frontiers in Physics

ORIGINAL RESEARCH
published: 28 September 2021
doi: 10.3389/fphy.2021.763081

COVID-19 Rumor Detection on Social Networks Based on Content Information and User Response

*Jianliang Yang and Yuchen Pan**

School of Information Resource Management, Renmin University of China, Beijing, China

The outbreak of COVID-19 has caused a huge shock for human society. As people experience the attack of the COVID-19 virus, they also are experiencing an information epidemic at the same time. Rumors about COVID-19 have caused severe panic and anxiety. Misinformation has even undermined epidemic prevention to some extent and exacerbated the epidemic. Social networks have allowed COVID-19 rumors to spread unchecked. Removing rumors could protect people's health by reducing people's anxiety and wrong behavior caused by the misinformation. Therefore, it is necessary to research COVID-19 rumor detection on social networks. Due to the development of deep learning, existing studies have proposed rumor detection methods from different perspectives. However, not all of these approaches could address COVID-19 rumor detection. COVID-19 rumors are more severe and profoundly influenced, and there are stricter time constraints on COVID-19 rumor detection. Therefore, this study proposed and verified the rumor detection method based on the content and user responses in limited time CR-LSTM-BE. The experimental results show that the performance of our approach is significantly improved compared with the existing baseline methods. User response information can effectively enhance COVID-19 rumor detection.

Keywords: rumor detection, COVID-19, social networks, social physics, user responses

OPEN ACCESS

Edited by:
Chengyi Xia,
Tianjin University of Technology, China

Reviewed by:
Zhan Bu,
Nanjing University of Finance and Economics, China
Yuan Bian,
University of Chinese Academy of Sciences, China

****Correspondence:***
Yuchen Pan
panyuchen@ruc.edu.cn

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *23 August 2021*
Accepted: *15 September 2021*
Published: *28 September 2021*

Citation:
Yang J and Pan Y (2021) COVID-19 Rumor Detection on Social Networks Based on Content Information and User Response. Front. Phys. 9:763081. doi: 10.3389/fphy.2021.763081

INTRODUCTION

Nowadays, the social network has become an indispensable tool in people's daily life. People carry out activities such as social communication, obtaining information, and expressing opinions on social network platforms. In the above activities, securing information and expressing opinions are particularly frequent on social networks. However, most of the content on social networks is user-generated content (UGC), and the veracity of UGC is challenging to be guaranteed. The net structure of a social network is convenient for the viral dissemination of information, which makes it easy to generate rumors in a social network, and rumors are easier to spread on a large scale. Rumors in social networks are particularly rampant when public incidents occur. During the COVID-19 epidemic outbreak in 2020, a large number of rumors spread widely on social platforms such as Twitter and Weibo, which aggravated people's fear and anxiety about the epidemic, and made people experience an "information epidemic" in the virtual space [1]. Rumor governance on social networks is essential and necessary work.

For social network users, removing rumors on social networks could effectively reduce people's anxiety and stress during COVID-19 and help people reduce wrong behavior (such as refusing vaccines) caused by misinformation, thus protecting their health. For social network platforms,

removing rumors could reduce the spread of false information and improve the platforms' environment and user experience. For public health departments, removing rumors could reduce the cost of responding to the epidemic by allowing truthful and correct policies and guidelines to be disseminated effectively. The effective detection of rumors is the key to rumor governance. If false rumors or fake news on social networks can be detected sooner, relevant measures (e.g., rumor refutation and timely disclosure of information) will be taken more timely.

For the detection of rumors, existing studies proposed methods from various perspectives. Most methods for rumor detection are based on rumor content information, rumor source, and propagation path. Rumor detection methods based on content information focuses on language style, emotional polarity, text and picture content features [1]. Rumor detection methods based on rumor source focuses on web address (e.g., the source URLs of rumors), website credit, and webpage metadata [2]. Rumor detection methods based on propagation focus on the propagation structural features during rumor propagation, such as the retweeting and commenting behavior by social platform users [3]. With the development of artificial intelligence, deep learning methods make a significant contribution to various tasks. Some studies had adopted artificial intelligence based methods in rumor detection and achieved decent performance [4]. With the advent of language models based on transfer learning like BERT [5] and GPT3 [6], the analysis ability of deep learning models for natural language is further improved, which indicates us to utilize the language models based on transfer learning on rumor detection.

Time constraints are an essential factor that needs to be taken into consideration. The timelier we detect the fake news on a social network, the less harm it will cause. Public health emergencies like COVID-19 epidemic-related information are radically concerned and could profoundly affect psychology and behavior. There is a stricter time constraint on COVID-19 rumor detection. With the time constraints, methods based on propagation path are not applicative. It takes time to form the propagation path of a rumor. This indicates that we pay more attention to the content of rumors and user comments, and retweets. Because the users of a social network can comment and retweet on a rumor, known as user responses, the user responses usually contain information on the rumor's veracity. However, most of the existing studies did not take the content of user responses. The responses from users can be considered as discussions or arguments around the rumor. By extracting user response features, we may be able to implement rumor detection better. Facing rumor detection on COVID-19 on social networks, this study proposes a novel deep learning method based on rumor content and user responses. Our method has the following contributions:

1. Our method incorporates user response sequence into the rumor detection system. On the one hand, the information contained in user responses is fully utilized; on the other hand, the sequence of user responses also contains a part of the features of the rumor propagation path.
2. Time limit is added in our study. Only user responses within 24 h of rumor release are used as model input for detection.
3. Our method is based on the language model with transfer learning to obtain content features. Moreover, to capture richer information about COVID-19 in the social context, we use post-training mechanism to post train BERT on the corpus of COVID-19 related posts on Twitter and Weibo.

The structure of this paper is as follows: *Related Work* introduces the research progress on this topic, especially the progress in methods development. *Methods* introduces the problem statement of COVID-19 rumor detection and the methods proposed in this study. *Experiments* introduces the experimental dataset, baselines, evaluation methods, experiment settings, and experimental results. In *Discussion*, the experimental results are deeply analyzed and discussed. *Conclusion* summarizes the research findings of our work and points out some future directions.

RELATED WORK

With the development of intelligent devices and mobile internet, human beings are experiencing an era of information explosion. At present, countless information is flooded in our lives. However, not all of this information is true, and even in the outbreak of a major public health crisis such as the COVID-19 epidemic, much of the information we have obtained is false rumors. Generally speaking, a rumor refers to a statement whose value can be true, false, or uncertain. Rumor is also called fake news [7]. Rumor detection means to determine whether a statement or a Twitter post is a rumor or non-rumor. The task of determining whether a statement or a Twitter post is a rumor or non-rumor is also called rumor verification [8]. According to recent studies, rumor detection refers to the veracity value of a rumor. Therefore, rumor detection is equivalent to rumor verification [9].

Since information is easier to spread on social networks, rumor detection on social networks is more complex than general fake news detection. For detecting fake news, text features, source URL, and source website credit can be considered [2]. The source of information is more complex on the social network, and information spreading is much faster and wider. Rumor detection on social media is critical. Existing studies show that rumor detection on social networks is often based on text content features, user features, rumor propagation path features. Among them, the text content features and rumor propagation path features are significant for rumor detection.

