



Text Mining on Movie Reviews

By

Yu Xin

2006647

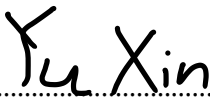
Supervisor: Dr Nik Whitehead

2021.11

DECLARATION

I, Yu Xin 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 : 2021.11.11

FORM OF CONSENT

I _____Yu_Xin_____ hereby consent that my Project, submitted in candidature for the M.Sc. Applied Computing, 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: 2021.11.11

Where necessary replace M.Sc. Applied Computing with your named award.

Ethics Approval Form

Before you complete this form, please take time to carefully consider the following questions:

Have you considered yet whether there are any problematic ethical issues in your proposed research project? If you have not you should talk to your Course Tutor or Supervisor.

Have you already completed an Ethics Approval Form? Yes – *then you do not need to complete this form*

Yes

No – please complete this form in as much detail as you can

Name:			
Project/Research Title:			
Name of Supervisor:			
School/Dept.			
Faculty:			
Proposed Start Date:		End Date:	

1. Summary of planned research (please indicate below the purpose of your planned project/research, together with your aims, main research questions and research design – you should continue onto a separate sheet if necessary)

--

2. Methodology (You need to be clear about the methodology you intend to use in your study; this could include any number of methods, so either tick those shown below (where appropriate) or put the details in the box marked *)

☐ Interviews	☐ Participant Observation	☐ Use of personal data
☐ Focus groups	☐ Questionnaire	☐ Literature Review
☐ Performance	☐ Presentation	☐ Other (state below)*

*

3. Participants - Does your proposed project/research involve human participants?

☐ 1. Yes, as a primary source*	☐ 2. Yes, as a secondary source*	☐ No- go to section 4
---	---	--

*If you have ticked yes, it is likely you will need an Advanced CRB check before undertaking your study

If yes, indicate who your participants are:

☐ Early years/Pre-School children

☐ Adults - give details below

☐ School age children

☐ Vulnerable people - *give details below*

☐ Young People aged 17-18

☐ Unknown at this stage

4. Ethical issues - you should tick all that apply

☐ Administration of drugs incl. alcohol

☐ Deprivation

☐ Unpleasant stimuli in any manner or form

☐ Active deception or withholding information

☐ Collection of highly personal information

☐ Payment

5. Are there any ethical concerns other than those listed above? (*continue onto a separate sheet if necessary*)

6. If there are any ethical concerns, please state how you intend to minimise any risk of harm or distress that could be caused (*continue onto a separate sheet if necessary*)

You should enclose any materials (e.g. questionnaire, interview schedule), plus the Consent Form, and the Debriefing Sheet when submitting the Ethics Approval Form to your supervisor.

Student Signature _____ Date _____

☐ Advanced CRB check required

☐ CRB confirmation received - Date: _____

☐ Recommendation of approval given at Faculty Level

☐ Approval not given at Faculty level – forwarded for discussion at the next meeting of the University Ethics Committee

Comments:

Signature: (Assistant Dean/Head of School) _____
Date: _____

Faculty recommendation endorsed by the Ethics Chair's Initials

Date

Abstract

In recent years, the film market has developed rapidly, and websites that score films have emerged one after another. These websites are full of netizens' comments and ratings on movies. These unstructured reviews contain a lot of valuable information. If we use this information, it will have a very good guiding significance for the film market.

Therefore, this paper studies the application method of text mining for movie reviews. Firstly, this paper mines the subject words of a certain category of film reviews, so as to find out what people are discussing about this category of films. Then, this paper trains the emotion polarity classification model to analyze the emotion polarity of comments, so as to achieve the purpose of public opinion monitoring. Finally, in order to make up for the deficiency of emotion polarity classification model in public opinion monitoring, this paper uses the pre trained semantic similarity calculation model to find the semantic related comments for a special comment. Specifically, it includes the following contents:

(1) **Theme words are extracted for a certain category of movies.** In this part, TF-IDF theme word extraction method is used to extract the theme words of high scoring films, low scoring films, Chinese animation films and Avengers Alliance II. For the extraction results of each problem, this part visually displays the results through word cloud, analyzes the results, and gives some valuable information for the film market.

(2) **The emotion polarity classification model is trained to analyze the emotion polarity of unlabeled text.** In this part, the ratings of comments are divided into positive, negative and neutral. The emotional polarity of the data is classified by fine-tuning the BERT pre training language model. In order to reflect the advantages of BERT, this part also uses LSTM, Transformers and CNN for comparative experiments. Finally, it can be seen from the results that the pre training language model does have the application value of public opinion monitoring in the classification of emotional polarity.

(3) **Similar comments are identified by calculating the semantic similarity between comments.** In order to make up for the deficiency of emotion polarity classification in public opinion monitoring, this part uses the SimCSE model to fine tune the model on the corpus to search for relevant comments for a single comment. Through the analysis of the results of this part, it is found that the comment semantic similarity monitoring based on SimCSE has better performance than the pre training language models such as Word2Vec and BERT.

Keywords: Text Mining, Subject Word Extraction, Emotion Analysis, Semantic Similarity, Film Reviews.

Table of Contents

Abstract.....	5
List of Tables.....	7
List of Figures.....	8
Literature Review.....	9
Introduction.....	9
1. Text mining on movie reviews.....	9
2. Early exploration of text mining	10
3. Topic extraction problem	10
4. Sentiment analysis.....	12
4.1. Introduction.....	12
4.2. Sentiment analysis based on sentiment lexicon	13
4.3. Sentiment analysis based on statistical machine learning	14
4.4. Sentiment analysis based on deep learning.....	14
5. Text similarity problem	15
5.1. Text similarity calculation based on keyword surface information.....	16
5.2. Text similarity calculation based on vector space model	16
5.3. Text similarity calculation based on semantic information	17
5.4. Conclusion of text similarity problem.....	18
6. Conclusion	18
CRISP-DM.....	19
1. Introduction for CRISP-DM	19
2. Business understanding.....	20
3. Data understanding	21
3.1. Data Collection	21
3.2. Data Description.....	22
3.3. Data exploration.....	23
3.4. Data quality inspection.....	24
4. Data preparation	24
4.1. Data Selection	24
4.2. Data Preprocessing.....	25
5. Methods Selection and Introduction	26
5.1. TF-IDF	26
5.2. Word2Vec.....	27
5.3. BERT.....	28
5.4. SimCSE.....	29
6. Experiment.....	30
6.1. Subject word extraction.....	30
6.2. Sentiment Analysis.....	34
6.3. Comment semantic similarity	40
Conclusion and Future Work.....	44
Personal Reflection	45
References.....	46

List of Tables

Table 1 .The Attribution of Douban Dataset	22
Table 2 . The general parameter settings of several models.....	36
Table 3 . The optimal experimental results of BERT	37
Table 4 . The experimental results of the four models on three different films.....	37
Table 5 . The training parameters related to SimCSE tuning	41

List of Figures

Figure 1 .The six steps in CRISP-DM	20
Figure 2 .The Screenshot of Douban reviews page.....	21
Figure 3 . The example of one data in dataset.....	22
Figure 4 . Time span of each film review release.....	23
Figure 5 . Number of reviews per film.....	23
Figure 6 . Number of comment lengths	24
Figure 7. The structure of CBOW	28
Figure 8. The structure of Skip-gram	28
Figure 9 . The basic structure diagram of BERT	29
Figure 10 . The structure diagram of unsupervised training model in SimCSE.....	29
Figure 11 . The word cloud of "What is good movie?".....	31
Figure 12 . The word cloud of "What is bad movie?".....	32
Figure 13 . The word cloud of "Monkey King: Hero is Back"	33
Figure 14 . The word clouds of "Avengers: Age of Ultron"	34
Figure 15 . The proportion of each category in the data set.....	35
Figure 16 . Change of loss value during training	36
Figure 17. Case study for Transformers 4.....	38
Figure 18. Case study for Free Guy	38
Figure 19. Experimental results of Word2Vec model	41
Figure 20. Experimental results of BERT model	42
Figure 21. Experimental results of SimCSE model	42

Literature Review

Introduction

The development of the film industry is of great significance. On the one hand, film can bring certain economic benefits, on the other hand, it also represents the degree of economic and cultural development of a country. In recent years, more and more scholars have devoted themselves to the research and creation of films. Since the 1950s, there have been a large number of studies on text mining. With the growth of various text data on the network, the significance of obtaining effective information from these texts is becoming more and more important. Text mining on large-scale film review texts can not only obtain the audience's views and attitudes towards the current film market, but also effectively supervise and manage public opinion. It is of great significance for film manufacturers and film practitioners.

1. Text mining on movie reviews

The application of text mining technology to the movie reviews is of great significance to the development of the film industry. Firstly, as a form of expression of the audience's attitude towards the film, comment not only includes the evaluation of the film, but also analyzes the reasons why you like or don't like the film. Secondly, the box office of released films is greatly affected by online comments. The audience who are ready to go to see the film can often be attracted by the films with good evaluation, and will also be discouraged by the comments with poor evaluation. Finally, for the film platform, film reviews are used as an important basis for user portraits. The platform can perform user clustering, advertising recommendation and other operations on users.

He [1] extracts subject words based on the positive and negative texts classified by naive Bayes according to the LDA subject model, so as to find the gap between Chinese animated films and animated films of other countries. The results show that most of the positive theme words in the United States and Japan are "special effects", "details", "Pixar", "Disney", "Hayao Miyazaki" and so on. These words show that the United States and Japan not only have advanced production technology, but also have large animation production companies or studios. In China, most of the negative theme words are "plagiarism", "imitation", "boredom", "young" and so on. This result shows that there is a lack of large animation production companies in Chinese animation films, and they fall into the dilemma of lack of excellent original works and good screenwriters. Xin et al. [2] also applied text mining technology and machine learning algorithm to film review analysis, and try to find the useful information behind user comments through descriptive analysis and emotional tendency. In descriptive analysis, Xin et al. [2] Used TF-IDF method to extract the theme words of "Avenger Alliance 4" and generate word clouds. They intuitively found that the plot and feelings of the film have become the key points to attract people to watch. On the other hand, the author combines naive Bayes classification with LDA subject word method for

emotional analysis, and finds that what attracts the audience of "Avenger alliance 4" lies in his special effects and heroic feelings behind him, while those who are dissatisfied think that the plot design of the film is general. Qu [3] quantizes the design features of film reviews, clusters them with k-means method, and then visually displays them through radar images. Through the analysis of the clustering results, the author comes to the conclusion that a good film often has excellent plot and good picture design, and the filmmaker generally does not attract the attention of the audience, but when the plot and filmmaker become the prominent shortcomings of the film, it will have a great negative impact on the total score of the film.

Through the above articles, we can see that text mining of film reviews is of great significance. The results presented by these studies have very valuable guidance and suggestions for the adjustment of film producers in the film market.

2. Early exploration of text mining

As early as the late 1950s, there have been relevant studies on text mining. In 1957, Luhn [5] put forward the idea of automatic text classification based on the statistics of word frequency in text. Shortly after that, in 1960, Maron et al. [6] proposed a probabilistic statistical method based on Naive Bayesian classification to automatically classify model text. In the 1970s, Salton et al. [7] proposed the vector space model, which provided the possibility for vector text classification. Around 1980, foreign text classification tended to the design of Feature Engineering, but it required a lot of human and material resources according to different tasks. In 1985, Zhu Qing et al. [8] introduced and studied the automatic document classification in combination with the development at that time. In 1992, Lewis [9] proposed a complete text classification system in detail. Since then, text classification methods based on statistics and machine learning have gradually sprung up. Hearst et al. [10] believe that after processing the comment text through a series of relevant methods through certain technical means, consumers' preferences for this kind of goods can be known through the description in the text language. In 2000, Eugene [50] and others of Columbia University designed snowball system to automatically perform data mining tasks in order to extract valuable structured data hidden in text documents. Snowball uses a supervised learning method to automatically extract the extraction mode according to the given training samples, and then applies this mode to plain text to extract valuable structured data in plain text on a large scale. After a comprehensive evaluation of the experiment, snowball achieved good results at that time, and the idea of automatic text mining proposed by snowball has a good reference significance.

