Deep learning based news text classification software design

Xiao Xu¹*, LiJuan Wang¹, RuFan Liu¹ and TianYu Xu¹
¹College of Internet of Things Engineering, Hohai University, Changzhou, Jiangsu Province, 213022, China
*Corresponding author’s e-mail: xx3361891@dingtalk.com

Abstract. New technologies such as artificial intelligence have developed at a rapid pace in recent years and are increasingly being used in the process of managing news in bulk. The development of deep learning has facilitated unprecedented progress in the field of computing and has opened our eyes to the possibility of using AI for news text classification. In this paper, based on the system requirements analysis, we describe the process of functional modules arising from the requirements analysis, design the internal details of functional modules, including algorithms and detailed principles, and finally obtain a prototype of news text classification software, which results in the pre-design expectations. The research in this paper makes the system development work more concrete, while providing software users, software developers, and analysts and testers with a unified and comprehensive understanding of the system's functional implementation.

1. Introduction
With the continuous development of Internet technology, the news industry has also ushered in a major impact and reform, and various types of news platforms have emerged. For today's society huge news data, when using the need to crawl a large amount of news data, the follow-up work will often have the need for data classification. For many search software stations, such as Baidu, Sogou, if you need to recommend news to users, which requires staff to carry out statistical classification. Nowadays, not only the statistical release time can meet the needs of readers, but also statistical news article categories, for the whole analysis to do a support, therefore, for the Internet search engine on the news classification is still very important. On this basis, we design software combined with neural network technology to extract the text information in the news, and then analyze the text information to accurately predict the news category from it. This saves the time of traditional manual judgment of news categories and also makes the news text system more flexible [2].

2. Module Design
First, we crawl from the web as well as download the data set required for training from the official website. Since the data obtained is in the form of html, we need to precisely match and extract the objects that conform to the laws through python regular expressions, get the classified news headlines and contents, and process the obtained data for subsequent analysis [3]. The four modules divided in the subsequent development phase of this system are described in detail as follows.

2.1. Text Analysis Module
The text analysis module addresses the fact that the word vector representation in the news text classification method does not retain the information of the word in the sentence and its polysemy well,
and enhances the semantic representation by fusing knowledge and linguistic semantic information. Using the knowledge-enhanced semantic representation ERNIE pre-trained model [4], the vector representation of the word is calculated according to the context, which can be adjusted according to the polysemy of the word while retaining the contextual information of the word. The semantic representation of the word is enhanced. It has good performance in Chinese text classification tasks.

2.2. Text Processing Module
The text processing module is a combination of a forward LSTM and a backward LSTM. It is used to model contextual information in natural language processing tasks. Longer distance dependencies can be better captured using LSTM models.

2.3. Result prediction module
According to the existing training model to process and predict the news words, the network prediction process will give a probability to each possible category of news, the category with the maximum probability is the prediction result, and then according to the accuracy calculation formula to obtain the prediction result of the batch operation [5]. The system monitors the metrics during the running of the code, determines the category of the news by the maximum probability of the prediction result for a single news item, calculates the accuracy of the validation set for each run, and ends the training when the validation set loss does not drop over a certain time. The F1-score mean of the training results is determined by calculating precision, recall, and accuracy [6].

3. Detailed Design and Implementation

3.1. Algorithm design

3.1.1. Text Analysis.
- BERT model: The BERT model is able to calculate the interrelationships between words and use the calculated relationships to adjust the weights and extract the important features in the text, using the structure of the self-attentive mechanism for pre-training, based on all layers fusing the left and right side contexts to pre-train deep bidirectional representations, which captures contextual information in a real sense than previous pre-training models and is able to learn the relationships between consecutive text fragments relationships, the model pre-training structure diagram is shown in Figure 1. One of the Transformer's network architect-ures is shown in Figure 2. Transformer is an encoder-decoder structure, formed by several encoders and decoders stacked on top of each other. The left part of the figure is the encoder, which consists of Multi-Head Attention and a full connection for transforming the input corpus into feature vectors [7]. The right part is the decoder, whose input is the output of the encoder and the predicted result, and consists of Masked Multi-Head Attention, Multi-Head Attention and a full connection to output the conditional probability of the final result.

