mathematics

MDPI

Article

Design of an Automatic Classification System for Educational Reform Documents Based on Naive Bayes Algorithm

Peng Zhang [1,2,3], Zifan Ma [1,2], Zeyuan Ren [1,2], Hongxiang Wang [1,2], Chuankai Zhang [1,2], Qing Wan [4,*] and Dongxue Sun [5]

[1] School of Mechanical Engineering, Hebei University of Technology, Tianjin 300401, China; zhangpeng@hebut.edu.cn (P.Z.); m202121203010@163.com (Z.M.); rzyhebut@126.com (Z.R.); 15822427129@163.com (H.W.); zuhom595@126.com (C.Z.)
[2] National Engineering Research Center for Technological Innovation Method and Tool, Hebei University of Technology, Tianjin 300401, China
[3] Yueqing Institute of Technological Innovation, Yueqing 325600, China
[4] Undergraduate School, Hebei University of Technology, Tianjin 300401, China
[5] Party Committee Organization Department, Hebei University of Technology, Tianjin 300401, China; 2016031@hebut.edu.cn
* Correspondence: 202221203005@stu.hebut.edu.cn

Abstract: With the continuous deepening of educational reform, a large number of educational policies, programs, and research reports have emerged, bringing a heavy burden of information processing and management to educators. Traditional manual classification and archiving methods are inefficient and susceptible to subjective factors. Therefore, an automated method is needed to quickly and accurately classify and archive documents into their respective categories. Based on this, this paper proposes a design of an automatic document classification system for educational reform based on the Naive Bayes algorithm to address the challenges of document management in the education field. Firstly, the relevant literature and document data in the field of educational reform are collected and organized to establish an annotated dataset for model detection. Secondly, the raw data are preprocessed by cleaning and transforming the original text data to make them more suitable for input into machine learning algorithms. Thirdly, various algorithms are trained and selected to determine the best algorithm for classifying educational reform documents. Finally, based on the determined algorithm, a corresponding classification software is designed to automatically classify and archive educational reform documents for analysis. Through experimental evaluation and result analysis, this research demonstrates the effectiveness and accuracy of the education reform document automatic classification system based on the Naive Bayes algorithm. This method can efficiently classify a large number of documents into their respective categories quickly and accurately, thereby improving the efficiency of educators and their information management capabilities. In the future, further exploration of feature extraction methods and machine learning algorithms can be conducted to optimize the classification performance and apply this method to practical management and decision-making in the education field.

Keywords: educational reform; manual classification; Naive Bayes algorithm; feature extraction

MSC: 68T09; 68T50

Citation: Zhang, P.; Ma, Z.; Ren, Z.; Wang, H.; Zhang, C.; Wan, Q.; Sun, D. Design of an Automatic Classification System for Educational Reform Documents Based on Naive Bayes Algorithm. *Mathematics* **2024**, *12*, 1127. https://doi.org/10.3390/math12081127

Academic Editor: Shih-Wei Lin

Received: 29 February 2024
Revised: 26 March 2024
Accepted: 8 April 2024
Published: 9 April 2024

1. Introduction

With the rapid development of artificial intelligence and machine learning, the education sector has also ushered in opportunities for transformation. Traditional education systems face many challenges, one of which is the management and utilization of educational resources. Educational institutions and educators often need to deal with a large number of documents, including course materials, student assignments, teaching materials, and so

on. The classification and archiving of these documents are crucial for improving educational efficiency, supporting personalized learning, and facilitating teaching assessment.

Text classification was proposed by American scholar Professor Luhn in 1957. In 1961, Maron [1] published the first paper on automatic text categorization. Text classification is to use the computer to automatically classify the text set according to a certain classification system or standard. Text classification technology has a wide range of application prospects. It is the technical support in the fields of information filtering, information retrieval, and document classification [2]. Text classification is generally divided into two categories: classification method based on knowledge engineering, and classification method based on machine learning. The classification method based on knowledge engineering refers to the classification based on expert experience and manual extraction rules. The classification method based on machine learning refers to the classification by computer self-learning and extracting rules. In view of the shortcomings of traditional text classification methods, many scholars have carried out research on text classification methods and revised and improved them. Based on the superiority of neural network algorithms in the field of natural language processing, Alina utilized a classification tool based on deep neural networks to automatically extract diagnostic and disease information at the time of surgery from electronic pathology reports monitored by the National Cancer Institute (NCI) and the Surveillance, Epidemiology, and End Results (SEER) population cancer registries, benefiting the cancer registries [3]. Liu trained a recursive neural network model to classify microblog posts related to urban flooding, establishing an online monitoring system for urban inundation triggered by flooding disasters [4]. Yang utilized artificial intelligence technology to improve the speed of encoding classroom dialogues, achieving automated classroom dialogue classification and instant feedback. They used neural network analysis models to evaluate and classify question level, answer level, and feedback level, constructing a comprehensive, rapid, and accurate method for evaluating classroom dialogue [5]. Vishaal established a recursive interleaved multi-task learning network that can be used for any general multi-label classification task related to the field of education [6]. Compared with the traditional classification methods, which mainly use supervised methods, relying on existing natural language processing tools can easily lead to error accumulation problems in the processing process. Chen proposed a Chinese text classification method based on appearance semantics and ASLA. Baidu Encyclopedia is used to extract the apparent semantics of Chinese text, and then PLSA is used to mine the potential semantics, and the correlation between the apparent semantics and the potential semantics and the category of the document is calculated. This method can deal with the classification of irregular texts such as Chinese network short texts [7]. In order to express text directly, Li proposed a short text classification model based on dense network, which uses one-hot coding, expands text feature selection by merging and random selection, and solves the problems of feature sparseness, dimensional text data, and feature representation [8]. Wang used the method of improving TF-IDF to modify the weight of word vector to optimize the text classification algorithm. Finally, the convolutional neural network was used to construct the classifier, which improved the accuracy of text classification. However, the high-order features were not properly disposed, resulting in the time complexity of learning much higher than the traditional machine learning method, which needs to be further improved [9]. With the research of deep learning for text classification, many researchers have found that the classification effect cannot be further improved by using only a single deep learning model, so scholars have proposed a method of mixing these deep learning models. Du proposed an emotion classification model based on convolution attention. This structure combines RNN with convolution-based attention model, and further superimposes the attention model to construct a hierarchical attention model for emotion analysis [10]. Peng proposed a hybrid model combining a deep ultra-deep convolutional neural network with a long short-term memory network, which is an application of deep CNN. The hybrid model shows better text effect than shallow CNN [11].

Although the existing common classification methods can meet the requirements of text classification in some aspects, there are still the following shortcomings for specific domain files (educational reform documents):

1. The classification standards are not uniform, and it is impossible to form a widely accepted and recognized classification.
2. The training sample data retrieval is difficult, the lack of standardized retrieval methods.
3. It is difficult to train pre-trained language models with domain background knowledge.

At the same time, many scholars have made contributions to the field of text classification. However, for educational-reform-related documents, manual classification by professional personnel from schools or educational institutions remains the prevailing method in today's educational environment. With a large number of educational policies, program proposals, and research reports constantly emerging, manually handling and organizing these documents is a tedious and time-consuming task. In order to solve the above problems, this paper proposes a text classification system for educational reform documents based on the Naive Bayes algorithm. The design involves several key steps: firstly, the classification standard is determined according to the "Guidelines for Application of Research and Practice Projects on Undergraduate Education and Teaching Reform at Hebei University of Technology in 2021". Next, develop a retrieval strategy to collect the relevant literature and document data in the field of education reform, and establish a well-annotated dataset for model testing. Again, preprocessing the raw data, cleaning and transforming the original text data to make them more suitable for input into machine learning algorithms. Further, training and selecting multiple algorithms to determine the most effective algorithm for classifying educational reform documents. Finally, based on the chosen algorithm, designing corresponding classification software to automatically classify and archive the analyzed educational reform documents. This system utilizes machine learning algorithms and natural language-processing techniques to automatically identify, classify, and archive different types of educational documents, greatly enhancing the efficiency and accuracy of educational management.

The development and application of this automatic classification system for educational reform documents based on the Naive Bayes algorithm will enable educational institutions and educators to better manage and utilize educational resources, thereby improving teaching effectiveness and learning outcomes. The advancement and application of this technology will bring about more intelligent and efficient management and teaching methods in the field of education, leading to overall enhancements in the quality of education delivery and outcomes.

2. Methods

In this section, the method designed for automatic classification of educational reform documents is introduced. To achieve the intended goals, it is necessary to follow several steps as illustrated in Figure 1. The key points of this method include:

1. Determine the categories for classification based on the "Guidelines for Application of Research and Practice Projects on Undergraduate Education and Teaching Reform at Hebei University of Technology in 2021".
2. Collect data from sources such as CNKI (China National Knowledge Infrastructure) and the school database for training and testing data.
3. Preprocess the data by segmenting the obtained Chinese text, and construct training and testing sets.
4. Train the model by using various algorithms to train the training set and calculate the probability analysis of each feature under each category.
5. Analyze the training results and error rates, select the optimal algorithm for further design.
6. Design corresponding software to achieve automatic classification.

