Text Mining for Mental Disease Screening for Social Network Users

Chen Wang*

School of Engineering, The University of Sydney, Sydney, 2050, Australia

*Corresponding author's e-mail:wchen19991218@163.com

Abstract. The widespread popularity of mobile internet has made the public more frequently share their lives and express their thoughts and emotions on social networks. In this context, Twitter has become the world's largest user base social platform, accumulating a large amount of user generated text information. In order to gain a deeper understanding of the emotional tendencies in these textual information, we used the Sentient140 dataset for analysis. Through exploratory data analysis of Twitter text information, we compared different feature extraction methods and classification algorithms. This process not only helps us understand the emotional expression of Twitter users, but also provides a foundation for subsequent analysis. By evaluating the F1 value, we determined the optimal feature extraction and classification parameters for this task. The optimization of this step makes our model perform well in sentiment classification tasks, providing an effective method for text sentiment mining. The final analysis process and results of this study not only provide strong support for sentiment classification tasks based on text information, but also provide useful references for the diagnosis of mental diseases, such as depression. This means that our research is not only limited to social network analysis, but also has practical application potential in a wider range of health fields.

Keywords: Text mining; Social networks; Mental illness screening

1 Introduction

Depression, as one of the most common psychological disorders in today's society, has a profound and widespread impact, and is also a huge challenge to the overall health and stability of society. Firstly, we need to gain a deeper understanding of the essence of depression and its global prevalence. The main characteristic of depression is prolonged and persistent depression, manifested as extreme emotional instability[1]. Patients often experience a series of negative emotions, such as sadness, inferiority, pain, pessimism, and worldliness, which gradually deepen and may eventually evolve into serious mental health problems, even life-threatening. Among the specific symptoms, patients with depression often have somatic symptoms, such as chest tightness, shortness of breath, etc., which adds some difficulty to the diagnosis and treatment of the disease. Globally, the number of depression patients is as high as 322 million, with

I. A. Khan et al. (eds.), *Proceedings of the 2025 4th International Conference on Humanities, Wisdom Education and Service Management (HWESM 2025)*, Advances in Social Science, Education and Humanities Research 934,
https://doi.org/10.2991/978-2-38476-422-8_4

a prevalence rate of 4.4%, ranking among the top causes of labor loss[2]. As a populous country in the world, China has over 90 million patients with depression, and the incidence rate has been increasing year by year, which has aroused great concern for public health and social psychological health[3].

However, the challenge of depression is not only reflected in its universality. Social factors have a significant impact on the diagnosis of depression, resulting in a relatively low diagnostic rate. Even with a clear diagnosis, the intervention and treatment of depression still face serious challenges, often unable to achieve ideal results due to insufficient timeliness. This is mainly because in society, there is still a significant lag in understanding mental health issues, and many people's understanding of depression remains superficial, unable to truly understand the patient's psychological condition, and lacking timely support and care[4]. For the diagnosis of depression, a comprehensive evaluation is required in medicine, including medical history, clinical symptoms, course of disease, physical examination, and laboratory tests[5]. However, the classification of depression and confusion with other mental illnesses increase the difficulty of diagnosis. This has also sparked a reflection on the medical system, and how to more accurately diagnose and distinguish various psychological diseases has become one of the important issues that need to be addressed in the current medical field.

Depression is not only a matter of individual health, but also a huge threat to social stability. The quality of mental health is directly related to the overall harmony and stability of society[6]. In this information age, people are facing increasing life pressure, and the complexity of social networks also increases the risk of mental health. Therefore, the prevention and treatment of depression is urgent and requires joint efforts from all sectors of society to increase awareness of mental health issues and promote the development of social mental health.

Taking into account these key points and conclusions, we urgently need to take effective measures, including improving public awareness of depression, establishing a more sound mental health service system, and strengthening research in the medical field on psychological diseases such as depression. Only through the efforts of the whole society can we better understand and respond to depression, and create a healthier, equal, and caring society[7].

