Combining joint statistical features with semantic features for text classification

Chuanshuai Ma [a], Junjie Peng [a,*], Cangzhi Zheng [a], Feng Cheng [b], Shuhua Tan [c], Fen Yi [c] and Huiran Zhang [a]

[a] *School of Computer Engineering and Science, Shanghai University, No. 333 Nanchen Rd., Shanghai 200444 China*
E-mails: machuanshuai@shu.edu.cn, jjie.peng@shu.edu.cn, cangzhizheng@shu.edu.cn, hrzhangsh@shu.edu.cn

[b] *Hasso Plattner Institute, University of Potsdam, Germany*
E-mail: feng.cheng@hpi.uni-potsdam.de

[c] *National Engineering Laboratory for Logistics Information Technology, YTO Express Co., Ltd., China*
E-mails: 00000226@yto.net.cn, 00001225@yto.net.cn

Abstract. With the rapid development of Internet technology, a large amount of electronic text information has been generated. As an important means of text information processing, text classification has become a research topic. At present, most text classification methods are devoted to the study of text features from different perspectives. However, few studies have focused on text classification in consideration of integrating multidimensional external information with the semantic features. To solve this problem, a text classification model that combines the joint statistical features with the semantic features is proposed. Specifically, it utilizes the word frequency and the part-of-speech (POS) frequency over different labels to generate joint statistical features. Furthermore, it selectively merges the joint statistical features with the semantic features through a gated mechanism to generate a highly discriminative feature representation. To verify the correctness and effectiveness of our model, extensive experiments are carried out on four commonly used classification datasets. Experiments show that our approach is effective and achieves better performance than state-of-the-art methods do.

Keywords: statistical features, part-of-speech, variational autoencoder, gated mechanism

1. Introduction

Text classification is one of the fundamental tasks in Natural Language Processing (NLP). It has been widely used in NLP tasks such as sentiment analysis, intent recognition, question answering and topic categorization. The key issue in text classification is to extract and learn text features and then assign one or more labels from a given label set to a text sequence.

Many studies have focused on text classification with different methods, among which approaches based on neural networks are mainly the most popular ones. For example, as a special feedforward neural network Convolutional Neural Networks (CNN) is widely used to extract N-gram features in text classification. The reason is that CNN has advantages in extracting semantic and syntactic information

[*] Corresponding author: School of Computer Engineering and Science, No. 333 Nanchen Rd., Shanghai 200444 China. Tel: 86-21-66135305; Fax: 86-21-66135517; E-mail: jjie.peng@shu.edu.cn.

from texts [1, 2]. Recurrent Neural Networks (RNN) based approaches are also extensively used for text classification [3], as it is well-known for processing sequential data. In addition, models based on Transformers [4] and BERT [5] are also the mainstream text classification methods nowadays. Such methods usually have complex models and huge resource consumption although they exploit the grammatical and syntactic information of the text and have high accuracy.

In text classification, besides deep learning models, introducing external information is also an effective way. For instances, statistical information, sentiment lexicon, external knowledge, entity knowledge base and label confusion [6–8] are commonly used external information. Take the most widely used statistical method Term Frequency-Inverse Document Frequency (TF-IDF) algorithm [9] as an example, it usually combines the statistical information with classifiers such as Bayes or Support Vector Machine (SVM) for text classification. By exploiting these external information, more discriminative text representations are obtained [10] to improve the text classification ability of the current popular text classification models such as CNN, RNN and BERT, etc. Recently, some scholars have proposed the method that combines the statistical information with neural networks, which has achieved remarkable progress. For example, Li et al. [11] merged statistical feature via adaptive gate for text classification. Yazdani et al. [12] proposed a bigram-based linguistic and statistical feature processing model for unstructured text classification. However, these studies only merge individual statistical features with semantics, instead of fusing different external information such as the word frequency information and the POS information to form a more discriminative statistical feature. When there are many text categories or data imbalance on the datasets, such a single statistical feature may not reflect the difference between different categories of texts.

In this paper, a text classification model based on the combination of semantic features and joint statistical features is proposed, whose statistical features consist of the word frequency and the POS frequency over all labels. Traditional statistical methods, such as TF-IDF, calculate the frequency of words throughout the entire dataset. However, due to the high occurrence of certain words across categories, the resulting statistical information may lack distinguishability. Our statistical approach differs from traditional methods in that it calculates the frequency of different words or POS under the corresponding label. This method yields unique distributions for each label category and distinctive distribution characteristics for each category of statistical information. Generally, the statistical features of word frequency and part-of-speech (POS) information of texts as well as their distribution over different labels are distinctive external information. Therefore, the word frequency information and the POS frequency information are incorporated into our model. By combining the word frequency features with the POS frequency features, a feature that is more discriminative than those of other statistical feature-based models is formed. Via merging the statistical features with the semantic features to enhance the differentiation of text features, the proposed model greatly improves the classification performance on the datasets such as AG's News and Yelp P., etc. As the statistical features may contain noise, not all statistical features are completely fused with the semantic features. Therefore, in the proposed model only selective feature fusion is adopted through a gated mechanism to optimize the model performance. The main contributions of this paper are summarized as follows:

- To enrich the statistical features, we count the distribution of the word frequency information and the POS frequency information over different labels. Moreover, we concatenate the word frequency distribution features to the POS frequency distribution features to form a more discriminative feature of the text sequence and prove it as an effective feature for classification.

