Research on news text classification based on improved BERT-UNet model

Zeqin Li
Neusoft Institute Guangdong Foshan
China
leezeqin@163.com

Jianwen Liu*
Neusoft Institute Guangdong Foshan
China
840118513@qq.com

Jin Lin
Neusoft Institute Guangdong Foshan
China
linjin@nuit.edu.cn

Deli Tan
Neusoft Institute Guangdong Foshan
China
tandeli@nuit.edu.cn

Ruyue Gong
Neusoft Institute Guangdong Foshan
China
gongruyue@nuit.edu.cn

Linglin Wang
Neusoft Institute Guangdong Foshan
China
2474599462@qq.com

ABSTRACT

The classification of news text is crucial for various applications such as information retrieval, sentiment analysis, and intelligence gathering. In light of the limitations of conventional convolutional neural networks in news text classification, this study introduces an enhanced BERT-UNet model for improved long-distance text feature extraction. Initially, the model leverages BERT for pre-training the text word vectors, followed by embedding and mapping them onto the UNet architecture to extract contextual key features. The Softmax function is then utilized for news opinion text categorization. To validate the model's performance, comparative experiments are conducted on the THUCNews dataset. The results indicate that the BERT-UNET model outperforms the standard TextCNN model and standalone BERT approach with a 3.11% and 0.29% increase in macro average F1 value, respectively. These findings demonstrate the effectiveness of the enhanced BERT-UNet model in capturing textual relationships, offering a fresh perspective on enhancing traditional news text classification methods.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Natural language processing; Information extraction.

KEYWORDS

BERT-UNet model, Deep learning, News text classification, Natural language processing

ACM Reference Format:
Zeqin Li, Jianwen Liu*, Jin Lin, Deli Tan, Ruyue Gong, and Linglin Wang. 2024. Research on news text classification based on improved BERT-UNet model. In *International Conference on Modeling, Natural Language Processing and Machine Learning (CMNM 2024), May 17–19, 2024, Xi'an, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3677779.3677780

CMNM 2024, May 17–19, 2024, Xi'an, China

ACM ISBN 979-8-4007-0976-0/24/05
https://doi.org/10.1145/3677779.3677780

1 INTRODUCTION

Text classification, a fundamental technique in Natural Language Processing (NLP), is employed for categorizing textual data into predefined categories or labels. With the rapid development of information technology, network information shows an exponential growth trend. As the carrier of information, text is an important medium of communication, covering speech, news, papers and other forms. Text classification algorithms need to extract feature information from sentences for model training, and quickly identify the semantic and emotional content of sentences. It has become an urgent need to extract valuable information from the massive network text to meet the needs of users. Traditional text classification methods mainly build classification models through machine learning. Common algorithms are as follows: Support Vector Machine [1] (SVM), Naive Bayes [2] (Naive Bayes, NB), k-Nearest Neighbor algorithm [3] (K-nearest Neighbor, KNN), etc. However, text classification is often difficult to achieve good results in machine learning. When processing text information, because the text itself is high-dimensional and has semantic and emotional characteristics [4], there are problems in common model training tasks such as sparse features and inability to capture network connections, resulting in classification tasks failing to process text data well. In order to adapt the text classification model to Chinese text data set, this paper introduces an improved BERT-UNet (Bidirectional Encoder Representations from Tranformers) model for news text classification. Through this model, the number of categories of text label information in the text base is successfully classified. The design of the system structure not only improves the adaptability of the text classification model to Chinese text, but also plays an auxiliary role in the monitoring of news public opinion.

