Analysis of User Behavior in the New Media Era

Rongzeng Hou*
Shandong Institute of Commerce and Technology, Jinan, China

Keywords: Data Analysis, Emotion Analysis, Text Classification, Bayesian Algorithm.

Abstract: With the rapid development of artificial intelligence in recent years, especially the rapid emergence of large natural language processing models represented by ChatGPT since March, people are increasingly aware of the importance of data to human society. Conclusions derived from data analysis have become an indispensable source of decision-making in academia and business. Bilibili, affectionately known as "Station B" by fans, is a leading youth culture community in China. The B station features a live comment function suspended above the video, which enthusiasts call "barrage". This paper uses Bayesian algorithm to analyze the sentiment of comments on a video on Bilibili website. Firstly, we collected a large amount of comment data and preprocessed it, including word segmentation and removal of stop words. We then used the naive Bayes algorithm to categorize each comment by emotion, including positive, negative, and neutral. Finally, we evaluated the classification results and came up with our sentiment analysis results.

1 INTRODUCTION

With the development of digital media and other technologies, bullet screen system, a new type of comment mode, appears and becomes popular gradually. It allows video viewers to post comments on the plot of the video in real time, and also helps viewers understand the content of the video. The generation of bullet screen text data provides new material for short text processing and real-time data processing. The study of the characteristics of bullet screen data and its expression of emotion can help us better understand the plot of video; By studying the similarity between bullet screen contents and analyzing the relationship between users, we can not only deeply understand the characteristics of bullet screen users and explore the potential relationship between different videos, but also provide more accurate solutions for the selection of audience groups in video production (Bourouis, S., 2021). At present, the two most famous video websites in China are AcFun and Bilibili, affectionately known as Station A and Station B. This paper uses published comment sentiment analysis data sets to train the model, and then conducts sentiment analysis around the comments of a popular video on Bilibili.

Through the emotional analysis of the video content and the viewers' real-time viewing experience, the emotional convergence and differences between the two can be found, and the overall attitude of the viewers can be clearly seen, thus providing a good statistical result for the evaluation, production and public opinion of the video. The design crawls comment data from Bilibili website, what kind of opinion the text data is, what kind of attitude people hold towards the current situation of young people -- positive, negative or neutral. The sentiment analysis of microblog comment data is carried out by establishing Bayesian classification model. In order to improve the accuracy of word emotion discrimination, this design uses data visualization based on "word cloud" for judgment.

Nowadays, as one of the two famous bullet screen website platforms in China, B Station has a large increase in video comments and a variety of comment content, which makes it difficult to achieve information acquisition. So it's important to collect and categorize these comments, especially sentiment analysis. The information of station b is widely disseminated among users. The information of Station b contains the subjective emotions of each user and has the characteristics of describing human subjective preferences, appreciation, dissatisfaction and other emotions. Presenting information in a visual way can help users to have a deeper understanding of the characteristics of b station, enable users to have an insight into the seemingly fragmented but actually mysterious data relations

Hou, R.
Analysis of User Behavior in the New Media Era.
DOI: 10.5220/0012286700003807

In *Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2023)*, pages 509-514
ISBN: 978-989-758-677-4

and their rules, and discover valuable emotional trends and communication trends, which has a very positive significance for public opinion guidance and news diffusion.

