Expert Systems With Applications 207 (2022) 117958

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multi-layer features ablation of BERT model and its application in stock trend prediction

Feng Zhao [a,*], Xinning Li [b], Yating Gao [b], Ying Li [a,c], Zhiquan Feng [c,d], Caiming Zhang [a,e,*]

[a] *School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China*
[b] *School of Statistics, Shandong Technology and Business University, Yantai, China*
[c] *Shandong Provincial Key Laboratory of Network Based Intelligent Computing, Jinan, China*
[d] *School of Information Science and Engineering, University of Jinan, Jinan, China*
[e] *School of Software, Shandong University, Jinan, China*

ARTICLE INFO

Keywords:
Network stock comments
BERT model
Sliding window strategy
Ablation study
Text topic recognition

ABSTRACT

Stock comments published by experts are important references for accurate stock trends prediction. How to comprehensively and accurately capture the topic of expert stock comments is an important issue which belongs to text classification. The Bidirectional Encoder Representations from Transformers (BERT) pretrained language model is widely used for text classification, due to its high identification accuracy. However, BERT has some limitations. *First*, it only utilizes fixed length text, leading to suboptimal performance in long text information exploration. *Second*, it only relies on the features extracted from the last layer, resulting in incomprehensive classification features. To tackle these issues, we propose a multi-layer features ablation study of BERT model for accurate identification of stock comments' themes. Specifically, we *firstly* divide the original text to meet the length requirement of the BERT model based on sliding window technology. In this way, we can enlarge the sample size which is beneficial for reducing the over-fitting problem. At the same time, by dividing the long text into multiple short texts, all the information of the long text can be comprehensively captured through the synthesis of the subject information of multiple short texts. *In addition*, we extract the output features of each layer in the BERT model and apply the ablation strategy to extract more effective information in these features. Experimental results demonstrate that compared with non-intercepted comments, the topic recognition accuracy is improved by intercepting stock comments based on sliding window technology. It proves that intercepting text can improve the performance of text classification. Compared with the BERT, the multi-layer features ablation study we present in the paper further improves the performance in the topic recognition of stock comments, and can provide reference for the majority of investors. Our study has better performance and practicability on stock trend prediction by stock comments topic recognition.

1. Introduction

The accurate prediction of stock trends not only helps investors avoid risks and obtain returns, but also enables the government to regulate the stock market. Stock comments published by experts play an important role in accurate prediction of stock trends (Ruan et al., 2018; Zhao et al., 2020). However, it is quite challenging to capture the comment topics comprehensively and quickly due to the large number of expert stock comments published at the same time and the time-consuming manual reading of the comments. It has become the focus of investors and researchers to effectively capture the topics from numerous expert stock comments, which is within the scope of this paper.

In essence, the topic recognition of expert stock comments can be formulated as the text classification problem (Kim, 2014). Recently, many text classification methods have been proposed which can be roughly divided into two categories. The first category is based on word frequency statistics (Liu, 2020; Sonam and Devaraj, 2020; Tripathy et al., 2016). The Part of Speech (POS) and Term Frequency - Inverse Document Frequency (TF-IDF) search the classified feature words by the word frequency difference of text, then news sentiments in stock market are recognized (Sonam and Devaraj, 2020; Tripathy et al., 2016). Liu (2020) first applied the Latent Dirichlet Allocation (LDA) to reduce the dimension of the word vectors of news documents, mine the implicit semantics of news by means of word frequency statistics, and then

[*] Corresponding author at: School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China (C. Zhang).
E-mail addresses: zhaofeng1016@126.com (F. Zhao), czhang@sdu.edu.cn (C. Zhang).

https://doi.org/10.1016/j.eswa.2022.117958
Received 7 July 2021; Received in revised form 27 May 2022; Accepted 21 June 2022
Available online 25 June 2022

analyze the relationship between hot news and stock market. Although LDA makes use of the associations between words, it fails to obtain the context information within adjacent words. Overall, these word frequency statistics methods do not fully explore the structural information in the text, which has a negative influence on the accuracy of text recognition.

The second category of methods are based on the distribution of document content, which takes full advantage of the context structure information to improve the accuracy of text recognition (Da'u et al., 2020; Liu et al., 2021; Otter et al., 2020; Rao et al., 2018; Sousa et al., 2019). To date, significant progress has been made with deep learning on financial text mining models (Da'u et al., 2020; Liu et al., 2021; Otter et al., 2020). For example, Long-Short Term Memory (LSTM) can continuously capture long-term context information in a one-way learning mode to obtain the local feature distribution of text, and implement document-level sentiment classification (Rao et al., 2018).

BERT, a model that bidirectional encodes input text, can achieve better performance than LSTM, and thus has been widely used in text topic recognition (Liu et al., 2021; Otter et al., 2020). Most studies directly apply BERT model to different fields (Araci, 2019; Liu et al., 2021; Pota et al., 2021; Tomihira et al., 2020). For example, Sousa et al. (2019) first applied the BERT model to analyze stock comments sentiment for stock market decision-making, and achieved prominent results due to fully capture the text structure information. FinBERT model proposed by Araci (2019) applies the financial corpus to re-pre-train the BERT model, so as to extend the application of the BERT model to the financial field. Liu's (2021) FinBERT uses general corpus and financial corpus to pretrain BERT in six new tasks, which are different from the original two pre-training tasks of BERT model. Pota (2021) directly applies BERT for sentiment analysis of tweets by avoiding noise in the pre-processing and processing the information hidden in emoticons into text as the input. Tomihira (2020) also directly utilizes BERT for the classification of tweets, which takes emoticons as text labels. It is a new form of classification. The above studies mainly aim to extend the application of BERT model to emotion analysis and text classification. They generally follow the original processing of BERT model: the truncation method is used for the input text exceeding the fixed length while the classifier token $[CLS]$ vector of the last layer is directly used as the extracted features for output.

However, current BERT model suffers the following drawbacks when applied for long text recognition, such as topic identification of expert stock comments. *Firstly*, the BERT model can only deal with fixed-length text, resulting in comprehending incomplete long text information. Since the length of long text is generally larger than the fixed length, the text features derived from part of the text rather than the whole text may lead to that only partial text structure information can be captured. *Secondly*, the extracted features in the BERT model are usually the classifier token $[CLS]$ vector of the last layer of the model, neglecting the multi-layer features information, resulting in capturing incomplete text identification features. Therefore, how to solve the problem of text length fixation and feature measurement in BERT model is of research value for text topic identification.

