Jurnal Ilmu Siber dan Teknologi Digital (JISTED)
ISSN 2986-7312, Vol 4, No 1, 2025, 15-31 https://doi.org/10.35912/ isted.v4i1.5547

Web-Based Research Article Classification Using The Random Forest Algorithm

Fiqqi Ahludzikri[1*], Riko Herwanto[2], RZ Abdul Aziz[3], Isnandar Agus[4], Suhendro Yusuf Irianto[5]
Institut Informatika dan Bisnis Darmajaya[1,2,3,4,5]
fiqqi.2111010034@mail.darmajaya.ac.id[1], rikoherwanto@darmajaya.ac.id[2], rz_aziz@darmajaya.ac.id[3], isnandaragus@darmajaya.ac.id[4], suhendro@darmajaya.ac.id[5]

Article History:
Received on 24 September 2025
1st Revision on 12 Oktober 2025
2nd Revision on 19 November 2025
Accepted on 20 November 2025

Abstract
Purpose: This study aimed to develop a web-based system that classifies research articles using the Random Forest algorithm to address mismatches between article content and journal scope.
Methodology/approach: The research employed the SDLC Waterfall model, using a dataset of 560 research articles published between 2019 and 2024 from an academic journal publisher across four classification categories. Text preprocessing included case folding, stopword removal, stemming, and tokenization, with TF-IDF applied for feature extraction. The random Forest was trained with 80% of the training data and 20% of the testing data.
Results: The model achieved 91% accuracy with high precision and recall across all categories. The system was successfully implemented as a web-based application, providing instant classification and journal recommendation.
Limitations: The dataset was limited to one publisher, and only Random Forest was applied, which may restrict the generalizability of the findings.
Contribution: This study contributes to the application of machine learning in scholarly publishing by offering a practical solution for editors to streamline article selection and improve efficiency.
Conclusion: Overall, the proposed web-based Random Forest classification system demonstrates strong performance and practicality in automatically classifying research articles, thereby supporting editors and authors in improving the efficiency, accuracy, and consistency of journal article selection.

Keywords: *SDLC, Classification, Random Forest, TF-IDF, Machine Learning*

How to Cite:. Ahludzikri, F., Herwanto, R., RZ , A. A., Agus, I., & Irianto, S. Y. (2025). WEB-BASED RESEARCH ARTICLE CLASSIFICATION USING THE RANDOM FOREST ALGORITHM. *Jurnal Ilmu Siber Dan Teknologi Digital*, *4*(1), 15–31. Retrieved from https://penerbitgoodwood.com/index.php/jisted/article/view/5547

1. Introduction

Scientific publications play a crucial role in the advancement of knowledge, where published articles contribute to the dissemination of information and discussion of the latest research findings. However, amid the growing number of publications, the process of selecting articles for publication in journals often faces significant challenges. One of the main issues is the misalignment between the article topic and the scope of the targeted journal, which frequently slows down the publishing process and results in articles being rejected or requiring multiple revisions. This creates a need for a system that can expedite and streamline the article classification process, allowing articles to be accepted and published more swiftly in appropriate journals.

The main issue that arises in the article selection process is the inefficiency of classifying articles into relevant categories. This process is often carried out manually, which is time-consuming and labor-intensive and can lead to errors in article classification. The gap in this research lies in the lack of application of automated and data-driven methods in scientific article classification. Although some studies have employed automated methods for classification, the use of machine learning algorithms that can manage large volumes of data and provide more accurate and efficient classification results is limited.

Previous studies have demonstrated the potential of machine learning algorithms for text classification. For example, research conducted by Meisty, Anggraeni, and Fatekurohman (Meisty, Anggraeni, & Fatekurohman, 2024) showed that the Random Forest algorithm could be used to classify reviews with a high level of accuracy, while Apriliah, Kurniawan, Baydhowi, and Haryati (Apriliah, Kurniawan, Baydhowi, & Haryati, 2021) applied the same algorithm for disease prediction with excellent results. However, the application of the Random Forest algorithm in the classification of scientific articles for journal publication has not been widely implemented, especially in web-based systems that can facilitate access and improve efficiency in article selection.

This study aimed to develop a web-based scientific article classification system using the Random Forest algorithm. This system is expected to improve the efficiency of the article selection process, allowing for the automatic classification of articles based on the title and abstract, as well as providing journal recommendations relevant to the topics discussed in the article. This research is also expected to accelerate the scientific article publication process and offer a technological solution that benefits both journal publishers and article authors, thereby enhancing the quality and efficiency of scientific publications.

2. Literature Review and Hypothesis Development

2.1 Data Mining

Data Mining is a process used to discover patterns or relationships from large datasets in complex relational databases. This process is a series of stages aimed at extracting valuable and previously unknown information. Such information is obtained by analyzing and recognizing interesting patterns within the existing data in a database (Siregar, Kom, Puspabhuana, Kom, & Kom, 2017). Data Mining is often used to uncover hidden knowledge within large databases and is therefore also referred to as Knowledge Discovery in Databases (Suhirman, 2023).

2.2 Text Preprocessing

Text Preprocessing is the initial stage in Text Mining aimed at transforming unstructured text data into structured data that can be further analyzed. The goal of preprocessing is to produce a term index that effectively represents the documents (Kurniasari, Santoso, & Prahutama, 2021). The stages involved are as follows:

1. Case Folding
   Case folding is the process of standardizing text by converting all uppercase letters in a document to lowercase. This transformation applies only to alphabetic characters from ‘a’ to ‘a' z to 'z', while other characters such as punctuation marks, commas, or spaces, known as delimiters or special characters, can be ignored or removed using Python commands (Rofiqi & Akbar, 2024).

