International Journal of Information Technologies and Systems Approach
Volume 17 • Issue 1 • January-December 2024

Social Network Public Opinion Analysis Using BERT-BMA in Big Data Environment

Hanqing Sun
Henan University of Animal Husbandry and Economy, China

Zheng Liu
Henan University of Animal Husbandry and Economy, China

Weimin Lian
Henan University of Animal Husbandry and Economy, China

Guizhi Wang
Henan University of Animal Husbandry and Economy, China

ABSTRACT

The existing social network public opinion analysis methods have problems such as poor semantic expression quality and weak detection ability in short texts. Therefore, a social network public opinion analysis method based on BERT-BMA is proposed. To normalize the comment text, the rumor text is initially transferred to a word vector matrix using the BERT (Bidirectional Encoder Representations from Transformer) model. The BiLSTM-based network architecture is subsequently employed to acquire the trace features of data transmission. Ultimately, this study employs the multi-head attention mechanism to extract feature information that is more significant in the analysis of online public opinion by mining the dependency relationships between users, resulting in increasing ability to detect public opinion emergencies. The experimental outcomes indicate that the results on the Twitter data set and Weibo dataset are superior to other comparative models.

KEYWORDS

Social Network, Big Data, Public Opinion Analysis, Emergency Detection and Search, BiLSTM

INTRODUCTION

Social network platforms are increasingly utilized by users to access, disseminate, and share a variety of information (Dong & Lian, 2021). Public opinion emergencies refer to news that erupts over a period of time and exhibits certain laws of dissemination and extinction. With the rapid and widespread dissemination of social network news, public opinion events can be comprehensively and widely fermented. Detecting and searching for unexpected public opinion events in social network big data is a meaningful and valuable research area (Anstead & O'Loughlin, 2015; Murphy et al., 2014a; Murphy et al., 2014b; Rim et al., 2020). Due to the limitations of social network platforms on publishing messages, however, the description information of emergencies is constrained by a certain number of words in the text format. At the same time, due to the popularity and universality of message publishers, the information presentation forms of text and images are both arbitrary (Chakraborty & Sharma, 2019; Han et al., 2020; Kim & Kim, 2014).

Social network cross media big data has certain semantic sparsity and heterogeneity of modal feature space, which makes detection and search tasks of social network public opinion emergencies

DOI: 10.4018/IJITSA.352509

difficult (Chen et al., 2022; Younus et al., 2011). Compared with news reports and other information appearing in plain text, social network messages have rich multi-modal representations such as images and videos, and specific social network multi-attribute features (Du et al., 2020; Liang et al., 2021; McGregor, 2019; Tavoschi et al., 2020; Zhang et al., 2022). These diverse forms of presentation and multi-attribute information can serve as auxiliary features for understanding the meaning of social network messages, complementing the representation of short text and images. Therefore, it is necessary to fully exploit and utilize the advantages of multi-attribute information on social networks; integrate text, images, and multi-attribute features; and research more effective methods for detecting and searching social network cross media emergencies, to improve the semantic representation quality of media public opinion emergencies and the accuracy of emergency detection and search (He et al., 2022).

In social networks that contain rich, multi-source, and multi-granularity descriptions of public opinion incident related messages, public opinion incident detection and search technologies face many research challenges (Gao et al., 2023Zhang et al., 2022): 1) how to combine multiple attribute information of social networks to achieve effective semantic acquisition and expression of public opinion emergencies on social networks; 2) how to construct a deep feature learning method with sufficient semantic expansion to conduct deep semantic analysis of sudden public opinion events (Li et al., 2019; Maynard et al., 2012; Skoric et al., 2020); and 3) how to effectively mine and integrate the semantic features of cross media public opinion emergencies on social networks from multiple perspectives, so as to achieve accurate detection and search of public opinion emergencies. In terms of semantic acquisition and expression of public opinion emergencies on social networks, due to the rich social and semantic associations contained in multiple attribute features such as topic tags and hyperlink URLs on social networks, it has an important impact on the semantic acquisition and expression of public opinion emergencies, and it plays a crucial role in the detection and search tasks (Karami et al., 2018; Trivedi & Patel, 2022; Urban & Bulkow, 2013). As a result, the foundation for social network public opinion emergency detection and search (Dong et al., 2018) is the means by which the semantics of such emergencies can be obtained, as well as the complete acquisition and expression of their texts and image information.

RELATED WORKS

The analysis of user behavior data in social networks can effectively intervene and control the information dissemination process (Feng 2019; Sobkowicz et al., 2012). Lagnier et al. (2013) proposed a propagation probability model that combines user profiles and user behavior. Wang et al. (2018) utilized the structural features of user relationship diagrams combined with cyclic neural networks to predict information dissemination. Wang et al. (2017) combined dynamic directed acyclic graphs with cyclic neural networks and proposed a topological structured cyclic neural network model for information propagation prediction. Bourigault et al. (2014) assumed that all users obtain information directly from the initial source of information. However, this does not reflect the true process of information dissemination. By calculating the distance in the user feature vector, the possibility of interaction between users is analyzed, and the information dissemination process is predicted. Sun et al. (2022) suggested a novel dual dynamic graph convolutional network that employs two graph convolutional networks to learn dynamic event representations at different time stages and progressively aggregates them to capture cascading effects for better rumor detection. By implementing representation learning in the context of information dissemination, Bourigault et al. (2016) offered a fresh viewpoint on issues associated with information dissemination. It achieved favorable results by mapping temporal information into a continuous space and combined diffusion kernel functions to predict information dissemination. Conversely, this approach operates under the evident presumption that every piece of information accessed by users originates from the primary source of information, which manifestly contradicts the reality of information distribution. Liu et al. (2021) proposed a wired equivalent privacy-convolutional neural network (WEP-CNN) model to analyze public opinion information. Li et al. (2021) developed an Ortony, Clore, and Collins (OCC)-CNN emotion rule system for labelling the online public opinion case library, and it used convolutional neural networks to construct an emotional tendency analysis model under online public opinion. A method for predicting opinions was suggested in Aiying and Zhang, (2023) which combined social media influence with a long short-term memory (LSTM) neural network and developed a post hot (PH)-LSTM neural network.

Existing methods, however, fail to adequately consider the various semantic nuances both within and outside the text, such as sarcasm or figurative language, and do not leverage the implicit social semantic relationships within multi-attribute features. This results in poor semantic acquisition and expression quality for short texts on social networks. To address these issues, we employed bidirectional encoder representations from transformers (BERT), a pre-trained language model capable of capturing complex contextual semantics. By pre-training on large-scale text data, the BERT model learns rich linguistic features, including expressions of sarcasm and figurative language.

