media, we can get viewing comments on a newly released movie which is published by other people, participate in discussions on current popular news events, and obtain feelings or views on any event. With this open concept of user-centered and user-participation, the role of the user has changed from a mere online reader to a network contributor and participant. This role change has brought about a blowout of the Internet's subjective sentiment tendencies.

Among them, Twitter is a very popular microblogging service that focuses on brief, real-time status updates. It attracts more and more people to express opinions and emotions on current hot topics. In the fourth quarter of 2019, the number of users that post a tweet in one day is 126 million. Users use the service provided by Twitter to create a kind of status information (tweet) to express their views on different topics. Twitter users can follow other users or be followed by other users without any authentication information. This has accelerated the speed of information dissemination.

Similarly, Weibo is a microblogging service in China. Weibo is a paragraph of text as the main content, while emojis, pictures, videos, etc. can be introduced to narrate and supplement the text. In this self-media era, Weibo plays a pivotal role in China. According to the latest data from the China Internet Center [28], as of June 2017, the number of Weibo users was approximately 291 million. Among the typical social applications in China, the number of users in Weibo ranks among the top three and is on the rise.

Compared with other social media, Weibo and Twitter have obvious advantages:

1. The form of release is simple and diverse. Twitter and Weibo can share their stories and publish their thoughts anytime and anywhere through the mobile client, and can also add emojis, pictures, videos, and links. The release form is simple and clear, and there are many ways to upload.
2. A wide range of information acquisition platforms. Twitter and Weibo cover the accounts of

various public figures, industry elites, news and information media, etc. Most celebrities and relevant industry platforms also have relevant accounts. We can also obtain all kinds of news on this platform in time and understand relevant information. By searching the topic that we are interested in on this platform, we can also get the general public's discussions and opinions on related topics.. Therefore, Twitter and Weibo can easily obtain the information that we want to know.

3. Information is updated and spread quickly and we can get the latest news.

In the era when the Internet is not popular, if government departments want to obtain citizens' opinions on a certain matter or film companies want to know the audience's evaluation and likes of a certain movie, they usually use questionnaire surveys or voting to obtain the information. This method only relies on manual tracking and analysis, so the work efficiency is very low, which not only will consume a lot of manpower and material resources, but the process of obtaining results will also be very long. With the birth of social media, these methods have gradually withdrawn from the stage. How to quickly obtain the information we need from short text data such as social media, and analyzing this information replaces the past methods.

Faced with such a large amount of unstructured emotional data, it is more and more difficult for people to obtain the required knowledge, it is also more and more difficult for people to obtain effective information, and it is also difficult for users to obtain an overall grasp of this information. Therefore, how to effectively process and analyze these massive amounts of information and quickly and accurately discover the required information has become more and more important.

In response to the above situation, sentiment analysis technology is more and more worthy of attention. Text sentiment analysis is to analyze the text and judge its sentiment polarity, that is, positive, negative, or neutral sentiment. Through sentiment analysis of massive social media data, we can understand the user's feedback on a certain event or a certain product, understand the user's true feelings and general opinions, to provide decision-making support to the service provider and give users better experience.

In summary, how to filter and analyze information on social media, and how to classify information on sentiment is of great practical significance. Therefore, this project selects Twitter and Weibo datasets for social media sentiment analysis.

1.2 Aim and objectives

The goal of this project is to perform sentiment analysis on social media texts. Its main task is to analyze the emotional tendency expressed in the text, that is, the positive or negative emotional tendency reflected in the text content. Since social media is a short text, this project focuses on sentence-level sentiment analysis. For example, "Today is a good day!", "A bit bad luck, today!" Through analysis, we can know that the first sentence is a positive emotional tendency, while the second sentence is a negative emotional tendency. Through sentiment analysis of the text, we can understand people's feedback on a certain event or a certain product, to better understand the real feelings and opinions of users.

The objectives of the project are as follows:

1. Research in the field of sentiment analysis. This objective is very important and understanding the related research and related theories of sentiment analysis will play an important role in the project.
2. Perform data analysis on the data set. Perform statistical analysis on the data set to better understand the data set and grasp the characteristics of related emotion types.
3. Analyze the emotional tendency expressed by the text, design different models, and conduct experiments. Research and analyze the experimental results and select the appropriate social media sentiment analysis model.
4. Compare the differences in sentiment analysis between Chinese and English social media.

Chapter 2 Literature Review

2.1 Concept

In this section, we will introduce some professional terms and concepts involved in sentiment analysis, as well as the definition of related concepts in the field of psychology.

2.1.1 Emotion and sentiment

At present, with the rapid development of Internet technology and the gradual popularization of smart phones, people increasingly rely on online social networks. At the same time, people's information exchange, sharing of life, and expressing their opinions and emotions are not limited to face-to-face communication, but more to express their thoughts on social networks anytime, anywhere. Social networks currently contain a large amount of unstructured data of user emotions. To better understand user emotions and behaviors, it is necessary to perform sentiment analysis on these unstructured data.

In daily life, we always confuse emotion and sentiment, but psychologists believe that they are different. Emotion is a more complex mental state, which is usually linked to actions⁰. Sentiment is a psychological attitude based on emotion [1]. Watson and Tellegen gave the famous mental model diagram, which shows the two-dimensional structure of emotions well, as shown in Figure 1[1]. It can be seen from the figure that emotions are divided into 8 categories, and then each category is subdivided into 3-6 categories, a total of 37 categories. In the field of psychology, emotion categories are

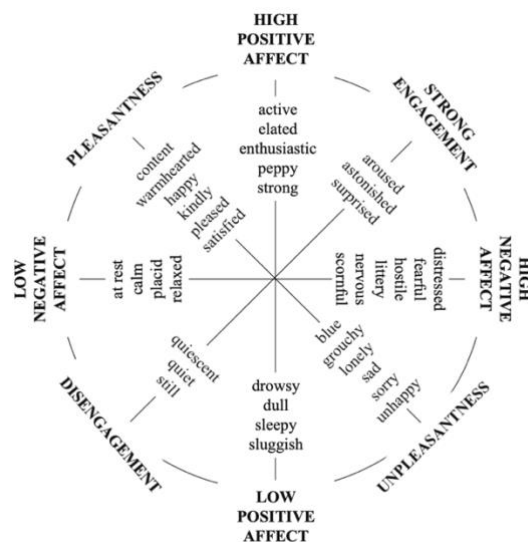


Figure 1 The two-dimensional structure of emotion[1]

more diverse and richer. American psychologist Ekman believes that there are six basic emotions, namely joy, sadness, anger, fear, disgust, and surprise[2]. Later other psychologists added two basic emotions, trust, and anticipation.

2.1.2 Sentiment analysis

Nesukawa et al.[4]proposed sentiment analysis and described the tasks and difficulties of sentiment analysis in detail. At present, Weibo and Twitter have become very popular communication tools among Internet users. Since the message format is unrestricted and the platform is easily accessible, people are increasingly inclined to switch from traditional communication tools such as traditional blogs, mailing lists, and telephones to these Internet social networking sites/software. Millions of messages appear on these websites every day. This information contains the author's life records, shared views on various events, and everyone will discuss together to form a topic circle. More importantly, more and more users will publish the products or services they use and give comments or express their political and religious views. With the user's permission, making full use of this information can be effectively used for marketing or social research. For example, sentiment analysis can allow customers to have a more comprehensive understanding of a product based on the buyer's emotions; at the same time, it can also allow businesses to better improve their products based on negative reviews. Sentiment analysis is also very important for political information. Before the European Parliament election, researchers will collect more than 1.2 million Twitter conversations (including three languages) for analysis within two weeks⁰. Therefore, sentiment analysis plays an important role in all aspects.

Researchers have given specific definitions of opinions/emotions. They believe that opinions/emotions need five parts, namely the name of the entity, the aspect of the entity, the sentiment on the aspect of the entity, the holder of the opinion, and the time of publication, which is expressed by a 5-tuple $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ [3]. [3] introduced and compared various views in sentiment analysis, including regular and comparative views, explicit and implicit views, subjectivity and emotion.

The current research on sentiment analysis can be classified from three perspectives: Task-oriented, Granularity-oriented and Methodology-oriented. Task-oriented mainly includes Polarity Classification, The Level of Valence or Arousal at a Specific Scale, Beyond Polarity, Subjectivity/Objectivity Identification and Feature/Aspect-based Sentiment Analysis. Granularity-oriented mainly includes Document Level Sentiment Analysis, Sentence Level Sentiment Analysis and Word Level Sentiment Analysis. Methodology-oriented mainly includes Supervised Learning Approach, Unsupervised Learning Approach and Semi-supervised Learning Approach⁰.

2.2 Dataset

This section will introduce the data sets used in this project. This project mainly uses three data sets, all of which are public data sets and do not involve privacy and other related issues. All three data sets can be used for sentiment analysis, two of which belong to the English data set (Twitter, IMDB), and one belongs to the Chinese data set (Weibo). Twitter and Weibo are social media sentiment analysis data sets, and IMDB is a movie review data set.

2.2.1 Twitter dataset

Sentiment140 can be obtained from <http://help.sentiment140.com/for-students>. The data set is in CSV format, and any traces of emoji have been deleted. The data set contains a total of 1,600,000 tweets extracted from the Twitter API. This data set labels 1,600,000 tweets as two emotions, with 0 representing negative and 4 representing positive.

This is a public data set that mainly collects tweets that people publish on Twitter. This data set performs preliminary data cleaning on tweets, deleting information such as emojis, and at the same time, this data set has sentiment annotations for each tweet. The sentiment tagged tweets can well verify the model's ability to perform sentiment analysis.

In social media, the information people publish often contains their own opinions and experiences, and this information often contains rich emotions. A reasonable sentiment analysis of this information can provide guidance for merchants and consumers in marketing strategies and consumption choices. At the same time, it can also provide convenience for the government and other official departments to understand the public opinion and public opinion orientation.

2.2.2 IMDB dataset

[5]published the data set used in the experiment, which provides a movie review data set, including 50,000 IMDB movie reviews, of which 25,000 are used as the training set and 25,000 are used as the test set. At the same time, the data set is a balanced data set with the same number of positive and negative. At the same time, the data set also provides 50,000 unlabeled movie reviews for unsupervised learning.

This is a public data set, collected mainly for people evaluate the film. And the data set has sentiment annotations for people's evaluations.

The distributor can predict the box office based on the audience's comments during the movie's release, and the audience can also judge whether to watch the movie based on the comments. Filmmakers can also analyze the strengths and weaknesses of movies based on audience comments and make better movies.

2.2.3 Weibo dataset

This is a Chinese data set. The Weibo dataset can be obtained from <https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/>. The data set contains more than 360,000 Sina Weibo, which contains a total of 4 emotions, 0 means joy, 1 means anger, 2 means disgust, and 3 means depression. Among them, there are about 200,000 for joy, and about 50,000 for anger, disgust, and depression respectively.

This is a Chinese data set, which is like the Chinese version of Twitter. Chinese users will post on Weibo some of their daily life and their views on some things. This data set performs more fine-grained sentiment analysis on user Weibo, which can better classify user sentiment.

The sentiment analysis of Weibo is as important as the sentiment analysis of Twitter.

2.3 Data preprocessing

In this section, we will introduce how to preprocess data. Data preprocessing is one of the basic tasks of sentiment analysis or natural language processing. Data preprocessing is to convert text into an intermediate form suitable for text processing and remove some irrelevant features, such as removing format tags in the text, filtering useless information, and removing stop words.

2.3.1 Format mark removal in data set

Some documents in the data set are stored in the form of web pages, and some are stored in the form of hypertext language. Documents stored in the form of web pages need to filter some useless information, such as URL links, Twitter usernames are the string following @, and special characters in Twitter such as "RT" and other abbreviations used exclusively by Twitter software. [6]

2.3.2 Remove stop words and stemming

Stop words generally refer to function words in the language. These stop words appear many times in the text but have no practical meaning. Generally called function words in Chinese, such as "的", "地", "啊", "吗" and so on. In English, it includes articles, pronouns, auxiliary verbs, copulas, etc.[18].

Stemming only means removing the prefixes and suffixes of words in English, and only keeping the stem parts of the words. Words in English often have prefixes, roots and suffixes. In sentences, words also have part-of-speech changes in gender, number, case, and tense, so they need to be stemmed[18]. Commonly used tools are Stanford coreNLP[8]

2.3.3 Word segmentation

Word segmentation is also an important step in preprocessing.

