A CNN-BiLSTM algorithm for Weibo emotion classification with attention mechanism

Xinyue Feng[1], Niwat Angkawisittpan[2], Xiaoqing Yang[3]
[1, 2]Research Unit for Electrical and Computer Engineering, Mahasarakham University, Maha Sarakham, Thailand
[3]Faculty of Computer Engineering, Shanxi Vocational University of Engineering Science and Technology, Taiyuan, China
[2]Corresponding author
E-mail: [1]*xinyue6570147@163.com*, [2]*niwat.a@msu.ac.th*, [3]*yangxqmty0907@163.com*

Received 12 March 2024; accepted 9 April 2024; published online 29 April 2024
DOI https://doi.org/10.21595/mme.2024.24076

Abstract. Weibo short text information contains a large amount of network language, emoticons, etc., and due to the long-time span of the content, the emotions of the posts posted by people often change due to time or the occurrence of certain special events. Therefore, traditional sentiment analysis methods are not suitable for this task. This article proposes a CNNs-Bi LSTM sentiment analysis method that integrates attention mechanism. It combines convolutional neural networks and bidirectional short-term memory networks to obtain keyword information in text through attention mechanism, efficiently and accurately realizing data temporal and semantic information mining. Through experimental verification using Weibo public opinion data, the results show that this method achieves higher accuracy compared to other benchmark models and can fully utilize multidimensional matrices to capture rich text features, with certain advantages.

Keywords: emotional analysis, attention mechanism, CNN network, Bi LSTM neural network.

1. Introduction

With the rapid development of mobile internet and the rise of online social media, social media such as Weibo is one of the effective channels to understand the public's thoughts. Social media, as a discussion platform for people to communicate and express their opinions, has accumulated a massive amount of digital information with user traces, providing a strong data foundation for text sentiment classification and topic extraction. In recent years, mobile communication terminals such as Weibo, Tiktok and WeChat have become important ways for people to communicate and express their feelings on the Internet [1]. Users have become producers and consumers of information, and can express and share opinions, interests, and feelings on different topics in a short time. Weibo data has been widely used.

Emotional classification refers to the analysis and judgment of the emotional polarity of a text, applied in areas such as opinion mining, emotion recognition, and public opinion analysis. Attention mechanism is widely used in the field of natural language processing and has high accuracy in many classification tasks. Recurrent neural networks and attention mechanisms are both end-to-end structures with the ability to combine context. Recurrent neural networks learn along the time direction and can remember sequential information. But when the sentence is too long, traditional methods cannot learn distant word information. The self-attention mechanism can serve as an encoder, enabling each word to obtain global contextual information. For sentiment classification algorithms, it is important to embed emotional information into the network to enrich text representation. Although self-attention mechanism can provide global attention, it also introduces noisy words. The more complex context is enough to confuse the auditory and visual aspects of the self-attention mechanism, and there is still much room for improvement in the self-attention mechanism, as every emotional word is important.

This article will be based on sentiment analysis to study the emotional situation of online public opinion, with blog posts related to information published on Weibo platforms as the research

object for sentiment polarity analysis. A CNN BiLSTM sentiment polarity analysis method integrating attention mechanism is proposed, which combines convolutional neural network and bidirectional long short-term memory network to obtain keyword information in text through attention mechanism, efficiently and accurately realizing data temporal and semantic information mining.

2. Text emotion classification

2.1. The types of different emotions

On various social platforms, there are many emotional texts involved. We will analyze the authenticity of the product based on its comments and consider whether it meets our own needs; We will write blogs, brief books, and even establish our own websites to share what we have learned and felt; We will express our opinions and opinions on social focal issues through our social circle and Weibo. It contains countless emotional messages. In this emerging era, people's ways of expressing emotions have undergone tremendous changes not only in channels, but also in language habits and structures.