Although these methods related to text mining appeared in the last century are relatively old. However, the ideas they left behind are very valuable and laid the foundation for the emergence of other methods.

3. Topic extraction problem

On the one hand, comment topic extraction can quickly find the current comment hot events. On

the other hand, it can find the user's own interest hot spots for subsequent product promotion, information push and other services. Therefore, theme word or sentence extraction is of great significance to tap the potential utilization value of users and maintain the sustainable development of the platform.

In 1976, Robertson [11] deeply explored the role of TF-IDF (Term Frequency–Inverse Document Frequency) algorithm in text feature extraction and weighting, and then creatively applied this calculation method to feature weighting of machine learning. It provides an important experimental basis for many later scholars to deal with the problem of feature weighting.

Deerwester [12] proposed LSA latent semantic model in 1998. This characteristic semantic model is considered to be able to effectively extract the association degree between language texts, so as to mine the subtle association between documents and words composed of a whole language text. However, due to the complexity of the LSA algorithm model, it is not suitable for large-scale commercial deployment. In 1999, Hofmann [13] expanded and sorted out the LSA model, mathematically simplified the calculation through the frequently used probability model, and greatly simplified the calculation process without losing accuracy. In 2001, Subasic et al. [14] proposed LSCrawler by analyzing text topics, and combine crawler technology with text topic extraction technology. It achieves the effect of self-learning language text grammar rules with machine learning method, and then crawling, which greatly improves the accuracy. In 2003, Blei [15] proposed the LDA topic model based on the PLSA probability model. The model creatively proposed that the text topic and the feature words corresponding to the topic have a logical prior distribution, which reduces the data dimension of the vocabulary corresponding to the topic and greatly improves the calculation speed.

Nowadays, PLSA and LDA, two classical topic word extraction algorithm models, have been able to do most of the topic word extraction tasks, and many scholars still improve them. In 2011, Zhang [16] proposed a microblog topic generation model based on MB-LDA. Li Jiyun [17] and Li Peng [18] improve LDA topic model by considering microblog comment context and forwarding factors respectively. In 2018, Feng et al. [19] proposed TIF-LDA topic extraction algorithm model with variable time window. In 2019, Cai [20] combined LDA and FCM algorithm to propose LDA-FCM algorithm, which not only reduces the dimension of text data, but also reduces the classification error rate caused by fuzziness. Zhang et al. [21] based on LDA topic model and divided according to user types, theme extraction was carried out for microblogs of different user types on microblog platform.

For LSA and other topic models, such models are only suitable for processing small text documents. With the growth of the number of texts, the number of words involved in language texts will increase exponentially, which greatly increases the amount of calculation of the model, which is beyond the range that the computer can bear. The LDA model reduces the data dimension of the vocabulary corresponding to the topic, which is more suitable for large-scale text computing. At the same time, it also achieves good results. However, the existing LDA models prefer long text, and the extraction effect of short text subject words is not ideal. Most topic word extraction algorithms suitable for short text topic mining mostly use TF-IDF method to measure the importance of words in comments only by word frequency, without considering the phenomenon

of synonyms and synonyms between words.

In the data set involved in this paper, after preliminary investigation, the vast majority of sentences are between 10-20 in length, that is, "short comments". So, this paper uses TF-IDF method suitable for short text to extract the subject words of comments, so as to find out where the audience is satisfied or dissatisfied with the film. TF-IDF is a statistical method used to evaluate the importance of a word to a document set or one of the documents in a corpus. Its core idea is that the importance of a word increases in proportion to the number of times it appears in the document, but decreases in inverse proportion to the frequency of its appearance in the corpus.

4. Sentiment analysis

4.1. Introduction

Text sentiment analysis began to enter the in-depth research in recent ten years, especially Chinese sentiment analysis.

There are many applications of emotional analysis of network comments for public opinion monitoring. For example, in the article [51], Wen used Bayesian method to conduct emotional analysis on the comment text on twitter, The emotional analysis results are visualized in the way of color depth. Wen's system provides a valuable reference for the monitoring of public opinion on the Internet of Public Security System.

At present, there are two commonly used text emotion analysis algorithms. One is the unsupervised method to match the emotional tendency by analyzing the sentence semantics, and the other is the supervised learning method to realize classification by training and learning using machine learning.

The unsupervised method of analyzing sentence semantics to match emotional tendency first carries out word segmentation and part of speech tagging of the sentence, and combines the tagged words in a certain way to form an emotional dictionary for judging sentence emotion. Then, according to the words in the sentence, score based on the emotional dictionary to judge the emotional intensity. In recent years, there are various methods to expand the emotional dictionary. The emotional dictionary based on these methods can achieve very good classification results. The classification process of this kind of method is artificially controllable and can also achieve a certain generalization. Although the use of emotion dictionary gets rid of the dependence on large-scale annotation data, it has a high dependence on emotion dictionary. However, the existing affective dictionaries are still poor in the representation of text semantics, especially for emerging words (OOV, Out Of Vocabulary) and polysemy. For Chinese text mining, the problem of polysemy is more significant because of the complexity of Chinese text and word segmentation without spaces.

The algorithm of sentiment analysis using machine learning is to simply classify complex emotions according to granularity. For example, in coarse granularity, they can be divided into

positive emotions and negative emotions, while in fine granularity, different categories can be defined according to different tasks, such as anger, despair, sadness and so on. In machine learning, first select the appropriate classification model and corresponding loss function according to the current data scale and data characteristics, and then represent the processed text as the input form of the model and input it into the model for supervised learning. Machine learning can be divided into traditional statistical machine learning methods and deep learning methods. Statistical machine learning supervision methods need a lot of manual feature design, which are excellent for specific tasks, but lack of generalization. The supervised learning method under deep learning can automatically learn the feature information in the text, avoiding the loss of time and energy caused by a large number of manual annotation. Whether traditional statistical machine learning methods or deep learning methods, supervised learning depends on a large number of labeled data, and the intermediate process of machine learning is more like a black box. It is difficult to carry out effective manual intervention for the found badcase.

Compared with English sentiment analysis, the effect of the same method is almost the same in Chinese text except that special processing is required in the preprocessing part. Chinese sentiment analysis has a good application effect and will have a very broad application prospect. However, most methods do not have much improvement in part of speech combination, still use the English model of part of speech combination, and do not adjust the Chinese semantics according to the unique usage habits of Chinese, so they need to be further improved.

4.2. Sentiment analysis based on sentiment lexicon

Since 2000, sentiment lexicon methods have been popular and widely used in various NLP tasks. Sentiment lexicon is to score the emotional polarity of each word in the text. For example, there is a sentence "I like this movie". After comparing each word with the words in the constructed emotional dictionary, we find that the emotional tendency expressed by the word "like" is mostly positive. Then, the sentence can be defined as positive meaning by weighted summation according to all emotions in the sentence.

In 2005, Kolchyna et al. [22] first proposed that unsupervised machine learning text classification method and point-to-point mutual information algorithm can be used to expand emotional vocabulary, Then, the reference word is used to expand the positive and negative emotion words to build an sentiment lexicon, and then the similarity between the word to be labeled and the reference word is calculated to judge whether the word has the same emotion type as the reference word. In 2006, Ghose [23] first proposed that the language information conveyed by the text and the emotion expressed by the text should be considered when using information technology to process the text. In 2007, Pan et al. [24] creatively defines the concept of the centripetalism of a word in the whole article based on the emotional polarity dictionary, and then uses this concept to express the correlation between words and categories. Such an affective lexicon can not only compare the intensity of different affective words, but also compare the correlation between word categories, which improves the analysis method of text data. After that, Ye [25] summarizes the subjective characteristics of Chinese text to a certain extent from the perspective of parts of speech, and preliminarily explores the method of Chinese text subjective automatic recognition. Based on

sentiment lexicon and word frequency, Zhou et al. [26] briefly introduced the main purpose, main task and mainstream technology of emotional dictionary in text mining. Sun et al. [27] believe that to judge the emotional category of a word, we can compare the degree of agreement between it and the benchmark word. On this basis, a method of text emotion analysis with words as the minimum analysis unit is proposed. In 2018, Yu Yang [28] also conducted text analysis based on sentiment lexicon.

4.3. Sentiment analysis based on statistical machine learning

Around 2010, statistical machine learning methods became the mainstream of research. Although the emotional dictionary can recognize the emotional information of many words, new words in comments emerge in endlessly, and it is difficult to update the emotional dictionary in real time. For example, for the comment "picture: (", the combination of a colon and an open parenthesis may not appear in the old emotion dictionary, but this symbol expression has a strong negative emotion. The statistical machine learning method can automatically identify these new emotional words according to the real-time updated corpus, and get rid of the time and energy consumption of manually constructing emotional dictionary.

In 2002, Pan et al. [29] obtained comment data related to the production, sales and cost of various products from Amazon and used LingPipe tool, a supervised learning model is established to calculate the score of emotion. In the same year, Bo et al. [30] conducted emotional classification on microblog data. They first distinguished positive and negative texts from microblogs, Support vector machine and Bayesian classifier are used to classify, and good results are achieved. Turney et al. [31] uses word bag distribution and machine learning model to classify text information based on twitter data. In 2006, Su et al. [32] made a detailed overview of text classification in machine learning, including the discussion of text emotion classification at that time. In 2014, Medhat et al. [33] used film review texts as corpora and adopted naive Bayesian algorithm Three machine learning methods, maximum entropy model and support vector machine, are used to classify the emotion of the corpus composed of relevant film reviews. Similarly, Zhu [34] carries out feature extraction and emotion classification of text based on SVM, and studies and analyzes this method. In 2017, Liu Shuang [35] analyzed and summarized the application of traditional machine learning in text emotion analysis.

4.4. Sentiment analysis based on deep learning

After the deep learning model is proposed, the low dependence on feature engineering and the high generalization of the problem make the deep learning method gradually replace the original machine learning method. For example, in some data science experiments, people use the deep learning method to get rid of the dependence on the manual construction of model functions and automatically learn the function expression between input X and output Y.

In the initial stage of deep learning, although various deep learning models can get rid of feature construction, the effect is still not as good as artificially extracting some key features as the input of the model. However, with the gradual development of deep learning, the static word vector

represented by Word2Vec makes it possible to completely get rid of the construction of input features. The function of word vector represents a word as a high-order and dense vector in vector space. Before deep learning, the construction of word vectors is often based on distributed assumptions or manually constructed features, and the semantic information is very lacking. Word2Vec, Glove and other static pre training word vectors obtain vectors with rich semantic representation through large-scale training, and transfer these vectors to specific tasks as the input of the model. Static word vector transfer learning has reached a new SOTA in various NLP tasks, and has become a milestone in the NLP field. In 2019, Wang et al. [37] used Word2Vec to convert text into word vector for film reviews, then used convolutional neural network to classify film reviews, and compared the final results with traditional machine learning methods such as SVM. They got a higher accuracy by that method. Gu Qian et al. [38] used deep learning method to learn word vector for text classification. The STL-FastText model she constructed has achieved good results in classification.

However, although the transfer learning method based on static word vector can bring sufficient semantic information representation to the input text, the vector representation of each word is constant and does not change with the change of context. This leads to some deficiencies in the representation of polysemy by static word vectors. On this basis, the dynamic word vector was born, which has become another milestone in the field of NLP. Different from the static word vector, the vector representation of each word of the dynamic word vector is calculated by its current context.

