Fake News Detection by Multichannel CNN-GRU with Attention

Sanjana Balagar
Department of Applied Data Science
San José State University
sanjana.balagar@sjsu.edu

Xueliu Fan
Department of Applied Data Science
San José State University
xueliu.fan@sjsu.edu

***Abstract*—The information explosion in the media has raised difficulties to detect true news from the fake one. The researchers have applied neural network approaches to classify text and help improve the information ecosystem such as recurrent neural network (RNN) and convolutional neural network (CNN). Though RNN can store the previous information in a sequence and CNN can capture the features and patterns, it is a challenge to integrate both advantages for text classification. In this paper, we proposed a multichannel CNN-GRU with attention model to detect fake news from the news dataset with varied topics. The CNN channels can extract the bigram, trigram, and four-gram features from the text and the bi-directional GRU channel can learn the phrase relationships between the words. Moreover, the attention can extract the most relevant element from GRU output. This hybrid model can process the data in both parallel and sequentially. The experiment results show that the multichannel CNN-GRU with attention model achieved high accuracy in detecting fake news and outperformed all the baseline models.**

***Index Terms*—classification deep learning multichannel model**

I. Introduction

A. Motivation

Human decision-making depends on information, which also affects daily habits. Early information exchanges took place in interactive contacts during normal discussions or were obtained via traditional media (e.g., books, newspapers, radio, and television). Such information is more reliable since it has either been independently reviewed or is under official control [1]. People today are exposed to a vast amount of information from several sources (such as web pages, blogs, and postings), especially given the popularity of social media and the Internet. In the same way that propaganda, satire, and false information have all been around for a very long time, so does the idea of fake news. Therefore, information that cannot be verified, lacks sources, and could not be factual is regarded as fake news.

In the twenty-first century, fake news's impact substantially increased, and the amount of information that is continually being added to the Internet has grown to unimaginable heights over time, making it possible for anyone to publish a great deal of unwanted, inaccurate, and misleading content on it [2]. Since such bogus news occasionally has fatal consequences, recognizing it immediately is essential to maintaining the credibility of online social networks. Fake news is everywhere from being distributed through emails; it has even started to evolve to target social media. This inspired us to detect fake news using artificial intelligence.

B. Background

Fake news is identified as information that is false or misleading in any publishing format [1]. It maliciously produces news, poses serious threats to society, undermines public confidence, influences people's decisions, and could even result in major catastrophes on a large scale.

Controlling the spread of false information is never simple. The determination of information reliability is a crucial issue for anti-misinformation. In the era of traditional journalism, respectable publishers and the refereeing procedure maintain the legitimacy of the information. The trustworthiness of information today varies greatly and might be real, false, or of varying degrees of reliability. Users of all backgrounds, whether famous or not, are allowed to post and share nearly any information online.

Traditional fake news detection relies on manual review. To reduce the bias and the cost, artificial intelligence is applied for automatic checking. Moreover, a multidisciplinary approach is used to evaluate the suitability (truthfulness / credibility / veracity / authenticity) of claims in a piece of information, according to a broad definition of disinformation detection, such as data mining, social media pragmatics, linguistic analysis, psychological experiments, and natural language processing (NLP) [3].

Despite decades of research, fake news identification is still a difficult challenge due to the variety of news contents and formats. Deep learning enhanced NLP has been developed for text categorization and classification and are widely used in daily life, including the classification of movie reviews, subjective and objective gradations of sentences, and the aiding of real-world decision-making [4]. Particularly, long short-term memory networks (LSTM) and convolutional neural networks (CNN) are frequently utilized in text classification. Deep learning has been proven to be a powerful tool for text classification.

II. Literature Review

Lin proposed a text classification model based on the hybrid of CNN and LSTM and developed its variant model without activation function in CNN [5]. Both models were trained and tested using 5,000 Rotten Tomatoes movie reviews dataset. They claimed the model without activation function outperformed the CNN-LSTM model and the conventional CNN or LSTM with precision of 99.1%, recall of 99.3%, and F1 score of 99.2%.

RNN-based models and CNN-based models can extract different information from input. In this paper, the author used CNN and RNN to obtain the short texts which are fed into a two layer feedforward neural network for classification. The model was trained and tested on different datasets separately. The model on Switchboard Dialog Act Corpus showed the highest accuracy of 84.0% [6].