2 Research Status

With the booming development of social networks, modern people are increasingly relying on the digital world. Meanwhile, the massive amount of online text information provides the possibility for various text-based applications, such as online public opinion analysis and assisted diagnosis of mental illness. In recent years, more and more research has focused on using social network information to assess users' emotional and mental states. From the perspective of sources, social networks come from various sources, such as Weibo in China, Twitter in the United States, and other sources such as Yahoo. From the perspective of specific research issues, the research topics have become increasingly diverse, including emotional analysis, diagnosis of mental illnesses such as depression, and multitasking[8].

GHOSH S, EKBAL A et al. designed a multitask classifier to simultaneously distinguish the emotional and depressive states of text.

SUMAN C, CHAUDHARI R, and others conducted multimodal fusion of image and text data and designed a multitask classifier that simultaneously discriminates between user gender and emotional state[9].

In addition, with the advent of the big data era, deep learning methods have also been increasingly used. For example, in Fang Zhenyu's research on "A Social Network Psychological Disorder Detection Method Based on Depression Dictionaries," he used Word2Vec to generate word vectors, construct depression dictionaries suitable for predicting depression scenarios, and then used short-term and short-term memory networks for classification tasks to predict and classify users' depression[10].

At the time of writing, this article thoroughly investigated the research situation of similar articles. According to the search results of relevant journal search engines, the Twitter dataset used by similar articles is generally at the level of ten thousand data. For example, GHOSH S et al. collected 10659 Twitter texts in their study, NASEEM et al. collected 90000 Twitter data in their study, and GUPTA et al. collected 12741 data in their study, Only the study by Zhou et al. collected 94707264 pieces of data reaching the level of tens of millions[11]. Based on the research content and results presented in similar articles, it can be seen that the larger the data volume, the more individuals it contains, and the richer the information it contains. Therefore, this article actively learns from the research of Zhou et al. based on the large data volume, and adopts the Sentient140 dataset. The final dataset reaches 2.1 million, providing rich materials for text mining in this article[12].

3 Experimental Methods

3.1 Data Set

This article uses the Sentient140 dataset, which includes a total of 2.1 million tweets from March 5, 2010 to July 3, 2010, covering English information published by 659775 users. In this dataset, various key fields provide rich information, including label (Polarity), Twitter ID, timestamp, identifier, username, and Twitter text. These fields provide researchers with detailed Twitter content and relevant context, providing a solid foundation for in-depth analysis of user emotional states. The label section is divided into positive and negative sections, containing 1.05 million positive and 1.05 million negative data, respectively[13]. The classification is very balanced, which helps to maintain the objectivity and accuracy of the research.

The main purpose of this dataset is to conduct emotional analysis, which involves annotating Twitter texts to reveal the emotional states contained within them. This is crucial for understanding the emotional expression of users on social media platforms[14-15]. By conducting emotional analysis on this vast dataset, researchers can track the fluctuations and trends of user emotions and gain a deeper understanding of emotional changes in different themes, events, or time periods on social media.

3.2 Pretreatment

Raw text data is not an ideal input for classification tasks, mainly because Twitter text has multi-source and heterogeneous features, including short links, emoticons, and non-standard spelling. This heterogeneity requires careful preprocessing of text data to better adapt to the needs of classification models.

In the data preprocessing process, the first step is data cleaning, which aims to remove semantic information and short words from the text. For the removal of short words, it is based on a reasonable assumption that words with fewer than 4 letters have a relatively low probability of carrying valid information. This decision helps to reduce noise and make subsequent feature extraction more accurate[16-17]. On the other hand, converting text to lowercase ensures that words with the same meaning are recognized as a semantic unit, thereby improving the effectiveness of subsequent processing steps.

The further processing stage involves natural language processing techniques, which use the NLTK toolkit. The removal of stop words is to eliminate common vocabulary that frequently appears in text but lacks practical semantics, such as "A", "the", "is", etc. This step helps to refine the text and make it more focused on information with classification value.

Symbolization is the process of separating text into sequences of words using spaces, which makes the text easier for computers to understand and process. The operations of word rooting and word form restoration are similar to lowercase conversion, aiming to normalize words with the same semantics but different forms, further improving data consistency.