To address the above issues, we propose a multi-layer features ablation method based on BERT model, which is suitable for long text analysis and can effectively capture text features comprehensively, for further using in stock text. Specifically, our work includes the following aspects. (1) Based on the sliding window technology, we divide a stock comment into several short texts with fixed length which are suitable for BERT model (Chu, 1995; Tao and Papadias, 2006; Xie et al., 2021; Yun et al., 2019). Then the topic of the stock comment is scored by the topic identification results of above several short texts. In this way, we can fully explore the information in the long text. At the same time, the enlarged sample size can alleviate the over-fitting problem. (2) In order to fully capture text topic recognition features, we make full use of each layer of BERT model, i.e., we extract features from each layer of BERT model and conduct ablation study on the resulting multi-layer features (Girshick et al., 2014; Wang and Neumann, 2018). Specifically, we do layer by layer weighted fusion of features from the last layer to the first and select the most discriminative multi-layer features of text topic recognition. This method preserves the valid features and removes the invalid features for text classification. Then it can also improve the efficiency of text topic identification.

Therefore, different from the previous works that applied BERT model directly, we identify the inherent drawbacks of BERT model and devote to improving BERT model from two aspects: input representation as well as feature fusion. On the one hand, we process the input expert stock comments by sliding window technology, not only enlarging the sample size but also capturing more information contained in long text. On the other hand, we propose to extract more effective features by multi-layer features ablation in the output of BERT model, so as to identify the theme of stock comments.

The rest of the paper is organized as follows. Section 2 briefly introduces the BERT model. In Section 3, we propose the multi-layer features ablation of BERT model and introduce the improvement of input and output in BERT model in detail. In Section 4, the effectiveness of the proposed method is demonstrated through experiments. Finally, the conclusions of our study are given in Section 5.

2. BERT model

The BERT model is one of the state-of-the-art pre-trained language models (Liu et al., 2021; Pota et al., 2021; Tomihira et al., 2020). In this section, we briefly introduce the BERT model from three parts. Specifically, we firstly introduce the architecture of the BERT model for text topic recognition in Section 2.1. Then, we also introduce the pre-training of the BERT model in Section 2.2. In Section 2.3, the merits and drawbacks of the BERT model are summarized.

2.1. The architecture of the BERT model

The basic architecture of BERT model is shown in Fig. 1. As shown in Fig. 1, the structure is divided into the input layer, the hidden layer and the output layer. The input of the BERT is the vectors $E_i, i = 1, 2, \cdots, N$, N is fixed size. The input E_i represents the sum of three embeddings: a marker word embedding, a position word embedding and a sentence word embedding. The components of the hidden layer are $T^j_{E_i}, j = 1, \cdots, M$, while $j = 1, 2$ in this paper. In hidden layer, the output of the upper layer is provided as input for the next one, and the multi-layer

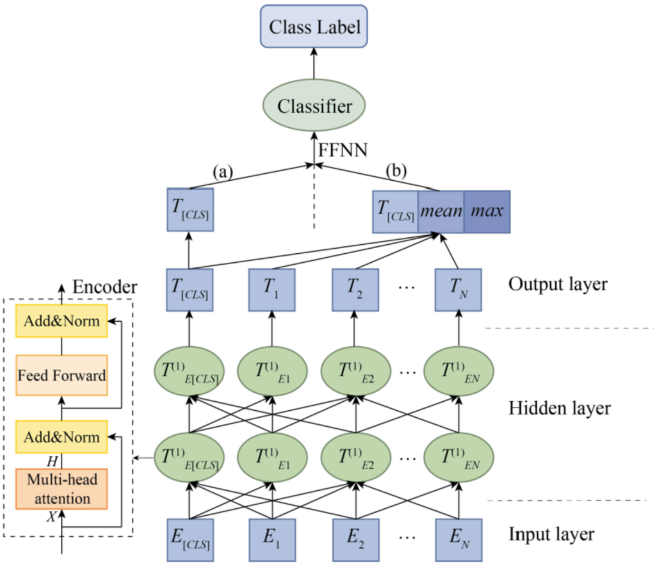

Fig. 1. Architecture of the BERT model for text topic recognition.

bidirectional coding obtains the implicit vector containing rich context characteristics. The output T_i vector is obtained from the E_i vector from input layer by BERT model processing. $E_{[CLS]}, T^j_{E[CLS]}, T_{[CLS]}$ of artificial token denoted $[CLS]$ are used for text classification, and $T_{[CLS]}$ is used to feed the classifier in Fig. 1(a). Then text class label is obtained.

The Encoder feature extractor of Transformer is applied to the hidden layer of the BERT model, where the j th layer Encoder corresponds to $T^j_{E_i}$ and $j = 1, \cdots, M$ (Vaswani et al., 2017). The unit structure of the Encoder is shown in Fig. 1 'Encoder', which is composed of 'multi-head attention' and 'Feed Forward'. The BERT model obtains contextual semantic information through calculation in the 'multi-head attention' part which contains multiple self-attention.

'Multi-head attention' is the key part of 'Encoder' and it works through self-attention mechanism. For each attention, suppose the input sequence is $X = [x_1, x_2, \cdots, x_N]$, and the output sequence is $H = [h_1, h_2, \cdots, h_N]$. And there are initialization matrices W^Q, W^K, W^V, the results of W^Q, W^K, W^V multiply with X are Q, K, V, which are query vector, key vector and value vector, respectively. The output sequence is calculated by multiplying the value vector and the attention weight, which is calculated through the dot product of the query vector and the key vector as:

$$H = softmax\left(\frac{Q \times K^T}{\sqrt{D_k}}\right) \times V \tag{1}$$

where D_k is the dimension of the query vector Q. After the operation of multi-layer self-attention, more context information can be learned.

On the basis of the output of BERT model, Gao et al. (2019) proposed the feature extraction and fusion of the output layer, and made full use of all T_i information to obtain new classification features. According to the above research, we can capture text topic by the connection of $[CLS]$ with the extracted fusion vectors *mean* and *max*. The vectors *mean* and *max* are extracted from all word vectors in the output layer of BERT model by pooling methods, which contain mean-pooling and max-pooling (Fujita and Cimr, 2019; Li et al., 2016; Tomihira et al., 2020).