An efficient model is proposed to improve text classification. It combined LSTM to extract features and introduced an attention mechanism to calculate the probability distribution from inputs and emphasize the effect of input on the output. The model was trained by 9,423 microblogs and tested by 2,426 microblogs and compared to traditional RNN, CNN, and LSTM. The model had 85% precision, 88% recall rate, and 86% F1 score [7].

Though previous work has shown the strength of text classification, few research worked on multi-topic datasets or large datasets. The simple model has high efficiency but not promising accuracy while the hybrid models on the large and complex dataset are time-consuming. In this work, we proposed an efficient deep-learning model for fake news detection. It can classify news into various topics. Multichannels are introduced to recognize the patterns behind the news. The embedding technique and the attention mechanism are applied to improve the model. We also compare this model with recurrent neural network-based models and convolutional neural network-based models to prove a better performance.

III. Data and Features

A. Data collection

In our project, we have collected data from three different domains namely entertainment, healthcare, and politics. The overall dataset size of the combined dataset is around 300MB with 105,000 rows. There are multiple features like title, context, web info, etc.

The entertainment dataset (FakeNewsNet) [8] is from a fake news data repository by Arizona State University. FakeNewsNet had multidimensional information including social, spatiotemporal, and news and labeled as fact or gossip. There are 23,196 news in the dataset that consists of the title of the article,news_url which contains the URL of the article, source domain which has the web domain where the article was posted,tweet_num which has the number of retweets for the articles, and finally the real column which is the column of the labels containing 0 and 1.

The healthcare dataset [9] with 10,179 news is from Fake News on COVID-19 Healthcare data repository by the National Institute of Technology Surat. They collected news from public websites such as CNN, BBC News, and The Atlantic.

The political dataset [10] (WELFake) contains 72,134 news stories, 35,028 of which are true and 37,106 false. To avoid bias, the authors combined four well-known news datasets (Kaggle, McIntire, Reuters, and BuzzFeed Political).

B. Data prepossessing

For preprocessing our data we have done several steps which include removing url/html, removing special characters, and punctuation, standardizing letters, removing stopwords, converting to root words, and removing missing values. These steps are performed using python. These steps are very much necessary to get good accurate results.

C. Data exploration

We performed exploratory data analysis to get key insights from the data.

Figure 1 demonstrates the amount of class imbalance in our data. We can see that the politics dataset didn't have any class imbalance whereas the other two datasets had a lot of data imbalance. So when we combined the dataset we got a fair amount of class imbalance which adds complexity to our data.

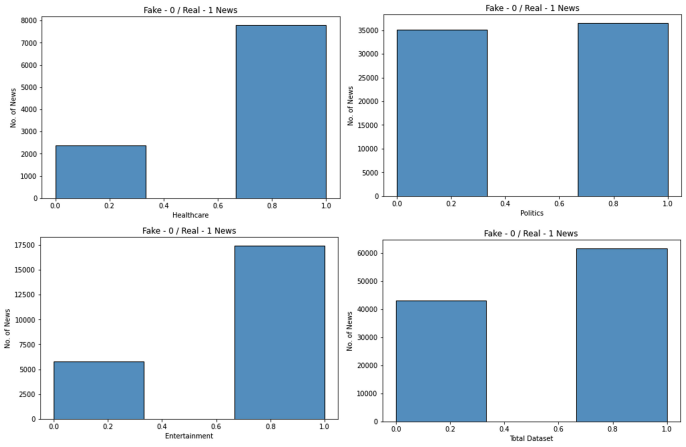

Fig. 1. Histogram of class distribution on fake/true news in politics, entertainment, healthcare, and mixed datasets

The most frequent words that are occurring in both fake and true news are shown in figure 2. This analysis gave us an understanding of how each word contributes to the news being fake or real. The histogram plot shows the distribution of popular words. For example, words like 'say' and 'trump' are in both sets and also occur the highest number of times.