- To eliminate the noise in the statistical features, we selectively merge the obtained joint statistical features with the semantic features through a gated mechanism to improve the performance of our model.
- Extensive experiments conducted on seven benchmarks demonstrate the effectiveness of our approach. The experiments show that our model achieves competitive results compared to these models on four commonly used datasets.

The remainder of this paper is organized as follows. Section 2 introduces the related work on text classification. Section 3 presents the method in detail. Section 4 shows the experimental setup and experimental results. Section 5 gives a conclusion.

2. Related work

In this section, the related work of text classification is introduced from three aspects: traditional text classification methods, text classification methods based on neural networks and additional knowledge-based text classification. Each aspect is presented in detailed below.

2.1. Traditional text classification methods

Traditional text classification methods adopt vector space model, N-grams or word bag model to extract text features. On this basis, the decision tree model, Support Vector Machine or Bayes model is used as the classifier. Bouboulis et al. [13] proposed a new framework for complex Support Vector Regression (SVR) as well as SVM for quaternary classification. Moreover, Ra et al. [14] presented adaptive boosting with uncertainty-based selective sampling for boosting the Naïve Bayes text classification. Although these methods are easy to understand and simple to implement, they ignore the latent semantic correlation between the words in text classification. This makes the text classification task face the problem of high dimensionality and high sparsity. Meanwhile, these methods usually extract few semantic information, resulting in low accuracy of classification results. Faced with such a dilemma, some scholars have turned their attention to neural network models.

2.2. Text classification methods based on neural networks

With the rapid development of neural networks, the application of neural network models in NLP has made great achievements. For instances, Convolutional Neural Networks (CNN) [2, 15] and Recurrent Neural Networks (RNN) [3] have been widely applied in text classification tasks and have achieved remarkable progress. Pu et al. [16] proposed a graph convolutional network to capture label dependency and label structure for multi-label text classification. In addition, Graves at al. [17] and Gangwar et al. [18] utilized the improved recursive model and a complex gated mechanism for text classification. However, such models fail to assign sufficient weights to some important words. Additionally, the attention mechanism introduced by Bengio et al. in machine translation [19] successfully solved this problem. Subsequently, the attention mechanism has been widely used in text classification tasks. Hu et al. [20] proposed a method based on multi-head attention, which achieves good performance for short text classification. Zhang et al. [21] proposed a hierarchical convolutional attention network using joint Chinese word embedding for text classification. These models mainly pay attention to architecture design for feature extraction. While our model merges additional information with semantic features to form a more informative representation.

2.3. Additional knowledge-based text classification

There has been a lot of research on external knowledge-based text classification methods in NLP. After so much research has been done, a class of effective additional information has been formed, such as label confusion, sentiment dictionary, external dictionary, external knowledge and so on. Wang et al. [6] adopted a label confusion model as an enhanced component of the current popular text classification models. Since numeric labels contain much less semantic information than textual labels, the effectiveness of this model decreases when the labels of the dataset are numeric. To address this problem, Liu et al. [7] combined a co-attention network with label embedding that jointly encoded the text and labels into their mutually attended representations. In addition, after incorporating the sentiment dictionary into the model framework for sentiment analysis, Rojas et al. [22] retrieved text definitions of all classes of hierarchical structure from an external dictionary and mapped them to the word vector space. Chen et al. [23] regarded conceptual information as a kind of knowledge and proposed deep short text classification with knowledge powered attention. Gao et al. [24] used the TF-IDF feature of text as the statistical feature and combined it with deep semantic features for legal text classification. However, these works do not take into account the necessity and compatibility of the added information, which may easily bring noise to the classifier. While, our model uses the gating mechanism to selectively fuse statistical information with semantic information to avoid the noise brought by the added information.

3. Joint statistical features and semantic features based classification method

Considering that the statistical information of the text classification methods based on statistical feature is all single, we combine the POS frequency and words frequency over the labels to enrich the statistical features. Therefore, a Word frequency and POS frequency based Network (WPN) that combines joint statistical features with semantic features controlled by a gated mechanism is proposed.

The overall architecture of our model is shown in Figure 1. The model consists of six main modules, which are the Word Frequency (WF) module, the POS Frequency (PF) module, the Variational AutoEncoder Network ($VAEN$), the Semantic Extraction Network (SEN), the Gated Mechanism (GM) and the $Classifier$ module. The WF module is the word frequency statistics module. In our model, the word frequency information is obtained by the Word Frequency Count (WFC) component. Then the word frequency information is mapped to a latent continuous space by the $VAEN$ to generate the global representation of Word Frequency Distribution (WFD) F^T. Besides, the POS Frequency Count (PFC) is performed on the POS tagged texts by the PF module, and the global representation of the POS Frequency Distribution (PFD) F^P is obtained by the $VAEN$, too. Then F^T and F^P are concatenated to get the representation of the joint statistical feature F^D. The eigenvector O^F is generated by F^D passing through a fully connected layer. In the $VAEN$ module, two encoders are employed to generate two sets of μ and σ as the mean and standard deviation of the prior distribution, respectively. The SEN is the semantic feature extraction module of the model, where S represents the semantic feature extracted via a BERT model. Furthermore, O^S is the feature vector after S passes through a dense layer, and O^G is the result of activation of O^S by a sigmoid function. The GM represents the gated mechanism of our model, whose main function is to filter the statistical features and integrate them with the semantic features. Moreover, the feature map O^E is generated by the combined statistical information with the semantic information. The $Classifier$ module is a classifier based on the attention mechanism for label prediction. $MatMul$ and $Dense$ in this component represents matrix multiplication and a dense layer respectively.