In English, the clear division between words and strict grammatical norms make English text processing more convenient. It is worth noting that it is necessary to make sure that some abbreviations are not divided, such as I'm, I'll, etc. Finally, some useless stop words in the text need to be removed, such as "a", "an" and so on[6].

In Chinese, there is no obvious boundary segmentation between words, which makes Chinese text

processing must first do Chinese word segmentation. For example, "我(I)/喜欢(like)/中国(China)". There are three methods for Chinese word segmentation: understanding word segmentation, string matching word segmentation, and statistical word segmentation. The string-matching method is to use a dictionary library to match strings. The more words in the dictionary library, the higher the accuracy of word segmentation. The comprehension method is to use the rule base to understand the string. The larger the rule base, the higher the accuracy of word segmentation, but the time complexity of processing the rule base algorithm is very high, and the speed of word segmentation is slower than the other two speeds. Statistics-based word segmentation requires a dictionary database for statistical training to obtain the probability of various words, which can be used as the basis for word segmentation. The size of the dictionary database and the size of the training set will affect the accuracy of word segmentation[18]. At present, the more popular word segmentation tools are jieba word segmentation[7] and Stanford coreNLP [8].

2.3.4. Part-of-speech tagging

The basic methods of part-of-speech tagging are rule-based part-of-speech tagging and statistics-based part-of-speech tagging. The rule-based word tagging method mainly relies on the context to judge the concurrent words. The part-of-speech tagging method based on statistics is to tag the part-of-speech based on the co-occurrence probability. It is generally believed that combining statistical methods and rule methods is the best way to solve the problem of part-of-speech tagging. The commonly used tool is Stanford coreNLP [8].

2.4 Techniques for Sentiment Processing

In this section, commonly used methods in sentiment analysis will be introduced.

2.4.1 Method based on sentiment dictionary

The method based on the sentiment dictionary needs to use a dictionary with sentiment labeling for different words and perform sentiment analysis based on the sentiment labeling words provided by the dictionary. This method first needs to preprocess the input text, then perform word segmentation, match the segmented words with the words in the dictionary, and finally output the results of sentiment analysis according to the rules of sentiment judgment. Most emotional dictionaries are artificially constructed. It takes a lot of cost to build an emotional dictionary. The tagger needs to read a lot of related materials and existing dictionaries. Based on this knowledge, combined with the tagger's understanding, summarize the words with emotional inclination, and mark the emotional degree of these words. SentiWordNet[11] is the earliest sentiment dictionary. It puts words of the same meaning together according to WordNet and marks the emotional tendency of these words. The dictionary mainly marks the positive or negative polarity scores of the words[10]. Chinese sentiment dictionaries mainly include NTUSD[12], HOW NET[13] and sentiment vocabulary ontology database[14]. These three dictionaries contain different numbers of Chinese praise and derogatory words[10].

2.4.2 Supervised learning approach

2.4.2.1 Rule-based algorithm

Rule-based algorithms, rules are formulated differently according to different emotions, and then match the rules according to the input text. The emotions corresponding to the matching rules are the emotions of the input text. Asghar et al.[15]analyzed the semantics of words, divided the words into +ive and -ive categories, and then identified and classified the expressions, modifiers, general words, and domain-specific words expressed in the public's feedback on the product0. This method solves the problem of low efficiency caused by data sparseness.

2.4.2.2 KNN algorithm

KNN was proposed by Cover and Hart in 1968. It is one of the earliest algorithms applied to text classification and has a good classification effect [29]. The basic idea is given the text to be classified, find the k texts closest to the new text in the training text set, and then judge the category of the new text according to the category of the k texts [16].

The KNN algorithm includes the following steps: first prepare the data and perform preprocessing, then calculate the distance between the test sample and other samples, then sort the calculated distance and select the k points with the smallest distance, and finally compare, test the sample The category of is the category with the highest proportion among the K points.

The KNN algorithm has high accuracy and good stability, but it can be seen from the steps that the value of k is very important. If the k value is too small, the true classification of the test sample may not be selected [29]. If the value of k is too large, the amount of calculation will increase, and the calculation overhead will be very large, which will affect the effect of classification.

2.4.2.3 Naive Bayes Algorithm

This is a probability-based machine learning algorithm, which is widely used in text classification. The main concern is the probability of the text belonging to each category. The main idea is to calculate the given text $d=(t_1, t_2, t_3, \dots, t_n)$ the probability that the feature item t belongs to each category, the probability that the text d belongs to the category C_i can be obtained, and the category of the text d is judged by these probability values, and the frequency of the feature item appearing in the category C_i is calculated The probability that the feature item belongs to the category C_i can be obtained[17].

Naive Bayes performs sentiment analysis, the process is simple and fast, and when the assumption of independence of distribution is established, the effect of Bayesian classification is very good, and it requires fewer samples [30]. Naive Bayes is highly efficient and robust. But it also has shortcomings. The good results are based on distribution independence. In fact, this assumption is

difficult to achieve, so it will affect the prediction effect[17].

2.4.2.4 Support vector machine algorithm

The support vector machine [19] started as a two-class classification model. The main idea of this method is to target the feature vector in a plane, including positive and negative categories, the SVM algorithm is to divide the two categories by finding a hyperplane to distinguish the two categories [31]. The specific classification is shown in the figure 2.

For such a two-classification model, when multi-classification of text is to be performed, multiple hyperplanes need to be constructed. The main methods are:

1. One-to-many, take one as a positive sample, and the other as a negative sample, and finally construct a multiple SVM model. In sample prediction, the final category is classified as the category with the largest SVM value [20].
2. One-to-one, each of the positive and negative samples selects one class. For K class samples, $K(K-1)/2$ classifiers are finally constructed [20].
3. DAG SVM method. Construct a directed acyclic graph and classify and train according to different categories [20].

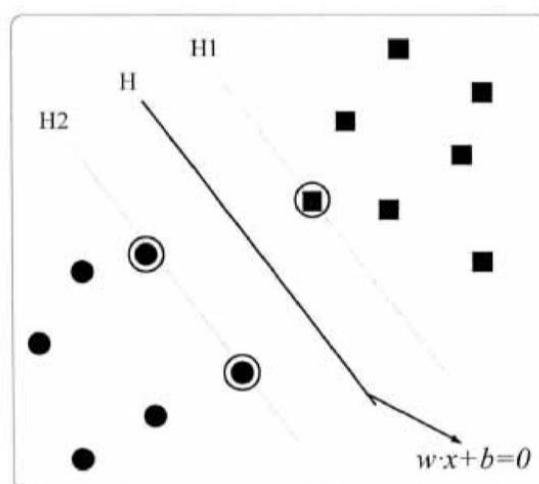


Figure 2 Classification algorithm process [20]

2.4.3 Deep learning

This section mainly introduces several basic neural network models.

2.4.3.1 Recurrent neural network

The cyclic neural network is a network model with an iterative structure. The input is a time series string. The intuitive understanding is similar to the way humans read or read text. With the help of the coding of interneurons in the computer, the history of interneurons Information can naturally be passed to the next interneuron, so it has obvious memory capabilities and can model serialized data.

However, in the RNN design, the back propagation of the gradient leads to the continuous accumulation of the gradient, and the intermediate eigenvalues will appear to move closer to the extreme value, which is manifested as infinity or close to 0. This phenomenon is called gradient explosion or gradient disappearance. Therefore, the biggest challenge RNN faces is the problem of long-term dependence. Long-term dependence means that the nodes of the neural network cannot obtain contextual information from long-distance time steps after multiple calculations over time. The specific performance is that the gradient of the weight matrix after multi-stage propagation tends to disappear or explode. The gradient is the value used to update the weight of the neural network. The disappearance of the gradient means that the gradient decreases as the backpropagation. If the gradient value becomes very small, the network will not continue to update the weight. Using RNN as a training model often fails to achieve expectations. The effect, the model structure diagram is shown in Figure 3.

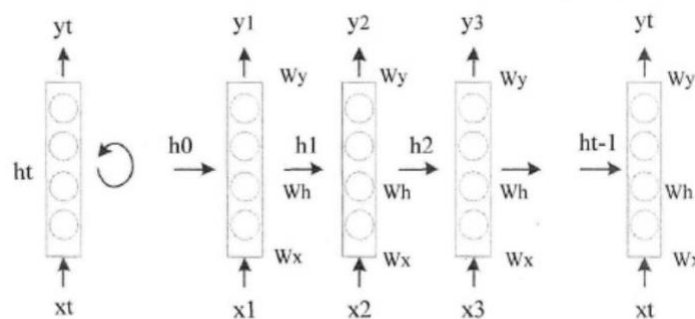


Figure 3 RNN model diagram

2.4.3.2 Long short-term memory network

In order to solve the problem of gradient disappearance and gradient explosion in the recurrent neural network, Hochreiter [21] et al. proposed a long and short-term memory neural network, referred to as LSMT. Compared with RNN, the neuron of LSTM is also a model for modeling serialized data, and its input includes the current node and the previous hidden layer [32]. The biggest difference from RNN is the change of internal structure, which is specifically manifested in the addition of input gates, forget gates, output gates and internal memory units to neurons [32]. The LSTM model structure is shown in Figure 4.

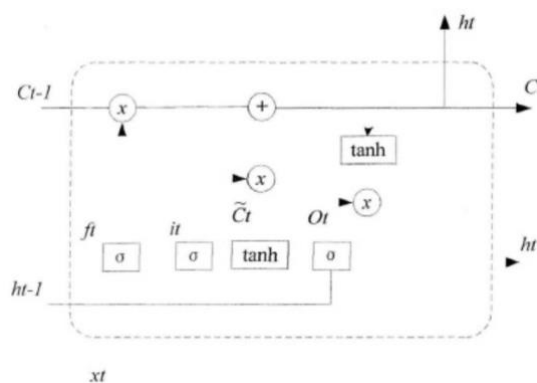


Figure 4 RNN model diagram [21]

2.4.3.3 Convolutional Neural Network

Convolutional neural networks were first used in the field of computer vision. Traditional convolutional neural networks consist of three structures: convolution, activation, and pooling. Kim [22] et al. applied convolutional neural network to text classification (Text-CNN) for the first time, training word vectors as weight variables to better capture the local relevance of text semantics and convert text into vectors when processing text tasks. The final output is obtained after a series of operations such as convolution, pooling, dropout regularization, etc., by multiplying different convolution kernels (size, number) and vectors by elements. The model structure of Text-CNN is shown in Figure 5.

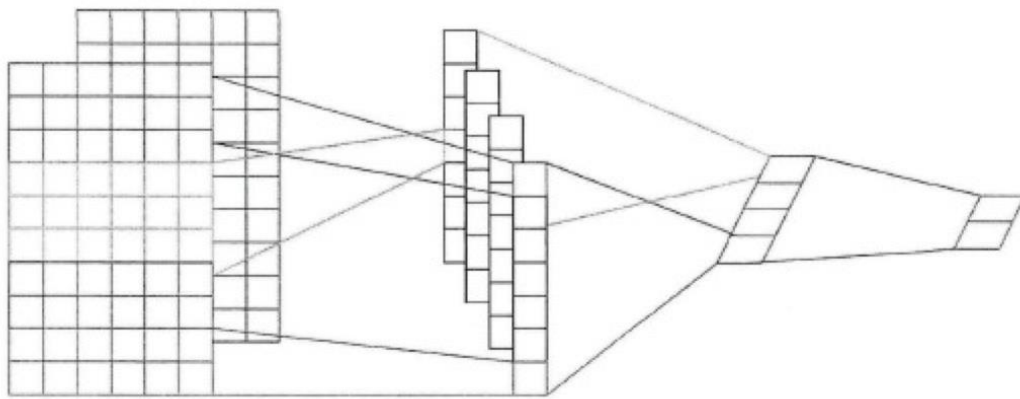


Figure 5 Text-CNN model diagram [22]

In addition, there are many joint deep learning models. Luo Fan et al. [23] used combined loop neural network and convolutional neural network to propose a multi-layer network model H-RNN-CNN, which uses a two-layer RNN to model text and introduces it into the sentence layer. Realize the sentiment classification of long texts; Xing et al. [24] introduced a new parametric convolutional neural network for aspect-level sentiment classification, which uses a parameterized filter (PF-CNN) and a parameter gate mechanism (PG-CNN), Han Jiansheng et al. [25] proposed a sentiment analysis model based on Bi-Temporal Convolutional Network (Bi-TCN). The model uses a unidirectional multi-layer cavity causal convolution structure to extract the forward and backward features of the text respectively and combines the sequence features in the two directions to perform sentiment classification. The experiment proves that it is compatible with the one-way time convolutional network sentiment analysis model. In comparison, the bidirectional temporal convolutional network model has a higher accuracy rate.