Additionally, we introduced a bidirectional LSTM (BiLSTM) network, which captures both forward and backward information in the text, providing a more comprehensive understanding of the sentence context and improving the identification of sarcasm and figurative language.

To further enhance the model's understanding capabilities, we integrated a multi-head attention mechanism. This mechanism allows the model to gather expression information from multiple positions within different representation subspaces, capturing the subtle differences in sarcasm and figurative language through multi-perspective analysis. During the model training process, we used annotated datasets that included various forms of linguistic expressions. By training on this diverse data, the model learned to recognize different types of linguistic phenomena, including sarcasm and figurative language. We also conducted detailed analyses of the model's errors in recognizing sarcasm and figurative language and made targeted improvements and fine-tuning to enhance the model's accuracy in handling such language.

Overall, we propose a BERT-Bayesian model averaging (BMA)-based method for social network public opinion analysis in a big data environment. The innovation of this method resides in:

1. Normalizing the comment text by employing the BERT model to transform social network public opinion text into a word vector matrix. By synthesizing contextual information, the BERT model facilitates comprehension of the contextual information.
2. Combining forward and backward information, the BiLSTM network architecture obtains the trajectory characteristics of information dissemination and thus more comprehensively encapsulates the contextual information in the input sequence, enhancing the efficacy of the model in analyzing public opinion on social networks.
3. Increasing the quality and efficiency of network public opinion analysis, the multiple attention mechanism is utilized to mine the dependency relationship between users, predict the actual process of information dissemination, precisely identify and extract feature information that is more beneficial for network public opinion analysis, and increase the utilization rate of feature information.

BERT-BILSTM-MULTI-HEAD ATTENTION METHOD

Overall Framework

The propagation process of information in a social network can be conceptualized as a sequence, where the set of users is denoted by U, and the propagation process of a certain information is represented by q,$q = \{(u_0, t_0), (u_1, t_1), \ldots, (u_{|q|-1}, t_{|q|-1})\}$, where $u_j \in U$ denotes that the *j-th* user is in a certain propagation sequence q, denotes that it is an event in the time of reception of user u_n, and $|q|$ denotes the length of the propagation sequence q. The model in this paper is to learn to give partial propagation sequences $\left(u_0, u_1, \ldots u_n\right)$, the propagation sequences are sorted by the time of reception of the message before time t_n, and the objective of the model is to forecast the users who are likely to receive the message at the next event.

The framework of the BERT-BMA-based social network public opinion method suggested in this article is depicted in Figure 1. First, the BERT model is used to embed users into vectors, and the characteristics of the information dissemination process are obtained through the BiLSTM network. Then, the multi-head attention diffusion module calculates the potential impact of historical users on the information dissemination process, captures the inherent importance of users in the information dissemination process, and predicts the information dissemination process. At time t_n, the input to the model is the information dissemination sequence $q_i = \left\{u_1, u_2, \ldots, u_n\right\}$. The original representation of each input user is $u_i \in \left\{u_1, u_2, \ldots, u_n\right\}$, and the users in the sequence are represented by vectors as $e_i = emb\left(u_i\right) \in R^d$, d denotes the dimension of the vector. It is worth noting that the multi-head attention diffusion module in the BERT-BMA network framework differs significantly from traditional multi-head attention modules. Traditional multi-head attention modules primarily enhance the model's ability to represent input sequences through parallel computation, with each attention head independently calculating attention weights. In contrast, the multi-head attention diffusion module in BERT-BMA further introduces collaborative interaction and information exchange mechanisms. This means that each attention head not only independently computes weights but also collaboratively explores different parts of the input sequence, thereby obtaining more comprehensive and in-depth semantic information. This diffusion mechanism enables the model to more effectively handle long-distance dependencies and complex semantic structures, thereby enhancing its application capabilities in complex tasks and multimodal data analysis.

BERT Model

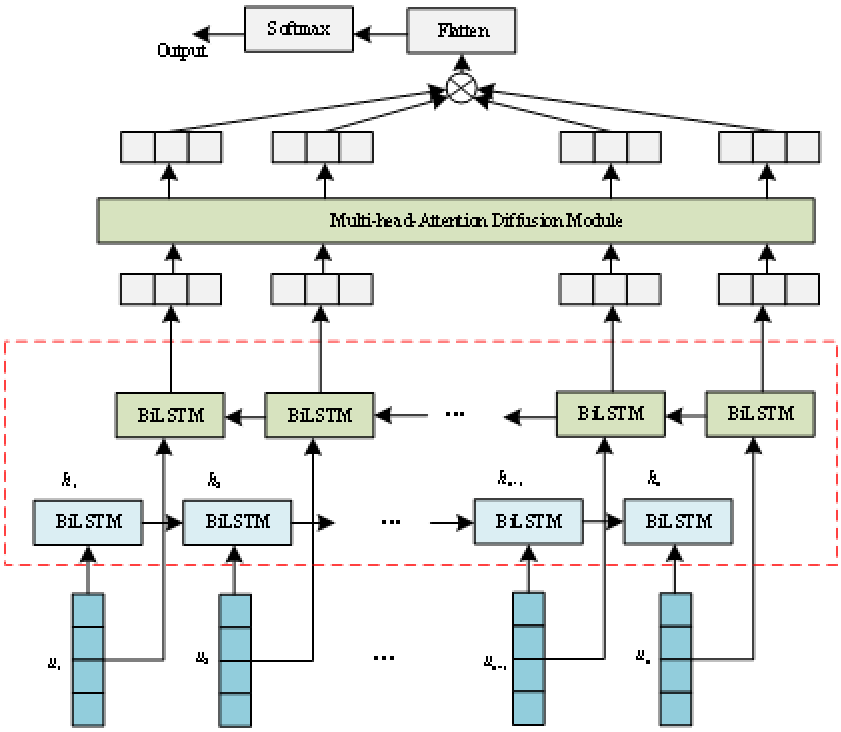
The process of generating word vectors by BERT is shown in Figure 2. The input encoding vector is the sum of three embedding features, namely, token embedding, position embedding, and segment embedding. The word embedding is a vector map of the vocabulary in the input text, the position embedding is the position encoding of the vocabulary, and the segmentation embedding is the index vector of the sentence where the vocabulary resides. Among them, CLS is the final hidden layer state corresponding to the sentence, which is mainly used to represent the entire sentence in the classification task, and SEP is a sentence separator symbol, which is used to segment the two input sentences. Each Trm unit is a transformer encoder and does not use the decoder in the transformer. The BERT input vector, which is customarily configured to have a length of 512, consists of three embedded features. Three varieties exist:

1. Token embedding: This layer is primarily responsible for assigning a 768-dimensional vector to each word via query.
2. Segment embeddings: The segment embeddings layer facilitates the classification computation of input sentence pairs for BERT.
3. Position embedding: As demonstrated by the analysis above, the transformer is incapable of capturing sequences. To address this issue, the implementation of position coding within the encoded word vector is suggested.