2. Stemming
   Stemming is the process of removing affixes from a word to obtain its root form, such as converting " running" to "run" or "cooking" to "cook." The purpose of stemming is to reduce unnecessary word variations, thereby improving the efficiency of text processing. (Albab & Fawaiq, 2023).

3. Stopword Removal

Stopword removal is the process of selecting important words by removing those considered unhelpful in text processing. Stopwords typically consist of frequently occurring words, such as pronouns, "and," "or," "also," "from," "to," and similar terms. This process was carried out by comparing the words in a document with a list of stopwords, and any matching words were removed. (Lestari & Saepudin, 2021).

4. Tokenizing
   Tokenization is the process of breaking a sentence into smaller units or individual words. This separation is typically performed based on the spaces between words. Additionally, numbers, punctuation marks, and special characters may be removed during this stage, as they are considered irrelevant to the text processing. (Mualfah, Prihatin, & Firdaus, 2023).

2.3 Data labeling
Data labeling is the process of assigning labels or annotations to unlabeled data to help computers recognize patterns. This labeling is crucial in machine learning, as it teaches the model to distinguish between different categories of data, such as image recognition, text processing, and other applications. (Alhaqq, Putra, & Ruldeviyani, 2022).

2.4 TF-IDF
TF-IDF is a measurement method used to determine the importance of words in a document. Term Frequency (TF) indicates how often a word appears in a document, as shown in Formula 1. The Inverse Document Frequency (IDF) measures how rare a word is across all documents, assigning more weight to words that are more unique or less frequently used, as shown in Formula 2. The TF-IDF calculation, as shown in Equation 3, combines these two concepts to generate an accurate weight. (Husain, Sukirman, & SAJIAH, 2024).

$$TF_{d,t} = \frac{\text{Number of times word t appears in the document}}{\text{Total number of word in the document}} \tag{1}$$

$$IDF = \log_2\left(\frac{D}{df}\right) \tag{2}$$

$$W_{d,t} = TF_{d,t} * IDF \tag{3}$$

2.5 Classification
Classification is a method of grouping data based on its characteristics and then predicting which group the new data belong to (Sihombing & Yuliati, 2021). In classification, a model is built using labeled data (supervised learning) to predict the class of new, unseen data (Mestika, Selan, & Qadafi, 2023). An example of classification is the use of an algorithm to categorize fruits for price inspection. This process involves learning patterns from training data so that the algorithm can make accurate predictions when provided with new data (Kristiawan, Somali, & Widjaja, 2020).

Classification typically involves two main steps: training and prediction. During the training phase, the model learns from a dataset containing both inputs and corresponding labels. In the prediction phase, the trained model applies this knowledge to predict the class of new, unlabeled data (Burhanuddin, 2024). Machine learning classification is widely used in various applications, such as image recognition, text analysis, and medical diagnosis (Pramann et al., 2023)

2.6 Random Forest
Random Forest is a popular ensemble learning algorithm in machine learning that combines multiple decision trees to improve accuracy and reduce overfitting. Random Forest works by building several decision trees from random subsets of training data and combining the prediction results from each tree through majority voting (for classification) or averaging (for regression) (Mario & Suryono, 2025). This process is known as bagging (bootstrap aggregating), which helps reduce variance and the risk of overfitting. Figure 1 provides an overview of the Random Forest workflow.

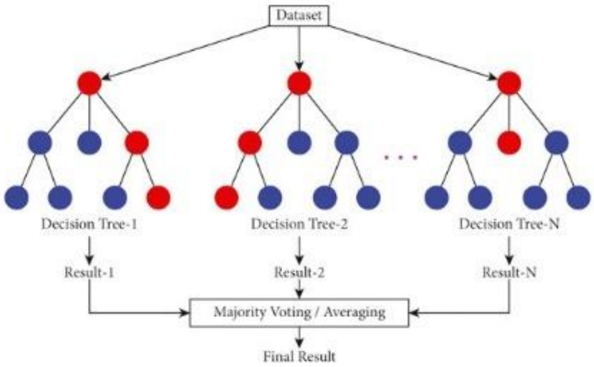

Figure 1. Random Forest Workflow.

In the formation of a Random Forest, each decision tree is built by randomly selecting subsets of features at each split node. This ensures that each tree has high diversity, making the combined predictions from all the trees more stable and accurate. Additionally, Random Forest can handle data with many features without the need for dimensionality reduction, as only randomly selected feature subsets are considered at each split. Another advantage of Random Forest is its ability to measure feature importance, which is useful for further analysis of each feature's contribution to the prediction process (Yulianto, Fanani, Affandy, & Aziz, 2024).

2.6.1 Model Evaluation Techniques

Model evaluation is a crucial step in assessing the performance of machine learning models. A commonly used evaluation tool is the confusion matrix, which displays the performance of a classification model by comparing the model's predictions with the actual labels. The confusion matrix consists of four main elements: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Based on the confusion matrix, several evaluation metrics can be calculated, such as accuracy, precision, recall, and F1-score.

2.6.2 Data visualization

Data visualization is the process of presenting data in graphical or visual formats, such as charts, diagrams, or maps, to facilitate the understanding and analysis of information. Data visualization helps identify patterns, trends, and anomalies within the data that may not be evident through regular numerical analysis. Common examples of data visualization include bar charts, line diagrams, heat maps, and scatter plots (Sudipa et al., 2023).

In scientific research, data visualization is crucial, especially when researchers work with large and complex datasets. Visualization aids researchers in observing the relationships between variables and understanding the results of machine learning models (Perkasa & Rahmatulloh, 2024). With visualization, the prediction results, such as research trends, can be easily understood by journal publishers to assist them in making better decisions regarding relevant publication topics.