Overall, this series of data preprocessing steps constitute a complete and detailed process, providing more informative and consistent input for subsequent classification tasks. By eliminating noise and standardizing text forms, this process provides a better data foundation for machine learning models, improving the accuracy and efficiency of classification tasks. This in-depth processing method not only makes the text more suitable for classification tasks, but also provides the model with better generalization ability, enabling it to better handle text data from different sources and styles(figure 1).

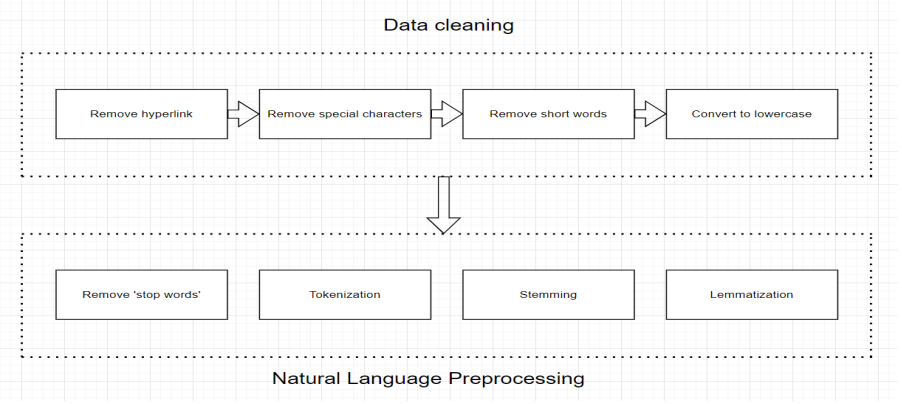

Fig. 1. Text preprocessing process

3.3 Feature Extraction

One of the key steps in document sentiment discrimination is to extract document features, and selecting appropriate methods can significantly improve classifier performance. Due to the fact that documents are essentially ordered and indefinite word lists, the commonly used feature description method is the word bag model. Although the word bag model is lossy, it is essential when dealing with the unstructured and ambiguous characteristics of natural language. The statistical indicators of the word bag model include word frequency, Boolean word frequency, TF-IDF, word vectors, etc. The specific selection depends on the characteristics of the dataset.

Neural network models, such as AutoEncoder and RBM, have the ability to generate unsupervised document vectors. Compared to word bag vectors, they perform better in terms of performance, but this also comes with greater computational overhead. In the feature extraction stage, we adopted two word bag models, CountVector and TF-IDF, respectively, considering the comprehensive indicators of word frequency and multiple documents.

In addition, the impact of monolingual and binary language models on the extraction performance was considered. The N-gram method is implemented by sorting documents and window segmentation, introducing more considerations into the feature extraction stage. Overall, in order to achieve these goals, this article chooses to use the tools provided by the Scikit learn library for feature extraction. This comprehensive method demonstrates excellent performance in handling document sentiment discrimination problems.

3.4 Pretreatment

This study compared the F1 values of XGBoost, Logistic Regression (LR), and Linear SVC algorithms under different feature extraction parameters, and obtained the following key points and conclusions.

As an optimized version of the GBDT algorithm, XGBoost has the advantages of high efficiency, scalability, and strong robustness. However, its ability to capture features is relatively weak when processing high-dimensional data. In contrast, deep learning algorithms perform better in situations with large amounts of data.

Logistic regression is widely used to solve binary classification problems, with significant advantages such as simple implementation, fast speed, and low resource demand. However, logistic regression is prone to underfitting and can only handle binary classification problems, and requires data to be linearly separable.

Linear SVC is a classifier based on SVM and using linear kernels, which has low sample size requirements and exhibits strong robustness. This gives it advantages in certain specific scenarios.

Overall, selecting an appropriate classifier based on task requirements is crucial. XGBoost performs well in terms of efficiency, scalability, and robustness, but may be limited in processing high-dimensional data. Logistic regression is suitable for simple, fast, and resource limited scenarios, but may perform poorly on complex data. How-

ever, linear SVC has lower requirements for sample size and is suitable for scenarios with strong robustness.

By comparing the F1 values of the three algorithms, the research results can be objectively evaluated and the credibility of this study can be improved.