BiLSTM Model

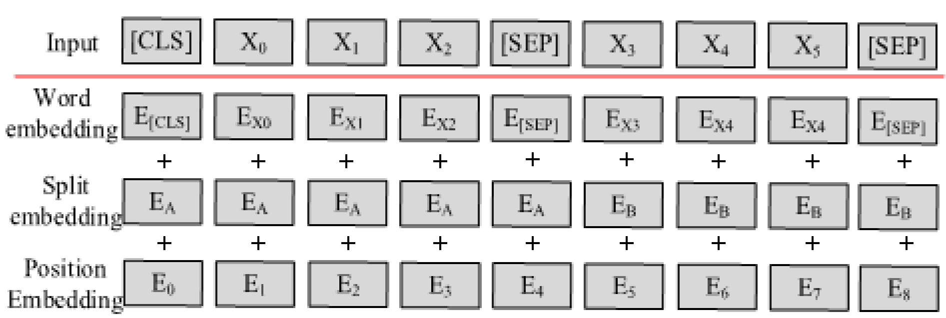
LSTM and BiLSTM are both iterations of the recurrent neural network (RNN) family. They are both employed to process sequential data and find widespread application in time series analysis and natural language processing. LSTM is an advanced variant of RNNs that addresses the issue

Figure 1. Model framework of the proposed method



of conventional RNNs becoming overly dependent on long-term data. Specifically, LSTM targets extended sequences in which conventional RNNs may struggle to capture information with significant time intervals. LSTM is an advanced variant of RNNs that addresses the issue of long-term dependence encountered by standard RNNs. Specifically, LSTM targets sequences of considerable length, where conventional RNNs may struggle to capture information over extended time periods. In order to regulate the flow of information, LSTM implements three gating units: the forget gate, the input gate, and the output gate. These gates enable the LSTM to retain or disregard lengthy sequences of information.

Figure 2. BERT pre-training model



BiLSTM is an extension of LSTM and is unique in that it processes both forward (left-to-right) and reverse (right-to-left) information of the input sequence. This helps to better capture contextual information, especially in natural language processing. BiLSTM has two sets of hidden states, one for forward propagation and the other for backward propagation. The final output is a splice of these two sets of hidden states. The formulation is similar to LSTM but with two sets of gating units and two sets of hidden states for forward and backward propagation, respectively. Forward and backward propagated information are included in the BiLSTM output, which aids the model in comprehending the input sequence's context.

BiLSTM has the following advantages over LSTM:

1. Bidirectionality: BiLSTM is capable of incorporating both past and future information, thereby enhancing its ability to capture contextual information.
2. Better sequence modeling: It can better handle long-distance dependencies and is suitable for language modeling and text generation tasks.
3. Temporal data: It is suitable for processing temporal data because it can consider future information.

In LSTM, C_t denotes a memory unit, h_t denotes a hidden unit, i_t denotes an input gate, f_t denotes a forgotten gate, and o_t denotes s an output gate. The following is the status update for the LSTM network:

$$\begin{aligned} C_t &= f_t * C_{t-1} + i_t * C_t' \\ C_t' &= \tanh h\left(W_c\left[h_{t-1}, X_{nm(t-1)}\right] + b_c\right) \\ f_t &= \sigma\left(W_f\left[h_{t-1}, X_{nmt}\right] + b_f\right) \\ i_t &= \sigma\left(W_i\left[h_{t-1}, X_{nmt}\right] + b_i\right) \\ o_t &= \sigma\left(W_o\left[h_{t-1}, X_{nmt}\right] + b_o\right) \\ h_t &= o_t * \tanh h\left(C_t\right) \\ \sigma(\cdot) &= \frac{1}{1 + e^{-(\cdot)}} \end{aligned} \tag{1}$$

In the formula, W_c, W_f, W_i, and W_o denote the weights of the memory unit, forgetting gate, input gate, and output gate, correspondingly, and b_c, b_f, b_i, and b_o denote the corresponding bias coefficients.

The BiLSTM calculation process is as follows:

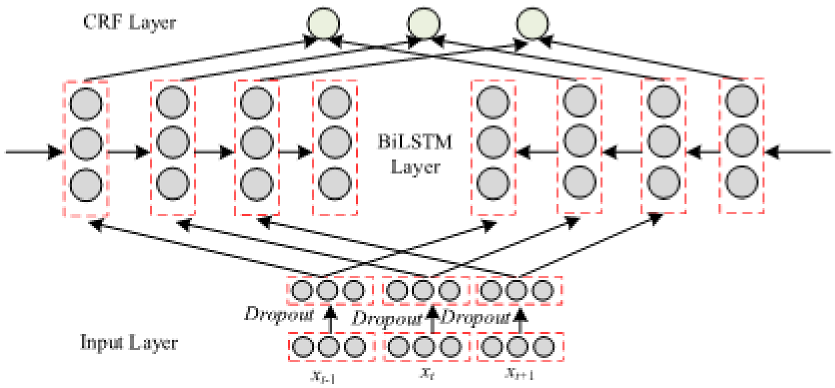
$$\vec{h}_t = f(\vec{W} \cdot x_t + \vec{W} \cdot \vec{h}_{t-1} + \vec{b}) \tag{2}$$

$$\overleftarrow{h}_t = f(\overleftarrow{W} \cdot x_t + \overleftarrow{W} \cdot \overleftarrow{h}_{t-1} + \overleftarrow{b}) \tag{3}$$

$$y_t = g(U \cdot [\overleftarrow{h};\vec{h}]+c) \tag{4}$$

Among them, $\vec{W}$ and $\overleftarrow{W}$ represent network hidden layer parameters, x_t represents input data, $\overleftarrow{h}_t$ and $\vec{h}_t$ represent the output results, and y_t represents the output of BiLSTM.

