The Design and Development of Electronic Text Natural Language Processing System Based on Deep Learning

Dan Liu[a], Jianmin Hao[b] and Jiayi Li[c]
School of big data and artificial intelligence, Dalian University of Finance and Economics, Dalian, Liaoning, China

Keywords: Natural Language Processing (NLP), Deep Learning Model, Fasttext, Hadoop, Electronic Text Classification.

Abstract: In this paper, the author takes the upstream task text classification of natural language processing as the research object, and completes the collection, cleaning and distributed storage processing of electronic texts with the help of HDHF and HBase components under Hadoop cluster. The MapReduce component is used to call the deep learning algorithm FastText to complete the supervised text classification operation, and finally Echarts technology is selected to show the classification results in the form of charts. Besides, the running environment of the whole text classification process is Java environment, and the server side will be built according to the Spring framework, so as to ensure that all API data interfaces can be encapsulated and called reasonably, thus forming the electronic text natural language processing system on the Web side. The system will quickly realize the capture, cleaning, classification and visualization of electronic text information, reduce the problem of information overload in the downstream task execution of natural language, and improve the information retrieval efficiency.

1 INTRODUCTION

With the rapid development and wide application of computer and network information technology, the information resources on the Web are becoming diversified and quantified. The Web service with electronic text as the main form appears in different types of cyberspace, and becomes the most important information acquisition channel in people's daily life. According to statistics, as of June, 2021, there are 4.22 million websites in China, and the text content accounts for more than 87% of the web content. The development level of the Internet industry is still at a high level, and the penetration rate of the industry is gradually increasing. At the same time, as users, our demand for network functions is increasing, and the single passive information service is gradually changing to the active search, analysis and interpretation service. Therefore, how to effectively and quickly acquire the required and available knowledge from the vast network information resources has become the main research topic in recent years, and a new technology: electronic text mining has been formed. (Li, 2015)

The main contents of electronic text mining include text information clustering, classification, sentiment analysis and interpretation, and visual expression of final results. The electronic text classification task is the main application direction of the application, and its realization process involves many steps such as text information collection, preprocessing, feature extraction, classification model construction, and classification result display. As a data object, Web electronic text has many characteristics, such as large volume, easy access, unstructured, dynamic and difficult to mark, and its complexity far exceeds that of general static text documents. The traditional electronic text classification is mostly done by combining text feature analysis with shallow machine learning, but there are some shortcomings in accuracy and convenience in the process of classification. In view of this, this paper holds that the essence of electronic text mining is a research direction of Natural

[a] https://orcid.org/0000-0003-3144-9401
[b] https://orcid.org/0000-0002-1283-1245
[c] https://orcid.org/0000-0003-0360-1752

Liu, D., Hao, J. and Li, J.
The Design and Development of Electronic Text Natural Language Processing System Based on Deep Learning.
DOI: 10.5220/0011909000003613
In *Proceedings of the 2nd International Conference on New Media Development and Modernized Education (NMDME 2022)*, pages 197-202
ISBN: 978-989-758-630-9

Language Processing and a sub-field of artificial intelligence. The deep learning model can be used to construct feature text vectors to accurately express the word meaning and semantic information in electronic texts, so as to effectively improve the classification accuracy of electronic texts. (Che, 2019) According to the actual application requirements, Hadoop cluster will be used to capture and distribute the electronic text data on the Internet, MapReduce will be used to call FastText, and Echarts technology will be used to present the classification results, so as to design and implement an electronic text classification system integrating data collection, preprocessing, data classification and visual display. The test and actual simulation show that the system can improve the efficiency of electronic text classification with excellent performance and convenient operation, and is suitable for various scenarios of large-scale electronic text classification.

2 OVERVIEW OF KEY TECHNOLOGIES

2.1 Natural Language Processing

The Natural Language Processing (NLP) is an important direction in the fields of computer science and artificial intelligence. As an interdisciplinary subject, the research content involves linguistics, computer science, mathematics, statistics and other fields, aiming at realizing human-computer interaction and communication with natural language as the medium. It means that all kinds of software applications are used to process the information of the form, sound, meaning and so on of natural language, through input, recognition, analysis, understanding, generation and output, so that computers can "understand and understand" human language, expand the application field of computers, and replace humans to complete some work. (He, 2020)

With the rapid development of artificial intelligence, the application scenarios and fields of natural language processing are constantly enriched. The common fields include text information retrieval, machine intelligent translation, text classification mining, information extraction and filtering, speech recognition and generation, automatic question answering and dialogue, etc. Among them, text classification is a typical problem in the field of natural language processing, and most of the tasks of natural language processing can be regarded as a classification task, which is in the upstream stage in the field of natural language processing research. Text classification can not only provide necessary preconditions for research in other fields, but also directly affect the practical application effect of natural language processing downstream.

