Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25787

BERT-Based Depressive Sentiment Analysis and Explainability for Chinese Texts

Hongfu Yu

Canada British Columbia International Schools-Chengdu, Chengdu, China
Bobbyxyu2008@outlook.com

Abstract. The extensive application of NLP technologies has greatly increased in recent years, for instance, in translation, text classification, and text generation. With the advancement of technology, models have become more complex, which raises the difficulty of model explanation. The paper explores sentiment analysis classification for identifying depressive text using the BERT model and demonstrates the explanation of model outputs through explainable artificial intelligence (XAI) technology. The pre-trained BERT model was fine-tuned by using a Chinese text dataset, which consists of Chinese texts with depression labels obtained from social media like Weibo, REDnote, and Twitter. The paper reveals that the deep learning model, leveraging contextual information and interpretability techniques, effectively analyzes depressive sentiment, and has endowed the model with certain interpretability.

Keywords: BERT, Depression Text Detection, Sentiment Analysis, SHAP

1. Introduction

Mental problems have become one of the most prevalent existing disorders all over the world. Among them, depression is a relatively common condition, which affects around 3.8% of the global population [1]. Nowadays, the development and prevalence of social media provide a new way for information sharing. Many people tend to share their lives, opinions, and feelings through texts on social media. Some of these texts can reflect users' mental conditions to some extent, providing an opportunity to detect depressive tendencies by analyzing emotional information in the texts [2]. The development of NLP technology has achieved remarkable improvement in recent years due to the involvement of deep neural networks, which are able to greatly improve models' accuracy. Increased model complexity makes outputs difficult to explain during learning and testing, a phenomenon often referred to as the black-box effect [3]. However, in highly sensitive fields such as mental health, there is an urgent need for user trust and transparency. To solve this problem, Explainable AI (XAI) offfers a potential solution. It helps humans understand and trust the decisions made by models and when combined with visualization techniques to provide a more comprehensible result [4].

Traditional methods for identifying depressive disorders largely depend on patients' subjective responses, which are limited by many factors. For instance, these factors include the environment for diagnosis and patients' subjective biases towards mental illness. To address the issues, this study

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ25787

fine-tunes a pre-trained BERT model using labeled depressive text data collected from platforms such as Weibo, REDnote, and Twitter. The goal is to perform binary classification of depressive content. Additionally, SHapley Additive exPlanations (SHAP) are employed to interpret the model's decision-making process. By integrating BERT with SHAP in the context of Chinese depressive texts, this research provides an interpretable technical framework for studies in the field of mental health.

2. Literature review

2.1 Advances in depressive sentiment analysis

Depression sentiment analysis technology can be classified into two categories including traditional machine learning methods and deep learning methods [5]. Traditional machine learning methods rely on artificially designed rules and feature engineering. According to Bokolo & Liu, the Term Frequency-Inverse Document Frequency (TF-IDF) technique is an effective way to achieve that [4]. But this dependency increases the time cost, and traditional methods perform poorly when dealing with complex emotions. Commonly used traditional machine learning classifiers in depression text detection include logistic regression, Naive Bayes, and random forests [5]. Methods based on deep learning and transformers use RNN, LSTM, and BERT. These models have a better performance in handling contextual features, but the increasing model complexity makes it impossible to understand how models work and trust the models, which raises the need of interpretation techniques. Another challenge that deep learning needs to face is the loss of authenticity caused by a large amount of unrelated noise in social media texts [5].

At the same time, Internet slang and homophonic puns are widely used in Chinese social media, which makes it more difficult to accurately capture the emotional tendency of the text. Traditional machine learning methods rely on fixed feature engineering and are difficult to dynamically adapt to such flexible and changeable language phenomena.