3.1. Statistical information based on word and POS frequency

This section mainly introduces the *WF* module and the *PF* module. As a matter of fact, both POS information and word frequency are important for analyzing text features. We enrich the statistical fea-

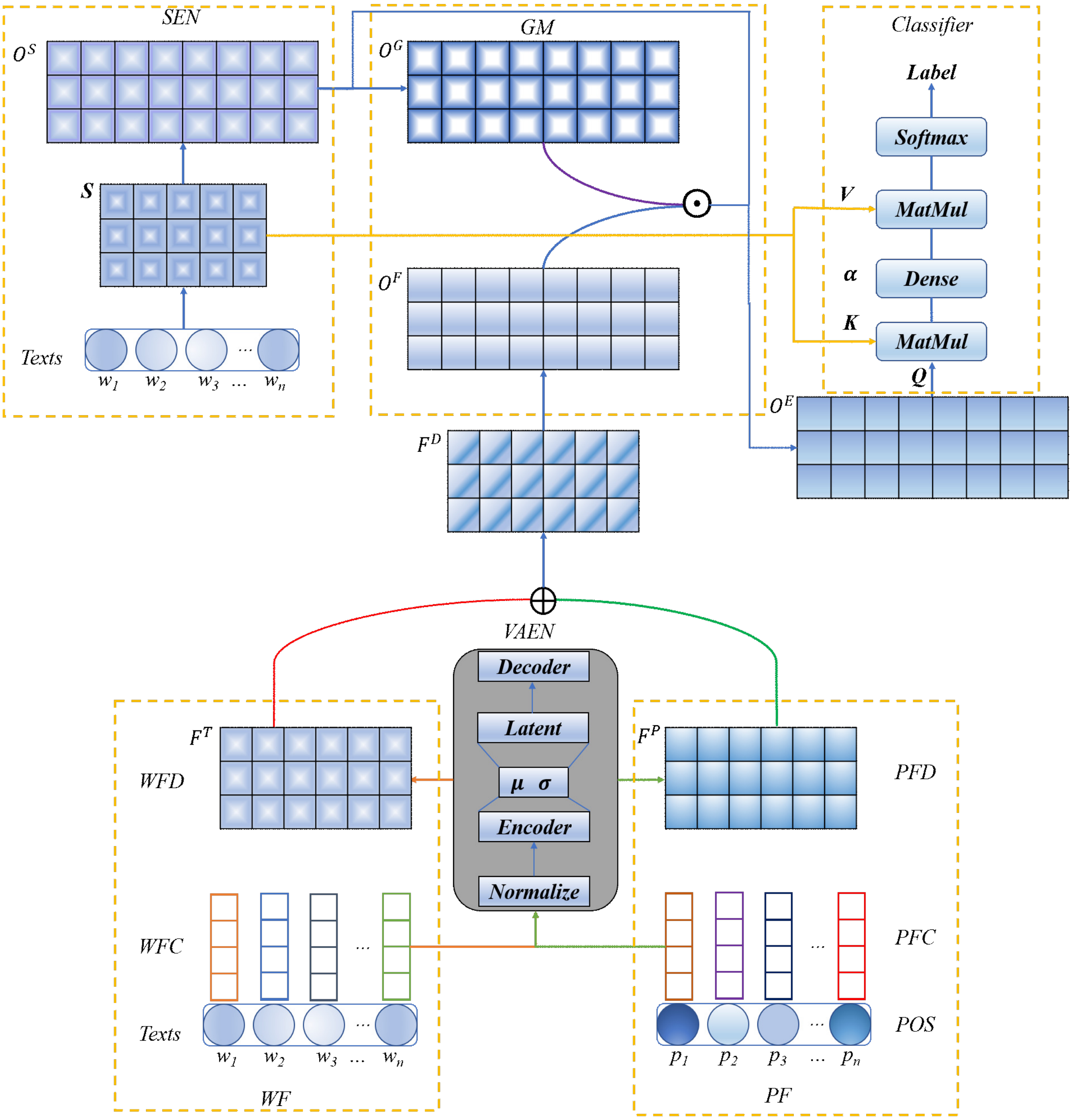

Fig. 1. The structure of WPN method, which is composed of six modules. ($\oplus$ represents concatenation operation. $\odot$ stands for an element-wise product.)

tures to increase the discrimination between different types of texts by combining the POS frequency with the word frequency of texts.

3.1.1. Word frequency count

For a given text sequence $T = [w_1, w_2, \cdots, w_n]$ ($w_i \in T, 0 < i \leqslant n$), w_i represents each word in the text, and n is the maxlength of the input text. Firstly, the WFC module is used for word frequency statistics [1]. Note that the total number of label categories in the dataset is l , the statistical representation of the word w over the label set is:

$$F^w = [c_1^w, c_2^w, \cdots, c_l^w] \tag{1}$$

where c_a^w is the count of word w of label $a(0 < a \leqslant l)$. Thus, the word frequency statistics for a given text sequence T is expressed as:

$$F^T = [F^{w_1}, F^{w_2}, \cdots, F^{w_n}] \tag{2}$$

where F^{w_i} ($F^{w_i} \in F^T, 0 < i \leqslant n$) represents the WFC of w_i. The word frequency is the distribution of the words over different labels across the entire dataset. Unfortunately, it is difficult to distinguish text sequences by counting only the frequency of the words in the dataset. It is only effective when a word has a high word frequency over the corresponding category.

