Jurnal Computech & Bisnis, Vol. 19, No. 2, December 2025, 105-113
ISSN (print): 1978-9629, ISSN (online): 2442-4943

Digital Revolution: Using the Naive Bayes Algorithm for Automatic Classification of Archives in a Public Institution in Bandung City

Haris Supriatna[1], Egi Badar Sambani[2], Ray Jati Kautsar[3]
STMIK Mardira Indonesia, Bandung [1,2,3]
haris.supriatna@stmik-mi.ac.id[1], egi.badars@stmik-mi.ac.id[2] , rayjati86@gmail.com[3]

Abstract

The application of the Naive Bayes algorithm in an automatic digital archive classification system at a public institution in Bandung City is covered in this study. The institution's biggest problem is the growing volume of records, which makes it challenging to manage and find archives rapidly. To automatically categorize digital archives into predefined categories, a system was developed. The Naive Bayes algorithm, which provides excellent accuracy, simplicity, and efficiency in processing text data, is the technique employed. Digital documents that have undergone text-preprocessing steps, including tokenization, stopword removal, and stemming, prior to categorization make up the archive data. This system uses MySQL, PHP, and the Laravel framework. According to test results, the system can satisfactorily classify archives while facilitating user location by category. As a result, this system can improve the institution's digital archive management effectiveness and efficiency.

Keywords : *Naive Bayes, Classification, Digital Archives, Information System*

INTRODUCTION

The number of digital archives in government institutions, including a public institution in Bandung City, has increased dramatically due to rapid growth in information technology. This institution still manually groups digital archives, which causes several problems, including lengthy processing times, potential errors in archive location, and reduced productivity. Staff interviews reveal that, because there is no automatic classification mechanism, searching for archives can take a significant amount of time.

An automatic classification system is required to solve this problem. Since the Naive Bayes algorithm is well-known for its simplicity and efficiency in text classification, it is a viable solution. (Zhang, P., Ma, Z., Ren, Z., Wang, H., Zhang, C., Wan, Q., & Sun, 2024; Zhao, 2023) To close the gap left by earlier research, which has primarily been conducted at the laboratory scale, this study aims to apply the Naive Bayes algorithm to an automatic classification system in a real-world setting at a public institution in Bandung City.

The rapid development of information technology has led to a significant increase in the number of digital archives within government entities, including a public institution in Bandung City. The way organizations handle, store, and access information has changed as a result of the explosion in digital documentation. Nonetheless, many organizations continue to classify and organize these archives using conventional techniques. Manual categorization of digital archives is still standard at one public institution in Bandung, posing severe operational difficulties. The sheer volume of documents frequently overwhelms staff workers, making it more challenging to maintain accessibility and structure.

Managing digital archives manually leads to several problems, such as long processing times and potential errors in record placement.

(Ernawati, S., Wati, R., Nuris, N., Marita, L. S., & Yulia, 2020; Wang, 2022) Because employees must spend too much time searching for specific documents, these inefficiencies can seriously impair productivity. Employees have complained in interviews about the sluggish search process, which can take a long time because there is no automatic classification system in place. This circumstance underscores the urgent need for a more dependable and effective way to maintain digital archives.

These issues could be significantly reduced by implementing an automated classification system. Institutions can automate the categorization process, improve document retrieval accuracy, and ultimately free up staff to concentrate on more important activities by using algorithms such as Naive Bayes. An excellent option for organizations seeking to enhance their archive management systems, the Naive Bayes algorithm is known for its simplicity and effectiveness in text classification.

Even though automatic categorization systems have shown great promise, most prior research has been conducted in controlled laboratory settings, leaving a gap in real-world applications. By using the Naive Bayes algorithm in an automatic classification system in a public institution in Bandung City, this study seeks to close that gap. By doing this, it aims to demonstrate how well the algorithm improves the administration of digital archives, opening the door to increased productivity and operational efficiency in government agencies.