2.4.3.4 Attention mechanism

The attention mechanism is derived from scientists' research on human vision, and the selective attention mechanism of the human visual system has attracted the attention of scientists [33]. Biologists have discovered that the attention of the human visual system is very limited, and it is impossible to carefully observe the information perceivable by the eyes at the same time. We tend to quickly locate the target area that needs to be focused on according to our own needs, so as to

devote limited visual attention to the target area, so that we will be able to obtain more detailed information of the target area, and we will also abandon the useless information in other areas, which greatly improves the efficiency of our visual system to process information. The attention mechanism in machine learning is inspired by the selective attention mechanism of the visual system. The core goal is to obtain more useful and critical information for the current task from a variety of information [34].

The attention mechanism model is shown in the figure.

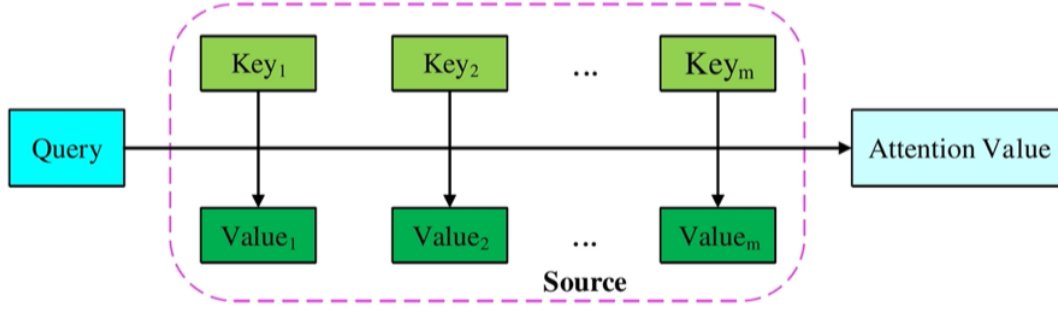


Figure 6 The attention mechanism model [34]

From the figure, we can see that the attention mechanism is composed of a series of <key, value> data pairs. When a certain query is given to perform a certain task, the Query and the weight score of each key are calculated (correlation), the value corresponding to the key is weighted and summed to obtain the final Attention vector value. The above process can be expressed by the following formula [2].

$$Attention(Query, Source) = \sum_{j=1}^m score(Query, Key_j) * Value_j$$

Among them, Key and Value are real number vectors, Query and Attention Value are also real number vectors, the positive integer m represents the length of the Source, and the score () is the similarity calculation function.

2.4.3.5 Bert

In 2018, the strongest NLP model released by the Google team-BERT [35] (Bidirectional Encoder Representation from Transformers) broke the best record of 11 NLP tasks in one fell swoop and became the most successful model in the NLP field. Bert's innovation is to use the Transformer structure in the language model and use the strategy of Mask part of the word in the pre-training language model [36]. This method can further enhance the generalization ability of the pre-training language model and can fully describe the character level and words [36]. The characteristics of the relationship between the levels.

The reason why Bert is so successful is that it uses a novel strategy in the pre-training language model. It uses the Masked Language Model when training word-level representation and uses Next Sentence Prediction for sentence-level representation. The Bert model uses the Transformer

architecture inside. The Transformer architecture is completely based on the above attention mechanism, and the residual connection and normalization methods are used, and the number of network layers can be very deep. The Bert model uses a two-way Transformer as the framework of its training model [36].

At the same time, we can also use the word vector extracted from the Bert pre-training language model to replace the traditional Word2Vec word vector features and transform from a context-free static word vector to a context-sensitive dynamic word vector, which can better solve the polysemy of a word problem and can also improve the classification performance of the model [36].

The two tasks of Bert are as follows:

(1) Masked Language Mode, which can be understood as cloze in the English test, it will randomly mask 15% of the words in each sentence, and then the model will predict the words that is masked based on the context information. For example, if the input sequence is "I am a student", the masked sequence may be "I am a [mask]", and then the masked word is predicted through model pre-training. At the same time, in order to solve the problem that some words may not appear in the model, a method of replacing according to probability is also adopted. In 15% of the words in the mask, 80% of the probability is directly replaced by the [mask] label, and 10% of the probability is used words that are randomly selected from the corpus to replace the mask words. For example, "I am a student" may be replaced with "I am a teacher", and there is a 10% probability that it will not be replaced. Using this strategy allows the Transformer model to better obtain the distributed representation of global information [36].

(2) Next Sentence Prediction can capture sentence-level features. In this task, the main objective is to predict whether two sentences are a continuous sentence pair relationship. When training the model, the input sentence pair information will be replaced, which means that the second sentence in the sentence pair will be replaced with a 50% probability. Bert's pre-trained language model can directly obtain a unique vector representation of an entire sentence. A marker "[CLS]" is added in front of each input sentence, and a marker "[SEP]" is also added at the end of the sentence, so that the Transformer can encode "[CLS]" during the training process. Therefore, the encoding information of "[CLS]" represents the semantic information of the entire sentence [36].

2.5 Feature selection

From the grammatical level, an article is generally composed of words, phrases, sentences and paragraphs, so these constituent elements can be used as the characteristics of the document. However, the combined features will increase exponentially as the level of grammatical elements increases. Therefore, classification methods based on the features of sentences and paragraphs are relatively rare. Common document features include words, N-gram items, phrases, concepts, and parts of speech.

2.5.1 Word

For a given document, the most direct method is to use words or phrases as feature items that represent the semantics of the document. Since Chinese is different from English in format, there is no obvious demarcation point between words. It is necessary to use dictionaries and special word segmentation technology to pre-process Chinese documents for word segmentation. If necessary, it is necessary to mark part of speech.

2.5.2 N-gram

The n-gram model was proposed by C.E. Shannon, the founder of information theory, when he was studying information coding[26]. It was later used to study the statistical characteristics of strings in English texts-information entropy. After that, N-gram was applied in many fields, including text processing related fields. The basic idea of the N-gram model is to operate a sliding window of size N in the unit of character bytes for the text content, and intercept a string of length N. Each string is called an N-gram item.

And n-gram item performs frequency statistics to form a text feature vector space. For Chinese, N-gram items generally consist of N adjacent Chinese characters.

2.5.3 TF-IDF

The TF-IDF weight function is a statistical method, which uses the inverse document frequency of the word to weight the word frequency [37]. The basic idea of TF-IDF is that the more the word appears in the text, the more important it is, and the higher the text frequency of the word. The lower the importance of the word [37]. The value of TFIDF is obtained by the product of two parts, which are term frequency (TF) and reverse document frequency. IDF is a value used to measure the universality of a word, and the IDF of a feature item is obtained by taking the logarithm of the reciprocal of the document frequency (DF) of the feature. TFIDF tends to keep the more special words in the document and filter the more commonly used words in the training set. The principle is that if the feature items are concentrated in some specific texts in the set, they often appear in some specific texts, and these texts account for a very large proportion of the entire training set. If it is small, then the feature can get a high-weight TFIDF. The formula is as follows:

$$w(t) = tf(t, d) \times idf(t) = tf(t, d) \times \log \left(\frac{N}{df(t)} \right) [37]$$

Among them, $tf(t, d)$ is the number of times the feature item t appears in the text d , N is the total number of texts in the training set, and $df(t)$ is the number of texts that contain the feature item t .

For example, if the total number of words in a document is 100, and the word "happy" appears three times, then the word frequency (tf) of the word "happy" in the document is $3/100=0.03$. Assume that there are 10,000,000 documents in total, and the word "happy" appears in 1,000 documents. Then the reverse document frequency (idf) is $\log(10,000,000/1,000)=4$. Then the TF-IDF score of the word is $0.03*4=0.12$.

2.5.4 Bag of words model

The BOW model is also a document representation method, which is often used in the field of information retrieval [38]. The BOW model ignores factors such as the order, grammar, and syntax of the words in the document, and regards the content in the document as a collection of several characters or words. The characters or words in the document maintain their own independence, and their appearance does not depend on the appearance of other words or words.

The specific practices of the BOW model are as follows [38]:

- (1) Word segmentation is performed on the sentences of all documents, and the sentences are cut into separate words [38].
- (2) Use all the words in the document to form a dictionary [38].
- (3) Construct an index to the words in the dictionary [38].
- (4) Encode the document and construct a vector representation of each document. The length of the vector is the length of the dictionary, and the value of each bit of the vector is the number of times the word represented appears in the document [38].

Below, gives an example, for example: There are two documents as follows:

- (1) I am a student.
- (2) I am a teacher.

According to the above two documents, the dictionary is constructed as follows:

`{"I":1, "am":2, "a":3, "student":4, "teacher":5}`

The two documents are represented as the following two vectors:

- (1) [1,2,3,4]
- (2) [1,2,3,5]

In this way, this vector is a kind of text representation and can be used as a feature of the text.

2.5.5 Word2Vec

Word2Vec [39][40] is a model for fast training word vectors proposed by the team led by Google researcher Mikolov. It contains two training models, one is CBOW and the other is Skip-gram. Both models include an input layer, a projection layer, and an output layer.

(1) CBOW

The CBOW model is to give a piece of text prediction for training, and then select a certain length (window) as the research object and use context vocabulary to predict the target vocabulary [39][40]. That is, the context of a certain characteristic word is used to predict the word vector of this word. Its input is the context word vector of a certain characteristic word, and its output is the word vector of the characteristic word. The CBOW model is shown in Figure 7.

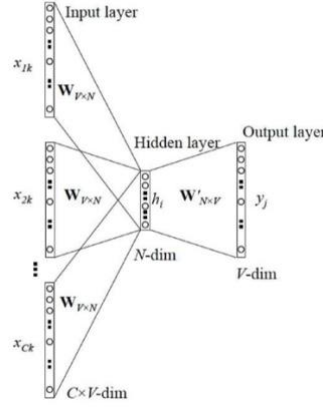


Figure 7 Word2vec CBOW model [39][40]

(2) Skip-gram model

The function of the Skip-gram model and the CBOW model is just the opposite. Given a text prediction for training, a certain length (window) is selected as the research object, and the target vocabulary is used to predict the context vocabulary [39][40]. That is, the model uses a certain feature word to predict the surrounding context. The skip-gram model is shown in Figure 8.

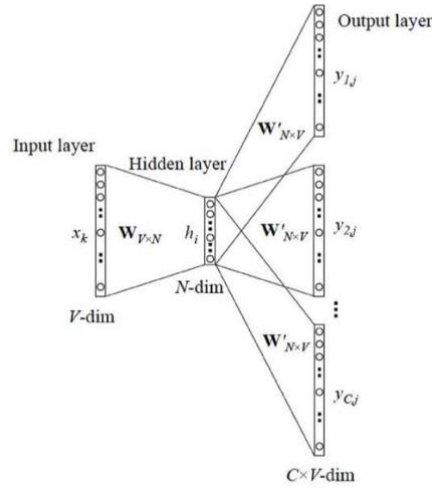


Figure 8 Word2vec skip-gram model [39][40]

2.6 Evaluation indicators

The evaluation indicators commonly used in classification problems are selected, which are called Precision, Recall, and F1 Value (F1-Measure). Among them, Precision refers to the proportion of samples that are predicted to be positive and correct in the final situation among all samples that are predicted to be positive; Recall refers to the proportion of all actual positive samples that are predicted to be positive and correct in the final situation. The calculation formulas of the two need to use the confusion matrix [27], the matrix is shown in the Table 1.

Table 1 Confusion matrix [27]

	Predicted as positive sample	Predicted as negative sample
Actual positive sample	TP	FN
Actual negative sample	FP	TN

TP represents the positive sample that is ultimately predicted to be positive. FP represents the negative sample that is ultimately predicted to be positive. FN represents the positive sample that is ultimately predicted to be negative. TN represents the negative that is ultimately predicted to be negative sample.

The calculation process of Precision is as follows [27].

$$Precision = \frac{TP}{TP + FP}$$

The calculation process of Recall is as follows [27].

$$Recall = \frac{TP}{TP + FN}$$

Precision is generally used to measure the accuracy of classification; Recall is generally used to measure whether the classification results are completely correct. In actual use, you will find that Precision and Recall are a pair of opposite measurement indicators, so usually the weighted harmonic average of the two is used as the corresponding comprehensive evaluation indicator, that is, the F1 value. The calculation process of F1 is as follows [27].

$$Recall = \frac{2 * Precision * Recall}{Precision + Recall}$$

2.7 Conclusions

The core of sentiment analysis task lies in the acquisition of text sentiment features and the construction of classification models. In recent years, with the popularity of deep learning, more and more deep learning techniques have been applied to sentiment analysis tasks. Using a variety of neural network architectures, the model can extract more and deeper information, and has achieved good results.