3.1.2. Part-of-speech frequency count

In view of the importance of the POS information in text classification, we add the statistical feature of POS frequency. The POS frequency is the distribution of the POS tag corresponding to the word over different labels across the entire dataset. In this work, we introduce the CoreNLP [25] as a POS tagging tool. As a tagging tool in NLP based on Java, the CoreNLP supports text tagging tasks in 8 languages, including Chinese, English, French, etc. The tagged POS sequence is denoted as $P = [p_1, p_2, \cdots, p_n]$, where p_i ($p_i \in P, 0 < i \leqslant n$) is the corresponding POS representation of w_i. Firstly, we use the PFC module to calculate the frequency distribution of p_i over different labels. The statistical representation of p over the label set is defined as follows:

$$F^p = [c_1^p, c_2^p, \cdots, c_l^p] \tag{3}$$

where c_a^p represents the count of POS p on label $a(0 < a \leqslant l)$. The POS frequency statistics for a given text sequence T is as formula 4 shows:

$$F^P = [F^{p_1}, F^{p_2}, \cdots, F^{p_n}] \tag{4}$$

where $F^{p_i}(F^{p_i} \in F^P, 0 < i \leqslant n)$ represents the PFC of w_i. As one of the most important methods to solve NLP tasks, POS tagging has been widely used in text classification. Similar to the word frequency based classification methods, most researchers simply combine the part-of-speech tags with the semantic features. However, our model combines the word frequency with the POS frequency to form a more discriminative statistical feature.

[1] If it is a Chinese dataset, then the word means a character, for the BERT model is a character-level pretraining mode

3.2. Variational autoencoder network

Although two dimensional statistical features have been obtained above, they are dimensionally incompatible with the semantic features. In order to solve this problem, our model adopts a variational autoencoder network [26] to transform the statistical features into the global representations. As shown in Figure 1, we map the word frequency information and the POS frequency information obtained in section 3.1 to a latent continuous space. By using the $VAEN$ module, we turn the discrete statistics into a more informative vector representation. We notice that the performance of the classifier are improved effectively by bounding the latent space with a multivariate Gaussian distribution. Thus, we adopt the variational autoencoder network rather than a vanilla autoencoder.

We utilize the word frequency count F^T and the POS frequency count F^P to represent the feature distribution $D^T = \{F^T_{(i)}\}_{i=1}^N$ and $D^P = \{F^P_{(i)}\}_{i=1}^N$ of each sentence in the dataset. The loss function of the multivariate Gaussian distribution is shown as Eq. (5):

$$\mathcal{L}(\theta, \phi, F) = \mathbb{E}_{q_\phi(D|F)}[log_{p_\theta(F|D)}] - D_{KL}(q_\phi(D|F))||p_\theta(D) \tag{5}$$

where $p_\theta(F|D)$ represents the latent representation of F^T or F^P and $p_\theta(D)$ is a prior distribution. We adopt $q_\phi(D|F)$ to approximate $p_\theta(D|F)$. $\mathbb{E}$ and D_{KL} indicates expectation and the KL divergence, respectively.

On this basis, we obtain the latent variables F^{D^T} and F^{D^P} through the probabilistic encoder. Moreover, we get the global variable F^D of the joint statistical by concatenating F^{D^T} and F^{D^P}, as shown in Eq. (6):

$$F^D = F^{D^T} \oplus F^{D^P} \tag{6}$$

where $\oplus$ stands for concatenation operator.

The global representation of the joint statistical feature is generated by combining the word frequency distribution with the POS frequency distribution. Note that, the POS frequency and word frequency are primitive primitive statistical features, and we concatenate them as additional information directly. Then we utilize a gated mechanism to selectively merge the statistical features with the semantic features in the GM module. Note that the word frequency count and part-of-speech frequency count modules are only applied during training. In the test set, the model collects tokens for each word in the sentence. Then the token and semantic information are passed into the model trained through the training set to predict the label.

3.3. Feature fusion

In this section, we detail the SEN module and the GM module in Figure 1. The function of the SEN is extracting semantic features from the texts. The input of the SEN is the text sequence T, whose maximum length is n which varies depending on the datasets. In our experiment, we adopt the BERT model to extract semantic features and project them into information space for confidence evaluation. The semantic features generated by the SEN is as formula (7) shows:

$$S = SEN(T) \tag{7}$$

Then we use a fully connected layer to map the semantic information S to an information space:

$$O^S = W^S \cdot S + b^S \tag{8}$$

where W^S represents the weight matrix of S and b^S is the bias. Moreover, we exploit a sigmoid activation function to activate O^S to obtain O^G, and carry out confidence evaluation of the semantic features. In addition, we have obtained the global representation F^D of the combined statistics in the above sections. In order to integrate the statistical features with the semantic features, we project F^D into an information space shared with the semantic features through a dense layer.

$$O^F = W^F \cdot F^z + b^F \tag{9}$$

We combine O^S with O^F through the GM module to output a semantic feature map O^E enhanced by the combined statistics:

$$O^E = ReLU(O^S) + GM(O^G, \varepsilon) \odot O^F \tag{10}$$

where $ReLU(\cdot)$ is an activation function, $\odot$ stands for an element-wise product. The value of O^G in the formula (11) is probabilistic. We use a gated mechanism to adjust the semantic information with poor confidence (with probability close to 0.5) to match the elements in O^F. Assume that $h \in O^G$:

$$GM(h, \varepsilon) = \begin{cases} h, & \text{if } 0.5 - \varepsilon \leqslant h \leqslant 0.5 + \varepsilon \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

where ε is a hyper-parameter that tunes the confidence threshold. When $\varepsilon = 0$, we drop all statistical information. If $\varepsilon = 0.5$, we accept all statistical information. In this way, we get the necessary statistical information through the GM function, so as to better integrate the statistical features with the semantic features.