N-gram analysis is conducted to illustrate the contiguous sequences. A text document that has n consecutive objects, including words, numbers, symbols, and punctuation, is known as an n-gram. Many text analytics applications where word sequences are important, such as sentiment analysis, text categorization, and text production, benefit from using N-gram

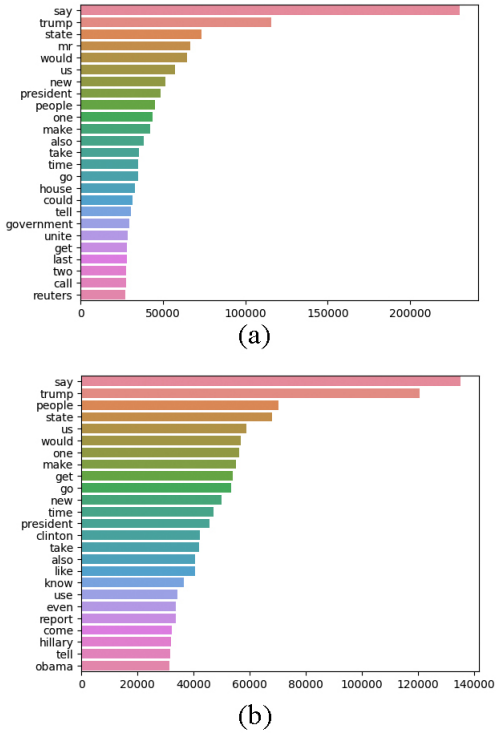

Fig. 2. Most popular words in a) fake news, and b) true news

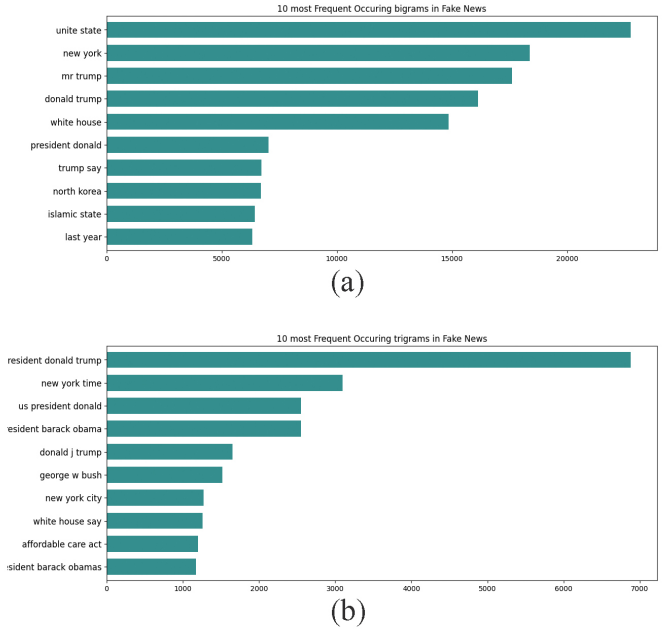

Fig. 3. Bigram and trigram analysis

models. Trigram gives us an understanding of 3 consecutive words which are used frequently in the data. It helps recognize the pattern and the phase relationship between the words. That will be reflected when we design the models.

In figure 3 we have plotted bigrams and trigrams where we can see that for bigrams we have 'unite state' as the most occurring 2 words together and for trigrams, we have 'president Donald Trump'.

The data exploration results indicate our dataset has multi-topics and multi-dimensional information.

IV. Methods

The proposed model is multichannel CNN-GRU with self-attention as illustrated in figure 4. It is inspired by the idea that the recurrent neural network can reserve past information of sequential data and CNN can extract feature maps from data.

A. Input layer

The input data are pre-processed text. Each row contains several sentences. The text data are encoded and tokenized, then padded to the fixed length, that is the maximum length of input sequences 236. Instead of converting the words to vectors randomly, encoding the words from a large corpus can make the model more generic. In this project, the unsupervised learning algorithm GloVe is applied to represent the words. It constructs the word matrix by learning the global word-word co-occurrence from 1.9 million vocabularies [11]. The size of the vocabulary is 10,001, and the embedding dimension is 300.

The architecture of the innovated multichannel CNN-GRU with attention model is demonstrated on the right of figure *. It consists of three convolutional neural networks and one bi-directional gated recurrent unit with self-attention.

B. Convolutional neural network

Convolutional neural network (CNN) can extract features map by sliding convolutional filers over the input data. The filter size and numbers are designed to capture the information and patterns from the data regardless of position [12].

