IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

www.ijitce.com

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

USING LLM DOCUMENT CLASSIFICATION AT LOCAL DISK

[1]Dr. M. Kiruthiga Devi, [2]Mr. Praveen Kumar Sah, [3]Mr. Rakesh Ranjan Kumar

[1]Professor, [2,3]UG Student
[1,2,3]Department of Information Technology, Dr. M.G.R Educational And Research Institute, Chennai, India

Using LLM Document classification at Local disk, Automated document classification is the machine learning fundamental that refers to assigning automatic categories among scanned images and files of the documents. It reached the state-of-art stage but it needs to verify the performance and efficiency of the algorithm by comparing. The objective was to get the most efficient classification algorithms according to the usage of the fundamentals of LLM. This project focuses on the development of an automated document categorization system for a local disk, leveraging a Large Language Model (LLM) and zero-shot classification techniques. The primary goal is to classify and organize documents based on both their content and file extension, automatically moving them to their corresponding folders. Users can either download the documents or pass them to the application through a command line or API, after which the system identifies the document's extension, analyses its content using a pre-trained LLM, renames the file based on its contents, and then moves it to the appropriate folder.

Keywords: Document classification; machine learning algorithms; LLM; Zero Short Technique; analysis

1. INTRODUCTION

Text classification is one of the most important tasks in natural language processing. The common goal of text classification methods is to assign a predefined label to a given input text, although this name can refer to various specialized methods applied in different fields. Text classification mainly has the following task standards[1] :

(1) Emotional Analysis (SA): The task of understanding the emotional states and subjective information contained in a text, usually classified based on the emotions of excitement[2];

(2) Theme Label (TL): The task of identifying one or more themes (i.e. their themes) of a paragraph of text[3];

(3) News Classification (NC): Task of assigning categories to news segments, such as topics or user interests[4];

Early text classification models largely relied on manually constructing feature engineering, converting text into a computer understandable feature space, and using methods such as support vector machines, random forests, and linear regression to classify these feature data. Although traditional machine learning methods such as SVM and random forest have certain accuracy in implementing text classification tasks, the application process of these models still requires a large amount of manual participation .In addition, these models still have the following shortcomings: the accuracy of classification mainly depends on the complexity of feature engineering processing; Feature engineering processing requires domain knowledge; Traditional machine learning algorithms can only obtain shallow features; The accuracy of the model needs to be further improved.

In order to solve the above problems, researchers have introduced deep learning methods, and deep learning models have been widely applied in various fields of artificial intelligence. Deep learning methods mainly include: Convolutional Neural Networks (CNN)[5] . Recurrent Neural Networks (RNN)[6], Transformer[7], etc. These models can achieve good results in feature extraction. Compared with CNN, RNN can process sequential information and perform better in long sentence classification, but there are problems with gradient loss and long-term dependence. Zhou[9] et al.proposed a text classification algorithm that combines BiLSTM and two-dimensional maximum pool, using the LSTM model to extract contextual features and avoid the problems of RNN. However, LSTM model training brings two obvious drawbacks: gradient explosion and complex model structure causing long training time. The above machine learning and deep learning algorithms are currently widely used in various fields of text classification tasks. At the same time, the widespread success of the Bert[10] model also provides a foundation for text classification. It benefits from the attention mechanism, which effectively captures the global features of sentences or documents by identifying relevant words in the text. At the same time, inputting these features into the existing deep learning model effectively improves the prediction accuracy of the model. However, existing topic classification and news classification can usually only handle shorter texts, mainly because the input length of Bert and other pre trained models is limited (usually 512). At the same time, as the text length increases, LSTM and other models have gradient attenuation and other problems, their ability to extract features will further decrease, and the accuracy of the model will quickly decrease.

Many studies have made significant efforts to address BERT's limitation on the length of the maximum input sequence (i.e., 512). Sun[11] et al. used a truncation method that only preserves the first few markers in long text. Pappagari[12] et al. segmented long text into multiple fragments and inputted them into the model. However, all of the above methods may compromise the semantic integrity of sentences. Pappagari et al. used CNN and LSTM for BERT based Chinese long text classification. However, they simply stacked the models, and the accuracy of long text classification did not improve significantly.