2 RELATED RESEARCH

Text categorization using CNNS has advanced significantly with the advancement of natural language processing (convolutional neural networks). Kim et al. made the initial proposal for the TextCNN model [5]. Based on the CNN model, the author enhanced the model. First, the text is transformed into a vector representation and fed into the model. Next, convolution kilns of varying sizes pick up local textual aspects, while pooling processes preserve important information. TextCNN offers clear advantages in local feature capture because it primarily uses the convolution operation of local features to capture text information. The text length restriction of the classification problem arises from the difficulty in capturing the long-distance dependency when processing text data due to the fixed length of the convolution kernel. On the 20News, Fuden, ACL, and SST data sets, the Recurrent Convolutional Neural Network (RCNN) model developed by Siwei Lai et al. achieves superior text classification impact than the CNN model [6]. The output layer, pool layer, and RCNN make up the model. A bidirectional recurrent neural network, which is based on CNN, is utilized to compensate for the difficulty in gathering text data with a fixed window size. Text input with complex structures can be processed by graph neural networks (GNN) as opposed to sequential recurrent neural networks. A text categorization approach based on text-level GNN was proposed by Huang et al. [7]. By establishing a shared network and sharing text parameters, this strategy gets rid of the reliance between specific texts and the corpus as a whole. Furthermore, a graph neural network is constructed with a smaller window to extract more precise local information. A text-GCN text categorization approach was proposed by Yao et al. [8]. The main idea behind this model is to create a text graph within the corpus, where words and documents are represented by nodes and edges. The graph model data is initialized using a single thermal coding technique, and embedded representations of words and documents are trained and supervised using labels for known document categories. As technology has advanced, deep learning models with a lot of parameters have been put forth. The Transformer [9] paradigm was put up by the Google team in 2017. Its main concept is to interpret input data solely via the Self-Attention mechanism. Conventional Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures are not used in it. The model is constrained by the length of the context, though, as it struggles with feature-dependent learning of long sequence text. To expand long-distance dependence, Dai et al. [10] presented the Transformer-XL architecture, which enhances the Transform model by adding relative position embeddings and loop mechanisms. This enhancement raises the text sequence limit in addition to improving the learning capacity of long-distance feature extraction of text. Several additional models have emerged in conjunction with the BERT model, which is based on the Transformer bidirectional encoder and has contributed to the development of natural language processing models. To further improve performance, researchers actively investigate various model designs in the text categorization problem. To complete sentence-level feature vector representation of text, Kaur K et al. suggested a text classification method that used a BERT pre-trained language model [11]. Local features in the text were captured using a CNN-based local feature convolution module. The method performed well on the PROMISE dataset. To address the issue of news classification of lengthy Chinese texts, Chen X et al. created a model called LFCN [12]. In the text-text Encoder (TTE) layer, the global feature relationships are extracted using BERT's multilayer bidirectional Transformer feature extractor, and the text is transformed into input vectors using the BERT embedding approach. The model has a convolutional function at the Local Feature Convolutional (LFC) layer. Local salient characteristics are extracted using convolution processes. Convolutional neural networks are employed in the aforementioned two techniques to achieve text classification following BERT. Spatial information will be lost during convolutional neural network computation, even

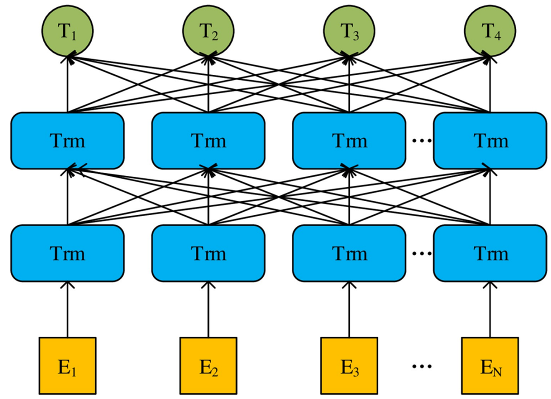

Figure 1: BERT model structure diagram.

though it can produce good local information feature extraction. Consequently, this research introduces a deconvolution layer for up-sampling in an attempt to reconstruct the spatial structure and features of the input sequence. to capture the text's feature information and lessen information loss. Thus, this work proposes the BERT-UNET model, which refines the BERT model based on prior research.

3 BERT-UNET MODEL DESIGN

3.1 BERT word vector model

The Google AI Language team proposed the BERT (Bidirectional Encoder Representations from Transformers) model, which is based on Transformers and was developed in response to the advancements in deep learning [13]. The BERT model structure is displayed in Figure 1 and the model is stacked with several identical Transformer encoders.