Figure 1. BERT model pre-training structure diagram.
Figure 2. Transformer network architecture diagram.

- **ERNIE Model**: Based on BERT, an enhanced language representation ERNIE model implementation was trained using the corpus and KG. The BERT model is trained using two pre-training tasks: Masked LM and Next Sentence Prediction. 15% of the words are masked during the BERT Mask process, and then the model is allowed to predict the masked words. The new strategy of the ERNIE model allows the model to learn some knowledge better, called Knowledge Mask. Knowledge Mask, which mainly consists of Phrase Mask and Entity Mask, can be composed of multiple words [8]. By masking some phrases in a sentence and predicting the whole phrase, it allows ERNIE to better capture the relationships between phrases, entities.

Figure 3. ERNIE model pre-training structure diagram.
3.1.2. Text Processing.
- BiLSTM model: The long short-term memory neural network LSTM is the most popular recurrent neural network, which is not only more sensitive to short-term inputs, but also better able to preserve long-term states. LSTM differs from a single neural network layer in that it has four units and interacts in a very special way, mainly by three switches to control the input and output of the units. The neural network can automatically extract features based on the distribution of words in the text. The forward LSTM is combined with the backward LSTM to form a bidirectional LSTM, after the bidirectional LSTM outputs the prediction layer, the global optimal news category is found by the result prediction module using the labels already predicted by the context.
- BiLSTM-Attention model: BiLSTM-Attention is to add an Attention layer on top of BiLSTM, in which we use the output vector of the last time series as the feature vector, and then perform Softmax classification. In the experiments, the addition of Attention does improve the results. The model structure is shown in Figure 4

![Figure 4. BiLSTM+Attention model structure.](image)

3.1.3. Result Prediction.
The dictionary of model parameters is first deserialized using state_dict to create a mapping between each layer and its corresponding parameter, which is used to load the training-loaded model parameters. Then in order to keep the inferred results consistent, some specific layers of the model are different during the training and evaluated behaviors, so these specific layers need to be turned off. Finally, the news words are processed and predicted based on the existing training model. The network prediction process assigns a probability to each possible category of news, and the category with the maximum probability is the prediction result.

3.2. Data Interaction Solutions
- PyTkinder front-end: This system deploys a computer-side application. No installation steps are required, users can unzip the zip package on the computer and run this system through python files.
- Algorithm - Service Backend: The service backend calls the deep learning algorithm. Deep learning classifies each text according to the news text passed by the backend and then performs front-end update and batch display.

4. News access and functional testing
Before conducting the feature test, we first obtained enough data from the Internet for our training and testing use. We obtained nearly 200,000 pieces of data from major news platforms, and we also used the official news classification files provided as a reference, which was enough for us to use.
4.1. News Access
Crawler program: With the rapid development of the Web, the World Wide Web has become a carrier of a large amount of information, and it has become a great challenge to extract and use this information effectively. A web crawler is a program that automatically extracts web pages from the World Wide Web for search engines and is an important component of search engines. Focused crawler is a program that automatically downloads web pages, which selectively visits web pages on the World Wide Web with related links to obtain the required information according to the set crawl target. The specific process is as follows.

- Obtain the initial URL: the URL address is used for the initial crawl of the web page developed by itself.
- When crawling the web page corresponding to the URL address, obtain the new URL address
- Putting the new URL address into the URL queue
- Read the new URL from the URL queue, then crawl the web page based on the new URL, and get the new URL address from the new web page, and repeat the above crawling process.
- Set the stopping condition. If the stopping condition is not set, the crawler will crawl until it cannot get the new URL address. If the stop condition is set, the crawler will stop crawling when the male stop condition is set.

Parsing Sogou news package: The original data processing of Sogou news mainly includes transcoding, extracting CONTENT and URL. The text used is the Sohu news data from Sogou Lab news resources [9]. According to the format of the data content in the text, regular expressions are used to extract the news category, title, and content in the text.