As shown in Fig. 2, *embedding-dimension* is the length of the word vector, *seq-length* is the fixed length of the BERT model text. Mean-pooling takes the average value of all elements in the *seq-length* direction. Its calculation formula is $s_j = \frac{1}{l}\sum_{i=1}^{l} t_{ij}$, suppose the j th value of the *mean* vector is s_j, the length of the text is l, and the value of j th position of the i th word is t_{ij}. Max-pooling takes the maximum value of all elements in the *seq-length* direction. Its formula is $s'_j = \max_{i\in[1,l]}(t_{ij})$, where s'_j denotes the j th value of the *max* vector.

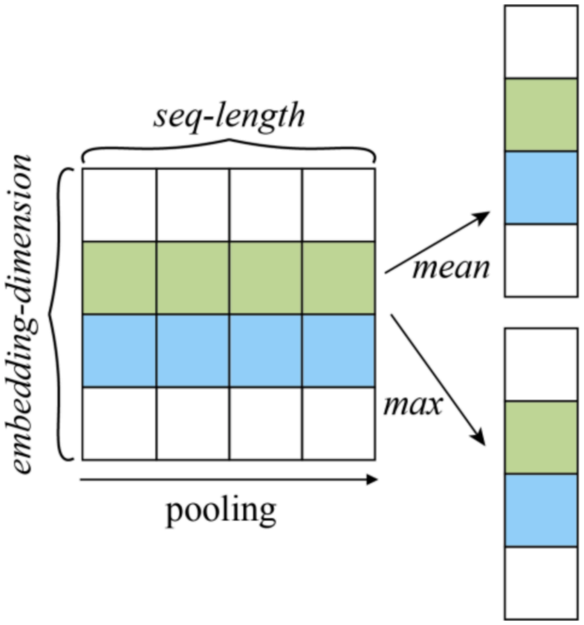

Fig. 2. The diagram of extracting the fusion information by the pooling method.

The mean-pooling and max-pooling methods take the average value and the maximum value of all elements in the *seq-length* direction of each layer in BERT model, respectively. And all elements of output vectors of each layer in BERT model are randomly distributed with positive and negative values. The effect of taking the maximum value and the minimum value of all elements in the *seq-length* direction is similar. Therefore, we take the max-pooling method in this paper.

As shown in Fig. 1, the three vectors *mean*, *max* and $[CLS]$ are fused in the output layer of BERT model to capture text topic. It is fed to the classifier to get the topic estimation.

2.2. *The pre-training of the BERT model*

The pre-training of the BERT model is carried out in two ways simultaneously, as shown in Fig. 3. The *first* way is masked language model, i.e., 15% of the characters are randomly masked when typed. Through the study of a large number of texts, the loss is reduced to the minimum, so that the specific character of the masked part can be correctly judged. For example, in Fig. 3, '[mask1]' and '[mask2]' are masked characters, and 'Token1′' and 'Token2′' are separately the correct 'Tok 2′ in 'Masked Sentence A' and 'Tok Q' in 'Masked Sentence B'. The *second* way is next sentence prediction, i.e., the relationships between sentences are learned while covering words. Specifically, a part of the sentences is replaced randomly, then the prediction is made of whether the next sentence will be the next based on the previous sentence. In Fig. 3, through learning a large number of texts in BERT model, the output corresponding to the $[CLS]$ position is used to predict whether 'Masked Sentence B' is the next sentence of 'Masked Sentence A' or not.

For deep learning frameworks, the pre-training of parameters is a common pre-application approach (Keneshloo et al., 2019). The pre-trained BERT model learns the relationship of all texts, and then the parameters are fine-tuned for specific downstream tasks. In this paper, we apply the pre-trained BERT model framework to identify the topic of stock comments.

2.3. *The merits and drawbacks of the BERT model*

The BERT model has extensive applications for its following merits. *Firstly*, BERT is a model that bidirectional encodes input text based on the Encoder part of Transformer, so the extraction of corpus features based on BERT model is more efficient (Devlin et al., 2018). *In addition*, the BERT model has achieved state-of-the-art performance in natural language processing tasks, and the pre-trained BERT model only needs supervised parameter fine-tuning for downstream tasks.

Despite its merits, we have also identified the following drawbacks of BERT model based on the experiments. (1) The length of the longest input sequence in the BERT model is fixed, thus the text will be truncated directly if it is beyond the longest length in the BERT model. The truncated text may still have unutilized information, indicating that the text which is longer than the limitation of BERT model requires further data processing. (2) The feature extracted from the BERT model is the vector corresponding to the marker $[CLS]$ or the connection of the three vectors, i.e., the *mean*, *max* and $[CLS]$ in the output layer, which is shown in Fig. 1. The feature extraction method ignores the information loss of multi-layer features which are effective for text topic recognition. To tackle these drawbacks, we propose an ablation study of BERT model features in next chapter.

3. Multi-layer features ablation of BERT model

In this section, we propose improvement methods to solve the deficiencies of the BERT model mentioned in Section 2.3, which are mainly

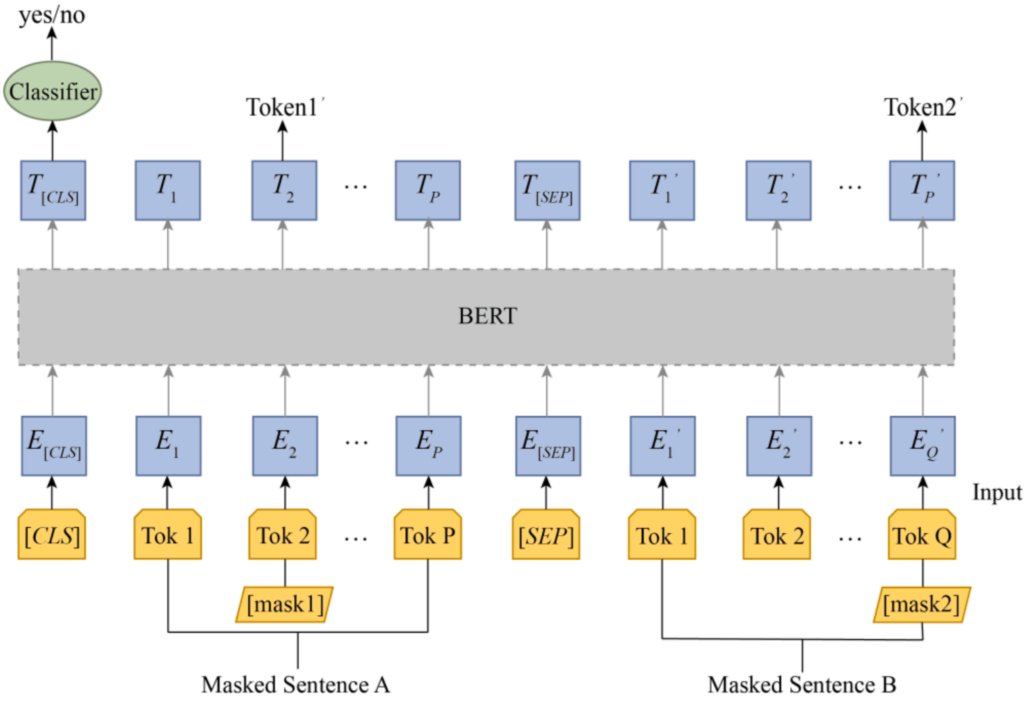
yes/no
Classifier
Token1′
Token2′
$T_{[CLS]}$ T_1 T_2 ⋯ T_P $T_{[SEP]}$ T_1' T_2' ⋯ T_P'
BERT
$E_{[CLS]}$ E_1 E_2 ⋯ E_P $E_{[SEP]}$ E_1' E_2' ⋯ E_Q'
Input
[CLS] Tok 1 Tok 2 ⋯ Tok P [SEP] Tok 1 Tok 2 ⋯ Tok Q
[mask1]
[mask2]
Masked Sentence A
Masked Sentence B