Transformer model consists of Encoder and Decoder. In BERT model, only Encoder part of Transformer model is concerned, so Decoder is not discussed in this paper. Figure 2 shows the structure of the encoder layer. Six identical encoder layers make up the encoder end, and each one has a Multi-Head Self-Attention mechanism that enables the model to focus on different parts of the input text simultaneously to better capture context dependencies [14]. Furthermore, every encoder layer incorporates a Feedforward Neural Network to do further processing on the self-attention mechanism's output. To maintain gradient stability and information flow, each sublayer's output is applied with a residual connection, meaning that the output is added to the sublayer's input. After residuals are connected, layer standardization is applied to each sublayer's output, helping to stabilize the data flow in each layer, including the Transformer model Encoder layer.

BERT Model adopts two pre-training methods, Masked Language Model (MLM) and Next Sentence Prediction (NSP). In the MLM task, BERT will choose random words to cover up, and then try to predict these marked words, helping the model understand lexical relationships in the context and how to fill in the missing information. The natural language processing (NLP) task requires the BERT model to determine whether there is a textual relationship between two input sentences, that is, whether they constitute coherent and

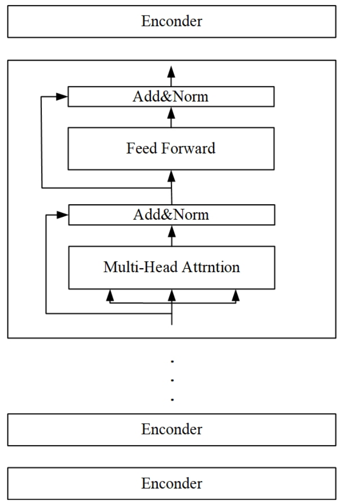

Figure 2: Encoder layer of the Transformer model.

related sentences. This task helps the model learn deep semantic connections between texts. The semantics of different sentences vary depending on the order in which they are arranged in the text sequence, so individual words need to be encoded.

As shown in Figure 3, [CLS] represents the identifier for text sequence input, while [SEP] is used to separate sentences. In a full word embedding process, a Token is the smallest unit of text data, typically represented as a word, subword, or character-level element. The BERT model breaks the input text into multiple tokens and encodes them, each Token having an embedded representation in a vector space, a process that transforms the text containing semantic information into a form that can be processed by a computer. Usually, a text consists of two or more sentences, which can be regarded as a collection of sentences, which may have interactive relationships with each other or may be independent of each other. Therefore, the concept of Segment is introduced so that each Token is assigned a Segment ID to identify whether it belongs to a particular paragraph, so that the model can better understand the relationship between different sentences or paragraphs in the text. In natural language text, in order to accurately record the Position information of words, additional position embeddings need to be introduced to assist the model in understanding the relative position information of different tokens in the text. Each Position embed is an independent vector that identifies position information in a sequence of text, and it plays a crucial role in the model's understanding of context and structure. Since the computer can only recognize the string vector composed of 0 and 1 [15], it is necessary to vector transform the original data text data. Therefore, in the complete text input of BERT model, there are three important Embedding vectors, namely Token Embedding, Segment Embedding, and Position Embedding.

3.2 UNet model

UNet neural networks initially achieved remarkable success in the field of medical image segmentation, showing excellent performance [16]. In this paper, by introducing UNet structure into the design of text classification model, as shown in Figure 4, the multi-dimensional convolution process in image segmentation is transformed into one-dimensional sequential convolution, so that the text classification model can capture the multi-level and multi-scale semantic information in the text, so as to enhance the model's sensitivity to the internal structure of the text. In the design of the convolution layer, a three-layer convolution structure is adopted, the size of the convolution kernel decreases layer by layer to 7, 5 and 3, and the padding is used for zero filling to ensure the consistency of the dimensions after the convolution operation of each layer, and then the ReLU function is used to activate the convolution layer. After the three-layer convolution operation, the feature layer after convolution is vertically spliced, so that the tensor is restored to the original 768 layers, so as to combine the feature information extracted from the convolution kernel of different sizes to form a richer representation. This multi-size convolution operation helps to capture semantic information at different scales, thus improving the model's ability to understand the input text. Deconvolution, also known as upsampling or deconvolution layer, aims to enlarge the spatial dimension of the feature map to the size of the original input. The Deconv layer and the Conv layer have similar operations. After the feature extraction and concatenation of convolutional kernel, the model performs three layers of deconvolution operation, with the size of convolutional kernel increasing layer by layer with values of 3, 5 and 7, respectively. ReLU function is used to activate the deconvolution layer. After deconvolution operation, the feature graph is also vertically spliced, and the data tensor is restored to 768. By upsampling technique, these reduced size feature maps are accurately restored to the original input space. This operation not only preserves more semantic information, but also improves the model's understanding of global and local features. By introducing multi-dimensional convolution and deconvolution layers, the