Emotional classification, proposed by Bo Pang et al., is an important task in the field of natural language processing. The research on sentiment analysis was first established on the basis of document level analysis. Pang and Snyder extended the bipolar classification of emotions to multi-level classification. Due to the continuous research, exploration, and innovation of many experts and scholars in this field, emotion classification related algorithms have become a hot research topic both domestically and internationally. Emotion classification algorithms mainly include sentiment classification algorithms based on dictionaries and rules, sentiment classification algorithms based on machine learning, and sentiment classification algorithms based on neural networks.

Emotional classification or opinion mining is a fundamental task in natural language processing (NLP), used to determine the polarity of text emotions. Binary classification, such as positive and negative emotions, can include fine-grained classification of emotions more generally. Natural language, like numbers, spreads and records information. The process of information dissemination and reception corresponds exactly to the encoding and decoding of information, and text encoding and decoding are important links in natural language processing. On various emerging social platforms, people express their opinions in a variety of sentences, many of which cannot be strictly analyzed according to grammar and syntax rules, and sometimes even contradictory analysis may occur; The same word may have vastly different meanings in different contexts. In different eras, people have different language expressions and preferences, and analyzing the meaning expressed by words based on contextual context is more flexible and feasible.

The most basic analysis unit for text sentiment classification is sentence, while the smallest unit of a sentence is word [2]. Because there are language processing regions in the brain, it is easy for humans to make judgments about emotions. However, for computers to make judgments on the emotions of text, algorithms and rules must be designed for text sentiment classification, and various constraints must be proposed [3]. Text is different from images in that behind images are pixel values, while text is a product of human history and modern culture [4]. Therefore, it is necessary to symbolize sentences first.

For a sentence M with a length of L, to determine whether the sentence belongs to emotion category a, it is necessary to calculate the probability that sentence M belongs to emotion category a. Abstract the classification of text emotions into the following formula:

$$P(a \forall M). \tag{1}$$

Among them, there are two important tasks: representing sentence M and, calculating the

probability P. The representation method of sentences directly affects the solution of classification probability, which relies on neural network design.

2.2. Discrete representation of words

To represent a sentence, the first step is to preprocess the sentence, including removing punctuation marks, segmenting the sentence into words, segmenting words based on spaces in English, and segmenting words in Chinese. After preprocessing, sentences are represented as a list of words. The initial representation method for words was single hot encoding, which required creating a word set for all words in the sentence, constructing a dictionary T for key value pairs, and recording $|T|$ as the number of words in the dictionary.

After unique encoding, each word is represented as a $|T|$ dimension vector, which is only 1 at the corresponding position in the word index, and all other positions are zero. Record the unique heat code as f, and the dimensional change process of sentence M is as follows:

$$R^L \xrightarrow{f} R^{L\times|T|}. \tag{2}$$

In actual text classification, due to the large size of the corpus and dictionary, it can be seen from Eq. (2) that the unique hot coding will make the sentence represented as a high-dimensional sparse matrix. Moreover, unique encoding cannot distinguish the importance of different words in a sentence, which makes it impossible to distinguish between high-frequency words and low-frequency words.

Using sentences as the basic unit code, a dictionary is established for sentences. The sentences are represented as $|T|$ dimensional vectors, and the frequency of corresponding words appearing in the dictionary is recorded at each position. The dimensional change process of sentence M is as follows:

$$R^L \xrightarrow{f} R^{|T|}. \tag{3}$$

2.3. Distributed table of words

The sentence representations constructed by DHC, BOW, and TF-IDF are sparse and high-dimensional. In the field of natural language processing, traditional machine learning algorithms are based on these representation methods. After the proposal of word2vec, sentence level sentiment classification algorithms have shifted from encoding with sentences as the basic unit to encoding with word embeddings as the basic unit. Word2vec is based on the distribution hypothesis: words with similar contexts have similar semantics.