For example, there are two sentences "I like this movie" and "I don't like this movie". In the static word vector, the expression of the word "like" in the two sentences is exactly the same. If we want to judge the emotional polarity of this sentence, we can only judge and identify it according to the negative word "no" in front of "like". In the dynamic word vector, the word vector of each word is generated dynamically according to the context. Similarly, in the above two sentences, the word vector of "like" in the first sentence will be different from that in the second sentence. The difference is that the negative word before "like" is found in the first sentence, and this information is embedded in the vector representation of the word. Compared with static word vector, dynamic word vector is obviously better, and each word contains richer contextual semantic information.

The representative ones are BERT, ElMo and GPT large-scale pre training word vectors. In 2020, Liu et al. [39] cited BERT to model the emotion of short conversations, so as to predict the emotion of the target characters in real time, which not only shows the possibility of judging the emotional fluctuation of the characters, but also shows the effectiveness of BERT in emotion analysis.

5. Text similarity problem

Semantic similarity calculation also plays an important role in text mining. The task of semantic similarity calculation in text mining is similar to text clustering, but their application scenarios are slightly different. The calculation of semantic similarity in the process of text mining can find

similar texts in the corpus according to the given text, so as to achieve the application purposes of recommendation, classification, public opinion monitoring and so on.

After years of efforts, many researchers at home and abroad have made rapid development in the research of text similarity. At present, there are three mature similarity calculation methods: text similarity calculation based on keyword surface information, text similarity calculation based on vector space model and text similarity calculation based on semantic information.

5.1. Text similarity calculation based on keyword surface information

In the above three calculation studies, the most widely used is the similarity calculation method based on keyword surface information. This method usually only considers the surface information such as word form similarity, word order similarity and sentence length similarity of words in text or sentence to calculate the similarity. Lv et al. [40] only considered the word form similarity and word order similarity in the text when calculating the text similarity, without using any other syntactic or semantic features. Wang et al. [41] not only consider word form similarity and word order similarity, but also consider the same number and length of word strings and corresponding weight information. Li et al. [42] used the information entropy of the public string to evaluate the text similarity. In addition to the above methods based on surface information such as word form and word order, many scholars have considered the syntactic features of sentences. Lan et al. [43] Designed the similarity calculation method between sentences according to the sentence structure.

However, these methods are based on word matching, and the methods that incorporate sentence syntactic information also ignore the semantic information of words in the text, which is not comprehensive and accurate. For example, the sentences "The picture of the film is very excellent" and "The special effects of the film are very good" express the same meaning, but we can't judge them as similar through keyword matching, because they don't intersect the keywords.

5.2. Text similarity calculation based on vector space model

The method based on feature vector space refers to using feature space vectors to represent the text, and then calculating the similarity between each vector in the space to represent the similarity between sentences. This method of space vector modeling is called Vector Space Model, VSM. In the vector space model, we first extract the features of the text data to obtain n feature items, then map these features to an n-dimensional vector space for representation, and then calculate the text similarity in the vector space. VSM method has been widely used in text classification, document duplicate checking, text clustering and so on. However, the traditional vector space model mostly uses distributed semantic hypothesis to represent text, which makes the vector space have dimension explosion and sparse data. Moreover, the distributed semantic hypothesis only considers the context information of words and does not consider the semantic information between words. To solve this problem, many scholars extend the text semantically by introducing external knowledge bases, such as WordNet and Wikipedia. Such as [44], [45] and [46]. However, this method is very dependent on the improvement of external knowledge base and the operation

is complex. For the problem of too high dimension, some scholars reduce the dimension by singular value decomposition of high-dimensional matrix, and the classic LSI hidden row semantic index. Some scholars ignore the low-frequency words and other information in the text through LDA and other topic models, so as to reduce the spatial dimension.

5.3. Text similarity calculation based on semantic information

The text similarity calculation method based on pre training language model is similar to the method based on vector space. It also maps the text to high-dimensional vector space for similarity calculation. The difference is that the language vector representation obtained by the method based on pre training language model is obtained by pre training on large-scale corpus. For example, there are the following two words in the data set, "actor acting" and "Will Smith's performance". In the only data set, it is difficult for us to determine that will smith is an actor through the training context, because Will Smith appears very few times in the data set. However, through transfer learning, we can directly obtain the pre trained word vector of "Will Smith", which contains the meaning of "actor". The word vector obtained by this transfer learning method has sufficient semantic information and has good performance in various downstream tasks, including semantic similarity calculation.

Representative are pre trained word vectors such as Word2Vec, BERT, ELMo and GPT. Word2Vec has fast computing speed and strong semantic representation ability, and has been applied to various fields of NLP. However, Word2Vec itself is a static word vector, and there will also be OOV (Out Of Vocabulary) problems and poor representation ability of polysemy. The dynamic word vector represented by BERT solves this problem well. The word vector generated by BERT will change dynamically according to the context. BERT has reached a new SOTA in all fields of NLP.

However, many researchers have found that BERT does not have a good effect on the representation of sentence vectors, and even the effect is not as good as Word2Vec and other static word vector models. The lack of representation ability of sentence vector directly leads to some deficiencies in the calculation of semantic similarity of BERT. Gao et al. [47] carried out dimensionality reduction visualization on the centralized vector representation and found that the transformer word vector is conical in the vector space compared with the Word2Vec word vector, which affects the representation ability of the word vector, The calculated sentence similarity has poor discrimination ability, which is also called representation degradation phenomenon. Li et al. [48] also found that the spatial distribution of BERT word vectors was anisotropic. In the word vector space of BERT the distribution of high-frequency words is dense and closer to the origin as a whole, while the distribution of low-frequency words is sparse, resulting in the overall semantic information in the sentence being dominated by some high-frequency words.

In 2021, Gao et al. [49] proposed SimCSE, which used the idea of contrastive learning to enlarge the distance of heterogeneous word vectors, thus enlarging the whole vector space and alleviating the problem of anisotropy in vector space. Meanwhile, Gao et al. made two dropouts in the encoder part of the model representation to enhance the data and achieve the SOTA on the semantic similarity task (SST).

5.4. Conclusion of text similarity problem

To sum up, when calculating text similarity, the method of constructing vector space based on feature representation has the problems of too high dimension and incomplete information. The emergence of pre trained word vector solves the problem of lack of semantics in sentence representation, while the static word vector represented by Word2Vec has poor ability to express the context of words. The dynamic word vector represented by BERT has achieved SOTA in major NLP problems and is widely used because it dynamically generates word vectors according to context. SimCSE solves the problem of BERT vector space anisotropy to a certain extent through the method of comparative learning, and surpasses BERT in the task of semantic understanding through data enhancement. Therefore, this paper also uses SimCSE model to calculate semantic similarity according to a specific review, and then find other similar comments.

6. Conclusion

A large number of review data are generated on the Internet every day, and there are more and more research methods for film review. Researchers have adopted many research methods for film reviews, from the analysis of the modeling idea of traditional statistics to the information extraction and visualization of the massive supervisor review text data obtained in combination with big data, showing a variety of processing methods. Efficient and excellent processing methods can provide data support for the development of the film market.

CRISP-DM

1. Introduction for CRISP-DM

Cross industry data mining standard process (usually known for its abbreviation CRISP-DM) is a data mining process model, which describes the common methods used by data mining experts to solve problems. CRISP-DM model is a kind of KDD model. It occupies a leading position in various KDD process models in recent years. It is jointly developed by analysts of DaimlerChrysler, SPSS and NCR. CRISP provides an open and freely usable standard process of data mining, which makes data mining suitable for the problem-solving strategy of commercial or research units. CRISP DM generally includes the following six standard stages:

- (1) Business understanding stage. Identify the business objectives to be achieved and transform them into data mining topics. Business understanding can also be said to be a stage of business / research understanding and determining analysis objectives.
- (2) Data understanding stage. It includes three steps. Firstly, find out the possible factors affecting the subject, and determine the data carrier, data embodiment and data storage location of these factors. Secondly, we should have a preliminary understanding and simple description of the data. Finally, we should analyze the potential information in the data and put forward reasonable assumptions.
- (3) Data preparation phase. Transform and combine the previously found data to establish the required format of data mining tool software.
- (4) Modeling phase. Apply software tools, select appropriate modeling methods, process the prepared data table, and find out the hidden rules in the data.
- (5) Model evaluation phase. Evaluate the model conclusion from the perspective of business and statistics.
- (6) Model release phase. Also known as model deployment. Deploy the model for practical commercial application.

The flow chart of the six stages is shown in the Figure 1:

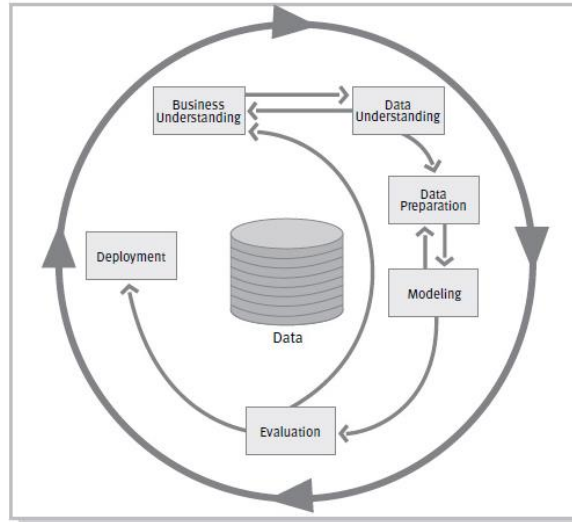


Figure 1 .The six steps in CRISP-DM

2. Business understanding

With the popularity of the Internet, more and more people tend to take the initiative to express their views, preferences, emotions and so on. In the film field, with the gradual expansion of the film market, film review websites such as rotten tomato and Douban appear. After registering, users can evaluate and score a movie on the platform. The platform will comprehensively calculate the word-of-mouth score of a film according to the scores given by all scoring users. More and more people will choose to refer to the movie scores on the platform and browse some popular comments before going to the cinema to see a movie, so as to judge whether they are willing to buy tickets to watch the movie. This forms a viewing mode of online reference, offline viewing and online scoring.

On the one hand, these comments are presented in the form of unstructured text, which is difficult to deal with directly. On the other hand, there is a huge amount of data in film reviews, and a simple processing mode can not deal with all situations. Therefore, the efficient modeling of these large-scale data to mine the hidden information is of great significance to the film review platform and even the film publishers and producers.

There are many applications for text mining of film reviews. These applications are designed for different problems and the information they want to mine. This paper mainly selects the following background problems for targeted application design:

- (1) A large number of comments contain the user's key evaluation information of film comments. These key information exists in unstructured text information. The film producer cannot get the overall feedback of netizens through a large number of existing film comments.
- (2) The comment text is often highly subjective. When the film is released, the malicious bad comments of various social platforms often have a negative impact on the box office of the film. Platforms or film publishers need to monitor online public opinion to prevent such phenomena.

- (3) Some movie review platforms will recommend users with similar preferences according to users' viewing preferences, so as to further recommend movies. On the other hand, in terms of network public opinion detection, platform side can also find the text of the same malicious comments through similar similarity search, and eliminate the similar malicious comments together.

Aiming at the above three background problems, we can set the application objectives of this time, that is, automatically extract text keywords according to categories, automatically analyze the emotion of comments without scoring text, and find similar comments according to a comment.

Because the comment data itself has the user scoring situation, the results of emotion analysis can be compared with the scoring situation to obtain an objective effect score, so as to measure the quality of the emotion analysis model. In addition, both keyword extraction and similarity calculation can only be evaluated subjectively. In other words, after the results are obtained, the subjective evaluation needs to be carried out according to whether the results meet the subjective cognition and whether the model is useful and efficient.