The essence of emotion analysis is a process of text classification, which is to analyze and excavate texts with certain emotional colors to find out the relevant emotional tendencies (Liu, K., 2019). They can be divided into three types: positive, negative and neutral. The design uses machine learning algorithm, machine learning is a branch of artificial intelligence in recent years more hot artificial intelligence, its main application for classification tasks, naive Bayes, support vector machine (SVM), maximum entropy and other algorithms in recent years continuous development: Some scholars improved the naive Bayes algorithm to improve the classification accuracy in view of the fact that the calculation of prior probability in text classification is relatively time-consuming and has little influence on the classification effect and the accuracy of classification is affected by the accuracy loss of posterior probability (Zhu, X., 2020). In the other research, the authors proposed a Dirichlet naive Bayes Swinburne classification algorithm based on Map Reduce, which significantly improves the accuracy and recall rate of traditional naive Bayes Swinburne classification algorithm and has excellent scalability and data processing ability (Rogers, D., 2022). Some scholars proposed a naive Bayes Swinburne classification algorithm with attribute weighted complement, and conducted comparative experiments with traditional naive Bayes and complementary naive Bayes algorithm. The results showed that the improved algorithm had the best performance when the distribution of sample sets was not balanced, and the classification accuracy, recall rate and G-mean performance were greatly improved (Abdalla, H. I., 2022). In the other study a new classification model based on naive Bayes, which can reduce the redundant attributes in the data set, calculate the weight of each reduced conditional attribute relative to the decision attribute, and integrate the weight into the naive Bayes classification model to improve the application scenario and classification accuracy of the naive Bayes classification model (Villa-Blanco, C., 2023).

Foreign scholars began to study text classification in the 1960s. In 1961, Maron published his first paper on automatic classification. In 1975, Salton built a vector space model based on information search, artificial intelligence and machine learning, which made text automatic classification obtain certain application results in different technical fields. H.P. Luhn proposed a classification based on word frequency statistics. The first paper on classification algorithm was published by Maron et al. after continuing the research and sorting of text classification based on this field. Later, scholars such as G. Stalton, K.Park and K.S. Ones also obtained many achievements in this field through the study of text classification. Under the extensive research of foreign scholars, text classification has been put into practice and widely used in the field of information resource organization and management. Sharma and Dey proposed the SVM mixed model based on Boosting, which improved the performance excellence of the SVM model (Han, M.- Gao, H.). The researchers have proposed a suicidal emotion prediction algorithm for social networks based on machine learning and semantic sentiment analysis in the journal Procedia Computer Science, and a WordNet-based algorithm for semantic analysis between tweets in the training set and tweets in the data set (He, J.- Hao, S. L.). The authors used machine learning methods for text classification in the International Conference on Bioinformatics and Computational Biology, In order to determine the contextual polarity of each call on the subject of the malaria bid, our data were used to harvest people's perceptions of malaria and understand the impact of research and recent development assistance on malaria aid on the subject of malaria (Cardenas, J. P., 2014). They collected, mined and analyzed college-related tweets through sentiment analysis based on machine learning algorithm (Li, L. F., 2019).

2 METHODS

2.1 Natural Language Processing Technology

Natural Language Processing (NLP) is a technology involving computer science, artificial intelligence, linguistics and other disciplines. It mainly involves taking information from human language and putting it into a form that a computer can process. Here are some examples of how natural language processing works:

Speech recognition: Speech recognition is the technology that converts audio signals of human speech into text form. The technology is usually implemented using acoustic models and language models, and can be applied to voice assistants, automatic translation and other aspects.

Text categorization: Text categorization is the technique of categorizing text data into predefined categories, which can be achieved by using machine learning algorithms. The technology is commonly used in spam sorting and sentiment analysis.

Named entity recognition: Named entity recognition is a technology that identifies entities in text and labels them as personal names, place names, organization names, etc. This technology can be applied to natural language question answering, information extraction and so on. Natural Language generation: Natural language generation is the technology of converting computer-generated information into natural language. This technology can be applied to machine translation, natural language dialogue system and so on.

Machine translation: Machine translation is the technique of translating one natural language into another. This technology is usually implemented using neural network model, which can be applied to cross-language communication, document translation and other aspects.In short, natural language processing technology is widely used in a variety of fields, and with the development of machine learning, deep learning and other technologies, its application will continue to expand.

2.2 Machine Learning Algorithm: Bayesian Algorithm

As a machine learning method with a long history and solid theoretical basis, Bayesian method can not only deal with many problems directly and efficiently, but also evolve many advanced natural language processing models from it. Bayes method is an excellent way to study natural language processing.