The byproduct of the word2vec task is $W_{|T|\times H} \in R^{|T|\times H}$, which is the word vector of the dictionary, and according to the index of the dictionary, the word embeddings corresponding to the words can be found. For the unique hot code θ_i, its word embedding is:

$$w_i = \theta_i^T \cdot W_{|T|\times H}. \tag{4}$$

The entire process embeds words from high-dimensional sparse representation to low-dimensional dense representation. In a single hot coding space, two vectors are linearly independent and orthogonal; In word2vec embedding space, there is a concept of distance between two vectors. This process maps words from high-dimensional, discrete, and sparse representations to low-dimensional, continuous, and dense representations.

3. The overview of CNN-BiLSTM algorithm

3.1. CNN Network

As a special type of feedforward neural network, CNN has been widely used in the field of natural language processing by scholars in recent years [5]. Its basic structure is divided into three parts: input layer, convolutional layer and pooling layer, and fully connected layer. Convolutional layer feature extraction first represents the text in the form of a word vector matrix, and then scans through convolutional check matrices of different sizes [6]. During the scanning process, the parameter values of the filter composed of convolutional kernels are fixed and unchanged. After filtering, a new feature map is mapped, and all elements on the feature map come from the filter with consistent parameters. The CNN network architecture is shown in Fig. 1.

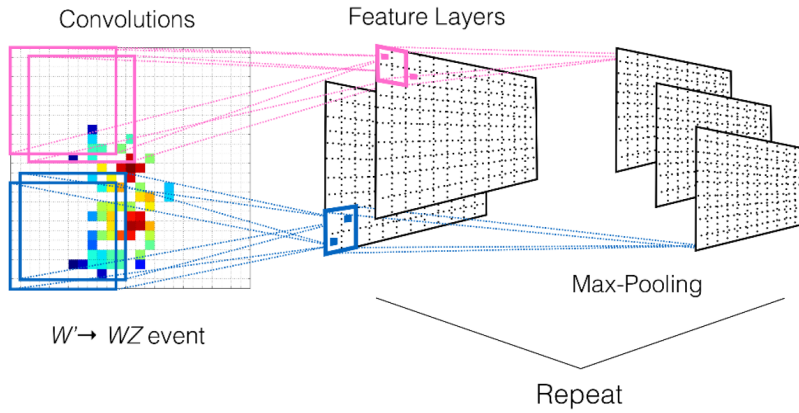

Fig. 1. CNN network architecture

3.2. Bi LSTM neural network

LSTM is a special type of recurrent neural network (RNN) composed of cell units and three gates. Cell units are the core computing power that records the current computing state, while forgetting gates, input gates, and output gates regulate the information flow in and out of the storage unit [7]. forgetting gates clear useless information from the storage unit, input gates select input information from the current storage unit, and output gates determine the final output of the information. When conducting sentiment analysis on Weibo texts, it is often necessary to consider the impact of contextual semantic features on the overall emotional state of the text. However, the general LSTM model only focuses on unidirectional semantic relationships in the text, which ignores the impact of semantic features on the overall state. The Bi LSTM model is composed of LSTM networks in both positive and negative directions, which fully capture contextual information based on two different orders and can mine more comprehensive text semantics. The Bi LSTM network architecture is shown in Fig. 2.

3.3. The joint mechanism of CNN-BiLSTM algorithm

This article constructs a Weibo sentiment classification model based on convolutional neural network CNN and bidirectional short-term memory network Bi LSTM. On the basis of Bi LSTM and CNN, an attention mechanism was introduced, and a multi-channel model combining convolutional networks and bidirectional long and short term networks was proposed for sentiment polarity analysis of Weibo texts [8]. CNN-BiLSTM Joint Mechanism is shown in Fig. 3.