2.7 Related Research

a. The study by Ferisa Dwi Alfia Meisty, Dian Anggraeni, and Mohamat Fatekurohman (2024) titled "Comparison of Naïve Bayes Classifier and Random Forest Methods in Predicting Korean Drama Review Ratings" aims to classify Korean drama reviews into categories such as Good, Not Good, or Fair, and to compare the performance of the Naïve Bayes Classifier and Random Forest methods. Data were obtained from the IMDB and underwent preprocessing stages, including text cleaning and labeling, followed by splitting into training and testing datasets. Evaluation was performed using accuracy, precision, recall, and F1-score, showing that Random Forest achieved a higher accuracy of 89% in predicting reviews compared to Naïve Bayes, which had an accuracy of 86%. In predicting ratings, Random Forest also outperformed with an accuracy of 41% compared to Naïve Bayes, which reached

40%. Therefore, it was concluded that Random Forest is more effective in predicting Korean drama review ratings (Meisty et al., 2024).

b. The study conducted by Juwariyem, Sriyanto, Sri Lestari, and Chairani (2024), titled "Prediction of Stunting in Toddlers Using Bagging and Random Forest Algorithms" aimed to improve the accuracy of stunting prediction in toddlers by applying the bagging technique and Random Forest algorithm. This study used a dataset of 10,001 records, with seven attributes as input variables and one class attribute as the prediction target. The Bagging technique was employed to address data imbalance, while Random Forest was applied to build a more accurate prediction model. The results showed that the developed model achieved an accuracy of 91.98%, with a precision of 91.72% for the "yes" class, a recall of 98.84% for the "yes" class, a precision of 93.55% for the "no" class, and a recall of 65.28% for the "no" class. Therefore, this method can be used as a decision-support tool for the prevention of stunting in toddlers (Lestari, 2024).

c. The study conducted by Saman Behrouzi, Zahra Shafaeipour Sarmoor, Khosrow Hajsadeghi, and Kaveh Kavousi (2020) titled "Predicting Scientific Research Trends Based on Link Prediction in Keyword Networks" aims to predict future scientific research trends by analyzing keyword networks in the field of computer science. The two main methods used are topology-based link prediction and machine learning algorithms, such as support vector machines (SVM) and random forests. The results show that the applied methods can predict research trends accurately, particularly within the keyword networks of scientific conferences (Behrouzi, Sarmoor, Hajsadeghi, & Kavousi, 2020).

d. The study conducted by Widya Apriliah, Ilham Kurniawan, Muhamad Baydhowi, and Tri Haryati (2021) titled "Prediction of Early Stage Diabetes Probability Using the Random Forest Classification Algorithm" aims to design a model that can predict the likelihood of developing diabetes at an early stage with high accuracy. This study used three classification algorithms: Random Forest, Support Vector Machine (SVM), and Naïve Bayes, along with a dataset from the UCI Repository. The evaluation results showed that the Random Forest algorithm provided the highest accuracy of 97.88% compared to the other algorithms, making it the recommended method for the early detection of diabetes (Apriliah et al., 2021).

e. The study conducted by Hendri Mahmud Nawawi, Agung Baitul Hikmah, Ali Mustopa, and Ganda Wijaya (2024) titled "Machine Learning Classification Model for Career Placement Accuracy Prediction" aims to predict appropriate career placement based on educational background and work experience data. Five algorithms were tested in this study: Random Forest, Decision Tree, Naive Bayes, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) using the Job Placement dataset from Kaggle. The evaluation results showed that Random Forest performed the best, with an accuracy of 87% and an AUC value of 0.93. The most influential factor in career placement is the "ssc_percentage" or the percentage of high school exam results (Nawawi, Hikmah, Mustopa, & Wijaya, 2024).

Table 1. Related Research

No	Author	Title	Objectives	Methods	Results
1	Ferisa Dwi Alfia Meisty, Dian Anggraeni, Mohamat Fatekuroh man (2024)	Comparison of Naïve Bayes Classifier and Random Forest Methods in Predicting Korean Drama Review Ratings	To classify Korean drama reviews and compare the performance of the Naïve Bayes and Random Forest algorithms	Data collected from IMDB; text preprocessing; evaluation using accuracy, precision,	Random Forest outperformed Naïve Bayes with an accuracy of 89% compared to 86%; Random Forest also achieved better

				recall, and F1-score	rating prediction accuracy (41%)
2	Juwariyem, Sriyanto, Sri Lestari, Chairani (2024)	Prediction of Stunting in Toddlers Using Bagging and Random Forest Algorithms	To predict stunting occurrence in toddlers by improving prediction accuracy using Bagging and Random Forest algorithms	Application of Bagging and Random Forest on a dataset of 10,001 records with 7 attributes and 1 class label	The model achieved an accuracy of 91.98%, with precision (yes) 91.72%, recall (yes) 98.84%, precision (no) 93.55%, and recall (no) 65.28%
3	Saman Behrouzi, Zahra Shafaeipour Sarmoor, Khosrow Hajsadeghi, Kaveh Kavousi (2020)	Predicting Scientific Research Trends Based on Link Prediction in Keyword Networks	To predict scientific research trends through keyword network analysis	Topology-based link prediction and machine learning algorithms (SVM and Random Forest)	The methods successfully predicted research trends with high accuracy
4	Widya Apriliah, Ilham Kurniawan, Muhamad Baydhowi, Tri Haryati (2021)	Prediction of Early-Stage Diabetes Probability Using the Random Forest Classification Algorithm	To predict the likelihood of early-stage diabetes with high accuracy	Dataset from the UCI Repository; Random Forest, SVM, and Naïve Bayes algorithms; accuracy evaluation	Random Forest achieved the highest accuracy of 97.88% and was recommended for early diabetes detection
5	Hendri Mahmud Nawawi, Agung Baitul Hikmah, Ali Mustopa, Ganda Wijaya (2024)	Machine Learning Classification Model for Career Placement Accuracy Prediction	To predict career placement based on educational background and work experience	Job Placement dataset from Kaggle; Random Forest, Decision Tree, Naïve Bayes, KNN, and SVM algorithms; AUC evaluation	Random Forest achieved the best performance with 87% accuracy and an AUC of 0.93; the most influential factor was secondary school exam percentage (ssc_percentage)