The goal of this study is to assess the accuracy of the Naive Bayes algorithm in classifying digital documents and to develop an automatic classification system for digital archives using the Naive Bayes algorithm in one of the public institutions in Bandung City. This will reduce reliance on manual processes prone to error.

The Rise of Digital Archives and the Digital Revolution

How information is managed and accessed has changed dramatically as a result of the digital revolution, especially in the public sector. Governmental organizations around the world, including those in Indonesia, have embraced digital systems to preserve and manage archives as information technology has advanced. The volume of digital archives has increased, calling for effective management and retrieval solutions. Numerous studies show that digitizing archives can lower operating costs, increase openness in public administration, and improve accessibility (Balaji, V. R., Suganthi, S. T., Rajadevi, R., Kumar, V. K., Balaji, B. S., & Pandiyan, 2020; Golub, K., Hagelbäck, J., & Ardö, 2020)

Archive Management's Automatic Classification

In the management of digital archives, automatic classification is an essential process, particularly as data volumes proliferate. This classification process makes it easier to search for and categorize documents into predetermined categories. Several techniques, such as Support Vector Machines (SVM), Decision Trees, and Neural Networks, have been investigated in the past for this classification job. However, because of its ease of use and efficiency in analyzing text data, the Naive Bayes method has become increasingly popular (Viet, T. N., Le Minh, H.,

Hieu, L. C., & Anh, 2021; Winarti, T., Indriyawati, H., Vydia, V., & Christanto, 2021).

Theory and Uses of the Naive Bayes Algorithm

A popular probability-based technique for text classification is the Naive Bayes algorithm. This algorithm assumes that all characteristics are independent, which often yields good results even though this is rarely the practice case. Naive Bayes can expedite the search and grouping operations in the context of archive management by classifying documents based on pre-existing keywords. Naive Bayes may achieve high accuracy in a variety of classification tasks, including document management, according to research by (Abdulameer, A. G., Hammood, A. S., Abdulwahed, F. M., & Ayyash, 2025; Rezaeian, N., & Novikova, 2020).

Applying Naive Bayes in Practical Environments

Although much research has shown that the Naive Bayes algorithm performs well across a variety of situations, it still needs to be tested in real-world environments, such as government buildings. There is currently little research on the use of Naive Bayes in public institutions, particularly in developing nations like Indonesia. At the same time, several studies have successfully implemented it in commercial document classification systems. This study aims to advance our understanding of digital archive management and to offer workable solutions to the problems government agencies confront by using Naive Bayes in an automated classification system at a public institution in Bandung.

METHOD

Because of its methodical and structured approach, this study uses the Object-Oriented Analysis and Design (OOAD) development technique in software engineering. To understand the fundamental theories of text classification, digital archive management, and the Naive Bayes algorithm, the study begins with a literature review. Next, training data is gathered from digital documents supplied by a Bandung City public institution. The next step in the research is data preprocessing, which entails cleaning the text data using techniques including tokenization, stopword removal, and stemming. The model is trained using the Naive Bayes technique to identify keyword patterns in documents and forecast their categories. The Laravel framework and MySQL database are then used to create a web application prototype (an upload interface and a categorization procedure). Lastly, the study assesses the model's performance using metrics such as accuracy and verifies the system's functionality.

RESULTS AND DISCUSSION

SIKLARAS (Digital Archive Classification System) is a web application developed as the outcome of this project. UML modeling was used to specify the structure and functions of this system.

System Design (UML)

The following primary UML diagrams are used to explain the system design:

Because of its methodical and structured approach, this study uses an object-oriented analysis and Design (OOAD) development

technique in software engineering. To understand the fundamental theories of text classification, digital archive management, and the Naive Bayes algorithm, the study begins with a literature review. The researchers then gather training data from digital documents that are made available by a public entity in Bandung City. The next step in the research is data preprocessing, which entails cleaning the text data using techniques including tokenization, stopword removal, and stemming. The model is trained using the Naive Bayes technique to identify keyword patterns in documents and forecast their categories. The researchers then use the Laravel framework and a MySQL database for the upload interface and classification procedure to create a prototype web application. Lastly, the study assesses the model's performance using metrics such as accuracy and verifies the system's functionality.