For rumor detection methods based on text content features, writing style and topic features are an essential basis for determining whether rumors are true or not [10]. In addition to the text content, postag, sentiment, and specific hashtags such as "#COVID19" and "#Vaccine" are also important content features [11]. Chua et al. summarized six features, including comprehensiveness, sentence, time orientation, quantitative details, writing style, and topic [12]. With the development of deep learning and artificial intelligence, deep learning models

such as CNN have been used to extract the text features of rumors and combined with word embedding generation algorithms such as Word2vec, GloVe. Deep learning models can automatically extract the features related to rumors detection through representation learning and have achieved decent performance in the rumor detection task. Using CNN to extract the features of rumor content has a good effect on limited data and early detection of rumors [13]. CNN is also applied to feature extraction of text content in multimodal fake news detection [4].

Rumor propagation path is another common and essential feature of rumor detection. Real stories or news often have a single prominent spike, while rumors often have multiple prominent spikes in the process of spreading. Rumors spread farther, faster, and more widely on social networks than real stories or news [14]. Focusing on the rumor recognition path, Kochkina et al. proposed the branch-LSTM algorithm, which uses LSTM to transform propagation path into a sequence, combines text features and propagation path features and conducts rumor verification through a multi-task mechanism [8]. Liu et al. regarded the rumor propagation path as a sequence and utilized RNN to extract propagation path information [15]. Kwon et al. combined text features, user network features, and temporal propagation paths to determine rumors [16]. Bian et al. transformed rumor detection into a graph classification problem and constructed the Bi-GCN from Top-Down and Bottom-Up two directions to extract the propagation features on social networks [3].

Because rumor detection needs a high-quality dataset as support, few studies are focusing on COVID-19 rumor detection. Glazkova et al. proposed the CT-BERT model, paying attention to the content features, and fine-tuned the BERT model based on other news and Twitter posts related to COVID-19 [17]. For the datasets, Yang et al. [18] and Patwa et al. [19] provided rumor datasets on COVID-19, which are mainly based on social network platforms such as Twitter, Facebook, and Weibo, and news websites such as PolitiFact.

Compared to routine rumor detection, COVID-19 rumor detection has a strict time constraint, especially during the outbreak stage of the epidemic. Once the rumor detection is not timely enough, the negative impact brought by rumor propagation is enormous. The damage caused by COVID-19 rumors can increase rapidly over time and have an even more significant and broader impact than other rumors. Therefore, early rumor detection on COVID-19 needs to be considered, and early detection and action should be taken. Most of the existing studies focus on the features of rumor content and propagation path but pay insufficient attention to user responses and rumor detection within a limited time. User responses to a rumor often include stance and sentiment toward the rumor. Particularly for false rumors, user responses are often more controversial [20].

In the existing studies, some suggested that user response can better assist systems in detecting rumors [9, 20]. However, more studies use user response to determine user stance and regard user stance classification as a separate task. User stance refers to users' attitudes toward rumors. Similar to sentiment polarity classification, user stance is generally a value of [−1,1], where one indicates full support for the rumor to be true, 0 indicates neutrality, and −1 indicates no support for the rumor to be true at all [21]. There are studies on implementing rumor verification and user stance simultaneously through a multi-task mechanism [8]. However, there are very few studies that directly use user responses to enhance rumor detection. Given the shortcomings of existing studies, this study proposes a rumor detection method based on rumor content and user response sequence in a limited time and uses the language model based on transfer learning to extract the features of rumor text.

METHODS

This section introduced the method based on rumor content and the user response sequence proposed in our study. *Problem Statement* presents the problem statement of rumor detection. *Rumor Content Feature Extractor* introduces the feature extracting method for the COVID-19 rumor content. User *Response Feature Extractor* introduces the feature extracting method for the user response of the COVID-19 rumor content.

Problem Statement

The problem of rumor detection on COVID19 on social networks can be defined as: for a rumor detection dataset $R = \{r_1, r_2, \ldots, r_n\}$. r_i is the i-th rumor event, and n is the number of rumors in the rumor dataset. $r_i = \{x_i, s_1^i, \ldots s_j^i, \ldots, s_{m_i}^i\}$, where x_i is the source post of rumor event r_i, and s_j^i is the response to the post x_i from other users within a certain period of time. Specifically, user responses $s_{1\ldots m_i}^i$ to the post x_i can be defined as a sequence. For each rumor events r_i is associated with a ground-truth label $y_i \in \{\mathrm{F}, \mathrm{T}, \mathrm{U}\}$ corresponding to False Rumor, True Rumor and Unverified Rumor. Given a rumor dataset on COVID-19, the goal of rumor detection is to construct a classification system f, that for any r_i, its label y_i can be determined. In many studies, this definition is the same as rumor veracity classification task [9, 22].

Rumor Content Feature Extractor

In this study, we implemented a deep learning model based on content features and user responses for COVID-19 rumor detection in limited time. Therefore, content features are an important basis for rumor detection. We need to extract the features for the rumor content and map the rumor content to embedding in a vector space. In the common representation learning process, for a rumor text x_i, pre-training models such as Word2vec or GloVe are generally transform the words $\{w_{i,1} \ldots w_{i,n}\}$ composed of rumor text x_i into word embedding, and then deep learning models such as RNN and CNN are used to extract features related to rumor detection and form rumor content feature C. For example, the last step h_n of RNN or the vector C from CNN pooling layer is normally used to represent the content feature of the whole rumor text x_i.

Along with the development of natural language processing technology, language models based on transfer learning, such as ELMo [23], BERT [5], and XLNet [24], have achieved excellent performance in text feature extraction. Benefit from the transfer learning mechanism, language models like BERT significantly improved backend tasks, including text classification, machine

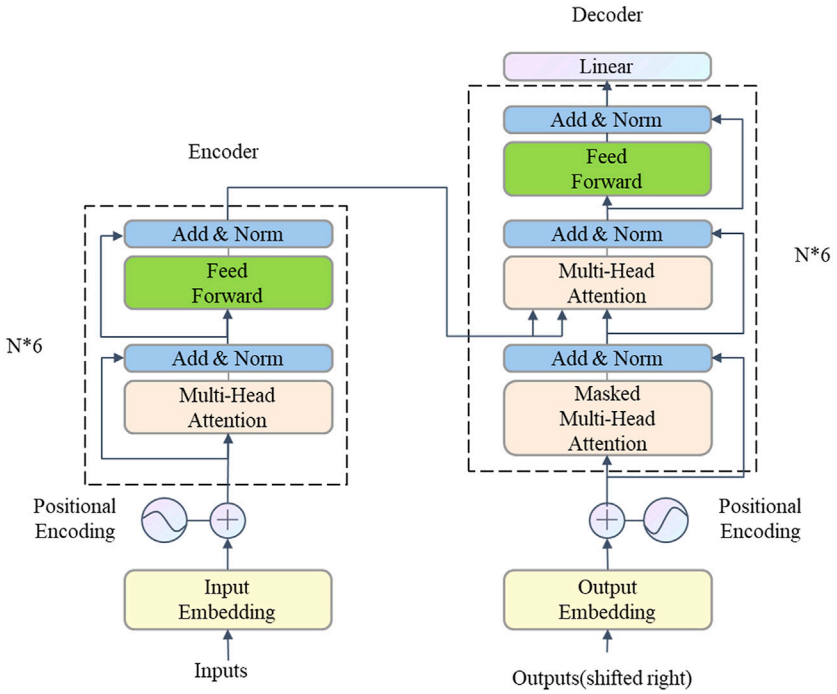

FIGURE 1 | The architecture of BERT.[1]

translation, named entity recognition, reading comprehension, and automatic question answering tasks. Since language models based on transfer training have better performance in natural language processing tasks, this study will use such models to extract the features of rumor content. Specifically, this study uses the post-trained BERT model to extract features from COVID-19 rumor post texts.