2.8 Conceptual Framework

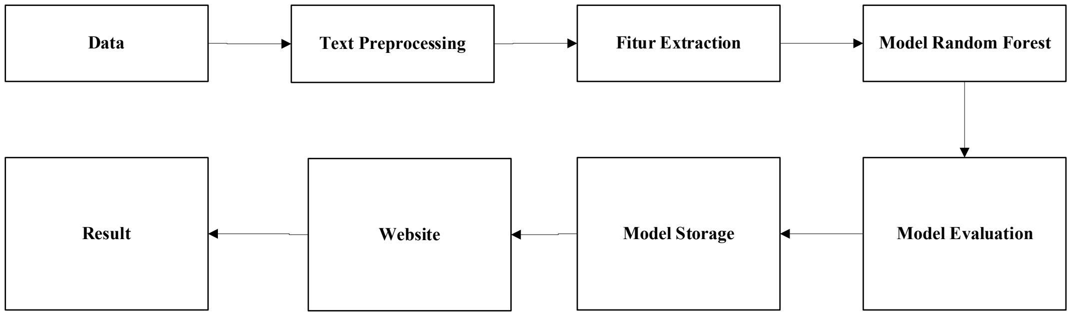

Figure 2. Conceptual Framework

2.8.1 Data
This stage was the initial step of the research. A dataset containing text documents (such as research titles and abstracts) and their categories/labels was collected. This dataset served as the foundation for building the text classification model. The data used must be relevant and representative to ensure that the resulting model performs well.

2.8.2 Text Preprocessing
After collecting the data, the next step was to clean and prepare the text for further processing. Preprocessing includes several stages.

1. Case folding: All text was converted to lowercase to ensure consistency.
2. Removing Non-Alphabet Characters: Eliminating symbols, numbers, or special characters that are not relevant.
3. Tokenizing: Splitting the text into individual words (tokens).
4. Stopword Removal: Removing common words that do not have significant meaning (such as "and," "in," "the").
5. Stemming: Converting words to their root forms (e.g., "writing" to "write").
6. This preprocessing is important for reducing noise and ensuring that the text is ready for feature extraction.

2.8.3 Feature Extraction
After cleaning the text, the next step was to convert the text into a numerical representation that could be processed by a machine learning model. The method used is Term Frequency-Inverse Document Frequency (TF-IDF), which calculates the weight of a word based on its frequency in the document and its uniqueness across all documents. The result is a numerical vector that represents the input text.

2.8.4 Model Random Forest
At this stage, a text classification model was built using the Random Forest algorithm. Random Forest was chosen because of its ability to handle text data and its resistance to overfitting. The dataset was split into two parts: training and testing data. The model was trained using the training data and tested using the testing data to ensure that the model performed well on new data.

2.8.5 Model Evaluation
Once the model is built, its performance is evaluated using several metrics, such as

1. Accuracy: The percentage of correct predictions was calculated.
2. Precision: The proportion of true-positive predictions.
3. Recall: The proportion of actual positives that were correctly predicted.
4. F1-Score: The harmonic mean of precision and recall.

Additionally, a confusion matrix was used to analyze the model's performance across each category. This evaluation helps determine whether the model is sufficiently good or requires improvement.

2.8.6 Model Storage
If the model meets the expected performance criteria, the model, vectorizer (TF-IDF), and label encoder are saved using either pickle or joblib. This storage allows the model and its supporting components to be reused without the need for retraining. These files were subsequently integrated into a web application.

2.8.7 Website
To enable the practical use of the model, a web application was built using the Flask framework. The application is designed with a simple user interface, where users can input (titles and abstracts) via a form. The web application processes the text using the stored model and displays the predicted category results.

2.8.8 Result
In addition to displaying the prediction results, the web application can also provide journal or category recommendations based on the prediction outcomes. For example, if the text is predicted to fall into the "Finance and Business" category, the application can recommend journals related to this field. This feature adds value to the proposed application.

3. Research Methodology

3.1 Research Flow Diagram

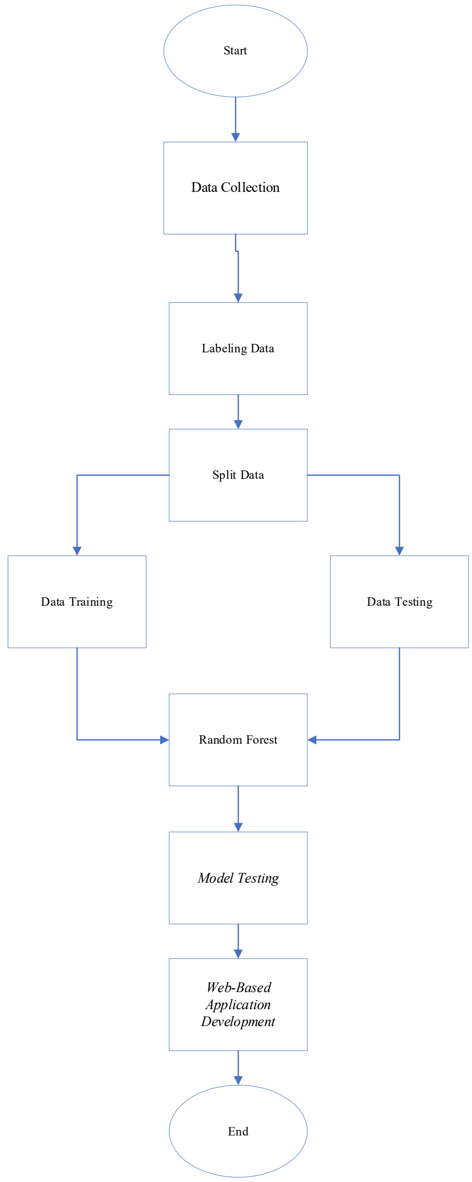

Figure 3. Research Flow Diagram

3.1.1 Data Collection

This study used a dataset of scientific articles published by a journal publisher between 2019-2024, with a total of 560 data points. The dataset includes attributes such as Author, Title, Category, Journals, Abstract, and Year. The collected data were stored in an Excel file in CSV format.