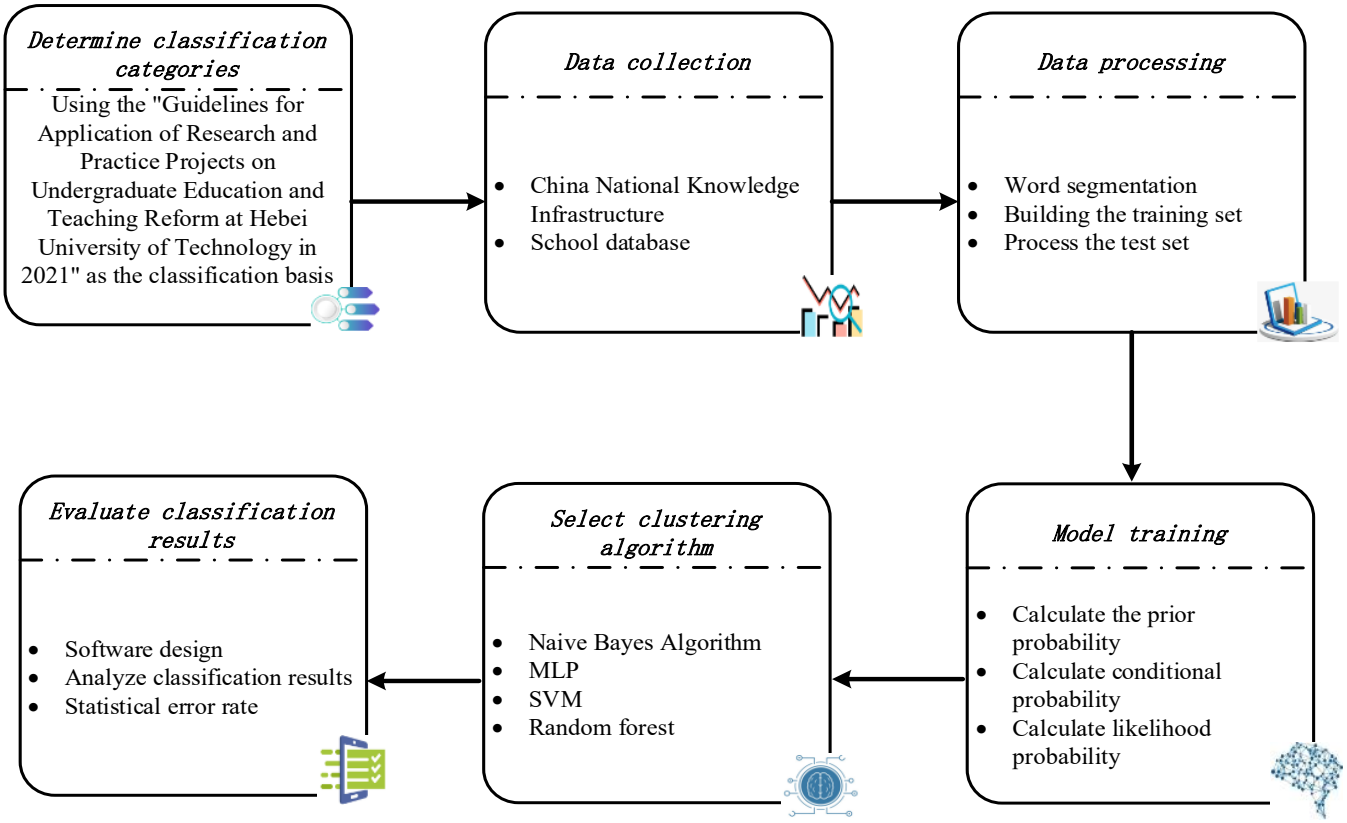

Figure 1. Representation of the methodology used to conduct this study.

The following parts of this paper are organized as follows: Section 2.1 identifies the categories for classifying documents related to educational reform; Section 2.2 determines the sources of the dataset and extracts the dataset; Section 2.3 selects various clustering algorithms, finalizes the algorithm to be used, pre-processes the initial data, and trains the selected algorithm; and Section 2.4 designs corresponding software to evaluate the model.

2.1. Categories of Documents Related to Educational Reform

To achieve precise classification of documents related to educational reform, it is essential to determine the categories of the documents. The "Guidelines for Application for Research and Practice Projects on Undergraduate Education Teaching Reform in 2021" from Hebei University of Technology provides detailed descriptions of the categories for educational teaching reform documents. Therefore, based on this guideline as the classification criteria, as shown in Figure 2, educational reform documents are divided into six categories: Special Topic on Theoretical and Practical Research of Emerging Engineering Education; Special Topic on Theoretical and Practical Research of New Liberal Arts; Research Topic of Collaborative Education Development; Research Topic of Curriculum Ideological and Political Construction; Labor Education, Aesthetic Education, Sports Research Topics; and Research on Teaching and Practice of Integrated Education. Keywords under each special topic are then counted for subsequent searching and extraction of the dataset.

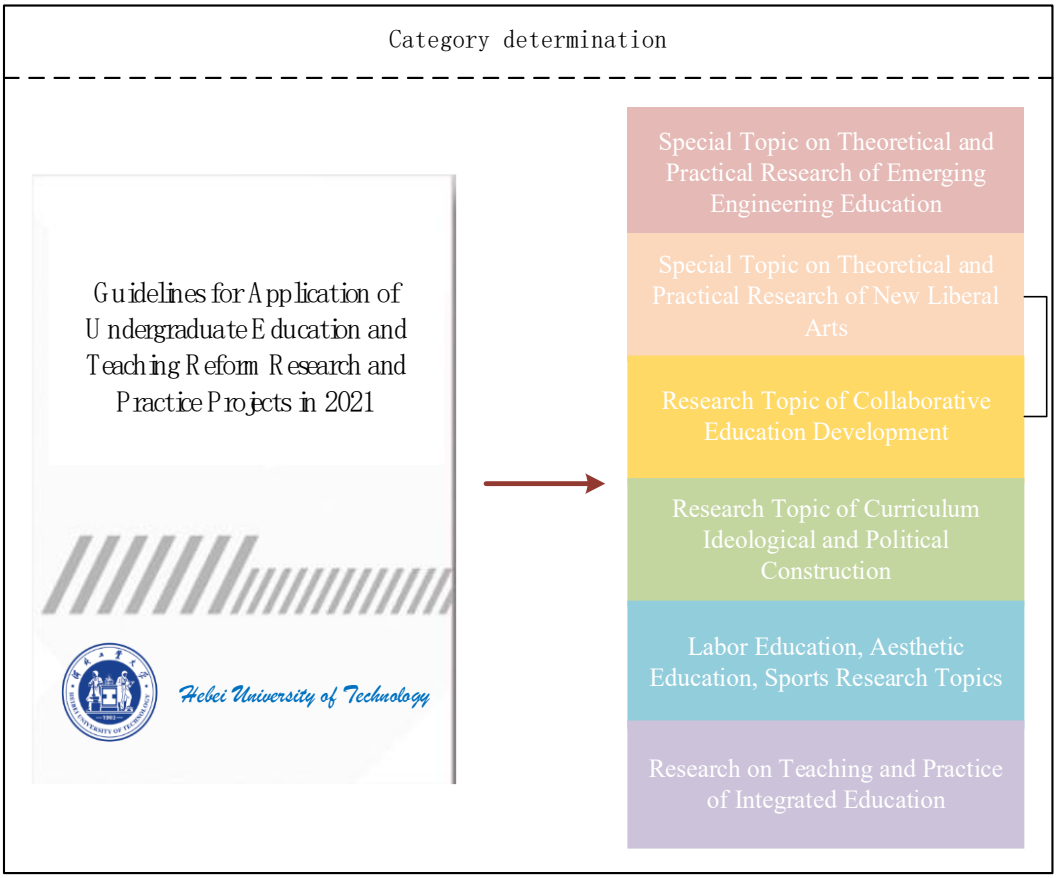

Figure 2. Category determination.

2.2. Data Set Extraction

China National Knowledge Infrastructure (CNKI) is one of the largest academic literature databases in China, gathering a vast amount of resources in the education field, including academic journals, conference papers, and master's and doctoral dissertations [12,13]. It covers a wide range of research areas related to educational reform. Its rich dataset provides a guarantee for the large training set required for the systematic design of this paper. At the same time, this paper is mainly aimed at classifying Chinese educational reform documents. As CNKI contains abundant Chinese literature, it can provide more specific information on practical applications, policies, and theoretical research related to educational reform in China. Therefore, this paper selects CNKI and the university's database as the sources of the dataset.

To accurately obtain each category of educational reform documents, this paper utilizes the advanced search mode of China National Knowledge Infrastructure (CNKI) and employs the keywords compiled in Section 2.1 as search terms. The paper uses "AND", "OR", and "NOT" operators to construct retrieval strategies for educational reform documents.

Taking the "Special Topics on Research on New Engineering Theory and Practice" as an example, a simple retrieval strategy is established as follows: "New Engineering" AND "Educational Reform" OR "Emerging Engineering" OR "New Type of Engineering" OR "Traditional Engineering" NOT "New Arts" NOT "Collaborative Education" NOT "Ideological and Political Education Construction" NOT "Labor Education, Aesthetic Education, Physical Education" NOT "Specialized and Innovative Education Teaching Integration". The goal is to retrieve educational reform documents related to "Special Topics on Research on New Engineering Theory and Practice" while being unrelated to the other five categories. The search results will be downloaded and saved into the corresponding folders for subsequent model training and evaluation. Finally, a Python program will be

used to count the number of documents in the training set, with each category containing hundreds of files, as depicted in Figure 3.