Figure 3. BiLSTM text emotion classifier



Multi-Head Attention Mechanism

Multi-head attention permits models to acquire knowledge of expression information from various locations within distinct representation subspaces. As illustrated in Figure 4, distinct semantic information can be acquired by projecting Q, K, and V via various h linear projections and subsequently executing the scaled dot-product attention computation. Equation (5) illustrates the calculation process of each multi-header module, while equation (6) depicts the output vector that transforms the outcomes of numerous self-attention headers into a particular dimension subsequent to splicing.

$$head_i = attention(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

$$MHead(Q,K,V) = Concat(head_i, ..., head_h) \tag{6}$$

Where, Q, K, V correspondingly denote the query matrix, key matrix, and value matrix; W_i^Q, W_i^K, VW_i^V respectively represent the transformation matrices of W_i^Q, W_i^K, VW_i^V. h denotes the number of self-attention heads. The output of the multi-header attention module, which has been spliced and transformed by multi-header information, is denoted by $MHead(Q,K,V)$.

Softmax

The Softmax function is employed to compute the probability of emotion categories in the subsequent manner:

$$P = Soft\max(W_p S + b_p), \tag{7}$$

where W_p and b_p are weight parameters and bias parameters, and S denotes the output of the sentence attention layer.

Figure 4. Multi-head attention model

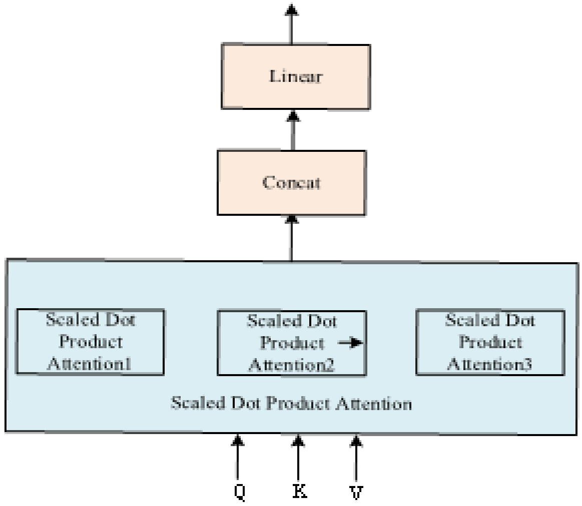


Table 1. Configuration of experimental environment

Experimental Environment	Specific Information
Operating system	Ubuntu 18.03
Memory	64GB
Language	Python3.5
Development tool	Pycharm
CPU	Intel Core CPU I7-9700K
Development platform	Tensorflowl.8.0
GPU	GeForce RTX 3090Ti

EXPERIMENT AND ANALYSIS

Experimental Environment

Keras 2.3.1, a neural network library written in the Python programming language, is utilized to conduct the experiment described in this article. Keras is a module-based neural network library that abstracts an entire neural network model as an unrestricted combination of numerous modules. As illustrated in Table 1, the hardware and software configuration of the experiment guarantees its seamless progression.

Datasets

As described in this article, we used two standard datasets for comparative experiments. These two datasets were derived from real data on Twitter and Weibo, respectively. The dataset comprises tweets accompanied by corresponding images, and there was no duplication of events between the test set and training set.

Table 2. Data set statistics

Data Set	Training Set	Test Set
Weibo	7532	1996
Twitter	12933	991

1. The Twitter dataset was mainly used to verify multi-media information tasks, that is, verify false images in social media. This dataset first collected false images and real images in related events. Then, fake images and real images were used to find relevant tweets on Twitter. As a consequence, the dataset exclusively comprised over 500 visual elements, from which an estimated 17,000 pertinent tweets were extracted. The events were distinct in the training set and test set from which this dataset has been subdivided. Each tweet contained pertinent visual and textual content.
2. The Weibo dataset included capturing all verified false rumor posts in the Weibo official rumor refutation system from May 2012 to January 2016. Dataset statistics are shown in Table 2.

Evaluating Indicators

In this study, we used accuracy, precision, recall, and F1 value to assess the method's performance indicators.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall=\frac{TP}{TP+FN} \tag{10}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

Where TP represents true positive; TN represents true negative; FP indicates false positive; FN indicates false negative.

Model Training

Model Accuracy During Training Under Different Datasets

To assess the sensitivity of BERT-BMA to datasets, this section verifies the impact of the number of different public opinion events and the proportion of social attributes in the dataset on the effectiveness of the BERT-BMA model. We conducted a dichotomous and multi-class performance test on BERT-BMA based on the number of events included in Twitter and Weibo datasets. Figure 5 shows the training process for BERT-BMA on two datasets with different event category numbers and multiple attribute feature ratios for social networks. The curves in different colors represent the recall rates obtained during 180 iterations of the test set over different datasets. The outcomes presented in Figure 5 indicate that the experimental data set exhibits a general upward trend in recall rate,

Figure 5. Training process of the proposed method under different datasets

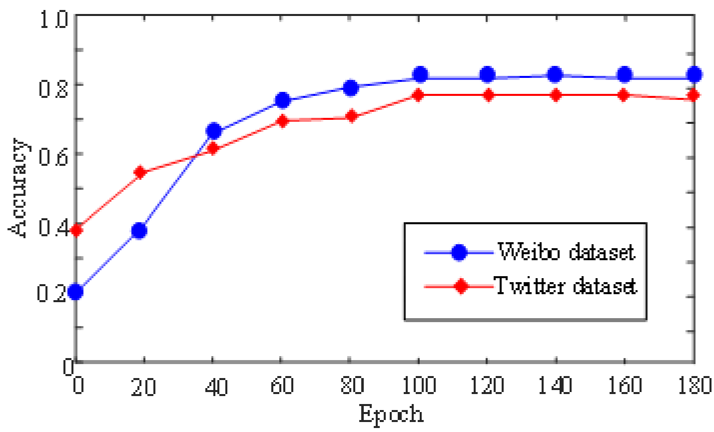
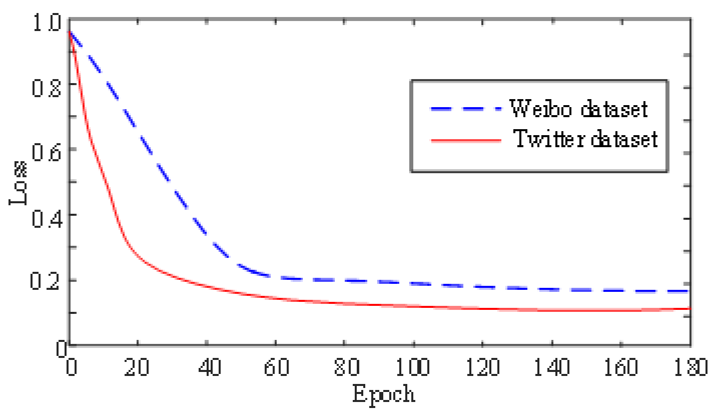


