A Study on Employment Problems and Sentiment Analysis of College Students Based on Bert-BiLSTM

Zihan Chen

Sdu-Anu Joint Science College, Shandong University,Shandong, 264209, China
202000700205@mail. sdu. edu. cn

Abstract. In recent years, Chinese college students have generally faced social problems such as fierce competition for employment and rising youth unemployment. Sentiment analysis of college students' employment attitudes helps them recognize the situation, accurately position themselves, and rationally arrange their college life. Therefore, this study suggests a BERT-BiLSTM-based sentiment classification model that can categorize comments' emotions into three groups: positive, neutral, and negative, and uses big data mining to explore factors contributing to college student employment issuess. The scheme employs the bidirectional encoder representations from transformers (BERT) pretraining model for sentence segmentation and word vectorization, and then feeds the bidirectional long short-term memory (BiLSTM) model with the processing results in order to do thorough feature mining for sentiment polarity categorization and aspect separation. The evaluation metrics are accuracy rate and loss rate. Considering the outcomes of the experiment, Bert-BiLSTM technique achieves 0. 7522 in accuracy rate and 0. 5891 in loss rate. In contrast to applying the Bert model (0. 7269, 0. 6138) and the LSTM model (0. 6252, 0. 6836) alone, it exhibits a certain advantage in the final results. The study's results can relatively accurately predict the emotional tendency of college students' employment in contemporary society, thus providing emotional guidance and suggesting directions for related social issues such as college students' mental health and employment choices.

Keywords: Sentiment analysis; BiLSTM; BERT; employment attitude text; Attention mechanism.

1 Introduction

The rapid development of the Internet and the rise of the big data era have accelerated the speed of information dissemination. With the popularization and application of social media, people prefer to use it as a tool for information exchange and sharing [1]. Some hot social issues are discussed in depth by users, such as the employment situation of college students, which is currently in the spotlight. In recent years, domestic college students have generally faced social problems such as fierce competition for employment, the lack of adaptation between college students' employment concepts and the actual job market, as well as the continuous rise of youth

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_71

unemployment, and the employment pressure on college graduates has increased. Physical and mental health problems of college students may arise as a result, which will have a serious impact on their physical and psychological health if they cannot be solved or released in time [2]. Compared with news propaganda and expert advice, people are more willing to learn to refer to the content and details of students' information, and the comment texts on social platforms can usually directly or indirectly reflect the viewpoints and attitudes, emotional tendencies, and planning suggestions of the main body of students [1]. On this basis, the employment attitude text sentiment analysis can help to improve the attention to the employment attitude of college students and personal psychological problems. In the era of higher education popularization, colleges, universities and the government can better guide policies through public opinion analysis to promote college graduates to achieve full and high-quality employment.

Sentiment analysis is a natural language processing method that looks for emotional patterns in written information [3]. Sentiment analysis is commonly used to assess users' emotional responses to products, services, or brands and detect emotional orientation in social media, texts, and comments [1]. Current research has produced several techniques that can be used to perform sentiment analysis tasks [4]. Large language models such as Bert's have been widely used to analyze textual data [5], and research on sentiment analysis in various domains is rapidly evolving.

Wang P, Xiong X [6] proposed a sentiment analysis technique for book reviews using the BERT-BiLSTM hybrid model to solve the problem of low accuracy of current book review sentiment analysis methods. In order to achieve fine-grained sentiment analysis, the researchers integrated deep learning techniques with sentiment language to build an overarching framework for sentiment analysis, including three phases of topic identification, sentiment recognition and topic emotion recognition, fully learning contextual information in text data through Bert-BiLSTM model architecture, utilizing the attention mechanism to draw attention to key elements and increasing resource consumption efficiency, so as to achieve book review sentiment analysis accurately. Si H, Wei X [1] team utilized LSTM and Bert for text sentiment analysis to solve inefficient information extraction when traditional deep learning methods deal with complex linguistic phenomena. It involves training word embeddings using skip-gram model, supplying the neural network with a 2D feature matrix and word embeddings together,and pre-training the language model with Bert to solve polysemy problems. Text classification is achieved by feeding the input items to LSTM for further feature learning. Fadhlurrahman J A M et al [7] investigated using BERT, BiLSTM and CRF models for sentiment analysis of game reviews on Steam platform. The researchers described in detail the methodology of the data preprocessing phase, which includes steps such as text conversion, denoising, and label normalization to ensure data quality. The outcomes of the experiment indicate that the BiLSTM model has problems in classifying certain positive sentiment sentences, which may lead to misclassification. However, by combining the BERT and CRF models, the paper's suggested technique performs better on the sentiment analysis challenge. Li X, Lei Y, Ji S [3] proposed a BERT-BiLSTM model for online Chinese buzzword sentiment analysis and conducted experimental evaluations. The experimental results show that the BERT-BiLSTM

model outperforms other models on the OCB dataset with higher F1 scores, recall and precision.