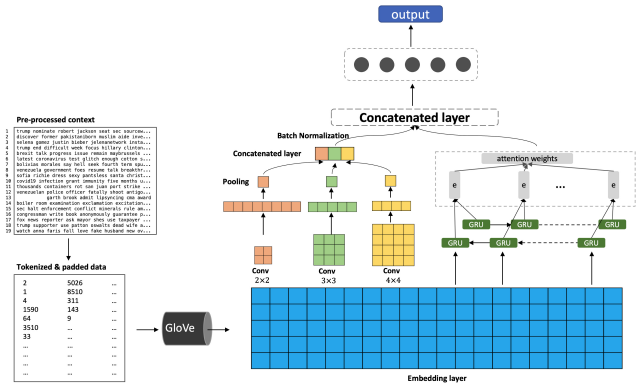

Fig. 4. Multichannel CNN-GRU with attention model architecture

Assume the sentences in the input embedding layer is represented as $C = [c_1, c_2, ..., c_n] \in R^{n+d}$, n is the number of words, and d is the embedded dimension for each word. In the convolutional operation, a filter w with the kernel size of $h \times h$ is sliding over the input data with 1 stride and

generates a feature map $O = [o_1, o_2, ..., o_{n-h+1}]$, in which $o_i = f(w \cdot c_i + b)$ and b is bias [4].

After convolutional operation over the entire embedding data, the output is passed into the max pooling layer to extract the dominating features.

In the MCNN-GRU-attention, three sizes of filters are applied to the embedding layers: 2×2, 3×3, 4×4. It aims to find the patterns of bigram, trigram, and four-gram from the sentences. Extracted feature maps from three filters and the pooling layer are concatenated into the vector.

C. Gated recurrent unit

In nature language processing, the recurrent neural network can process the sequential data by reserving the past words' information and passing it to the hidden state for prediction [13]. As one kind of RNN, GRU can transfer the selected past information sequentially as follows [4]:

$$z_t = \sigma(W^z \cdot [h_{t-1}, x_t]) \tag{1}$$

$$r_t = \sigma(W^r \cdot [h_{t-1}, x_t]) \tag{2}$$

$$z_t = tanh(W \cdot [r_t \cdot h_{t-1}, x_t]) \tag{3}$$

Where z_t is the update gate, r_t is the reset gate, and h_t is the hidden state at time step t. σ is sigmoid activation function. Using this method, we can avoid gradient exploding or vanishing for long sentences. The bidirectional GRU processes the sentences forward and backward to learn the position relationship between the words.

D. Attention layer

The self-attention layer is added to enhance GRU as RNN-based models have memory constraints and limitations. It will weigh the most relevant element in the sequences. A self-attention mechanism can be described as mapping the input X from GRU output into query Q and key K, and value V. The similarity between the query and the key is calculated to obtain the scores for each input data. The SoftMax function is applied to calculate the weights from scores. Finally, the weights multiply the values and sum up to obtain the scalar output [4].

$$Attention(Q, K, V) = softmax(\frac{QK^t}{\sqrt{d_k}})V \tag{4}$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$

The output from attention layer is merged with the output from multichannel CNN and passed to the fully connected layer and the last dense layer with sigmoid activation function.

E. Modeling training parameters

In the multichannel CNN-GRU with attention model, the adjustable hyperparameters include kernel size k and filter numbers n in CNN channels, the hidden state size h in GRU channels, the learning rate, training steps, etc. Based on varied hyperparameter experiments results, the final setting of hyperparameters is as follows:

- Three CNN channels
  - Kernel size: 2×2, 3×3, 4×4
  - Filter numbers n: 128
  - Activation functions: relu
- GRU channels
  - Hidden state size h: 128
  - Activation functions: tanh
- Fully connected layers
  - Dense(16,activation='relu')
  - Dense(1,activation = 'sigmoid')
- Learning rate: 1e-3 with decaying factor of 1e-5
- Batch size: 1024

Besides, the regularizations are applied including the spatial dropout (0.2) for the embedding layer, and the dropout layers (0.5 and 0.2) are applied on the fully connected layers.

The model is trained on GPU with Adam optimizer and stopped at 14 epochs. It is evaluated by binary cross-entropy loss function.