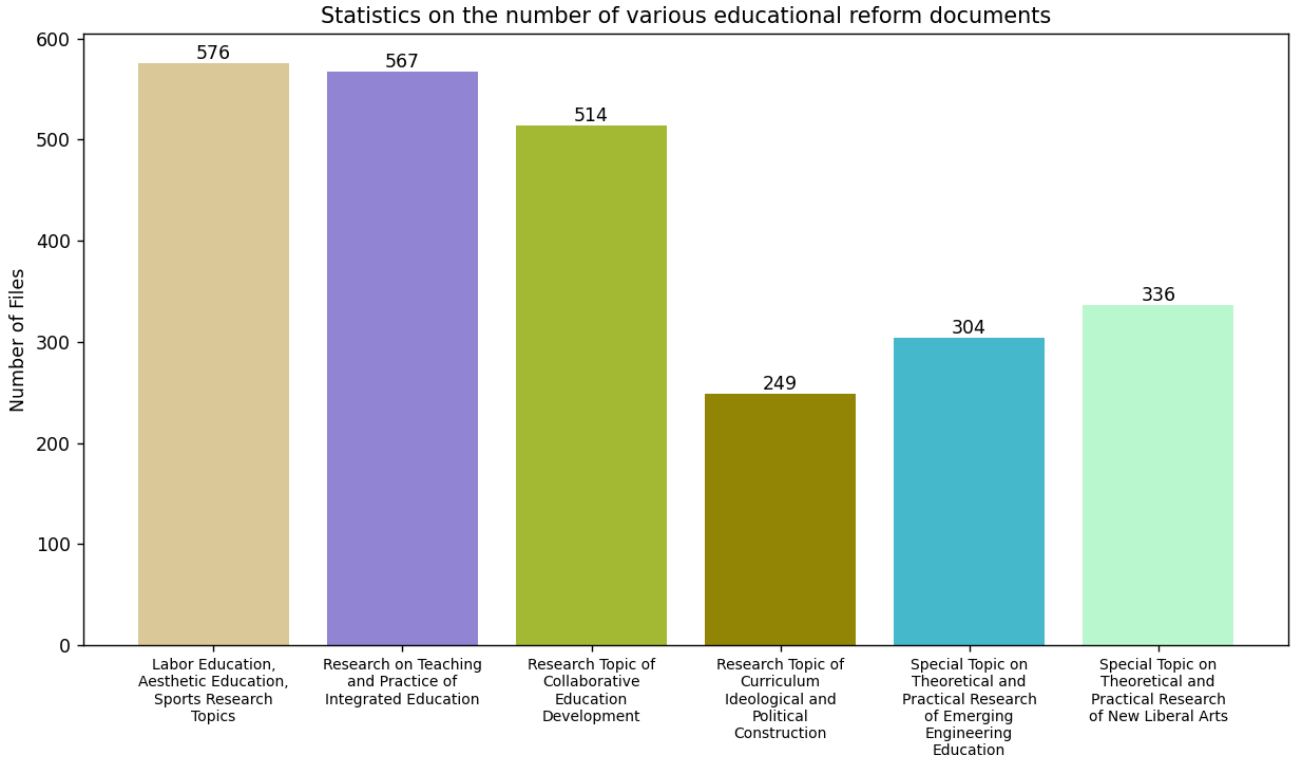

Figure 3. Statistics on the number of various educational reform documents.

2.3. Algorithm Selection and Application

2.3.1. Data Preprocessing

Text preprocessing plays a crucial role in text classification tasks. Its main purpose is to clean and transform raw text data to make them more suitable for input into machine learning algorithms [14,15]. Through text preprocessing, raw text data can be converted into more standardized, clean, and structured data, providing a better foundation for feature extraction and model training. The preprocessing process can reduce data noise, and improve the accuracy and efficiency of classification tasks.

For the text types involved in this article, the preprocessing mainly consists of the following steps. Firstly, Python-based natural language processing (NLP) technology is used to segment the original text data set [14,15], each sentence is processed as a unit, and the resulting sentences are segmented into word sequences. Secondly, the words after word segmentation are compared with a predefined list of stop words to remove the meaningless or too common words, such as 'the', 'is', 'in', etc. Thirdly, the part-of-speech tagging of words after word segmentation is carried out; that is, the part-of-speech of each word in the sentence is determined, such as nouns, verbs, adjectives, etc., which is helpful for the model to understand the sentence structure, the relationship between words, and semantic information. Finally, the data set after word segmentation is constructed and saved as a binary file.

After preprocessing the initial dataset, it is necessary to construct the TF-IDF (term frequency-inverse document frequency) term frequency space vector. The TF-IDF algorithm is a common statistical method used to measure the importance of words in information retrieval and data mining. The TF-IDF algorithm is used to measure the importance of a word in a corpus, and it calculates the TF-IDF value of a word using the product of

term frequency (TF) and inverse document frequency (IDF) [16,17]. The specific formula is as follows:

$$TF_{w,D_i} = \frac{count(w)}{|D_i|} \quad (1)$$

$$IDF_w = \log \frac{N}{\sum_{i=1}^{N} I(w, D_i)} \quad (2)$$

$$TF - IDF_{w,D_I} = TF_{w,D_i} \times IDF_w \quad (3)$$

In the above formula, $count(w)$ represents the frequency of the term in the document, and $|D_i|$ is the total number of words in the document. TF represents the frequency of the term in the document, and a higher TF reflects that the term appears more frequently. IDF reflects the commonness of a term in the document. When a large number of documents contain a word, the IDF is lower, indicating that the word is more common [18].

The TF-IDF value of a word is equal to the product of its TF and IDF. A higher TF-IDF value for a word indicates that the word is more important in the document and better represents the document [19]. In text feature extraction, TF-IDF performs well, and is easy to implement and use. Therefore, this article selects the TF-IDF algorithm to calculate the importance of each word in the text dataset and constructs a word frequency matrix. Finally, the same processing is performed on the test set.

2.3.2. Algorithm Selection

After completing the above tasks, the next step is to select a text classification algorithm. Common algorithms used for Chinese text classification include Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest Algorithm, and Naive Bayes Algorithm. This article will introduce these four algorithms, conduct model testing, and ultimately determine the algorithm to be used.

Multilayer Perceptron (MLP), also known as Artificial Neural Network (ANN), is built upon a single-layer neural network by introducing one or more hidden layers, creating a neural network with multiple layers, hence the name "multilayer perceptron". The hidden layers are located between the input layer and the output layer [20,21]. The simplest MLP contains only one hidden layer, making it a three-layer structure, as shown in Figure 4.

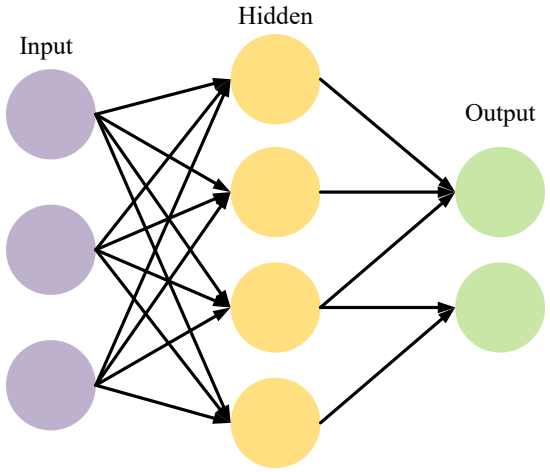

Figure 4. Schematic diagram of MLP.

MLP has powerful learning capabilities and can automatically extract features from training data for classification, making it very effective in handling complex text classification problems. Additionally, MLP can utilize multiple hidden layers to enhance the model's representational capacity. It can also leverage different activation functions and regularization methods to optimize model performance, giving it good scalability when

dealing with large-scale text classification problems. However, due to the large number of parameters in MLP, the training process tends to be slow and may struggle with processing large amounts of text data. Moreover, when dealing with small amounts of training data, it is prone to overfitting.

SVM is a supervised learning algorithm widely used for classification and regression tasks [22]. The core idea of SVM is to find a hyperplane that maximizes the margin between two classes, thus achieving good classification performance. Its advantages lie in its ability to perform well in classifying data in high-dimensional spaces, making it suitable for handling high-dimensional features, as demonstrated in fields such as text classification and image recognition. SVM is suitable for small sample datasets, as it determines the decision boundary based on support vectors rather than the entire dataset, giving it an advantage in handling small sample data. Additionally, SVM can use kernel functions to map data to high-dimensional spaces, enabling nonlinear classification and providing good flexibility. However, SVM also has some drawbacks. Firstly, it incurs significant computational costs and longer training times when dealing with large-scale datasets. Secondly, SVM is sensitive to missing data and requires additional handling methods. Lastly, the original SVM algorithm is only applicable to binary classification problems, requiring extensions for multi-class problems, which increases the complexity of the algorithm.

Random Forest is a classifier that trains and predicts samples using multiple trees. It repeatedly randomly samples k samples with replacement from the original training sample set N to generate a new training sample set, and then uses the bootstrap sample set to generate k classification trees to form a random forest [23–25]. The classification result of new data is determined by the score based on the voting of the classification trees. Random Forest has the following advantages: firstly, it can handle high-dimensional data and performs well in feature selection, automatically selecting important features and reducing the burden of feature engineering. Secondly, Random Forest is efficient in handling large-scale datasets and can quickly build a large number of decision tree models. Furthermore, Random Forest exhibits good robustness to missing data and outliers, capable of handling incomplete datasets. Additionally, Random Forest uses a voting mechanism for classification or regression, which can reduce the risk of overfitting and is less susceptible to the influence of individual decision trees. However, Random Forest also has some drawbacks. Firstly, due to the ensemble of multiple decision trees, the model has poor interpretability compared to a single decision tree. Secondly, Random Forest may overfit on datasets with high noise and may require significant memory space. Moreover, Random Forest may lean towards categories with more samples when dealing with highly imbalanced datasets.