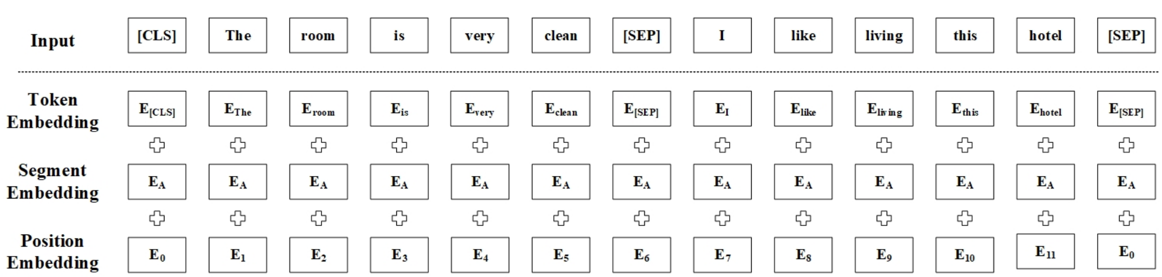

Figure 3: BERT input sequence structure diagram in the model.

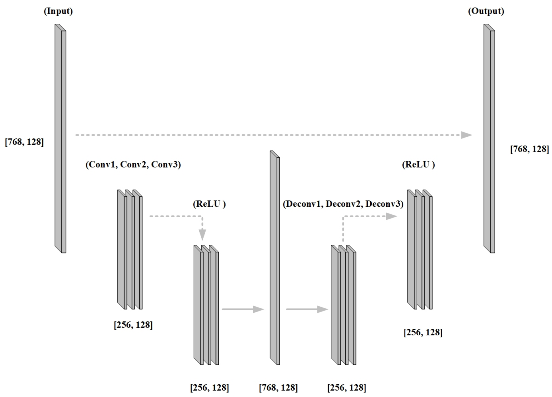

Figure 4: UNet model structure diagram.

model structure designed in this paper enables the model to understand the text at different levels of abstraction. At the same time, the use of up-sampling operation effectively retains more spatial information, and provides strong support for capturing multi-scale semantic information in text, which has important practical significance and application value for text classification and other tasks.

3.3 BERT-UNet model

The BERT-UNet model, displayed in Figure 5, initially decomposes text data into multiple tokens through word embedding using the BERT model. After passing through the multi-layer Transformer model Encoder, mapping relationships are trained between words. The model then extracts feature relationships of the training text sequence and finally converts it into vector expressions [17]. To enhance feature extraction ability for deep text, the initial vector data is embedded into the UNet model within the BERT pre-training corpus. The dimension of BERT word vector output is 768 layers, so the input channel of the UNet model is also designed as 768 layers. Following convolution and deconvolution operations in UNet, the word vector enters a fully connected layer through global average pooling. The tensor at this point is [hidden_size, num_classes]. Finally, the word vector passes through the Softmax function layer for information extraction, yielding the predictive relationship between text and label. This enhanced model structure improves BERT performance in text classification tasks by better capturing correlation and dependency information in texts, adapting well to such tasks.

3.4 Architecture Overview Model

This work builds a BERT-UNet -based news text categorization model. An enhanced model based on BERT is called the BERT-UNet model. The word vector representation module and the deep text feature extraction classification module make up the majority of the model [17]. The word vector conversion module is a BERT-based pre-training model for Chinese that translates text into word vectors and carries out basic feature extraction. Using the UNet network, the deep text feature extraction classification module extracts local key information and contextual deep features before obtaining the classification results. Figure 6 depicts the system design model's structure. To ensure that the system can receive and analyze the most recent news data promptly, the BERT-UNet model uses a time series-based data-gathering method during the database extraction stage. This helps to realize the effectiveness of public opinion monitoring. With the use of this strategy, the system is better able to gather social hot news, resulting in timely information that supports risk management and decision-making.