In this study, we aim to achieve accurate classification of news, policies, reports, and other data in the oil and gas industry. These category labels include "energy related, energy unrelated, natural gas related, infrastructure construction, LNG related, investment related, Chinese, other countries', energy prices, policy documents, research reports, accidents", etc. This is a multi-label classification task, and each text may include one or more labels. We hope that each article can accurately obtain all the labels, and the length of these text data ranges greatly, ranging from dozens of words of news bulletins to thousands of words of policy documents. Therefore, this research process expects the model to have the ability to recognize local and global features.

To achieve this research task, this paper proposes a new text classification method for the oil and gas industry based on GLM-Bert-BiLSTM. This method first implements the extraction of long text news abstracts, filters data on short texts, inputs the results into the Bert model for encoding, and trains for classification. Abstract extraction based on large models overcomes the limitation of BERT algorithm on the maximum input sequence length. Similarly, it reduces training time and resources by setting the maximum input sequence length to 256. On this basis, this article uses BERT to complete the text level feature vector representation of news texts and learns its contextual representation (global features of sentences and documents). Then, we use the BiLSTM model to capture the detailed features and determine the final classification label.

2. MODELING

2.1 BERT

In 2015, DAI[13] et al. proposed the concept of pre training models, and in 2018, DEVLIN[14] et al. proposed the Bert pre training model based on the Transformer model. The Bert model uses massive data to achieve deep representation of input statements through unsupervised learning. This model solves long-distance context dependencies, and the word vectors generated by this model fuse sentence context information, enabling effective participation in the next research work.

2.2 BiLSTM

LSTM[8] (Long Short Term Memory) is an improved RNN, where each LSTM unit consists of an input gate, a forgetting gate, and an output gate. The LSTM storage unit can not only store short-term input information, but also store long-term input status. Among the three gates of the LSTM core, the input gate controls the size of the new storage content input, the forgetting gate determines the amount of storage that needs to be forgotten, and the output gate modulates the quantity of output memory content. The problem of gradient vanishing in traditional RNN models for feature extraction of long sequences can be overcome through three gates.

Due to the fact that LSTM models are all encoded from front to back, sentences can only grasp contextual information from front to back, and many words after sentences have strong correlation with the preceding words. Therefore, the BiLSTM model combines the forward LSTM model and the backward LSTM model to learn bidirectional contextual information.

2.3 GLM

ChatGLM[15] is a continuously developing family of large language models. It is a different architecture from BERT, GPT-3, and T5, and is an autoregressive pre training model that includes multiple objective functions. So far, the GLM-4[16] model has been pre trained on 10 trillion mainly Chinese and English labels. Evaluation shows that GLM-4 is close to or better than GPT-4 in general metrics such as MMLU, GSM 8K, MATH, BBH, GPQA, and HumanEval, and is close to GPT-4 Turbo in instruction following measured in IFEval. In short, the GLM model has superior text processing capabilities, so this study is based on the GLM-4 model.

2.4 INTRODUCTION TO EXPERIMENTAL DATA

In this experiment, we obtained 10000 news text data from the internet, covering various aspects related to oil and natural gas. These data were obtained from multiple news websites, which included both English and Chinese news. These news data were manually annotated by researchers using the Doccano tool,

In order to ensure that the data does not have a significant imbalance in the number of labels, we need to manually adjust the data distribution. In the end, 3500 news data were labeled as energy related, energy unrelated, natural gas, infrastructure construction, LNG related, investment, China, other countries, energy prices, policy documents, research reports, accidents, etc. When labeling at the same time, it should also be noted that each news article may describe multiple events, so a news article may have multiple labels. The distribution of labeled data labels is shown in Table 1.

Table 1: Number of classification labels.

Id	Label	Number
1	energy related	2631
2	energy unrelated	869
3	natural gas	912
4	infrastructure construction	766
5	LNG related	610
6	investment	230
7	China	1780
8	other countries	851
9	energy prices	543
10	policy documents	410
11	research reports	328
12	accidents	129
2.5	Experimental Plan	

The main technical route of this experiment is as follows: extract the original text data through the GLM model for summarization, input the extraction results into the Bert model for text vector encoding, and finally input the vector encoding into the BiLSTM model for text classification. The technical route is shown in Figure 1.