This paper aims to explore college students' attitudes towards employment issues and the main factors that lead to employment pressure. The comments of popular videos on the topic of college students' employment on the platform of BiliBili were crawled as a dataset using python. The keywords extracted from the preprocessed comments were visualized using big data mining techniques to generate a word cloud map. Finally, the sentiment analysis was used to predict the college students' sentiment tendency towards employment. Unlike the traditional English text analysis and single model training, this study adopted the Bert-BiLSTM sentiment classification model and utilized Bert-base-chinese for Chinese text analysis to classify the comment sentiment into three emotion classifications: positive, neutral and negative. The method utilized the Bert pretraining model to segment sentences and convert word vectors. After the BiLSTM model had been re-trained using the findings, the results were predicted and categorized using the trained model. The accuracy and loss rate were used as the evaluation criteria.

2 Method

2.1 Data Preprocessing

No public dataset is available for reviewing texts on college students' employment attitudes, and certain platforms employ anti-crawling techniques and encryption methods to safeguard users' confidential data. As a gathering place for the young generation and a leading media community in China, BiliBili has rich and diverse contents in each video section, and its audience is mainly students, who are more willing to comment on hot issues in society, and their views on the current situation of college students' employment are more authentic and timely. Therefore, the dataset chosen for this study is the popular videos on college employment topics in BiliBili. The researcher used Python to filter and crawl the relevant comments while keeping only the content field of the comments under the condition of fully guaranteeing the user's privacy.

In this study, the sentiment of the comment sentences was categorized into three categories, i. e. , predicting whether the college students' attitudes towards employment were positive, neutral, or negative. The corresponding dataset labels were negative '0', neutral '1', and positive '2'. Data preprocessing is an important step in model input, focusing on deleting invalid characters, deactivated words and emoticons, special symbols, numbers, blanks, and irrelevant data, de-emphasizing data with the same content, predicting the filling of missing text and further and standardizing semantics and sentence length. Finally, more than 8000 valid comments were processed, and the test set and training set were manually divided according to the ratio of 2:8. Table 1 displays the schematic table of the preprocessing data of college students' employment attitude comments, the first column is the label of the dataset, the second column is the

example of the original text, and the third column is the corresponding text after preprocessing.

Table 1. Schematic representation of comment data preprocessing

Dataset labels	text	The processed text
0	Or is it the high level of job classification and unequal pay for work that is the ultimate employment problem!!!?	A high degree of job stratification and unequal pay are at the root of the employment problem.
0	In fact, this is even more pathetic is that nowadays no one feels that this is very outrageous 666, this is the most outrageous, the involution to so, we already do not take this thing as something new, think of the year when Tsinghua University to sell pork is more sensational, nowadays the top universities to play the screws are very common, hey ~	The inward rolls have gotten to such a point that it's numbing. It's now commonplace for students at top universities to work part-time.
0	I'm afraid not just any blue collar job is better than this, right (>^ <;) a little bit of skill are mixed better than this, graduate students as a dormitory manager, how long it takes to earn back the cost of the investment in education? And the lost money is still a small matter, young time is precious!	Getting a blue-collar job is better, graduate school as a boarding house, the cost of the educational investment is a minor one, and the younger years are more valuable.

2.2 Bert model

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained deep learning model based on the Transformer architecture, which learns generalized language representations through bi-directional context modeling and large-scale unsupervised pretraining [8,9]. Many tasks in the field of natural language processing require a deep understanding of textual semantics and context, such as text categorization, question and answer systems, and named entity recognition [10]. Traditional word vector-based models suffer from problems such as unidirectional context restriction and inability to handle complex dependencies in sentences [11]. To resolve these issues, Transformer-based pretraining is suggested by the BERT concept, which achieves bidirectional context modeling and better learning of language representations. Transformer is designed solely through the self-attention mechanism (self-attention) and Feed Forward Neural Network (FFNN), and does not It achieves better results simply through self-attention and Feed Forward Neural Network, without the use of sequence-aligned recurrent architectures [12]. Its advantages are twofold: it abandons the network structure model of RNN and can parallelize the computation well. Its attention mechanism can help the current word to obtain better contextual information [13]. The BERT model has improved significantly in various natural language processing tasks, including text categorization, named entity recognition, question-answer systems, etc [14].

