Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/38/20230551

Research on sentiment analysis based on the Bilibili video barrage

Junyu Bai

Computer Science Department, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei Province, 0335-066000, China

202112165@stu.neuq.edu.cn

Abstract. Based on the analysis of the emotions and mentality of site B users watching videos, this paper proposes a method to visualize important attributes such as high-frequency words, the proportion of positive and negative comments, and word cloud diagrams. In the context of the rise of the Internet and the increasing application of Web 2.0, this paper took the Bilibili barrage text as the research object. Because of the large amount of barrage data, only individual barrages were selected as the analysis objects according to the requirements. The crawler was used to preprocess the crawled video barrage. Then machine translation knowledge and four algorithms such as the word segmentation algorithm and sentiment analysis algorithm were used to analyze the sentiment of the video barrage from three different dimensions and compare the results. Through the analysis of the visualization results, the differences in the emotional distribution of different video barrages were compared, and two important conclusions were drawn: First, the mentality of Bilibili users watching videos is positive; second, there is a certain correlation between the content of the video and the emotional orientation of the barrage, and mutual prediction can be made between the two. However, the research in this paper is only the tip of the iceberg in the research of public opinion analysis. At present, the application of sentiment analysis in public opinion still faces difficulties. How to optimize the algorithm model according to the current situation requires researchers to conduct deeper research and more extensive thinking.

Keywords: Sentiment Analysis, Machine Learning, Algorithms, Comparison.

1. Introduction

The efficiency of Chinese information processing mainly depends on the key technology of Chinese word segmentation. This paper comprehensively explained the current various Chinese algorithms and analyzed them comparatively [1]. Weibo's massive data has gradually become a research hotspot in many fields in recent years. This paper introduced Weibo's sentiment analysis algorithm from three research directions: emotional dictionary, machine learning, and feature fusion [2]. This paper combined mechanical word segmentation and statistics-based word segmentation methods. Researchers proposed a statistical Chinese word segmentation algorithm that can detect uprising fields and resolve ambiguity in text [3]. In this paper, a new Chinese text word-cutting method combining the N-gram model and an effective Viterbi search algorithm was proposed. A variety of Chinese word segmentation methods were tested from the two quantitative indicators of accuracy rate and recall rate, proving that

DOI: 10.54254/2755-2721/38/20230551

this experiment's algorithm is more efficient [4]. Taking data analysis in a big data environment as the background, this paper examined and compared the implementation effects of five sentiment analysis algorithms under different data scales from the aspects of accuracy and scalability [5]. This paper introduced a technical means to monitor negative energy public opinion on the Internet and filter and analyze sensitive words on data web pages. Researchers provided a correct orientation and reliable method for Internet public opinion control [6]. Based on the background of finding and restoring the characteristic word variation of Chinese spam messages, this paper improved a new Bayesian mail filtering algorithm with better performance based on the ordinary Bayesian algorithm [7]. This paper proposed a statistical machine translation method based on the SMT model that used chapter context information to improve the accuracy of rule selection. This method effectively improved the translation quality of English to Chinese [8]. In this paper, three mainstream crawling techniques based on Python language were analyzed for efficiency and accuracy. The results showed that to achieve the most ideal crawling results, crawler technology should be reasonably selected according to business needs and technical characteristics [9]. This paper discussed several visualization algorithms for image recognition, which have been widely used and had good results. Accordingly, researchers provided reference examples for wind power operation and maintenance monitoring [10]. This paper mainly reviewed the application of sentiment analysis methods in tourism and looks forward to the future development direction of sentiment analysis methods. In this way, more insights can be gained to promote further innovation in tourism research in theory [11].

Sentiment Analysis is a common application of natural language processing (NLP) methods, which is the process of analysis, processing, induction, and reasoning of subjective text with emotional color. In recent years, with the development of the Internet, especially WEB 2.0 applications, netizens have become more convenient to comment on various products and hot events [1]. The barrage text in videos on the platform of station B (Bilibili) can intuitively, quickly, and effectively reflect the user's emotions and mentality when watching the video. So the barrage sentiment analysis was selected as the project content. However, due to the large amount of data collected from all videos of Bilibili, the analysis is difficult. Therefore only individual video barrages were selected as analysis objects according to requirements.