Preparatory work stage: This stage mainly preprocesses the text, first marks the samples, and then screens the feature words according to word frequency. At this stage, all samples to be classified are input, and then the characteristic attributes and training samples are obtained. The accuracy of naive Bayes classifier is mainly determined by the selected feature attributes.

Classifier training stage: According to the frequency in the sample, then calculate the prior probability of each category by each feature. This stage is mainly based on the formula of mechanical calculation. This stage is the most important part of naive Bayes classification.

Application stage: In this stage, the test samples are mainly input, and then the classification demerit is calculated by the classifier.

2.3 Data Acquisition

In order to accurately capture the most authentic content of emotional tendencies, Using comment sentiment analysis data set (https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/ simplifyweibo_4_moods/intro.ipynb). Then 100,000 pieces of data related to text analysis were selected, including 40,000 positive, 40,000 negative and 20,000 neutral, and the model was trained with the selected data. Then I found a video with a large number of comments and meaningful analysis from Bilibili website, and the comment content should have a certain emotional tendency. A total of 30,000 comment data were obtained under this video, and the garbled and dirty data and invalid data without emotional orientation were removed, and finally 21,760 effective information was obtained. Finally, this data is classified by sentiment analysis to get our emotional statistical results.

Data Preprocessing: After obtaining the data used in the experiment, it may not be easy to process the comments because the format of the data does not agree, so the format and form of the data should be unified first. The specific steps are as follows:

1) The effects of text de-duplication include improving the efficiency of text processing, reducing storage space, avoiding information redundancy and improving the accuracy of text analysis. If a large number of duplicate texts exist in the text set, a lot of time and computing resources will be wasted, storage space will be occupied, information redundancy will be caused, and the accuracy of text analysis will be affected. By deweighting, we can ensure the uniqueness of each text in the text set and avoid these problems. Figure 1 shows the python code and its execution.

```
#!/usr/bin/python
# -*- coding: UTF-8 -*-
readPath='../source.txt'
writePath='qvchong.txt'
lines_seen=set()
outfiile=open(writePath,'a+',encoding='utf-8')
f=open(readPath,'r',encoding='utf-8')
for line in f:
if line not in lines_seen:
outfiile.write(line)
lines_seen.add(line)
```