2.3 Bi-LSTM model

As an improved system of RNN, Bi-LSTM (Bidirectional Long Short-Term Memory) is a neural network model based on the structure of Long Short-Term Memory (LSTM), which is able to efficiently capture long-distance dependencies in sentencesin problems involving natural language processing [6]. The Bi-LSTM model combines a forward propagation LSTM (left-to-right) and backward-propagation LSTM (right-to-left), which can take into account the contextual information of each word in the sentence at the same time to understand the textual context [6,13] better. Each LSTM unit contains input gates, forgetting gates, output gates, and memory units that can efficiently process sequential data and learn long-term dependencies [6]. Compared with unidirectional LSTM, Bi-LSTM is able to consider both pre- and post-textual information and better capture dependencies in sequential data. It can provide more comprehensive contextual information and improve model performance when dealing with natural language processing tasks. It usually achieves better results in tasks such as text classification and sequence labeling [14].

2.4 The overall architecture of the Bert-BiLSTM model

This paper uses a bidirectional encoder combined with a bidirectional long and short-term memory network (Bert-LSTM) model for the predictive construction of sentiment analysis of college students' employment attitudes. The model's overall architecture is depicted in Figure 1, the model adopts a four-layer structure, with the first layer being the preprocessed data input stratum, and the second stratum being the Bert downstream module for segmenting the data fields and converting the word vectors, i. e. , fine-tuning the pre-trained Bert model. The output word vectors are then passed through the Bi-LSTM upstream module in the third stratum for further feature extraction and finally transferred to the data output stratum in the fourth stratum. The output stratum uses the softmax function to output the text classification training results of college students' employment attitude comments, and finally completes the sentiment analysis task.

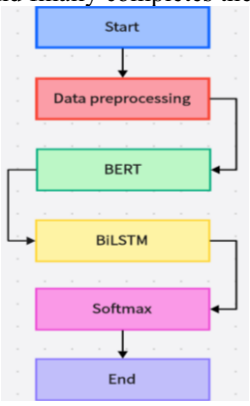

Fig. 1. Bert-BiLSTM model training structure

2.5 Assessment of indicators

In deep learning, the accuracy and loss function values during model training are two very important evaluation metrics [3], and monitoring their changes can help us understand the model's performance and the training progress.

Accuracy is a metric used to assess the performance of a classification model, indicating the ratio of the number of samples correctly predicted by the model to the total number of samples. The accuracy rate can intuitively tell us how accurately the model predicts the whole, and it is one of the important metrics for assessing the performance of classification models [6]. Usually, a higher accuracy rate tending to 1 indicates a better model performance. However, in some cases, the accuracy rate may have some limitations, such as in the case of category imbalance, it's possible that the accuracy rate underrepresents the model's capabilities.

The loss function value is a metric used to measure the difference between a model's predicted and true values. In this study, the cross-entropy loss function is used to optimize the model training process. In order to improve the model's ability to fit the data, the model parameters are improved by minimizing the loss function, which calculates the model's prediction error during training. The smaller the loss value tends to 0 indicates that the model fits the data better. Still, to fully evaluate the model's performance, it must be paired with accuracy and additional metrics [15].

3 Analysis

3.1 Big data analytics

After data preprocessing of the college students' employment attitude comment text, it is also necessary to examine the content of the data and excerpt the keywords in it to assess the factors that lead to the college students' employment problems by using high-frequency words. In this project, the TF-IDF algorithm is used to extract keywords. The extracted elements include keywords, word frequencies and weight values, totaling more than 300 keywords. The extracted keywords are modified to retain only the nouns related to college student employment factors and delete the keywords with word frequency less than 10, leaving 91 keywords in the end.

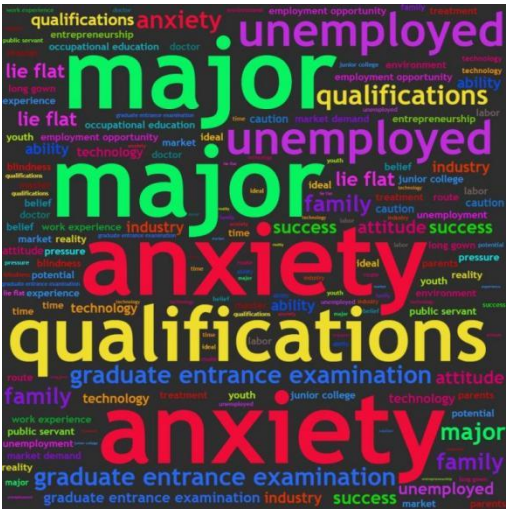

Fig. 2. Keyword word cloud map

Figure 2 shows the word cloud map produced after 91 keywords data visualization, which has a stronger visual effect compared to other statistical maps, and is more capable of highlighting the theme and attracting readers' attention. It can be inferred from the high-frequency words that the main factors leading to the employment problems of college students include anxiety, professional prospects, academic threshold, unemployment pressure, and family internal conflict.