In the initial stage of text classification, expert rules are mostly used to complete the classification operation, which requires a lot of human work to reason and judge, and human factors are uncontrollable, so it does not have good expansibility. However, with the rise of machine learning, text classification has entered the statistical era, relying on the method of text feature analysis combined with shallow-level machine learning. Although the work efficiency, cost control and application expansibility have been significantly improved, it still can't keep up with people's demand for fineness and accuracy. Until the emergence of deep learning technology, coupled with the substantial improvement of computer hardware capabilities, it has greatly promoted the development of natural language processing and further expanded the application scope of text classification.

2.2 Deep Learning Model

The deep learning technology based on neural network architecture is a branch of machine learning. Its essence is to make computers perform specific tasks by imitating the way humans acquire and apply knowledge. (Han, 2021) At present, deep learning model has gradually become the mainstream technology for text classification. The method of constructing feature text vectors based on deep learning analysis model can accurately express the word meaning and semantic information in the text, and automatically acquire the feature expression ability by virtue of its excellent network structure, thus avoiding the tedious work of manually designing rules and features, and realizing end-to-end problem solving. In this paper, according to the characteristics of electronic text, FastText deep learning model is selected to complete text classification. The application advantage of FAST is that it is suitable for a large amount of data samples and supports multilingual expression, and the overall training speed is far FastText than that of the same type model. The core principles include model architecture, hierarchical SoftMax and N-gram features. Among them, the FastText model architecture can predict and classify the whole text as a feature, as shown in Figure 1, which is its model architecture.

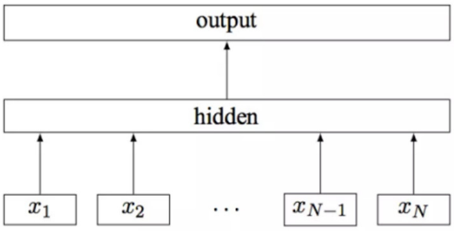


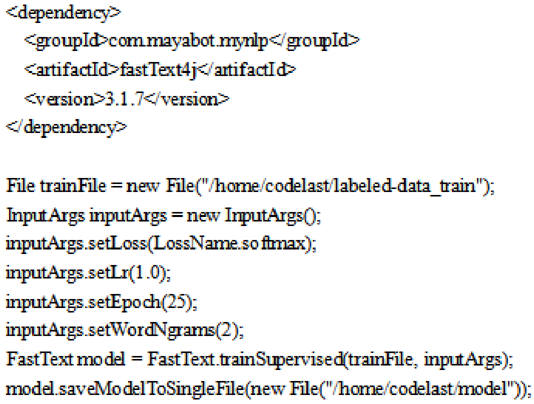
Figure 1: FastText model architecture

2.3 Development Process

According to the above application requirements, complete the configuration and deployment of the development environment of electronic text natural language processing system. The development content of the system is divided into two parts. One is the collection, pre-processing and distributed storage of electronic data under Hadoop cluster architecture, and the MapReduce is used to call FastText model to complete the electronic text classification operation. The other is to use Spring framework to complete the development of the server side of the system under the Java development environment, realize the encapsulation of each functional module of the system, and visually display the final classification results.

Firstly, Hadoop cluster architecture needs a lot of software and hardware as support. The operating system is Linux, CentOS 6.7 and JDK 1.8. According to the application requirements of the system, Hadoop cluster will be set to three nodes, named Master, Slave1 and Slave2 respectively. Hadoop version is 2.7.7, and components such as Yarn, HDFS, Zookeeper, HBase, etc. are also deployed in each node. Under Hadoop architecture, the system uses Scrapy crawler tool to grab the web page content to realize text information collection. After the collection, the tasks of MapReduce will be segmented, and the semi-structured or unstructured web page text information will be reasonably preprocessed and converted into structured data, which will be stored in HDFS or HBase.