Table 2. Number of data points in each category.

No	Category	Number
1	Human Resource Management and Leadership	140
2	Finance and Business	140
3	Community Empowerment and Business	140
4	Education and Learning	140

Total	560

Table 2 presents the data distribution based on the research categories used in this study. Overall, 560 data points were evenly distributed across four main categories, with 140 data points in each category. The first category, Human Resource Management and Leadership, covers research on human resource management strategies, leadership, and organizational aspects of the workplace. The second category, Finance and Business, includes data related to financial aspects, investments, and company management. The third category, Community Empowerment and Business, includes research on business development strategies, entrepreneurship, and efforts toward community empowerment across various sectors. Finally, the Education and Learning category covers studies on teaching methods, educational innovation, and the learning process. With a balanced data distribution across these four categories, the research can be conducted proportionally, ensuring that the analysis is more representative of the various fields being studied.

3.1.2 Data Preprocessing

3.1.2.1 Case Folding
All text in columns, such as titles and abstracts, was converted to lowercase to eliminate inconsistencies caused by the use of uppercase letters. Below is a comparison of the document before and after case folding, as shown in Section 4.3.

Table 1. Example of case folding results for titles.

Before	After
The Role of Corporate Social Responsibility on the Performance of Indonesian Banking Corporation	the role of corporate social responsibility on the performance of indonesian banking corporation

3.1.2.2 Stopword Removal
Common words that do not provide significant meaning (such as "and," "which," "in") are removed to improve the relevance of the text features. Table 4 shows a comparison of the document before and after stop word removal.

Table 4. Example of stopword removal results

Before	After
The Role of Corporate Social Responsibility on the Performance of Indonesian Banking Corporation	role corporate social responsibility performance indonesian banking corporation

3.1.2.3 Stemming
Words are converted to their root form to reduce word variation, for example, "banking" becomes "bank." Table 5 shows a comparison of the documents before and after stemming.

Table 5. Example of stemming results

Before	After
role corporate social responsibility performance indonesian banking corporation	role corporate social responsibility perform indonesia bank corporation

3.1.2.4 Tokenizing/Tokenisasi
Text is broken down into individual words or tokens; for example, "scientific article" becomes ["scientific," "article"]. Table 6 shows a comparison of the documents before and after tokenization.

Table 6. Example of tokenization results.

Before	After
role corporate social responsibility perform indonesia bank corporation	['role', 'corporate', 'social', 'responsibility', 'perform', 'indonesia', 'bank', 'corporation']

3.1.3 Labeling Data

Data are labeled based on specific categories, such as research topics, citation count, or potential reader interest. This process ensured that the data were ready for use as input for the classification model.

3.1.4 Data Split

The dataset was split into training and testing data with a ratio of 80:20. The training data were used to train the Random Forest model, whereas the testing data were used to evaluate the model's accuracy.

3.1.5 Random Forest Implementation

The application of the Random Forest algorithm includes:

1. Bootstrap sampling: Random samples are created from the dataset.
2. Random feature selection: Attributes are randomly selected to form decision trees.
3. Model Training: Growing many decision trees without pruning.
4. Majority Vote: Classification results are determined based on the majority vote from all the trees formed.

3.1.6 Model Testing

The model was tested using a Confusion Matrix to obtain the evaluation results. The metrics used included the following:

a) Accuracy: The correctness of the model is evaluated.
   Formula: $\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$

b) Precision: The accuracy of classifying a specific class.
   Formula: $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$

c) Recall: The ability of the model to find all relevant data.
   Formula: $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$

d) F1-Score: The harmonic balance between Precision and Recall.
   Formula: $\frac{2\ \mathrm{X\ TP}}{2\ \mathrm{X\ TP+FP+FN}}$

3.1.7 Web-Based Application Development

The Random Forest model was integrated into a web-based application. The main function of this application is to classify relevant topics and predict research publications that are likely to interest readers.

4. Results and Discussion

4.1 Model Performance Evaluation

The Random Forest model used in this study was evaluated based on several metrics, such as precision, recall, f1-score, accuracy, and Root Mean Square Error (RMSE). In this test, an 80:20 ratio was used, that is, 80% of the data (448 data points) were used for training and 20% (112 data points) for testing. Each category has a balanced number of data points, with 140 articles in each category: Human Resource Management and Leadership, Finance and Business, Community Empowerment and Business, and Education and Learning. The classification results of the test using the confusion matrix are shown in Figure 4, and the confusion matrix display is shown in Figure 5.

```
Classification Report:
                                              precision    recall  f1-score   \
Keuangan dan Perusahaan                            0.93      0.89      0.91
Manajemen Sumber Daya Manusia dan Kepemimpinan     0.87      0.93      0.90
Pemberdayaan Masyarakat dan Bisnis                 0.90      0.96      0.93
Pendidikan dan Pembelajaran                        0.96      0.86      0.91
accuracy                                           0.91      0.91      0.91
macro avg                                          0.91      0.91      0.91
weighted avg                                       0.91      0.91      0.91

                                              support
Keuangan dan Perusahaan                         28.00
Manajemen Sumber Daya Manusia dan Kepemimpinan  28.00
Pemberdayaan Masyarakat dan Bisnis              28.00
Pendidikan dan Pembelajaran                     28.00
accuracy                                         0.91
macro avg                                      112.00
weighted avg                                   112.00

Akurasi Keseluruhan: 0.91

Root Mean Square Error (RMSE): 0.43

R² Score: 0.85
```