3.4. Classifier

In this section, as shown in the $Classifier$ module in Figure 1, we use the attention mechanism to combine the feature representation O^E with the original feature map S:

$$Attention(O^E, S) = softmax(O^E S^U)S \tag{12}$$

where $softmax(\cdot)$ is the normalization function, and S^U is the transpose matrix of matrix S. Furthermore, after passing through fully connected layers and a softmax layer, the feature vectors obtained afore are mapped to the label space for label prediction and loss calculation. To maximize the probability of label Y_{True}, we introduce an optimizer to minimize the cross entropy loss L:

$$L = CrossEntropy(Y_{True}, Y_{Pred}) \tag{13}$$

where Y_{True} is the real label of the sequence, and Y_{Pred} represents the predicted value of the model.

4. Experiments and results

4.1. Datasets

To verify the correctness and effectiveness of the proposed method, we adopt 4 commonly used benchmark datasets, including three English datasets and a Chinese dataset. The summary of the datasets is shown in Table 1.

Table 1
Statistics of the four text classification datasets used in our experiments.

Dataset	Class	Length	Train Samples	Test Samples	Task
AG's News [2]	4	45	120000	7600	Topic
THUCNews [27]	13	944	26169	12831	Topic
Yelp P. [2]	2	153	560000	50000	Sentiment
Yelp F. [2]	5	159	650000	38000	Sentiment

AG's News [2] is an English News classification dataset, which contains a total of 127,600 samples with four classes.
THUCNews[27] is a Chinese news classification dataset published by Tsinghua University. We randomly select 39,000 samples and divide them into 13 categories according to the method proposed by Guo at el. [6].
Yelp Review Polarity (Yelp P.) [2] is a binary sentiment classification dataset whose class is either positive or negative. It is obtained from Yelp Dataset challenge 2015.
Yelp Review Full (Yelp F.) [2] is also from Yelp dataset challenge 2015. It is a sentiment classification dataset with five classes, including very negative, negative, neutral, positive and very positive.

4.2. Models in comparison

Our model is compared with seven benchmarks in performance, which contain not only traditional popular methods but also state-of-the-art methods.
BiLSTM [17]: It is a Bidirectional Long Short-Term Memory (BiLSTM) model which extracts sequence features from forward and reverse directions. Moreover, it not only solves the problem of long-term dependence, but also the problem of gradient explosion or gradient disappearance in neural networks.
CNN [1]: CNN is a popular classifier, which uses convolution operation and maximum pooling layer to classify sentences. The convolution operation of the model extracts n-gram features from text well.
BERT [5]: It is a method of pre-training language representations which obtains competitive results in NLP tasks.
LCM [6]: It is a text classification method based on label confusion. The model improves the classification performance by calculating the similarity between instances and labels.
AGN [11]: AGN integrates semantic information with text statistical features through an adaptive network to enhance the classification capability of the model.
BGCapsule [18]: BGCapsule is a novel hybrid model, which is proposed for text classification through an ensemble of Bidirectional Gated Recurrent Units.
CNLE [7]: It is a co-attention network with label embedding, which jointly encodes the text and labels into their mutually attended representations.

4.3. Word embedding and parameter settings

Both Chinese BERT and English BERT are used as the pretrained word embedding, since the experiment involves Chinese and English datasets. Besides, the preprocessing of textual data on all datasets follows that of Kim (2014) [1]. The hyper-parameter settings involved in the experiments are as follows. For the CNN-based models, we use three filters with size 3, 4 and 5, and the number of filters in each convolution block is 100. Dimension of hidden layer of BiLSTM model is set to 128. In addition, the Chinese and English BERT models are BERT-wwm [28] and Bert-base Uncased respectively. Both of them contain 12 layers, 768 hidden units and 110M parameters. For experiments using the BERT model we set the batch size to 16 or less. While, for the non-Bert model the batch size is set to 64. As for the optimizer we exploit the Adam optimizer and set the dropout rate to 0.5.

Table 2

Accuracy comparison among WPN and all competing methods. Higher metric value indicates better model performance. Best results are in bold.

Model	Dataset				Year
	AG's News	THUCNews	Yelp P.	Yelp F.	
BiLSTM [17]	89.25	90.08	94.80	66.91	2013
CNN [1]	89.59	93.31	94.53	66.98	2014
BERT [5]	90.24	95.59	94.53	68.17	2019
LCM [6]	90.42	96.07	93.15	67.64	2021
AGN [11]	93.82	96.86	96.79	70.74	2021
BGCapsule [18]	92.59	94.37	96.62	66.93	2022
CNLE [7]	94.00	96.98	97.13	68.15	2022
WPN(Ours)	**94.21**	**97.18**	**97.22**	**71.28**	2022

4.4. Results and analysis

The results of accuracy and F1-score on AG's News dataset, THUCNews dataset, Yelp P. dataset and Yelp F. dataset are shown in Table 2 and Table 3. The results in the tables show that the performance of our model is superior to that of the 7 benchmarks on these datasets.

Firstly, compared with classical models such as BiLSTM, CNN and BERT, our model WPN outperforms these models in accuracy and F1-score on benchmark datasets. As far as the BERT model is concerned, the results in Table 2 and Table 3 show that the performance of WPN is improved most obviously on AG's News. The accuracy rate is increased by 3.97%, and the F1-score is improved by 3.9%. The least improvement is shown in these two tables, with a 1.59% increase in accuracy and a 2.84% increase in F1-score on THUCNews dataset. According to the statistics in Table 1, the data in THUCNews dataset has the longest average length. And AG's News has the shortest average length of the data. As our model intercepts a sequence of length n as input, part of the original text information is lost. Since the model loses more information on THUCNews dataset than that of other datasets, it has the least performance improvement on this dataset.

