Analysis of Emoticon based on BERT model

Pengfei Dai [1] , Chenhao Kong [2] and Boxiang Zeng [3,*]

[1] The School of Computer Science, South-Central Minzu University, Wuhan City, Hubei Province, China
[2] The School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou, Zhejiang, China
[3] School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai, China
211042y324@mail.sit.edu.cn

Abstract. With the widespread adoption of digital communication platforms, emojis have become an integral part of conveying subtle emotions and expressions within written content. This paper delves into the application of BERT and its foundational Transformer technology in processing texts enriched with emojis, underlining their indispensable role in contemporary communication. This paper aims to showcase the enhanced capabilities of AI in gratifying the subtleties of emoji-inclusive text, thereby broadening the horizon of text analysis within the sphere of computer science. In this paper, Bert analyzes and classifies the long and short sentences containing emojis, hoping to infer the corresponding emotional relationship between emojis and sentences and let the computer learn to infer the meaning of different emojis from symbols. At present, the results show that the effect of text sentiment analysis based on BERT model is relatively average, the accuracy does not reach the expected standard, and it can only accurately identify the emotion tendency in the text.

Keywords: Emotion analysis, Natural Language Processing, BERT.

1 Introduction

In the past, the association products between text and the so-called emojis or emojis were often relatively one-sided, and emojis were mostly triggered by keywords in a single sentence. In contrast, the keywords of emojis were associative only when part of the text contained in the picture itself was exactly the same. Chinese is a very complex system, its meaning depends on the context and context much more than other languages, it should have a more complex and sophisticated text and emoji association system. Integrating emoticons into written communication signifies a transition towards more expressive and subtle forms of interaction within academic discourse. Emotion analysis plays a crucial role in analyzing Chinese text, as the expression of emotions in Chinese is diverse and abundant, encompassing vocabulary, word order, sentence structure, and more. It enables accurate understanding and interpretation of the emotional nuances embedded within Chinese texts.

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_70

With the development of machine learning and deep learning, researchers began to explore the use of machine learning algorithms and neural network models for sentiment analysis. This data-driven approach can automatically learn the ability to extract emotional features and patterns from text, significantly improving the accuracy and efficiency of sentiment analysis. Particularly with the advent of recurrent neural networks (RNNS) and long and short-term memory networks (LSTM), innovative deep learning techniques have enabled remarkable advancements in sentiment analysis in recent years. These models can model text sequences, capture contextual information and achieve excellent performance in sentiment analysis tasks. Some people explore the use of BERT, trained using domain adaptation techniques, for multi-turn dialogue systems. With the rise and popularity of social media, researchers have begun to focus on the challenges and methods of sentiment analysis on social media. A paper utilizes both BERT and Bi-LSTM models to analyze a large amount of text data from social media platforms for detecting signs of depression or anxiety. But data on social media often contains a lot of non-standard language, emoticons, and abbreviations, which creates additional difficulties for sentiment analysis. Therefore, the researchers have proposed several social media-specific preprocessing and model design methods to address these challenges.

There are related studies that discuss this topic. Yang X, Liu M [1] proposes a deep learning-based method to address the problem of association between Chinese text and emojis. This research establishes a semantic representation space for emojis and Chinese text, enabling more accurate and flexible association recognition. Besides, Li D, Rzepka R, Ptaszynski M, et al [2]delves into the semantic analysis and application of Chinese emojis. This study presents a Chinese emoji classification method based on semantic features, which has been successfully applied to sentiment analysis and emoji prediction in Chinese social media texts.

Therefore, BERT will be utilized to examine data collection and preprocessing in order to investigate the emotional correlation between emojis and text. This project aims to recognize emojis that correspond to identifiable emotions within various Chinese sentences, with the dataset divided into two parts for accuracy assessment.

2 Content

2.1 Data

The data set of Chinese comments on Taobao products is selected from the data website. The text emotion in this data set is more prominent and not implicit, and the data set can be better trained. Moreover, there are many of this data set, and 14,822 comments can be allocated to the training and test sets. The data set is preprocessed, and the for loop and regular expression are used to remove non-Chinese characters, Spaces, and unique labels, and then the text is classified.

Table 1. The emoji to the emotion.

Emoji	Emotion

Emoji1	(๑•́₃•̀๑)
Emoji2	~ (̿▽ ̿ ~)~
Emoji3	(๑^o^๑)

As shown in Table 1, the text is divided into three kinds of emoji: ordinary Emoji1, happier Emoji2, and very happy Emoji3. The parts of a comment that are good and bad are ordinary. Primarily suitable is glad, and only sound is happy.