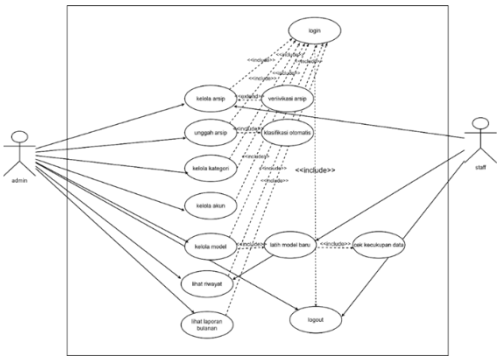

Figure 1. Use Case Diagram

Activity Diagram: This diagram shows the system's workflow procedures. The login flow, for instance, shows how the system verifies user credentials and routes users to the appropriate dashboard based on their roles.

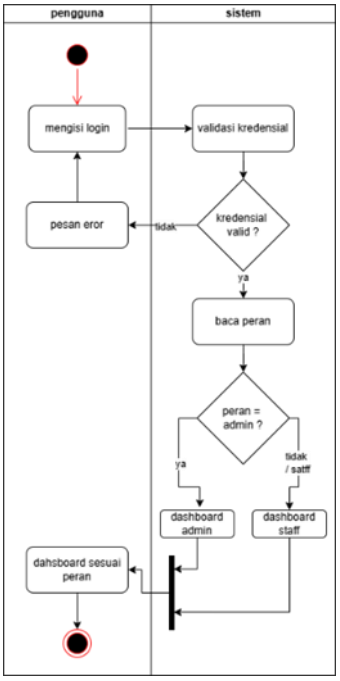

Figure 2. Activity Diagram Login Per Role

The primary function of the system is the classification of digital archives, which is depicted in Figure 3. Document uploading initiates the process, followed by preprocessing and Naive Bayes-based category prediction.

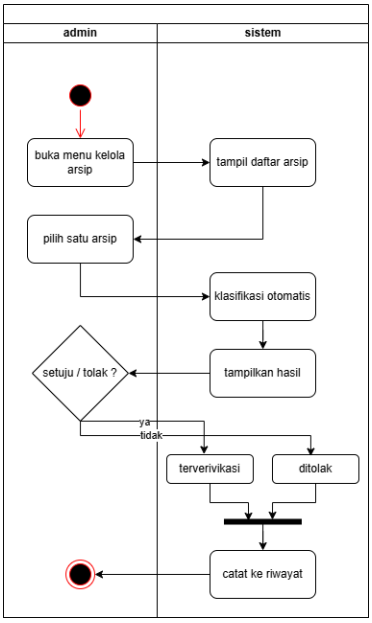

Figure 3. Activity Diagram for Managing Archives (Admin)

4) Class Diagram: Specifies the data structure of the system. Major classes, including User, Archive, Category, Classification, and ClassificationModel, are included, along with their characteristics and relationships.

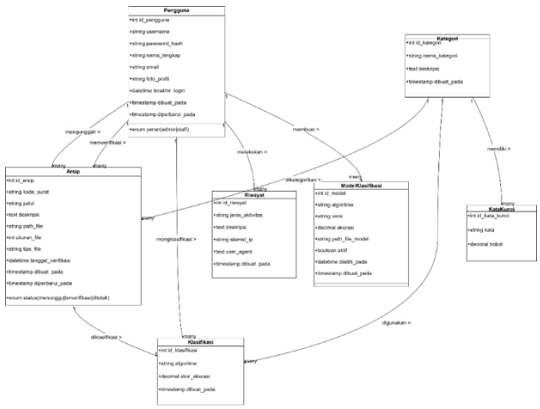

Figure 4. Class Diagram

System Functionality

Class Diagram: Specifies the data structure of the system. Major classes, including User, Archive, Category, Classification, and ClassificationModel, are included, along with their characteristics and relationships.