2. Method

2.1. Collection method of sample dataset

Multi-dimensional sampling was used for video sampling, which was mainly divided into three dimensions:

(1) Several videos were randomly selected and combined into a sample set, and then the average result was obtained.

(2) Different types of videos were selected. The test results were obtained and recorded separately. Then they were compared and analyzed.

(3) Videos of the same type but with different emotional tones were selected. The test results were obtained and recorded separately. Then they are compared and analyzed.

2.2. Text Acquisition

Crawler is used to obtain barrage text content. The web crawler is a program or script that automatically crawls World Wide Web information according to rules. It usually crawls the initial URL and continues to find links from the page to establish URL queues. Then it crawls information on the web until it meets users' needs. Different web crawler technologies have different advantages and disadvantages and are suitable for different business scenarios, so they should be reasonably selected [2].

DOI: 10.54254/2755-2721/38/20230551

2.3. Relevant algorithms of the code module

2.3.1. Statistical and rule-based machine translation methods

Broadly speaking, statistical machine translation can also be seen as using rules to translate, and the general translation rules include the source side and the target side [3]. Technologies and algorithms related to machine translation are used to optimize and simplify the processing and analysis process of bullet screens, improve the accuracy and efficiency of text analysis, and further automate text analysis tasks.

2.3.2. Word segmentation algorithm

According to the characteristics of Chinese word segmentation, the existing word segmentation algorithms can be divided into four categories: word segmentation method based on string matching, word segmentation method based on understanding, word segmentation method based on statistics, and word segmentation method based on semantics [4]. The barrage text is converted into a dataset composed of data structures of word units through Jieba Chinese segmentation, which makes facilitating subsequent analysis and processing steps more convenient. For example, "tai(very) li hai(good) le(a certain modal particle)" is cut into "tai \ li hai \ le". In short, Chinese word segmentation is to divide Chinese sequences into meaningful word sequences for computer understanding [5].

2.3.3. Sentiment analysis algorithm

Sentiment analysis is a process that includes steps such as data retrieval, data processing, and feature extraction. A variety of different analysis methods can be derived from this, namely supervised learning, dictionary-based oriented methods, semantic methods, etc. Text sentiment analysis refers to the process of using natural language processing and text mining technology to analyze, process, and extract subjective text with emotional color. Emotional polarity analysis can be methodologically divided into binary, ternary, etc. This paper adopts binary classification so that the polarity of comments is defined as positive and negative under the premise that the user barrage is subjective. In this experiment, the SnowNLP library was used to determine the emotions of the barrage text, and the judgment results were classified into three categories: "positive emotion", "negative emotion" and "neutral emotion", with values of 0.5-1.0, 0-0.5 and 0.5, respectively.

2.3.4. Abnormal word filtering algorithm

In the barrage text, anomalous words may appear in any barrage. Therefore, after roughly browsing the content of the barrage sample, the abnormal words should be summarized as much as possible. Artificially customizing the "stop words" list to filter meaningless or abnormal words, such as punctuation marks, stop words, etc., to normalize the barrage text is convenient to facilitate subsequent processing. At present, the most commonly used word filtering algorithm is to use a tree structure to store abnormal words or sensitive words. However, considering that a large amount of space will be occupied by the storage tree, which will take a long time to be built, and also considering that the sample size taken in this study is small, so the tree structure is not used.

2.3.5. Visualization algorithm

The application of visualization can effectively enhance visual sense, and the direct perception of non-visual information in the form of numbers, text, etc. by human vision lags far behind the understanding of visual symbols [6]. The visualization algorithm used in this article is an algorithm that presents the test results of various algorithms by visualizing the computing environment. In this experiment, Matplotlib was used to draw pie charts to show the proportion of emotional polarity of the barrage, and the WordCloud library was used to plot word clouds to show keywords that appear more frequently in the barrage.

DOI: 10.54254/2755-2721/38/20230551

3. Results

3.1. Sentiment distribution ratio chart

From the first dimension, multiple obvious videos were randomly selected and integrated into a sample set, and the average result was obtained. It can be seen that in the videos with different emotional tones, the barrage emotions of Bilibili users are generally positive.