Figure 4. Classificaion Report

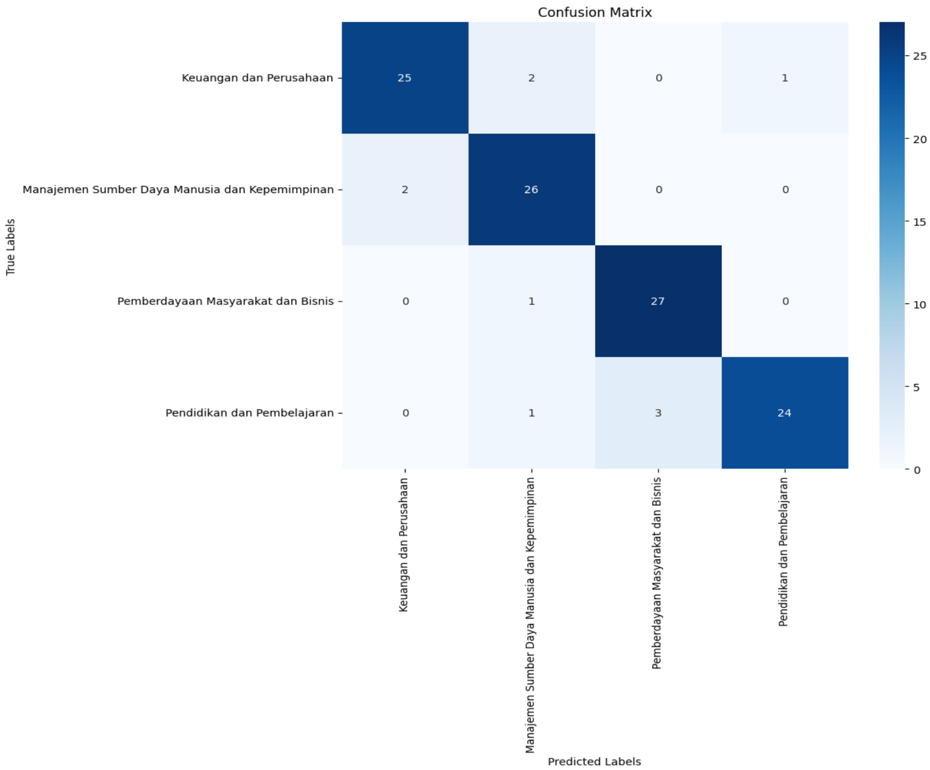

Figure 5. Confusion Matrix

$$\text{Accuracy} = \frac{25+26+27+24}{112} \times 100\% = 91\%$$

Table 7. Confusion Matrix for each category

No	Category	Confusion Matrix				Testing
		TP	FP	FN	TN	
1	Finance and Company	25	2	3	82	28
2	Human Resources and Leadership	26	4	2	80	28
3	Community Empowerment and Business	27	3	1	81	28
4	Education and Learning	24	1	4	83	28

Table 7 shows the performance of the classification model based on the confusion matrix for each category of journal articles. Each category had the same number of test data (112 articles), which were used to evaluate the classification results. Overall, the model showed good performance in classifying articles into the four main categories, with relatively high True Positive (TP) values and varying False Positive (FP) and False Negative (FN) values in each category.

Table 8. Model Performance Evaluation Results

No.	Category	Precisio n	Recall	f1-Score	TPR	FPR
1	Human Resources and Leadership	93%	89%	91%	93%	2%
2	Finance and Company	87%	93%	90%	79%	10%
3	Community Empowerment and Business	90%	96%	93%	82%	5%
4	Education and Learning	96%	86%	91%	89%	2%
Average		**91%**	**91%**	**91%**		

Table 8 presents the model performance evaluation based on the Precision, Recall, F1-Score, True Positive Rate (TPR), and False Positive Rate (FPR) for each journal article category.

4.2 Discussion

4.2.1 Random Forest Model Performance

Based on the evaluation results, the Random Forest model performed well in classifying scientific article categories. This report includes the precision, recall, and f1-score metrics, which were used to measure the model's performance in classifying articles. For the Finance and Business category, the model achieved a precision of 0.93, recall of 0.89, and f1-score of 0.91, indicating that the predictions were quite accurate, although there were some misclassifications. Meanwhile, for the Human Resource Management and Leadership category, the model had a precision of 0.87, recall of 0.93, and f1-score of 0.90, indicating that the model was more likely to recognize data well (high recall), although the precision was slightly lower. For the Community Empowerment and Business category, the model showed excellent performance, with a precision of 0.90, recall of 0.96, and f1-score of 0.93, indicating the model's strong ability to identify articles in this category. In the Education and Learning category, the model recorded a precision of 0.96, recall of 0.86, and f1-score of 0.91, which showed that the model was fairly accurate but still had some chance of misclassifying articles into this category. Overall, the model achieved an accuracy of 91%, indicating good performance in classifying research articles.

4.2.2 Confusion Matrix Analysis

The confusion matrix (Table 7) shows the performance of the classification model in predicting the article categories. On the axes of True Labels (actual labels) and Predicted Labels, there are four categories: Finance and Business, Human Resource Management and Leadership, Community Empowerment and Business, and Education and Learning.