For the FastText model, the Python package supported by it can't be directly installed under Hadoop architecture, so it is necessary to use "FastText4j" in Java version to load MapReduce-job in Java environment. The introduction and training of the FastText4j model is shown in Figure 2. During the implementation process, it is necessary to introduce corresponding dependencies into Maven project and complete model training. The key code is shown in Figure 1.



```
<dependency>
  <groupId>com.mayabot.mynlp</groupId>
  <artifactId>fastText4j</artifactId>
  <version>3.1.7</version>
</dependency>

File trainFile = new File("/home/codelast/labeled-data_train");
InputArgs inputArgs = new InputArgs();
inputArgs.setLoss(LossName.softmax);
inputArgs.setLr(1.0);
inputArgs.setEpoch(25);
inputArgs.setWordNgrams(2);
FastText model = FastText.trainSupervised(trainFile, inputArgs);
model.saveModelToSingleFile(new File("/home/codelast/model"));
```

Figure 2: Introduction and training of the FastText4j model

Secondly, after the FastText model completes the text classification, the system will support the users to realize the visual display of the classification results through the operation on the Web side. The Web server operating system is Windows 10.0, the JDK version of the development kit is 1.8, the Web server is Apache Tomcat 9.0, the Java integrated development tool is IntelliJ IDEA 2019, the project management tool Maven 3.5.0, and the database is MySql 8.0. Choose to create a maven-archpetype-webapp in IntelliJ IDEA, and introduce several jar packages such as J2EE, Mysql and Spring Framework. Then select Add Spring Framework to complete the server-side construction. For the visual display of the final result, it will be realized by Echarts technology, that is, Echarts.js will be introduced into IntelliJ IDEA, and the module loader configuration and chart path setting will be completed. With the introduction of the above key technical theories, the overall environment of the system development, the configuration of related software and tools are determined, and the technical feasibility of the overall project of the electronic text natural language processing system is also clarified.

3 REQUIREMENT ANALYSIS

3.1 System Function Analysis

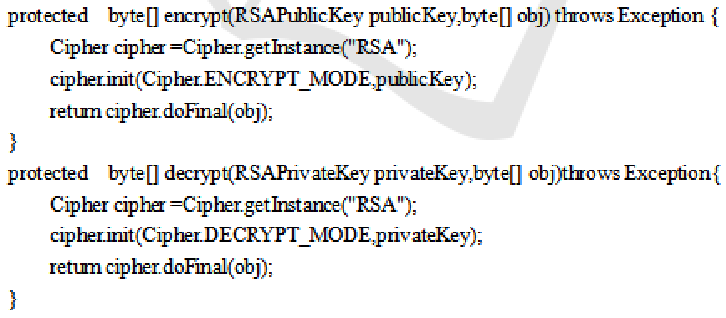
Aiming at the shortcomings of traditional machine learning classification algorithm in electronic text classification, this system will complete the collection, cleaning, storage, classification and processing of electronic text by taking advantage of the advantages of FastText, a deep learning classification algorithm model, and combining HDHF, HBase and MapReduce components under Hadoop framework. And the whole process is integrated and encapsulated in Java environment to form a standard Web application, which is convenient for users to quickly, conveniently and conveniently realize the classification task of a large number of electronic texts. The system not only effectively improves the efficiency of electronic text classification, but also further expands the application scenarios of electronic text classification technology. The system will support users to complete account registration by submitting information, and complete the login and use of the system with unique identification information. According to the actual needs of users and the process of text mining, each functional module is designed and implemented.

3.2 Overall Design

The electronic natural language processing system adopts B/S architecture and is developed based on MVC design pattern. According to the layered design idea, the system will be divided into standard five-layer architecture by using the Spring framework under J2EE specification, which are view layer, control layer, business logic layer, data access layer and data storage layer. The view layer, which is composed of several Web pages, is the interface of human-computer interaction, and is mainly realized by JSP technology. The core function of the control layer is to accept the user's request, call the business processing object according to the request, and then return the corresponding processing result to the view layer for display after the business processing object is executed. The business logic layer is the key to realize the functions of the whole system, and is responsible for the processing of the specific business logic of the system. The design of data access layer aims to realize the interaction between business processing logic and data storage layer, so that business logic layer can obtain data services better. (Zhou, 2016)

4 FUNCTIONAL IMPLEMENTATION

When the system users log in for the first time, they need to complete identity registration and real-name registration system authentication, and use the unique account information to log in and use the system. The system supports hashing algorithm to encrypt user password, which is used as authentication method to complete user login authentication. The key code of implementing RSA encryption algorithm in Java language is shown in Figure 3. After successful login, the system will automatically jump to the home page of the system, and select the corresponding function module in the navigation bar of the system function.



```
protected   byte[] encrypt(RSAPublicKey publicKey,byte[] obj) throws Exception {
      Cipher cipher =Cipher.getInstance("RSA");
      cipher.init(Cipher.ENCRYPT_MODE,publicKey);
      return cipher.doFinal(obj);
}
protected   byte[] decrypt(RSAPrivateKey privateKey,byte[] obj)throws Exception{
      Cipher cipher =Cipher.getInstance("RSA");
      cipher.init(Cipher.DECRYPT_MODE,privateKey);
      return cipher.doFinal(obj);
}
```