F. Evaluation

Several metrics are utilized to evaluate the model on the test data: overall accuracy, precision, recall, and F1 score. The overall accuracy is calculated with a threshold of 0.5. The project is aiming to detect fake news as 'fake', the recall for the negative class is the most important.

G. Experiments

The baseline models are designed based on the studies from the literature. RNN-based models and CNN models are the two main categories. The combination models of RNN and CNN are also studied. However, those research are conducted on independent topics which would lead to a bias in classification. We built several baseline models on the mixed-topics-based datasets and compare them with the innovative models in two aspects:

1) We compare RNN-based models with CNN-based models on accuracy and efficiency. The studies of RNN and CNN on nature language processing show that the hybrid of RNN-based models and CNN-based models can lead to high performance.

2) We compare the model with and without embedding techniques. It has been proved that the model can be trained on the pre-trained word vectors to achieve high accuracy in the classification.

The models in this project are RNN, LSTM, GloVe-enhanced LSTM, GloVe-enhanced BiLSTM, GloVe enhanced CNN, and the innovated model GloVe enhanced multichannel CNN_GRU_attention. They are marked as RNN, LSTM, G_LSTM, G_BiLSTM, G_CNN, G_MCNN-GRU in the figures and tables.

V. Results and analysis

A. Compare with baseline models

The overall performances for all the models are in table I. The results of the comparison of deep learning models for fake news classification show that the hybrid multichannel

CNN_GRU with attention and GloVe embedding performed the best, with an accuracy of 0.9162. This model also converged the fastest among the tested models. The CNN with GloVe embedding was the second-best model, with an accuracy of 0.8911. The BiLSTM with GloVe embedding was the third-best model, with an accuracy of 0.8704. The worst performing model was the simple RNN, which had an accuracy of 0.6254. This suggests that more complex models, such as the hybrid model and the CNN with GloVe embedding, are more effective at detecting fake news than simpler models like the RNN.

TABLE I
MODEL COMPARISION

Model	Metrics			
	Test accuracy	*Precision*	*Recall*	*F1 score*
G_MCNN-GRU	0.9162	0.92	0.91	0.91
G_CNN	0.8911	0.90	0.88	0.89
G_BiLSTM	0.8704	0.87	0.87	0.87
G_LSTM	0.8685	0.87	0.86	0.86
LSTM	0.7422	0.80	0.70	0.70
RNN	0.6254	0.61	0.59	0.61

From the validation loss curves in figure *, we can see the multichannel CNN-GRU with attention model outperformed the other models by fast convergence due to the ability to process data in parallel. It reached the lowest value at around 12 epochs.

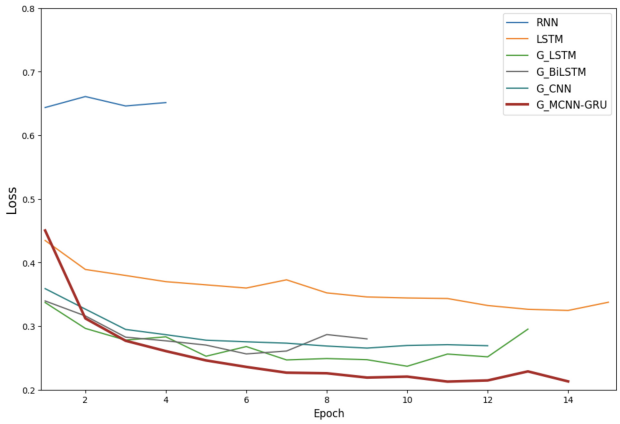

Fig. 5. Validation loss comparison

The hybrid model of CNN and GRU with attention outperformed the other models because it combines the strengths of both CNN and GRU. The CNN can extract useful features from the text data, while the GRU can capture the long-term dependencies in the data. The attention layer on top of the GRU further improves the performance by allowing the model to focus on the most important parts of the text.

The self-attention layer on the top of the GRU allows the model to focus on the most relevant parts of the text for the task of fake news classification. This improves the performance by allowing the model to better capture the key features that are indicative of fake news.

Figure * shows the comparison of recall of negative class. It demonstrates the ability to detect fake news and label it as 'fake'. The multichannel CNN-GRU with attention model has a recall of 0.85, GloVe_LSTM model with a recall of 0.84, and GloVe_CNN model with a recall of 0.82. The results indicated the models can detect fake news from the phrase relationships and the combination of the words.