3. Data understanding

3.1. Data Collection

Douban film is a Chinese website that allows netizens to share their comments and views on the film. Users can post short or long comments on movies and rate them. The screenshot of the short comment section of Douban film network is as follows:



Figure 2 .The Screenshot of Douban reviews page

As can be seen from the above figure, similar to most film review platforms, netizens can post their comments on the film in the short comment part of Douban film and give corresponding scores. The short comment is exactly what we need. With the gradual improvement of Douban web crawler technology, it is not easy to crawl the latest comments from Douban in real time. Therefore, we found the Douban film review dataset previously crawled by other developers from Kaggle [56]. All users registered on Kaggle can download the dataset for free.

3.2. Data Description

The data set contains more than 2 million short Chinese comments on 28 films on Douban film website. There are many kinds of films selected, including Chinese films such as “Operation Mekong”, foreign films such as the “Marvel's The Avengers”, high-score films such as “Zootopia” and low-score films such as “Transformers 4”. The attribute information contained in each data set is as follows:

Column	Description	Data type
ID	Unique token of data	number
Movie_Name_EN	English name of the film	text
Movie_Name_CN	Chinese name of the film	text
Crawl_Date	Time of crawling this comment	date
Number	Comment number under a movie	number
Username	The name of the user who commented	number
Date	Time of comment	date
Star	Comment score given (1 for worst, 5 for best)	number
Comment	Comment text	text
Like	Number of likes obtained	number

Table 1 .The Attribution of Douban Dataset

An example of complete data is given below:

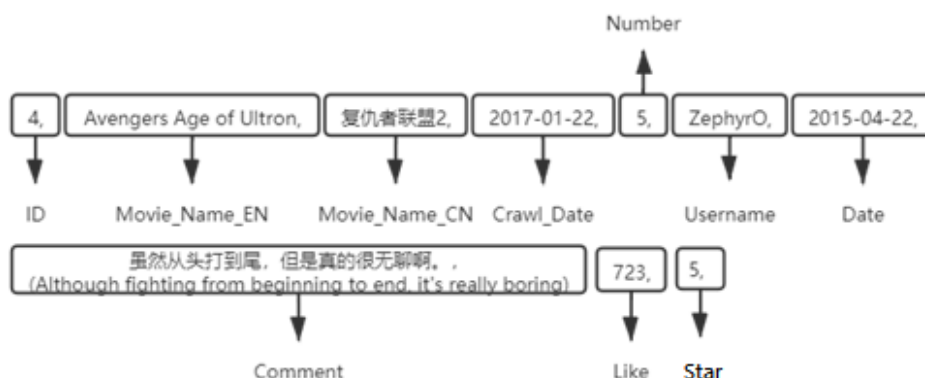


Figure 3 . The example of one data in dataset

3.3. Data exploration

As can be seen from the above figure, each data gives the corresponding movie name to be evaluated, so we can classify and mine and analyze according to the movie name, so as to obtain the advantages and disadvantages of each movie in the eyes of users. In addition, each data gives the score given by the user. We can use these numerical scoring and comment texts for text mining to obtain the emotional features.

Now perform statistical visualization on the data as a whole. First, the time distribution of each film and the number distribution of film reviews are as Figure 4 and Figure 5.

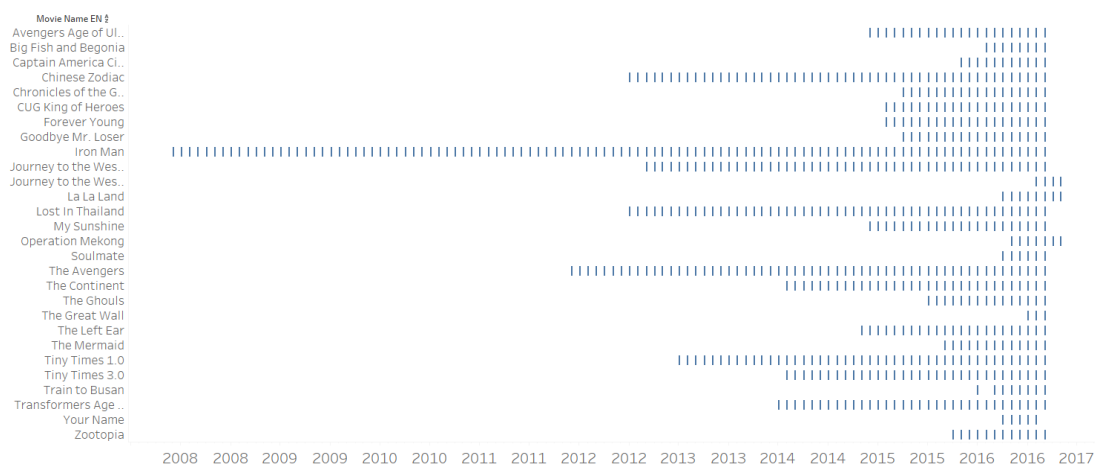


Figure 4 . Time span of each film review release

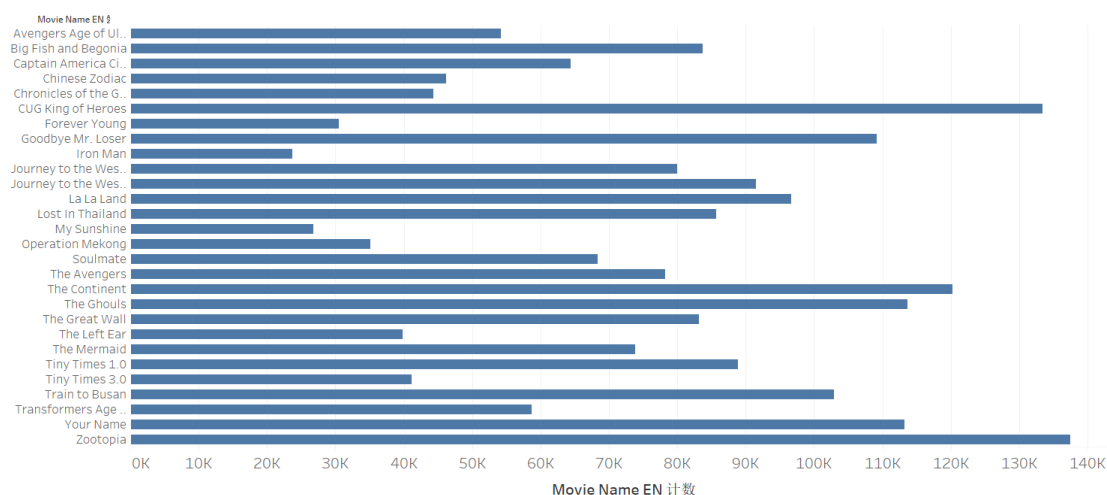


Figure 5 . Number of reviews per film

As can be seen from the Figure 4 and Figure 5, the latest comment data contained in the dataset was published in early 2017, which is also the time point when the author crawled to make the dataset. A small number of films are relatively new, so the comment interval is more concentrated after the release of film reviews, while for some films that appear earlier, such as “The Iron Man”, the comment covers all comments since 2008. On the other hand, we can see that although “The

Iron Man” has a large time span, the number of comments is the least. This shows that there is no absolute correlation between the release time of the film and the number of comments, and the number of comments is often determined by the popularity of the film.

Next, we need to analyze the length information of the main part of the data, that is, the comment part. The specific analysis results are shown in the figure below:

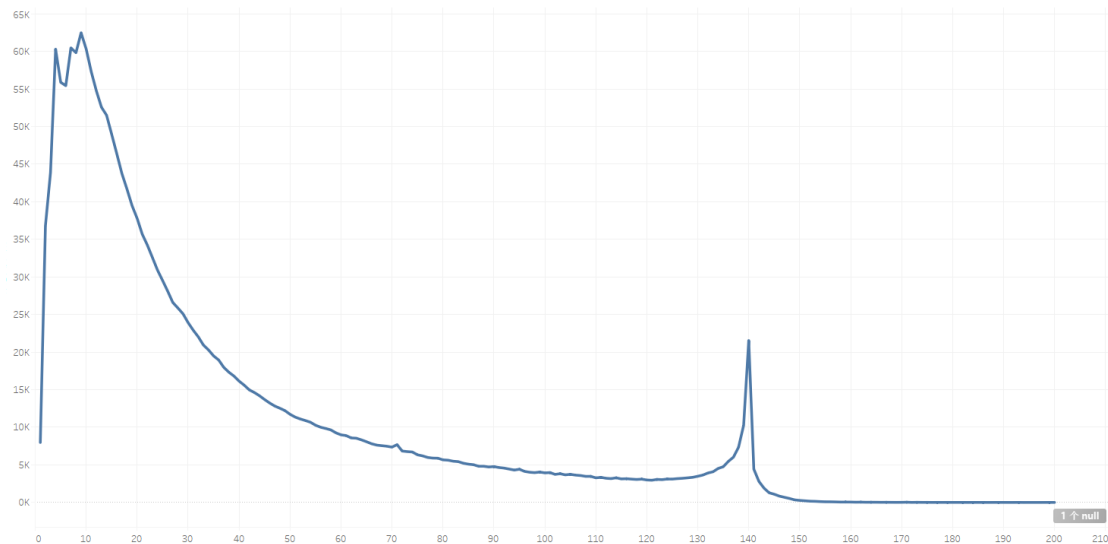


Figure 6 . Number of comment lengths

As can be seen from the Figure 6, most of the comments are distributed within 20 Chinese characters, which is what we call "short comments". A small number of comments will exceed 20 Chinese characters, up to 210 Chinese characters. It is worth noting that the curve in the figure has a high point, because Douban website had a word limit for short comments, that is, 140 words. Later, this restriction was lifted.

3.4. Data quality inspection

Because the quality of the comment text itself cannot be inspected, the data quality value mentioned here is the integrity of the data. After traversing 2 million pieces of data and judging whether they contain empty attributes one by one, it is found that the data set has high data integrity and there is no empty data phenomenon. Some of these users' names are displayed as "logged off", which indicates that the user no longer exists and is not a problem with the data collection itself.

4. Data preparation

4.1. Data Selection

Select the data of the following specific attributes according to the differences of the three tasks.

(1) Theme word extraction task:

The main purpose of theme word extraction task is to divide movies into multiple categories, and then extract keywords according to categories, so as to compare various movies. For example, when comparing the difference between high score films and low score films, it is necessary to select film reviews with higher scores and film reviews with lower scores. Then remove many films with neutral scores, because we can assume that the emotional polarity of the reviews of these films is not obvious. After selecting data, irrelevant attributes need to be eliminated, and only the "comment" text data can be retained here.

(2) Emotion analysis task:

The emotion analysis task mainly classifies the emotion polarity according to the comment text. All data is retained here. For each piece of data, only the "comment" and "score" attributes can be retained.

(3) Semantic similarity task:

The main purpose of semantic similarity task is to find several comments with the closest meaning to a comment. In this task, a query review needs to be selected, and a matching library is needed. In order to avoid too much data in the matching library leading to too long calculation time, it can be assumed that a query may have semantic relationship only with the same comments as its movies, but not with the comments of other movies. Therefore, for a query, only the comments under the movie to which the query belongs are selected as the matching library. In this step, after filtering according to the movie name, the two attributes of "comment" and "score" are also retained.

4.2. Data Preprocessing

Data preprocessing is mainly for the comment text. It includes two parts: word segmentation and removing stop words.

(1) Word segmentation

Because the Chinese dataset is used, there is no space between each word in the English dataset for segmentation, so it needs to be segmented manually. As a necessary preprocessing step of Chinese NLP problem processing, word segmentation has been proposed by more and more Chinese scholars. The existing methods can be divided into two kinds: Dictionary forward matching algorithm and machine learning method. Machine learning method has high precision, but the speed is slow, depends on a large number of data sets, and the process is complex, and the intermediate process is uncontrollable. Many existing applications based on forward matching algorithm can achieve good results, which is enough for industrial applications. Here, I use the commonly used Chinese word segmentation tool Jieba [57] to segment words.