4 EXPERIMENTS AND RESULT

4.1 Experimental data set

Comparative experiments on the BERT-UNet model with the THUC-News data set, which is accessible to the public, are presented in this work. This dataset, which includes around 830,000 news documents, was created by filtering historical data from Sina News RSS subscription channels between 2005 and 2011, as shown in Table 1. There are fourteen categories in the data set: horoscopes, games, entertainment, education, technology, society, fashion, sports, real estate, lottery, and home. All the data will be organized into data sets for this experiment, and the training, verification, and test sets will be split in an 8:1:1 ratio. Ultimately, the information will be loaded into the corresponding files train.txt, val.txt, and test.txt.

4.2 Experimental environment

The operating system of this experiment is Ubuntu-20.04, and the GPU cloud server with NVIDIA A10 computing card resources is selected as the computing operation platform in the hardware environment. The Python version is 3.10, the deep learning framework PyTorch is Torch2.0.1, and the CUDA version is 11.8. The detailed

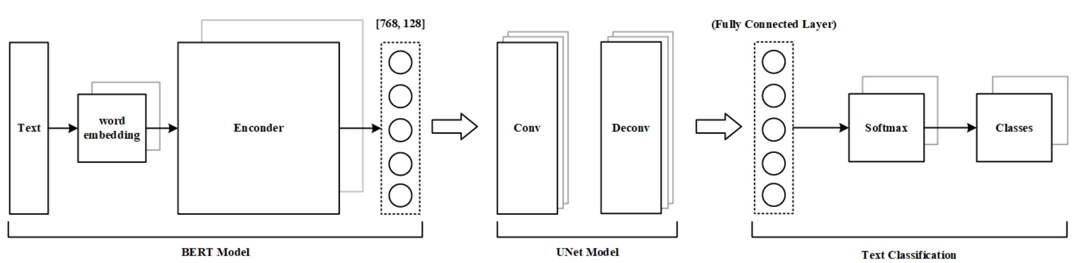

Figure 5: BERT-UNet model structure diagram.

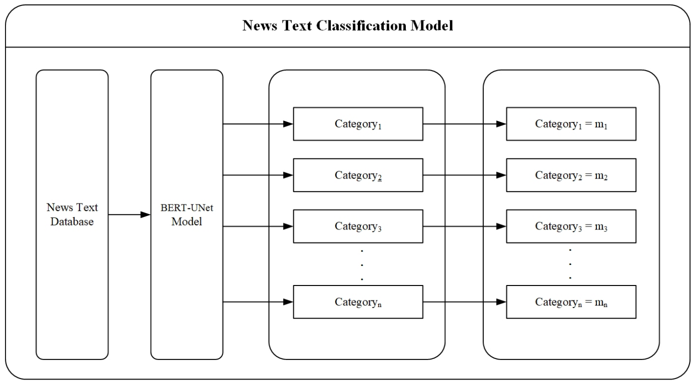

Figure 6: Overall model structure diagram.

Table 1: Dataset Classification Category Sample Quantity

Sample Category	Sample Quantity
Finance	37098
Lottery	7588
Real Estate	20050
Stocks	154398
Furniture	32586
Education	41936
Technology	162929
Society	50849
Fashion	13368
Politics	63086
Sports	131604
Horoscope	3578
Games	24373
Entertainment	92632

Table 2: Experimental Environment Configuration

Experimental Environment	Configuration
Operating System	Ubuntu-20.04
GPU	NVIDIA A10
Python	3.10
PyTorch	2.0.1
CUDA	11.8

configuration of the experimental environment is shown in Table 2.

4.3 Parameter selection

By changing the values of the aforementioned parameters in this experiment, a model with high classification accuracy is produced without taking running time or memory into account. The crucial decisions made during model training in the BERT-UNet pretraining model are exemplified by the parameter configuration presented in Table 3. These choices-which include the word embedding dimensionality of 768, batch_size of 128 and two epochs, learning_rate of 5e-5, use of the Adam optimizer, pad_size of 32, and dropout rate of 0.1-reflect careful thought processes focused on maximizing generalization performance, computational efficiency, and model convergence. By adjusting the model's parameters, the training procedure is customized to strike a balance between computational resources and model complexity, resulting in reliable and efficient learning outputs.