Secondly, our model achieves better performance than BGCapsule on all datasets. Specifically, BGCapsule is an ensemble of Bidirectional Gated Recurrent Units, which is devoted to extracting semantic features. It can be seen from Table 2 and Table 3 that the performance difference between this model and our model in the test metrics is the minimal on Yelp P. dataset. While the performance is poor on the

Table 3
F1-score comparison among WPN and all competing methods. Higher metric value indicates better model performance. Best results are in bold.

Model	Dataset				Year
	AG's News	THUCNews	Yelp P.	Yelp F.	
BiLSTM [17]	89.13	91.25	93.59	57.17	2013
CNN [1]	89.10	92.37	92.96	58.22	2014
BERT [5]	90.30	94.33	93.04	61.59	2019
LCM [6]	91.46	95.77	92.35	52.47	2021
AGN [11]	93.79	97.00	95.85	63.01	2021
BGCapsule [18]	92.47	92.76	96.65	62.43	2022
CNLE [7]	93.31	96.33	96.73	62.15	2022
WPN(Ours)	**94.20**	**97.17**	**96.99**	**63.37**	2022

other three multi-class datasets compared to other neural network based methods. The reason is that BG-Capsule only considers the semantic information. It is difficult to perform well with limited information on multi-class tasks.

Thirdly, the performance of our model is superior to that of AGN. AGN adaptively merges the semantic information with the statistical features, making the performance of this model superior to that of the above benchmarks on most datasets. However, just like the traditional statistical feature-based classification methods, AGN only makes use of the word frequency information. Whereas, our model not only utilizes the word frequency information but also adds the POS frequency information and both of these statistical information are collected according to corresponding labels. As POS is an important feature information of text classification, when POS frequency features are combined with word frequency features, a more discriminative statistical feature is formed. This makes the accuracy and F1-score of our model outperform those of AGN on all datasets.

In addition, our method achieves better performance than LCM and CNLE. Both of them integrate label information with the semantic features of text for text classification. Table 2 and Table 3 show that while LCM performs well on news datasets like AG's News and THUCNews, it performs poor on sentiment datasets like the Yelp P. dataset as well as the Yelp F. dataset. The reason is that the model needs to extract the "title" information in the text when incorporating label information. But the above sentiment datasets only contain text sequences and numerical label sets, which leads to poor results. On the contrary, on AG's News dataset and THUCNews dataset, each sample contains title information. As for CNLE, the results show that it achieves better performance than LCM on both the Yelp P. dataset as well as the Yelp F. dataset. However, it only focuses on the semantic information of labels and texts. While our model not only extracts the semantic information of texts, but also selectively integrates multidimensional statistical information with the semantic information. Therefore, our model outperforms CNLE in accuracy and F1-score on the above four datasets.

4.5. Ablation experiment

Considering that WFD and PFD are the most foundational statistical features, our ablation experiments are performed only on WF and PF modules. To validate the effectiveness and contribution of the word frequency information and the POS frequency information in our model, we conduct the ablation study on all four datasets. The results are shown in Table 4. In this table, WPN(N) means WPN model without statistical information. And WPN(W) indicates WPN model with the word frequency information. In additon, WPN(P) indicates WPN model with the POS frequency information. WPN(F) means

Table 4

Ablation study on four datasets. Higher metric value indicates better model performance. Best results are in bold.

Model	AG's News		THUCNews		Yelp P.		Yelp F.	
	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
WPN(N)	93.29	93.30	95.62	95.33	95.97	95.97	68.93	61.74
WPN(W)	93.85	93.79	96.86	96.94	96.58	96.42	70.25	62.33
WPN(P)	93.51	93.50	96.64	96.78	96.19	96.03	69.92	62.06
WPN(F)	**94.21**	**94.20**	**97.18**	**97.17**	**97.22**	**96.99**	**71.28**	**63.37**

WPN model with both the word frequency information and the POS frequency information. As shown in Table 4, compared with WPN(N), the accuracy rate of WPN(W) is improved by 0.56%, and the F1-score is increased by 0.49% on AG's News dataset. Generally, the results on these four datasets show that both the word frequency information and the POS frequency information improve the performance of our model. However, the improvement made by the word frequency or the POS frequency on both datasets is limited. As shown in Table 4, after the word frequency information is combined with the POS frequency information, compared with WPN(N), the accuracy rate is improved by 1.56%, and the F1-score is improved by 1.84% on THUCNews dataset. Moreover, in comparison to WPN(N), the accuracy of our model is improved by 2.35% and the F1-score is improved by 1.63% on Yelp F. dataset. Obviously, as we merge the word frequency with the POS frequency to form a more discriminative feature, the performance of our model is further improved.

4.6. *Efficiency comparison*

Table 5 shows a comparison of different models in terms of execution time. It can be seen from the table that BiLSTM and CNN spend the least training time because their model structure is relatively simple and the number of parameters is small compared with other models. BGCapsule and CNLE require slightly more execution time than BiLSTM and CNN do because they have more parameters than BiLSTM and CNN. The rest of the models are based on BERT so the parameters are the same. Since our model needs to count word frequency and part-of-speech frequency information, it requires additional time compared with other BERT-based models. However, our model has a richer feature representation and achieves the best experimental results compared to the few time increases. For tasks with large text classification applications, it is worth spending a small amount of time to achieve better results.