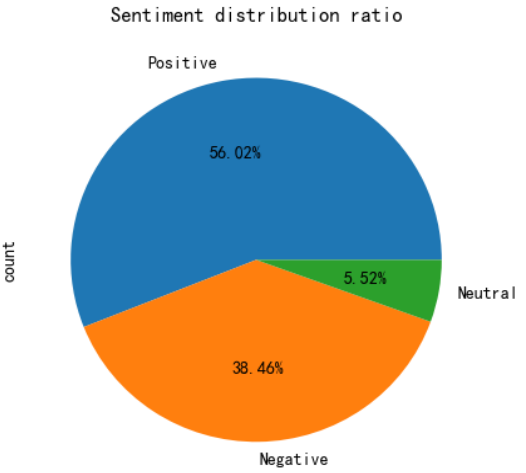

Figure 1. Pie chart 1.

From the second dimension, different types of videos were selected. The test results were obtained and recorded separately, and they were compared and analyzed. Here is the visualization result of beautiful song MV videos.

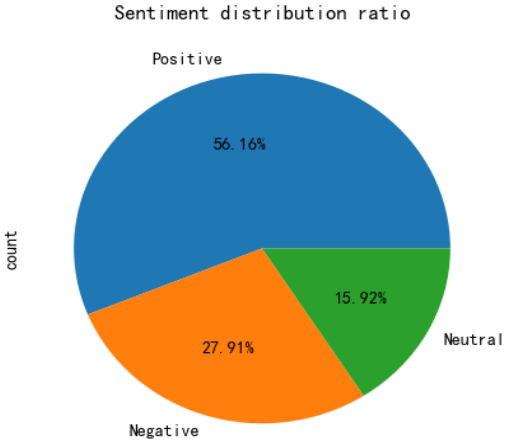

Figure 2. Pie chart 2.

First of all, in the pie chart, it can be seen that positive emotional barrage accounts for the majority, about 56%, which is in line with the emotional tone of "beautiful song video". Most of the positive emotional barrages contain "hahaha" phrases. Meanwhile, the neutral emotional barrages are mostly phrases composed of English letters or meaningless combinations of numbers. The above results are basically in line with the facts. Here is the visualization result of a popular science video with relatively serious emotions.

DOI: 10.54254/2755-2721/38/20230551

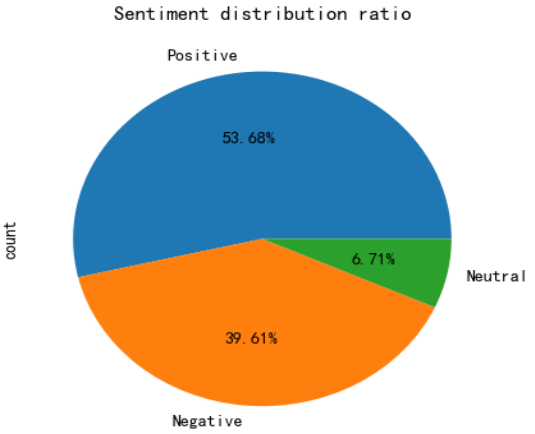

Figure 3. Pie chart 3.

It is easy to know that for videos with relatively serious emotions, the user's barrage accounts for about half of the positive emotions and negative emotions. Combined with people's realistic cognition, it can be judged that the video content should be relatively rational. For example, barrages containing phrases such as "congratulations" and "wonderful ah" are classified as positive, while negative barrages mostly have emotions like sarcasm. The above results are basically in line with the facts.

From the third dimension, videos of the same type but with different emotional tones were selected. The test results were obtained and recorded separately, and they were compared and analyzed. Here is the visualization result of hilarious news with positive emotions.