Figure 6. Loss value during training

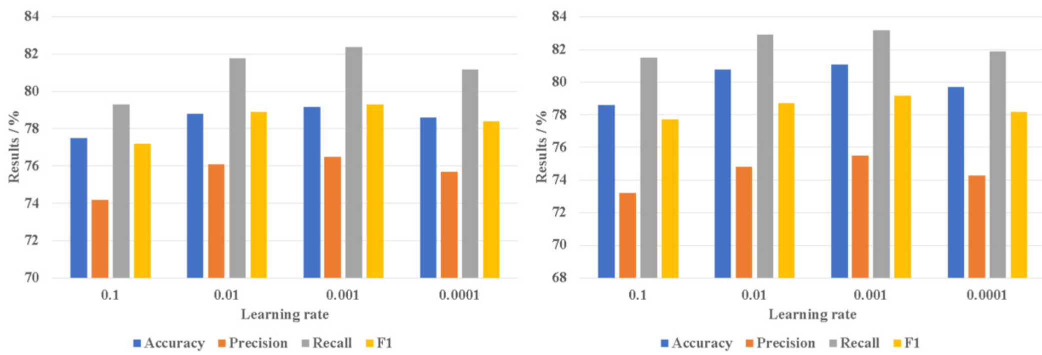


accompanied by minor fluctuations. The crawled Weibo cross media emergency dataset can exhibit similar classified recall rates compared to the public Twitter dataset with a standard category, indicating that the BERT-BMA model has better stability within the social network cross media data set.

Model Loss Training Under Different Datasets

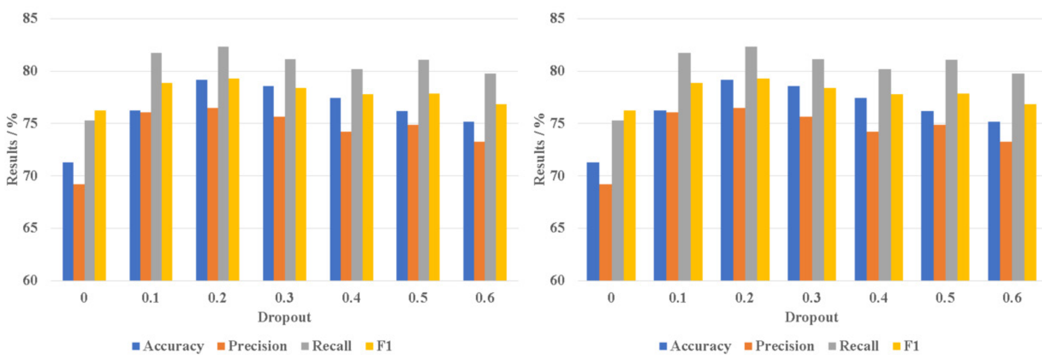
In this section we calculated the loss function value of the BERT-BMA based social network public opinion analysis method, and further analyzed the effectiveness of algorithm training through the loss changes of the algorithm. The loss value curve of the method based on BERT-BMA shows the change trend of the algorithm target value at different training stages. The method for analyzing public opinion on social networks based on deep learning that is developed in this article was trained using the 4.2 dataset. Throughout the 180 epochs of training and testing, the process diagram of the acquired loss value is depicted in Figure 6. In the initial 60 epochs, the loss rate of the training set decreased significantly as the number of training times increased in both datasets. Subsequently, as the number of training times increased in the final 120 epochs, the loss rate progressively stabilized. The model achieved convergence on the Twitter and Weibo datasets after 64 epochs, with loss rates remaining between 0.12 and 0.18, respectively.

Figure 7. The effect of learning rate on indictors value



Note. The first graph illustrates the Twitter dataset; the second graph illustrates the Weibo dataset.

Figure 8. Trends in indicators values with different dropout



Note. The first graph illustrates the Twitter dataset; the second graph illustrates the Weibo dataset.

Hyperparameter Analysis

In order to study the value of learning rate and dropout, comparative experiments were conducted by setting its value to different values to observe its effect on the model. The experimental results of indictors value with the value of learning rate and dropout are shown in Figures 7 and 8, respectively.

Figure 7 illustrated that the value of the indicators demonstrates a progressive increase with the learning rate from 0.0001 to 0.001 on both datasets. Nevertheless, a decline in performance became apparent once the learning rate increased. As a result, the learning rate for this experiment was set at 0.001.

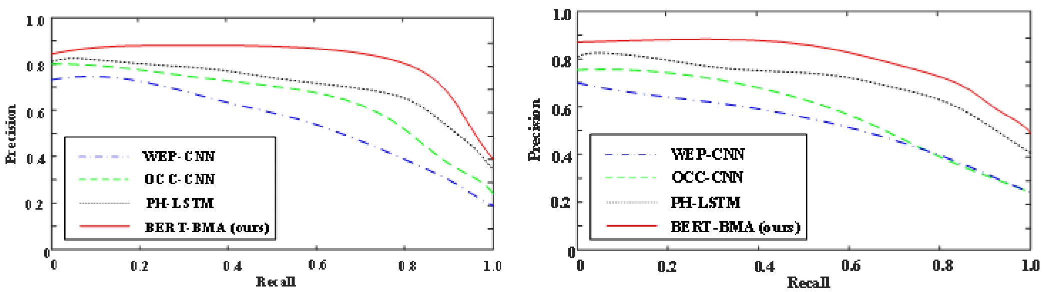
As depicted in Figure 8, for both Twitter and Weibo datasets, the indictors value exhibited an upward trend as the dropout value increased from 0 to 0.2. However, as the dropout continued to increase, the indicator value showed a downward trend. Therefore, 0.2 was chosen as the dropout value for this experiment.

Comparative Experiment

Comparison of Precision-Recall Values of Different Models

We presented a method for analyzing public opinion on social networks in the context of big data using BERT-BMA. We also developed a pre-training language model based on BERT and employed the BiLSTM model to acquire semantic information in both positive and negative directions

Figure 9. Comparison of precision-recall curve of social network cross-media public opinion emergency search



Note. The first graph illustrates the Twitter dataset; the second graph illustrates the Weibo dataset.

simultaneously. To assess the efficacy of the BERT-BMA model in complex tasks of multi-category public opinion event search, under the same experimental conditions, the suggested method was compared with WEP-CNN (Liu et al., 2021), OCC-CNN (Li et al., 2021), and PH-LSTM (Aiying & Zhang, 2023). Figure 9 depicts the precision-recall curve of the BERT-BMA model in the emergency search experiment.