Figure 3: RSA algorithm implementation key code

4.1 Text data collection

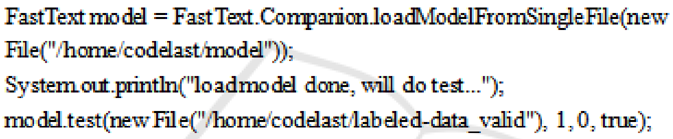
After entering the system, users will gradually complete all the operations of electronic text classification according to the operation guide. In the process of text data collection, the system supports Scrapy crawler tool to achieve hierarchical crawling of electronic texts on the Web. In the address box of the user interface, enter the website address, click to get electronic text data, and the system can automatically crawl the relevant information, and after screening, write it into HDFS or HBase under Hadoop framework to realize the storage of electronic text information.

4.2 Pretreatment

In the preprocessing stage, the system will perform text segmentation, stop words removal, text feature extraction and other operations on the acquired text data. The words in a sentence are aggregated by N-gram after word segmentation, and the corresponding token is established for each phrase, and the combination between phrases also has its own token. While in text feature extraction, FastText uses Word2vec combined with N-gram model to train Word embedding to produce feature word vectors. (Ma, 2019)

4.3 Text Classification

Under this function module, the system takes the preprocessed text sequence as the input content of the classification algorithm, and carries out classified storage operation. The system automatically loads the trained model through MapReduce module, and puts the model into distributed cache for distribution. The key code of model loading and test effect is shown in Figure 4. The classification results will output the Label label, and the storage of the classification results will be completed at the same time. After the simulation test, the FastText classification algorithm model achieves an accuracy rate of 80.447% when dealing with the whole content or single sentence of electronic text. Compared with the traditional machine learning algorithms such as Random Forest, Naive Bayes and Support Vector Machine, the overall effect is in line with the system research expectation, and the specific results are shown in Table 1.



```
FastText model = FastText.Companion.loadModelFromSingleFile(new
File("/home/codelast/model"));
System.out.println("load model done, will do test...");
model.test(new File("/home/codelast/labeled-data_valid"), 1, 0, true);
```

Figure 4: The MapReduce Job loads the FastText model and tests the effect

Table 1: Performance comparison of FastText and other algorithms

	Random Forest	Naive Bayes	Support Vector Machine	FastText
Precision (%)	71.821	66.181	77.411	80.447
Training time	6.553	1.081	44.631	110.633
Forecast time	0.258	0.399	0.281	5.094

4.4 Visual Presentation

We use the visual drawing tool of Echarts to complete the rendering of electronic text classification results. According to different feature dimensions, the classification results are sorted according to the quantity, and the sorting results are displayed in a suitable chart. There are many kinds of charts of Echarts, including line charts, pie charts, radar charts, etc. Besides, in the front-end page design, Ajax will also be used to implement the data refresh function. Ajax asynchronous refresh enables the data to be loaded dynamically, so as to present the dynamic refresh effect of data charts.

5 CONCLUSIONS

In this paper, electronic text classification is taken as the research object, aiming at many shortcomings of the current traditional electronic text classification in the practical application process, and with the help of the functional advantages of deep learning model, the electronic text natural language processing system is built under the Java language environment with FastText model as the core and Hadoop cluster processing framework as the auxiliary. It can quickly capture, clean, classify and visualize electronic text information, reduce the problem of information overload caused by the downstream task execution of natural language, and improve the information retrieval efficiency.

REFERENCES

Che Qingxun (2019). The Design and Implementation of Text Processing System Based on Deep Learning. Huazhong University of Science and Technology.05.

Han Chong, Wang Junli and so on (2021). A Survey of Deep Learning Models Based on Neuroevolution. Acta Electronica Sinica.02:373-378.

He Kai (2020). The Research and Application of Text Classification Based on Natural Language Processing. Nanjing University of Posts and Telecommunications.12.

Li Xiaodi (2015). The Research and Application of Web Mining Technology. Beijing Jiaotong University06.

Ma Yu. The Research of News Text Classification Based on fastText and Its Application in Agricultural News. Jilin University.2019.05.

Zhou Yanling (2016). Exploration and Research of WEB Application Development Based on Spring MVC Framework. Science Mosaic.06:25-28.