4 Data Analysis

4.1 Time Latitude Distribution of Data

The visualization of data presented in terms of hours of the day reveals some key points and conclusions of Twitter's emotional distribution. Throughout the entire day, we observed clear time distribution trends in both positive and negative tweets.

Specifically, we noticed that Twitter sentiment showed a downward trend between 8:00 and 20:00, and began to rise thereafter. It is worth mentioning that there are relatively few positive tweets during this time period, while there are relatively more negative tweets. This may be related to the characteristics of working hours, as during this time range during the day, people may express negative emotions due to work pressure.

However, surprisingly, as we entered between 22am and 5am the next day, the situation reversed. During this period, the number of positive tweets was much higher than that of negative tweets. We speculate that this may be because most people use this period of time for rest and entertainment activities, and these positive experiences may lead to more positive emotional expression.

Although we have proposed these possible explanations, it should be noted that these speculations lack further data support. Therefore, in order to gain a deeper understanding of the reasons for this trend, we need to obtain more data in order to conduct a more comprehensive and accurate analysis. This will help reveal the deeper reasons behind these time trends observed in Twitter's emotional distribution(figure 2).

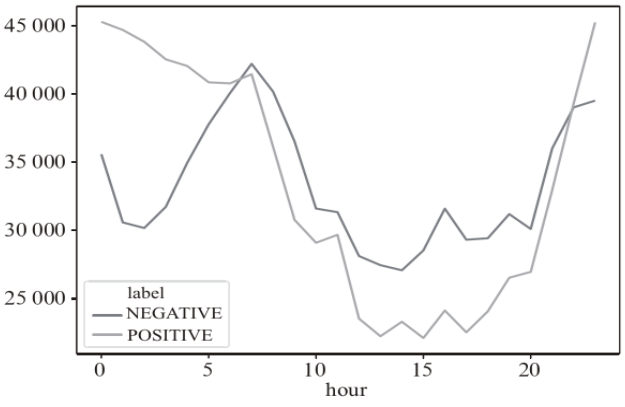

Fig. 2. Time distribution

4.2 Word Cloud Analysis

Through in-depth analysis of Twitter data, we have identified some key points and conclusions. Firstly, in positive tweets, high-frequency word cloud analysis shows that 'Love' is a prominent word, indicating that positive emotions dominate these tweets. On the contrary, in negative Twitter, "miss" is a prominent word that implies a feeling of loss or longing.

Secondly, tag analysis revealed an important finding that tags starting with "#" play a crucial role in the topic or related topic of Twitter information. This means that we can better understand the content and underlying themes of Twitter by carefully analyzing these tags.

Finally, in the cloud of negative tweets, 'fail' is one of the high-frequency words with a strong emotional color. Therefore, we can conclude that analyzing the information in tags is crucial for text sentiment analysis, as these tags may contain key clues about the emotional state of Twitter authors. By integrating high-frequency word cloud analysis and tag analysis, we can have a more comprehensive understanding of Twitter data and delve deeper into the emotions and themes contained within it(figure 3).

Fig. 3. All Twitter Words Cloud

5 Conclusion

We use F1 values to evaluate the performance of the algorithm and suppress random factors by conducting 10 fold cross validation to ensure that the results we obtain reflect

the average performance of the algorithm. In order to further improve the reliability of the results, we used hierarchical random partitioning, setting the ratio of the training set to the test set to 80% and 20%.

By analyzing the results in Table 1, we found that logistic regression achieved the best results under various parameter configurations. This indicates that logistic regression is a powerful algorithm choice in our task.

Especially when using the univariate TF-IDF method for feature extraction, logistic regression performed well, reaching the optimal F1 value of 76.1%. This result emphasizes the superiority of logistic regression in our specific dataset and task, providing strong support for further practical applications.