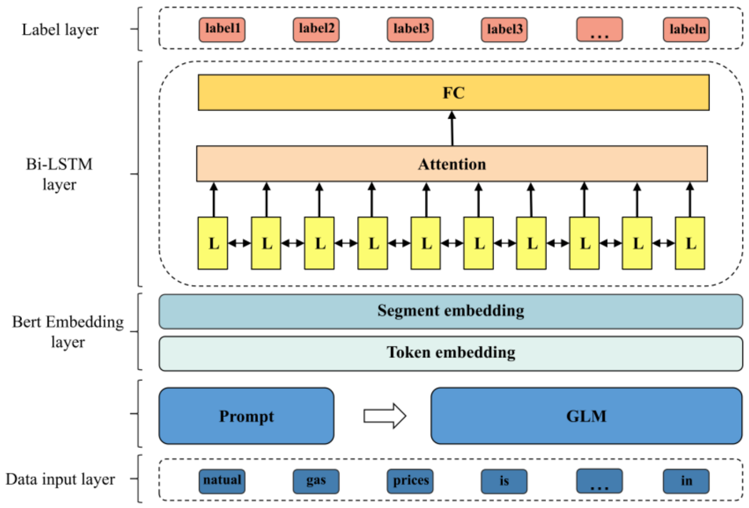

Figure 1: Model Technology Route.

In order to compare the accuracy of this classification model, we conducted two sets of comparative experiments. (1) Control group 1: text truncation was performed on long news texts, while retaining a certain length of text for training. (2) Control group 2: When the text is too long, in order to avoid missing the feature information of the text as much as possible, while ensuring that the paragraph length is acceptable to the model, the news of the long text is segmented into multiple paragraphs, and each paragraph is treated as a piece of data to participate in model training. Of course, this increases the workload of researchers, as they must re-examine whether the labels of each paragraph match the description of that paragraph.

3. RESULT ANALYSIS

3.1 Prompt

The focus of this study is to obtain accurate textual abstracts, which must be validated, reliable, and accurate in expressing the content of the entire news article. For large models such as GLM, abstract extraction is a very important task in the model fine-tuning process during the extraction process. Therefore, their abstract extraction ability is strong, but a reasonable Prompt is still needed to inspire the model to accurately complete our work. This study manually scored different prompt results and retained high scoring prompts that were more in line with the actual situation. Table 2 shows the different prompt contents designed in this study and whether they were selected

Table 2: Prompt Design.

Number	Prompt	Weather to adopt
1	Extract a summary of the following text	False
2	Extract a summary of the following text, keeping the word count at around 250 words and keeping the original meaning as much as possible	False
3	Extract a summary of the following text, with the word count floating up and down, not exceeding 250 words. Pay more attention to the location and events described in the text	True
4	Extract a summary of the following text, with a word count that can fluctuate up or down, not exceeding 250 words. Focus on domestic and foreign, policy documents, key events, and other aspects	True
5	Extract a summary of the following text, with the word count floating up and down, not exceeding 250 words, and try to maintain the original meaning as much as possible	True

3.2 Text Classification Results

When conducting accuracy analysis based on label categories, the comparison of results is shown in Table 3.The accuracy and F1 values of the classification based on the large model are 0.9623 and 95.87, respectively, with overall accuracy higher than that of control group 1 and control group 2. Therefore, the long text classification after abstract extraction based on the large model is effective in improving the accuracy of the original model. Comparing the accuracy and F1 values of control group 1 and control group 2 at the same time, the results showed that the accuracy and F1 values of control group 2 were higher than those of control group 1. Therefore, it can be concluded that directly truncating long texts lost the effective information of news texts, which affected the effectiveness of text classification. Classifying text into segments can help alleviate information loss caused by text truncation, but it can also make the model more one-sided in its focus on news. Therefore, we know that generating news abstracts based on LLM effectively solves the above problems and improves the accuracy of text classification.

Table 3: Results of comparative experiments (%).

Model	Accuracy	F1
GLM-Bert-BiLSTM	96.23	95.87
Bert-BiLSTM+Data Truncation	90.12	89.41
Bert-BiLSTM+Data Segmentation	91.92	91.75

In order to more accurately compare the differences in the predicted results of the three models in this experiment, we further focus on the indicator of whether a single news article can be recognized by the model for all labels. The calculation method for this indicator is: Accuracy=Number of news identified for all tags/Total number of news. The statistical results are shown in Figure 2.

