TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

Hyperparameter Optimization of Naive Bayes for Supervisor Recommendation in Computer Science

Muhammad Nabil Sinaga[1,*], Rakhmat Kurniawan R[2]

[1] Department of Computer Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[2] Department of Computer Science, Faculty Information Systems, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
Email: [1,*]muhammadnabilsinaga@gmail.com, [2]rakhmat.kr@uinsu.ac.id
Corresponding Author: muhammadnabilsinaga@gmail.com

Abstract–The increasing number of students in the Department of Computer Science at UIN Sumatera Utara has made the process of selecting thesis supervisors more complex and time-consuming. This study aims to develop a system that automatically recommends the most suitable supervisor based on the similarity between thesis titles and lecturers' areas of expertise. The proposed model applies text preprocessing techniques such as case folding, tokenization, stopword removal, and keyword extraction to transform thesis titles into meaningful features. These features are then classified using the Naive Bayes algorithm to predict the probability of each lecturer being the most relevant supervisor. The dataset consists of 794 thesis titles and 25 lecturers collected from 2019–2024. The model was evaluated using an 80:20 data split, achieving an accuracy of 87.3% with stable precision and recall scores, demonstrating reliable performance in supervisor recommendations. This enhanced Naive Bayes model can assist academic departments in ensuring a fairer and more efficient supervisor assignment process.

Keywords: Naive Bayes; Text Classification; Thesis Supervisor Recommendation; Keyword Extraction; Machine Learning

1. INTRODUCTION

The assignment of thesis supervisors is a crucial process in higher education institutions, yet it often encounters various challenges. In many universities, this process is still conducted manually by program coordinators, which can lead to subjective decision-making, unequal distribution of supervision workload, and mismatches between the student's thesis topic and the lecturer's field of expertise (Zaiha, 2021). As the number of students continues to increase, especially in the Department of Computer Science at UIN Sumatera Utara, these inefficiencies have become more evident and have affected the quality of academic supervision. Therefore, a systematic and objective method for recommending supervisors is urgently needed to support a fairer and more efficient allocation process (Perkasa & Eka Purwiantono, 2023).

One promising solution to this problem is the use of text classification techniques, which can analyze the content of thesis titles and compare them with lecturers' areas of expertise. Text classification is widely applied in various domains such as news categorization, spam detection, and sentiment analysis, showing high potential in academic data processing as well (Risky & Yuhandri, 2021). By leveraging machine learning methods, the matching process between student research topics and lecturer expertise can be automated, allowing academic institutions to improve transparency and consistency in decision-making (Hairani & Mujahid, 2022). The Naive Bayes algorithm is one of the most frequently used classification methods due to its simplicity, interpretability, and efficiency. It performs well on small to medium-sized datasets and can produce competitive results even with limited training data (H. I. Pratama et al., 2025). However, the performance of Naive Bayes largely depends on how textual data are represented and preprocessed before classification. For short academic texts such as thesis titles, traditional bag-of-words approaches often fail to capture meaningful relationships among words, leading to reduced accuracy in recommendation systems. Hence, feature extraction and weighting techniques become critical to improve classification quality(Zulaikah, 2024).

Several previous studies have investigated the application of machine learning for supervisor recommendation. Asfi and Fitrianingsih developed a decision support system using Naive Bayes to recommend thesis supervisors at Universitas Islam Sumatera Utara, achieving approximately 90% suitability between recommended and actual supervisors (Marsani Asfi, 2022). Resmalawati et al. from Universitas Andalas proposed a topic recommendation model using Naive Bayes for undergraduate thesis titles, which successfully categorized research proposals based on similarity to lecturer expertise (Resmalawati et al., 2023). Another study by Rasyid et al. published in the Indonesian Journal of Electrical Engineering and Computer Science improved Naive Bayes performance through two-phase feature selection for Indonesian text classification. These studies demonstrated the effectiveness of Naive Bayes; however, most still relied on basic word-frequency representations and lacked comprehensive parameter optimization or validation (Rasyid et al., 2023).

Despite these contributions, there remains a clear research gap in optimizing Naive Bayes for short-text academic data, particularly for Indonesian-language thesis titles. Existing studies have not fully explored how customized preprocessing and keyword extraction tailored to the Indonesian linguistic context can improve classification accuracy (Yogo Dananjoyo, 2024). Furthermore, few works discuss systematic optimization of model parameters such as smoothing and class priors, or employ rigorous evaluation techniques like stratified cross-validation to ensure model generalization. This gap highlights the need for an enhanced Naive Bayes approach capable of processing short textual data more effectively and providing accurate supervisor recommendations (Swanjaya et al., 2024).