Naive Bayes algorithm is a classification algorithm based on Bayes' theorem, assuming that features are independent of each other. Under a given category, the joint probability of features equals the product of the independent probabilities of each feature. By calculating the posterior probability of each category, the category with the highest posterior probability is chosen as the prediction result [26]. Bayes' theorem is an important theorem in probability theory used to calculate the probability of an event occurring given certain conditions [27]. Its calculation formula is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4}$$

In the formula, $P(A|B)$ represents the probability of event A occurring given that event B has occurred, $P(B|A)$ represents the probability of event B occurring given that event A has occurred, and $P(A)$ and $P(B)$ represent the probabilities of events A and B occurring, respectively.

The Naive Bayes algorithm has the following advantages. Firstly, it has relatively fast training and prediction speeds because it assumes independence between features, reducing the complexity of parameter estimation. Secondly, the Naive Bayes algorithm performs well on small sample datasets and can handle high-dimensional feature data, making it suitable for tasks such as text classification and spam filtering. Additionally, the

Naive Bayes algorithm is robust to missing data, and can handle partially missing sample data. Finally, the Naive Bayes algorithm is not easily influenced by noisy data, and can provide good classification results even for datasets with significant noise. However, the Naive Bayes algorithm also has some disadvantages. Firstly, it assumes independence between features, which may not hold true in some cases, leading to a decrease in classification performance. Secondly, the Naive Bayes algorithm may exhibit higher error rates for datasets with large feature spaces or strong feature correlations. Furthermore, the Naive Bayes algorithm cannot handle continuous features and requires discretization.

The four algorithms mentioned above each have their own advantages and disadvantages in text classification, making it difficult to directly choose a classification algorithm. Therefore, it is necessary to build and test models for them. Four algorithm testing models were built separately, and the classification process of the testing models is shown in Figure 5.

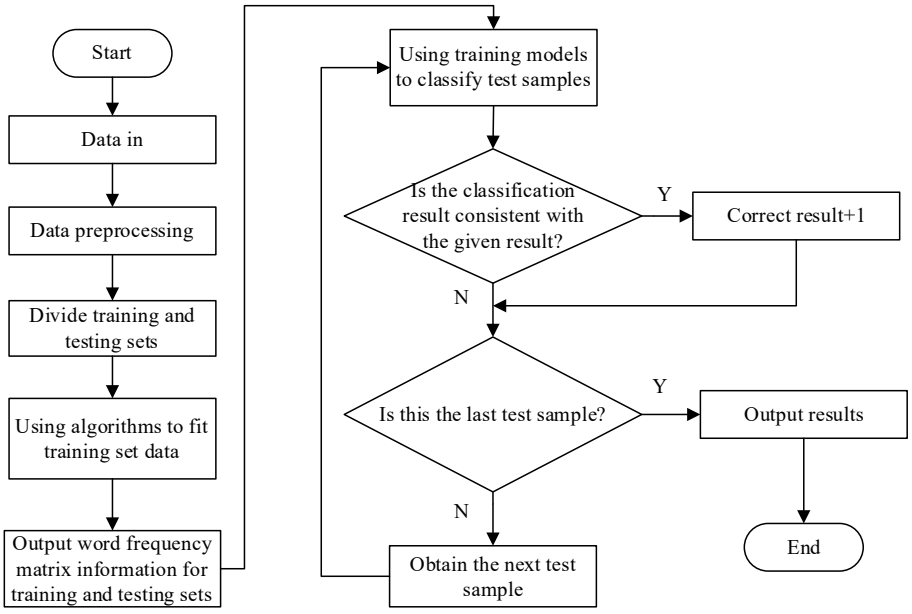

Figure 5. Classification process of the testing models.

The classification process of the test model is as follows:

1. Data preprocessing: Pre-process the data, including steps such as data cleaning, denoising, and standardization.
2. Divide the training and testing sets: Divide the pre-processed data into training and testing sets, and divide the dataset into training and testing sets in an 8:2 ratio.
3. Use algorithms to fit the training set data: Select the above four algorithms to train the data and establish a model.
4. Output word frequency matrix information: Perform word frequency statistics on the training and testing sets to generate word frequency matrix information.
5. Use training models to classify test samples: Use the established model to classify test samples.
6. Determine whether the classification result is correct: Evaluate the classification result to determine whether it is consistent with the given result. If the classification is consistent, the correct result is+1; if not correct, proceed to the next step.
7. Determine if it is the last test sample: Determine if there are still unclassified test samples, and if so, continue to classify the next test sample; if not, output the result and the classification ends.
8. Obtain the next test sample: Obtain the next test sample for classification testing, and then repeat steps 5–7.

After completing the model building and testing, it is necessary to evaluate the performance of the four models to determine the final model to be used. The evaluation index is a quantitative indicator of the performance of a model. A single evaluation indicator can only reflect a portion of the model's performance. If the selected evaluation indicator is unreasonable, it may lead to incorrect conclusions. Therefore, this study used commonly used evaluation indicators in classification tasks, including Precision, Recall, F1 score, Accuracy, Macro average (represented by Macro avg in the table), and Weighted average (represented by Weighted avg in the table) to comprehensively evaluate the model. The results are shown in Table 1, where Topic 1–Topic 6 represents the six categories from left to right in Figure 3. R-F stands for Random Forest Algorithm. N-B represents the naive Bayesian algorithm. The dataset is divided into training and testing sets in an 8:2 ratio. Therefore, Topic 1–Topic 6 obtained 58, 57, 51, 25, 30, and 33 test sets in sequence for model evaluation.

Table 1. Calculation results of various indicators.

	Precision				**Recall**				**F1-Score**				**Support**
	MLP	**SVM**	**R-F**	**N-B**	**MLP**	**SVM**	**R-F**	**N-B**	**MLP**	**SVM**	**R-F**	**N-B**	
Topic 1	0.93	0.90	0.75	0.96	0.89	0.90	0.74	0.90	0.91	0.90	0.75	0.93	58
Topic 2	0.93	0.96	0.82	0.97	0.91	0.88	0.79	0.98	0.92	0.92	0.80	0.97	57
Topic 3	0.92	0.92	0.86	0.96	0.86	0.90	0.73	0.96	0.89	0.91	0.79	0.96	51
Topic 4	0.70	0.83	0.63	0.88	0.84	0.96	0.60	0.88	0.76	0.89	0.61	0.88	25
Topic 5	0.84	0.90	0.61	0.85	0.90	0.90	0.83	0.97	0.87	0.90	0.70	0.91	30
Topic 6	0.91	0.86	0.71	0.93	0.88	0.91	0.73	0.91	0.89	0.88	0.72	0.92	33
Accuracy									0.89	0.90	0.75	0.94	255
Macro avg	0.87	0.90	0.73	0.93	0.88	0.91	0.74	0.93	0.87	0.90	0.73	0.93	255
Weighted avg	0.89	0.90	0.75	0.94	0.88	0.90	0.74	0.94	0.89	0.90	0.75	0.94	255

From Table 1, it can be seen that the Naive Bayes algorithm leads the other three algorithms in most indicators in the classification process of educational reform documents. Although the SVM algorithm is leading in some indicators in the Topic 4 classification process, the lead is not significant, and from a comprehensive performance perspective, the Naive Bayes algorithm is still significantly better than the SVM algorithm. Therefore, this study selected Naive Bayes algorithm as the classification algorithm for subsequent software design.

2.4. Software Design

In order to enhance the usability of the classification algorithm, this article designed auxiliary software based on relevant Python3.9 technologies for multiple steps in the process, ultimately creating an automated classification assistance system for education reform documents using the Naive Bayes algorithm.

2.4.1. Teaching Reform File Information Extraction System

The teaching reform file information extraction interface is shown in Figure 6, which is mainly used for extracting the required information from the dataset and extracting information from the documents to be classified later. The information retrieval process is divided into five steps:

1. Obtain relevant education reform documents based on the classification category from patent retrieval systems such as CNKI (China National Knowledge Infrastructure).
2. When the user clicks the "Browse" button, the system will open a file dialog box allowing the user to select the path where the files to be classified are located. The program will then display the selected directory path in the text box corresponding to that button.
3. When the user clicks the "Reading files" button, the program will read all file names in the previously selected directory and display these file names in a text box.

4. When the user clicks the "Save" button, the program will open a file dialog box allowing the user to choose a directory. The program will then iterate through all PDF files in the previously selected directory, extract the abstracts and keywords from these files, and save the abstracts and keywords to a TXT file named after the PDF file. These TXT files will be saved in the user-selected directory.
5. Implement the functionality to navigate to the next interface, which is bound to the "NEXT" button. When the user clicks this button, the program will close the current interface and call the run_next_script method to execute the next script. In the run_next_script method, the program will start the next script using the subprocess Popen method.