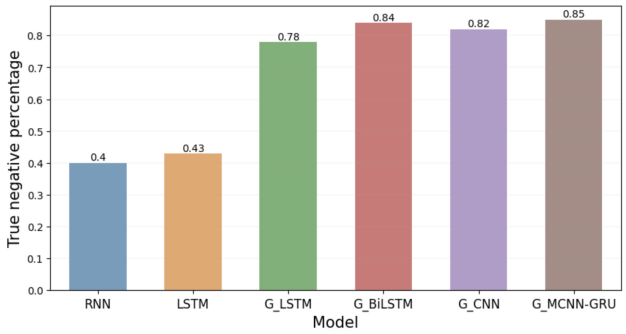

Fig. 6. Recall of Negative Class (Fake news)

One interesting aspect of the results is that the use of GloVe embeddings appears to improve the performance of the models. All the top performing models (the hybrid model, the CNN with GloVe embedding, and the BiLSTM with GloVe embedding) used GloVe embeddings, while the worst performing model (the RNN) did not use GloVe embeddings.

B. Compare with models in literature

Table II shows the comparison of multichannel CNN-GUR with attention models with the models from literature reviews. The results demonstrate the high accuracy of the multi-topic news dataset. Since the dataset contains multi-dimensional information including the varied length of sentences, several resources, and different subjects and topics, our model shows a significant improvement in text classification and achieved high accuracy and efficiency.

TABLE II
STATE-OF-ART MODEL COMPARISON

Model	*Test Data*	*Result*
G_MCNN-GRU (our model)	26,228 multi-topic news	precision 92%, recall 91%, F1 score 91
NA-CNN-COIF-LSTM [5]	1,000 movie reviews	precision 99.1%, recall 99.3%, F1 score 99.2%
RNN/CNN-ANN [6]	1,155 telephone conversations	84% accuracy
LSTM-attetion [7]	2,426 microblogs	precision 85%, recall 88%, F1 score 86%

VI. CONCLUSION

In conclusion, the comparison of different deep learning models for fake news classification showed that more complex models, such as the hybrid model and the CNN with GloVe embedding, are more effective at detecting fake news than

simpler models like the RNN. The use of GloVe embeddings also appears to improve the performance of the models.

In the future, we would expand the models from word-level prediction to character-level prediction and apply the methodology to the free-process data. We could also explore the topic-based model to classify text.

REFERENCES

[1] Institute of Medicine (US) Committee on Assuring the Health of the Public in the 21st Century. The Future of the Public's Health in the 21st Century. Washington (DC): National Academies Press (US), 2002, 7.

[2] A. Flanagin, "Online Social Influence and the Convergence of Mass and Interpersonal Communication," Human Communication Research, 2017, pp. 450–463.

[3] N. R. de Oliveira, P. S. Pisa, M. A. Lopez, D. S. V. de Medeiros, and D. M. F. Mattos, "Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges," Information, 2021, p. 38.

[4] I. Goodfellow, Y. Bengio, A. Courville, Deep learning. MIT press, 2016.

[5] Y. Luan, S. Lin, "Research on Text Classification Based on CNN and LSTM," ICAICA, 2019, pp. 352-355.

[6] J. Lee, F. Dernoncourt, "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks," arXiv:1603.03827, 2016.

[7] X. Bai, "Text classification based on LSTM and attention," ICDIM, 2018, pp. 29-32.

[8] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, "Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media," arXiv:1809.01286, 2018.

[9] I. Agarwal, D. Rana, N. Daivik, C. Teja, "Fake News on COVID-19 Healthcare", IEEE Dataport, 2021.

[10] P. Verma, P. Agrawal, I. Amorim, R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection," IEEE Trans. Comput. Soc., 2021, pp. 881-893.

[11] J. Pennington, R. Socher, C. Manning, "GloVe: Global Vectors for Word Representation," EMNLP 2014.

[12] X. Zhang,, J. Zhao, Y. LeCun, "Character-level Convolutional Networks for Text Classification," Adv Neural Inf Process Syst, 2015, 28.

[13] C. Li, G. Zhan, Z. Li, "News Text Classification Based on Improved Bi-LSTM-CNN," ITME, 2018, pp. 890-893.