Fig. 3. The pre-training of the BERT model.

from two aspects. Specifically, we firstly present the sliding interception method of text in Section 3.1. Next, the feature extraction method of multi-layer features ablation based on BERT model is introduced in Section 3.2.

3.1. The method of text interception by sliding window

The maximum length of the input sequence in the BERT model is fixed, while the lengths of stock comments are mostly more than this fixed length. To overcome the drawback of incomplete topic information capture in long text caused by truncation text in BERT model, we apply the sliding window technique to intercept text, which is shown in Fig. 4. We set the sliding window size to m and the sliding step size to n. Let 1 denote the first character of text, N denote the N th character of text, then the length of text is N, and $i_1, i_2, \cdots, i_j$ denote the subsets of text that is captured by the sliding. The number of subsets of text is $j = \left\lceil \frac{N-m}{n} \right\rceil + 1$.

The method of text interception by sliding window technology has two advantages. It can not only fully capture the thematic information of the long text, but also reduce the over-fitting problem through increasing sample size.

3.2. Multi-layer features ablation method based on BERT model

The accuracy of text topic recognition can be improved by extracting more representative text features. Therefore, we propose a multi-layer characteristic ablation method for BERT model. This method deals with all outputs of the BERT model layers. The ablation study with all information is not only for maximizing the valid feature information and discarding the invalid feature information of text classification, but also for reducing the information loss between the layers of the BERT model. Thus, we can improve the accuracy of text topic recognition.

We extract features from all the layers of the BERT model for ablation study. Specifically, on account of classification feature ([*CLS*]) in BERT model basic architecture for text topic identification in Fig. 1, we firstly extract features from the ablation study of the multi-layer [*CLS*] vectors in Fig. 5(a). Then in the whole Fig. 5, we ablate the multi-layer [*CLS*] vectors, the multi-layer *mean* vectors and the multi-layer *max* vectors, respectively, and fuse the above three extracted feature vectors based on the idea of feature connection.

Firstly, when we ablate multi-layer [*CLS*] vectors in BERT model, features in each layer [*CLS*] can be extracted, and features can be added layer by layer starting from the last layer. In other word, the features can be added layer by layer for fusion according to the rules of the last layer, the last two layers and the last three layers et al. We can learn the weights of each layer in BERT model and obtain new features by

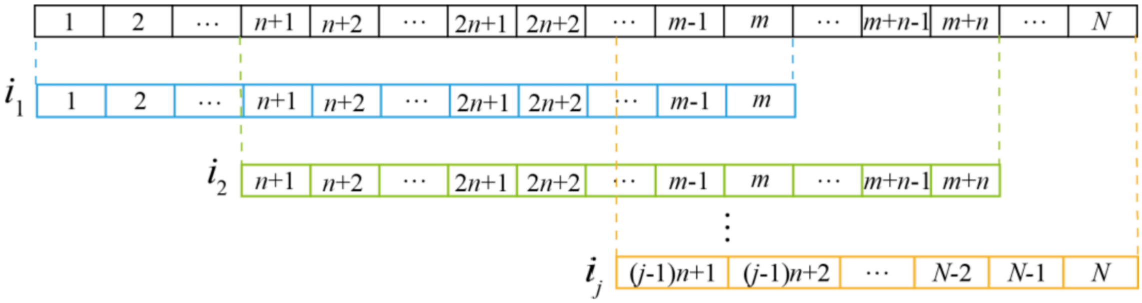
1 2 ⋯ n+1 n+2 ⋯ 2n+1 2n+2 ⋯ m-1 m ⋯ m+n-1 m+n ⋯ N
i_1
1 2 ⋯ n+1 n+2 ⋯ 2n+1 2n+2 ⋯ m-1 m
i_2
n+1 n+2 ⋯ 2n+1 2n+2 ⋯ m-1 m ⋯ m+n-1 m+n
⋮
i_j
(j-1)n+1 (j-1)n+2 ⋯ N-2 N-1 N

Fig. 4. The method of text interception by sliding window.

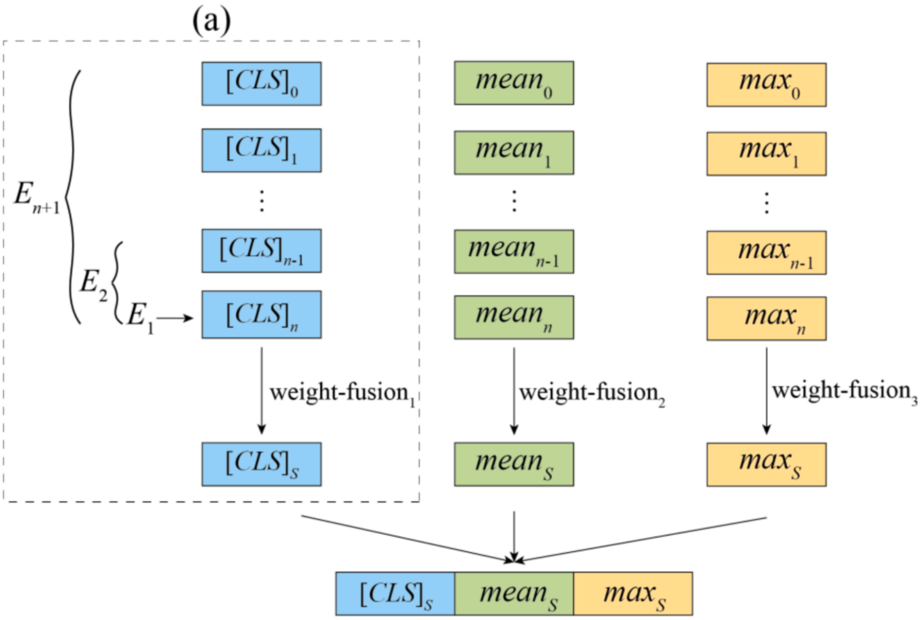

Fig. 5. Multi-layer features ablation method based on BERT model containing two classification features.