ISSN PRINT 2351-5279, ISSN ONLINE 2424-4627

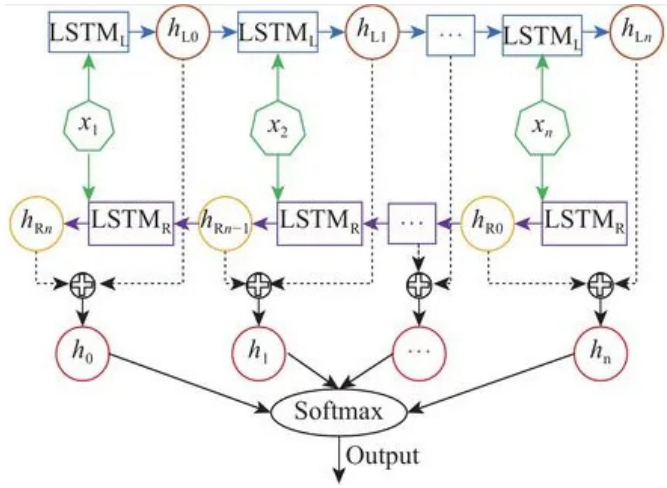

Fig. 2. Bi LSTM network architecture

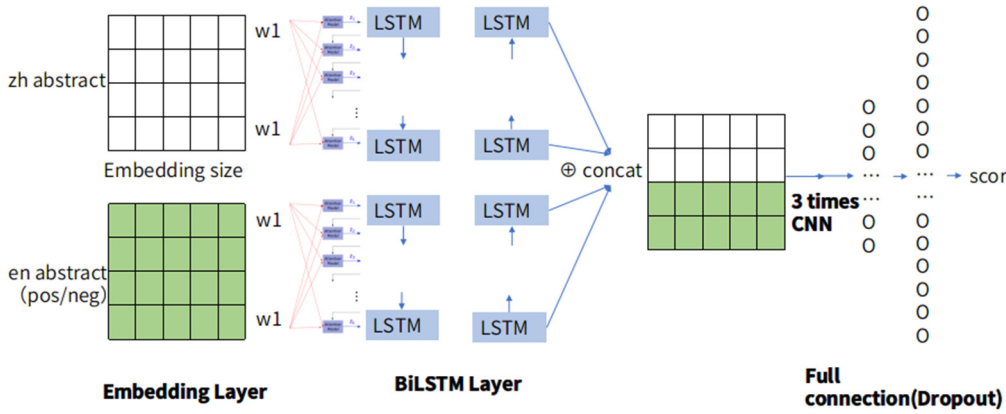

Fig. 3. CNN BiLSTM joint mechanism

4. Attention mechanism

Emotional classification, as an important task in the field of NLP, aims to determine the emotional polarity of text, including positive and negative emotions [9]. After reading a large number of references, this article found that improving the accuracy of emotion classification algorithms and determining whether the classification algorithm has good performance can be achieved from three aspects: whether the algorithm has the ability to combine context, and how much context can be learned; Does the algorithm have attention and can it focus on learning; Does the algorithm have memory and store sequential information.

The attention mechanism originates from human research on vision [10]. Due to the bottleneck in information processing, humans selectively focus on some information while ignoring others, a mechanism known as attention mechanism. When used for sentence level text sentiment classification, the attention mechanism is usually placed at the decoding end, which weights and sums the input sequence over the entire time range.

The existing sentiment classification algorithms based on neural networks mainly include sentiment classification algorithms based on recurrent neural networks, sentiment classification algorithms based on convolutional neural networks, and sentiment classification algorithms based on attention mechanisms. The self-attention mechanism belongs to a special attention mechanism, which can serve as both an encoder and a decoder. When the sentence length is long, even LSTM

may experience the phenomenon of gradient vanishing and information dissemination being hindered. In this case, using self-attention mechanism as the encoder can achieve cross distance learning, and the classification effect is better than RNN. The attention mechanism can be expressed as the following formula:

$$A(Q,S) = \sum_{i=1}^{L} S(Q,K_i) \times V_i, \tag{5}$$

where, S refers to the input information, including sentence M and label y, Q as the query vector, K and V appear in pairs, $S(Q,K)$ refers to the attention weight value of V. $A(Q,S)$ is the attention force vector weighted and summed by the attention mechanism. This formula represents the way the attention mechanism decodes sentence M. The attention mechanism architecture is shown in Fig. 4.