BERT is short for Bidirectional Encoder Representations from Transformers proposed by Jacob et al. (2018). Through the mechanism of the transformer network and the transfer learning mechanism, BERT contains vibrant text lexicon information and semantic information. BERT model has been trained on more than 110M corpus and can be directly loaded and used. It is pre-trained by MLM (Masked Language Model) and NSP (Next Sentence Prediction) task. The basic architecture of BERT is shown in **Figure 1**. Rumor text first goes through the BERT tokenizer and creates token embedding, segment embedding, and position embedding in the BERT model. Then the embedding of the text enters the encoder of BERT. The encoder is composed of multi-head attention layers and a feed-forward neural network. After six layers of encoding, the encoded text is embedded into the decoder, composed of a multi-head attention layer and feed-forward neural network. After six layers of decoding, the feature of rumor content is extracted. The multi-head attention mechanism is the critical process to extract text features. It can be formulated as:

$$Q_i = QW_i^Q, K_i = KW_i^k, V_i = VW_i^V$$

$$Head_i = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_K}}\right)V_i$$

$$MultiHead\,(Q,K,V) = Concat\,(Head_1, Head_2, \ldots, Head_n)W^O$$

where Q represents the input of the decoder in a step, the K and V represent the rumor text embedding. W_i^Q, W_i^k, and W_i^V are the weight parameters of Q, K, and V. d_K is the number of dimensions in K to scale the dot product of Q_i and K_i. $Head_i$ represents the output of the i-th attention head layer. W^O is the weight parameters for concatenated outputs. $MultiHead\,(Q,K,V)$ represents the final output of the multi-head attention layer.

Existing studies have shown that post-train on BERT by domain-specific corpus can significantly improve the performance on the natural language processing task in specific domains [25]. In combination with the COVID nine rumor detection task, post-training on BERT was carried out through a COVID 19 Twitter dataset [26] and a COVID19 Weibo dataset [27], respectively. Specifically, we use the MLM task to post-train BERT so that our BERT model contains more semantic and contextual information on COVID 19-related posts from social networks. This study uses BERT and Chinese BERT in the PyTorch version released by Huggingface[2] as our primary model.

[1]The figure is modified based on: Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin "Attention is all you need." In Advances in neural information processing systems, pp. 5998–6008. 2017.

[2]https://github.com/huggingface/pytorch-transformers

An Example of a Rumor Post and Its User Responses:

Twitter Post:

"CDC is preparing for the 'likely' spread of coronavirus in the US, officials say https://t.co/cm9pRyVTcU Do we have anyone left in the CDC who knows what the fuck they are doing," Mon Feb 24, 2020.

User Responses:

– "Georgia Doctor Appointed Head Of The CDC: Health News Dr. Brenda Fitzgerald, who leads the Georgia Department of Public Health, has been appointed CDC director. She'll take over as the Trump administration seeks big cuts to the CDC's budget." Mon Feb 24 10:40:44 + 0000 2020, 0, 1, 1, 49:33.4.

– "@NikitaKitty @PerfumeFlogger We used to. This a travesty.", Mon Feb 24 10:47:35 + 0000 2020, 1, 0, 0, 49:33.4.

–"As we've reported, that would include a $186 million cut to programs at the CDC's center on HIV/AIDS, hepatitis and other sexually transmitted diseases.", Mon Feb 24 10:43:18 + 0000 2020, 1, 2, 2, 49:33.4.

– "The CDC's chronic disease prevention programs, such as those for diabetes, heart disease, stroke and obesity, would be cut by $222 million. What will she do stave the fucking virus?", Mon Feb 24 10:43:18 + 0000 2020, 1, 4, 4, 49:33.4.

After the pre-training of 20 epochs on the COVID-19 dataset, we post-trained the original BERT and original Chinese BERT to the COVID-19 Social Network BERT (CSN-BERT) model.

User Response Feature Extractor

Users on social networks would reply or forward a Twitter, whether it is a true rumor or a false rumor. These responses and retweets contain users' views. Some of these views are to the rumor, and others are to other users' responses or retweets. The user responses and retweets can be considered as discussions or arguments around the rumor. An example of a Twitter post's user responses is shown below. Typically, the responses and retweets can be seen as a tree structure. Rumors and their responses and retweets are called conversational threads. Many studies focused on the tree structure consisting of user responses and retweets and determine rumor veracity based on its structure known as propagation path. However, they do not pay much attention to the content of user responses. Because COVID-19 rumors are more likely to cause panic, there are stricter time constraints for discovering these rumors. In limited time, the structure of responses and retweets, the propagation path, may not be comprehensive enough to determine the veracity of rumors. This indicates that we need to dig into the user responses for essential features on rumor detection.

In this study, we focus on the opinions expressed from user responses. We think of user responses as a sequence, $R_i = \{s_1^i, \ldots s_j^i, \ldots, s_{m_i}^i\}$. The sequence is arranged by response time. To be sure, the original rumor post is not recorded in the sequence. This responses sequence is constructed with time limits. We start with the time of the first responses or retweets and only record responses within 24 h. For the response sequence R_i, we need to extract features from the user response sequence R_i for rumor detection. In order to extract features from the user response sequence, we proposed COVID-19 Response-LSTM (CR-LSTM) to learn about the user response sequences. We implemented the post-trained BERT model (CSN-BERT) mentioned in *Rumor Content Feature Extractor* and a textCNN extractor to learn the sentence embedding of each user response. To be specific, BERT's [CLS] vector is used to represent the feature of user responses. The structure of the entire model is shown in **Figure 2**.

For a user response, its sentence embedding firstly generated through the CSN-BERT. Then, the sentence embedding enters a bidirectional LSTM layer in the order of release time. Each hidden layer in the LSTM layer corresponds to a response, denoted as $\vec{h}_t^i$. After encoding by two LSTM layers, the vector is weighted by an attention layer. We use the multi-head self-attention mechanism to find the responses that have more influence on the results. This process can be represented as:

$$\vec{h}_t^i = \overrightarrow{LSTM}_i\left(\vec{h}_{t-1}^i, e_t^i\right)$$

$$\overleftarrow{h}_t^i = \overleftarrow{LSTM}_i\left(\overleftarrow{h}_{t+1}^i, e_t^i\right)$$

$$L_i = Concat_1^n\left(Softmax\left(\left[\frac{\vec{h}_t^i, \overleftarrow{h}_t^i\right]e_t^i}{\sqrt{d_K}}\right)e_t^i\right)W^O$$

where $\overrightarrow{LSTM}_i$ indicates the encoding operation in the forward direction, and $\overleftarrow{LSTM}_i$ in the backward direction. $\vec{h}_t^i$ represents the forward hidden state of the t-th embedding in R_i, also corresponding to word s_j^i in R_i, which is calculated by its previous hidden state $\vec{h}_{t-1}^i$ and current post sentence embedding e_t^i. $\overleftarrow{h}_t^i$ represents the backward hidden state of the t-th embedding in R_i. The hidden state of the t-th embedding is obtained by concatenating $\vec{h}_t^i$ and $\overleftarrow{h}_t^i$, denoted by $h_t^i = [\vec{h}_t^i, \overleftarrow{h}_t^i]$. L_i is the final embedding of the CR-LSTM. $Concat_1^n\left(Softmax\left(\left[\frac{\vec{h}_t^i, \overleftarrow{h}_t^i\right]e_t^i}{\sqrt{d_K}}\right)e_t^i\right)W^O$ indicates the multi-head self-attention.