3.2 Analysis of the results of model preprocessing

To assess the BERT-BiLSTM University Employment Attitude Sentiment Analysis Model in greater detail, three sets of trials using the Bert, LSTM, and Bert-BiLSTM models were put up for this investigation, each set of experiments was trained and adjusted 10 times, and the experimental model with the highest accuracy was saved at last. The values of ACCURACY, LOSS were obtained on the test set respectively. According to Figure 3, the Bert-BiLSTM method achieved 0. 7522 in terms of the highest accuracy and 0. 5891 in terms of loss. in contrast, using the Bert model (0. 7269, 0. 6138) and the LSTM model (0. 6252, 0. 6836) independently was slightly inferior regarding the final results. This indicates that the capacity of Bert-BiLSTM model to learn textual contextual features is more prominent in the task of textual sentiment analysis of college students' employment attitudes, probably because on the one hand, It incorporates the Bert dynamically encoded-word vector model, which has the ability to distinguish between synonyms according to their context, and also has the excellent ability of the BiLSTM model to learn textual contextual features, thus The word vectors obtained from the BERT-BiLSTM pretraining model contain more complete contextual information, and thus for training, the BERT-BiLSTM model performs well.

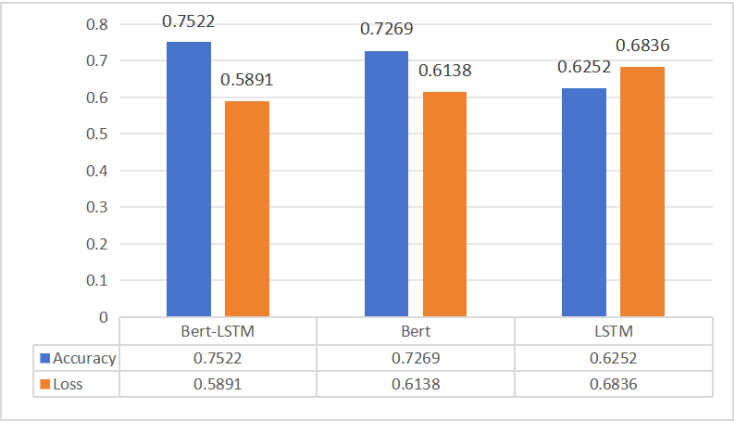

Fig. 3. Contrast between each model's experimental outcomes

4 Discussion

In this research, the use of the language model Bert-BiLSTM model to realize the college students' employment problems and sentiment analysis, in general, got good results, but there are still worthy of improvement, such as more detailed preprocessing operations on the comment data, and more kinds of multiclassification divisions for sentiment classification. The dataset has too few samples, and there is also a possibility that a lot of ambiguous comments were introduced leading to errors in model training. In addition, there are too few model evaluation metrics, and future work pre-adds recall and model F1 values to analyze the strengths and weaknesses of the model performance by combining the sum of the accuracies. Finally, an attention layer can be added after the BiLSTM layer to focus on the important parts of the context by assigning different weights, and then the softmax function outputs the training results to form a five-layer structured model.

Clip model is a multimodal model based on vision and language, which is able to process image and text data simultaneously [16]. The next step of the research will try to apply the Clip model to the sentiment analysis task, especially in image and text matching. Collecting datasets containing both image and text comments, the Clip model is trained to identify the sentiment associations between images and text, thus analyzing the sentiment content contained therein more accurately. This multimodal approach to sentiment analysis can help us gain a more comprehensive understanding of what people are posting on social media, leading to a better understanding of their emotional states and attitudes. This is valuable for both companies and research organizations as it helps them better understand the needs and feedback of their research subjects so that they can make better decisions and improvements. Therefore, future work will experiment more with this multimodal sentiment analysis task toenhance the analysis's efficacy and accuracy.

5 Conclusion

Aiming at the current employment status and mental health problems of college students, this study suggests a BERT-BiLSTM-based approach for text sentiment analysis for college students' employment comments and uses big data mining to extract keywords to predict the factors affecting college students' employment sentiment tendency, and proves the feasibility of the study by designing experiments.