When the results are combined, it can be found that the BERT-BMA model achieved the best search results. Compared to the proposed method, WEP-CNN and OCC-CNN simply use convolutional machines over networks, there is no fusion of more meaningful semantic information in the learning of cross modal data features. The proposed method utilizes multiple attention mechanisms to mine dependencies between users to forecast the actual information dissemination process, accurately identify and extract feature information that is more helpful to network public opinion analysis, resulting better performance on both datasets. Compared with PH-LSTM, the suggested method uses the BiLSTM network structure to derive the trajectory characteristics of information propagation by combining forward and backward information, thus encompassing the contextual information in the input sequences in a more comprehensive way, enhancing the efficacy in the analysis of online public opinion.

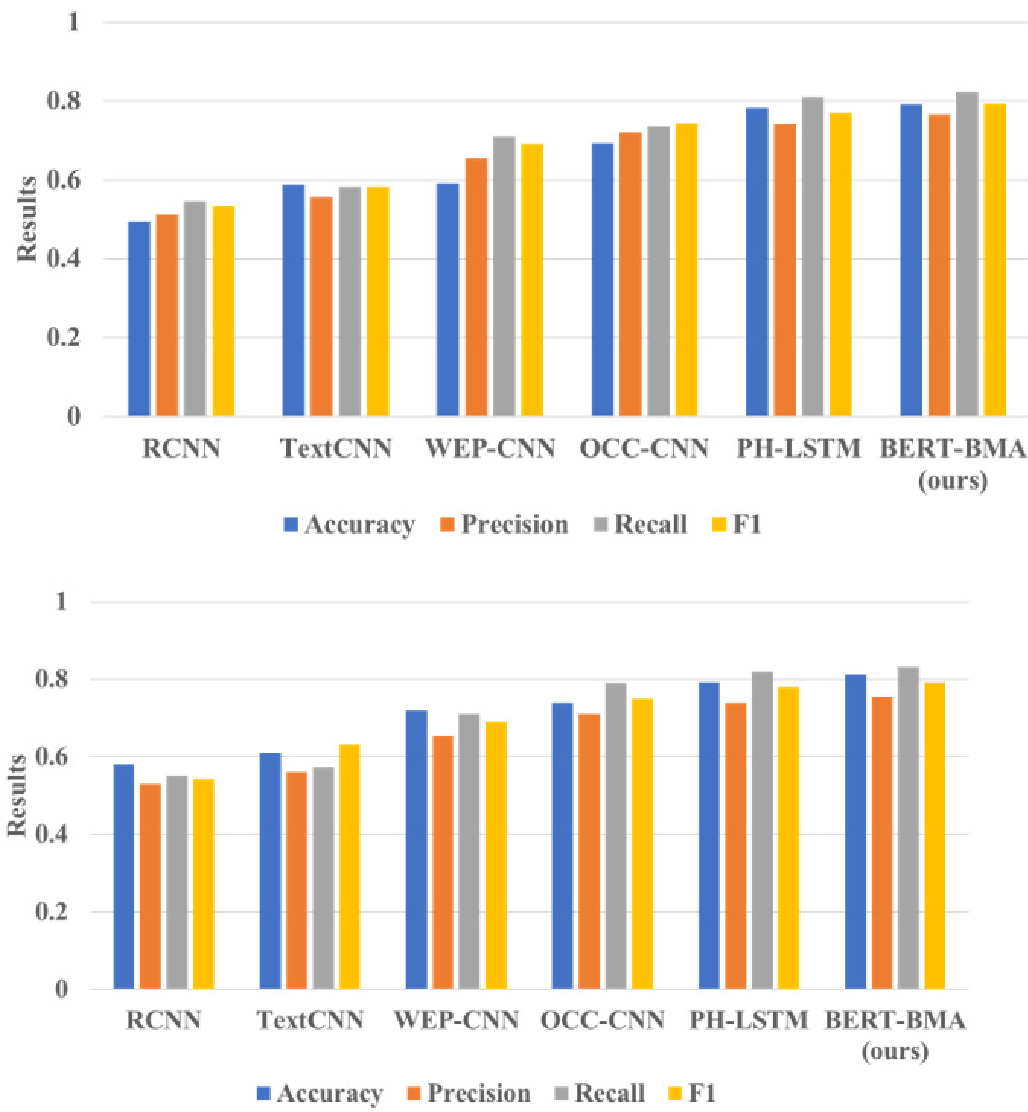
Comparison of Model Performance Under Different Datasets

To assess the advantages of BERT-BMA model in the analysis of public opinion events, under the same experimental conditions, several baseline models were selected for comparison including WEP-CNN (Liu et al., 2021), OCC-CNN (Li et al., 2021), PH-LSTM (Aiying & Zhang, 2023), recurrent CNN (RCNN) (Maheshwarkar et al., 2021), and TextCNN (Wang et al., 2022), the comparison results are shown in Figure 10.

As demonstrated in Figure 10, across all different models, the suggested method exhibits a generally superior performance in social network public opinion analysis tasks when compared with the comparative reference. Overall, among all methodologies tested in the Twitter dataset experiment, text accuracy was the lowest. This is due to the fact that the Twitter dataset encompasses a multitude of events, and words of text differ considerably. Local characteristics of text cannot therefore be shared, leading to a low rate of accuracy.

The accuracy, precision, recall, and F1 values of the social network public opinion analysis method that was suggested are as follows on the Twitter dataset: 0.792, 0.765, 0.824, and 0.793, correspondingly. The suggested social network public opinion analysis method achieved accuracy, precision, recall, and F1 values of 0.811, 0.755, 0.832, and 0.792, correspondingly, on the Weibo dataset. These values surpassed those of the compared methods. The reason is that the suggested method standardizes comment text using the BERT model and mines dependencies between users

Figure 10. Comparison of model performance under different datasets



Note. The first graph illustrates the Twitter dataset; the second graph illustrates the Weibo dataset.

using BiLSTM to acquire feature information that is more significant to social network opinion analysis, resulting in improved outcomes.

Ablation Experiment

To validate the efficacy of the suggested method in social network public opinion analysis, a corresponding ablation experiment was designed using the control variable method.

Experiment 1: BiLSTM method.
Experiment 2: BERT-BiLSTM method.
Experiment 3: BiLSTM-Multi-head attention method.