Jieba also uses the forward matching algorithm. The inter process is very simple, that is, the maximum forward matching is carried out according to the existing manually constructed

common thesaurus. The pseudo code of forward matching word segmentation is as follows. The advantage of Jieba is that it has a very sufficient Chinese dictionary.

```
begin = 0
end = min(begin + max_len_word, len(sentence))
word_seg_list = []
while begin < end:
    word = sentence[begin:end]
    if word in word_set or end - begin == 1:
        word_seg_list.append(word)
        begin = end
        end = min(begin + max_len_word, len(sentence))
    else:
        end -= 1
return word_seg_list
```

(2) Removing stop words

Removing stop words is to remove our common meaningless Chinese words to reduce the redundancy of data representation. Removing a stop word is to find out whether it exists in the general word thesaurus according to the result of word segmentation. If it exists, remove the word. This project selects the Chinese stop word thesaurus released by Baidu [52].

(3) Building word-list

The function of word-list is to complete the one-to-one mapping of words to numbers. For example, for a word sequence {A, B, C, D}, it needs to be mapped to the corresponding ID, such as {23, 12, 43, 54}, as the input part of the model. The purpose of this is because the model needs numbers as input rather than words. Here, we directly call the dictionary method of Gensim [55] toolkit to build a corpus dictionary, and complete the use of the mapped vocabulary through `dictionary.token2id()`. It should be noted that when using BERT for emotion analysis, we do not need the construction of thesaurus. In the BERT pre training language model, there is already a specified vocabulary.

So far, we have completed the data preprocessing and can further model or mine the data.

5. Methods Selection and Introduction

5.1. TF-IDF

As mentioned above, the vast majority of comments in the data set are between 10 and 20 words in length. This paper uses TF-IDF (term frequency – inverse document frequency) [19] method suitable for short text to extract the subject words of comments, so as to find out where the audience is satisfied or dissatisfied with the film. TF-IDF is more stable than LDA [15] and LSA [13], and will not be affected by the number of texts and sentence length.

TF-IDF is a common weighting technique for information retrieval and data mining, which is used to evaluate the importance of a word to a file set or one of the files in a corpus. His core idea is that the importance of a word increases in proportion to the number of times it appears in the document, but decreases in inverse proportion to the frequency of its appearance in the corpus. His calculation formula is as follows:

$$TF = \frac{\text{The number of times a word appears in the text}}{\text{The number of words that appear the most in the article}}$$

$$IDF = \log \frac{\text{Number of documents in Corpus}}{\text{Number of documents containing the word} + 1}$$

$$TF-IDF = TF \times IDF$$

It can be seen from the above formula that TF represents the number of times a given word appears in the article, similar to word frequency. If we judge the subject word of a sentence only based on word frequency, it is easy to be affected by general high-frequency words. Therefore, IDF weights the TF value by calculating the inverse document frequency of the word. The calculation of TF-IDF can filter out common words, retain important words belonging to the text, and finally serve as our subject words.

5.2. Word2Vec

Word2vec is a group of correlation models used to generate word vectors. These models are shallow and double-layer neural networks used for training to reconstruct linguistic word texts. The network is represented by words, and the input words in adjacent positions need to be guessed. Under the assumption of word bag model in Word2Vec, the order of words is not important. After training, Word2Vec model can be used to map each word to a vector, which can be used to represent the relationship between words. This vector is the hidden layer of neural network. Word2vec includes Skip-Gram and CBOW.

The Continuous Bag Of Words model (CBOW) predicts the current word through the words in the surrounding window. In this model, words such as sentences or documents can be expressed in a bag with these words, which does not consider grammar and word order. The model structure is shown in the Figure 7.

Skip-gram model is a simple but very practical model. Contrary to CBOW model, it uses neural network to predict surrounding words through current words. The model structure is shown in the Figure 8.

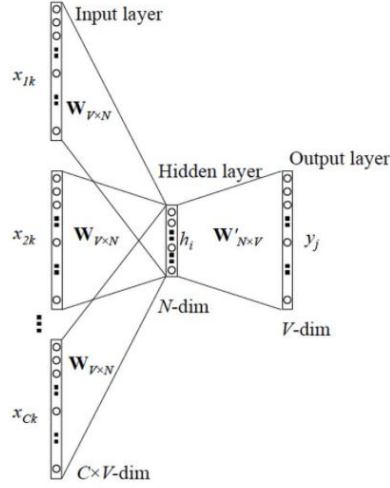


Figure 7. The structure of CBOW [54]

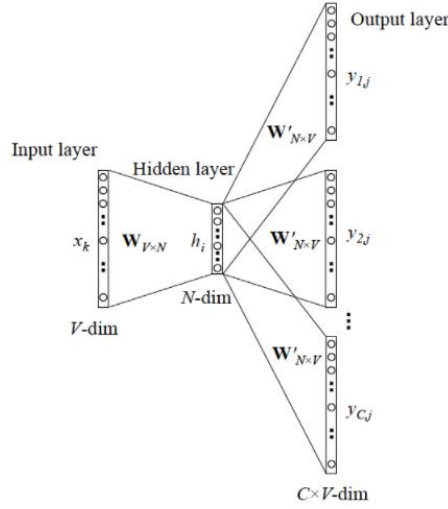


Figure 8. The structure of Skip-gram [54]

5.3. BERT

BERT (**B**idirectional **E**ncoder **R**epresentation from **T**ransformers) [53] is a pre training language model based on deep transformer proposed by Devlin et al. In 2018. BERT not only uses large-scale unlabeled text to mine rich semantic information, but also further deepens the depth of natural language processing model. The basic model structure of BERT consists of multi-layer transformer, including two pre training tasks: masked language model (MLM) and next sense prediction (NSP). MLM task is to cover some words in the sentence through the [mask] mask during training, and then restore them to the original words through the deep transformer model. NSP task is a binary classification task, which is used to predict what the next sentence of the sentence is. The basic structure diagram of BERT is as follows:

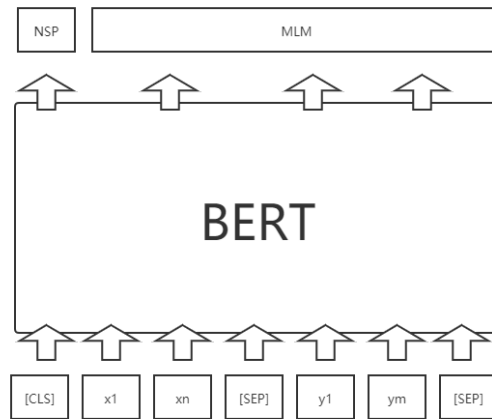


Figure 9 . The basic structure diagram of BERT

As can be seen from the figure, the input of the model is composed of two pieces of text x and y , and then the context semantic representation is obtained through BERT modeling, and finally MLM and NSP are learned.

5.4. SimCSE

SimCSE (**S**imple **C**ontractual Learning of **S**entence **E**mbeddings) [49] model is a model for sentence similarity calculation proposed by Gao et al. in April 2021. SimCSE model is based on the pre trained BERT word vector and finetuned on a specific corpus. Fine tuning adopts the method of comparative learning to improve the performance of BERT in sentence similarity task by enlarging the vector representation between dissimilar sentences and narrowing the vector representation between similar sentences. SimCSE adopts the unsupervised method. For each sentence, by adding noise to construct positive examples, SimCSE predicts the sentence itself in the process of comparative learning, and achieves the SOTA of STS task. SimCSE model supports two training methods, unsupervised training and supervised training. Unsupervised training does not need marked corpus, which meets the conditions of our existing corpus. The structure diagram of unsupervised training model is as follows:

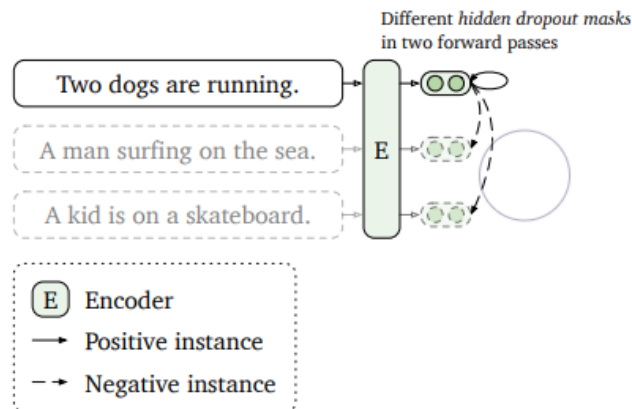


Figure 10 . The structure diagram of unsupervised training model in SimCSE

As can be seen from the figure, after the sentence is input, the sentence will be transformed into vector representation through the encoder part of BERT. In encoder, the model uses two different dropouts to artificially make noise and obtain two different representations of the same sentence. In contrast learning, the artificially constructed noise enhances the data in a disguised manner, which makes the training of the model more sufficient.

6. Experiment

6.1. Subject word extraction

Text theme word is the smallest unit of understanding text. Through the extracted keywords, the meaning of text expression can be grasped faster. For example, for the previous comment "Although fighting from begin to end, it's boring", we can extract two keywords "fighting" and "boring". Through these two words, we can quickly know the main meaning of the comment: playing more and boring. By extracting theme words from a large number of texts and selecting representative words, it can help people quickly grasp the theme of some texts and reduce the waste of time.

Firstly, this paper preprocesses the text, including word segmentation and stop word removal. Then, according to different needs, the text to be extracted is extracted, and the TF-IDF method is used for keyword extraction. Finally, the extracted keywords are visualized by word cloud.

This part mainly studies the following problems:

- (1) What are the characteristics of good movies? Or what kind of movies are more popular with people?
- (2) What are the characteristics of films with low ratings? Why are they not liked by people?
- (3) Compared with foreign animation films, what are the shortcomings of Chinese animation films?
- (4) Why does the "Avengers II" score lower than the "Avengers I"?

For the four problems, the text preprocessing part is the same, but the difference is that TF-IDF is used for theme word extraction. This paper uses the TF-IDF modeling tool in Gensim [55] toolkit to model TF-IDF. Gensim is a natural language processing toolkit, which contains most of the modeling methods required for basic natural language processing. Each problem needs to select the preprocessed qualified corpus and documents.

6.1.1. What kind of movies are more popular with people?

This experiment is to explore what kind of movies are more popular on Douban film online. Douban film, as the largest film review website in China, has a very large user base. The score of each film on the website is calculated by the score given by netizens. It can truly reflect the real feelings of the majority of the audience without manual modification.

Corpus in this part of the experiment is all film reviews, while document selects all five-star reviews given by the audience. The experimental results are as follows. For ease of understanding, all the extracted results have been translated from Chinese to English.

what is good movie?



Figure 11 . The word cloud of "What is good movie?"

From the above results, we can generally draw the following conclusions:

- (1) Most of the extracted keywords are praise words for the film, such as "nice" and "like". These words often have strong emotional information and fully express the audience's love for good works.
- (2) There are many keywords describing the characteristics of the film, such as "plot", "law", "moved" and so on. According to these key words, we can see that the films loved by the audience are often films with strong emotions. For example, for comedies, movies are funny enough; For feature films, his plot is very rich and moving. This also shows the importance of good scripts for movies.
- (3) There are also many keywords describing the characteristics of the film, such as "concept", "human nature", "exception", "hope" and so on. These films often have sufficient practical significance, can bring people thinking, or give people positive energy. This also shows the importance of the plot to the film. The plot can not be too illusory and needs to have a certain practical significance.
- (4) There are also many words describing the characteristics of the film, such as "music", "picture" and "special effects". These often describe the effect of film post production. Although these words account for less than the plot, it can also be seen that films with real pictures, beautiful music and sufficient special effects can add to the icing on the cake.
- (5) It can be seen from the words "acting", "actor", "Stephen Chow", "director" and so on. Good actors and good acting skills are also important factors in making films popular.