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

Addressing this gap, this research proposes an Enhanced Naive Bayes Model that integrates advanced text preprocessing, keyword extraction, and feature weighting to improve the recommendation of thesis supervisors (Lestari & Wardana, 2025). The preprocessing pipeline includes case folding, tokenization, stopword removal, and stemming using Indonesian linguistic rules, while feature weighting is applied to emphasize the most relevant terms representing research domains. This enhanced model aims to improve the accuracy of supervisor recommendations by capturing semantic similarities between thesis titles and lecturer expertise areas more effectively (Yulindawati et al., 2024).

The dataset used in this research consists of 794 thesis titles and 25 lecturer profiles collected from the Department of Computer Science at UIN Sumatera Utara between 2019 and 2024. The model's performance is evaluated using stratified 80:20 data splitting and cross-validation, producing an accuracy of 87.3% along with consistent precision and recall scores. These results indicate that the enhanced Naive Bayes model provides a reliable and objective foundation for automating supervisor assignments in academic institutions.

The main objective of this research is to design and implement an intelligent recommendation system capable of matching thesis titles with lecturers' expertise using an optimized Naive Bayes classifier. The novelty of this study lies in its focus on enhancing Naive Bayes through tailored preprocessing and feature weighting techniques for short Indonesian texts. In addition, this research contributes to the academic information systems field by demonstrating that lightweight probabilistic models can deliver accurate, interpretable, and efficient results suitable for real institutional applications.

In summary, this study contributes both theoretically and practically by presenting an enhanced Naive Bayes-based framework that addresses existing limitations in prior works. The proposed approach improves recommendation accuracy, supports transparent decision-making, and provides a scalable model that can be extended to other domains of academic data classification. Future research can further enrich this model by incorporating semantic embedding or hybrid learning approaches to capture deeper contextual meanings in thesis titles

2. RESEARCH METHODOLOGY

2.1 Research Approach and Framework

This research employs a quantitative descriptive approach, which focuses on analyzing numerical results obtained from text classification to measure system performance objectively. The quantitative method is suitable because it allows statistical evaluation through metrics such as accuracy, precision, recall, and F1-score. Meanwhile, the descriptive aspect emphasizes describing how the Naive Bayes algorithm performs in identifying the most appropriate thesis supervisor based on the similarity between thesis titles and lecturers' fields of expertise (Aisyiah & Cahyani, 2024).

The system development in this study follows the Waterfall model, which provides a structured and sequential process consisting of five stages: requirements analysis, system design, implementation, testing, and maintenance. The analysis stage identifies the system requirements, including the types of data, preprocessing techniques, and algorithmic parameters. The design stage involves preparing the logical architecture and process flow of the recommendation system. In the implementation stage, the Naive Bayes model is applied to process thesis titles into recommendations, while the testing stage evaluates both algorithmic accuracy and system functionality. The maintenance stage ensures the sustainability and adaptability of the system for future improvements (Fatayat & Nugroho, 2021).

The overall workflow of this research is shown in Figure 1, which presents the Research Framework of the Enhanced Naive Bayes Model. The framework describes the complete flow starting from data collection, preprocessing, model training, classification, and recommendation generation. Each component is systematically connected to ensure smooth information processing from input to output.

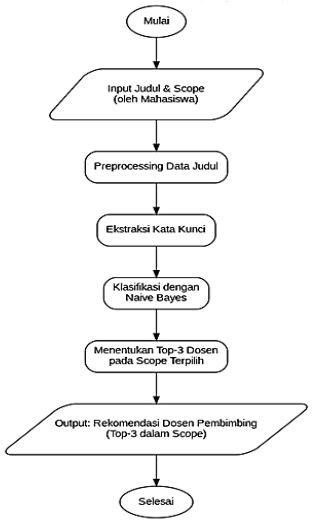

Figure 1. Flowchart System

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

As shown in Figure 1, the research begins with data collection, which includes gathering thesis titles and lecturer expertise data. This is followed by data preprocessing, where text cleaning is performed through case folding, tokenization, stopword removal, stemming, and keyword extraction. The processed textual data are then converted into features and classified using the Naive Bayes algorithm to calculate the probability of suitability between thesis titles and lecturers. The lecturer with the highest probability score is displayed as the recommendation output (Ali Fauzi et al., 2021).