weighted fusion to verify its accuracy in text topic recognition. Multi-layer ablation with $[CLS]$ characteristics is shown in Fig. 5(a). Let $[\mathrm{CLS}]_0$ denote the initial $[CLS]$, $[\mathrm{CLS}]_i(i = 1, 2, \cdots, n)$ denote the output vector for i th layer $[CLS]$, $[\mathrm{CLS}]_S = \sum_{j\in T}\alpha_j[\mathrm{CLS}]_j$ denote the feature for text subject recognition, where T denotes the number of layers corresponding to the features selected for the ablation study, $0 \leq \alpha_j \leq 1$, and $\sum_{j\in T}\alpha_j = 1$. Meanwhile, we set the ablation of the last layer as E_1, the last two layers as E_2, the ablation of all layers as E_{n+1}, and so forth.

Secondly, in addition to presenting the above multi-level ablation study of $[CLS]$, we also propose ablation studies on multi-layer $[CLS]$, *mean* and *max* vectors together, as shown in the Fig. 5. We fuse the vectors of multiple layers into one vector respectively and then connect them. Thus, we can get a combined feature with more information. The fusion of multi-layer $[CLS]$, *mean* and *max* vectors is similar to the fusion in Fig. 5(a). Let $mean_i(i = 0, 1, 2, \cdots, n)$ and $max_i(i = 0, 1, 2, \cdots, n)$ denote the output vector for i th layer *mean* and *max* respectively, $mean_S = \sum_{m\in T}\alpha_m mean_m$ and $max_S = \sum_{k\in T}\alpha_k max_k$ respectively denote the part fusion features for text subject recognition, $0 \leq \alpha_m \leq 1$, $\sum_{m\in T}\alpha_m = 1$, $0 \leq \alpha_k \leq 1$, and $\sum_{k\in T}\alpha_k = 1$. The connection of $[\mathrm{CLS}]_S$, $mean_S$ and max_S is the feature of text topic classification. The layer that achieves the highest accuracy of text topic recognition can be found. Here, the ablation of the three features of the last layer is E_1, the ablation of the last two layers is E_2, and the ablation of all layers is E_{n+1}, which are linked together corresponding to different layer fusion features and used for text topic classification.

The multi-layer features ablation method based on BERT model has two merits. This method not only extracts text structure information through bidirectional coding, but also obtains fusion features with higher classification contribution rate through making full use of discriminative information from multi-layer features and discard invalid information. Thus, we can improve the accuracy of text topic recognition.

4. Experiments and analysis

To verify the performance of the proposed method, we conduct four different experiments. Specifically, we firstly introduce the experimental data in Section 4.1. Then, in Section 4.2, we report the classification results of our model and some related baseline models. In Section 4.3, we analyze the influence of features from different layers on the text recognition. Afterwards, Section 4.4 compares the text identification results with different feature ablation methods. Finally, Section 4.5 presents the application of our method in predicting stock trends. The results demonstrate that text interception based on sliding window and multi-level ablation study of $[CLS]$, *mean* and *max* vectors can improve the effective utilization of features and further increase the accuracy of text subject recognition. Meanwhile, the application of predicting stock trends by multi-layer features ablation of BERT model can achieve the best result.

4.1. Experimental data

The data for our experiments is from 'East Money'(https://www.eastmoney.com/). We extract the experts stock comments from 'Market Analysis' Sector by crawler technology (Fang et al., 2017). The extracted stock comments are from July 2, 2020 to September 30, 2020. Subsequently, we manually annotate the subject of the 603 comments containing 'positive' and 'negative' after the screening, where 'negative' topic means 'not positive'. At present, some studies have classified the topics of expert stock comments into two categories, namely, positive and negative (De Albornoz et al., 2010). Therefore, in this paper, we follow these studies and divide negative and neutral comments into 'not positive', and positive comments are 'positive'. Then the comprehensive results of stock comments with the above two labels are counted for stock trend prediction. The number of 'positive' and 'negative' comments for our experiments are basically equal.

Considering the length of most experimental text data exceeds the maximum length of input limited by the BERT model, we do preprocess to intercept text. The sliding step size is 300, and the intercepted comments are divided into training set, validation set and test set with the ratio of 8:1:1. Details of the experimental data are illustrated in Table 1.

Table 1
Explanation of experimental data.

Data declaration	Data value
Time range of stock comments	2020.7.2—2020.9.30
Categories of comments	'positive': 314 'negative': 289
Parameters of sliding window	sliding widow size: 512 sliding step size: 300

For performance evaluation, we adopt classification accuracy (ACC) and F1-score (F1) as the metrics. The larger value of ACC and F1, the better performance of text topic recognition.

Some important hyper-parameter settings of the pre-trained BERT model are listed in Table 2. The '*num-epochs*' is 4, the 'max-*position-embeddings*' is 512, i.e., the maximum fixed length of input text is 512. The '*num-attention-heads*' is 12 and '*num-hidden-layers*' is 12, which means that the number of all layers of features in BERT model is 13. We use the experimental data to fine-tune the pre-trained BERT model and testing.

4.2. The performance of different methods on text identification

To verify the effectiveness of the proposed method, we compare the text topic recognition performance by our model with several baseline methods. The baseline methods include LDA, TextCNN, LSTM and BERT model. For LDA, the implicit topic distribution is first learned from text (Mehrotra et al., 2013). Then, SVM is used to classify the text based on the implicit topic distribution. TextCNN applies one-dimensional convolution to obtain the n-gram feature representation of the sentences of text, which is able to extract shallow text features (Guo et al., 2019). LSTM is a one-way learning mode for capturing long-term context information (Rao et al., 2018).