6.1.2. What kind of film is called "bad film"?

The corpus of this part of the experiment is all film reviews, while the document selects all one-star reviews given by the audience. The experimental results are as follows:

what is bad movie?



Figure 12 . The word cloud of "What is bad movie?"

From the above results, we can generally draw the following conclusions:

- (1) Similarly, the extracted keywords contain many critical words for the film with strong negative emotions, such as "rotten", "nausea", "silly" and so on. There are also words such as "hours", "time" and "waste", which express that the audience thinks watching these films is a complete waste of time.
- (2) From "plot", "story" and "logic", we can see that the lack of plot or logic often causes the audience's disgust. Combined with the conclusion (2) of the previous experiment, it can be seen that the plot is an important influencing factor of film word-of-mouth.
- (3) It can be seen from the words "line", "actor", "director" and so on. Actors with poor acting skills are also an important factor affecting the box office of the film. "Guo Jingming" is a director who has made a lot of junk films with poor reputation before, "Yang Mi" is an actor with poor acting skills. From these two words, we can see that the actor's word-of-mouth and wind review will also have a certain impact on the film's word-of-mouth. Further combined with words such as "money" and "box office", it can be concluded that traffic films, that is, films without substantive content and only trying to make money, tend to get relatively poor evaluation.

6.1.3. What are the shortcomings of Chinese films?

The domestic animation film market has been performing poorly. Although a few works have gained good reputation, they still have a certain gap compared with several excellent film works

6.1.4. “The Avengers I” and “The Avengers II”

Compared with the Avenger I, the score of the Avenger II has decreased. The reasons for the decline can be found through short comments. The experiment is similar to the operation of the previous experiment. Document is selected as all the bad reviews of Avenger alliance 1, while corpus is selected as all the bad reviews of the two films. The experimental results are as follows:

Avengers: Age of Ultron

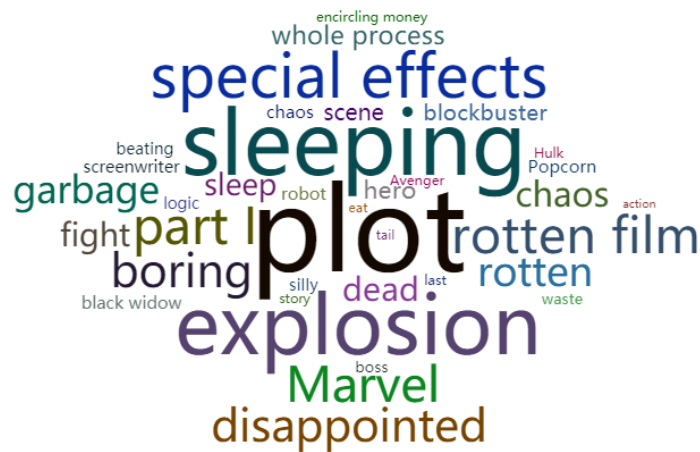


Figure 14 . The word clouds of "Avengers: Age of Ultron"

From the keywords "plot", "chaos", "fight", "part I", "excluded" in the above figure, it can be reflected that the audience thinks that they have high expectations for the second film compared with the first film. The lack of plot and too much play make the audience look very confused, which makes the audience disappointed and give comments on the business trip.

6.2. Sentiment Analysis

6.2.1. Introduction and Process

From the keyword extraction experiment above, we can see that many comments on the Internet will affect the real-time reputation of the film. Obviously, bad comments may have a negative impact on the box office of the film. Because of commercial competition, other box office competitors will hire a large number of sailors to maliciously give a lot of bad reviews on the network platform, so as to crack down on the box office revenue of a film. For Douban and other film comment platforms with user scoring, the platform operator can screen malicious comments according to the user's scoring. Other social platforms without scoring mechanism, such as twitter and Weibo, can only identify malicious comments through the meaning of the text.

This part will build an emotion classification model for the existing tagged text data. In order to

verify the effectiveness of the model, this paper divides the data set into training set, verification set and test set according to 8:1:1, and uses the test set to test the effectiveness of the model. After the model reaches a certain recognition effect, the trained model can be applied to the comment text without labels on other platforms to realize the supervision of film comments.

In this part of the experiment, BERT pre training model is selected for transfer learning. At the same time, in order to prove the effectiveness of BERT model, Transformer, CNN and LSTM are also selected for comparison and reference. In two different experiments, in order to make it easy to distinguish, the two BERT models are named BERT_{all} and BERT_{per} respectively, which represent that the model is trained on all data sets and on a single data set.

6.2.2. Experimental Data

The experimental data included 2125055 film reviews, including up to 28 films. Due to the huge amount of data, the comparative experiment selects the comments of specific films instead of all the comments for the experiment. In the comparative experiment, three films in the data set were selected for the experiment, including "La La Land" with high score (8.4 points), "Tiny Time 3" with low score (4.6 points), and "Transformers 4" with medium score (6.7 points).

Whether all data is used or a small part of data is selected according to specific conditions, it is divided into training set, verification set and test set according to 8:1:1. In this experiment, labels are divided into three categories: positive category (4-5 stars), neutral category (3 stars) and negative category (1-2 stars). The proportion of each category in the data set is shown in the figure below:

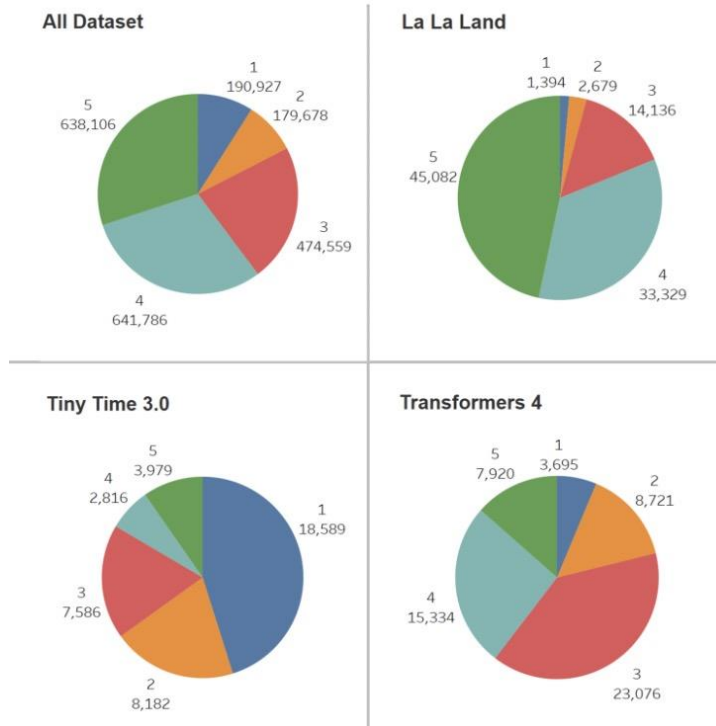


Figure 15 . The proportion of each category in the data set

The experimental data are described in the part 3.2.

It is the same as the theme word extraction part, the experimental data need to be preprocessed, including word segmentation and stop word removal. It is worth noting that this step is not required when using the BERT pre training model. BERT pre training model has its own set of word segmentation thesaurus, which only needs to encode the original text according to the thesaurus.

6.2.3. Training Parameters

The general parameter settings of several models are as follows:

max_length	batch	learning_rate	epochs
144	64	1e-5	100

Table 2 . The general parameter settings of several models

It should be noted that in order to find the best fitting point of the model, the training epoch of all models is set to 100, and then the optimal epoch is determined by observing the loss images of the model on the training set and verification set, so as to determine the score of the model on the test set.

In addition, LSTM, CNN and Transformer models all use the embedded layer as the first layer of the model, and the embedded dimensions are 512. The convolution kernel of one-dimensional CNN model is 5, and the number of convolution kernels is set to 64. BERT adopts Chinese pre training model (to be quoted).

6.2.4. Experimental Result

Firstly, this paper uses the pre trained BERT model to fine tune all the data in the data set. Because of the huge amount of data, in order to avoid the waste of time and computing resources, only the training rounds are set to 10. The training process is shown in the figure below:

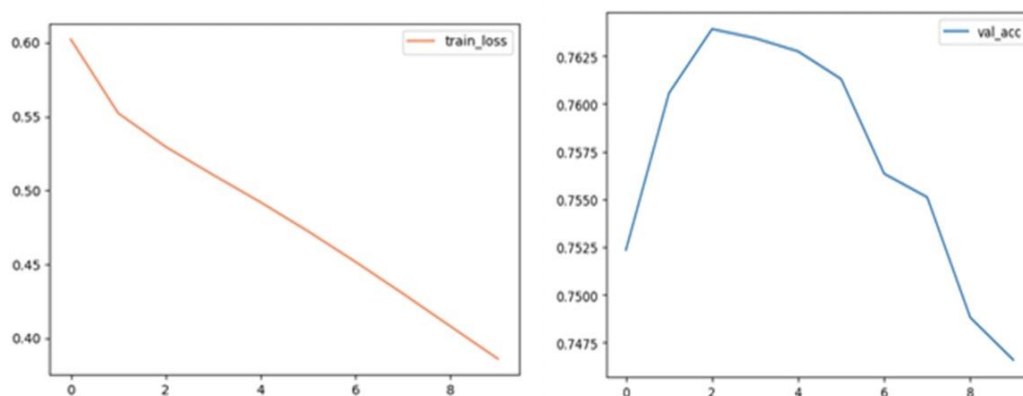


Figure 16 . Change of loss value during training

It can be seen that although the round of the model is only 10, the model soon reached the best fitting state in the second round because of the sufficient amount of data. The optimal experimental results of BERT_{all} model are as follows:

Model	Accuracy_train	Accuracy_validation	Accuracy_test
BERT _{all}	0.8854	0.7639	0.7637

Table 3 . The optimal experimental results of BERT

In order to prove the of BERT pre training language model compared with other pre training language models, LSTM, one-dimensional CNN and transformers coding layers are selected for comparative experiments. The experimental results of the four models on three different films are shown in the table below:

Model	La La Land	Transformers 4	Tiny Time 3.0
BERT _{per}	0.8460	0.7261	0.7410
CNN	0.8177	0.5208	0.6182
LSTM	0.8316	0.6873	0.6747
Transformers	0.8156	0.7056	0.6651

Table 4 . The experimental results of the four models on three different films

As can be seen from the table, BERT_{per} has the best effect in four models and three data sets. Compared with the transformers model, BERT has been pre trained in other general expectations, so it can better express the semantic information in comments. Compared with LSTM and CNN models, it is not only the semantic information of pre training, but also the mask mechanism of BERT itself, which ignores distance, has also improved the effect. In terms of effect, LSTM model is obviously more suitable for text classification task than one-dimensional CNN.

At the same time, the performance of all models on different film data sets is different, and the accuracy fluctuates in a certain range. However, compared with other models, BERT has the most stable effect on the three films, while the accuracy of other models such as CNN fluctuates by nearly 30%. This also shows that the stability of the pre training language model on each data set is better than the model with other embedded layers as input.