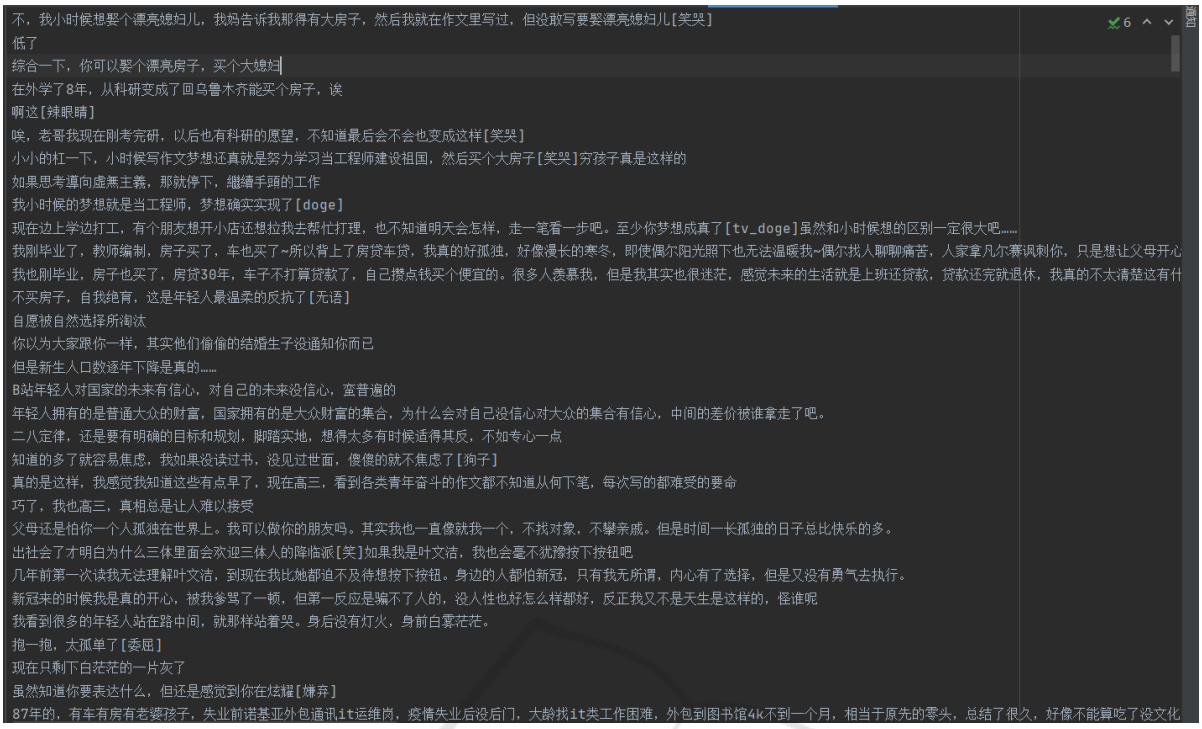

Figure 1: Dereprocessing text.

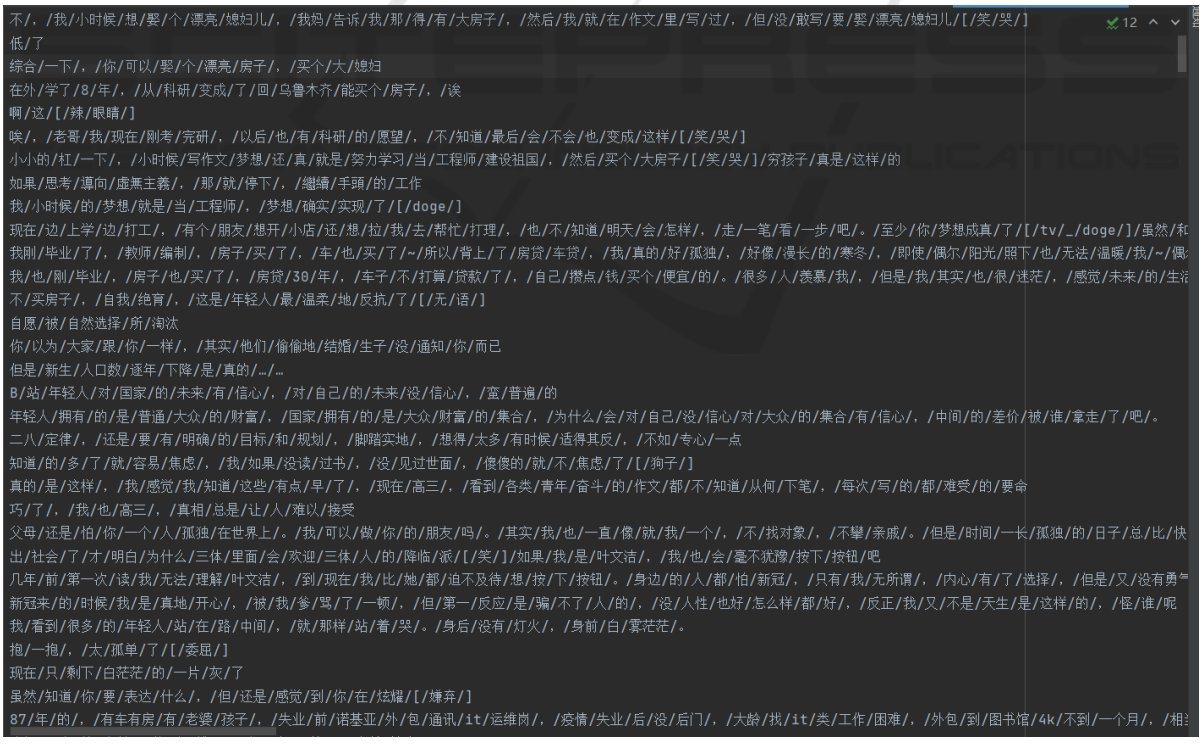

Figure 2: Word segmentation result.