2.2 Model

The model at the heart of the study is BERT, a groundbreaking natural language processing framework introduced by Google in 2018. BERT stands for a major departure from earlier language models because of its special capacity to comprehend a word's context in a phrase from both sides. Using Next Sentence Prediction (NSP) and Masked Language Modeling (MLM), the model is pre-trained on a sizable corpus of unstructured text, enabling it to capture a profound semantic grasp of language. Deven Shah, H. Andrew Schwartz, and Dirk Hovy propose a unified conceptual framework called the predictive bias framework for Natural Language Processing (NLP) in their paper. Along with a generic mathematical description of predictive bias in NLP, they give an overview of the literature in the field. Label bias, selection bias, model overamplification, and semantic bias are the four primary causes of biases in NLP models, according to their conceptual framework. The authors provide their framework as a leading outline for comprehending predicted biases in NLP and talk about how prior research has handled each form of bias.

By analyzing the words before and after a given word, BERT is able to consider its full context, differentiating itself from traditional sequential models, which process text in a single-directional manner. By capturing more nuanced meanings, this bidirectional knowledge helps the model perform far better when it comes to machine learning tasks like text categorization, entity recognition, and question answering.

The introduction of BERT marked a paradigm shift in natural language processing, prompting an influx of research to explore its capabilities and extend its application. For instance, the Robustly Optimized BERT Pretraining Approach model (RoBERTa) introduced by Liu et al. (2019) builds upon BERT by optimizing its pre-training conditions, including training on a larger dataset and for a longer duration, which leads to even higher performance across several benchmarks.

Further research has demonstrated BERT's adaptability to a diverse range of NLP tasks. Studies, the paper "Aspect-Based Sentiment Analysis using BERT" successfully demonstrated the ability of BERT to capture the contextual representation of words under pre-training, along with its capability to generate additional text, have shown its effectiveness in sentiment analysis, where BERT's deep understanding of contextual nuances allows for more accurate detection of sentiment in text, highlighting its superiority over traditional models that often misinterpret the context or the sentiment of words when they are used in different ways (Reference to specific studies).Moreover, the adaptability of BERT for specialized tasks has been showcased through fine-tuning, where the pre-trained BERT model is further trained on a smaller, task-specific dataset.

Some papers demonstrate the potential of the BERT model for analyzing text sentiment in specific field, such as electronic commerce. These papers show that BERT can be trained and yield good results in various specialized domains. This process has enabled researchers and practitioners to leverage the model's deep linguistic understanding for various applications, demonstrating its versatility and effectiveness across different domains.

In conclusion, BERT represents a significant advance in natural language processing, offering unprecedented capabilities for understanding and generating human language. Its impact is evidenced not only by performing at the highest level on different NLP tasks but also by the vast body of research it has inspired, pushing the boundaries of what is possible in language understanding and processing.

2.3 Performance evaluation

The metric utilized in this study is accuracy, providing an intuitive indication of the success of the research findings.

It is a good result if the test set's accuracy rate exceeds 80%. If the accuracy rate reaches 80%, it indicates that the model is highly applicable to the analysis of text emotion and which emoji should be selected after the text. The model is successful for this training and prediction.

3 Result

Table 2. The result of the first run.

Times of iterations for test set	Accuracy	Loss
First	0.7445	0.5936
Second	0.7749	0.4785
Third	0.7934	0.4310
Fourth	0.8019	0.3957
Fifth	0.8075	0.3682
Results of final test set	0.7319	

The parameter BATCH_SIZE is set to 16, indicating the number of input samples in each model training, while the learning rate LEARNING_RATE is set to 2e-5, as demonstrated in Table 2. After 5 rounds of traversal, the final accuracy of the test set is only 73.19%, which is a low accuracy.

After resetting the parameters, the partition of the training and test sets was modified from a fixed partition to a random 8,2 ratio partition of random. Again five passes were made.

Table 3. The result of the second run.

Times of iterations for test set	Accuracy	Loss
First	0.7387	0.6009
Second	0.7815	0.4820
Third	0.7929	0.4324
Fourth	0.7997	0.3983
Results of final test set	0.7502	

As shown in Table 3, the final accuracy of the test set was 75.02%, showing a slight improvement compared to the initial result. However, it still falls below the desired threshold of 80%. Consequently, the paper proceeded to reconfigure the parameters by modifying LEARNING_RATE from 2e-5 to 5e-5, resulting in an increased accuracy rate of 76.03%.

The accuracy rate is relatively high but still falls short of meeting expectations. The parameters will be further adjusted, or the model will be replaced to achieve the targeted accuracy rate of 80%.