4.4 Experimental criteria

In the task of text classification in natural language processing (NLP), Accuracy, Precision, Recall and F1 Score are often used to

Table 3: Parameter configuration table.

Parameters	Values
Learning Rate	5e-5
Word Embedding	768
Batch Size	128
Epoch	2
Optimizer	Adam
Pad Size	32
Dropout	0.1

evaluate the text classification model. Taking into account the possible class imbalance in the actual problem, we further use Macro Average and Weighted Average as evaluation indicators to ensure that the impact of each class is fully considered in the multi-class classification task. Macro averaging calculates the index values individually for each category and then averages them. It treats each category equally and is suitable for situations where the number of different samples varies greatly. Weighted average The index value of each category is weighted by the proportion of the number of samples in each category to the total number of samples. This method is applicable to the situation where the number of different samples is uneven. The comprehensive evaluation of each index aims to provide a comprehensive evaluation of the model's performance and ensure the data performance of the text classification model in all aspects.

4.5 Analysis of experimental results

In this section, the experimental results of different text classification models on the THUCNews dataset are analyzed in detail. Table 4 shows the performance metrics of each model in Accuracy, Macro Average, and Weighted Average, including Precision, Recall, and F1 scores. Objects for comparison of models include: BERT-CNN, BERT, Transformer, TextCNN, TextRNN, TextRCNN. BERT-UNet, BERT-CNN, and BERT all achieved over 96% accuracy, as well as excellent accuracy, recall, and F1 scores on macro and weighted averages, demonstrating their strong performance on text classification tasks. Notably, BERT-UNet improved its macro average F1 score by 0.19 percentage points and weighted average F1 score by 0.15 percentage points relative to BERT-CNN. Compared with the original BERT model, BERT-UNET improves the macro average F1 score by 0.29 percentage points and the weighted average F1 score by 0.23 percentage points, respectively, showing better overall performance. The Transformer model exhibits relatively low performance compared to BERT-UNet based models. There was a slight decrease in accuracy and F1 scores, with macro average F1 and weighted average F1 scores decreasing by 3.72 percentage points and 2.82 percentage points, respectively. Although TextCNN, TextRNN and TextRCNN are slightly less accurate than BERT-based models, they still achieve good performance. On the macro average F1 score, TextCNN reaches 93.32%, TextRNN reaches 94.12%, and TextRCNN reaches 94.40%. On the weighted average F1 score, TextCNN reached 93.34%, TextRNN reached 95.02%, and TextRCNN reached 95.27%. This shows that the traditional model is still competitive in the task of text classification, and this experimental result provides useful insights for selecting a suitable text classification model, and also provides a reference for subsequent research.

5 CONCLUSION

This research proposes an enhanced BERT-UNet news text categorization model. To achieve text classification of Chinese data, the BERT-UNet model should be based on Chinese corpus data. This would address the issue of classic convolutional neural networks' inefficiency in classifying news texts. After initial training, the word vector data is first obtained by training the BERT model's mapping relationship between words. The word vector data is then transferred to the graph UNet model to obtain the word vector data following the extraction of feature information. Lastly, the Softmax function layer is used to retrieve the label information. In this way, the model can extract text feature information and global text information in news text classification tasks, and improve the accuracy of text classification. Compared with the traditional model, the BERT-UNet model is more effective in news text classification and provides strong support for the task of news text classification.

ACKNOWLEDGMENTS

This work was supported by Guangdong Youth Innovative Talent Project (2023KQNCX109), Guangdong College Students Science and Technology Innovation Cultivation Special Fund Project (pdjh2022b0607).

REFERENCES

[1] Y. Wang, S. Shanshan, L. Chang, *et al.* 2020. Classification model based on support vector machine for Chinese extremely short text. Application Research of Computers, 37(02), 347-350. DOI: 10.19734/j.issn.1001-3695.2018.06.0514.

[2] Wen Y, Liang Y, Zhu X. 2023. Sentiment analysis of hotel online reviews using the BERT model and ERNIE model—Data from China. PLoS ONE 18(3): e0275382. https://doi.org/10.1371

[3] C. Xin and Q. Zhan zhi. 2021. Research on Chinese Text Classification Based on WAE and SVM. In 2021 3rd International Conference on Natural Language Processing (ICNLP) (Beijing, China, 2021), 14–19. DOI: 10.16208/j.issn1000-7024.2022.01.024.