Table 1. F1 value of each classifier

Feature extraction/classifier	XGBC	LR	Linear SVC	Feature extraction/classifier
One dollar CountVector	0.720	0.755	0.738	One dollar CountVector
Binary CountVector	0.676	0.714	0.699	Binary CountVector
Univariate TfidfVector	0.721	0.761	0.746	Univariate TfidfVector
Binary TfidfVector	0.676	0.722	0.710	Binary TfidfVector
Feature extraction/classifier	XGBC	LR	Linear SVC	Feature extraction/classifier
One dollar CountVector	0.720	0.755	0.738	One dollar CountVector
Binary CountVector	0.676	0.714	0.699	Binary CountVector
Univariate TfidfVector	0.721	0.761	0.746	Univariate TfidfVector
Binary TfidfVector	0.676	0.722	0.710	Binary TfidfVector
Feature extraction/classifier	XGBC	LR	Linear SVC	Feature extraction/classifier
One dollar CountVector	0.720	0.755	0.738	One dollar CountVector
Binary CountVector	0.676	0.714	0.699	Binary CountVector
Univariate TfidfVector	0.721	0.761	0.746	Univariate TfidfVector
Binary TfidfVector	0.676	0.722	0.710	Binary TfidfVector

References

1. GHOSH S,EKBAL A,BHATTACHARYYA P(2022)What Does Your Bio Say? Inferring Twitter Users Depression Status From Multimodal Profile Information Using Deep Learning.IEEE Transactions on Computational Social Systems(05),1484-1494.
2. SUMAN C,CHAUDHARI R,SAHA S,et al(2022).Investigations in Emotion Aware Multimodal Gender Prediction Systems From Social Media Data.IEEE Transactions on Computational Social Systems(06).1-10.
3. NASEEM U,RAZZAK I,KHUSHI M,et al(2021).COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis.IEEE Transactions on Computational Social Systems(4).1003-1015.

4. ZHOU J,ZOGAN H,YANG S,et al(2021).Detecting Community Depression Dynamics Due to COVID-19 Pandemic in Australia.IEEE Transactions on Computational Social Systems(4).982-991.
5. Zhang Yixiang (2023) Master's thesis on the research of online public opinion in the 14th National Games based on text mining technology, Xi'an Institute of Physical Education.
6. Xu Xiang, Zhang Lingyuan&Wang Yuchen (2023) Empirical Study on the Phenomenon and Effects of Content Convergence Gravity in Social Networks - Weibo Data Mining Based on Word2vec Frontiers of Data and Computing Development (02), 136-149.
7. Wu Shuai, Shi Yi, Yang Xiuzhang, and Xiang Meiyu (2022) Research on Short Text Mining Based on Social Network Analysis and LDA Topic Mining Modern Electronic Technology (20), 124-128.
8. Huang Xifeng and Liao Hongzhou (2022) Mining and predicting group events based on social networks Computer Technology and Development (06), 39-44.
9. Chen Guanshan (2021) Master's thesis on emotional transmission and its influencing mechanisms in the evolution of overseas public opinion on social networks, Guangdong University of Foreign Studies.
10. Sun Yu and Qiu Jiangnan (2022) Research on the Influence of Opinion Leaders Based on Network Analysis and Text Mining Data Analysis and Knowledge Discovery (01), 69-79.
11. Xu Xiang (2021) The phenomenon and mechanism of "standard idols" in social network content production: Weibo text mining based on latent semantic analysis Journal of South China University of Technology (Social Sciences Edition) (04), 109-121.
12. Wang Xiaoqian (2021) Master's thesis on the evolution and prediction of online public opinion popularity based on text mining, Dalian Maritime University.
13. Zhou Haichen (2021) Doctoral Dissertation on Academic Innovation Knowledge Mining for Academic Evaluation, Nanjing Agricultural University.
14. Li Zhengguang (2021) Doctoral Dissertation on Semantic Enhancement Based Biomedical Text Mining Research, Dalian University of Technology.
15. Jiang Runlian (2020) Master's thesis on automated recognition and visualization of academic research trends based on text mining technology, Shenzhen University.
16. Wang Ying, Ku Tingting, Xu Shuping, Li Weiqiang&Yuan Bo (2020) Analysis of the emotional components of awe: Text mining based on social networks Psychological Technology and Applications (04), 235-242.
17. Zhao Cunxiu (2019) Enterprise innovation knowledge integration and management based on social networks Information Technology (12), 135-140.