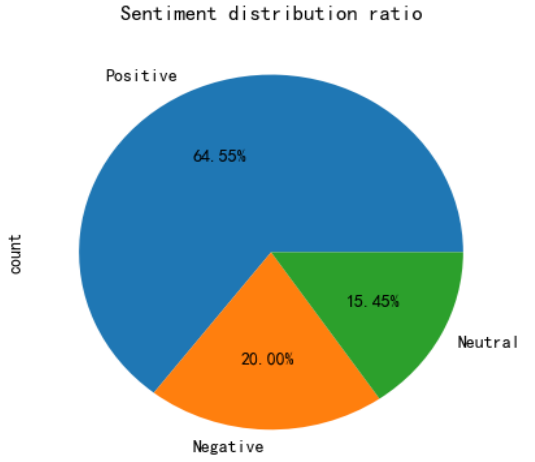

Figure 4. Pie chart 4.

Obviously, from the pie chart, about 64% of the barrage emotions are positive, in line with the emotional tone of "hilarious videos", and most positive emotions contain typical positive emotional phrases such as "hahaha" and "good", which is in line with the facts. Here is the visualization result of negative news with negative emotions.

DOI: 10.54254/2755-2721/38/20230551

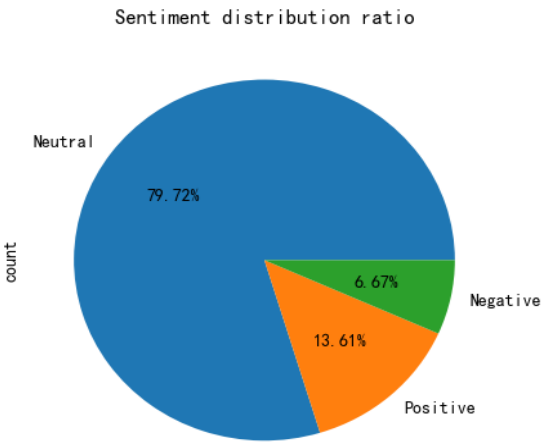

Figure 5. Pie chart 5.

As can be seen from the pie chart, positive emotional barrages account for only about 6.7%, which is in line with the emotional tone of "negative news". Relatively, the large number of barrages containing negative emotional phrases such as "horror", "silent mourning" and "uncomfortable" are classified as neutral and negative, which is generally in line with the facts.

3.2. Summary

Based on the analysis of the emotions and mentality of Bilibili users watching videos, this experiment proposes a method to visualize important attributes such as high-frequency words, the proportion of positive and negative comments, and word cloud diagrams. Specifically, the crawled video barrages were first preprocessed in this experiment. Then the sentiment of the video barrage is analyzed from three different dimensions. So the results were compared. Through the analysis of the visualization results, **Figure 1.** Pie chart 1, **Figure 2**. Pie chart 2, **Figure 3.** Pie chart 3, **Figure 4.** Pie chart 4, **Figure 5.** Pie chart 5, the differences in the emotional distribution of different video barrages were compared, and two conclusions were drawn: First, the general emotional orientation of Bilibili users when watching videos is generally positive; second, the results show that there is a certain correlation between the content of the video and the emotional orientation of the barrage, and a mutual prediction can be made between the two. For example, when a popular science video is opened, its positive barrage and neutral and negative barrage can be predicted to account for half each; when a video's barrage list is obtained, the sum of its negative barrage and neutral barrage ratio exceeds 50%, then the emotional tone of the video can be predicted as a negative emotion.

4. Discussion

In terms of text processing, considering that the content of the barrage text is complex, casual, and colloquial, these barrage texts need to be preprocessed. Stop words, such as the particle "of", and the pronouns "she", "he", etc. all need to be removed. After the preprocessing is completed, the barrage text becomes a string of words concatenated. For instance, "Chinese athlete Wu Dajing won the short track speed skating men's 500m final." It will become {China, athlete, Wu, Dajing, in, short track speed skating, men, 500, meters, final, middle, win the championship}. This algorithm improves accuracy. However, there is still interference. The self-assembling Chinese word segmentation method based on the N-gram model can be considered to first automatically generate a word segmentation dictionary through machine learning of the raw corpus. It can greatly reduce the interference caused by word mutual interference and high-frequency sharpening [7].