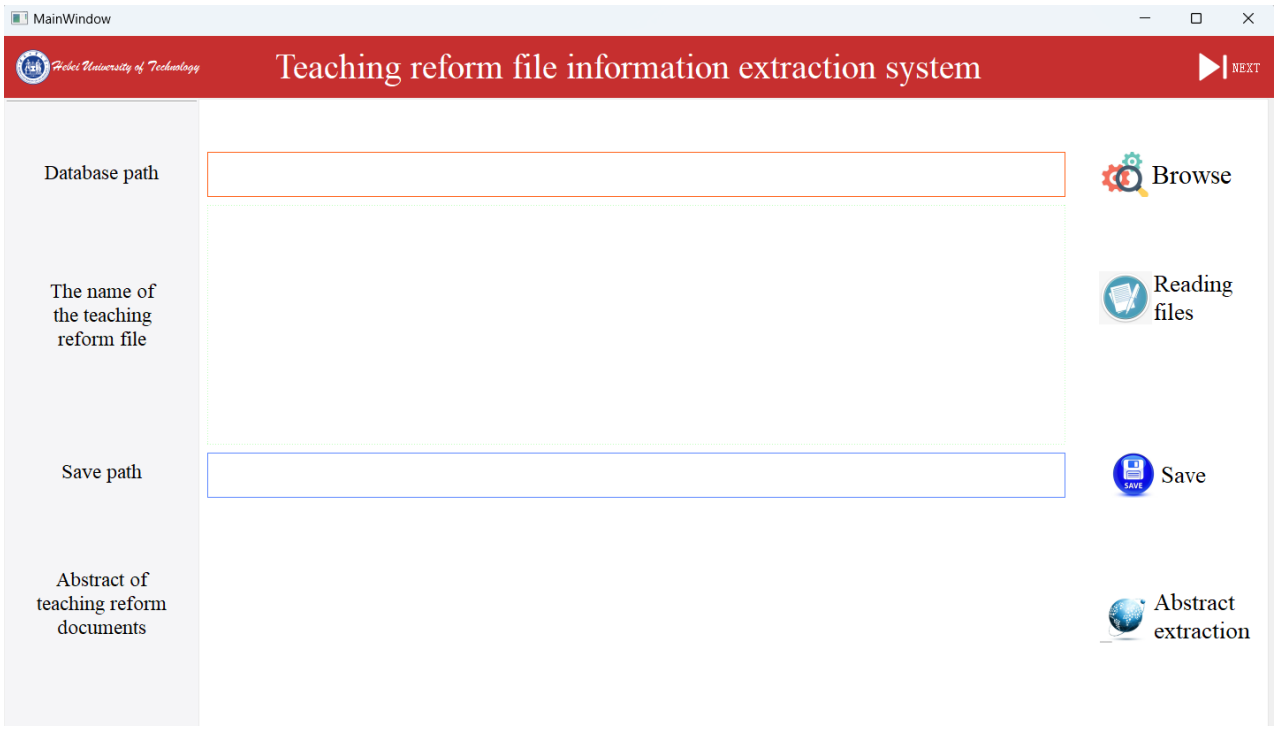

Figure 6. Teaching reform file information extraction system.

2.4.2. Automatic Classification System of Educational Reform Documents

The automatic classification system of educational reform documents consists of two modules: the "Classification" module, and the "Result statistics" module.

The "Classification" module, as shown in Figure 7, achieves the classification of the files to be processed by utilizing a model trained with the Naive Bayes algorithm. The specific implementation steps are as follows:

1. Initialize the interface, load the UI file, and connect the click events of the buttons.
2. After clicking the "Browse" button, a folder selection dialog will pop up, allowing the user to choose the folder containing the files to be classified, and display the selected path in the corresponding text box.
3. After clicking the "Reading files" button, read all the files in the selected folder and display the file names in the connected text box.
4. After clicking the "Articles classification" button, the system will start classifying the files in the selected folder. First, it will iterate through all the TXT files in the folder; next, it will open and read the content of each file, tokenize the content, and remove stop words. Then, it will construct the TF-IDF term frequency vector space, convert the tokenized text into a TF-IDF frequency matrix, and save it as a binary file. Finally, it will utilize the model built using the Naive Bayes algorithm to calculate the conditional probabilities of different categories based on the TF-IDF vectors, thus achieving

the classification of the test files. The classification results will be output and written into an Excel file. The classification results will be displayed in the text box associated with the "Articles classification" button.

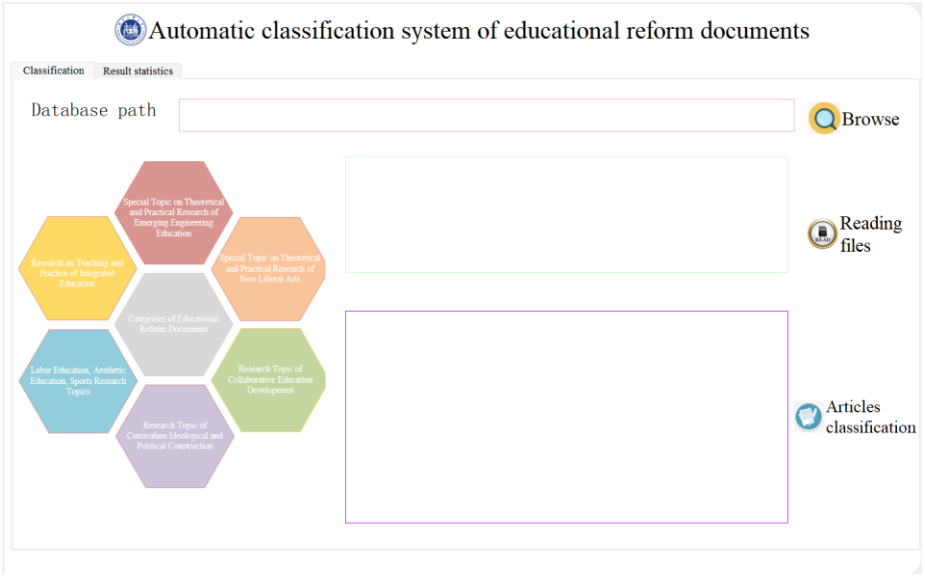

Figure 7. The classification module of Automatic classification system of educational reform documents.

The "Result statistics" module, as shown in Figure 8, mainly focuses on the statistical analysis of the classification results. The implementation steps are as follows:

1. After clicking the "Result statistics" button, read the classification results saved in the Excel file from the above interface, count the number of each category, and display the number of each category and the corresponding file names in their respective text boxes.
2. After clicking the "Pie chart" button, generate a pie chart based on the classification results to visualize the proportion of each file category within the processed folder. The corresponding results will be displayed in the output box associated with the "Pie chart" button.

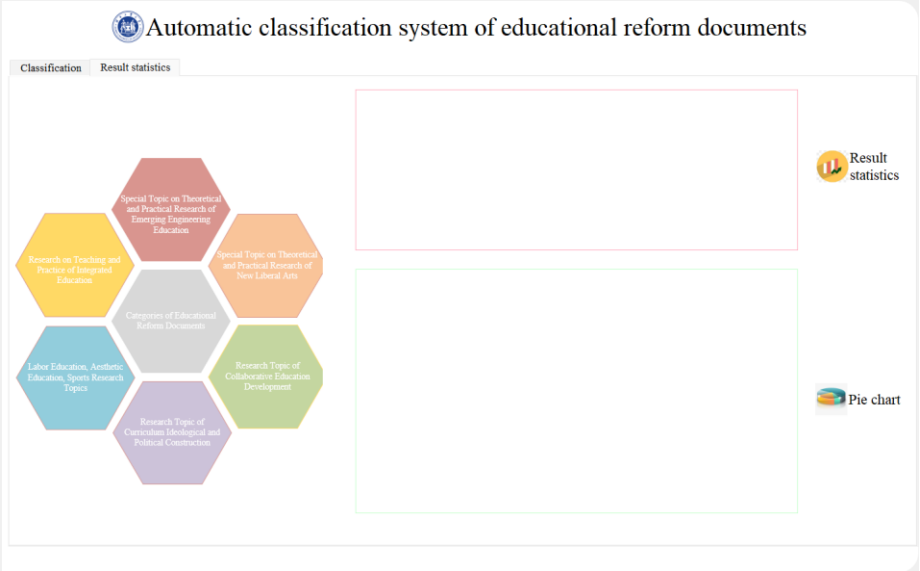

Figure 8. The Result statistics module of Automatic classification system of educational reform documents.

3. Experimental Verification, Taking the Educational Reform Documents of Hebei University of Technology as an Example

3.1. Background

In recent years, the deepening reforms in education and teaching have been continuously advancing, leading to a surge in the number of educational reform documents. This has resulted in challenges, such as a massive quantity of educational reform documents and a wide variety of document types. However, traditional manual classification methods require significant human and time costs, resulting in low efficiency and potential classification errors. Moreover, as the number of documents increases and the types become more diverse, manual classification becomes increasingly difficult. To address these issues, this paper proposes an automatic classification system for educational reform documents based on the Naive Bayes algorithm. This system is designed to assist education administrators in automatically classifying a large number of educational reform documents.

3.2. Extracting Information from Educational Reform Documents of Hebei University of Technology

Using the educational reform documents related to Hebei University of Technology from the past two years as samples, we obtained 52 educational reform documents awaiting classification from the school database and stored them in the corresponding folder for further processing.

After obtaining the files to be classified, we used the Educational Reform Document Information Extraction System to extract the abstract sections of the pending documents, as shown in Figure 9. Firstly, click "Browse" to select the folder to be processed, and the system will automatically display the path of the folder in the corresponding text box. Then, click "Reading files", and the names and file types of the files to be processed in this folder will be automatically displayed in the text box corresponding to this button. Furthermore, use the "Save" button to determine the location for saving the processed text. Finally, use the "Abstract extraction" button to extract the abstracts from the files to be processed and save them as TXT files in the location selected by the "Save" button for future processing.