From this matrix, it can be observed that the model correctly classified most articles. For example, in the Finance and Business category, there were 25 correct predictions, while two articles were misclassified as Human Resource Management and Leadership, and one article was misclassified as Education and Learning. Similarly, in the Human Resource Management and Leadership category, the model successfully classified 26 articles correctly, but two articles were misclassified as Finance and Business.

In the Community Empowerment and Business category, the model had 27 correct predictions, with only one misclassification as Human Resource Management and Leadership. Meanwhile, in the Education and Learning category, the model correctly classified 24 articles but made errors with three articles, which were misclassified as Community Empowerment and Business, and one article misclassified as Human Resource Management and Leadership.

Overall, this confusion matrix indicates that the model has a high accuracy rate but still has some misclassification errors, particularly between categories with similar contexts.

4.2.3 Website Testing Results

Table 9. Testing Results

No	Title	Predicted Category	Journal Recomendation	Accuracy (%)
1	Credit risk management: An imperative for profitability of Centenary Bank Kabale Branch	Human Resource Management and Leadership	Annals of Human Resource Management Research	82.00%
2	Analysis of Factors Affecting Company Capital Structure	Finance and Business	International Journal of Financial, Accounting, and Management	88.00%
3	The Influence of Good Corporate Governance on Company Financial Performance	Finance and Business	International Journal of Financial, Accounting, and Management	76.00%
4	The Role of MSMEs in Community Economic Empowerment	Community Empowerment and Business	Yumary: Jurnal Pengabdian kepada Masyarakat	61.00%
5	Effectiveness of Using Project-Based Learning Models in Improving	Education and Learning	Journal of Social, Humanity, and Education	76.00%
6	The Impact of Transformational Leadership on Employee Performance in the Manufacturing Industry	Human Resource Management and Leadership	Annals of Human Resource Management Research	45.00%
7	Developing a Human Resource Competency Model to Enhance Company Competitiveness	Human Resource Management and Leadership	Annals of Human Resource Management Research	42.00%
8	The Role of MSMEs in Community Economic Empowerment	Community Empowerment and Business	Yumary: Jurnal Pengabdian kepada Masyarakat	56.00%
9	Effectiveness of Using Project-Based Learning Models in Improving Student Learning Outcomes	Education and Learning	Journal of Social, Humanity, and Education	73.00%
10	The Influence of Organizational Culture on Employee Engagement and Retention in the Tech Industry	Human Resource Management and Leadership	Annals of Human Resource Management Research	63.00%

The testing results with 10 new data points show that the model performs well in classifying scientific articles into the appropriate categories. The model achieved high accuracy in several categories, such as Finance and Business, with the highest accuracy of 88%, indicating its ability to effectively recognize patterns in articles related to finance and business management. Additionally, articles in the Education and Learning category also showed solid results, with accuracies of 76% and 73%, indicating that the model can accurately identify topics related to education.

Meanwhile, the Human Resource Management and Leadership and Community Empowerment and Business categories still have areas that could be improved, with accuracies ranging from 42%–82%. Nevertheless, these results suggest that the model is still capable of providing useful predictions, especially for articles with clearer and more specific language patterns. Overall, this classification model showed promising results and has the potential for further refinement. With an increase in the size and variation of the dataset and optimization of the model parameters, the accuracy can be further enhanced, allowing the system to better support an automatic and efficient article selection process.

5. Conclusion

Based on the research results regarding scientific-article classification using the Random Forest algorithm, several conclusions can be drawn as follows:

1. Development of the Prediction Model with the Random Forest Algorithm
   In this study, a scientific article classification model was developed using the Random Forest algorithm with an accuracy of 91%. The model can classify articles into appropriate categories based on their titles and abstracts, assisting in the scientific publication selection process.

2. Evaluation of Random Forest Algorithm Performance
   The evaluation results show that the Random Forest algorithm performs well in classifying articles based on predetermined categories. The high precision, recall, and f1-score values in each category indicate that this model can be effectively used to assist in the classification of scientific articles.

3. Development of the Web-Based Application
   The developed web-based application was successfully implemented and tested to facilitate the automatic classification of articles. This system allows users to input the title and abstract of articles and obtain classification results instantly, thereby improving the efficiency of the article-selection process.

Limitations and Future Research

Although this study has shown good results, there are several suggestions for further development.

1. Use of a Larger Dataset
   Future studies could use a larger and more diverse dataset from various journal publishers to make the model more accurate in classifying the articles.

2. Testing with Other Algorithms
   It is recommended to try other algorithms besides Random Forest to compare their performance in classifying scientific articles.

3. Development of Application Features
   The web application can be further developed by adding more accurate journal recommendation features and research trend analysis to assist editors in decision making.

REFERENCES

Albab, M. U., & Fawaiq, M. N. (2023). Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic. *Jurnal Transformatika, 20*(2), 1-12. doi:https://doi.org/10.26623/transformatika.v20i2.5374

Alhaqq, R. I., Putra, I. M. K., & Ruldeviyani, Y. (2022). Analisis Sentimen terhadap Penggunaan Aplikasi MySAPK BKN di Google Play Store. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 11*(2). doi:https://doi.org/10.22146/jnteti.v11i2.3528

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi kemungkinan diabetes pada tahap awal menggunakan algoritma klasifikasi Random Forest. *Sistemasi: Jurnal Sistem Informasi, 10*(1), 163-171.