[4] Y. Lin and J. Wang, "Research on text classification based on SVM-KNN," 2014 IEEE 5th International Conference on Software Engineering and Service Science, Beijing, China, 2014, pp. 842-844.

[5] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Doha, Qatar, 2014), 1746–1751. Association for Computational Linguistics.

[6] S. Lai, L. Xu, K. Liu, *et al.* 2015. Recurrent convolutional neural networks for text classification. In Proceedings of the AAAI conference on artificial intelligence (Vol. 29, No. 1, 2015).

[7] L. Huang, D. Ma, S. Li, *et al.* 2019. Text level graph neural network for text classification. arXiv preprint arXiv:1910.02356 (2019).

Table 4: Experimental Results of THUCNews Dataset.

Model	Accuracy(%)	Macro Average			Weighted Average		
		Precision(%)	Recall(%)	F1(%)	Precision(%)	Recall(%)	F1(%)
BERT-UNet	96.87	96.46	96.41	96.43	96.89	96.87	96.87
BERT-CNN [17]	96.72	96.37	96.11	96.24	96.73	96.72	96.72
BERT [13]	96.63	96.23	96.06	96.14	96.65	96.63	96.64
Transformer [9]	94.05	92.48	93.04	92.71	94.06	94.05	94.05
TextCNN [5]	94.34	93.95	92.73	93.32	94.36	94.34	94.34
TextRNN [18]	95.02	94.29	93.97	94.12	95.03	95.02	95.02
TextRCNN [6]	95.27	94.44	94.39	94.40	95.29	95.27	95.27

[8] L. Yao, C. Mao, Y. Luo. 2019. Graph convolutional networks for text classification. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, 2019), 7370–7377.
[9] Li, Q ; Xiao, Z ; Zhao, Y. Research on the Classification of New Energy Industry Policy Texts Based on BERT Model. Sustainability 2023, 15, 11186. https://doi.org/10.3390/su151411186
[10] Krosuri, L. R., & Rama, S. A. 2023. Novel heuristic-based hybrid ResNeXt with recurrent neural network to handle multi class classification of sentiment analysis. Machine Learning : Science and Technology, 4(1), 015033. doi:https://doi.org/10.1088/2632-2153/acc0d5
[11] Talaat, A. S. 2023. Sentiment analysis classification system using hybrid BERT models. Journal of Big Data, 10(1), 110. doi:https://doi.org/10.1186/s40537-023-00781-w
[12] Xia, X., Cao, H., Li, Y., & Song, Q. 2023. Tibetan text classification based on graph convolutional network. Journal of Physics: Conference Series, 2577(1), 012012. doi:https://doi.org/10.1088/1742-6596/2577/1/012012
[13] Xu, X., Chang, Y., An, J., & Du, Y. 2023. Chinese text classification by combining Chinese-BERTology-wwm and GCN. PeerJ Computer Science, doi:https://doi.org/10.7717/peerj-cs.1544
[14] Imad, J., Rachid, E. A., & Mohamed, B. 2023. Automated arabic news classification using the convolutional neural network. International Journal on Electrical Engineering and Informatics, 15(2), 277-290. doi:https://doi.org/10.15676/ijeei.2023.15.2.7
[15] Wang, H., & Li, X. 2022. Chinese news text classification based on convolutional neural network. Journal on Big Data, 4(1), 41-60. doi:https://doi.org/10.32604/jbd.2022.027717
[16] Rizinski, M., Jankov, A., Sankaradas, V., Pinsky, E., Mishkovski, I., & Trajanov, D. 2024. Comparative analysis of NLP-based models for company classification. Information, 15(2), 77. doi:https://doi.org/10.3390/info15020077
[17] Ge, B., He, C., Xu, H., Wu, J., & Tang, J. 2023. Chinese news text classification method via key feature enhancement. Applied Sciences, 13(9), 5399. doi:https://doi.org/10.3390/app13095399
[18] Hu W, Gu Z, Xie Y, *et al.* Chinese text classification based on neural networks and word2vec [C]//2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC). IEEE, 2019: 284-291.