Text segmentation methods contain 'without text interception' and 'with text interception', where 'without text interception' means the whole text is taken as model input and "with text interception" means the short texts obtained by intercepting the text by sliding window are used as model input. The accuracy and F1-score of text topic recognition obtained by different methods are shown in Table 3. In Table 3, LDA, TextCNN, LSTM and BERT model represent identifying topic by the feature extracted from LDA, TextCNN, LSTM and BERT model respectively. And our model represents identifying topic by the improved BERT model proposed in this paper.

From the results shown in Table 3, we can make the following conclusions: (1) Compared with the method of 'without text interception', the accuracy and F1-score of text topic identification by the 'with text interception' method are significantly improved. The reason lies in that dividing text by sliding window can comprehensively capture the long text information. (2) The results of feature extraction utilizing BERT model based on the distribution of document content are better than those of LDA method based on frequency statistics. And the results of feature extraction utilizing BERT model with bidirectional encoder are better than LSTM that encodes text in a one-way learning mode. This is because the BERT model can fully capture the text structural information. (3) Our model is better than those baseline approaches. This is because the multi-layer features ablation we put forward can obtain more information of text topic recognition than original BERT model.

4.3. The influence of the layers of fusion features on text identification

To explore the influence of different feature ablation layers on text topic recognition, we compare the topic recognition accuracy of all possible ablation layers. Specifically, we conduct experiments of two multi-layer characteristic ablation methods based on the BERT model. Firstly, we ablate multi-layer [*CLS*] vectors in BERT model to measure the efficiency from the last layer, the last two layers to all the layers. Then, the performance of the ablation of the multi-layer [*CLS*], *mean* and *max* vectors together in BERT model is measured from the last layer, the last two layers to all the layers.

Table 2
Hyper-parameters of the pre-trained BERT model.

Parameter names	The parameter value
num-epochs	4
max-*position-embeddings*	512
num-hidden-layers	12
num-attention-heads	12

Table 3
The text topic recognition performance of different methods.

Text Classification Methods	Text segmentation without text interception ACC	F1	with text interception ACC	F1
LDA	0.551	0.531	0.58	0.545
TextCNN	0.783	0.626	0.801	0.634
LSTM	0.779	0.616	0.797	0.625
BERT	0.818	0.628	0.835	0.635
Our model	**0.863**	**0.643**	**0.89**	**0.659**

The experimental results are shown in Fig. 6 which demonstrates the topic recognition efficiency of different combination layers by using two different multi-layer characteristic ablation methods based on BERT model. The value of abscissa denotes a superposition of the output features from the last layer to the first layer, the value of ordinate denotes the accuracy of topic recognition, the blue line denotes the efficiency of multi-layer [*CLS*] vectors ablation, and the green line denotes the efficiency of ablation of the multi-layer [*CLS*], *mean* and *max* vectors together.

Based on Fig. 6, we can make the following conclusions: (1) Among the classification results of two multi-layer feature ablation strategies, the recognition accuracy of the last six-layer [*CLS*], *mean* and *max* vectors fusion is the highest, which is 0.89, while the recognition accuracy of the last three-layer [*CLS*], *mean* and *max* vectors fusion is the lowest, which is 0.83. Their difference is 0.06. This phenomenon shows that the effect of fusion features with different layers changes greatly in text topic classification. Therefore, it is important to select the features with appropriate layers for fusion. In the future work, we will continue to research the method for selecting the fusion features with the best text topic recognition performance. (2) Choosing the features from appropriate layers in BERT model is the key to improve the topic recognition accuracy. The last three layers fusion are the best of [*CLS*] vectors fusion and the last six layers fusion are the best of [*CLS*], *mean* and *max* vectors fusion. The reason may be that these fusions make full use of the features that are valid for text classification and remove the features that are invalid for text classification. (3) For the 13 fusion layers in the ablation experiment, the green line and the blue line correspond to 13 results of text identification respectively. In the case of the same layer, 8 results of the green line are better than the blue line and only 4 results of the green line are worse than the blue line. Therefore, in most cases, the text identification accuracy of the green line is better than that of the blue line. In other words, the performance of the multi-layer [*CLS*], *mean* and *max* vectors ablation is better than the performance of multi-layer [*CLS*] vectors ablation. This can be explained by that multi-layer [*CLS*], *mean* and *max* vectors ablation utilizes classification features that contain more information than multi-layer [*CLS*] vectors ablation. In addition, we observe that the worse accuracy among all the results is observed in the green line. We think this phenomenon may be due to the interference of information redundancy on the effect of text subject recognition.

4.4. The influence of feature ablation method on text topic recognition

In order to verify the effectiveness of the proposed feature ablation method, we compare the topic recognition results of different feature ablation schemes of BERT model. Among them, original BERT and improved BERT model in different ablation schemes are compared in different text segmentation methods.

'Without text interception' and 'with text interception' are two different text segmentation methods of model input. 'Without text interception' means a text without segmentation and 'with text interception' means a long text segmented by sliding window technology. And feature ablation schemes contain original BERT model and different improved feature ablation schemes of BERT model. In Table 4, the

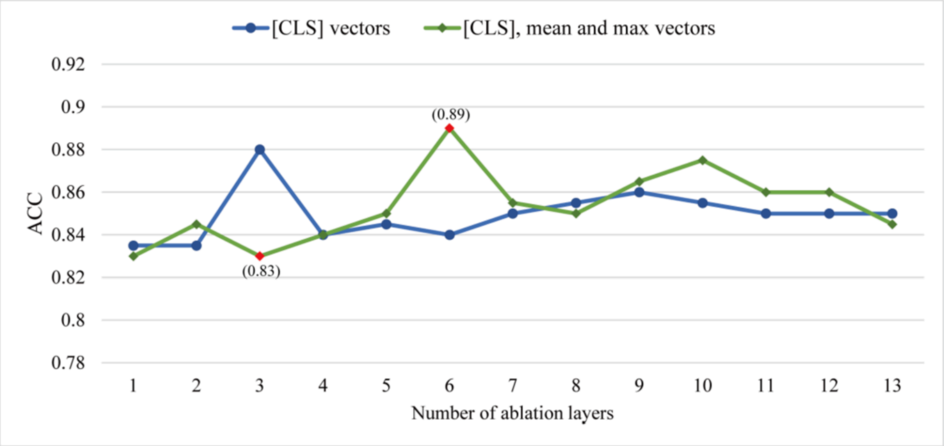

Fig. 6. Efficiency of two multi-layer characteristic ablation methods.

Table 4
The text topic recognition performance of different feature ablation schemes.