At the same time, more and more model combination techniques are also applied to sentiment analysis tasks. Ouyang et al. [41] proposed a method of combining CNN network and Word2Vec vector and added dropout technology to the model to prevent overfitting. Bradbury et al. [42] combined CNN and RNN networks to construct a brand-new structure—QRNNs. The model is much faster than the simple RNN structure in training speed and achieves better performance [42]. Xue et al. [43] used CNN to combine with the gate mechanism and used the Tanh-Relu activation function as the threshold unit and achieved good results [43]. Zhang et al. [44] combined the convolutional neural network with the attention mechanism, used CNN to extract features, and added the features extracted by CNN to the attention mechanism, thereby improving the performance of the model [44]. Ambartsoumiandeng et al. [46] applied the self-attention mechanism to sentiment analysis tasks. Wang Wei and others [45] integrated the attention mechanism into the two-way gated recurrent neural network, used BiGRU (Bidirectional Gate Recurrent Unit) [47] to extract the emotional features in the text, and then calculated and assigned the corresponding weights to the emotional features through the attention mechanism. The classification ability of the model was also improved [45].

Chapter 3 Research and Development Methodology

This chapter will mainly discuss the research and development methods of the project.

3.1 Research Methodology

This project mainly uses two research methods.

(1) Collection of literature and theoretical research.

In response to the problem of social media sentiment analysis, by consulting relevant information such as network information, journal literature, etc., we learned about the background of sentiment analysis and related theoretical and technical research knowledge and studied the feasibility of the project design plan.

(2) Data analysis method

Perform statistical analysis and experiments on the acquired data according to certain methods.

3.2 Experimental Design

At present, there are two most widely used data mining models, namely SAS's SEMMA and CRISP-DM [48]. The two models will be briefly introduced in the following.

(1) SEMMA

The whole process of SEMMA is divided into 5 parts: Sample, Explore, Modify, Model and Assessment. SEMMA reflects the essence of data mining [49].

Data sampling. The task of this step is to sample the data. We need to extract a sample data subset related to the exploration problem from a large amount of data. Through the selection of data samples, this can reduce the amount of data processing, save system resources, and at the same time make the regularity reflected by it more prominent [49].

Data processing. This step can be further divided into three parts: feature exploration, analysis and preprocessing. After obtaining a sample data set, we need to check whether it meets the envisaged requirements, whether it has obvious laws and trends, and whether there are unexpected data states. In addition, in order to improve the effect of data mining, the data needs to be preprocessed [49].

Data analysis. This step continues to clarify the problem and whether the data set fits the problem. In all stages, data mining products must make the data used and the model to be established in a state of easy adjustment, modification and change to ensure the effective progress of data mining. Therefore, at this stage we need to clarify the problem, adjust the data, and select the technology [49].

Model establishment. This step is the key and core part of data mining. We must provide sufficient alternative technical means at this stage [49].

Model evaluation. This step is to comprehensively explain and evaluate the model and knowledge [49].

(2) CRISP-DM

CRISP-DM is the world's first data mining industry methodology, jointly launched by the world's leading data mining experts SPSS, NCR, Daimler Chrysler and OHRA [50]. According to the CRISP-DM model, the development process of a data mining project includes the following six interconnected links: business goal value, data understanding, data preprocessing, model building, evaluation model and model implementation [50]. The CRISP-DM model is shown in the Figure 9.

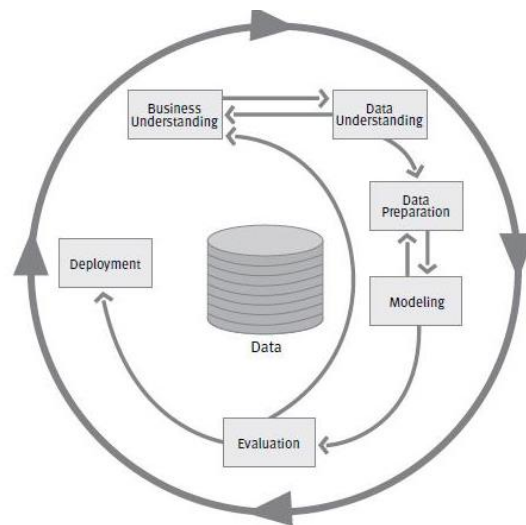


Figure 9 The CRISP-DM model [50]

At the same time, the six links also have their own tasks. The stages and their tasks are shown in the Figure 10.

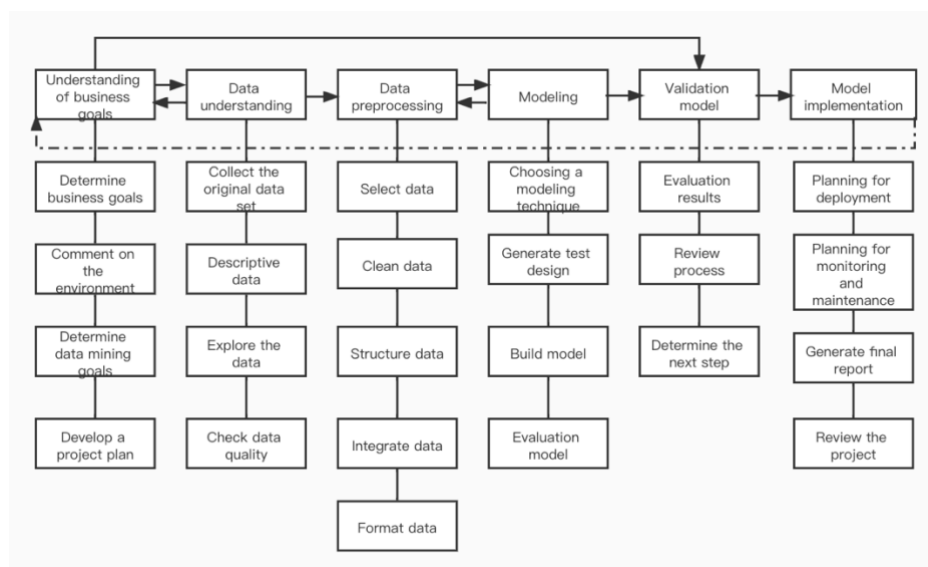


Figure 10 The stages and their tasks

It can be seen from the figure that the CRISP-DM model provides an overall description of the life cycle of a data mining project, including the six phases that the data mining project will go through, the tasks in each phase, and the connections between each phase. It is worth noting that the order of these stages is not fixed. There is a cycle between stages (as shown by the direction of the realization arrow), and at a certain stage, it may be necessary to return to the previous stage. Which stage or task to be executed in each step depends on the result of the previous step. As indicated by the direction of the dashed arrow, the overall cycle proceeds in stages. At the same time, the model shows that the data mining work does not end at once, it is a process of multiple iterations, multiple adjustments, continuous loops, continuous iterations, and continuous revision and improvement [50].

In summary, CRISP-DM describes the data mining process from the perspective of support function and methodology, and it is more appropriate to describe the data mining process at one time. Therefore, this project uses the CRISP-DM model for data mining and analysis.

Chapter 4: Design and Implementation

This chapter will carry out experimental design and implementation according to the CRISP-DM model mentioned in the previous chapter.

4.1 Understanding of business goals

At this stage, we need to understand our needs, and then focus on understanding the needs of the project from a business perspective, and then convert this knowledge into problem definitions for data mining.

With the rapid development of the Internet, people are more and more inclined to post some of their subjective feelings on social media. This information includes comments on certain products, participation in discussions on current popular news events, and emotions or views on any event. Among these social media, very popular microblogs, such as Twitter and Weibo, which mainly focus on brief, real-time status updates, attract more and more people to express their opinions and emotions on current hot topics.

At the same time, Twitter and Weibo can use some specific signs to indicate which event/item they are commenting on when they post comments or their own emotions. Therefore, when we need to get people's views on a specific event or product or movie, we only need to search according to a specific logo to get people's comments. This greatly facilitates us to collect people's views on certain things.

According to different specific signs, different institutions can obtain relevant information according to their own needs. For example, government departments need to obtain people's views on a recently promulgated policy, so government departments can search for comments on tags containing the policy to get people's ideas, so as to better improve the policy; If a business wants to get people's opinions on a certain product, they can also search for the tag that contains the product to obtain feedback on the product, and continuously improve their product based on the feedback to improve their reputation.

After we have collected the information released by people, we need to perform sentiment analysis and classification of the released information. Faced with a large amount of unstructured emotional data, it is very difficult to draw conclusions by only relying on one-by-one manual reading. Therefore, modeling and sentiment analysis are necessary, which are of great significance to different organizations such as government departments and businesses.

According to the above description, this project mainly conducts data mining based on the following background:

1. The text information of different emotions will contain different emotion words, and the text

information of the same emotion will contain many words of the same emotion. It is difficult for different institutions to obtain key emotional words of different emotions only from text information.

2. After we obtain a large amount of text information, it is difficult for institutions to directly obtain the emotional tendency of the text information.

3. For different language text information, different methods are needed to process the text and analyze the sentiment of the text.

4.2 Data understanding

In the data understanding stage, we first need to collect the initial data sources, and then do a preliminary analysis of the data, including identifying data quality problems, and obtaining a preliminary understanding of the data, implicit information, and its assumptions and feelings through basic observations of the data. Interested data subset [50].

4.2.1 Preliminary analysis of the data set

The source of the data set was introduced in 2.2, so this section will initially analyze the data set.

1. Twitter data set

The data set has only one csv file. The CSV file contains a total of 6 fields, which are the sentiment polarity of the tweet, the id of the tweet, the date of the tweet, the flag, the id of the user, and the text of the tweet. This project mainly uses the first and last columns, namely the emotional polarity of the tweet and the text of the tweet. For this data set, it mainly analyzes the polarity of Twitter text. In this dataset, tweet texts mainly have two polarities, negative (0) and positive (4). For example, the negative "Im heartbroken", the positive "have a great time in rio".

The data set mainly contains six attributes, as shown in the following Table 2 .

Fields	Data type	Description
target	number	The polarity of the tweet (negative:0, positive:4)
id	integer	The id of tweet
date	string	The date of the tweet
flag	string	The query. If there is no query, then this value is NO_QUERY.
user	string	The user that tweeted.
text	string	The text of the tweet.

Table 2 Fields in the Twitter data set

Some negative emotional tweets are shown in the Figure 11.

0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Biah!
0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here? because i can't see you all over there.
0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
0	1467811592	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	mybirch	Need a hug
0	1467811594	Mon Apr 06 22:20:03 PDT 2009	NO_QUERY	coZZ	@LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , i'm fine thanks , how's you ?

Figure 11 Negative emotional tweets

Some positive tweets are shown in the Figure 12

4	1573112204	Mon Apr 20 23:46:09 PDT 2009	NO_QUERY	jadamson	@Anshin D'aw. Awesome though.
4	1573112205	Mon Apr 20 23:46:11 PDT 2009	NO_QUERY	Brunap	And Beirut is a great band, @leonmaia
4	1573112232	Mon Apr 20 23:46:09 PDT 2009	NO_QUERY	klein2	My very first astrophoto http://www.klein2.de/index.php?p=70
4	1573112282	Mon Apr 20 23:46:10 PDT 2009	NO_QUERY	LaylaElayyadi	@Columbus_Short Lol i see you trying to help the cause as well.Goodluck. I donated in behalf of both of yall.
4	1573112286	Mon Apr 20 23:46:10 PDT 2009	NO_QUERY	brookehaskins	@Dutchruder All good headache nearly gone loved the pic of ur little one cleaning the windows! What's it like in notts today?
4	1573112300	Mon Apr 20 23:46:10 PDT 2009	NO_QUERY	ashley_k8	might be regretting starting this.....hopefully i convince people to get on

Figure 12 Positive tweets

In fact, we just need the first and last columns, so the first and last columns of the data need to be extracted, and then the text in the Twitter text needs to be preprocessed. Because the data set is not divided into training set, test set, and validation set. Therefore, the data set needs to be divided at a ratio of 8:1:1.

2. IMDB data set

This data set is a movie review data set, which contains two folders, train and test. The train folder is mainly used for movie reviews for training, and two folders, neg and pos, need to be used. Neg contains all the negative comments, which contains 12,500 txt files, and each txt file contains a negative comment. The same pos contains all the positive comments, which contains 12,500 txt files, and each txt file contains a positive comment. The Test folder is mainly used for testing movie reviews, and its structure is the same as the train folder.