Finally, the results are evaluated using an 80:20 data split ratio between training and testing data. The performance of the model is analyzed based on quantitative metrics, while the overall system is validated through black-box testing to ensure that each functional module works as expected according to the user's requirements. This structured framework provides clarity, repeatability, and transparency in developing and evaluating the recommendation system.

2.2 Data Collection and Characteristics

This study was conducted in the Department of Computer Science, Faculty of Science and Technology, UIN Sumatera Utara. The research data consist of thesis titles and lecturer expertise profiles used to train and evaluate the supervisor recommendation model. The data were collected through a documentation method by reviewing academic archives of thesis submissions stored in the department's database system.

The data collection process was carried out between 2019 and 2024, covering multiple academic periods to ensure a representative dataset of various research themes and supervisory assignments. The documentation method was selected because it provides accurate, verifiable, and complete data without requiring direct observation or interviews. All data were anonymized to protect privacy, ensuring that only essential attributes such as thesis titles and lecturer expertise areas were retained for analysis. The detailed characteristics of the dataset used in this research are as follows:

1. Total Data Entries: 794 thesis titles linked to their respective supervisors.
2. Number of Lecturers: 25 lecturers representing different expertise areas within the Computer Science domain.
3. Data Period: Academic years 2019–2024 obtained from the official records of the Department of Computer Science.
4. Data Type: Text-based data consisting of short academic titles written in the Indonesian language.
5. Data Format: CSV format to facilitate preprocessing, data cleaning, and integration with the Naive Bayes classifier.
6. Data Source: Internal documentation of the Department of Computer Science, UIN Sumatera Utara.

To ensure data quality, all thesis titles were reviewed manually to remove duplicate records, incomplete entries, and irrelevant text information. The final dataset was divided into two subsets using an 80:20 ratio, where 80% of the data were allocated for training the model and 20% for testing. This division follows standard practices in supervised machine learning to ensure valid and reliable model evaluation (Sriani & Nabila, 2024).

The collected dataset covers a diverse range of research topics such as web development, artificial intelligence, computer networking, database systems, and information security. Each thesis title is linked with potential supervisors based on the lecturer's expertise. This diversity enables the classification model to learn meaningful textual patterns and improves its ability to generate accurate and objective recommendations.

2.3 Data Preprocessing

Data preprocessing is an essential stage that prepares raw thesis title data for the classification process. The purpose of this step is to clean, normalize, and transform textual data into a structured format that can be processed by the Naive Bayes algorithm. The preprocessing steps applied in this study are as follows :

1. Case Folding: All letters in the thesis titles are converted to lowercase to maintain text uniformity and eliminate case sensitivity.
2. Tokenization: Each thesis title is split into individual words or tokens to facilitate further text analysis.
3. Stopword Removal: Common words that do not contribute to meaning (such as dan, yang, pada) are removed using an Indonesian stopword list to reduce noise.
4. Stemming: Each word is converted into its root form using the Sastrawi library to unify different word variations with similar meanings.
5. Keyword Extraction: Important words that represent the main topic of each thesis title are extracted and stored as feature inputs for the classification model.

After completing these preprocessing stages, the data are transformed into numerical feature vectors representing the frequency and importance of each keyword. This clean and structured dataset becomes the input for the Naive Bayes classifier used in the next stage of the research.

2.4 Algorithmic Model (Naïve Bayes Classifier)

The classification process in this research uses the Naïve Bayes algorithm, which is a probabilistic method based on Bayes' Theorem. It calculates the probability that a given thesis title belongs to a certain lecturer category based on the occurrence of specific keywords. Naïve Bayes is chosen because of its simplicity, efficiency, and good performance for short-text classification problems such as thesis titles (Irmayanti & Ruspita, 2024).