4 Discussion

Based on the provided results and the adjustment process, several discussions and recommendations can be made:

The initial model accuracy was 73.20%, which improved slightly to 75.02% by changing the data set segmentation method and increasing the learning rate. This highlights the significant impact of hyperparameter selection on model performance, although the current adjustments have not yet reached the target accuracy of 80%.

Some aspects need to be limited and improved and discrepancies in model performance between training and test data need thorough examination, with focus on potential enhancements in:

Data augmentation to enhance model generalization through increased diversity and quantity.

Model tuning by exploring different structures or depths for a better fit to the task.

Feature engineering for deeper analysis and extraction to improve data understanding.

Hyperparameter search for optimal configuration through systematic exploration of different combinations. - Ensemble learning by combining predictions from multiple models.

Despite falling short of expectations, current accuracy is approaching desired levels. Continuous experimentation and refinement offer opportunities to gradually enhance model performance. Future efforts may concentrate on improving data quality, optimizing model structure, and adjusting hyperparameters for higher accuracy and improved generalization.

The inclusion of subsequent expressions, based on a thorough analysis of the text's emotional content, can enhance the overall conveyance of its emotional inclination.

Meanwhile, there is a certain matching relationship between different types of emotions and specific expressions. Through the comparative analysis of emotion classification results and added expressions, it is possible to find that some emotional tendenci.

5 Conclusion

In this experiment, sentiment classification tasks were conducted using text data extracted from Taobao product purchase records. A large amount of text data containing intense emotions was collected and preprocessed by removing irrelevant characters and special tags, resulting in a dataset of 14,822 entries.

For the text sentiment classification task, the texts were categorized into three groups: neutral, moderately positive Emoji1, more happy Emoji2, and highly positive Emoji3. Reviews with both positive and negative aspects were considered neutral, predominantly positive reviews were classified as moderately happy, and exclusively positive reviews were labeled as very happy. The Bert model was selected for the sentiment classification task, and multiple rounds of training and fine-tuning were performed. Initially trained with parameters that the batch size is 16 and the learning rate is 2e-5, while the percentage of pupils is 5. It resulted in a test set accuracy of only 73.20%, falling short of expectations.

In an effort to improve accuracy, adjustments were made to parameters along with changing the division ratio between the training set and test set from fixed to random at an 8:2 proportion for five additional rounds of training, which yielded an improved accuracy rate of 75.02%. But the experiment still didn't live up to expectations.

Further efforts to enhance accuracy included adjusting the LEARNING_RATE parameter from 2e-5 to 5e-5, leading to an increased accuracy rate reaching up to76.03%.

Although expectations have not been met, current accuracy is approaching desired levels, continued experimentation and improvement provide opportunities to gradually improve model performance, and future efforts are likely to focus on improving data quality, optimizing model structure, and adjusting hyperparameters for higher accuracy and improved generalization.

In summary, this experiment has explored various parameter settings alongside model selections through multiple rounds of training & tuning. Although there have been improvements in terms of accuracy rates, challenges remain. We are committed to continued efforts to optimize the model further to achieve higher levels of satisfaction that better align with actual needs.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

References

1. Yang X, Liu M: The pragmatics of text-emoji co-occurrences on Chinese social media[J]. Pragmatics, 31(1): 144-172 (2021).
2. Li D, Rzepka R, Ptaszynski M, et al.: Emoticon-aware recurrent neural network model for Chinese sentiment analysis, 018 9th International Conference on Awareness Science and Technology (iCAST). IEEE, 161-166 (2018).
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V..:RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019)
5. Shah D, Schwartz H A, Hovy D. Predictive biases in natural language processing models: A conceptual framework and overview[J]. arXiv preprint arXiv:1912.11078, (2019).
6. Hoang M, Bihorac O A, Rouces J. Aspect-based sentiment analysis using bert[C]//Proceedings of the 22nd nordic conference on computational linguistics. 187-196 (2019)..
7. Xu H, Liu B, Shu L, et al. BERT post-training for review reading comprehension and aspect-based sentiment analysis[J]. arXiv preprint arXiv:1904.02232, (2019).
8. Singh M, Jakhar A K, Pandey S. Sentiment analysis on the impact of coronavirus in social life using the BERT model[J]. Social Network Analysis and Mining, 11(1): 33 (2021).
9. Zeberga K, Attique M, Shah B, et al. A novel text mining approach for mental health prediction using Bi-LSTM and BERT model[J]. Computational intelligence and neuroscience, (2022).
10. Whang T, Lee D, Lee C, et al. An effective domain adaptive post-training method for bert in response selection[J]. arXiv preprint arXiv:1908.04812, (2019).