1) Admin: System data, including total archives, most recent categories, and system correctness, is shown on the admin dashboard page. The administrator can manage users, verify staff-uploaded archives, manage archives (CRUD), and manage classification models (including training new models).

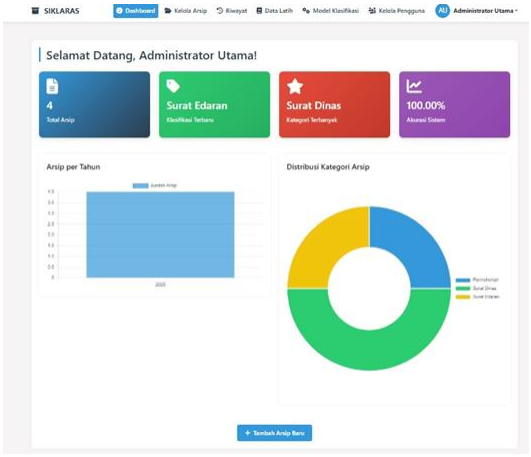

Figure 5. Admin Dashboard Page

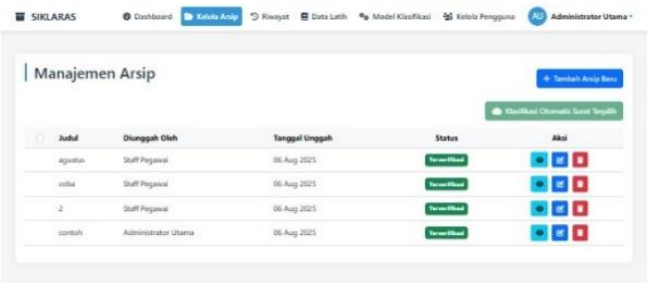

Figure 6. Manage Archives Page

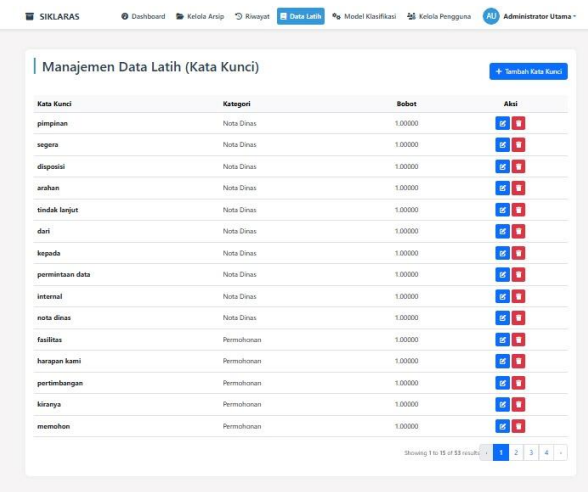

Figure 7. Training Data Page

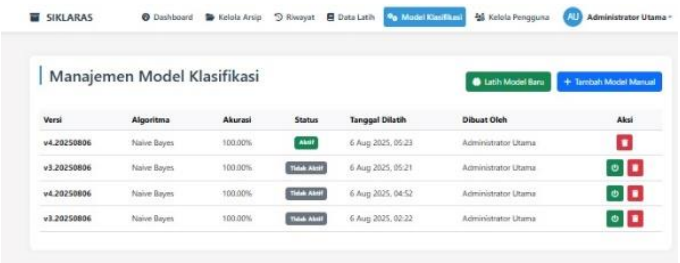

Figure 8. Train Model Page

System Test Results

The Black-Box Testing approach is used to test functionality. Without considering internal procedures, this testing focuses on assessing the system's functionality based on its inputs and outputs. According to the test results, every test scenario—including the classification process, CRUD for Archives, and login—was executed correctly and yielded the expected results. Additionally, the system demonstrated its ability to handle faulty inputs by providing unambiguous error messages and rejecting prohibited file formats.