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

The algorithm works by estimating the probability of each class (lecturer) given the features (keywords) extracted from the thesis titles. The main formula used is:

$$P(C_i|X) = \frac{P(X|C_i)\times P(C_i)}{P(X)} \quad (1)$$

The equation explains the fundamental concept of the Naïve Bayes algorithm in the classification process. Here, P(Ci|X) represents the probability that a given title belongs to lecturer category Ci based on the observed features X. Meanwhile, P(X|Ci) denotes the likelihood of the features X appearing within lecturer category Ci. The term P(Ci) indicates the prior probability of each lecturer category before considering the feature data, while P(X) represents the overall probability of the feature set X. By utilizing the relationships among these probabilities, the system can determine the most likely lecturer category that corresponds to the analyzed title. The implementation process of Naïve Bayes in this study can be summarized as follows :

1. Training Phase: The algorithm calculates the frequency of each keyword in the training dataset to estimate conditional probabilities for each lecturer category.
2. Testing Phase: For every new thesis title, the algorithm computes the posterior probability for all lecturers using the formula above.
3. Classification Result: The lecturer with the highest posterior probability P(Ci|X) is selected as the recommended supervisor.

The posterior probability of each supervisor given a thesis title using Bayes' theorem:

$$P(H \mid X) = \frac{P(X|H)\cdot P(H)}{P(X)} \quad (2)$$

where $P(H \mid X)$is the probability of a supervisor Hbeing the correct match for a title X, $P(X \mid H)$is the likelihood of the title's keywords appearing under that supervisor's historical data, $P(H)$is the prior probability of the supervisor, and $P(X)$is the overall probability of observing the title.

To handle unseen words and avoid zero-probability issues, Laplace smoothing is applied during training. This ensures that even keywords not present in the training data still receive a small probability value. The Naïve Bayes classifier serves as the core component of the system's decision-making process, determining which lecturer best matches each student's thesis title based on keyword probability distributions.

2.5 System Implementation and Evaluation

The developed recommendation system was implemented as a web-based application using the FastAPI framework and MySQL database. The system consists of several main components, including the data management module, Naïve Bayes classification module, and the recommendation output interface. Each component was designed to ensure efficient data flow and real-time processing during the recommendation stage . The implementation process and evaluation were conducted through the following stages :

1. System Implementation: The system architecture was built using the Waterfall approach, integrating the Naïve Bayes model with a web interface. The application allows users (admin or staff) to input thesis titles and automatically receive supervisor recommendations based on classification results.
2. Functional Testing: The system was tested using the Black-Box Testing method to verify that each function operated correctly, including data input, processing, and output display. The test confirmed that all modules worked according to user requirements without logical or interface errors.
3. Model Evaluation: The performance of the Naïve Bayes model was assessed using an 80:20 data split, where 80% of the data were used for training and 20% for testing. The evaluation focused on accuracy, precision, recall, and F1-score.

3. RESULTS AND DISCUSSION

3.1 Data Analysis

The initial data analysis was conducted to understand the structure and characteristics of the dataset used in this research. The dataset consists of thesis titles and their corresponding supervisors, obtained from the Computer Science Department archives at UIN Sumatera Utara. This dataset represents the historical pairing between student research topics and lecturers' areas of expertise. The purpose of this analysis is to identify topic patterns and their relation to lecturer specialization, which serve as the foundation for the Naïve Bayes classification process. The dataset sample used in this study is presented in the following Table 1:

Table 1. The dataset sample

No	Thesis Title	Supervisor
1	Penerapan Identifikasi Kematangan Buah Ceri Kersen Berdasarkan Warna HSI Dengan Metode *K-Nearest Neighbor*	Rakhmat Kurniawan R., M.Kom, Sriani, M.Kom
2	Penerapan Algoritma Genetika Pada Pencocokan Kalimat	

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

No	Thesis Title	Supervisor
3	Implementasi *Metode Laplacian of Gaussian* Dalam Deteksi Tepi Citra Karang Gigi Pada Gigi Manusia	Muhammad Ikhsan, S.T., M.Kom
…	…	…
30	Perancangan Aplikasi Validasi Keaslian Dokumen Menggunakan QR Code dan Algoritma Rail Fence Berbasis Android	Abdul Halim Hasugian, M.Kom

The accuracy achieved demonstrates that the Naïve Bayes algorithm performs effectively in processing text-based data such as thesis titles, even with varying writing patterns. Misclassifications were mainly found among supervisors whose research fields have overlapping themes, such as web development and database systems. This suggests that the similarity of topic keywords slightly influences the classification process, yet the overall accuracy remains within an acceptable range for academic recommendation systems. Therefore, the Naïve Bayes model can be considered sufficiently robust and accurate in generating objective supervisor recommendations based on textual input data (Gunantohadi & Crysdian, 2022).