Table 3. Results of ablation experiment

Method	Accuracy	Non Rumor(F1)	False Rumor(F1)	True Rumor(F1)	Unconfirmed Rumor(F1)
Twitter Dataset					
BiLSTM	0.633	0.633	0.658	0.675	0.598
BERT-BiLSTM	0.635	0.647	0.673	0.682	0.624
BiLSTM-multi-head-attention	0.648	0.658	0.684	0.684	0.657
BERT-BMA (ours)	0.792	0.685	0.689	0.697	0.686
Weibo Dataset					
BiLSTM	0.642	0.668	0.663	0.679	0.612
BERT-BiLSTM	0.658	0.678	0.678	0.686	0.665
BiLSTM-multi-head-attention	0.667	0.686	0.686	0.698	0.679
BERT-BMA (ours)	0.811	0.707	0.712	0.721	0.703

Table 4. Training time to complete one iteration for different models

Model	Twitter Dataset(s)	Weibo Dataset(s)
RCNN	68	51
TextCNN	85	65
WEP-CNN	369	138
OCC-CNN	285	107
PH-LSTM	512	265
BERT-BMA (ours)	476	234

Experiment 4: BERT-BMA (ours).

All of the aforementioned investigations underwent training in an identical hardware and software environment, employing identical training parameters. The suggested method was compared to the base network in the ablation experiment. The accuracy and F1value indicators of the suggested method were both optimal for the section 4.1 dataset, as illustrated in Table 3. This is due to the fact that a language model for the suggested method has been developed utilizing BERT pre-training word vectors rather than conventional word vectors. The generation of semantic vectors is dynamic and determined by the words' context. The BiLSTM model simultaneously captures semantic information by utilizing a combination of positive and negative directions to measure the emotional polarity information of each text in its entirety. Additionally, the model employs multiple attention mechanisms to interact with different features.

Network Training Time Analysis

Under the identical hardware environment, the training time of different models to complete one iteration on two datasets were compared and analyzed, and the outcomes are depicted in Table 4.

As outlined by the data in Table 4, RCNN and TextCNN used a fast model to acquire text features, so the training time of these two networks was much lower than other models. Compared with OCC-CNN, WEP-CNN employed a two-channel convolutional neural network, so it required more training time. The BERT-BMA model was more complex compared to WEP-CNN and OCC-CNN, and the invocation of multi-head attention mechanism increased the training time of the model to some extent. In comparison with other models, the PH-LSTM took the longest in training time because it considered the changes in long- and short-term network opinions. Overall, this training time was acceptable for the proposed BERT-BMA in obtaining the best accuracy rate.

CONCLUSIONS

In view of the existing semantic acquisition and expression methods for public opinion analysis on social networks, they do not fully consider multiple semantic information inside and outside the text, nor do they rely on effective multi-attribute features to imply social semantic relationships. We propose a social network public opinion analysis method based on BERT-BMA in a big data environment and combine multiple attention mechanisms to mine dependencies between users. Experiential outcomes demonstrate that the suggested method is superior to the comparison method. Furthermore, the detection of public opinion emergencies can be enhanced by leveraging the real-time nature of social network data and the rapidity of message dissemination. Considering further studying the mining of the evolution rules of public opinion emergencies and the prediction of development trends based on the integration of semantic analysis and expansion of cross media emergencies, so as to conduct real-time tracking and discovery of public opinion emergencies.

The proposed BERT-BMA model, however, still has some limitations that need to be improved. In future work, we plan to integrate BERT-BMA with advanced artificial intelligence (AI) technologies such as image recognition, knowledge graphs, user behavior analysis, and time series analysis to enhance its overall capability in social network public opinion analysis. This will enable a comprehensive understanding of multi-modal data and more accurate public opinion predictions. Additionally, we will incorporate more factors reflecting the occurrence of public opinion emergencies, such as user behavior and news release space, to identify the spatial scope of user behavior and information dissemination. By learning and expressing the comprehensive features of sudden events in deep neural network models, more effective social network public opinion analysis can be achieved. We will also use larger and more comprehensive datasets to train the BERT-BMA model, enhancing its generalization ability and providing decision-makers with more accurate public opinion insights.

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING STATEMENT

No funding was received for this work.

PROCESS DATES

07, 2024

This manuscript was initially received for consideration for the journal on 05/28/2024, revisions were received for the manuscript following the double-anonymized peer review on 06/25/2024, the manuscript was formally accepted on 07/15/2024, and the manuscript was finalized for publication on 07/29/2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Zheng Liu; lz211215@163.com

REFERENCES

Aiying, B., & Zhang, Z. Y. (2023). Analysis of public opinion on the network of burst events based on the PH-LSTM neural network. *Frontiers in Science and Engineering*, 1-6.

Anstead, N., & O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 UK general election. *Journal of Computer-Mediated Communication*, *20*(2), 204–220. 10.1111/jcc4.12102

Bourigault, S., Lagnier, C., & Lamprier, S. (2014). Learning social network embeddings for predicting information diffusion. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. 10.1145/2556195.2556216

Bourigault, S., Lamprier, S., & Gallinari, P. (2016). Representation learning for information diffusion through social networks: an embedded cascade model. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 10.1145/2835776.2835817

Chakraborty, P., & Sharma, A. (2019). Public opinion analysis of the transportation policy using social media data: A case study on the Delhi odd–even policy. *Transportation in Developing Economies*, *5*(3), 1–9. 10.1007/s40890-019-0074-8

Chen, L., Chen, J., & Xia, C. (2022). Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, *64*(1), 1030–1040. 10.1016/j.jisa.2021.103060

Dong, H. W., Qian, W., & Jie, C. (2018). Tracing public opinion propagation and emotional evolution based on public emergencies in social networks. *International Journal of Computers, Communications & Control*, *13*(1), 129–142. 10.15837/ijccc.2018.1.3176

Dong, X., & Lian, Y. (2021). A review of social media-based public opinion analyses: Challenges and recommendations. *Technology in Society*, *67*(3), 101724. 10.1016/j.techsoc.2021.101724

Du, X., Kowalski, M., Varde, A. S., de Melo, G., & Taylor, R. W. (2020). Public opinion matters: Mining social media text for environmental management. *ACM SIGWEB Newsletter*, *4*(2), 1–15. 10.1145/3352683.3352688

Feng, Z. (2019). Hot news mining and public opinion guidance analysis based on sentiment computing in network social media. *Personal and Ubiquitous Computing*, *23*(3-4), 373–381. 10.1007/s00779-018-01192-y

Gao, Y., Wang, J., Li, Z., & Peng, Z. (2023). The social media big data analysis for demand forecasting in the context of globalization: Development and case implementation of innovative frameworks. [JOEUC]. *Journal of Organizational and End User Computing*, *35*(3), 1–15. 10.4018/JOEUC.325217

Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China. *International Journal of Environmental Research and Public Health*, *17*(8), 2788–2796. 10.3390/ijerph1708278832316647

He, M., Ma, C., & Wang, R. (2022). A data-driven approach for university public opinion analysis and its applications. *Applied Sciences (Basel, Switzerland)*, *12*(18), 9136–9139. 10.3390/app12189136

Karami, A., Bennett, L. S., & He, X. (2018). Mining public opinion about economic issues: Twitter and the us presidential election. [IJSDS]. *International Journal of Strategic Decision Sciences*, *9*(1), 18–28. 10.4018/IJSDS.2018010102

Kim, D. S., & Kim, J. W. (2014). Public opinion sensing and trend analysis on social media: A study on nuclear power on Twitter. *International Journal of Multimedia and Ubiquitous Engineering*, *9*(11), 373–384. 10.14257/ijmue.2014.9.11.36

Lagnier, C., Denoyer, L., Gaussier, E., et al. (2013). Predicting information diffusion in social networks using content and user's profiles. In *Proceedings of the Advances in Information Retrieval: 35th European Conference on IR Research (ECIR 2013).* Springer Berlin Heidelberg. 10.1007/978-3-642-36973-5_7

Li, D., Zhang, Y., & Li, C. (2019). Mining public opinion on transportation systems based on social media data. *Sustainability (Basel)*, *11*(15), 4016–4023. 10.3390/su11154016

Li, J., Wang, Y., & Wang, J. (2021). An analysis of emotional tendency under the network public opinion: Deep learning. *Informatica (Vilnius)*, *45*(1), 2212–2222.

Liang, Q., Hu, C., & Chen, S. (2021). Evaluation of the optimal topic classification for social media data combined with text semantics: A case study of public opinion analysis related to COVID-19 with microblogs. *ISPRS International Journal of Geo-Information*, *10*(12), 811–821. 10.3390/ijgi10120811

Liu, X., Li, Z., & Lv, J. (2021). Research on key technologies of public opinion analysis of power operation and maintenance based on deep learning. In *Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Information Systems*. 10.1145/3469213.3470420

Maheshwarkar, A., Kumar, A., & Gupta, M. (2021). Analysis of written interactions in open-source communities using RCNN. In *Proceedings of the 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. 10.1109/ICAC3N53548.2021.9725610

Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of the@ NLP Can U Tag# usergeneratedcontent, 7*(11) 15-22.

McGregor, S. C. (2019). Social media as public opinion: How journalists use social media to represent public opinion. *Journalism*, *20*(8), 1070–1086. 10.1177/1464884919845458

Murphy, J., Link, M. W., & Childs, J. H. (2014b). Social media in public opinion research: Report of the AAPOR task force on emerging technologies in public opinion research. *American Association for Public Opinion Research*, *2*(11), 1752–1763.

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M., & Harwood, P. (2014a). Social media in public opinion research: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, *78*(4), 788–794. 10.1093/poq/nfu053

Rim, H., Lee, Y. A., & Yoo, S. (2020). Polarized public opinion responding to corporate social advocacy: Social network analysis of boycotters and advocators. *Public Relations Review*, *46*(2), 2023–2032. 10.1016/j.pubrev.2019.101869

Skoric, M. M., Liu, J., & Jaidka, K. (2020). Electoral and public opinion forecasts with social media data: A meta-analysis. *Information (Basel)*, *11*(4), 187–196. 10.3390/info11040187

Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, *29*(4), 470–479. 10.1016/j.giq.2012.06.005

Sun, M. Z.. (2022). DDGCN: Dual dynamic graph convolutional networks for rumor detection on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 10.1609/aaai.v36i4.20385

Tavoschi, L., Quattrone, F., & Andrea, E. D. (2020). Twitter as a sentinel tool to monitor public opinion on vaccination: An opinion mining analysis from September 2016 to August 2017 in Italy. *Human Vaccines & Immunotherapeutics*, *16*(5), 1062–1069. 10.1080/21645515.2020.171431132118519

Trivedi, S., & Patel, N. (2022). Mining public opinion about hybrid working with RoBERTa. *Empirical Quests for Management Essences*, *2*(1), 31–44.

Urban, J., & Bulkow, K. (2013). Tracing public opinion online–An example of use for social network analysis in communication research. *Procedia: Social and Behavioral Sciences*, *10*(3), 108–126. 10.1016/j.sbspro.2013.10.703

Wang, J., Zheng, V. W., & Liu, Z. (2017). Topological recurrent neural network for diffusion prediction. In *Proceedings of the 2017 IEEE international conference on data mining (ICDM)*. IEEE. 10.1109/ICDM.2017.57

Wang, Z., Chen, C., & Li, W. (2018). A sequential neural information diffusion model with structure attention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 10.1145/3269206.3269275

Wang, Z., Wang, M., Shen, H., & Han, Y. (2022). Application of sentiment classification of Weibo comments based on TextCNN model. In *Proceedings of the 2022 International Conference on Computer Network, Electronic and Automation (ICCNEA)*. 10.1109/ICCNEA57056.2022.00057

Younus, A., Qureshi, M. A., & Asar, F. F. (2011). What do the average twitterers say: A Twitter model for public opinion analysis in the face of major political events. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*. IEEE. 10.1109/ASONAM.2011.85

Zhang, X., He, X., Du, X., Zhang, A., & Dong, Y. (2023). Supply chain practices, dynamic capabilities, and performance: The moderating role of big data analytics. [JOEUC]. *Journal of Organizational and End User Computing*, *35*(3), 1–26. 10.4018/JOEUC.325214

Zhang, Y., Chen, F., & Rohe, K. (2022). Social media public opinion as flocks in a murmuration: Conceptualizing and measuring opinion expression on social media. *Journal of Computer-Mediated Communication*, *7*(4), 2312–2321.

Hanqing Sun, assistant professor, Master, Graduated from the Zhengzhou University in 2009.Studied in Information Engineering,Worked in Henan University of Animal Husbandry and Economy. He research interests include information processing

Zheng Liu, University Lecturer,Master,Graduated from the Zhengzhou University in 2009.Studied in Information Engineering,Worked in Henan University of Animal Husbandry and Economy.He research interests include information processing and data analysis.

Weimin Lian,professor,Master.Graduated from the Zhengzhou University in 2003.Studied in Information Engineering,Worked in Henan University of Animal Husbandry and Economy. He research interests include Software Engineering and data analysis.

Guizhi Wang,professor,Master.Graduated from the Zhengzhou University in 2008.Studied in Information Engineering,Worked in Henan University of Animal Husbandry and Economy.Her research interests include software technology