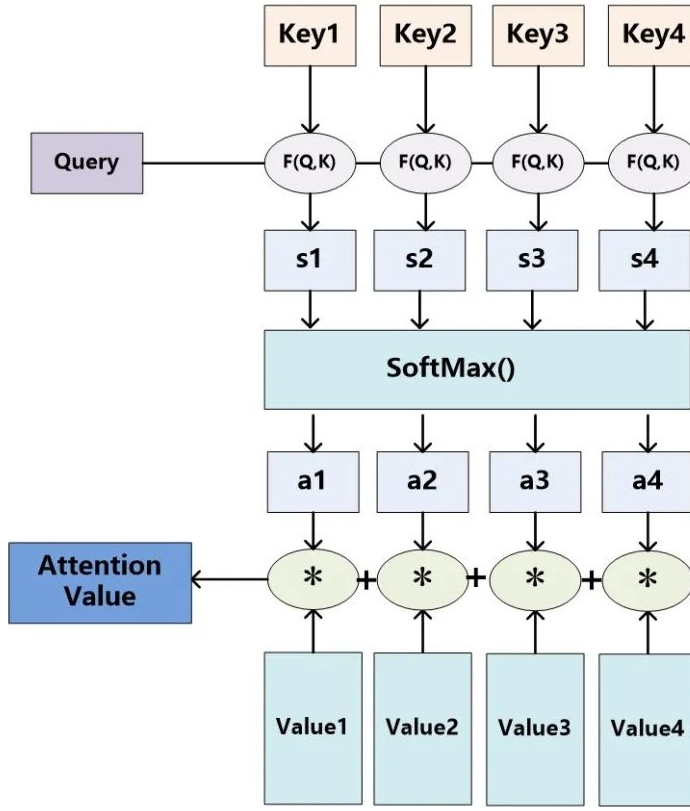

Fig. 4. The attention mechanism architecture

5. Data sources and processing

5.1. Construction of Weibo dataset

The original Weibo public opinion corpus data used in this article comes from the social media dataset (Weibo COV V2) provided in reference [11]. This public opinion data contains large-scale epidemic related Weibo, mainly targeting active Weibo users as crawling objects. Active users tend to use Weibo as a tool for online communication and discussion. Therefore, the obtained Weibo information is more real-time and emotional compared to the posts of inactive Weibo users.

This dataset includes Weibo content data and user information data. In order to meet the experimental requirements, these two pieces of data were first merged based on user name information. Then, Weibo data related to lottery, certain entertainment celebrity super talk, and incomplete Weibo information with the same content were deleted from the dataset. Based on this dataset, a new Weibo sentiment polarity dataset was constructed. Partial Weibo text data was selected from the dataset and sentiment polarity was annotated manually. The labeling team consisted of 7 people. To ensure the quality of data annotation, cross and repeated labeling was used, and text with inconsistent labeling in two rounds was deleted. A total of 63975 pieces of data were organized. In order to avoid unsatisfactory experimental results due to imbalanced sample sizes, a total of 31987 text data with positive sentiment expressions and 31988 data with negative sentiment expressions were collected. The labeled corpus dataset achieved a uniform distribution of positive and negative sentiment polarity in the samples.

5.2. Dataset preprocessing

Data preprocessing is an indispensable preliminary work for text sentiment polarity analysis, which mainly includes data filtering, text segmentation, and removing stop words. In the constructed labeled corpus dataset, there are many non-textual data, labels, and special characters. Considering optimization efficiency, storage space is saved, and word vector representation accuracy is improved. Therefore, it is necessary to remove these useless information and filter it mainly through regular expression matching.