6.2.5. Case Study

This section further illustrates the effectiveness of the model by using the trained BERT model to analyze the emotion of several randomly selected comments. For the model, only the trained BERT model is selected in this part, because it has been proved in Table 4 that the BERT model is the best model. In order to prove the effectiveness of the BERT_{per} model for a single film, ten latest reviews were randomly selected from the "Transformers 4" reviews of Douban film for verification. The reason for choosing "Transformers 4" is that it can be seen from Table 4 that the performance of the BERT_{per} model on the film data is the worst. On the other hand, in order to prove the robustness of BERT_{all} model in the emotional analysis of all film reviews, ten reviews of the latest film "Free Guy" are randomly selected for verification. The results of the two experiments are shown in Figure 17 and Figure 18 respectively.

The whole is very good. So many Chinese elements of top American IP look great 😊

Although there are many clothes and advertisements, the four stars all give special effects 😊

Remember to pee and shit before you see it. It's too fucking long. 😊

Why can Michael Bay still be a director? 😊

I think 4 is the best. It's so fucking awesome. It's just special effects 😊 (model give 😊)

I remember super long, Li BingBing, and endless climax, giving enough special effects 😊

I think it's general. There's no logic in some places 😊

Good business shit 😊

Smelly and long 😊

Good guy, it's almost three hours. It's really smelly and long... 😊 (model give 😊)

😊 = Positive comments from netizen
 😊 = Negative comments from netizen
 😊 = Neutral comments from netizen
 123 = The model gives the correct answer
 = The model gives the wrong answer

Figure 17. Case study for Transformers 4

7 points. I haven't laughed so much in the cinema for a long time. 😊

Game version of "Truman's World". The viewing process was very pleasant. 😊

I never dreamed that the captain America gave the shield to the death waiter. 😊 (model give 😊)

The narrative context is very clear, more complete and organized than I expected, and finally a sweet ending. 😊

The funniest romance I've ever seen. The only interesting thing is the third act of the 21st century and Disney's linkage eggs and CE. 😊

Two hours on pins and needles, the mask is my pain mask 😊

Artistic intelligence version of Truman's world, some stems are good, but the whole plot framework is relatively old-fashioned. 😊

In fact, it's quite old-fashioned. Many logical things can't be studied deeply. Don't pull the stem of Truman. It's not a way at all 😊

I feel like I've watched Tik Tok for 114 minutes. It's not even a popcorn movie. 😊

Captain America is great 🍌 😊

😊 = Positive comments from netizen
 😊 = Negative comments from netizen
 😊 = Neutral comments from netizen
 123 = The model gives the correct answer
 = The model gives the wrong answer

Figure 18. Case study for Free Guy

As can be seen from Figure 17, BERT_{per} performed well in the latest ten randomly selected comments, and only two comments were mispredicted. For the fifth comment on misprediction, some words with negative emotions are used to enhance the mood, which may be the reason for the misprediction of the model. In addition, it can be seen from the last comment with wrong prediction that the comment does have obvious negative emotions, but netizens gave a neutral score.

As can be seen from Figure 18, BERT_{all} also has a good effect in predicting the comments of new movies after training on multiple movie datasets, and there is only one comment prediction error. The wrong prediction comment only states some plots in the film “Free Guy”. It is difficult for both models and humans to give a correct prediction of the comment.

Some problems can be further found from these wrong predictions:

- (1) The literal mood of many online comments is inconsistent with the label. On the one hand, some comments have the phenomenon of "irony". For example, there is a comment "*It's really, really, really a good film*". From the literal meaning alone, we can easily characterize the comment as a positive comment, but netizens actually want to satirize through such comments and give the lowest score. On the other hand, some comments, such as comment 5 in Figure 17, are mixed with positive and negative words, and the prediction ability of the model is still insufficient.
- (2) The model is lack of prediction for neutral statements. Compared with the positive and negative sentences with obvious emotional words, the expression of neutral sentences is often vague, which makes the model lack of recognition of this kind of sentences.
- (3) Some comments have no obvious emotional tendency. For example, the third comment with wrong prediction in Figure 18 only describes some plots in the film. It is often difficult for the model to give a correct prediction of this comment.

In short, it can be seen from the results of this part that it is difficult for both models and humans to give accurate judgment on the problematic data. For most other comments, the model has a good performance. This shows that the model has the application value of predicting film reviews in large quantities.

6.2.6. Conclusion

This part uses the BERT model to classify emotions on all data sets, which proves the possibility of applying BERT to the public opinion monitoring module of the platform for preliminary comment screening and filtering. At the same time, in order to prove that the performance of BERT is better than other common models, LSTM, CNN and Transformers are selected for comparative experiments. Finally, in order to further verify the application value of the model and carry out some bad case analysis, case study is carried out in the last section of this part and the results are analyzed. Affective analysis has important applications not only in the classification of film reviews, but also in other fields. Excellent emotion classification model can quickly classify

large-scale text data, which is of great significance.

6.3. Comment semantic similarity

6.3.1. Introduction and Process

In the experiment of sentiment analysis, the sentences are preliminarily divided according to sentiment polarity. However, in practical application, the classification goal is often not as simple as positive or negative. For example, the film review platform found that there was a malicious comment "Director Guo Jingming is short and ugly". There are such malicious comments on the network all the time, which will have a bad impact on the overall network environment. If we only use sentiment classification, we can't directly retrieve the comments with relevant meanings. By calculating the semantic similarity between the target comments and the sentences in the database, we can quickly find the comments related to the target comments, so as to achieve the purpose of public opinion monitoring.

In this part of the experiment, SimCSE is used as the embedded representation model of sentences. SimCSE is a comparative learning model based on pre trained BERT, which greatly improves the effect of BERT on sentence vector representation. The overall process of the experiment is as follows:

- (1) The SimCSE model is used to train on the existing corpus to obtain the fine tuned language model.
- (2) The sentences in the corpus are vectorized, and all vectors are stored in the "database".
- (3) The same model will be vectorized with query comment text.
- (4) The similarity between the vector and the vectors in the "database" is calculated one by one, and the sentences with the highest similarity are selected as the final query results.

6.3.2. Experimental Data

In order to speed up the training speed in step (1) and the retrieval speed in step (4), this part only selects all the review experimental data in the film "Tiny Times 3" (Douban film score 4.6). The reason for choosing the film is that the overall reputation of the film is poor, and the comment area is full of a large number of negative comments, which is in line with the business background of "public opinion monitoring".

6.3.3. Training Parameters

In addition to the SimCSE model mentioned above, Word2Vec and the unmodified BERT model are selected for comparison. The window of Word2Vec model is set to 5, the epoch is set to 8, and the word vector dimension is set to 768. The BERT model adopts 768 dimensional Chinese pre training model "BERT base Chinese".

The model uses the toolkit of hugging face for parameter management. The Chinese pre training language model "BERT base Chinese" in Hugging Face [58] is selected as the pre training language model. All parameters related to BERT adopt the default parameters. The training parameters related to SimCSE tuning are as follows:

Epochs	Batch_size	Learning_rate	Max_length
1	64	3e-5	32

Table 5 . The training parameters related to SimCSE tuning

6.3.4. Experimental Results and Analysis

Because we cannot evaluate the overall effect of the unmarked data set, we can only manually estimate the performance of the model by customizing query and displaying the results. This section takes " Yang Mi is not good at acting" as the query in step (3). Yang Mi is a young actress in China. There are often discussions about her poor acting skills on the Internet. According to the query, the model accountant calculates the similarity of all " Tiny Times 3" film reviews, and sorts the final results according to the similarity. As can be seen from Figure 19, Figure 20 and Figure 21, ten most similar sentences were selected for each model, and their relevance was expressed by color.

It is worth noting that Word2Vec uses cosine similarity as the standard for sentence similarity calculation. The SimCSE model will fine tune the sentence similarity as a comparative learning task, and the final output is a percentage score. The two are only different in presentation form, and the actual evaluation will not have an impact on the result ranking.

Rank Number	Comment	Cosine Similarity
1	Yang Mi's acting skills are worse than Chen Xuedong	0.786
2	The acting skills of these actors are really hard to watch	1.926
3	I really don't want to talk about the actor's acting	2.665
4	Hey, I just think Yang Mi's acting is explosive	2.852
5	Yang Mi's acting, hehe	3.298
6	It's just slandering the acting skills of actors	3.428
7	Yang Mi's acting is really good. Several others feel that his acting is not good...	3.904
8	The actors look good, the plot and acting skills...	3.932
9	Bad acting and plot	4.062
10	Yang Mi's acting is too bad	4.118

RED=Completely Irrelevant/Opposite Meaning

YELLOW=Partial Correlation

GREEN=Completely Correlation

Figure 19. Experimental results of Word2Vec model

Rank Number	Comment	Cosine Similarity
1	Yang Mi's acting is too bad	0.9810
2	Yang Mi's acting skills are better than those of "breakup master"	0.9796
3	Yang Mi's total performance is very fake, none of the three is true	0.9774
4	Yang Mi's acting skills are worse than Chen Xuedong	0.9756
5	Wu Yifan misses you	0.9753
6	Yang Mi garbage	0.9714
7	Guo Jingming is good for bad films. Easter egg is good.	0.9714
8	Can Yang Mi's performance be worse?	0.9708
9	Yang Mi's acting skills are too bad. To be honest, I think the plot of the TV drama version is more in line with the original.	0.9706
10	Really annoy Yang Mi's crying voice and fat face	0.9697

RED=Completely Irrelevant/Opposite Meaning

YELLOW=Partial Correlation

GREEN=Completely Correlation

Figure 20. Experimental results of BERT model

Rank Number	Comment	Semantic Similarity
1	Yang Mi's acting is too bad	0.9051
2	Mi Mi's performance was too bad	0.8807
3	Can Yang Mi's performance be worse?	0.8731
4	I don't like it. Ke Zhendong's performance is OK.	0.8616
5	The plot is bad and the actors are good looking	0.8604
6	Good looking actors	0.8588
7	I don't like Yang Mi's acting	0.8575
8	Yang Mi's performance was terrible, and Guo Caijie was great	0.8527
9	It's really boring. The actors' looks and clothes are good	0.8490
10	Yang Mi's acting skills are worse than Chen Xuedong	0.8466

RED=Completely Irrelevant/Opposite Meaning

YELLOW=Partial Correlation

GREEN=Completely Correlation

Figure 21. Experimental results of SimCSE model

The experimental results of Word2Vec model, BERT model and SimCSE model are shown in Figure 19, Figure 20 and Figure 21 respectively. Based on the above three figures, we can draw the following conclusions:

- (1) In terms of the number of related sentences, SimCSE > BERT > Word2Vec. On the whole, the Word2Vec model captures the point of "acting" in query. Ten comments are words related to

"acting". This is actually determined by the Word2Vec model itself. Because Word2Vec is not pre trained on a large corpus, most of the selected sentences with high similarity are sentences with the same words as query (ten sentences have "acting"), and there is still a certain lack of semantic information capture. From the results of BERT and SimCSE model, it can be seen that the model can well identify "acting" and "performance" as synonyms after pre training. The second comments in the results of SimCSE model identifies "Yang Mi" and "Mi Mi" (Yang Mi's nickname) as the same person, which also proves this.

- (2) In terms of irrelevant sentences, BERT is unstable. Irrelevant sentences are completely irrelevant to query. Unlike irrelevant sentences in Word2Vec, at least "acting" is discussed. From several irrelevant sentences in SimCSE model, it can be seen that several sentences contain the point of "actors look good". For bold speculation, SimCSE model thinks "Yang Mi" equals to "good-looking actor" during training, so it recalls many similar irrelevant sentences. If more epoch training, this problem should be improved.
- (3) Comparing the semantic similarity between BERT and SimCSE, we can find that the semantic similarity of each sentence of BERT is more than 0.96, while the semantic similarity of SimCSE is distributed between 0.84-0.9. This phenomenon verifies the problem mentioned in the literature review, that is, the anisotropy of BERT in semantic space vector. The relevant research on this problem can be seen in the [47][48].