For the IMDB data set, it is necessary to match the movie reviews in the data set with their polarities. Each review in train and pos needs to be placed in the same csv file, and each review is given a polarity label. Pos is marked as 1, and neg is marked as 0. Therefore, the csv file should contain two columns, one column is polarity label, and the other column is movie reviews.

The processed data set contains 2 attributes, as is shown Table 3.

Fields	Data type	Description
target	number	The polarity of the tweet (negative:0, positive:1) .
comment	string	The movie reviews.

Table 3 Fields in the IMDB data set

One negative comment is as follows.

Story of a man who has unnatural feelings for a pig. Starts out with a opening scene that is a terrific example of absurd comedy. A formal orchestra audience is turned into an insane, violent mob by the crazy chantings of its singers. Unfortunately, it stays absurd the WHOLE time with no general narrative eventually making it just too off putting. Even those from the era should be turned off. The cryptic dialogue would make Shakespeare seem easy to a third grader. On a technical level it's better

than you might think with some good cinematography by future great Vilmos Zsigmond. Future stars Sally Kirkland and Frederic Forrest can be seen briefly.

A positive comment is as follows.

Although this was obviously a low-budget production, the performances and the songs in this movie are worth seeing. One of Walken's few musical roles to date. (He is a marvelous dancer and singer and he demonstrates his acrobatic skills as well - watch for the cartwheel!) Also starring Jason Connery. A great children's story and very likable characters.

3. Weibo data set

The data set has 360,000 pieces of data, including 4 kinds of emotions: joy, anger, disgust, and depression. Among them, there are about 200,000 pieces of joy, and 50,000 pieces each of anger, disgust and depression. The data set also only contains a csv file, and also needs to be divided into training set, test set and validation set according to the ratio of 8:1:1.

The data set contains two attributes, as is shown in Table 4.

Fields	Data type	Description
label	number	The polarity of the tweet (joy:0, angry:1, disgust:2, depression:3)
Weibo	string	The text of Weibo.

Table 4 Fields in the Weibo data set

An emotionally positive Weibo.

太萌了好可爱哦。(So cute, so cute)

An angry Weibo.

吃了一个巨大的馅饼居然不觉得饱~~~(I did not feel full after eating a huge pie~~~)

A disgust Weibo.

我表示压力很大。(I said it was very stressed)

A depressing Weibo.

昏昏噩噩,混混沌沌~~我的世界充满了浆糊~~(Faint, chaotic~~My world is full of paste~~)

From the above example, we can see that the first two data sets and the third data set are different languages. At the same time, the tweets in the first data set are mostly shorter English sentences, while the second data set are mostly longer English sentences.

4.2.2 Data quality

First, the completeness of the data set is analyzed. Through the traversal of the three data sets, it is found that the integrity of the data set is high, and there is no data with null values.

Then, I made basic observations on the data set and found that on social media such as Twitter and

WeiBo, people are more casual when expressing their emotions and thoughts, do not pay special attention to language norms, and use some abbreviations and some special symbols. For example, in the Twitter data set "hope your ok." As a non-indigenous Chinese, he is not very clear about some colloquial expressions, such as "Is not going to sleep tonite." It is known that tonite is tonight by checking the information. In the same WeiBo data set, "悬疑小说 0.0"("Suspense Novel 0.0"), "0.0" may be an emoji, "太靚 拉!!!!廣告拍攝 中" (It's so pretty! We're Shooting a commercial.) The translation may be more grammatical, but in fact, from the Chinese point of view, this sentence uses a lot of spaces, and the expression is more colloquial. At the same time, it is worth noting that the text content of tags 2 and 3 in the WeiBo dataset, that is, disgust and depression, are the same, so the dataset has only 3 emotion types instead of the 4 types of emotion described in the dataset introduction. Therefore, the subsequent preprocessing stage is very important.

4.3 Data preprocessing

The data preprocessing stage covers all the activities used to construct the final data set from the original data, and the final data set will be used as the input or sample data of the data mining model [50]. Each task in the data preprocessing stage may need to be executed repeatedly. As mentioned in the previous chapter, the order in which these tasks are executed is not fixed. In this stage, we need to select the attributes of the data set, and at the same time, we need to clean, transform, and organize the data set according to our needs and the characteristics of the data. The quality of data is critical to data mining, and it often affects our prediction results. Therefore, in the process of data mining, we cannot ignore data preprocessing.

During this stage, we need to process the data. Common processing includes selection of attribute sets, processing of abnormal data, processing of missing values, data standardization, data type conversion, etc. We need to process the data appropriately according to our needs.

This project focuses on sentiment analysis of social media, while this project will focus on sentiment analysis of short texts. The IMDB data sets are mostly long texts and only relevant text information of movie reviews. Therefore, this project finally chose the Twitter and Weibo datasets for social media sentiment analysis according to the needs and goals of the project and the actual situation. The following will introduce the preprocessing work of this project on the two data sets with Twitter and Weibo.

4.3.1 Twitter data set

From the introduction of the previous data set, we know that the data set contains six attributes, as is shown in the Figure 13, and we only need the first attribute, which is sentiment polarity, and the last attribute, which is tweets text. Therefore, these two related attributes in the data set need to be extracted, as is shown in the Figure 14.

	target	ids	date	flag	\
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	

	user	text
0	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	scotthamilton	is upset that he can't update his Facebook by ...
2	mattycus	@Kenichan I dived many times for the ball. Man...
3	ElleCTF	my whole body feels itchy and like its on fire
4	Karoli	@nationwideclass no, it's not behaving at all....

Figure 13 Original Twitter data set

	target	ids	date	flag	\
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	

	user	text
0	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	scotthamilton	is upset that he can't update his Facebook by ...
2	mattycus	@Kenichan I dived many times for the ball. Man...
3	ElleCTF	my whole body feels itchy and like its on fire
4	Karoli	@nationwideclass no, it's not behaving at all....

[1600000 rows x 2 columns]

Figure 14 Two-attribute Twitter data set

After obtaining the tweets text and its polarity, the tweets text needs to be further cleaned. First, we need to remove the @user mark in tweets. As mentioned earlier, the tweets text contains many Twitter tags, and these tags are user information on Twitter, because this information is not helpful for sentiment analysis.

According to the characteristics of the Twitter tag, regular expressions are used to clean up, the regular expression is "@[\w]*", and the cleaned data set is shown in the Figure 15.

	sentence_length	text_tidy
0	19	http://twitpic.com/2y1zl - Awww, that's a Bu...
1	21	is upset that he can't update his Facebook by ...
2	18	I dived many times for the ball. Managed to ...
3	10	my whole body feels itchy and like its on fire
4	21	no, it's not behaving at all. i'm mad. why a...
...	...	...
1599995	11	Just woke up. Having no school is the best fee...
1599996	11	TheWDB.com - Very cool to hear old Walt interv...
1599997	11	Are you ready for your MoJo Makeover? Ask me f...
1599998	12	Happy 38th Birthday to my boo of alll time!!! ...
1599999	5	happy #charitytuesday

Figure 15 Twitter data set after cleaning

The comparison shows that @switchfoot in the first sentence, @Kenichan in the third sentence, and @nationwideclass in the fifth sentence have been removed.

At the same time, it can be noticed that there is a link/website in the first sentence, which is not very helpful for sentiment analysis, so it also needs to be removed. Similarly, according to the characteristics of the link/URL, regular expressions are used to clean up, the regular expression is "((http://)[^]*(https://)[^]*(www\.)[^]*)", the data set after cleaning is shown in the Figure 16.

	sentence_length	text_tidy
0	19	- Awww, that's a bummer. You shoulda got ...
1	21	is upset that he can't update his Facebook by ...
2	18	I dived many times for the ball. Managed to ...
3	10	my whole body feels itchy and like its on fire
4	21	no, it's not behaving at all. i'm mad. why a...
...	...	...
1599995	11	Just woke up. Having no school is the best fee...
1599996	11	TheWDB.com - Very cool to hear old Walt interv...
1599997	11	Are you ready for your MoJo Makeover? Ask me f...
1599998	12	Happy 38th Birthday to my boo of alll time!!! ...
1599999	5	happy #charitytuesday

Figure 16 Twitter data set after cleaning

Punctuation marks, numbers and special characters are also meaningless, and these characters also need to be eliminated. Similarly, use regular expressions to eliminate these characters, the regular expression is "[^a-zA-Z']", and the cleaned data set is shown in the Figure 17.

	sentence_length	text_tidy
0	19	Awww that's a bummer You shoulda got ...
1	21	is upset that he can't update his Facebook by ...
2	18	I dived many times for the ball Managed to ...
3	10	my whole body feels itchy and like its on fire
4	21	no it's not behaving at all i'm mad why a...
...	...	...
1599995	11	Just woke up Having no school is the best fee...
1599996	11	TheWDB com Very cool to hear old Walt interv...
1599997	11	Are you ready for your MoJo Makeover Ask me f...
1599998	12	Happy th Birthday to my boo of alll time ...
1599999	5	happy charitytuesday

Figure 17 Twitter data set after cleaning

In addition, some stop words and words with characters less than 3 are meaningless words. The English stop words used in this project come from the English stop words provided by Harbin Institute of Technology [51]. The method used is mainly to determine whether the vocabulary in the sentence is in the stop word list, whether the character of the vocabulary is less than 3, and if it is, it will be eliminated. The cleaned data set is shown in the figure.

	sentence_length	text_tidy
0	19	Awww bummer shoulda David Carr Day
1	21	upset can't update Facebook texting cry result...
2	18	dived times ball Managed save rest bounds
3	10	body feels itchy fire
4	21	behaving mad can't
...	...	...
1599995	11	woke school feeling
1599996	11	TheWDB cool hear Walt interviews
1599997	11	ready MoJo Makeover details
1599998	12	Happy Birthday boo alll time Tupac Amaru Shakur
1599999	5	happy charitytuesday

[1600000 rows x 4 columns]

Figure 18 Twitter data set after cleaning

4.3.2 WeiBo data set

From the introduction of the previous data set, it can be known that the data set contains 2 attributes. As shown in the Figure 19, the first attribute is the emotion category, and the other attribute is the Weibo text. Since the data format is only introduced here, the sentence is not translated.

	label	review
0	0	啊呀呀! 要死啦! 么么么! 只穿外套就好了, 我认为里面那件很多余啊周小伦喜歡 你各種 五角星的...
1	0	嗯.....既然大姚通知了.....那我也表示下收到.....姚, 你知道吗? 假如外星人入侵地球, 只要摧毁我们的...
2	0	风格不一样嘛, 都喜欢! 最喜欢哪张?
3	0	好呀, 试试D .I .Y .去死皮面膜1.将燕麦片加水浸泡6小时, 加入木瓜牛奶搅拌。2.放...
4	0	张老师, 谢谢依的信任! 粉丝多少无所谓重在质地近日发现一个现象——他加了你关注, 你回加后, 他立...
...	...	...
361739	3	德国队就那样输了.....
361740	3	人也一样, 沒 钱沒 地位, 谁会关注你? 大庆和上海都是被人遗忘的角落, 都是被遗弃的孩子。。。当...
361741	3	巴西要回家啦
361742	3	派啦! 諗 咩吖! 撈 呀! 十五十六之家已經 差不多十一月仲派唔派歌好呢? 定係 出年先派呢?...
361743	3	发觉呢个期末特别懶~死硬啦死硬啦, 我要静心复习啊!!

[361744 rows x 2 columns]

Figure 19 WeiBo data set

It is worth noting that, as mentioned in the previous data quality analysis, the data set has only three emotion categories: 0 for joy, 1 for anger, and 2 for disgust. The depressive sentiment text which the value of label is 3 will not be discussed in the subsequent analysis

Next, we will count the number of microblogs contained in each type of sentiment in the microblog data set and the number of microblogs contained in the entire data set. The statistical results of the experiment are shown in the Figure 20.

Number of microblogs (total) : 361744
Number of microblogs (happy) : 199496
Number of microblogs (anger) : 51714
Number of microblogs (disgust) : 55267
Number of microblogs (depression) : 55267

Figure 20 The number of microblogs

Since the above statistical result is added to the number of microblogs in the depression category, the above statistical result is not too accurate. In fact, the data set contains a total of 306,477 microblogs, of which the number of microblogs in the happy emotion category is 199,469, the number of microblogs in the angry category is 517,14, and the number of microblogs in the disgusting emotion category is 55, 267.