5.3. Training process

In the model proposed in this article, the Input input layer is the first layer, which reads the cleaned dataset. The Weibo text data processed by word segmentation is passed into the second layer Embedding layer. During the vectorization process, the Word2Vec model is used, and the Embedding layer embeds the corresponding vector based on the incoming words, which is then transmitted into the multi-channel model. The third layer of the model is a multi-channel feature extraction layer, which enters the BiLSTM model channel and can simultaneously capture semantic information in both positive and negative directions. Multiple CNN models are designed to extract local features of sentences at different scales, and using convolutional kernels of different sizes can better extract features from different dimensions to achieve semantic information extraction.

The word vector dimension trained using the Word2Vec model is 350, and the filters for the three convolutional channels are 1×350, 2×350, and 3×350, respectively. The pooling layer adopts Max Pooling operation to eliminate weak features. The BiLSTM model adopts L_2 regularization processing to control the model complexity and avoid overfitting as much as possible. Introducing attention mechanism in the fourth layer to extract more important feature information from each channel; The splicing layer summarizes the features output by all channels to obtain richer feature information; The summarized features are passed into the fully connected layer, and a random deactivation mechanism is added to reduce redundancy and improve the model's generalization ability. Finally, the sentiment polarity category of the text is determined based on the Softmax classifier in the output layer.

6. Experimental results and analysis

6.1. The benchmark model

To evaluate the effectiveness of the algorithm proposed in this article, some benchmark models will be selected for comparison in the same experimental environment. The benchmark model selected in this article is as follows.

1) Logistic Regression Model (LR): It mainly solves binary classification problems and is a classic machine learning classification model.

2) Support Vector Machine (SVM): Use TF-IDF to represent text words, and use SVM algorithm to determine emotional polarity.

3) LSTM and Bi LSTM models: Use LSTM to capture unidirectional semantics or Bi LSTM to extract sentence context features, and use softmax classifiers for sentiment analysis.

4) CNN-Att and Bi LSTM-Att models: CNN extracts local feature information from text, or uses Bi LSTM to obtain contextual semantic information, and then calculates the attention weight of the model's output features using Attention. Finally, it enters the fully connected layer and outputs with the classifier.

5) CNN-Bi LSTM model: The segmented sentences are vectorized through word embedding, and then input into the CNN. The extracted local features are further input into Bi LSTM, and the final result is obtained through a classifier.

Table 1. Experimental comparison results

Algorithm	Accuracy	Recall	F-value
LR	76.53	75.92	76.06
SVM	80.82	80.47	80.65
LSTM	82.89	82.56	82.73
CNN-Att	83.55	83.21	83.38
Bi LSTM	84.67	84.18	84.51
Bi LSTM-Att	86.78	86.70	86.72
CNN-Bi LSTM-Attention	88.25	88.15	88.20

The experimental results of the proposed algorithm and several other neural network model methods on Weibo sentiment dataset are shown in Table 1. From the experimental results, it can be seen that traditional machine learning models such as LR and SVM have poor experimental results. LR models are relatively simple, and their models themselves cannot perform feature filtering, resulting in poor data fitting performance; Although the classification performance of the SVM model is better than that of the LR model, it only performs a single weighted average for the word vector information in the sentence. Therefore, traditional machine learning methods are not suitable for the current needs. In the Weibo sentiment polarity analysis experiment, it is evident that the test results of the deep learning model are superior to traditional machine learning models. The experimental results of the Bi LSTM model have higher accuracy compared to the LSTM model, indicating that bidirectional units have significant advantages in serialized data processing. After incorporating attention mechanism into the CNN network, its classification performance is better than that of the LSTM model. The CNN network learns the word vector representation of text through convolutional window sliding, which can effectively extract the overall features of sentences. Moreover, the attention mechanism can focus on semantically related vocabulary in declarative memory, better capturing emotional words and assigning them higher weight values.