Feature Ablation Schemes	Text segmentation without text interception ACC	F1	with text interception ACC	F1
BERT	0.818	0.628	0.835	0.635
BERT-E_1'	0.824	0.628	0.83	0.643
BERT-E_3	0.858	0.635	0.88	0.65
BERT-E_6'	**0.863**	**0.643**	**0.89**	**0.659**

accuracy and F1-score of text topic recognition by different feature ablation schemes are shown respectively. And BERT represents identifying topic by BERT model, BERT-E_1' represents subject classification by the last layer $[CLS]$, *mean* and *max* fusion feature in BERT model, BERT-E_3 represents topic recognition by the fusion feature of the last three layers $[CLS]$ vectors in BERT model, BERT-E_6' represents the method of using the fusion feature of the last six layers $[CLS]$, *mean* and *max* vectors to identify text topic.

Based on Table 4, we can make the following conclusions: (1) The multi-layer features ablation of BERT model can effectively increase the performance of text topic recognition. It can be explained by that the fusion features of the last six layers $[CLS]$, *mean* and *max* vectors contain the most effective categorization information. Therefore, it is necessary to apply multi-layer features ablation to extract the most effective features and improve the efficiency of text topic recognition. (2) The text topic recognition results by the 'with text interception' method is significantly better than the 'without text interception' method. This is because the text interception by sliding window helps model comprehend the whole text information.

4.5. The influence of expert stock comments on the stock trend prediction

In order to explore the significance of the proposed method for predicting stock trend, we analyze the relationship between stock comments and stock trends. The experimental steps are as follows. (a) We extract expert stock comments from 'East Money' between October 1, 2020 and November 30, 2020 by crawler technology. This period is termed as the back testing period. (b) We apply the multi-layer features ablation with the best result of BERT model to identify the intercepted comments. (c) Considering that stock comments have delay effects on the stock trend prediction, and the different time steps of the statistics topic of stock comments also have an impact on the stock trend prediction. We predict the trend on day *d* according to topics of all text in the previous *t* days. Among them, the prediction time step size *t* and delay day *d* of the statistics topic of stock comments are two important parameters. The performance of stock trend prediction by expert stock comments can be optimized by adjusting the *t* and *d*, where $t \in [1, 3, 5, 7]$ and $d \in [1, 2, 3]$. We choose the best parameters to predict stock trend through experiments. Specifically, we apply the method of voting to get the topic of each comment from intercepted comments, obtain the topic of each day in the back testing period and obtain the topic of *t* days in different groups to predict the stock trend of day *d*. For example, we count the number of days with 'positive' topic (P) and days with 'negative' topic (N) in *t* days. *NP* denotes the trend is up and *NN* denotes it is not up. If P > N, then the predicted trend of the day of *d* is *NP*. (d) The Shanghai Composite Index and the Shenzhen Component Index are obtained to test the efficiency of predicting trends of stock market indexes in the back testing period. The changes of the Shanghai Composite Index and the Shenzhen Component Index are measured by the difference between the closing price of the day and the previous day. We only calculate the two indexes of the opening day of the stock, and use the stock comments on the opening day and non-opening day to forecast the trend of them. If the difference is positive, it is *NP*, otherwise, it is *NN*. (e) In the back testing period, we compare the predicted trend of the day with the real trend of the Shanghai Composite Index and the Shenzhen Component Index in the same day.

The prediction effect of stock trend by proposed method of stock comments recognition is shown in Table 5 and Table 6. Table 5 and Table 6 respectively show the statistical results of expert stock comments on the trend prediction effect of the Shanghai Composite Index and the Shenzhen Component Index. The results are ACC and F1 that change along with different combinations of parameters of prediction time step size *t* and delay day *d*.

From the prediction results of the Shanghai Composite Index trend and the Shenzhen Component Index trend shown in Table 5 and Table 6, we can make the following conclusions: (1) The parameters for the best prediction effect in Table 5 and Table 6 are prediction time step size $t = 5$ and delay day $d = 1$, i.e., when we apply the comprehensive stock comments of the first five days to predict the stock trend of the sixth day, the prediction effect can achieve the best result by the multi-layer features ablation of BERT model in this paper. The reason may be that the

Table 5
Prediction results of the Shanghai Composite Index trend by stock comments.

t	*d*					
	1		2		3	
	ACC	F1	ACC	F1	ACC	F1
1	0.595	0.622	0.541	0.605	0.486	0.585
3	0.648	0.622	0.568	0.585	0.514	0.585
5	**0.703**	**0.653**	0.622	0.638	0.541	0.585
7	0.622	0.638	0.541	0.636	0.514	0.604

Table 6
Prediction results of the Shenzhen Component Index trend by stock comments.

t	*d*					
	1		2		3	
	ACC	F1	ACC	F1	ACC	F1
1	0.595	0.609	0.514	0.572	0.486	0.55
3	0.622	0.609	0.541	0.591	0.514	0.591
5	**0.676**	**0.625**	0.622	0.609	0.541	0.571
7	0.595	0.591	0.568	0.59	0.514	0.572

comprehensive stock comments in the first five days can enable the proposed model to obtain more abundant and effective information, so as to improve the performance of stock trend prediction. When the prediction time step size is too short, the information acquisition will not be perfect. And if the time step is too long, the model will obtain redundant information, which will reduce the prediction performance. The information effectiveness is higher when delay day $d = 1$. However, the information effectiveness will decrease if the delay time is too long. (2) The above results show that automated topic recognition of expert stock comments by multi-layer features ablation of BERT model can provide an important basis of predicting stock trends. It may help the investors to gain more wealth returns and help the government to regulate the development of the stock market.

5. Conclusion

In this paper, we conduct a study based on the BERT model, which is called a multi-layer features ablation to improve the accuracy of stock comments topic recognition. In order to validate the effectiveness of the proposed method, we firstly collect expert stock comments by the crawler technology. Then, intercepting expert stock comments by sliding window and finally identifying comments topic by multi-layer features ablation of BERT model. The experimental results show that: (1) Intercepting text with sliding window cannot only capture all the thematic information of the text, but also increase the sample size to reduce the over-fitting problem. (2) The multi-layer features ablation method can extract all the informative features of the multi-layers output in BERT which can be used to further improve the recognition accuracy. (3) The automatic topic recognition of stock comments has the guiding significance to stock investment. Through the analysis of the test results and the algorithm, it is verified that our method is effective in stock trend prediction.