Behrouzi, S., Sarmoor, Z. S., Hajsadeghi, K., & Kavousi, K. (2020). Predicting scientific research trends based on link prediction in keyword networks. *Journal of Informetrics, 14*(4), 101079. doi:https://doi.org/10.1016/j.joi.2020.101079

Burhanuddin, R. R. (2024). KLASIFIKASI PENYAKIT PADI MELALUI CITRA DAUN MENGGUNAKAN METODE NAIVE BAYES. *Jurnal Informatika dan Teknik Elektro Terapan, 12*(2). doi:https://doi.org/10.23960/jitet.v12i2.4012

Husain, N. P., Sukirman, S., & SAJIAH, S. (2024). Analisis Sentimen Ulasan Pengguna Tiktok pada Google Play Store Berbasis TF-IDF dan Support Vector Machine. *Journal of System and Computer Engineering (JSCE), 5*(1), 91-102. doi:https://doi.org/10.61628/jsce.v5i1.1105

Kristiawan, K., Somali, D. D., & Widjaja, A. (2020). Deteksi Buah Menggunakan Supervised Learning dan Ekstraksi Fitur untuk Pemeriksa Harga. *Jurnal Teknik Informatika dan Sistem Informasi, 6*(3). doi:https://doi.org/10.28932/jutisi.v6i3.3029

Kurniasari, R., Santoso, R., & Prahutama, A. (2021). Analisis Kecenderungan Laporan Masyarakat Pada “LAPORGUB..!” Provinsi Jawa Tengah Menggunakan Text Mining Dengan Fuzzy C-Means Clustering. *Jurnal Gaussian, 10*(4), 544-553. doi:https://doi.org/10.14710/j.gauss.10.4.544-553

Lestari, S. (2024). Prediction of Stunting in Toddlers Using Bagging and Random Forest Algorithms. *Sinkron: jurnal dan penelitian teknik informatika, 8*(2), 947-955. doi:https://doi.org/10.33395/sinkron.v8i2.13448

Lestari, S., & Saepudin, S. (2021). *Analisis sentimen vaksin sinovac pada twitter menggunakan algoritma Naive Bayes.* Paper presented at the Prosiding Seminar Nasional Sistem Informasi dan Manajemen Informatika Universitas Nusa Putra.

Mario, C., & Suryono, R. R. (2025). Public Sentiment Analysis on Dirty Vote Movie on YouTube using Random Forest and Naïve Bayes. *INOVTEK Polbeng-Seri Informatika, 10*(1), 111-122. doi:https://doi.org/10.35314/ev9j2g33

Meisty, F. D. A., Anggraeni, D., & Fatekurohman, M. (2024). Perbandingan Metode Naïve Bayes Classifier dengan Metode Random Forest pada Prediksi Rating Review Drama Korea. *ESTIMASI: Journal of Statistics and Its Application*, 84-95. doi:https://doi.org/10.20956/ejsa.v5i1.26942

Mestika, J. C., Selan, M. O., & Qadafi, M. I. (2023). Menjelajahi Teknik-Teknik Supervised Learning untuk Pemodelan Prediktif Menggunakan Python. *Buletin Ilmiah Ilmu Komputer dan Multimedia (BIIKMA), 1*(1), 216-219.

Mualfah, D., Prihatin, A., & Firdaus, R. (2023). Analisis Sentimen Masyarakat Terhadap Kasus Pembobolan Data Nasabah Bank BSI Pada Twitter Menggunakan Metode Random Forest Dan Naïve Bayes. *JURNAL FASILKOM, 13*(3), 614-620. doi:https://doi.org/10.37859/jf.v13i3.6478

Nawawi, H. M., Hikmah, A. B., Mustopa, A., & Wijaya, G. (2024). Model Klasifikasi Machine Learning untuk Prediksi Ketepatan Penempatan Karir. *Jurnal Saintekom: Sains, Teknologi, Komputer dan Manajemen, 14*(1), 13-25. doi:https://doi.org/10.33020/saintekom.v14i1.512

Perkasa, M. A. P., & Rahmatulloh, A. (2024). Perbandingan Skor Prediksi dalam Pembelajaran Mesin pada Keselamatan Penumpang Kapal: Comparison of Prediction Scores in Machine Learning on Ship Passenger Safety. *Jurnal Mahasiswa Teknik Informatika, 3*(1), 60-74. doi:https://doi.org/10.35473/jamastika.v3i1.2739

Pramann, S., Ajanakitti, K., Pohma, H., Wutichat, S., Tanphanich, T., & Pichainarongk, S. (2023). *Factors Affecting Product Support to Promote Ideas and Improve Life through Movement Workshop Project.* Paper presented at the ACPES 2022: Proceedings of the 8th ACPES (ASEAN Council of Physical Education and Sport) International Conference, ACPES 2022, October 28th–30th, 2022, Medan, North Sumatera, Indonesia.

Rofiqi, L., & Akbar, M. (2024). Analisis Sentimen Terkait RUU Perampasan Aset dengan Support Vector Machine. *JEKIN-Jurnal Teknik Informatika, 4*(3), 529-538. doi:https://doi.org/10.58794/jekin.v4i3.824

Sihombing, P. R., & Yuliati, I. F. (2021). Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 20*(2), 417-426. doi:https://doi.org/10.30812/matrik.v20i2.1174

Siregar, A. M., Kom, S., Puspabhuana, M. K. D. A., Kom, S., & Kom, M. (2017). *Data Mining: Pengolahan Data Menjadi Informasi dengan RapidMiner*: CV Kekata Group.

Sudipa, I. G. I., Sarasvananda, I. B. G., Prayitno, H., Putra, I. N. T. A., Darmawan, R., & WP, D. A. (2023). *Teknik Visualisasi Data*: PT. Sonpedia Publishing Indonesia.

Suhirman, S. (2023). Data Warehouse dan Mining. *GAES-PACE Book Publisher*, 14-25.

Yulianto, S. P. R., Fanani, A. Z., Affandy, A., & Aziz, M. I. (2024). Analisis Metode Smoote pada Klasifikasi Penyakit Jantung Berbasis Random Forest Tree. *JURNAL MEDIA INFORMATIKA BUDIDARMA, 8*(3), 1460-1471. doi:https://doi.org/10.30865/mib.v8i3.7712