It can be seen that the CNN Bi LSTM algorithm has more advantages over traditional machine learning algorithms in grasping Weibo text features. The text features extracted by this algorithm are more comprehensive than the information captured by a single CNN or Bi LSTM network. The introduction of attention mechanism makes the model pay more attention to the emotional part of the text, and the allocation of weight information makes the text representation more rich in emotional feature information, This advantage is also reflected in the comparison between Bi LSTM and Bi LSTM-Att models, where the text representation is consistent, but the model has been effectively improved under the attention mechanism.

6.2. The optimality of the architecture

To further explore the advantages of the design of each part of the model proposed in this

article, ablation experiments were conducted. The proposed model was decomposed and BiLSTM network, multi-channel CNN network, and Attention mechanism were removed to verify the effectiveness of the removed part. To control the influence of parameters on the experiment, the hyperparameters were set the same for each group of experiments. The experimental results are shown in Table 2 (w/o represents without).

Table 2. Ablation test results

Model	Accuracy (%)	Recall (%)	F-value (%)
w/o word2vec	87.61	87.60	87.59
w/o CNNs	86.54	86.54	86.54
w/o BiLSTM	88.97	88.78	88.78
w/o Attention	88.96	88.85	88.84
CNN-Bi LSTM-Attention	89.14	89.1	89.09

From the experimental results, it can be seen that each part of the model's structure significantly improves the performance of the model. Under the same conditions, the word embedding layer trained with Word2Vec has an improved F-value compared to the method of randomly initializing word vectors, indicating that Word2Vec has a good effect on the training of word vectors. When the attention mechanism is not added, the model can achieve good results by only utilizing the feature extraction advantages of CNNs and BiLSTM. However, the attention mechanism can amplify the differences in text features in the final results, achieving further mining of text features. The embedding of CNNs and BiLSTM is beneficial for improving model performance. The introduction of BiLSTM model can increase the richness of semantic information, effectively capturing long-distance dependencies and contextual information in text; After removing the CNN network with multiple channels, local features of multi-scale text were ignored, resulting in a significant decrease in F-value. This also indicates that increasing the structure of CNNs has a positive effect on sentiment polarity analysis models.

Therefore, the analysis results of the MCCB model in this article are better than other models. On the one hand, it fully utilizes the advantages of CNN and BiLSTM models, which can extract multi-scale local feature information in the text while parsing contextual semantic information; On the other hand, by constructing an attention mechanism to obtain more implicit information, the impact of non-key vocabulary on the model is reduced. Multiple features are fused in a multi-channel form to achieve sentiment enhancement of word vectors, making semantic features more correlated with sentiment polarity labels, thus achieving the best sentiment classification performance.

7. Conclusions

A Weibo sentiment analysis method is proposed based on CNN, Bi LSTM, and attention mechanism. This method focuses on multi-scale text input features and achieves the importance allocation of different dimension text features through attention mechanism. The experimental results demonstrate that the multi-channel CNNs Bi LSTM sentiment classification algorithm based on attention mechanism performs better than other benchmark models on Weibo datasets and can effectively identify and analyze the emotions of netizens' Weibo posts in online public opinion.

There are still some shortcomings in the data and methods of this study. The research data is limited to Weibo short text, and the classification effect of the model on long text data is not discussed. At the same time, neutral text is not considered to be added.

The next stage of research work is to deploy the model to long text classification and explore the effect of three classification sentiment analysis.

Acknowledgements

This research project was financially supported by Mahasarakham University.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Xinyue Feng contributed conceptualization, methodology, writing-original draft preparation, writing – review and editing. Niwat Angkawisittpan contributed formal analysis, data curation, investigation, resources. Xiaoqing Yang contributed software, supervision, validation, visualization.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] H. Ikeda, D. Malgazhdar, T. Shionoiri, B. Bino Sinaice, T. Adachi, and Y. Kawamura, "Development of a communication system for underground mining informatization leading to smart mining: a comparison of Wi-Fi Ad Hoc and Wi-Fi direct," *International Journal of the Society of Materials Engineering for Resources*, Vol. 25, No. 2, pp. 218–223, Oct. 2022, https://doi.org/10.5188/ijsmer.25.218