The Full View

Combining the rumor content feature extractor and the user response feature extractor, we can extract the integrated rumor feature. For a rumor $r_i = \{x_i, s_1^i, \ldots s_j^i, \ldots, s_{m_i}^i\}$ in a rumor dataset $R = \{r_1, r_2, \ldots, r_n\}$, the rumor content feature extractor (CSN-BERT) can extract the rumor content feature C_i from x_i. The user response feature extractor (CR-LSTM) can extract the user response feature L_i from $\{s_1^i, \ldots s_j^i, \ldots, s_{m_i}^i\}$.

We concatenate the user response feature L_i extracted by CR-LSTM with the rumor content feature C_i extracted by CSN-BERT into the integrated rumor feature. The rumor detection feature then goes through a fully-connected layer dimension, activated by Relu function, and at last output the probability distribution of

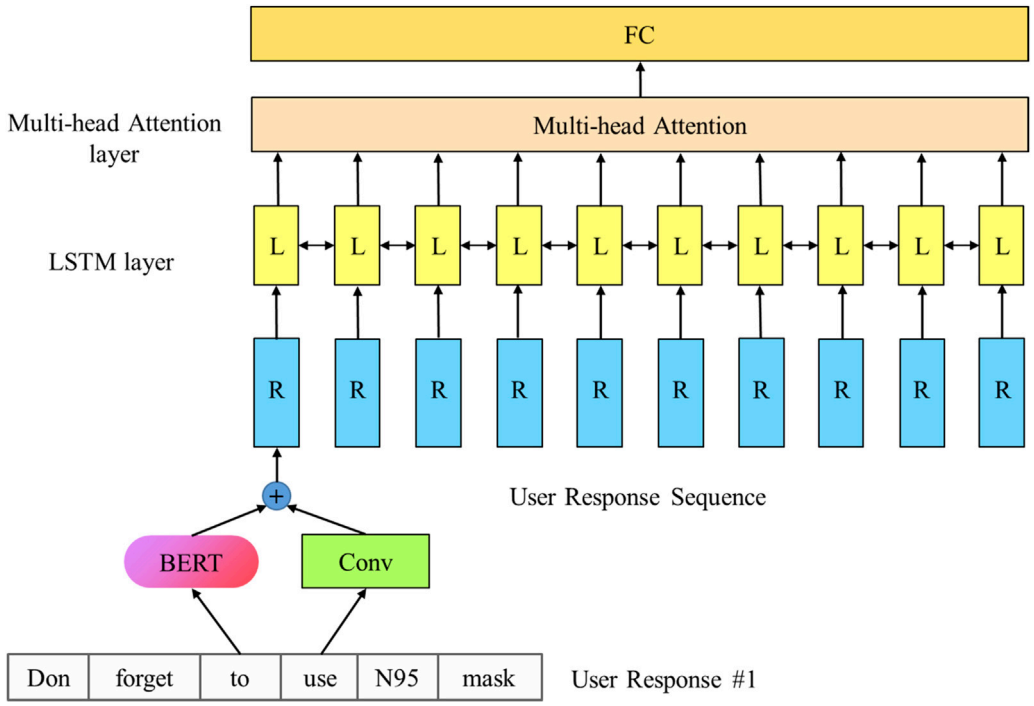

FIGURE 2 | The architecture of CR-LSTM

rumor detection by a Softmax function. The total model is called CR-LSTM-BE (COVID-19 Response LSTM with BERT Embedding). The full view of our model is shown in **Figure 3**.

EXPERIMENTS

In this section, we introduced the experimental preparation and the experimental results. *Dataset* introduces the two datasets for the COVID-19 rumor detection task used in our experiments. *Evaluation Metrics* introduces the evaluation metrics with the computing methods. *Experiment Settings* introduces the experiment settings, especially the hyperparameters selected in the experiments. *Results* presents the experimental results and compares and analyzes the results with baseline methods.

Dataset

To confirm the performance of the CR-LSTM-BE model proposed by us on the COVID-19 rumor detection task. Since there are not many datasets for COVID-19 rumors and considering the data requirements, this study conducted experiments on two datasets. The datasets selected to conduct experiments are the COVID-19 rumor dataset and the CHECKED dataset. The experimental results and related indicators tested the performance of the CR-LSTM-BE model.

The COVID-19 rumor dataset is provided by Cheng et al. [28] and consists of rumors from two types of sources. One is news from various news sites, and the other is from Twitter. There are 4,129 news and 2,705 Twitter posts in this dataset. This study focuses on COVID-19 rumor detection on the social network, so only the Twitter post part of the dataset is selected as the experimental data. The Twitter part of the dataset contains rumor Twitter post id (Hashed), Twitter post content, rumor label (True, False or Unverified), number of likes, number of retweets, number of comments, user responses over a while, user response time and stance of user response. This study mainly used the Twitter post content in the dataset and the user responses of each Twitter post within 24 h to conduct experiments.

The CHECKED data set was provided by Yang et al. [18], and the data came from the Chinese Weibo social network. This dataset contained 2,104 tweets. The dataset contains the rumor microblog's post id (hashed), microblog's post id content, rumor label (True or False), user id (hashed), the time the microblog was posted, number of likes, number of retweets, number of comments, user responses over some time, user retweet over some time, user response time, and user retweet time. This study mainly used the contents of the rumor microblog and the responses and retweets of each microblog within 24 h to conduct experiments. Statistics of the relevant data are shown in **Table 1**. We randomly split the two datasets into the training set, validation set, and testing set with the proportion of 70, 10, and 10%, respectively.

Due to the uncontrollable quality of user response data, we performed resample on the data while preprocessing the user response data. Specifically, we removed user responses that are very concise (less than three words), contain more emoji (over 80%), and have only one hyperlink without other information.