Table 5

Comparison in terms of execution times.

Model	Parameter amount	Total time(second)
BiLSTM [17]	0.19M	20
CNN [1]	0.44M	**20**
BERT [5]	110M	90
LCM [6]	110M	95
AGN [11]	110M	110
BGCapsule [18]	8.20M	58
CNLE [7]	2.15M	25
WPN(Ours)	110M	115

5. Conclusion

In this paper, we propose a text classification model based on the combination of semantic features and joint statistical features, whose statistical features consist of the word frequency and the POS frequency over all labels. In this model the word frequency statistical features and POS frequency statistical features are mapped to a latent continuous space by the $VAEN$. Subsequently, the global representation of the joint statistical features is obtained. Furthermore, the statistical features are selectively fused with the semantic features through a gated mechanism. Since our model incorporates both multivariate statistical features and semantic features of the text, our model outperform these 7 benchmarks in performance on all 4 datasets. In the future, we will add prior knowledge to our method to improve the performance of the model on small datasets.

Acknowledgements

We appreciate the Open Project Program of Shanghai Key Laboratory of Data Science (No. 2020090600004), and thank the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600) for providing the computing resources. We also thank YTO express company for supporting the data and industry knowledge.

References

[1] Y. Kim, Convolutional Neural Networks for Sentence Classification, in: *EMNLP*, 2014, pp. 1746–1751. doi:10.3115/v1/d14-1181.

[2] X. Zhang, J.J. Zhao and Y. LeCun, Character-level Convolutional Networks for Text Classification, in: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, 2015, pp. 649–657.

[3] P. Liu, X. Qiu and X. Huang, Recurrent Neural Network for Text Classification with Multi-Task Learning, in: *IJCAI*, 2016, pp. 2873–2879. http://www.ijcai.org/Abstract/16/408.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, Attention is All you Need, in: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[5] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

[6] B. Guo, S. Han, X. Han, H. Huang and T. Lu, Label Confusion Learning to Enhance Text Classification Models, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 12929–12936. https://ojs.aaai.org/index.php/AAAI/article/view/17529.

[7] M. Liu, L. Liu, J. Cao and Q. Du, Co-attention network with label embedding for text classification, *Neurocomputing* **471** (2022), 61–69. doi:10.1016/j.neucom.2021.10.099.

[8] Y. Zhang, Z. Zhang, M. Chen, H. Lu, L. Zhang and C. Wang, LAMB: A novel algorithm of label collaboration based multi-label learning, *Intell. Data Anal.* **26**(5) (2022), 1229–1245. doi:10.3233/IDA-215946.

[9] D. Kim, D. Seo, S. Cho and P. Kang, Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec, *Information Sciences* **477** (2019), 15–29. doi:10.1016/j.ins.2018.10.006.

[10] M. Hu, J. Peng, W. Zhang, J. Hu, L. Qi and H. Zhang, Text Representation Model for Multiple Language Forms in Spoken Chinese Expression, *International Journal of Pattern Recognition and Artificial Intelligence* (2021). doi:10.1142/S0218001422530044.

[11] X. Li, Z. Li, H. Xie and Q. Li, Merging Statistical Feature via Adaptive Gate for Improved Text Classification, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press, 2021, pp. 13288–13296. https://ojs.aaai.org/index.php/AAAI/article/view/17569.

[12] S.F. Yazdani, Z. Tan, M. Kakavand and A. Mustapha, NgramPOS: a bigram-based linguistic and statistical feature process model for unstructured text classification, *Wireless Networks* **28**(3) (2022), 1251–1261. doi:10.1007/s11276-018-01909-0.

[13] P. Bouboulis, S. Theodoridis, C. Mavroforakis and L. Evaggelatou-Dalla, Complex Support Vector Machines for Regression and Quaternary Classification, *IEEE Transactions on Neural Networks and Learning Systems* **26**(6) (2015), 1260–1274. doi:10.1109/TNNLS.2014.2336679.

[14] H.-J. Kim, J.-U. Kim and Y.-G. Ra, Boosting Naïve Bayes text classification using uncertainty-based selective sampling, *Neurocomputing* **67** (2005), 403–410. doi:https://doi.org/10.1016/j.neucom.2004.09.003.

[15] P. Li, P. Zhong, K. Mao, D. Wang, X. Yang, Y. Liu, J. Yin and S. See, Act: an attentive convolutional transformer for efficient text classification, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 13261–13269.

[16] T. Pu, S. Yin, W. Li and W. Xu, Graph Convolutional Network Exploring Label Relations for Multi-label Text Classification, in: *PRICAI*, Vol. 13032, 2021, pp. 127–139. doi:10.1007/978-3-030-89363-7_10.

[17] A. Graves, N. Jaitly and A.-r. Mohamed, Hybrid speech recognition with deep bidirectional LSTM, in: *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, 2013, pp. 273–278.

[18] A.K. Gangwar and V. Ravi, A Novel BGCapsule Network for Text Classification, *SN Computer Science* **3**(1) (2022), 81. doi:10.1007/s42979-021-00963-4.

[19] D. Bahdanau, K. Cho and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, in: *ICLR*, 2015. http://arxiv.org/abs/1409.0473.