[2] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, No. 9, pp. 429–439, Jan. 2020, https://doi.org/10.1109/taslp.2019.2957872

[3] K. Anuratha and M. Parvathy, "Twitter sentiment analysis using social-spider lex feature-based syntactic-senti rule recurrent neural network classification," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 30, No. Supp01, pp. 45–66, Jun. 2022, https://doi.org/10.1142/s0218488522400037

[4] R. Duwairi, M. N. Al-Refai, and N. Khasawneh, "Feature reduction techniques for Arabic text categorization," *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 11, pp. 2347–2352, Jul. 2009, https://doi.org/10.1002/asi.21173

[5] M. C. Q. Farias, P. H. de Castro Oliveira, G. Dos Santos Lopes, C. J. Miosso, and J. A. Lima, "The Influence of magnetic resonance imaging artifacts on CNN-based brain cancer detection algorithms," *Computational Mathematics and Modeling*, Vol. 33, No. 2, pp. 211–229, Jan. 2023, https://doi.org/10.1007/s10598-023-09567-4

[6] Y. V. Koteswararao and C. B. R. Rao, "Multichannel KHMF for speech separation with enthalpy based DOA and score based CNN (SCNN)," *Evolving Systems*, Vol. 14, No. 3, pp. 501–518, Nov. 2022, https://doi.org/10.1007/s12530-022-09473-x

[7] J. Shobana and M. Murali, "An improved self attention mechanism based on optimized BERT-BiLSTM model for accurate polarity prediction," *The Computer Journal*, Vol. 66, No. 5, pp. 1279–1294, May 2023, https://doi.org/10.1093/comjnl/bxac013

[8] K. Vo, S. Truong, K. Yamazaki, B. Raj, M.-T. Tran, and N. Le, "AOE-Net: entities interactions modeling with adaptive attention mechanism for temporal action proposals generation," *International Journal of Computer Vision*, Vol. 131, No. 1, pp. 302–323, Oct. 2022, https://doi.org/10.1007/s11263-022-01702-9

[9] S. Endo, T. Takahashi, and S. Satori, "Emotional pattern classification by image analysis of camera using machine learning," *Journal of The Color Science Association of Japan*, Vol. 41, No. 3, pp. 95–98, Jan. 2017, https://doi.org/10.15048/jcsaj.41.3__95

[10] H. Kour and M. K. Gupta, "AI assisted attention mechanism for hybrid neural model to assess online attitudes about COVID-19," *Neural Processing Letters*, Vol. 55, No. 3, pp. 2265–2304, Dec. 2022, https://doi.org/10.1007/s11063-022-11112-0

[11] H. Madhu, S. Satapara, S. Modha, T. Mandl, and P. Majumder, "Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments," *Expert Systems with Applications*, Vol. 215, No. 4, p. 119342, Apr. 2023, https://doi.org/10.1016/j.eswa.2022.119342

Xinyue Feng is now a doctoral student majoring in electrical and computer engineering at Mahasarakham University, Mahasarakham, Thailand. She works as a Lecturer at Foshan Polytechnic, Foshan city, Guangdong province, China. Her main research interests include big data crawler technology, machine learning algorithms.

Niwat Angkawisittpan is an Associate Professor works in Faculty of Engineering in Mahasarakham University, MahaSarakham, Thailand. He is the supervisor of doctoral students. His main research interest is electrical and computer engineering.

Xiaoqing Yang is now a doctoral student majoring in Electrical and Computer Engineering at Mahasarakham University, Mahasarakham, Thailand. She works as a Lecturer at Shanxi Vocational University of Engineering Science and Technology, Taiyuan city, Shanxi province, China. Her main research interests include big data technology, machine learning algorithms, encrypted traffic identification and anomaly detection technology.