2.2 A review of SHAP

In recent years, XAI techniques have been applied in medical fields due to the need for intelligibility requirements [4]. In depression text detection tasks, the involvement of SHAP can provide both local explanations and global explanations. Bhuiyan et al combined CNN-BiLSTM with SHAP interpretation technique to detect English depression texts. As a result, SHAP enhanced the model's transparency, resulting in high accuracy in depression detection [6]. Nevertheless, research on applying SHAP to depressive texts detection is mainly based on English corpora, fewer studies investigate the feasibility of applying it to non-English languages. Furthermore, SHAP only provides the contribution value of terms, suggesting that SHAP may be limited in explaining long texts or complex emotions [6]. This limitation is more prominent in the context of Chinese media. In Chinese social media, phenomena such as Internet slang and polysemous words are widespread. SHAP's contribution value calculation for polysemous words may be ambiguous, making it difficult to accurately capture the emotional depth in the context. These deficiencies highlight the necessity of verifying the feasibility of the SHAP method in the interpretation of Chinese depressive texts.

DOI: 10.54254/2755-2721/2025.BJ25787

3. Methodology

3.1 Research dataset

This study uses the dataset from Kaggle data repository to fine-tune the BERT-Base Chinese model provided by Google in 2019 [7]. The dataset includes about 10,000 Chinese reviews obtained from popular social media platforms, including Weibo, REDnote, and Twitter [8]. Data cleaning involves filtering duplicate texts and applying noise reduction techniques. These texts have been classified into five depression levels. The level 0 represents non-depression, and the level 1 to 4 represent mild to severe depression. However, in this study, only level 0 is considered non-depression, and other levels are classified into depression category.

3.2 BERT

In the field of NLP-based depression text detection, researchers have already extensively utilized multiple models. In this study, the BERT-based model is applied to depression text detection process, which has 12 layers, 1024 hidden dimensions, and 12 self-attention heads. BERT model is a type of pre-trained model introduced by Devlin et al., and it is a unified architecture that can be applied to multiple tasks [7]. The architecture of BERT consists of two stages, including pre-training and fine-tuning.

In the pre-training step, BERT does not use traditional unidirectional methods, which are left-to-right or right-to-left. Instead, BERT's pre-training step is processed by two tasks, which are both unsupervised. The first, named Masked LM, allows the model to be trained bidirectionally, meaning it may perform better on complex texts. However, the limitation of this method is that masked words in pre-training step do not appear in fine-tuning step. To address this issue, researchers have introduced a probability mechanism, which consists of three components. First, 80% of masked words remain. Second, 10% of words are replaced by random words. Otherwise, the words remain in their original form [7]. The second task is Next Sentence Prediction (NSP). The NSP is included mainly to understand relationships between multiple sentences in many downstream tasks. NSP operates through a simple binary classification task that classifies the relationship between two sentences into IsNext and NotNext [8].

The fine-tuning step of BERT is based on its Transformer self-attention mechanism [8]. In this study, the pre-trained BERT-base-Chinese model is fine-tuned by Chinese texts with depression and non-depression labels, enabling it to be specifically adapted to identify depression texts with a higher accuracy.

3.3 SHAP

In this study, the involvement of interpretation technique - SHAP aims to mitigate the problem of poor intelligibility. Lundberg and Lee introduced the concept of SHAP. It is based on the concept of game theory, which was introduced by Shapley in 1953 [9]. Originally, it focused on how to fairly distribute rewards among a group of players who contribute to a certain outcome [10]. When applied to model interpretability, the important features are equivalent to players, and outcomes become the decisions made by the model. SHAP can provide both local and global interpretations. Local interpretation mainly focuses on explaining single sample, whereas global interpretation shows the global decision-making pattern [3,10]. For this research SHAP mainly focuses on local interpretability, which tries to highlight words that are most contributive to depression detection.

DOI: 10.54254/2755-2721/2025.BJ25787

In the interpretation process, SHAP is a framework that unifies six explanation methods. The core concept of SHAP is the additive feature attribution methods:

$$g\left(z'\right) = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i' \tag{1}$$

This equation means the output of the interpretation model is the sum of contributions of each feature. A unique property of SHAP compared to other additive feature attribution explanation methods that it satisfies three desirable properties. It includes local accuracy, missingness, and consistency [9].

SHAP value calculation usually involves approximation methods to get a greater efficiency. For instance, Kernel SHAP is applied to any models, and other algorithms that aim at specific model such as Linear SHAP and Deep SHAP [9]. In this study, the depression texts are interpreted locally by using Kernel SHAP.