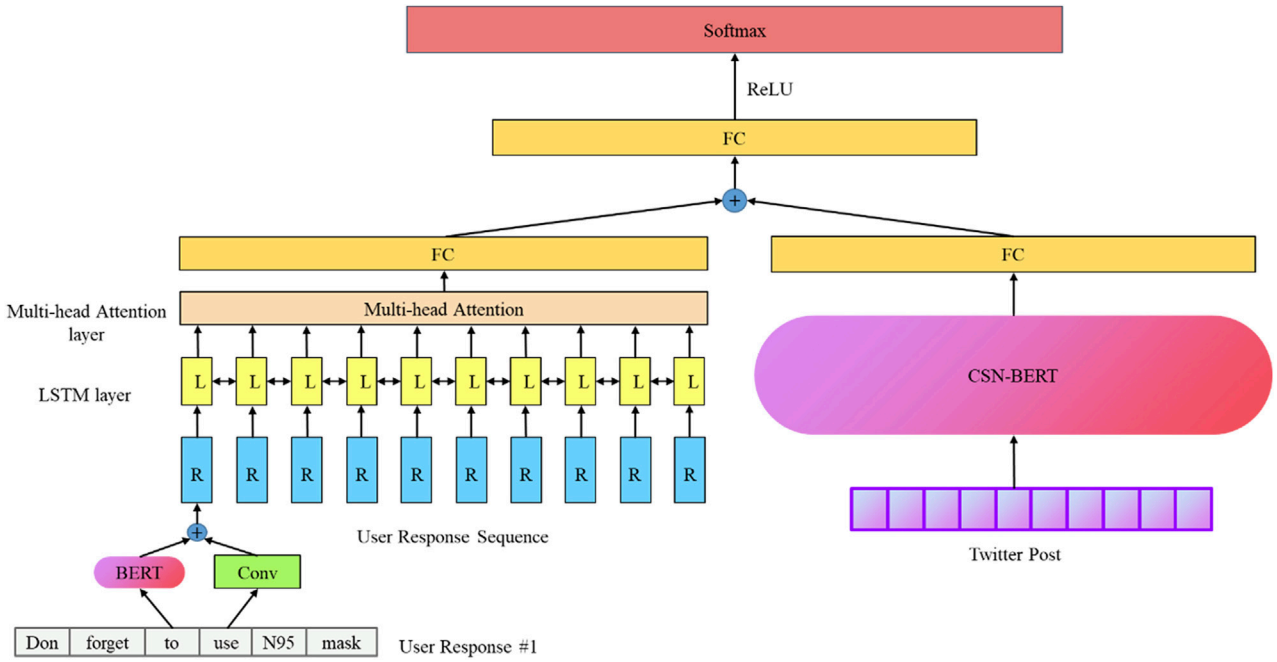

FIGURE 3 | The full view of CR-LSTM-BE.

Evaluation Metrics

The evaluation metric followed most of the existing studies, which regards rumor detection as a classification task. We used the Macro F1, precision score, recall score, and accuracy to evaluate the performance of our model. Macro F1 is used because the labels of rumor posts are imbalanced, which means the distribution is skewed. Marco F1 allows us to evaluate the classifier from a more comprehensive perspective. The precision and recall score in our evaluation is also macro. The definitions of precision, recall, Marco F1, and accuracy are shown below:

$$Precision_c = \frac{TP_c}{TP_c + FP_c}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c}$$

$$\mathrm{F1}_c = \frac{2*(Recall_c * Precision_c)}{Recall_c + Precision_c}$$

$$\text{Marco F1} = \sum_{c=1}^{n} \mathrm{F1}_c / n$$

$$\text{Accuracy} = \frac{Correct\ Predictions}{all\ samples}$$

where c is the label of a rumor, which could be True, False, or Unverified. TP_c stands for the true positives of rumor label c, which means that the actual label of this rumor is c, and the predicted label is also c. FP_c stands for the false positives, which means that the actual label of this rumor is not c, but the predicted one is c. FN_c stands for false negatives, which means that the actual label c, but the predicted label is not c. Macro F1 was used to integrate all $\mathrm{F1}_c$.

Experiment Settings

In our experiments, we fine-tuned the CSN-BERT on rumor veracity classification task. To prevent overfitting, we disabled backpropagation of CSN-BERT while training the CR-LSTM-BE model. We implemented our model by Pytorch, and the bias was initialized to 0. We used the dropout mechanism to prevent the model from quickly overfitting, the dropout rate was set to 0.5. Random Search method [29] was used to find the optimum hyperparameters. For post training the BERT model and fine-tuning the CSN-BERT, AdamW optimizer [30] was applied with an initial learning rate 1e-5 for model updating, and a mini-size batch of 16 was set. Early stopping is used, and the patience was set to five epochs. In the CR-LSTM-BE model, the optimum number of RNN layers is one and the optimum hidden size is 512. the one optimum number of attention head is 8, and the optimum attention size is 512. We used the Word2vec [31] embedding to initialize word embedding vectors in the textCNN part of the CR-LSTM-BE model, the word embedding vectors were pretrained on English corpus provided by Google. The dimension of word embedding vector was set to 300. For training the CR-LSTM-BE model, Adam optimizer [32] was applied with an initial learning rate 1e-3 for model updating, and a mini-size batch of 16 was set. Early stopping is used, and the patience was set to 15 epochs. All the experiments were done on a GeForce TITAN X.

RESULTS

The datasets adopted in this study do not provide detailed rumor detection results based on different methods. The COVID-19 rumor dataset provides rumor detection results for all data,

TABLE 1 | Statistics of the datasets.

	COVID-19-rumor dataset	CHECKED dataset
Sentence per tweets	1.39	4.76
Words per sentence	11.4	25.96
Words per tweets	15.87	123.67
Total words	42,939	260,197
Total tweets	2,705	2,104
Total responses	34,963	2997063

including news and Twitter data. However, only the Twitter dataset was used in this study. The CHECKED dataset includes benchmark results of FastText, TextCNN, TextRNN, Att-TextRNN, and Transformer methods, but the test only gives Macro F1 score, which lacks more specific indicators such as accuracy and F1 scores on different labels. In order to compare and analyze the performance of our model. We set up several baseline methods based on rumor content features. Referring to related studies and the CHECKED dataset, baseline methods in this study include SVM classifier with word bags, textCNN with word2vec embedding, TextRNN with word2vec embedding, AttnRNN with word2vec embedding, Transformer with word2vec embedding, and BERT-base. We used the Word2Vec embedding pretrained on the English corpus published by Google and the Word2Vec embedding pretrained on the Chinese corpus published by Sogou.

We repeatedly conducted experiments with each method ten times in our study. With the results of the ten experiments, the median of Macro F1 in each group was selected as the experimental results for comparison. We conducted the t-test to confirm if the proposed model performed significantly differently from the baseline methods. The results of the t-test show a significant improvement (p-value<0.05) between CSN-BERT and the baseline methods, CR-LSTM-BE and the baseline methods, and CR-LSTM-BE and CSN-BERT. The experimental results of this study in the COVID-19 rumor dataset are shown in **Table 2**. According to the experimental results, the best-performed method in the baselines is the BERT-base, of which the precision, recall, Marco F1, and accuracy score achieved 55.22, 55.53, 55.34, and 55.42, respectively. In our methods, the post-trained CSN-BERT model showed significant improvement on the data set. Its precision, recall, Marco F1, and accuracy score achieved 58.47, 58.64, 58.55, and 58.87, respectively. Compared to the best-performed baseline, the CSN-BERT showed a 5.8% improvement on Macro F1. The CR-LSTM-BE method based on rumor content feature and user responses proposed in this study has achieved the best performance in the COVID-19 rumor dataset. The precision, recall, Marco F1, and accuracy score of the CR-LSTM-BE achieved 63.15, 64.39, 63.64, and 63.42, respectively. Compared to the best-performed baseline, the CR-LSTM-BE improves 15.0% on Macro F1. Compared to the post-trained CSR-BERT method, this is an 8.7% improvement on Macro F1.