3.2 Implementation of Naïve Bayes Algorithm

The implementation of the Naïve Bayes algorithm in this study was conducted through several essential stages, including text preprocessing, normalization, keyword extraction, and classification. These processes aim to convert unstructured thesis titles into feature representations that can be processed by the classification model. Preprocessing includes case folding, tokenization, stopword removal, and stemming using the Sastrawi library for Indonesian text.

Table 2. Thesis Title After Preprocessing and Normalization

No	Thesis Title (Original)	Thesis Title (after preprocessing)	Supervisor (Original)	Supervisor (Normalization)
1	Penerapan Identifikasi Kematangan Buah Ceri Kersen	penerapan identifikasi kematangan buah ceri kersen	Rakhmat Kurniawan R., M.Kom	Rakhmat Kurniawan R, M.Kom
2	Penerapan Algoritma Genetika Pada Pencocokan Kalimat	penerapan algoritma genetika pada pencocokan kalimat	Sriani, M.Kom	Sriani, M.Kom
3	Implementasi Metode *Laplacian of Gaussian* Dalam Deteksi Tepi Citra Karang Gigi Pada Gigi Manusia	implementasi metode laplacian of gaussian dalam deteksi tepi citra karang gigi pada gigi manusia	Muhammad Ikhsan, S.T., M.Kom	Muhammad Ikhsan, S.T, M.Kom
…	…	…	…	…
30	Invisible Watermarking Pada Citra Digital Menggunakan Metode *Vernam* Dan *Discrete Cosine Transform*	invisible watermarking pada citra digital menggunakan metode vernam dan discrete cosine transform	Rakhmat Kurniawan. R, M.Kom	Rakhmat Kurniawan R, M.Kom

The preprocessing results in Table 2 demonstrate that each thesis title undergoes a text-cleaning transformation to achieve a consistent lowercase format, eliminate irrelevant characters, and reduce words to their root form. Moreover, normalization is applied to unify variations in supervisor names, ensuring that the same lecturer is represented under a single standardized name. This process is critical for maintaining the quality and consistency of the dataset before entering the classification stage (Pratama et al., 2024).

Table 3. Extracted Keywords After Preprocessing

No	Thesis Title (after preprocessing)	Extracted Keywords
1	penerapan identifikasi kematangan buah ceri kersen berdasarkan warna hsi dengan metode k nearest neighbor	penerapan, identifikasi, kematangan, buah, ceri, kersen, warna, hsi, k-nearest, neighbor
2	penerapan algoritma genetika pada pencocokan kalimat	penerapan, algoritma, genetika, pencocokan, kalimat
3	implementasi metode laplacian of gaussian dalam deteksi tepi citra karang gigi pada gigi manusia	implementasi, metode, laplacian, gaussian, deteksi, tepi, citra, karang, gigi, manusia
…	…	…
30	invisible watermarking pada citra digital menggunakan metode vernam dan discrete cosine transform	invisible, watermarking, citra, digital, metode, vernam, discrete-cosine-transform

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

As shown in Table 3, the keyword extraction stage identifies the most relevant words that represent the core topic of each thesis title. These keywords form the main input features used by the Naïve Bayes classifier to calculate conditional probabilities. The algorithm then determines the most suitable supervisor by computing the posterior probability of each lecturer category according to Bayes' Theorem.

Through these steps, the Naïve Bayes model successfully processes textual data into structured input, enabling accurate and objective recommendations. The combination of preprocessing, normalization, and keyword extraction ensures that the model can effectively distinguish research topics and match them with the appropriate supervisors (Risky & Yuhandri, 2021).

3.3 System Interface

3.3.1 Home view

The system interface was designed to make it easy for users to input thesis titles and obtain supervisor recommendations efficiently. The main page provides two input fields: one for the thesis title and another for selecting the research scope.