In Chinese text processing, word segmentation is a very important preprocessing step [51]. In the English text, the English text has clear separator spaces to divide words, so the word segmentation of the English text will be easier. For example, "我是一名学生。" (I am a student.) Chinese word segmentation result is "我(I)/是(am)/一个(a)/学生(student)". It can be seen that English texts can generally be segmented by spaces or other punctuation marks, while Chinese texts need to understand their meaning to be able to be accurately segmented. In fact, sometimes in Chinese texts in different contexts, the same sentence may have different participles representing different meanings.

This project mainly uses the jieba [7] tool for word segmentation. The specific code is as follows:

```
x_all_data = list(map(lambda x: list(filter(lambda x: len(x)>1,
                                           jieba.lcut(x))), all_data['review'].values))
```

The word segmentation results of the data set are shown in the Figure 21.

```
[['啊呀呀', '要死', '外套', '认为', '里面', '很多', '周小伦', '喜欢', '各种', '五角星', '衣服', 'XD', '伦伦', '真的', '越来越'], ['既然', '大姚', '通知', '表示', '收到', '知道', '假如', '外星人', '入侵', '地球', '只要', '摧毁', '我们', '网络', '地球', '完蛋', '恐惧', '觉得', '阿粥', '就是', '外星人', '一定', '微薄', '通知', '一下', '大家'], ['风格', '一样', '喜欢', '喜欢', '哪张'], ['试试', '去死皮', '面膜', '燕麦片', '加水', '浸泡', '小时', '加入', '木瓜', '牛奶', '搅拌', '放入', '一个', '压缩', '面膜', '敷于', '脸上', '10', '20', '分钟', '清水', '洗净', '可以', '改善', '粗糙', '肌肤', '去除', '死皮', '肌肤', '光滑'], ['张老师', '谢谢', '信任', '粉丝', '多少', '无所谓', '重在', '质地', '近日', '发现', '一个', '现象', '他加', '关注', '回加', '立即', '取消', '关注', '这是', '为什么', '回应', '立即', '取消', '关注']]
```

Figure 21 The word segmentation results

Since only the content of the word segmentation is displayed, the above results are not translated. Comparing the content of the review tag in the above Figure 19, it can be seen that the tool has completed the word segmentation work.

At the same time, it can be found through comparison that some characters were removed during the word segmentation process. Like the Twitter data set preprocessing, the Weibo data set also removed meaningless stop words in Chinese and segmentation words that is less than one character. Chinese stop words also use Chinese stop words provided by Harbin Institute of Technology [51]..

At this point, we have completed the work of data cleaning, the following will further analysis and modeling of the data set.

4.4 Data analysis

This section will conduct a detailed statistical analysis of the Twitter and WeiBo data sets. In this process, we will count the number of samples and the number distribution of each type of label. At the same time, the distribution of sentence length and word frequency will also be analyzed. Through these analyses, we can obtain the keywords of each emotion, and solve the first goal in the understanding of business goals, that is, text information of different emotions will contain different emotion words, and text information of the same emotion will contain many of the same emotion words. It is difficult for different institutions to obtain key emotional words of different emotions only from text information. Not only that, but the analysis can also help the group judge the reasonable truncation length of the sentence, which plays a vital role in the construction of the subsequent model.

4.4.1 Twitter data set

In the first step, we first need to understand the distribution of the number of emotional polarity labels in the dataset. In this project, the seaborn toolkit is used for statistical analysis, and the statistical results are shown in the Figure 22.

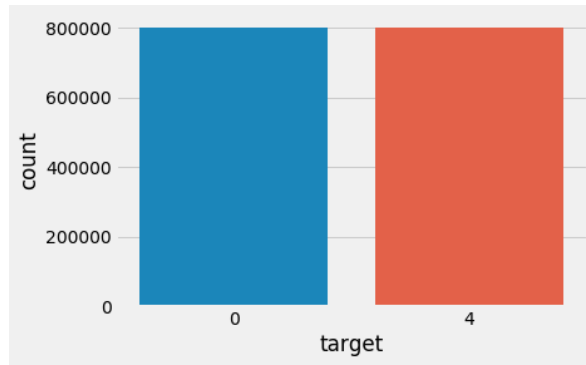


Figure 22 The distribution of the number of emotional polarity labels

It can be seen from the figure that 0 represents a tweet with a negative attitude, and 4 represents a tweet with a positive attitude. The tweets of the two attitudes are evenly distributed, each with 800,000 tweets.

The second step is to analyze the sentence length distribution. The sentence length distribution diagram is shown in the Figure 23.

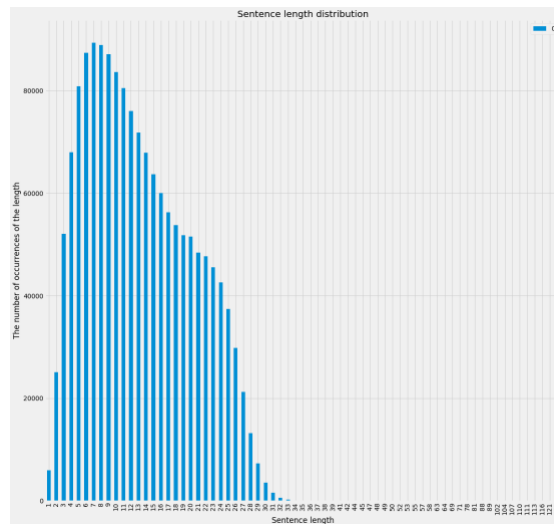


Figure 23 The sentence length distribution

From the above analysis, we can see:

1. The length of the sentence is between 1-33, which reflects that the content of the social media Twitter Tweets is short, and there is no exchange content like other forums, which never ends, and the length is longer.
2. At the same time, it can be seen from the figure that most sentences are concentrated between the lengths of 4-27. Therefore, in subsequent experiments, the length of 30 can be considered as a reasonable truncated alignment length, that is, it will not cause a large number of sentences to be truncated and lose the main information, and it can effectively avoid too many complementary features, resulting in excessive model parameters.

The third step is to analyze word frequency.

The picture below shows the most frequent words in positive comments. The positive emotion tweets word cloud is shown in the **Figure 27**

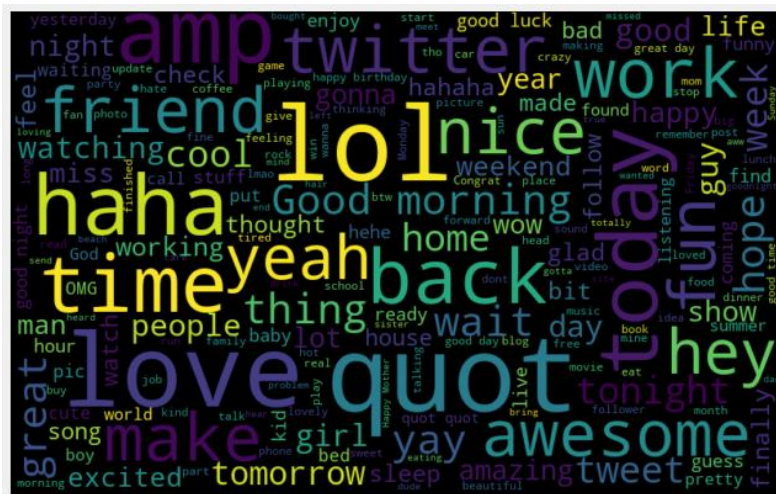


Figure 27 Positive emotion tweets word cloud

It can be seen from the above word frequency analysis:

1. The number of occurrences of 'LOL' in this data set is greater. 'LOL' is a competitive game, indicating that the collection time of this data set is likely to encounter a major game of the 'LOL' game, so people discussed the game on Twitter at that time, and it also shows that sentiment analysis of social media can well grasp people's views on a certain event and social media has a good real-time nature.
2. From the negative emotion tweets, we can find that 'sad', 'sick', 'bad', and 'hate' appear more frequently, indicating that people use some negative vocabulary in a negative attitude, which is also in line with our cognition. At the same time, it can be noticed that some positive words such as 'good' and 'love' also appear in negative emotion tweets, but it should be noted that there are more negative words such as 'don't' and 'can't' in negative words than the positive attitude and emotion tweets, so it may be a negative of some positive words. Through the positioning search, it was found that such collocations as "good morning" and "good luck" lead to more negative emotions with the word "good". At the same time, many sentences were also found to negate some positive words, such as "spencer is not a good guy.", "@DonnieWahlberg I hope i can make it to the auburn show. but its not looking good for me " etc.
3. From positive emotional tweets, we can find that words such as "good", "great", "love", "fun", etc. appear more frequently, indicating that people use more positive words in a positive attitude, which is also in line with our perception. At the same time, it can be found that negative words rarely appear in tweets with positive attitudes, indicating that people rarely negate negative words to express positive emotions.

4.4.2 Weibo data set

In the first step, we first need to understand the number distribution of sentiment classification labels in the dataset. The statistical results are shown in the Figure 28.

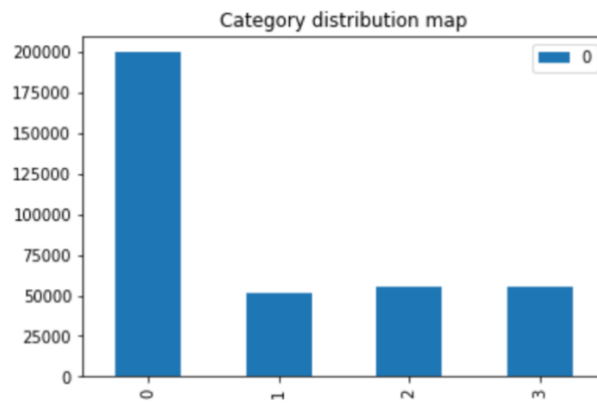


Figure 28 Category distribution map

First, the label which the value is 3 should be ignored. At the same time, it can be found that the number of data labels is not balanced. In order to conduct experiments better, the number of labels needs to be balanced. Therefore, in the follow-up experiments, 1,000 pieces of data were randomly selected from each label for related experiments.

The second step is to analyze the sentence length distribution. The sentence length distribution diagram is shown in the figure below.

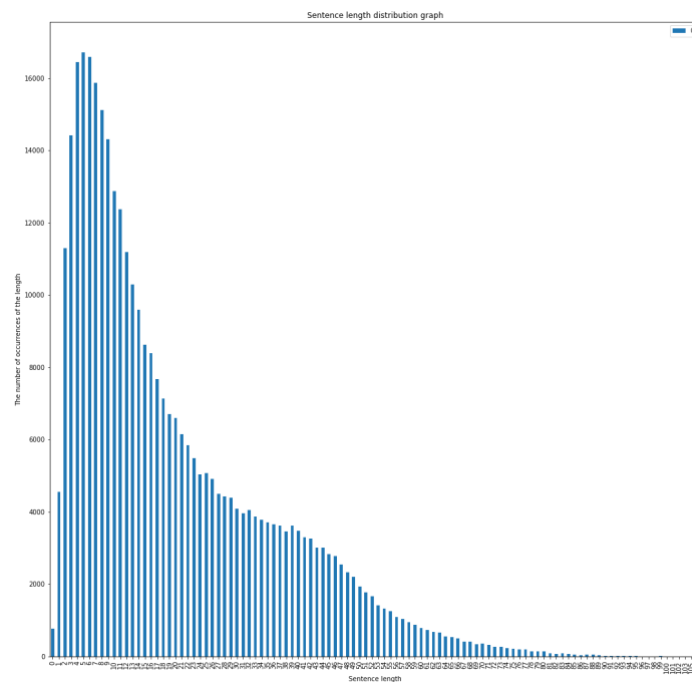


Figure 29 The sentence length distribution

As can be seen from the above figure, the length of the sentence is between 0-99, and most sentences are concentrated in the length of 3-60. Therefore, in subsequent experiments, the length of 60 can be considered as a reasonable truncated alignment length.

Finally, the word frequency of the three types of sentiment microblogs will be analyzed and counted. Since the sentiment of Chinese sentiment analysis is mainly reflected by adjectives, the jieba tool was used for part-of-speech tagging in this project, and the adjectives were extracted for analysis and statistics.

First, count the 50 words with the highest frequency of occurrence of happy Weibo, as is shown in the Figure 30.