Table 1. Black Box Testing

No	Tested Features	Testing Scenario	Expected results	Test Results
1	User Login	Login with valid and invalid credentials	The system successfully logs in to the dashboard if the data is correct, and displays an error message if it is incorrect.	Succeed
2	Archive Management (CRUD)	Add, modify, delete, and search archives	The system can add, change, delete, and display archives according to user input.	Succeed
3	Archive File Validation	Uploading files of this type is not permitted.	The system rejects the upload and displays a validation error message.	Succeed

4	Archive Search	Search with relevant and irrelevant keywords	The system displays results according to keywords or the message "Data not found"	Succeed
5	Model Management	Training and activation of new models	A new model can be trained and activated as the primary model	Succeed
6	Automatic Classification	Single and bulk file classification	The system is able to classify archives automatically according to the correct category.	Succeed

Algorithm Accuracy Test Results

Every time the administrator trains a new model, the system automatically evaluates the Naive Bayes algorithm's accuracy. The system splits the confirmed archive dataset into training (80%) and testing (20%) data. By contrasting the actual categories of the Testing Data with the model's projected outcomes, accuracy is determined. Tests with the supplied dataset showed that the system achieved 75% accuracy. This outcome shows that three of the four test papers were accurately classified by the model. As a starting point, 75% accuracy is considered relatively good. The limited amount of training data available and the lexical similarities between document categories are two factors that may contribute to prediction errors.

CONCLUSION

The SIKLARAS system effectively applies the Naive Bayes algorithm to automatically classify digital archives at the Kesbangpol of Bandung City, according to the research findings. The outcomes of the accuracy and functionality testing indicate that the research goal of creating an automated system has been achieved. This technology can reduce errors in the human classification process and increase productivity.

Suggestion

Optical Character Recognition (OCR) technology should be added in the future to allow the system to read and categorize scanned documents using images. Furthermore, evaluating stability and improving model correctness will be facilitated by testing the system with a much larger, more varied dataset. The system's capabilities will be further strengthened by adding a more sophisticated search function, including semantic search, and by improving data security through file encryption.

REFERENCES

Abdulameer, A. G., Hammood, A. S., Abdulwahed, F. M., & Ayyash, A. A. (2025). Naïve Bayes algorithm for timely fault diagnosis in helical gear transmissions using vibration signal analysis. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *19*(5), 3695–3706.

Balaji, V. R., Suganthi, S. T., Rajadevi, R., Kumar, V. K., Balaji, B. S., & Pandiyan, S. (2020). Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier. *Measurement*, *163*.

Ernawati, S., Wati, R., Nuris, N., Marita, L. S., & Yulia, E. R. (2020). Comparison of Naïve

Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application. *Journal of Physics: Conference Series*, *1641*.

Golub, K., Hagelbäck, J., & Ardö, A. (2020). Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches. *J. Data Inf. Sci*, *5*(1), 18–38.

Rezaeian, N., & Novikova, G. (2020). Persian text classification using naive bayes algorithms and support vector machine algorithm. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, *8*(1), 178–188.

Viet, T. N., Le Minh, H., Hieu, L. C., & Anh, T. H. (2021). The Naïve Bayes algorithm for learning data analytics. *Indian Journal of Computer Science and Engineering*, *12*(4), 1038–1043.

Wang, R. (2022). Automatic classification of document resources based on Naive Bayesian classification algorithm. *Informatica*, *46*(3).

Winarti, T., Indriyawati, H., Vydia, V., & Christanto, F. W. (2021). Performance comparison between naive bayes and k-nearest neighbor algorithm for the classification of Indonesian language articles. *IAES International Journal of Artificial Intelligence*, *10*(3), 452.

Zhang, P., Ma, Z., Ren, Z., Wang, H., Zhang, C., Wan, Q., & Sun, D. (2024). Design of an automatic classification system for educational reform documents based on naive bayes algorithm. Mathematics. *Mathematics*, *12*(8), 1127.

Zhao, Z. (2023). Classification tree algorithm and its application in general archives management system. *Procedia Computer Science*, *228*, 946–951.