TABLE 2 | Performance on the COVID-19 rumor twitter dataset.

Methods	Precision	Recall	Macro F1	Accuracy
WB-SVM	43.20	43.50	43.33	43.35
textCNN	52.87	52.82	52.80	53.45
textRNN	51.18	51.83	51.45	51.35
attnRNN	51.79	53.04	52.23	51.97
Transformer	52.85	52.63	52.72	52.22
BERT-base	55.22	55.53	55.34	55.42
CSN-BERT	58.47	58.64	58.55	58.87
CR-LSTM-BE	63.15	64.39	63.64	63.42

TABLE 3 | Performance on the CHECKED dataset.

Methods	Precision	Recall	Macro F1	Accuracy
WB-SVM	62.35	69.21	62.33	70.09
textCNN	81.99	89.08	84.76	89.87
textRNN	71.57	81.00	73.77	80.54
attnRNN	81.98	91.44	85.32	89.87
Transformer	84.84	92.36	87.82	91.93
BERT-base	95.74	98.16	96.89	98.10
CSN-BERT	97.13	99.32	98.18	98.89
CR-LSTM-BE	100.00	100.00	100.00	100.00

The experimental results of this study on the CHECKED dataset are shown in **Table 3**. According to the experimental results, the best-performed method in the baselines is the BERT-base, of which the precision, recall, Marco F1, and accuracy score achieved 95.74, 98.16, 96.89, and 98.10, respectively. In our methods, the precision, recall, Marco F1, and accuracy score of the post-trained CSN-BERT model achieved 97.13, 99.32,98.18, and 98.89, respectively. Compared to the best-performed baseline, the CSN-BERT slightly improved Macro F1 (1.3%). The CR-LSTM-BE method based on rumor content feature and user responses proposed in this study has achieved the best performance in the CHECKED dataset. The precision, Recall, Marco F1, and accuracy score of the CR-LSTM-BE all achieved 100. Compared to the best-performed baseline, the CR-LSTM-BE improves 3.2% on Macro F1. Compared to the post-trained CSN-BERT method, this is a 1.9% improvement on Macro F1.

DISCUSSION

In this section, we discussed the performance and the characters of our proposed models. *Improvement Analysis* analyzes the improvements of CSN-BERT and CR-LSTM-BE compared with the baseline methods. Number of Responses Analysis analyzes the effect of the number of responses to rumor detection.

Improvement Analysis

Among the methods experimented in this study, CSN-BERT has a particular improvement than the baseline methods according to the experimental results, which indicates CSN-BERT has a better performance in the feature representation of rumor content by post-training COVID-19 twitter dataset than the original BERT (BERT-base). Compared with general deep learning models (such as textCNN and LSTM), it is not surprising that the BERT model, which is based on transfer learning, performs better in the

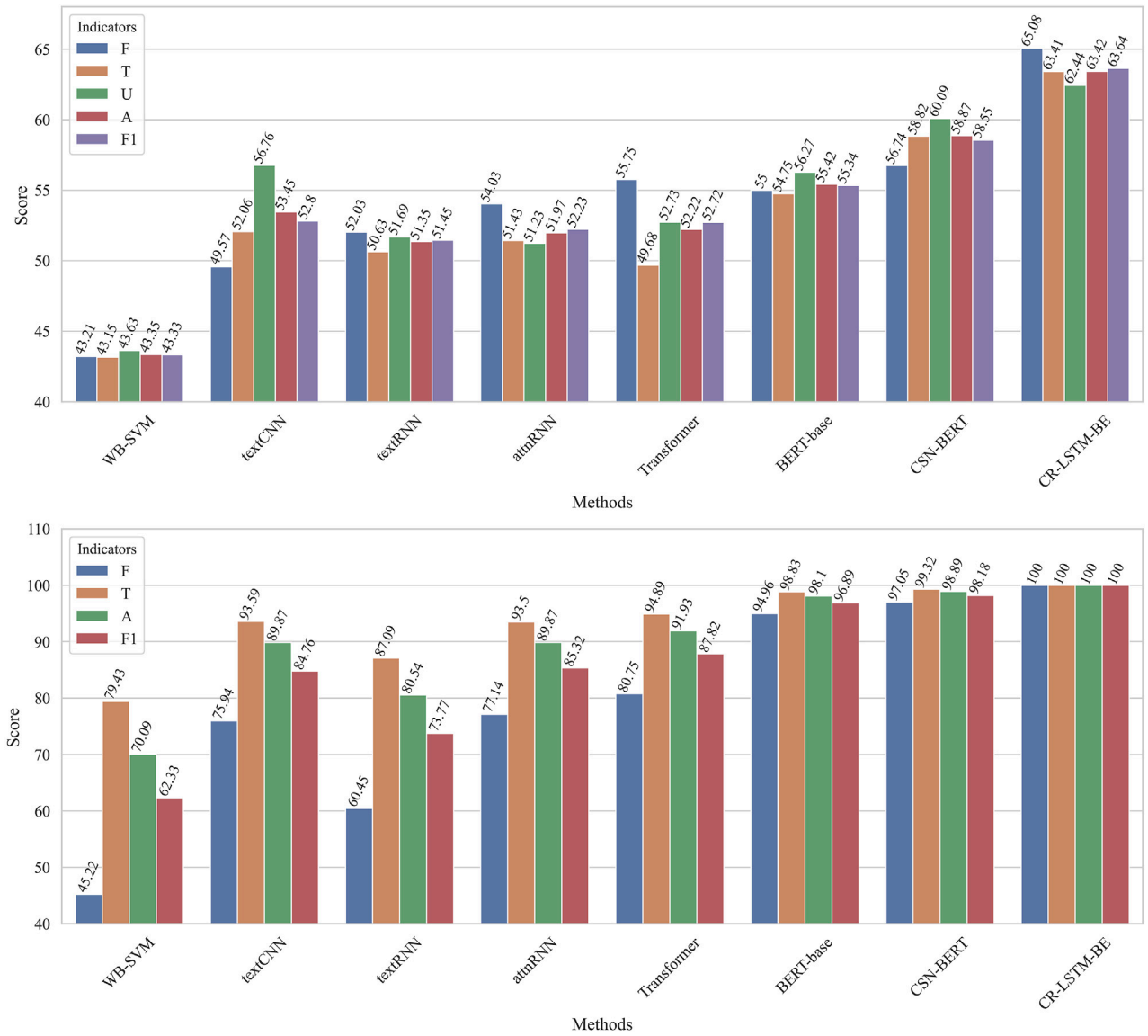

FIGURE 4 | The performance of methods on different rumor labels (F: False, T: True, U: Unverified, A: Accuracy, F1: Macro F1).

problem of rumor detection because the model is based on transfer training has more contextual semantic information—continuing with the idea of allowing the model to acquire more contextual semantic information, CSN-BERT allowing the BERT model to learn more information on COVID-19 discussed by users in the social network in advance. Compared with the original BERT, BERT after post-training is more suitable for COVID-19 rumor detection.