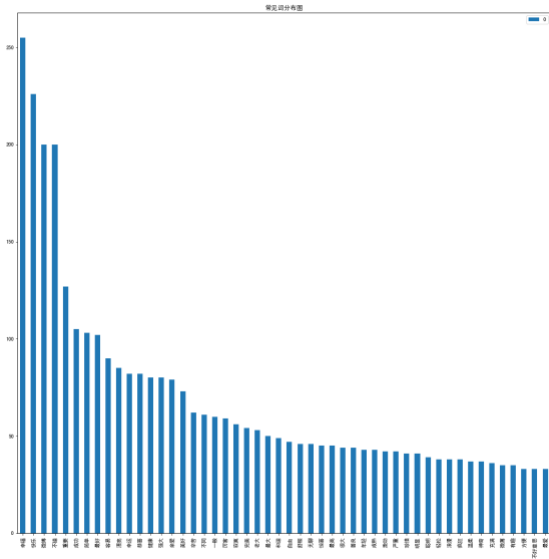


Figure 30 Words with the highest frequency of occurrence

The first ten words are "Happiness", "Happy", "Weibo", "Good", "Important", "Success", "Simple", "Best", "Easy", "Pretty".

The picture below shows the word cloud of this emotion type.



Figure 31 Word Cloud

From the picture, we can also see positive emotion words such as "快乐" (happy), "健康" (health), and "恭喜" (congratulations).

Then, count the 50 words with the highest frequency of occurrence of anger Weibo, as is shown in the Figure 32.

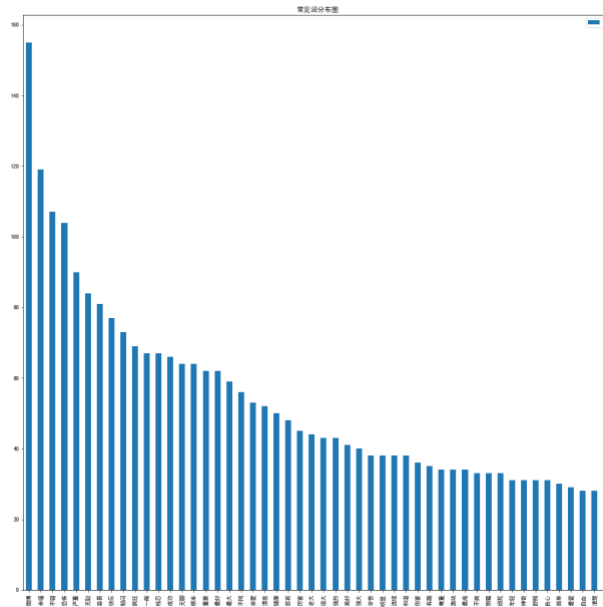


Figure 32 Words with the highest frequency of occurrence

The first ten words are 'Weibo', 'Happiness', 'Good', 'Horror', 'Serious', 'Shameless', 'Easy', 'Happy', 'Depressed', 'Crazy'. What is more surprising is that there are many positive words in Weibo classified by angry emotions, but on the whole, there are more negative or neutral words such as boredom(无聊), sadness(悲哀), bad luck(倒霉), irritability(烦躁), etc. which rarely appear in happy emotion Weibo. vocabulary.

The Figure 33 below shows the word cloud of this emotion type.



Figure 33 Word cloud

In the picture, we can see words that do not appear in happy emotions such as depression (郁闷), horror (恐怖), boredom (无聊), and sadness (悲哀), as well as positive words such as happiness (快乐) and beauty (漂亮).

Finally, count the 50 words with the highest frequency of occurrence of anger Weibo, as is shown in the Figure 34.

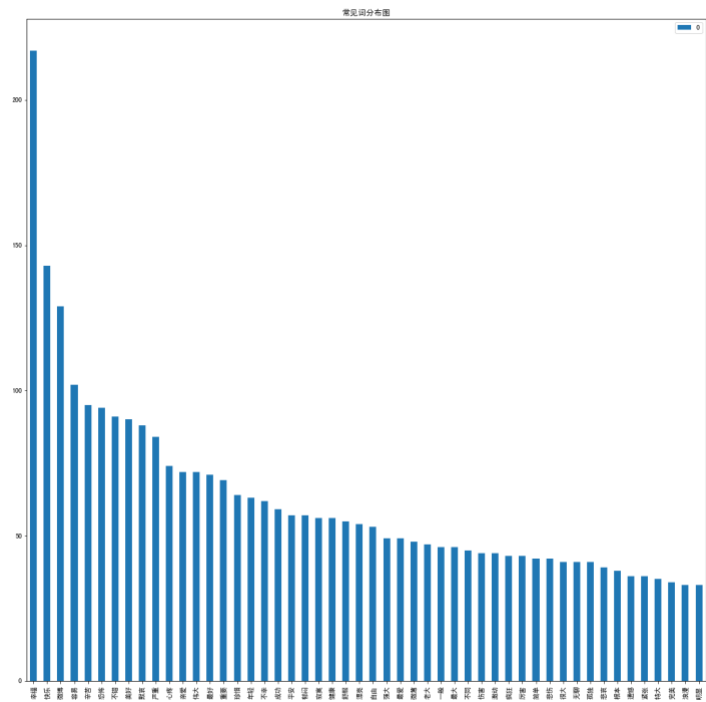


Figure 34 Words with the highest frequency of occurrence

The first ten words are 'Happiness', 'Happy', 'Weibo', 'Easy', 'Hard', 'Horror', 'Good', 'Great', 'Silence', 'Serious'. The high-frequency words appearing in Weibo with disgust and anger are similar. There will be words of positive emotions, as well as words of negative emotions.

The Figure 35 below shows the word cloud of this emotion type.



Figure 35 Word cloud

In the picture, we can see words that do not appear in joyful emotions such as hardship (辛苦), trouble (烦恼), regret (遗憾), and misfortune (不幸). At the same time, we can also see some positive words such as happy (快乐) and happiness (幸福).

4.4.3 Conclusions

From the above analysis, it can be seen that whether it is Chinese or English social media microblogs, microblogs with positive attitudes and emotions contain more words with positive attitudes, and generally do not show words with negative attitudes. For microblogs with negative attitudes and emotions (including microblogs with angry emotions and disgusting emotions), there are often words with negative attitudes, and words with positive attitudes appear more often. This is more surprising to me. I learned through analysis. It may be because mantras or idioms contain more positive words, and people also use negative positive words to express their negative emotions.

4.5 Model

This stage is mainly based on the previous analysis of business value goals, modeling in the pre-processed data, and data mining. At this stage, TF-IDF and Bag of Words are mainly used to extract features and encode the data. At the same time, logistic regression and Bert are used to classify the corresponding sentiment of the data set.

4.5.1 Data set

This project mainly uses two data sets, which are Twitter and Weibo. At the same time, according to the previous description, the data sets have been preprocessed accordingly.

4.5.2 Experiment

First, we need to perform feature extraction on the cleaned data set. This project mainly uses two methods to extract the features of the micro-blog content in the data set, and then vectorize it for text representation.

The first method is Bag of Words for feature extraction, mainly using the CountVectorizer toolkit in sklearn in python for feature extraction. The corresponding parameter settings are as follows:
CountVectorizer (max_df=0.90, min_df=2, max_features=1000000, stop_words='english')

After extracting the vocabulary features in the data set, the data set is divided into training set and test set according to the ratio of 8:2. Then it is necessary to vectorize the obtained features by passing the training set and the test set.

The second method is TF-IDF for feature extraction, mainly using the TfidfVectorizer toolkit in sklearn in python for feature extraction. The corresponding parameter settings are as follows:
TfidfVectorizer (max_df=0.90, min_df=2, max_features=1000000, stop_words='english')

After the features are also extracted, the proportioned training set and the data set are vectorized by the acquired features.

Then, after obtaining the vectorized representation, select the corresponding machine learning algorithm for training and prediction. This project mainly chose logistic regression and naive Bayes algorithm for training and prediction.

Logistic regression mainly uses the sklearn. linear_model. LogisticRegression toolkit in python to build the algorithm model. Experiment with the default parameters in sklearn, which is `logisticModel = LogisticRegression ()`.

Naive Bayes also uses the sklearn. naive_bayes. MultinomialNB toolkit in python to build the algorithm model. Use the default parameters of the tool to experiment, that is, `nb = MultinomialNB ()`.

Finally, this project also conducted experiments using the Bert pre-training model. The English data set mainly uses the `wwm_uncased_L-24_H-1024_A-16` [52] pre-trained model for experiments. In order to save time and waste of resources, epoch is set to 5. According to the experimental results, the best state is often reached in the 2-3 rounds. The Chinese data set mainly uses the `chinese_wwm_L-12_H-768_A-12`[52]pre-training model for experiments. Similarly, the epoch is set to 5. According to the experimental results, the best state is often reached in less than 5 rounds. It is worth noting that Bert has its own set of standards for word segmentation. We only need to input our sentences and emotional labels in the corresponding format to start training the model and test data. Unlike the previous data set division, this round of experiments is divided into training set, validation set, and test set according to 8:1:1.

Chapter 5 Results and discussion

In this chapter, the experimental results will be given and analyzed.

5.1 Results

According to the experimental procedures in the previous chapter, this project uses the accuracy rate in section 2.6 as the evaluation index. The experimental results are shown in the Table 5. It is worth noting that due to the large amount of WeiBo data, this project carried out a small sample experiment with about 4,000 experimental data, and then conducted experiments with about 30,000 experimental data.

Table 5 The experimental results

	BOW+ LogisticRegression	TF-IDF+ LogisticRegression	BOW+ naive_bayes	TF-IDF+ naive_bayes	Bert
Twitter	0.7295	0.7576	0.7250	0.7440	0.7829
Weibo (4,000)	0.2437	0.2675	0.2662	0.2850	0.2950
Weibo (30,000)	0.4524	0.4879	0.5025	0.5467	0.6047

5.2 Discussion

From the above experimental results, it can be seen that the results of social media sentiment analysis of Chinese texts are worse than those of English texts. This may mainly come from three aspects: First, the Chinese text involves word segmentation. By analyzing the results of word segmentation, it can be seen that the results of some word segmentation are not very consistent with Chinese grammar. At the same time, there are multiple word segmentation in a sentence. And the results of different word segmentation express different meanings. Second, the social media sentiment classification of Chinese text is a three-category problem. At the same time, it can be learned from the previous analysis that the keywords of disgusting emotion and angry emotion are relatively similar, so it is relatively difficult to extract their respective characteristics. Third, in the data set of Chinese text, the sentiment analysis classification is incorrect, and there are texts that are difficult to be classified as definite sentiment classification. Therefore, the results of sentiment analysis of Chinese text social media are poor.

At the same time, it can also be seen that when using the same method to extract features, Chinese text is more suitable for sentiment classification using Naive Bayesian algorithm, while English text is more suitable for sentiment classification using logistic regression algorithm. It can also be seen that Bert can indeed extract the features in the data better, thus showing the best accuracy on all data sets.

Finally, through the comparative experiment of the data volume of the Chinese data set, it can be seen that any algorithm needs more data to achieve better results and performance. Therefore, when performing data analysis, not only the quality of the data needs to be paid attention to, but also the quantity of the data.

Chapter 6 Conclusions and recommendations

6.1 Conclusions

This project carried out sentiment analysis on Twitter and Weibo according to the CRISP-DM method and solved the three goals of this project at the same time.

First of all, this project cleaned up the data set. In this work, it was discovered that sentiment analysis of Chinese texts and sentiment analysis of English texts are very different in word segmentation.

Then, through statistical analysis of the data set, the emotional words often used by different emotional tendencies were obtained. At the same time, it was unexpectedly discovered that people frequently use positive emotional tendencies in negative emotions.

Finally, through modeling, the sentiment analysis and classification of the data set is completed. At the same time, according to the experimental results, it is found that when Chinese text sentiment analysis and English sentiment analysis use the same features, different algorithms should be used.

6.2 Recommendations

Although this project has been completed, there are still many areas for improvement in this project.

Although Bert's performance in social media sentiment analysis is relatively good, the training time is particularly long. Because of the many parameters, it takes a long, long time to train. Therefore, in future work, can the model be modified to reduce the training time?

Secondly, it can be seen that the effect of social media sentiment analysis of Chinese text is not satisfactory, whether it is possible to use more complex and deeper neural networks to extract the more features of Chinese text, so as to improve the effect of social media sentiment analysis of Chinese text.

Chapter 7 Reflection

Just like an old Chinese saying, the sun and the moon fly by, and time flies like an arrow, which means that time passes very quickly. This project is nearing its end. Through this project, I have learned a lot and have a deeper understanding of data mining related knowledge.