2) Word segmentation must be performed for vectorization of the text after it has been de-duplicated. Word segmentation is the segmentation of text into meaningful words according to rules and algorithms for the convenience of text processing and analysis. It can improve text structure, improve

processing efficiency, and optimize classification and information retrieval effect. In this task, use the jieba Chinese word segmentation kit, Jieba.cut () method to segment the text, and use/to cut off the jieba words, and the code analysis is as follows:

```
Use the jieba Chinese Word Segmentation kit
importjieba
The text path of the word to be divided
sourceTxt = '.. /1 Text deduplication /qvchong.txt'
Text path after word segmentation
targetTxt = 'fenci.txt'
Manipulate the text
with open(sourceTxt, 'r', encoding='utf-8') as sourceFile, open(targetTxt, 'a+', encoding='utf-8') as targetFile:
for line in sourceFile:
seg = jieba.cut(line.strip(), cut_all=False)
Use/partition between words
output = '/'.join(seg)
targetFile.write(output)
targetFile.write('\n')
print(' Write successfully! ')
sourceFile.close()
targetFile.close()
```

The result of word segmentation is shown in Figure 2.

According to the results of Bayesian analysis, the emotional trend of the comment data was finally obtained, which was then visually processed to make a pie chart and the final conclusion was obtained, as shown in Figure 3 below:

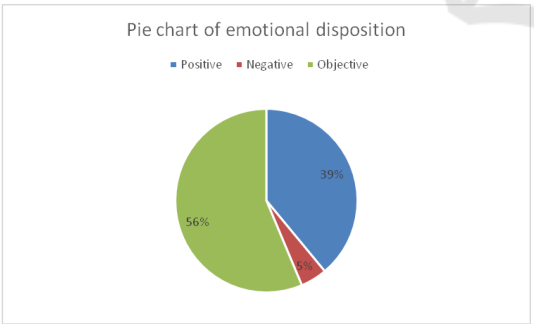

Figure 3: Pie chart.

According to the pie chart, we can see more clearly the emotional state of contemporary young people towards the social status quo. Only a small part of them hold a negative attitude.

3 CONCLUSION

The main result of this design completed the sentiment analysis of comments on a hot topic on Bilibili website. Mainly use familiar development tools for research, and combined with basic knowledge for detailed design and implementation. Machine learning plays an important role in sentiment analysis of comments. Comment classification is closely related to word segmentation, data source, feature selection and parameter selection. At the initial system development level, it is necessary to be familiar with the process of review analysis work and have a basic knowledge of appropriate software programs. From the very beginning, a thorough understanding of the whole, although there are many problems in the whole process, the final result, detailed design and final testing are still acceptable. In this process of exploration, I encountered many problems, but at the same time, I also got a lot of professional solutions and good suggestions.

Today's Internet era has driven the development of entertainment platforms, and people express their views and opinions reasonably under the restriction of rules. Emotional analysis of these statements can complete the grasp of the social group's views and attitudes, and improve the management of public opinion monitoring. Emotion analysis is the process of analyzing, processing, concluding and reasoning the subjective text with emotion. According to different types of texts processed, sentiment analysis can be divided into news comment based sentiment analysis and product comment based sentiment analysis. Among them, the former is mostly used for public opinion monitoring and information forecasting, while the latter can help users understand the reputation of a certain product in the eyes of the public. At present, there are two common methods of emotion polarity analysis: the method based on emotion dictionary and the method based on machine learning. This paper uses sentiment analysis based on news comments for public opinion monitoring.