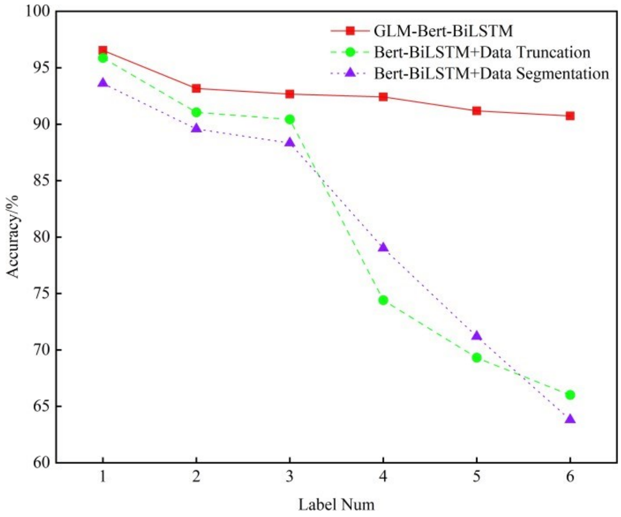

Figure 2: The proportion of all labels being recognized.

In this news classification, there may be multiple labels for a single news article. From the results shown in Figure 2, it can be seen that the proportion of news with a single label accurately recognized in the three experiments is high, and the effectiveness of the three models is similar. However, it can also be clearly found that the accuracy of GLM-Bert-BiLSTM is higher than that of the other two models. As the number of news labels increases, the proportion of all three models identifying the classified labels of news gradually decreases, which is consistent with the basic situation that the difficulty of multi label classification is higher than that of single label classification. The difference is that as the number of single news tags increases, the accuracy of the GLM-Bert-BiLSTM model decreases slowly, while the accuracy of the control group model decreases very quickly to around 50%. It can be seen that the accuracy of the algorithm in this study is mainly reflected in the accurate classification of multi label texts.

4. CONCLUSION

This article focuses on the classification of long news texts in the field of oil and gas. The GLM large model can extract abstracts from long news texts through Prompt, effectively reducing the length of the text while retaining the main content of the news. The Bert-BiLSTM model was used to train the abstract, and the results showed that compared with control group 1 and control group 2, the model effectively improved the accuracy of text classification. The accuracy of this model in multi label classification is much higher than that of the model that uses long text truncation for classification.

REFERENCES

[1] Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A Survey on Text Classification: From Shallow to Deep Learning. arXiv 2020, arXiv:2008.00364

[2] Gangamohan, Paidi, Sudarsana Reddy Kadiri, and B. Yegnanarayana. "Analysis of emotional speech—A review." Toward Robotic Socially Believable Behaving Systems-Volume I: Modeling Emotions (2016): 205-238.

[3] Jayady, Siti Hajar, and Hasmawati Antong. "Theme Identification using Machine Learning Techniques." Journal of Integrated and Advanced Engineering (JIAE) 1.2 (2021): 123-134.

[4] Minaee, Shervin, et al. "Deep learning--based text classification: a comprehensive review." ACM computing surveys (CSUR) 54.3 (2021): 1-40.

[5] Gu JX, Wang ZH, Kuen J, et al. Recent advances in convolutional neural networks. Pattern Recognition, 2018, 77: 354–377. [doi: 10.1016/j.patcog.2017.10.013]

[6] McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology, 1986, 18(1): 1–86. [doi: 10.1016/0010-0285(86)90015-0]

[7] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[8] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[9] Zhou P, Qi ZY, Zheng SC, et al. Text classification improved by integrating bidirectional LSTM with twodimensional max pooling. Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 3485–3495.

[10] Kenton J D M W C, Toutanova L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of naacL-HLT. 2019, 1: 2.

[11] Sun C, Qiu X, Xu Y, et al. How to fine-tune bert for text classification? [C]//Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18. Springer International Publishing, 2019: 194-206.

[12] Pappagari R, Zelasko P, Villalba J, et al. Hierarchical transformers for long document classification[C]//2019 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, 2019: 838-844.

[13] Dai A M, Le Q V. Semi-supervised sequence learning [J]. Advances in neural information processing systems, 2015, 28.

[14] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

[15] Zeng et al. GLM-130B: An Open Bilingual Pre-Trained Model. ICLR 2023.

[16] GLM, Team, et al. "ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools." arxiv preprint arxiv: 2406. 12793 (2024).