From the day when the topic was released, I was very excited to choose Professor Nik Whitehead's topic on social media sentiment analysis. Because this topic is very attractive to me, I think the topic is very interesting. First of all, the content of social media is very rich. We can almost get people's views and comments on any things or articles that we want to know, and analyze the text information published by people, so that I can better understand people's views and thoughts. Secondly, the amount of information on social media is huge. I get what I want from the massive amount of information and perform sentiment analysis. This topic is very challenging and can also improve my ability and understanding of knowledge. Finally, sentiment analysis on social media is also of great commercial and practical value. Different types of text data can be selected for sentiment analysis for different needs. Social media allows us to obtain almost any required data. And social media has classified different text data according to the relevance of events or things, we can obtain them according to the needs easily. After the data is obtained, the sentiment classification can be used to understand people's opinions and emotions. This is very important, because people's opinions and emotions can often affect the development of things, such as the subsequent purchases of goods, policy changes, and subsequent movie box office. Therefore, when I saw the topic selection, I did not hesitate to choose social media sentiment analysis.

After identifying the topic, I started to read a lot of papers, and then found that sentiment analysis is a very complex topic. It includes sentiment analysis of long texts, sentiment analysis of short texts, explicit sentiment analysis and implicit sentiment analysis. Then I started to read about the related technologies of sentiment analysis, and finally chose to use machine learning and deep learning to experiment.

After writing a literature review and communicating with my supervisor, I started my data analysis and experiments. During the data analysis, the experiment progressed relatively smoothly, but later my server failed and the experiment could not be performed, and the experiment results could not be obtained for a long time. I am very grateful to my mentor, because when she learned that I did not get the results of the experiment in the later communication, she analyzed my project for me every time, and at the same time, he continued to encourage me. Therefore, in the later group meeting, I was not too satisfied with my own performance.

At the same time, in the later exchanges and discussions, I fine-tuned the goal of this topic, only to perform sentiment analysis on short texts in social media, so in the end only two data sets of Twitter and Weibo were selected for experimentation, and the sentiment of IMDB long texts was removed from my project. I am very grateful to my supervisor for his understanding and help, and constantly revising every part of the paper I submitted.

In the process of completing the project, in addition to server failures, I also encountered many problems with writing papers and implementing experimental codes. For the thesis writing problem, my supervisor gave me the greatest help and support, and smoothly solved my problems in the thesis writing process. Problems related to the experimental code were successfully solved by reading the information and checking the blog. Although there were many difficulties in the process of completing the project, my ability has been greatly improved.

In the end, with the help and encouragement of my mentor, I completed the project relatively smoothly. In this project, I completed the cleaning of the two data sets of Twitter and WeiBo. At the same time, I extracted the keywords of different emotional text information. At the same time, I modeled the emotional analysis of social media, which can effectively analyze the social media text and effectively classify the emotional tendency of social media information texts. At the same time, due to the selection of data sets in two different languages, the differences in the emotional analysis process of social media in two different languages are also compared.

Although I have completed this project, I think I can do better. First of all, in project analysis and goals, I can narrow down the scope of the goal and only conduct sentiment analysis for a certain type of social media topic, instead of analyzing the entire Twitter and WeiBo data, so that the problem has a small entry point, and the subsequent experimental settings may not require too much computing power.

Then due to the narrowing of the scope, the reduction of the amount of data, and the reduction of the time and computing power of the experiment, it may be possible to select more models for experimentation, and thus it is possible to select more suitable models for data sets in different languages. So as to better demonstrate the third goal of this project.

REFERENCES

- [1] Yue, Lin, Weitong Chen, Xue Li, Wanli Zuo and Minghao Yin. "A survey of sentiment analysis in social media." Knowledge and Information Systems (2018): 1-47.
- [2] Ekman P. An argument for basic emotions[J]. Cognition and Emotion, 1992, 6(3/4): 169-200
- [3] LiuB(2012)Sentimentanalysisandopinionmining.SynthLectHumLangTechnol5(1):1–167
- [4] Nasukawa T, Yi J. Sentiment analysis: Cap turing favorability using natural language processing[C]. Proceedings of the 2nd international conference on Knowledge capture. ACM, 2003:70-77
- [5] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," in The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, 2011, pp. 142–150, [Online]. Available: <https://www.aclweb.org/anthology/P11-1015/>.
- [6] Appel O, Chiclana F, Carter J, et al. A hybrid approach to the sentiment analysis problem at the sentence level[J]. Knowledge-Based Systems, 2016, 108: 110-124.
- [7] <https://pypi.org/project/jieba/>

- [8] <https://stanfordnlp.github.io/CoreNLP/>
- [9] Li Chuzhen, Yu Yuwen. Research on common methods of Chinese microblog data preprocessing[J]. Journal of Science & Technology Economics, 2019, 27(33):23.
- [10] Wang Ting, Yang Wenzhong. A review of text sentiment analysis methods[J]. Computer Engineering and Applications, 2021, 57(12): 11-24.
- [11] BACCIANELLA S, ESULI A, SEBASTIANI F. Senti-WordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining[C]//Proceedings of International Conference on Language Resources and Evaluation (LREC 2010) .Malta: LREC, 2010: 83-90.
- [12] KU L W, CHEN H H. Mining opinions from the Web: beyond relevance retrieval[J]. Journal of the American Society for Information Science and Technology, 2007, 58 (12): 1838-1850.
- [13] HowNet. Sentiment analysis vocabulary [EB/OL]. (2013-10-16) [2019-11-10]. http://www.keenage.com/html/c_index.html.
- [14] Xu Linhong, Lin Hongfei, Pan Yu, et al. The construction of emotional vocabulary ontology[J]. Journal of Information, 2008, 27(2): 180-185.
- [15] Asghar MZ, Khan A, Ahmad S, Qasim M, Khan IA (2017) Lexicon-enhanced sentiment analysis frame- work using rule-based classification scheme. PLoS ONE 12(2):e0171649
- [16] <https://baike.baidu.com/item/%E9%82%BB%E8%BF%91%E7%AE%97%E6%B3%95/1151153?fromtitle=Knn&fromid=3479559>
- [17] Su Yuyu. Research on Feature Selection and Weight in Chinese Text Sentiment Analysis[D]. Nanjing University of Posts and Telecommunications, 2017.
- [18] 王帅. 情感分析的特征提取算法与观点的聚类算法研究[D]. 昆明理工大学, 2015. (Wang Shuai. Research on the feature extraction algorithm of sentiment analysis and the clustering algorithm of opinions [D]. Kunming University of Science and Technology, 2015.)
- [19] Neumann J, Schnorr C, Steidl G. Combined SVM-Based Feature Selection and Classification. Machine Learning, 2005
- [20] 杨文维. 基于微博特征的情感分析方法研究[D]. 东北师范大学, 2018. (Yang Wenwei. Research on Sentiment Analysis Method Based on Weibo Features [D]. Northeast Normal University, 2018.)
- [21] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8):1735-1780
- [22] Kim, Y. Convolutional neural networks for sentence classification[C]. In Proceedings of the 2014 Conference on Empirical Methods for Natural Language Processing, 2014: 1746-1751.
- [23] 罗帆, 王厚峰. 结合 RNN 和 CNN 层次化网络的中文文本情感分类[J]. 北京大学学报(自然科学版), 2018, 54 (3): 459-465. (Luo Fan, Wang Houfeng. Chinese text sentiment classification combining RNN and CNN hierarchical network[J]. Journal of Peking University (Natural Science Edition), 2018, 54(3): 459-465.)
- [24] XING Y, XIAO C, WU Y, et al. A convolutional neural network for aspect-level sentiment classification[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(14).
- [25] 韩建胜, 陈杰, 陈鹏, 等. 基于双向时间深度卷积网络的中文文本情感分类[J]. 计算机应用与软件, 2019, 36 (12): 225-231. (Han Jiansheng, Chen Jie, Chen Peng, et al. Chinese text sentiment classification based on two-way time deep convolutional network[J]. Computer

Applications and Software, 2019, 36(12): 225-231.)

- [26] C.E.Shannon : A mathematical theory of communication, Bell System Tech.J., Vol.27, pp. 623—656,(1984)
- [27]D. A. Asamoah and R. Sharda, “Adapting CRISP-DM Process for Social Network Analytics: Application to Healthcare,” 2015, [Online]. Available: <http://aisel.aisnet.org/amcis2015/BizAnalytics/GeneralPresentations/33>.
- [28] 中国互联网络信息中心.第 38 次中国互联网络发展状况统计报告 [EB/OL]. <http://www.cnic.net.cn/>. 2016-08-03(China Internet Network Information Center. The 38th Statistical Report on Internet Development in China [EB/OL]. <http://www.cnic.net.cn/>. 2016-08-03)
- [29] 唐慧丰, 谭松波, 程学旗. 基于监督学习的中文情感分类技术比较研究[J]. 中文信息学报, 2007, 21(6):88-94. (Tang Huifeng, Tan Songbo, Cheng Xueqi. Comparative Research on Chinese Sentiment Classification Technology Based on Supervised Learning[J]. Journal of Chinese Information Processing, 2007, 21(6): 88-94.)
- [30] Troussas C, Virvou M, Espinosa K J, et al. Sentiment analysis of Facebook statuses using Naïve Bayes classifier for language learning [C]. IISA, 2013:1-6.
- [31] Mullen T, Collier N. Sentiment analysis using support vector machines with diverse information sources[C]. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004: 412-418.
- [32] Schmidhuber J. Deep learning in neural networks: An overview[J], Neural networks, 2015, 61:85-117.
- [33] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. computer science, 2014.
- [34] 张俊林, 深度学习中的注意力机制 [OL], <https://blog.csdn.net/malefactor/article/details/78767781>, 2017. (Zhang Junlin, Attention Mechanism in Deep Learning [OL], <https://blog.csdn.net/malefactor/article/details/78767781>, 2017.)
- [35] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.
- [36] 王金浩. 基于 BERT 嵌入的文本情感分类方法研究[D].山东科技大学,2020.(Wang Jinhao. Research on text sentiment classification method based on BERT embedding[D]. Shandong University of Science and Technology, 2020.)
- [37] 朱婧. 基于 Twitter 的情感分析相关问题研究[D].武汉理工大学,2014.(Zhu Jing. Research on related issues of sentiment analysis based on Twitter[D]. Wuhan University of Technology, 2014.)
- [38] 尹璐. 面向临床电子病历的实体链接方法研究[D].北京交通大学,2020.(Yin Lu. Research on entity link method for clinical electronic medical records[D]. Beijing Jiaotong University, 2020.)
- [39] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. 2013: 3111-3119.
- [40] Mikolov T, Chen K, Corrado G S, et al. Efficient estimation of word representations in vector space[C]// International Conference on Learning Representations, 2013.
- [41] Outang xi, Zhou Pan, Li ChengHua, et al. Sentiment analysis using convolutional;

Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 2015: 2359-2364.

- [42] Bradbury J, Merity S, Xiong C, et al. Quasi-recurrent neural networks[C]. In Proceedings of the International Conference on Learning Representations, 2017:568-577
- [43] Xue W, Li T. Aspect Based Sentiment Analysis with Gated Convolutional Networks[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2514-2523.
- [44] Zhang Y, Er M J, Venkatesan R, et al. Sentiment classification using Comprehensive Attention Recurrent models[C]. International Joint Conference on Neural Networks. IEEE, 2016: 1562-1569.
- [45] 王伟, 孙玉霞, 齐庆节. 基于 BiGRU- Attention 神经网络的文本情感分类模型[J/OL]. 计算机应用研究, 2019, 36(12): 56-60 (Wang Wei, Sun Yuxia, Qi Qingjie. Text sentiment classification model based on BiGRU-Attention neural network[J/OL]. Application Research of Computers, 2019, 36(12): 56-60)
- [46] Ambartsoumian A, Popowich F. Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers[C]//Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2018: 130-139.
- [47] Cho K, van Merriënboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.
- [48] 徐志远. 基于数据挖掘方法的电信客户流失分析[D]. 北京邮电大学, 2016. (Xu Zhiyuan. Analysis of Telecom Customer Churn Based on Data Mining Method[D]. Beijing University of Posts and Telecommunications, 2016.)
- [49] 潘无名, 潘云鹤. 数据挖掘过程的多维视图[J]. 计算机应用研究, 2004, 8. (Pan Wuming, Pan Yunhe. Multi-dimensional view of data mining process[J]. Computer Application Research, 2004, 8.)
- [50] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colla Shearer (SPSS) and Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0 Step-by-step data mining guide [M]. Clifton: CRISP-DM consortium, 2000.8.
- [51] https://github.com/dweiii/stop_words
- [52] <https://github.com/google-research/bert>