The CR-LSTM-BE proposed in this study adds user responses information into the deep learning model and encodes user responses through the LSTM network with multi-head attention. Use responses contains much information to the original twitter post [33]. In our hypothesis, adding user responses into the model can provide richer information standing for user feedback for the learning process and enable the model to determine the veracity of rumors based on user feedback. The experimental results show that CR-LSTM-BE achieves the best results on both datasets. The experimental results confirmed our hypothesis. In **Figure 4**, we compare the F1 scores of all methods on the various rumor labels (F: False, T: True, U: Unverified). The legend "A" in **Figure 4** is the accuracy, and legend "F1" is the Macro F1. It can be seen that the F1 score on each rumor label of CR-LSTM-BE is better than other methods. In addition, this method can still get a more balanced classification result from unbalanced training data.

Number of Responses Analysis

Interactions on social networks could help better represent user profiles. More user responses can be seen as connections on social network and will provide richer information to describing an

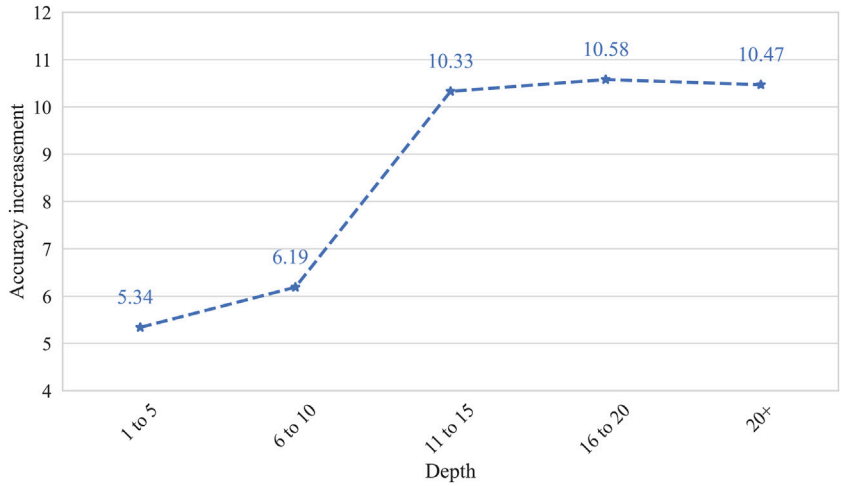

FIGURE 5 | The accuracy increasement of twitter post with various number of responses.

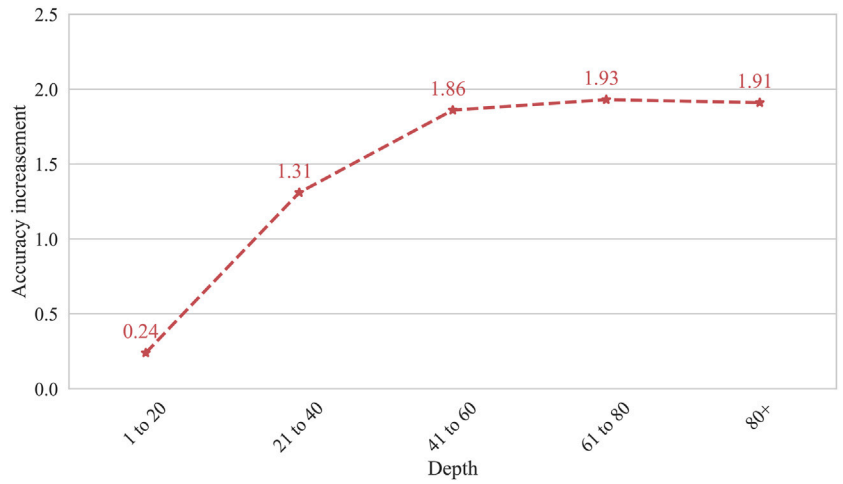

FIGURE 6 | The accuracy increasement of microblog post with various number of responses.

event from a more abundant perspective [34–39]. To further understand the effect of user responses on rumor detection, we compared the accuracy of a different group of Twitter and microblog posts with various responses within 24 h. **Figure 5** shows the rumor detection accuracy improvements of a different group of Twitter and microblog posts with various responses tested on CR-LSTM-BE and CSN-BERT. While the number of user responses is 0, CR-LSTM-BE will degenerate into CSN-BERT, and the accuracy will not be improved. As shown in **Figure 5**, while the number of user responses is 1–5, the accuracy of rumor detection increased by 5.34%. While the number of user responses is 6–10, the accuracy of rumor detection increased by 6.19%. While the number of user responses is more than 11, the accuracy improvement of rumor detection is stabilized at about 10%. This indicates that we should consider including more than 11 user responses for COVID-19 rumor detection on Twitter. For Weibo, due to a large number of retweets and responses, we use another category scheme in the division of the number of user responses. As can be seen from **Figure 6**, the curve of accuracy promotion is similar to that of Twitter (**Figure 5**). While the number of user responses is more than 41, the improvement of rumor detection accuracy tends to be stable. This suggests that we should consider including more than 41 user responses for COVID-19 rumor detection on Weibo.

CONCLUSION

In this study, we proposed rumor detection methods based on the features of rumor content and user responses because of the rapid propagation and prominent domain characteristics of COVID-19 rumor detection on social networks. In order to better capture

and extract rumor content features, we combined the language model based on transfer learning with a post-training mechanism to construct CSN-BERT based on COVID-19 user posts on social networks. In order to make better use of the information in user responses, we further proposed CR-LSTM-BE, which incorporated the information of user responses into the learning process through LSTM. The experimental results show that the post-trained CSN-BERT model can better extract the content features of COVID-19 rumors on social networks than other deep learning models. The CR-LSTM-BE model that integrates user responses achieves the best performance on both datasets. In addition, we found that more user responses can help the CR-LSTM-BE model to achieve better results. On the Twitter network, more than 11 user responses can help to achieve the best performance. On the Weibo network, more than 41 user responses can help to achieve the best performance.

This study focuses on exploring the enhancement of user responses information on rumor detection. Limited by the experimental data, this study did not consider the structural features of user responses and retweets, known as propagation path. Future research will focus on the structural features of user response and retweets and implementing deep learning methods to implement rumor detection better. One direction is to utilize the GCN or hierarchical attention model to incorporate and extract structural and user response features simultaneously.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: DATASET1: https://github.com/MickeysClubhouse/COVID-19-rumor-dataset DATASET2: https://github.com/cyang03/CHECKED

AUTHOR CONTRIBUTIONS

JY and YP conceived and designed the study. JY and YP conducted the experiments. YP reviewed and edited the manuscript. All authors read and approved the final manuscript.