[20] J. Hu, J. Peng, W. Zhang, L. Qi, M. Hu and H. Zhang, Intention Multiple-Representation Model for Logistics Intelligent Customer Service, in: *Knowledge Science, Engineering and Management - 13th International Conference, KSEM 2020, Hangzhou, China, August 28-30, 2020, Proceedings, Part I*, Vol. 12274, 2020, pp. 186–193. doi:10.1007/978-3-030-55130-8_16.

[21] K. Zhang, S. Wang, B. Li, F. Mei and J. Zhang, Hierarchical Convolutional Attention Networks Using Joint Chinese Word Embedding for Text Classification, in: *PRICAI 2019: Trends in Artificial Intelligence*, A.C. Nayak and A. Sharma, eds, Cham, 2019, pp. 234–246.

[22] K.R. Rojas, G. Bustamante, A. Oncevay and M.A.S. Cabezudo, Efficient Strategies for Hierarchical Text Classification: External Knowledge and Auxiliary Tasks, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2252–2257. doi:10.18653/v1/2020.acl-main.205.

[23] J. Chen, Y. Hu, J. Liu, Y. Xiao and H. Jiang, Deep Short Text Classification with Knowledge Powered Attention, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 6252–6259. doi:10.1609/aaai.v33i01.33016252.

[24] J. Gao, H. Ning, Z. Han, L. Kong and H. Qi, Legal text classification model based on text statistical features and deep semantic features, in: *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020*, P. Mehta, T. Mandl, P. Majumder and M. Mitra, eds, CEUR Workshop Proceedings, Vol. 2826, CEUR-WS.org, 2020, pp. 35–41. https://ceur-ws.org/Vol-2826/T1-7.pdf.

[25] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard and D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 55–60. doi:10.3115/v1/p14-5010.

[26] D.P. Kingma and M. Welling, Auto-Encoding Variational Bayes, in: *ICLR*, 2014. http://arxiv.org/abs/1312.6114.

[27] M. Sun, J. Li, Z. Guo, Y. Zhao, Y. Zheng, X. Si and Z. Liu, THUCTC: An Efficient Chinese Text Classifier (2016). http://thuctc.thunlp.org.

[28] Y. Cui, W. Che, T. Liu, B. Qin and Z. Yang, Pre-Training With Whole Word Masking for Chinese BERT, *IEEE ACM Trans. Audio Speech Lang. Process.* **29** (2021), 3504–3514. doi:10.1109/TASLP.2021.3124365.

[29] P. Li and K. Mao, Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts, *Expert Syst. Appl.* **115** (2019), 512–523. doi:10.1016/j.eswa.2018.08.009.

[30] D. Tang, B. Qin and T. Liu, Document Modeling with Gated Recurrent Neural Network for Sentiment Classification, in: *EMNLP*, 2015, pp. 1422–1432. https://doi.org/10.18653/v1/d15-1167.

[31] Z. Yang, D. Yang, C. Dyer, X. He, A.J. Smola and E.H. Hovy, Hierarchical Attention Networks for Document Classification, in: *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489. doi:10.18653/v1/n16-1174.

[32] S.S. Samant, N.L.B. Murthy and A. Malapati, Improving Term Weighting Schemes for Short Text Classification in Vector Space Model, *IEEE Access* **7** (2019), 166578–166592. doi:10.1109/ACCESS.2019.2953918.

[33] T. Pranckevicius and V. Marcinkevicius, Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification, *Balt. J. Mod. Comput.* **5**(2) (2017). doi:10.22364/bjmc.2017.5.2.05.

[34] S.R. Gunn et al., Support vector machines for classification and regression, *ISIS technical report* **14**(1) (1998), 5–16.

[35] S. García-Méndez, M. Fernandez-Gavilanes, J. Juncal-Martínez, F.J. González-Castaño and Ó.B. Seara, Identifying banking transaction descriptions via support vector machine short-text classification based on a specialized labelled corpus, *IEEE Access* **8** (2020), 61642–61655.

[36] Q. Lu and Y. Wang, Latent semantic text classification method research based on support vector machine, *International Journal of Information and Communication Technology* **15**(3) (2019), 243–255.

[37] A. Mccallum and K. Nigam, A comparison of event models for Naive Bayes text classification, 1998, pp. 41–48.

[38] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural computation* **9**(8) (1997), 1735–1780.

[39] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv preprint arXiv:1412.3555* (2014).

[40] F. Yin, Z. Yao and J. Liu, Character-Level Attention Convolutional Neural Networks for Short-Text Classification, in: *Human Centered Computing - 5th International Conference, HCC 2019, Čačak, Serbia, August 5-7, 2019, Revised Selected Papers*, D. Milosevic, Y. Tang and Q. Zu, eds, Lecture Notes in Computer Science, Vol. 11956, Springer, 2019, pp. 560–567. doi:10.1007/978-3-030-37429-7_57.

[41] M. Tezgider, B. Yildiz and G. Aydin, Text classification using improved bidirectional transformer, *Concurrency and Computation: Practice and Experience* **34**(9) (2022). doi:10.1002/cpe.6486.

[42] D. Muñoz-Valero, L. Rodriguez-Benitez, L. Jimenez-Linares and J. Moreno-Garcia, Using Recurrent Neural Networks for Part-of-Speech Tagging and Subject and Predicate Classification in a Sentence, *International Journal of Computational Intelligence Systems* **13** (2020), 706–716. https://doi.org/10.2991/ijcis.d.200527.005.

[43] J. Hu, J. Peng, W. Zhang, L. Qi, M. Hu and H. Zhang, An intention multiple-representation model with expanded information, *Computer Speech and Language* **68** (2021), 101196. doi:10.1016/j.csl.2021.101196.