Our future research will focus on the following aspects. (1) For a large number of parameters, a lot of text are required to fine-tune the BERT model. Although our method is applied to the long text of stock comments, which increases the sample size, it is still not enough to fine-tune the parameters to the best state. Therefore, we will collect more data to enlarge the sample size of expert stock comments to improve the model performance. (2) In this paper, we only apply the method to identify comments topic and predict stock trend. This is just based on the expert stock comments. Thus, we will study how to extend multi-layer features ablation of BERT model to other fields, such as portfolio selection (Guan and An, 2019).

CRediT authorship contribution statement

Feng Zhao: Conceptualization, Methodology. **Xinning Li:** Conceptualization, Software, Writing – original draft, Methodology, Formal analysis, Investigation, Validation. **Yating Gao:** Validation. **Ying Li:** Writing – review & editing. **Zhiquan Feng:** Writing – review & editing. **Caiming Zhang:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (61773244, 61976125, 61272319, 61873117, 61972235 and 82001775), Yantai Key Research and Development Program of China (2017ZH065, 2019XDHZ081), Shandong Provincial Key Research and Development Program of China (2019GGX101069), and Opening Fund of Shandong Provincial Key Laboratory Of Network based Intelligent Computing: Study on the Presentation Path of Folk Life along the Grand Canal Based on Immersive Experience.

References

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Chu, C. S. J. (1995). Time series segmentation: A sliding window approach. *Information Sciences, 85*(1–3), 147–173. https://doi.org/10.1016/0020-0255(95)00021-G

Da'u, A., Salim, N., Rabiu, I., & Osman, A. (2020). Recommendation system exploiting aspect-based opinion mining with deep learning method. *Information Sciences, 512*, 1279–1292. https://doi.org/10.1016/j.ins.2019.10.038

De Albornoz, J. C., Plaza, L., & Gervás, P. (2010). A hybrid approach to emotional sentence polarity and intensity classification. In *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning* (pp. 153–161).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL)* (pp. 4171–4186).

Fang, B., Jia, Y., Li, X., Li, A., & Wu, X. (2017). Big search in cyberspace. *IEEE Transactions on Knowledge and Data Engineering, 29*(9), 1793–1805. https://doi.org/10.1109/TKDE.2017.2699675

Fujita, H., & Cimr, D. (2019). Computer aided detection for fibrillations and flutters using deep convolutional neural network. *Information Sciences, 486*, 231–239. https://doi.org/10.1016/j.ins.2019.02.065

Gao, Z., Feng, A., Song, X., & Wu, X. (2019). Target-dependent sentiment classification with BERT. *IEEE Access, 7*, 154290–154299. https://doi.org/10.1109/ACCESS.2019.2946594

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587). https://doi.org/10.1109/CVPR.2014.81

Guan, H., & An, Z. (2019). A local adaptive learning system for online portfolio selection. *Knowledge-Based Systems, 186*, Article 104958.

Guo, B., Zhang, C., Liu, J., & Ma, X. (2019). Improving text classification with weighted word embeddings via a multi-channel TextCNN model. *Neurocomputing, 363*, 366–374.

Keneshloo, Y., Shi, T., Ramakrishnan, N., & Reddy, C. K. (2019). Deep reinforcement learning for sequence-to-sequence models. *IEEE Transactions on Neural Networks and Learning Systems, 31*(7), 2469–2489.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *In Proceedings of the Empirical Methods in Natural Language Processing*, 1746–1751.

Li, J., Mei, X., Prokhorov, D., & Tao, D. (2016). Deep neural network for structural prediction and lane detection in traffic scene. *IEEE transactions on neural networks and learning systems, 28*(3), 690–703. https://doi.org/10.1109/TNNLS.2016.2522428

Liu, C. (2020). Analysis of relationship between hot news and stock market based on LDA model and event study. *In Journal of Physics: Conference Series, 1616*, 012–048. https://doi.org/10.1088/1742-6596/1616/1/012048

Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2021). FinBERT: A pre-trained financial language representation model for financial text mining. In *In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence* (pp. 4513–4519).

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 889–892).

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems, 32*(2), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Pota, M., Ventura, M., Fujita, H., & Esposito, M. (2021). Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. *Expert Systems with Applications, 181*, 115–119. https://doi.org/10.1016/j.eswa.2021.115119

Rao, G., Huang, W., Feng, Z., & Cong, Q. (2018). LSTM with sentence representations for document-level sentiment classification. *Neurocomputing, 308*, 49–57. https://doi.org/10.1016/j.neucom.2018.04.045

Ruan, Y., Durresi, A., & Alfantoukh, L. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems, 145*, 207–218. https://doi.org/10.1016/j.knosys.2018.01.016

Sonam, Devaraj, M. (2020). Analyzing news sentiments and their impact on stock market trends using POS and TF-IDF based approach. In 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET),1-6.

Sousa, M. G., Sakiyama, K., de Souza Rodrigues, L., Moraes, P. H., Fernandes, E. R., & Matsubara, E. T. (2019). BERT for stock market sentiment analysis. In *In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 1597–1601).

Tao, Y., & Papadias, D. (2006). Maintaining sliding window skylines on data streams. *IEEE Transactions on Knowledge and Data Engineering, 18*(3), 377–391. https://doi.org/10.1109/TKDE.2006.48

Tomihira, T., Otsuka, A., Yamashita, A., & Satoh, T. (2020). Multilingual emoji prediction using BERT for sentiment analysis. *International Journal of Web Information Systems, 265–280*. https://doi.org/10.1108/IJWIS-09-2019-0042

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications, 57*, 117–126. https://doi.org/10.1016/j.eswa.2016.03.028

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, W., & Neumann, U. (2018). Depth-aware cnn for rgb-d segmentation. In *In Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 135–150).

Xie, J., Hu, K., Li, G., & Guo, Y. (2021). CNN-based driving maneuver classification using multi-sliding window fusion. *Expert Systems with Applications, 169*, Article 114442. https://doi.org/10.1016/j.eswa.2020.114442

Yun, U., Lee, G., & Yoon, E. (2019). Advanced approach of sliding window based erasable pattern mining with list structure of industrial fields. *Information Sciences, 494*, 37–59. https://doi.org/10.1016/j.ins.2019.04.050

Zhao, F., Zhang, J., Chen, Z., Zhang, X., & Xie, Q. (2020). Topic identification of text-based expert stock comments using multi-level information fusion. *Expert Systems.* https://onlinelibrary.wiley.com/doi/10.1111/exsy.12641.