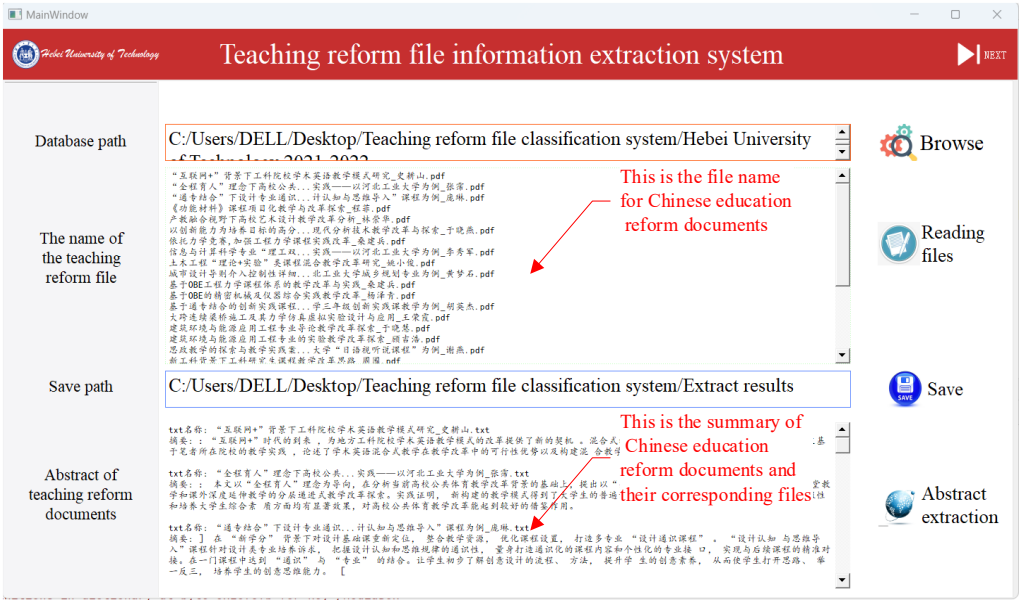

Figure 9. Information extraction results of teaching reform file from Hebei University of Technology.

3.3. Automatic Classification of Educational Reform Documents at Hebei University of Technology

First, use the "Classification" module to classify the educational reform documents of Hebei University of Technology to be processed, as shown in Figure 10.

1. Click "Browse" and select the file path saved in the previous step.
2. Click "Reading files", and the names and file types of the files to be processed in this folder will be automatically displayed in the text box corresponding to this button.
3. Utilize the model trained using the Naive Bayes algorithm to classify the documents to be processed. The classification results of the educational reform documents at Hebei University of Technology will be displayed in the text box corresponding to the "Articles classification" button.

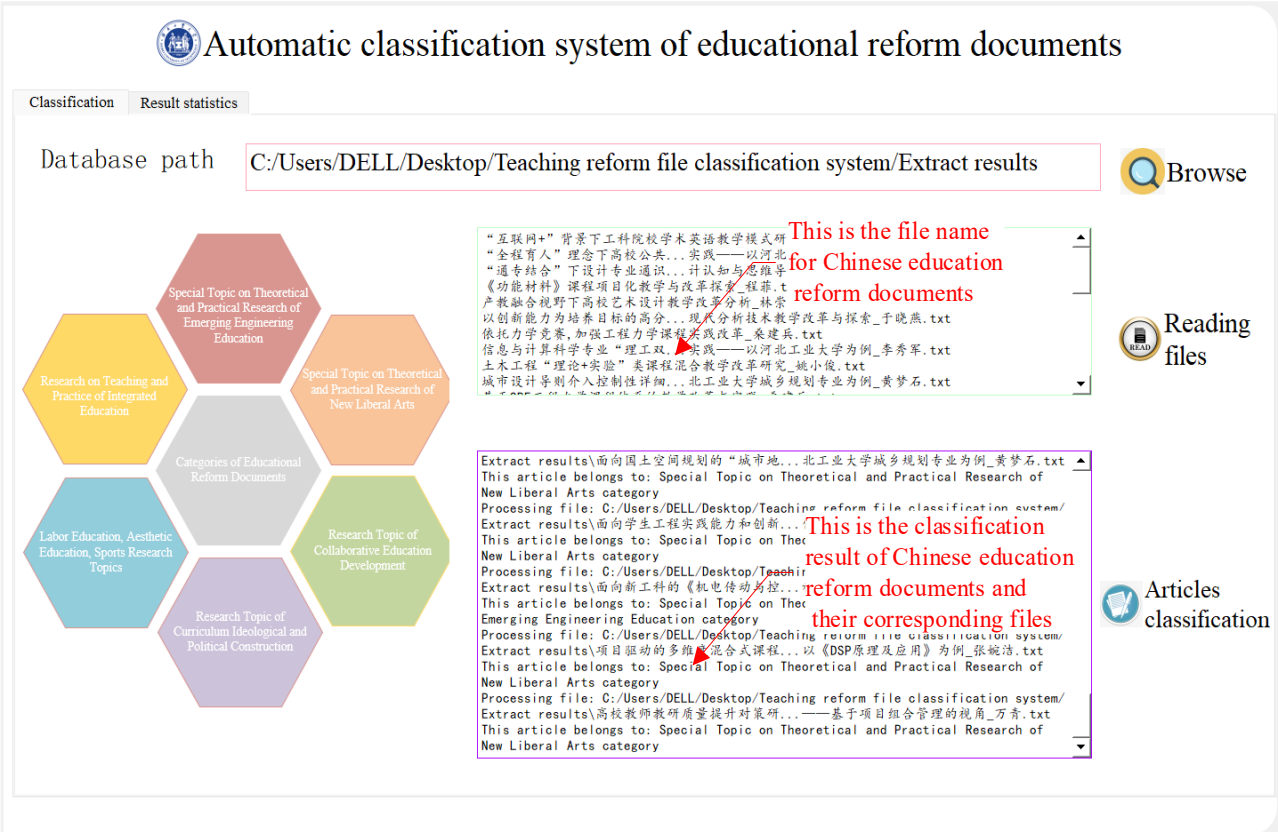

Figure 10. The classification process of teaching reform file from Hebei University of Technology.

Next, use the "Result statistics" module as shown in Figure 11. This module is mainly used to provide statistical analysis of the classification results of the educational reform documents at Hebei University of Technology. The number and names of documents in each category will be displayed in the text box corresponding to the "Result statistics" button. Additionally, a pie chart will be generated based on the classification results, allowing for an easy analysis of the distribution of each type of document within the processed folder. The corresponding results will be displayed in the output box corresponding to the "Pie chart" button. From the results, it can be seen that Hebei University of Technology has invested heavily in educational reform in the past two years, with a focus on the "Special Topic on Theoretical and Practical Research of New Liberal Arts" and the "Special Topic on Theoretical and Practical Research of Emerging Engineering Education", accounting for 46.2% and 36.5%, respectively. Therefore, this system can be used to analyze the key directions of educational reform in a certain area for the current year.

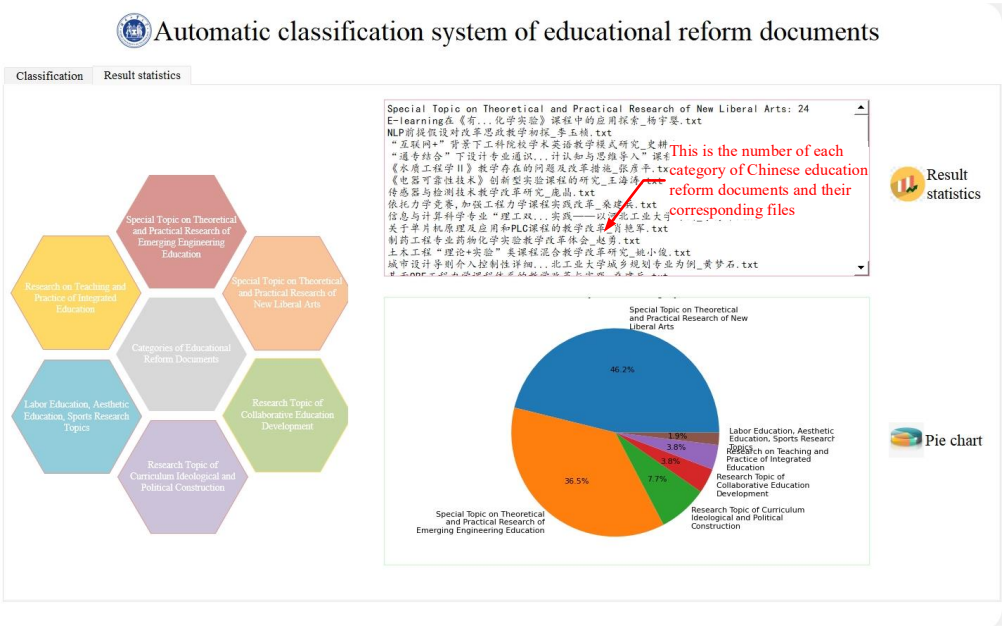

Figure 11. Statistics on the classification results of teaching reform file from Hebei University of Technology.

Finally, the educational reform documents at Hebei University of Technology from the past two years will be manually classified by professional personnel. The results of the manual classification will then be compared with the classification results of the automatic educational reform document classification system based on the Naive Bayes algorithm, and the comparison is presented in Figure 12. From Figure 12, it can be observed that the classification results of the automatic educational reform document classification system based on the Naive Bayes algorithm are largely consistent with the classifications made by professional personnel, with only a few instances of misclassification.