REFERENCES

1. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID-19 social media infodemic. *Sci Rep* (2020) 10(1): 16598–10. doi:10.1038/s41598-020-73510-5
2. Mazzeo V, Rapisarda A, and Giuffrida G. Detection of Fake News on COVID-19 on Web Search Engines. *Front Phys* (2021) 9:685730. doi:10.3389/fphy.2021.685730
3. Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, et al. Rumor detection on social media with bi-directional graph convolutional networks. *Aaai* (2020) 34(01):549–56. doi:10.1609/aaai.v34i01.5393
4. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, et al. Eann: Event adversarial neural networks for multi-modal fake news detection. In: *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. Stroudsburg: ACL Press (2018). p. 849–57.
5. Devlin J, Chang M, Kenton L, and Kristina T, Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HTL*. Stroudsburg: ACL Press (2019). p. 4171–86.
6. Brown TB., Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. "Language models are few-shot learners." arXiv [Preprint].14165 (2020). Available at https://arxiv.org/abs/2005.14165 (Accessed September 20, 2021).
7. Qazvinian V, Rosengren E, Radev D, and Mei Q. Rumor has it: Identifying misinformation in microblogs. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg: ACL (2011). p. 1589–99.
8. Kochkina E, Liakata M, and Zubiaga A. "All-in-one: Multi-task learning for rumour verification." arXiv preprint arXiv:1806.03713 (2018).
9. Cao J, Guo J, Li X, Jin Z, Guo H, and Li J. "Automatic rumor detection on microblogs: A survey." arXiv [Preprint] (2018). Available at https://arxiv.org/abs/1807.03505 (Accessed September 20, 2021).
10. Ma J, Gao W, and Wong K. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. *Proc 55th Annu Meet Assoc Comput Linguistics* (2017) Vol. 1:708–17.Long Papers)
11. Castillo C, Mendoza M, and Poblete B. Information credibility on twitter. *Proc 20th Int Conf World wide web* (2011) 675–84. doi:10.1145/1963405.1963500
12. Chua AYK, and Banerjee S. Linguistic predictors of rumor veracity on the internet. *Proc Int MultiConference Eng Comp Scientists* (2016) 1:387–91.
13. Yu F, Liu Q, Wu S, Wang L, and Tan T. A Convolutional Approach for Misinformation Identification. *IJCAI* (2017) 3901–7. doi:10.24963/ijcai.2017/545
14. Vosoughi S, Mohsenvand MN, and Roy D. Rumor Gauge. *ACM Trans Knowl Discov Data* (2017) 11:1–36. doi:10.1145/3070644
15. Liu Y, and FangWu YB. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: *32nd AAAI Conference on Artificial Intelligence*. California: AAAI press (2018). p. 354–61.
16. Kwon S, Cha M, and Jung K. Rumor Detection over Varying Time Windows. *PloS one* (2017) 12:e0168344–1. doi:10.1371/journal.pone.0168344
17. Glazkova A, Glazkov M, and Trifonov T. g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection. *Commun Comput Info Sci* (2021) 116–27. doi:10.1007/978-3-030-73696-5_12
18. Yang C, Zhou X, and Zafarani R. CHECKED: Chinese COVID-19 fake news dataset. *Soc Netw Anal Min* (2021) 11:58–8. doi:10.1007/s13278-021-00766-8
19. Patwa P, Sharma S, Pykl S, Guptha V, Kumari G, Akhtar MS, et al. Fighting an infodemic: COVID-19 fake news dataset. In: *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Cham: Springer (2021) p. 21–9. doi:10.1007/978-3-030-73696-5_3
20. Li Q, Zhang Q, Luo S, and Liu Y. Rumor Detection on Social Media: Datasets, Methods and Opportunities. In: *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship*. Snyder: Disinformation, and Propaganda (2019) p. 66–75. doi:10.18653/v1/d19-5008
21. Zubiaga A, Kochkina E, Liakata M, Procter R, Lukasik M, Bontcheva K, et al. Discourse-aware rumour stance classification in social media using sequential classifiers. *Inf Process Manag* (2018) 54(2):273–90. doi:10.1016/j.ipm.2017.11.009
22. Zhou X, and Zafarani R. "Fake news: A survey of research, detection methods, and opportunities." arXiv [Preprint] (2018). Available at https://arxiv.org/abs/1812.00315v2 (Accessed September 20, 2021).
23. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. *Proc NAACL-HLT* (2018) 2227–37. doi:10.18653/v1/n18-1202
24. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, and Quoc V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv Neural Inf Process Syst* (2019) 32:5753–63.
25. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (2020) 36(4):1234–40. doi:10.1093/bioinformatics/btz682
26. Lamsal R. Design and analysis of a large-scale COVID-19 tweets dataset. *Appl Intell* (2021) 51(5):2790–804. doi:10.1007/s10489-020-02029-z

27. Leng Y, Zhai Y, Sun S, Wu Y, Selzer J, Strover S, et al. Misinformation during the COVID-19 outbreak in China: Cultural, social and political entanglements. *IEEE Trans Big Data* (2021) 7(1):69–80. doi:10.1109/tbdata.2021.3055758
28. Cheng M, Wang S, Yan X, Yang T, Wang W, Huang Z, et al. A COVID-19 Rumor Dataset. *Front Psychol* (2021) 12(2021):644801. doi:10.3389/fpsyg.2021.644801
29. Bergstra J, and Bengio Y. Random Search for Hyper-Parameter Optimization. *J Machine Learn Res* (2012) 13:281–305.
30. Loshchilov I, and Hutter F. *Fixing weight decay regularization in adam*. arXiv [Preprint] (2012). Available at https://arxiv.org/abs/1711.05101 (Accessed September 20, 2021).
31. Mikolov T, Chen K, Corrado G, and Dean J. *Efficient estimation of word representations in vector space* (2013). arXiv preprint arXiv:1301.3781.
32. Kingma DP, and Jimmy B. *Adam: A method for stochastic optimization* (2014). arXiv preprint arXiv:1412.6980.
33. Tuzón P, Fernández-Gracia J, and Eguíluz VM. From Continuous to Discontinuous Transitions in Social Diffusion. *Front Phys* (2018) 6:21. doi:10.3389/fphy.2018.00021
34. Omodei E, De Domenico M, and Arenas A. Characterizing interactions in online social networks during exceptional events. *Front Phys* (2015) 3:59. doi:10.3389/fphy.2015.00059
35. Bellingeri M, Bevacqua D, Scotognella F, Alfieri R, Nguyen Q, Montepietra D, et al. Link and Node Removal in Real Social Networks: A Review. *Front Phys* (2020) 8:228. doi:10.3389/fphy.2020.00228
36. Lou J, Xu Z, Zuo D, Zhang Z, and Ye L. Audio Information Camouflage Detection for Social Networks. *Front Phys* (2021) 9:715465. doi:10.3389/fphy.2021.715465
37. Bu Z, Li H, Zhang C, Cao J, Li A, and Shi Y. Graph K-means based on leader identification, dynamic game, and opinion dynamics. *IEEE Trans Knowledge Data Eng* (2019) 32(7):1348–61.
38. Wang Z, Xia C, Chen Z, and Chen G. Epidemic Propagation with Positive and Negative Preventive Information in Multiplex Networks. *IEEE Trans Cybern* (2020) 51 (3):1454–62. doi:10.1109/TCYB.2019.2960605
39. Wang Z, and Xia C. Co-evolution Spreading of Multiple Information and Epidemics on Two-layered Networks Under the Influence of Mass Media. *Nonlinear Dyn* (2020) 102:3039–52. doi:10.1007/s11071-020-06021-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.