6.3.5. Conclusion

In conclusion, from the above results, we can see that although semantic similarity models such as SimCSE are unstable, the overall effect is good. For the bad comment " Yang Mi is not good at acting ", we can quickly recall the comments expressing the actor's poor acting skills in batch through this method, so as to achieve the purpose of "public opinion monitoring" based on semantic similarity. It should be noted that although Word2Vec is slightly inferior to SimCSE in performance, Word2Vec model is significantly better than SimCSE in training and sentence coding speed. In practical application, we need to comprehensively consider the advantages and disadvantages of both.

Conclusion and Future Work

In this paper, three different forms - film theme word extraction, comment sentiment analysis and semantic based comment recall - are used to mine the Douban film comment data. In the selection of methods, considering different application backgrounds, this paper sets up several groups of mainstream methods for comparative experiments to prove the effectiveness of the method. Aiming at three different problems and application backgrounds, the three methods achieve the goal of text mining to obtain valuable information from a large number of texts.

Firstly, in the theme word extraction part, compared with the previous articles [12][13][15], this method comprehensively analyzes the application background of several topic word extraction methods, and selects the more classic TF-IDF topic word extraction method according to the data selected in this paper. Then, using the visualization method in paper [2] for reference, this paper shows the results of the data by means of word cloud.

In the part of emotion analysis, compared with some classical model methods [34][37][38], the model based on BERT word vector fine-tuning obviously has better performance. Compared with some effective models [39], the highlight of this experiment is that the amount of training data is very large, and the corpus is the real sentences people will see on the Internet.

Finally, in the similarity calculation experiment, this paper uses the latest SimCSE model. In the comparative experiment, it can be seen that SimCSE has better effect than some classical similarity calculation methods in the previous paper [44][45][46].

At the end of this paper, some deficiencies are put forward as points that can be improved in the future.

- (1) Although many models have achieved good results, because of the huge amount of parameters, the model consumes a lot of computing resources and time. If we want to apply it to industry, we need to accelerate or adjust it according to the demand.
- (2) As a classic problem in NLP, sentiment analysis has a variety of new methods every year. Especially after the BERT model, the models based on BERT emerge in endlessly. These models do have better performance in emotion classification than only BERT. If I have time in the future, I am willing to try these models.
- (3) As a kind of online review, film review has its unique novelty and complexity. For example, it is difficult to guarantee that we can understand some comments on the Internet ten years ago. At the same time, if we do not understand, it is difficult to understand the meaning of some symbolic expressions on the Internet. The understanding of these texts has always been a challenging work, and the processing of these texts will be very interesting.

Personal Reflection

In this experiment on film review text mining, I personally gained a lot and learned a lot of new knowledge and methods. First of all, I think the experiment itself is very interesting. It attracts me very much. I can't wait to know some of the results of the experiment to see if they are in line with our usual cognition of the film. Secondly, to my surprise, the model not only draws some conclusions in line with our usual cognition, but also gets many points we usually don't notice (for example, in the subject word extraction experiment, it is found that many people think that the "Monkey King: Hero Is Back" is over marketed). Finally, the correct and rapid processing of large quantities of text by the model makes me feel that the model does have industrial application value.

I also encountered many difficulties in the experiment. For example, at the beginning of the experiment, I didn't know how to start "text mining" for these large quantities of data. Thanks to my mentor, under his guidance, I came up with some valuable application directions of text mining. In addition, when conducting various experiments, I also encountered many unexpected problems because I had no experience before. In order to solve these problems, I consulted many books and blogs. But thanks to these difficulties, I have expanded my knowledge and enhanced my learning ability.

Finally, I think this experiment is a rare and valuable experience for me. At the same time, I also thank my tutor and classmates for their help during the period.

References

- [1] He Yan. Analysis of animated film network reviews based on text mining [D]. Lanzhou University of Finance and economics.2020
- [2] Xin Yuxuan, Wang Xiaodong. Research on Emotional Analysis of film reviews based on text mining [J]. Journal of Mudanjiang Normal University (NATURAL SCIENCE EDITION), 2021 (01): 25-28.
- [3] Qu Xiaoxiao. Research on clustering analysis and visualization methods of film review data [D]. Shandong University, 2018.
- [4] Li Xiang. Breakthrough of domestic animation films from big fish and Begonia [J]. News lovers, 2017 (09): 64-66.
- [5] Luhn H.P. A statistical approach to mechanized encoding and searching of literature information. IBM Journal of research and development, 1957,1, 309-317
- [6] Maron M E, Kuhns J L. on relevance, probabilistic indexing and information retrieval [J]. Journal of the ACM, 1960, 7 (3): 216-244.
- [7] Salton G, Wong a, Yang C S. a vector space model for automatic indexing [J]
- [8] Zhu Qing, Yu Yiming. Analysis and Research on automatic document classification [J]. Modern library and information technology, 1985 (04): 40-46 + 39
- [9] Lewis D. Representation and learning in information retrieval [J]. Ph.D.Thesis Department of Computer Science University of Massachusetts, 1992, 12 (2): 201-210
- [10] Hearst m A. Direction based text interpretation as an information access refining. Text based intelligent systems. 1992:178-180.
- [11] RoBERTson S F, Sparck Jones K. relevance weighting of search terms. Journal of the American Society for information science. 1976:456-458
- [12] Scott Deerwester. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science(1986-1998);sep 1990;41,6; ABI?INFORM Global
- [13] Hofmann, T. Probabilistic latent semantic indexing [J]. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999: 50-57.
- [14] Subasic P., A. Huettner. Effect analysis of text using fuzzy semantic typing. IEEE Transactions on fuzzy systems. 2001:98-100.
- [15] Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation [J]. Journal of machine Learning research, 2003:993-1022.
- [16] Zhang Chenyi, sun Jianling, Ding Yiqun. Microblog topic mining based on MB-LDA model [J]. Computer research and development, 2011,48 (10): 1795-1802.
- [17] Li Jiyun, Huang Yun, Chen Jie. CGRMB-LDA: topic mining for implicit microblog [J]. Computer application, 2016,36 (S1): 67-71.
- [18] Li Peng, Yu Yan, Li Yingle. Improved LDA microblog topic model based on weighted microblog chain [J]. Computer application research, 2016,33 (7): 2018-2021.
- [19] Feng Yong, Qu Bohao, Xu Hongyan, Wang Rongbing. TIF-LDA microblog topic model with variable time window [J]. Small microcomputer system, 2018, 39 (9): 2067-2071
- [20] CAI Gaobadou. Research on text classification method based on FCM clustering [D]. Guangzhou University, 2019

- [21] Zhang Yusha, Luo Lixia. Construction of microblog topic model for big V and grass-roots users based on LDA [J]. Modern computer (Professional Edition), 2019, (2): 3-6,11.
- [22] Kolchyna O, TTP Souza, P Treleven, T Aste, Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination.2005: 5-7.
- [23] Ghose A , Ipeirotis P G . Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews[C]// Proceedings of the Ninth International Conference on Electronic Commerce. ACM, 2007.
- [24] Bing Pan, Ph.D.et.al. Travel Blogs And The Implications For Destination Marketing. Journal of Travel Research.2007: 16-18.
- [25] Ye Qiang, Zhang Ziqiong, Luo Zhenxiong. Research on Chinese subjective automatic discrimination method for Internet comment emotion analysis [J]. Journal of information systems, 2007 (01): 79-91
- [26] Zhou Lizhu, He Yukai, Wang Jianyong. Review of emotion analysis [J]. Computer applications, 2008 (11): 2726-2728.
- [27] Sun Bingyu, Li Jiuyong, Wu Desheng, Zhang Xiaoming, Li Wenbo. Kernel discriminant learning for ordinal regression. IEEE Transactions on Knowledge and Data Engineering. 2010:88-90.
- [28] Yu Yang. Research on emotion analysis and topic extraction of Douban film review text [D]. Yunnan University of Finance and economics, 2018
- [29] Bing Pan and Daniel R. Fesenmaier, Semantics of Online Tourism and Travel Information Search on the Internet: A Preliminary Study. Information & Communication Technologies in Tourism.- ota.cofc.edu. 2002: 8-10.
- [30] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up. Sentiment Classification Using Machine Learning Techniques. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. 2002:120-123.
- [31] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the Association of Computational Linguistics. 2002:348-351.
- [32] Su Jinshu, Zhang Bofeng, Xu Xin. Research progress of text classification technology based on machine learning [J]. Journal of software, 2006 (09): 1848-1859.
- [33] W Medhat, A Hassan, H Korashy, Sentiment analysis algorithms and applications: Ain Shams Engineering Journal.2014, 5(4):1093-1113
- [34] Zhu Shaojie. Research on text emotion classification based on deep learning [D]. Harbin Institute of technology, 2014.
- [35] Liu Shuang, Zhao Jingxiu, Yang Hongya, Xu Guanhua. Overview of text emotion analysis [J]. Software guide, 2018,17 (06): 1-4 + 21
- [36] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.
- [37] Wang Ding, Du Hong. Research on emotion classification of film review based on deep neural network [J]. Computer and information technology, 2019 (4): 18-20
- [38] Gu Qian. Research on text classification method based on feature vector [D]. Xi'an University of technology, 2019.
- [39] Liu Yong, Wang Zhen. Emotion prediction of target characters based on deep learning [J]. Computer system application, 29 (6): 7
- [40] Lv Xueqiang, Ren Feiliang, Huang Zhidan, etc. sentence similarity model and the most

- similar sentence search algorithm [J]. Journal of Northeast University, 2003 (06): 531-534
- [41] Wang Rongbo, Chi zheru, Chang Baobao, etc. Chinese sentence similarity measurement based on word string granularity and weight [J]. Computer Engineering, 2005 (13): 142-144.
- [42] Li Shengwen, Ling Wei, Gong Junfang, et al. An entropy based text similarity calculation method [J]. Computer application research, 2016,33 (03): 665-668
- [43] Lan Yanling, Chen Jianchao. Sentence structure similarity calculation based on part of speech and part of speech dependence [J]. Computer Engineering, 2011,37 (10): 47-49
- [44] Wang Qihe. Cluster analysis of agricultural information based on Wikipedia semantic model [J]. Information science, 2013,31 (12): 101-104
- [45] Christos Bouras, Vassilis Tsogkas. A clustering technique for news articles using WordNet[J]. Knowledge-Based Systems,2012,36.
- [46] Zhu Zengyu. A Text Hybrid Clustering Algorithm Based on HowNet Semantics [A]. Intelligent Information Technology Application Association. Proceedings of 2011 International Conference on Advanced Materials and Computer Science (ICAMCS 2011 Part3) [C]. Intelligent Information Technology Application Association:
- [47] Gao J, He D, Tan X, et al. Representation Degeneration Problem in Training Natural Language Generation Models. 2019.
- [48] Li B, Zhou H, He J, et al. On the Sentence Embeddings from Pre-trained Language Models[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.
- [49] Gao T, Yao x, Chen D. Simple contractual learning of sense embeddings [J]. 2021.
- [50] Agichtein E. Snowball ; Extracting relation from large plain-text collections[C]// Acm International Conference on Digital Libraries. 2000.
- [51] Wen Junwei. Research on Twitter oriented analysis system [a]. Proceedings of the 31st national computer security academic exchange conference of China Computer Society [C]. China Computer Society: Computer Security Professional Committee of China computer society, 2016:6
- [52] <https://wenku.baidu.com/view/98c46383e53a580216fcfed9.html>
- [53] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
- [54] <https://zhuanlan.zhihu.com/p/26306795/>
- [55] <https://radimrehurek.com/gensim/>
- [56] <https://www.kaggle.com/utmhikari/doubanmovieshortcomments>
- [57] <https://pypi.org/project/jieba/>
- [58] <https://huggingface.co/>