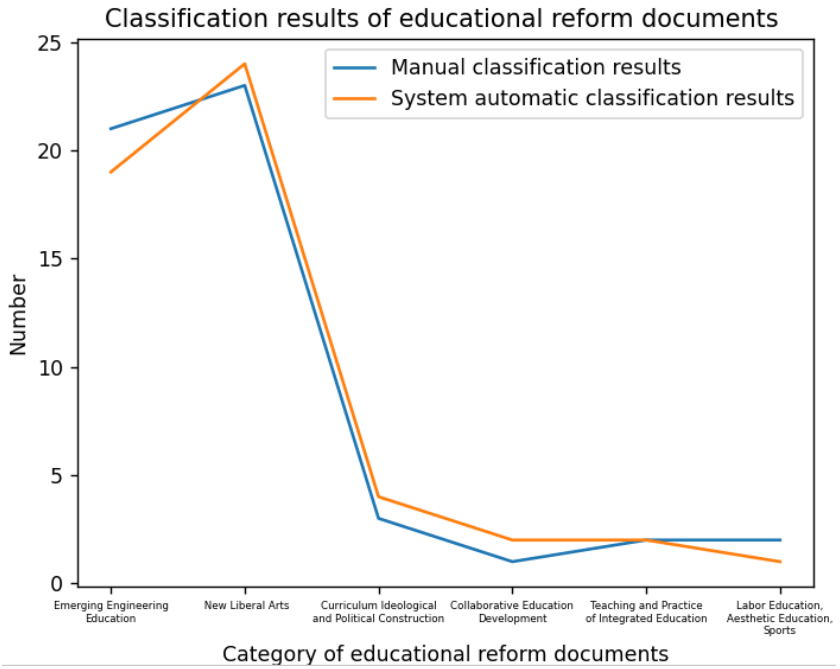

Figure 12. Comparison of classification results.

4. Verification and Discussion

Through in-depth analysis of questionnaire surveys and systematic evaluations, the universal applicability and effectiveness of the educational reform document automatic classification system based on the Naive Bayes algorithm designed in this paper have been validated.

To further demonstrate the universality and efficiency of the current design of the educational reform document automatic classification system based on the Naive Bayes algorithm, 20 relevant education professionals were invited to evaluate the system and fill out a questionnaire. The Chinese version and English version of the questionnaire are provided in Table 2. The 20 invited workers were divided into two groups based on their experience level: experienced workers, and novices. Figures 13 and 14 summarize the survey results, where EW-A represents experienced workers choosing option A, and N-A represents novices choosing option A.

Table 2. Questionnaire.

Chinese Version	**English Version**
您的身份**:**	**Identity:**
Q1:使用该系统是否对相应文件分类有帮助?	**Q1**: Does the system help in classifying the corresponding files?
☐很有帮助☐有一定帮助☐没有帮助	**A** Very helpful. **B** Some help. **C** No help.
Q2:使用该系统是否能解决由于专业知识不足而导致分类错误?	**Q2**: Can the use of the system solve the classification errors caused by lack of professional knowledge ?
☐能☐一定程度上可以☐不能	**A** Yes. **B** To some extent. **C** No.
Q3:使用该系统是否能提高文件分类速度?	**Q3**: Can the system improve the speed of file classification ?
☐能☐一定程度上可以☐不能	**A** Yes. **B** To some extent. **C** No.
Q4:使用该系统是否能提高文件分类准确性?	**Q4**: Can the use of the system improve the accuracy of file classification ?
☐能☐一定程度上可以☐不能	**A** Yes. **B** To some extent. **C** No.
Q5:该系统是否有实用价值?	**Q5**: Does the system have practical value ?
☐有☐有一定价值☐没有	**A** Yes. **B** To some extent. **C** No.
Q6:使用该系统是否能得到期望的分类效果?	**Q6**: Can the system be used to get the desired classification results ?
☐能☐一定程度上可以☐不能	**A** Yes. **B** To some extent. **C** No.
Q7:使用该系统时是否遇到困难,使用效果如何?	**Q7**: Do you encounter difficulties when using the system, and what is the effectiveness of its usage?
☐使用简单,没有任何困难	**A** User-friendly, no difficulty.
☐有一些困难,使用效果一般	**B** Some difficulties, average performance.
☐完全不会使用,非常困难	**C** Not used at all, very difficult.
Q8:是否推荐其他人员使用?	**Q8**: Do you recommend other personnel to use it
☐是的!我推荐☐可以尝试☐不推荐	**A** Yes. **B** Probably. **C** No.
其他:	Other:
_____	_____
	Q is question

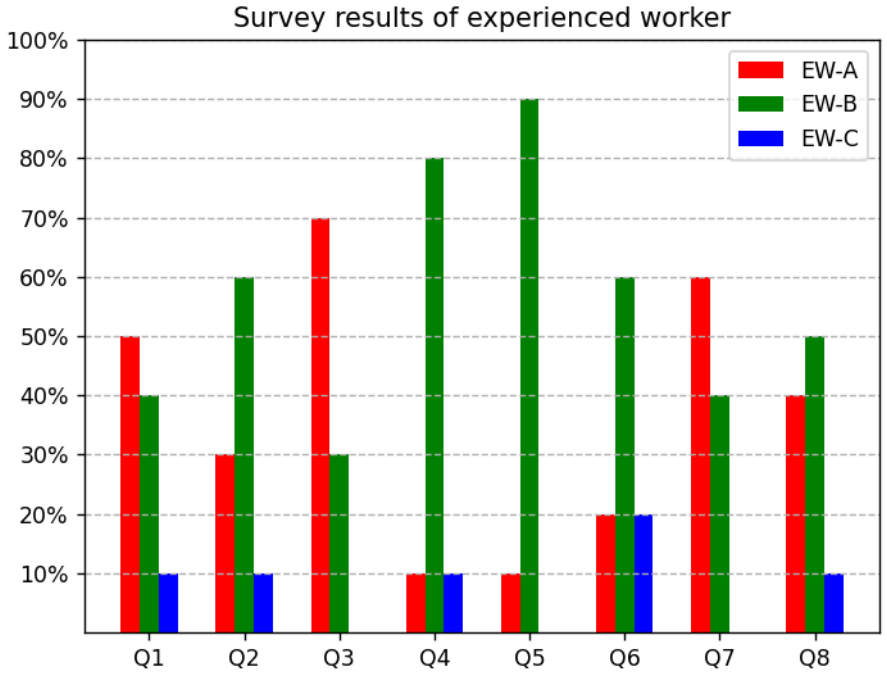

Figure 13. Survey results of experienced workers.

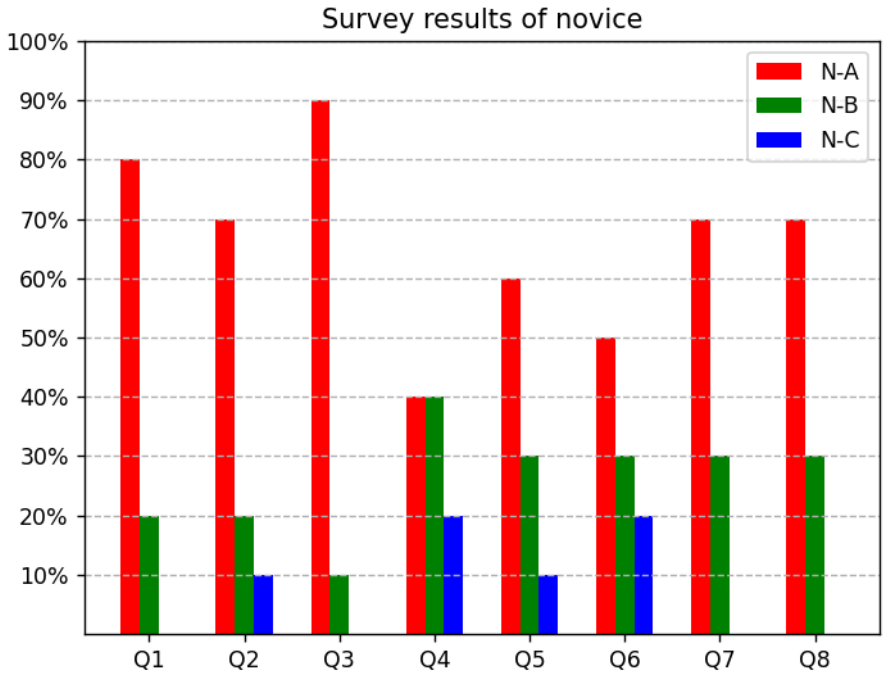

Figure 14. Survey results of novice.

4.1. Analysis of Survey Results from Experienced Worker

As shown in Figure 13, from Q1 and Q2, it can be seen that 90% of experienced workers believe that the current design of the classification system is useful and can compensate for the lack of professional knowledge reserves among workers.

According to the results of Q3, Q5, and Q7, 100% of experienced workers believe that the current design of the classification system can improve document classification speed, has practical value, and performs well in its usage.

From Q4 to Q8, it is evident that 90% of experienced workers believe that the system can improve document classification accuracy and are willing to recommend it to others.

Based on the results of Q6, 80% of experienced workers believe that the current design of the classification system can achieve the expected classification results.

4.2. Analysis of Survey Results from Novice

As shown in Figure 14, from Q1 and Q3, it can be seen that 100% of novices believe that the current design of the classification system is useful and can improve document classification speed.

According to the results of Q2 and Q5, 90% of novices believe that the current design of the classification system can compensate for the lack of professional knowledge reserves among workers and has practical value.