The outcomes demonstrate the deep learning foundation of the BERT pre-trained language model and its ability to perform downstream natural language processing tasks. In the meanwhile, BiLSTM is able to effectively extract text features and obtain contextual information in its entirety. The accuracy of the final sentiment classification feature vectors can be successfully increased by combining the BERT model with BiLSTM. The Bert-BiLSTM technique obtains an accuracy of 0. 7522 and a loss rate of 0. 5891. This combination method not only improves the performance of the sentiment classification task, but also provides a new idea and methodology for other natural language processing tasks in the research by providing new ideas and methods.

By combining BERT and BiLSTM, we can make full use of BERT's advantages in understanding language and at the same time be able to utilize BiLSTM's advantages in extracting text features to achieve better results in sentiment classification tasks. This approach improves the accuracy of sentiment classification and provides new inspirations and methods for other natural language processing tasks such as named entity recognition and text generation. Therefore, the combination of BERT and BiLSTM has broad applications in natural language processing and is expected to bring new breakthroughs in the research and development of this field.

The Clip model is a multimodal model that combines image and text data for sentiment analysis, which can help to understand sentiment information in social media content more accurately and provide valuable decision support to companies and research organizations.

References

1. Si, H. , Wei, X. : Sentiment Analysis of Social Network Comment Text Based on LSTM and Bert. Journal of Circuits, Systems and Computers, 32(17), 2350292 (2023).
2. Guo, Z. : The psychological status of employment stress and suggestions for coping with it among Chinese college students in the post epidemic situation. International Education, 3(2), 19-22 (2021).
3. Li, X. , Lei, Y. , Ji, S. : BERT-and BiLSTM-based sentiment analysis of online Chinese buzzwords. Future Internet, 14(11), 332 (2022).
4. Murfi, H. , Gowandi, T. , Ardaneswari, G. , et al. : BERT-based combination of convolutional and recurrent neural network for Indonesian sentiment analysis. Applied Soft Computing, 151, 111112 (2024).
5. Srivastava, S. , Sarkar, M. K. , Chakraborty, C. : Machine Learning Approaches for COVID-19 Sentiment Analysis: Unveiling the Power of BERT. In: 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, pp. 0092-0097 (2024).

6. Wang, P. , Xiong, X. : A BERT-ABiLSTM Hybrid Model-Based Sentiment Analysis Method for Book Review. Journal of Circuits, Systems and Computers, 33(02), 2450039 (2024).
7. Fadhlurrahman, J. A. M. , Herawati, N. A. , Aulya, H. R. W. , et al. : Sentiment Analysis of Game Reviews on STEAM using BERT, BiLSTM, and CRF. In: 2023 International Conference on Electrical Engineering and Informatics (ICEEI). IEEE, pp. 1-6 (2023).
8. Devlin, J. , Chang, M. W. , Lee, K. , et al. : Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810. 04805 (2018).
9. Vaswani, A. , Shazeer, N. , Parmar, N. , et al. Attention is all you need. Advances in neural information processing systems, 30 (2017).
10. Shaik Vadla, M. K. , Suresh, M. A. , Viswanathan, V. K. : Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT. Algorithms, 17(2), 59 (2024).
11. Wankhade, M. , Annavarapu, C. S. R. , Abraham, A. : MAPA BiLSTM-BERT: multi-aspects position aware attention for aspect level sentiment analysis. The Journal of Supercomputing, 79(10), 11452-11477 (2023).
12. Liang, H. , Tang, B. , Cao, S. : Sentiment analysis of comment texts on converged media platforms based on BERT model. In: 2021 International Conference on Culture-oriented Science & Technology (ICCST). IEEE, pp. 552-555 (2021).
13. Zhou, X. : Sentiment Analysis of the Consumer Review Text Based on BERT-BiLSTM in a Social Media Environment. International Journal of Information Technologies and Systems Approach (IJITSA), 16(2), 1-16 (2022).
14. Tang, H. , Zhang, R. : The Research on Aspect-Level Sentiment Analysis for Online Reviews Based on BERT-BiLSTM Model. CAIBDA (2023).
15. Dickson, M. C. , Bosman, A. S. , Malan, K. M. : Hybridised Loss Functions for Improved Neural Network Generalisation. In: Ngatched, T. M. N. , Woungang, I. (eds) Pan-African Artificial Intelligence and Smart Systems. PAAISS 2021. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 405. Springer, Cham (2022).
16. Radford, A. , Kim, J. W. , Hallacy, C. , et al. : Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp. 8748-8763 (2021).