4. Experiment

4.1 Model performance

In the fine-tuning step, the training dataset was put into the model for 5 epochs and a relatively low learning rate was set to prevent overfitting problems. According to Table 1, the model performs quite well, with precision, recall, F1-score, and accuracy of 0.9687, 0.9340, 0.9510, and 0.9703, respectively.

Table 1. Model performance metrics

Precision	Recall	F1-Score	Accuracy
0.9687	0.9340	0.9510	0.9703

After the fine-tuning step, the involvement of SHAP local interpretation was applied to the model to visualize some random samples. According to statistics, sentences with words like restlessness, loneliness, and gloom tend to have a higher probability of being classified as depressive texts.

4.2 Summary

According to the results of this experiment, the possibility of combining the BERT-based depression text detection model with SHAP local interpretation has been confirmed. The precision, recall, F1-score, and accuracy of the model demonstrate the high efficiency of detecting Chinese texts that are related to depression. This is largely attributed to the extensive pre-training step of BERT-base-Chinese which is based on a large amount of Chinese texts. This ensures it has relatively strong ability in handling complex texts and can be fine-tuned for detection of depressive texts. Meanwhile, SHAP can highlight the words that most contribute to depression detection when combined with visualization, which solves the incomprehensible issue in a certain extent of black-box models.

5. Discussion

The model also has several limitations. First, the dataset imbalance exists with a 2:1 ratio of depressive to non-depressive texts. As a result, this may cause the model to tend to define a text with depression traits. Second, BERT struggles with long texts as it has a length limitation of 512 tokens.

DOI: 10.54254/2755-2721/2025.BJ25787

SHAP's interpretive approach determines its high calculation cost, especially when dealing with complex structures. Furthermore, SHAP focuses on single words, which means it cannot directly reflect contextual dynamics.

6. Conclusion

This research explores that the development of machine learning has allowed models with complex architecture to be applied to depression text detection by using sentiment analysis techniques. However, models with black-box traits in highly sensitive fields tend to face trust issues from users. In this case, the involvement of XAI techniques may alleviate this problem to a certain extent. Social media has become a common platform that allows people to share their experiences, feelings, and emotions, which provides a chance to detect depression tendencies using NLP techniques. This paper uses pre-trained model BERT-base-Chinese provided by Google, which tries to gain a better performance in handling context and sentences with complex emotional information. The performance of the model is proved by four metrics, including precision, accuracy, F1-score and recall. Additionally, the involvement of SHAP aims to reduce the negative effect caused by black-box models. In conclusion, the feasibility of application of pre-trained model and interpretation method in mental health fields has been demonstrated. Furthermore, it indicates the limitations including model structure limitation, imbalanced dataset, and high calculation cost. Future directions can be expanded to multimodal depression monitoring, with the establishment cross-platform dynamic tracking systems.

References

[1] World Health Organization. (2023). Depression disorder (depression). https: //www.who.int/news-room/fact-sheets/detail/depression
[2] Teck Kiong, Y., Cheah, W. S., Pradana, M., Sharbiniung, H., & Yanto, I. T. R. (2016). Depression detection on Mandarin text through BERT model. Journal of Advanced Research in Applied Sciences and Engineering Technology. 60: 295-311.
[3] Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. https: //doi.org/10.48550/arXiv.2006.11371
[4] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys, 55: 1-33.
[5] Bokolo, B. G., & Liu, Q. (2023). Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques. Electronics, 12(21), Article 4396.
[6] Bhuiyan, M. I., Kamarudin, N. S., & Ismail, N. H. (2025). Enhanced suicidal ideation detection from social media using a CNN-BiLSTM hybrid model. https: //arxiv.org/abs/2501.11094
[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis. pp. 4171-4186.
[8] Deok, K. (2023). Chinese depression text sentence corpus. https: //www.kaggle.com/datasets/kyharndeok/dpreesion
[9] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In: Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., Red Hook. pp. 4765-4774.
[10] Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). SHAP-based explanation methods: A review for NLP interpretability. In: Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022). Gyeongju. pp. 4593-4603.