REFERENCES

Bourouis, S., Alroobaea, R., Rubaiee, S., Andejany, M., Almansour, F. M., & Bouguila, N. Markov Chain Monte Carlo-Based Bayesian Inference for Learning Finite and Infinite Inverted Beta-Liouville Mixture Models[J]. *IEEE Access*, 2021, 9, 71170-71183. http://doi.org/10.1109 /access.2021. 3078670

Liu, K., & Chen, L. Medical Social Media Text Classification Integrating Consumer Health Terminology [J]. *IEEE Access*, 2019, 7, 78185-78193. http://doi.org/10.1109/access. 2019.2921938

Zhu, X. H., Xu, Q. T., Chen, Y. S., Chen, H. C., & Wu, T. J. A Novel Class-Center Vector Model for Text Classification Using Dependencies and a Semantic Dictionary [J]. *IEEE Access*, 2020, 8, 24990-25000. http://doi.org/10.1109/access.2019.2954106

Rogers, D., Preece, A., Innes, M., & Spasic, I. Real-Time Text Classification of User-Generated Content on Social Media: Systematic Review [J]. *IEEE Transactions on Computational Social Systems*, 2022, 9(4), 1154-1166. http://doi.org/10.1109/tcss.2021.3120138

Abdalla, H. I., & Amer, A. A. On the integration of similarity measures with machine learning models to enhance text classification performance [J]. *Information Sciences*, 2022, 614, 263-288. http://doi.org/10.1016/j.ins.2022.10.004

Villa-Blanco, C., Bregoli, A., Bielza, C., Larranaga, P., & Stella, F. Constraint-based and hybrid structure learning of multidimensional continuous-time Bayesian network classifiers [J]. *International Journal of Approximate Reasoning*, 2023, 159. http://doi.org/10.1016/j.ijar.2023.108945

Han, M., Wu, H. X., Chen, Z. Q., Li, M. H., & Zhang, X. L. A survey of multi-label classification based on supervised and semi-supervised learning [J]. *International Journal of Machine Learning and Cybernetics*, 2023, 14(3), 697-724. http://doi.org/10.1007/s13042-022-01658-9

Ajitha, P., Sivasangari, A., Rajkumar, R. I., & Poonguzhali, S. Design of text sentiment analysis tool using feature extraction based on fusing machine learning algorithms[J]. *Journal of Intelligent & Fuzzy Systems*, 2021, 40(4), 6375-6383. http://doi.org/10.3233/jifs-189478

Gao, H. Y., Zeng, X., & Yao, C. H. Application of improved distributed naive Bayesian algorithms in text classification [J]. *Journal of Supercomputing*, 2019, 75(9), 5831-5847. http://doi.org/10.1007/s11227-019-02862-1

He, J., Du, C. Y., Zhuang, F. Z., Yin, X., He, Q., & Long, G. P. Online Bayesian max-margin subspace learning for multi-view classification and regression [J]. *Machine Learning*, 2020, 109(2), 219-249. http://doi.org/10.1007/s10994-019-05853-8

Hao, S. L., Zhang, P., Liu, S., & Wang, Y. H. Sentiment recognition and analysis method of official document text based on BERT-SVM model [J]. *Neural Computing & Applications*, 2023. http://doi.org/10.1007/s00521-023-08226-4

Cardenas, J. P., Olivares, G., & Alfaro, R. Automatic text classification using words networks [J]. *Revista Signos*, 2014, 47(86), 346-364. http://doi.org/10.4067/s0718-09342014000300001

Li, L. F., Li, W. X., & Gong, D. Q. Naive Bayesian Automatic Classification of Railway Service Complaint Text Based on Eigenvalue Extraction [J]. *Tehnicki Vjesnik-Technical Gazette*, 2019, 26(3), 778-785. http://doi.org/10.17559/tv-20190420161815