On large-scale datasets, statistical learning is usually used directly to solve sentiment analysis problems. It focuses on the machine learning algorithm and its parallelized implementation [8]. The method of machine learning is simple, which saves a lot of manpower. The database can be used to

DOI: 10.54254/2755-2721/38/20230551

update the vocabulary in time. However, machine learning requires manual annotation of sequences, resulting in insufficient context utilization and insufficient accuracy. Deep learning can make full use of the context information to retain the order of sentences, to achieve sentiment classification and word polysemy. Multi-layer neural networks can be used to extract data features and learn better. In this experiment, the machine learning method was used to train the model, which led to an inaccurate judgment of emotional polarity.

In terms of the filtering algorithm, on the one hand, the Bayesian algorithm can be used to improve the research. Bayesian algorithm is well applied in spam filtering techniques. It can automatically learn from spam, automatically adapt to user habits, automatically filter out spam, and has high accuracy [9]. On the other hand, sensitive words should be taken into consideration. So the dictionary tree can be used to filter the bad information in the text from the total number of sensitive words, and also upgrade the category of bad information [10].

5. Conclusion

This paper taking the barrage text of Bilibili as the research object, showed the application research of the algorithm based on sentiment analysis in public opinion analysis. Then two important conclusions were drawn. However, the research in this paper is only the tip of the iceberg in the research of public opinion analysis. Public opinion will continue to change with the development of the times, especially in a subcultural gathering place loved by young people such as Bilibili. As we all know, the barrage and comment text of its users are changing with each passing day. Therefore, in the future of the increasing popularity of big data technology, the application of artificial intelligence, particularly natural language processing, must be gradually improved. And the corpus that meets the characteristics of different video platforms should be updated constantly. Also, a sentiment analysis model that is more in line with the progress of the times should be designed. The above three aspects can help video production teams and related marketers monitor users' reactions and opinions on video content. So that the test results can provide a reference for merchants' marketing promotion. The sentiment analysis made in this article is mainly an analysis of the user's subjective comments, which usually reflect the user's judgment and emotions about an entity or event. Tapping into these large numbers of subjective e-word of mouth is of great value to tourism organizations seeking to improve customer management and business profitability [11]. At present, the application of sentiment analysis in public opinion still faces many difficulties. How to optimize the algorithm model according to the current situation requires researchers to conduct deeper research and more extensive thinking.

References

[1] Yang Wenting. Research and implementation of sentiment analysis algorithm based on microblog [D]. Southwest Jiaotong University, 2015.

[2] YANG Jian, CHEN Wei. Research on Three Web Crawler Technologies Based on Python [J]. Software Engineering, 2023, Vol. 26(2): 24-27,19.

[3] Yu Hui, Xie Jun, Xiong Hao, et al. Statistical machine translation method based on chapter context [J]. Journal of Chinese Information Technology, 2013, 27(2): 86-91.

[4] Zhang Qiyu, Zhu Ling and Zhang Yaping. Review of Chinese word segmentation algorithm [J]. Information Exploration, 2008(11): 4.

[5] Lin Dongsheng. Research and implementation of Chinese word segmentation algorithm [D]. Northwest University, 2011.

[6] Cai Jinzhu, Lai Youfu, Han Lulu, et al. Research on image recognition algorithm of wind turbine based on visualization [J]. Modern Industrial Economics and Informatization, 2023, Vol. 13(5): 277-280.

[7] Wu Yingliang, Wei Gang, Li Haizhou. A Chinese word segmentation algorithm based on the n-gram model and machine learning [J]. Journal of electronics & information technology, 2001, 23(11): 6.

DOI: 10.54254/2755-2721/38/20230551

[8] Yu Chuanming, Yuan Sai, Wang Feng, et al. Research on the scale adaptation of text sentiment analysis algorithm in big data environment: using Twitter as data source [J]. Library and Information Service, 2019, 63(4): 11.
[9] Wang Xia, Zheng Ning, Xu Ming, et al. Bayesian mail filtering model based on Chinese inflection word matching [J]. Computer Applications and Software, 2010(1): 4.
[10] Wei Xinyu, Li Binyong, Chen Hongdou et al. Filter and Analysis of Sensitive Words on Web Page Based on Public Opinion Data [J]. Cybersecurity Technology and Application, 2022, (7): 38-39.
[11] Liu Tengjie. Application of sentiment analysis in tourism research: review and prospect [J]. Tourism Overview, 2023, (1): 48-52.