News collectors: We use numerous data collectors, they come with crawler function and data de-noise function, the best one is Houyi collector [10]. Its working process can be multiple processes in parallel, which has the advantages of simplicity and efficiency compared with other data collectors. We only need to select the URL of the news to be collected, the software can help us automatically extract the text, we can select the text data we want according to our needs, click to confirm and then you can delete the duplicate data or advertisements in it, which is convenient for users.

4.2. Functional Testing
Purpose of functional testing:
- Determine the basic operating condition and basic functionality of the system.
- Evaluate the performance of commonly used text classification models.
- List the recommended high-level testing requirements that will benefit the system at a later stage of improvement.
- Recommend test strategies that can be used and describe these strategies.
- Determine the required resources and estimate the testing effort.

| Model        | Accuracy | Description                                      |
|--------------|----------|--------------------------------------------------|
| TextCNN      | 91.22%   | Kin2014 Classic CNN Text Classification          |
| TextRNN      | 91.12%   | BiLSTM                                          |
| TextRNN_Att  | 90.90%   | BiLSTM+Attention                                |
| TextRCNN     | 91.54%   | BiLSTM+Pooling                                  |
| FastText     | 92.23%   | bow+bigram+trigram, the effect is surprisingly good |
| DPCNN        | 91.25%   | Deep Pyramid CNN                                |
| Transformer  | 89.91%   | Poorer effect                                   |
| Bert         | 94.12%   | Bert+fc                                         |
| ERNIE        | 94.23%   | Slightly better than Bert                       |
| ERNIE_Att    | 94.61%   | ERNIE + BiLSTM+Attention                        |
The accuracy of the prediction results is shown in the table above, and we can see that the prediction effect of Transformer model is the worst, and the effect of Bert and ERNIE model is the best.

5. Conclusion
Through the study of this paper, all the functions of the system have been realized and can be used normally, and the scalability and usability of the system are relatively good. In the test of the basic performance and functions of the system, the program can carry out human-computer interaction normally, and each basic function is complete, which can meet the basic requirements of the application. Among the commonly used Chinese text classification models, the Bert model and ENIRE model have greater advantages, but the ENIRE_Att model performs better, so the ENIRE_Att model is used for news text classification in this paper. Our system combines the current cutting-edge natural language processing technology to achieve automatic recognition of news text in order to improve the inefficiency faced by the current news text system, which plays an important role in the intelligent development of news text classification system, and can even be said to be imperative. However, there are still some problems in this system, such as the interface is still not beautiful and simple and generous enough, and the batch recognition speed of news is about 5 items/second, which still takes a relatively long time when predicting a large number of news, so the solution for this field needs to be further improved in future research.

References
[1] JingSu, Qingyun Dai, FrankGuerin, MianZhoud. (2021) BERT-hLSTMs: BERT and hierarchical LSTMs for visual storytelling. Elsevier Ltd.
[2] Qi Jiaqi, Chi Chengying, Zhan Xegang. (2021) ERNIE-CNN text classification model. Journal of Liaoning University of Science and Technology.
[3] Yu Wang, Yining Sun, Zuchang Ma, Lisheng Gao, and Yang Xu. (2020) An ERNIE-Based Joint Model for Chinese Named Entity Recognition. Applied Sciences.
[4] Liu Gaojun, Yuan Qiuxia, Duan Jianyong, Kou, Jie, Wang Hao. (2020) Chinese Question Classification Based on ERNIE and Feature Fusion (Conference Paper). Lecture Notes in Computer Science.
[5] Yan Chunman, Wang Cheng. (2021) Convolutional neural network model development and applications. School of Physics and Electronic Engineering, Northwest Normal University.
[6] Paolo Galeone (Italy), Yan Longchuan, Bai Dongxia, Guo Yonghe, Li Yan. (2020) Tensor-Flow 2.0 neural networks in practice. North Machinery Industry Press.
[7] Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, (2019) Haoyu Zhang, Pengjun Xie, Gongshen Liu Hierarchy-Aware Global Model for Hierarchical Text Classification.
[8] Zhang Jinjin. (2019) Attention Mechanism in Deep Learning.
[9] Adam Paszke, Sam Gross, Francisco Massa. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library.
[10] Ji Ma, Kuzman Ganchev, David Weiss. (2018) State-of-the-art Chinese Word Segmentation with Bi-LSTMs.