From Q4 to Q6, it is evident that 80% of novices believe that the system can improve document classification accuracy and achieve the expected classification results.

Based on the results of Q7 and Q8, 100% of novices believe that the current design of the classification system performs well in its usage and are willing to recommend it to others.

4.3. Discussion

To select the most suitable algorithm for classifying educational reform documents, we trained classification models based on multilayer perceptron, support vector machine, random forest algorithm, and the Naive Bayes algorithm. After data testing, we found that the classification model based on the Naive Bayes algorithm performed the best, with a prediction accuracy of about 94%.

Based on the analysis of the experimental results, our designed automatic classification system for educational reform documents based on the Naive Bayes algorithm was generally consistent with the classification of professional personnel, with only a few misclassified files. The classification results were acceptable. At the same time, applying this software can reduce the demand for professional knowledge reserves of management personnel and reduce classification errors caused by subjective factors. Therefore, this system is feasible for classifying educational reform documents.

Finally, we verified the universal applicability and effectiveness of our designed automatic classification system for educational reform documents based on the Naive Bayes algorithm through survey questionnaires. The results showed that this system is helpful for education workers, especially for novices.

5. Conclusions

Through the organization of educational reform documents and algorithm reuse, we have designed an automatic classification system for educational reform documents based on the Naive Bayes algorithm. Experimental verification has shown that the system can effectively classify educational reform documents automatically, filling the gap in the application of text classification in this category of documents and reducing the demand for professional knowledge of management personnel. The main contributions of the current research can be summarized as follows:

1. Determined the classification direction for educational reform documents based on the "Guidelines for the Declaration of Undergraduate Education Reform Research and Practice Projects in 2021".
2. Constructed retrieval strategies for educational reform documents using "AND", "OR", and "NOT" combinations based on the keywords obtained from the guidelines.
3. Conducted data cleaning, algorithm construction, training, and selection for educational reform documents, identifying the most suitable classification algorithm for this type of document.
4. Developed an automatic classification software for educational reform documents based on the naive Bayes algorithm, which significantly reduces classification time

while lowering the requirement for professional knowledge storage of classifiers, all while ensuring accuracy.

Despite the contributions of this system, its limitations are evident. Educational reform documents often involve multiple topics and fields, and sometimes a document may belong to multiple categories, making it difficult for the system to handle multi-label classification problems. This system is designed for Chinese documents only, and supports the classification of single-language documents, making it unable to accurately classify documents in other languages. Additionally, education reform is a constantly evolving field with new themes and concepts emerging regularly. Therefore, the classification system requires regular updates and maintenance to adapt to new developments in educational reform. These limitations point to future research directions. It is necessary to explore the use of more scientific algorithms for document analysis to make the results more objective and effective. Designing the software to be scalable to accommodate new educational reform topics and concepts is also important. Future research will also focus on designing classification systems for different language types and continuously improving the analysis of educational reform documents while enriching the corresponding training databases.

Author Contributions: Conceptualization, P.Z., Z.M. and Q.W.; methodology, P.Z. and Z.M.; writing—original draft preparation, Z.M.; data curation, Z.R. and Q.W.; writing—review and editing, H.W. and D.S.; visualization, C.Z.; supervision, P.Z.; funding acquisition, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was funded by the National Natural Science Foundation of China (Grant No. 51975181), the National Innovation Method Fund of China (Grant No. 2020IM020500) and Hebei Province Teaching Reform Project (Grant No. 2022GJJG042 and No. 2022GJJG038).

Data Availability Statement: The data is contained in the article and is available from the corresponding authors on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Maron, M.E. Automatie indexing: An experimental in-quiry. *J. ACM* **1961**, *8*, 404–417. [CrossRef]
2. Zhang, Z.J.; Wang, Z.Q. Summary of text classification and algorithm. *Comput. Knowl. Technol.* **2012**, *8*, 825–828.
3. Alina, P.; Ioana, D.; Yoon, H.J.; Mohd, Y.J.; Tanmoy, B. Deep learning uncertainty quantification for clinical text classification. *J. Biomed. Inform.* **2024**, *149*, 104576.
4. Liu, H.; Hao, Y.; Zhang, W.H.; Zhang, H.Y.; Gao, F. Online urban-waterlogging monitoring based on a recurrent neural network for classification of microblogging text. *Nat. Hazards Earth Syst. Sci.* **2021**, *21*, 1179–1194. [CrossRef]
5. Yang, X.Z.; Wang, Q.Q.; Jiang, J.L. Analysis of classroom teacher-student dialogue based on artificial intelligence:automatic classification and sub-level construction of lRE. *E-Educ. Res.* **2023**, *44*, 79–86.
6. Vishaal, U.; Abhishek, A.; Anubha, G.; Tanmoy, C. InPHYNet: Leveraging attention-based multitask recurrent networks for multi-label physics text classification. *Knowl.-Based Syst.* **2021**, *211*, 106487.
7. Chen, Y.W.; Wang, J.L.; Cai, Y.Q.; DU, J.X. A method for Chinese text classification based on apparent semantics and latent aspects. *J. Ambient Intell. Humaniz. Comput.* **2015**, *6*, 473–480. [CrossRef]
8. Li, H.M.; Huang, H.N.; Cao, X.; Qian, J.G. Falcon: A novel Chinese short text classification method. *J. Comput. Commun.* **2018**, *6*, 216–226. [CrossRef]
9. Wang, G.S.; Huang, X.J. Convolutional neural network text classification model based on Word2vec and improved TF-IDF. *J. Chin. Comput. Syst.* **2019**, *40*, 1120–1126.
10. Du, J.C.; Gui, L.; He, Y.L.; Xu, R.F.; Wang, X. Convolution-Based Neural Attention with Applications to Sentiment Classification. *IEEE Access* **2019**, *7*, 27983–27992. [CrossRef]
11. Peng, Y.Q.; Song, C.B.; Yan, Q.; Zhao, X.S.; Wei, M. Research on Chinese text classification based on Hybrid Model of VDCNN and LSTM. *Comput. Eng.* **2018**, *44*, 190–196.
12. Yun, X.P. Research progress and prospect of emergency management based on CNKI and CiteSpace. *China Saf. Sci. J.* **2022**, *32*, 185.
13. Nan, M.Y.; Chen, J. Research Progress, Hotspots and Trends of Land Use under the Background of Ecological Civilization in China: Visual Analysis Based on the CNKI Database. *Sustainability* **2022**, *15*, 249. [CrossRef]
14. Li, X.P.; Zhou, Y. Research on information text extraction and analysis technology based on natural language processing. *Wirel. Internet Technol.* **2023**, *20*, 157–159.

15. Zhu, Y.H. Medical text mining and knowledge extraction based on natural language processing and knowledge graph. *China Comput. Commun.* **2023**, *35*, 1–3.
16. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]
17. Paulsen, D.; Yash, G.; AnHai, D. Sparkly: A Simple yet Surprisingly Strong TF/IDF Blocker for Entity Matching. *Proc. VLDB Endow.* **2023**, *16*, 1507–1519. [CrossRef]
18. Wan, Q.; Xu, X.H.; Han, J. A dimensionality reduction method for large-scale group decision-making using TF-IDF feature similarity and information loss entropy. *Appl. Soft Comput.* **2024**, *150*, 111039. [CrossRef]
19. González, F.; Torres-Ruiz, M.; Rivera-Torruco, G.; Chonona-Hernández, L.; Quintero, R. A Natural-Language-Processing-Based Method for the Clustering and Analysis of Movie Reviews and Classification by Genre. *Mathematics* **2023**, *11*, 4735. [CrossRef]
20. Jitchaijaroen, W.; Keawsawasvong, S.; Wipulanusat, W.; Kumar, D.R.; Jamsawang, P. Machine learning approaches for stability prediction of rectangular tunnels in natural clays based on MLP and RBF neural networks. *Intell. Syst. Appl.* **2024**, *21*, 200329. [CrossRef]
21. Xiang, M.; Zhou, B.T.; Cheng, S.Q.; Liu, S. MCMP-Net: MLP combining max pooling network for sEMG gesture recognition. *Biomed. Signal Process. Control* **2024**, *90*, 105846.
22. Sun, H.L.; Lu, Y.F. A novel approach for solving linear Fredholm integro-differential equations via LS-SVM algorithm. *Appl. Math. Comput.* **2024**, *470*, 128557. [CrossRef]
23. Chen, C.F.; He, Q.X.; Li, Y.Y. Downscaling and merging multiple satellite precipitation products and gauge observations using random forest with the incorporation of spatial autocorrelation. *J. Hydrol.* **2024**, *632*, 130919. [CrossRef]
24. Lauzon, D.; Gloaguen, E. Quantifying uncertainty and improving prospectivity mapping in mineral belts using transfer learning and Random Forest: A case study of copper mineralization in the Superior Craton Province, Quebec, Canada. *Ore Geol. Rev.* **2024**, *166*, 105918. [CrossRef]
25. Li, C.; Managi, S. Mental health and natural land cover: A global analysis based on random forest with geographical consideration. *Sci. Rep.* **2024**, *14*, 2894. [CrossRef] [PubMed]
26. Wang, L.; Li, Z.W.; Zhu, C.D.; Li, Y.J. Research on spam filtering based on NB algorithm. *Transducer Microsyst. Technol.* **2020**, *39*, 46–48.
27. Yuan, L.H.; Li, X.W.; Xu, J. An improved anti-spam filtering method based on bayesian. *Comput. Digit. Eng.* **2020**, *48*, 513–516.