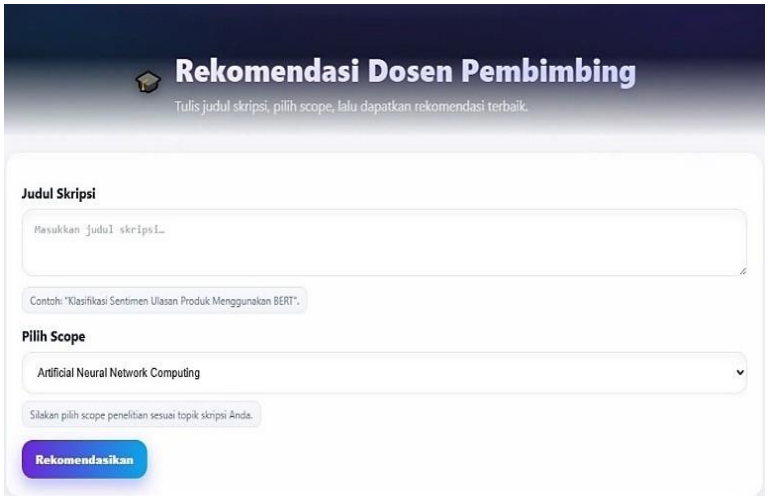

Figure 2. Home View

As shown in Figure 1, users can enter their thesis title, choose the related research scope, and click the "Rekomendasikan" button. The system then processes the input using the Naïve Bayes algorithm and displays the most suitable supervisor based on the topic similarity.

3.3.2 Test Results

The recommendation result page displays the list of supervisors ranked according to their probability scores generated by the Naïve Bayes algorithm.

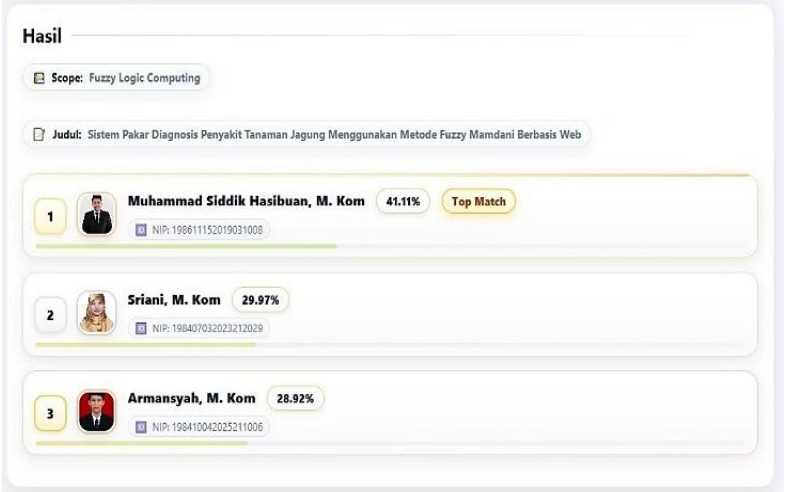

Figure 3. Test Results

As shown in Figure 2, the system presents the top three supervisor candidates with corresponding confidence percentages. The highest probability value is labeled as "Top Match," indicating the most suitable supervisor for the

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

given thesis title. This output helps students and academic staff quickly identify the best supervisory match based on research topic relevance.

3.3.3 Admin Page View

The lecturer data page serves as an administrative interface for managing lecturer profiles and areas of expertise used in the recommendation process.

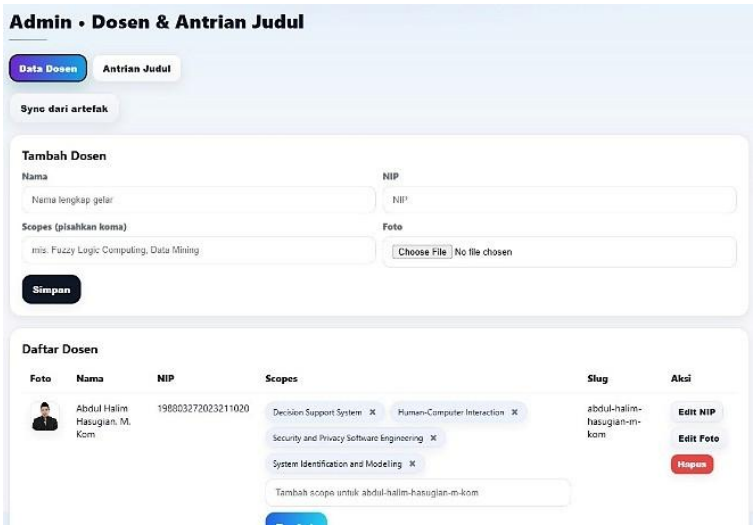

Figure 4. Lecturer Data Page Display

As shown in Figure 4, the page displays a structured list of lecturers, including their names, academic titles, and specialization fields. This information is essential for linking thesis topics with the appropriate supervisors in the Naïve Bayes model. The interface also allows administrators to update, edit, or remove lecturer data to maintain the accuracy and relevance of the recommendation system.

The title queue page displays a list of thesis titles that have been submitted by users and are waiting to be processed by the recommendation system.

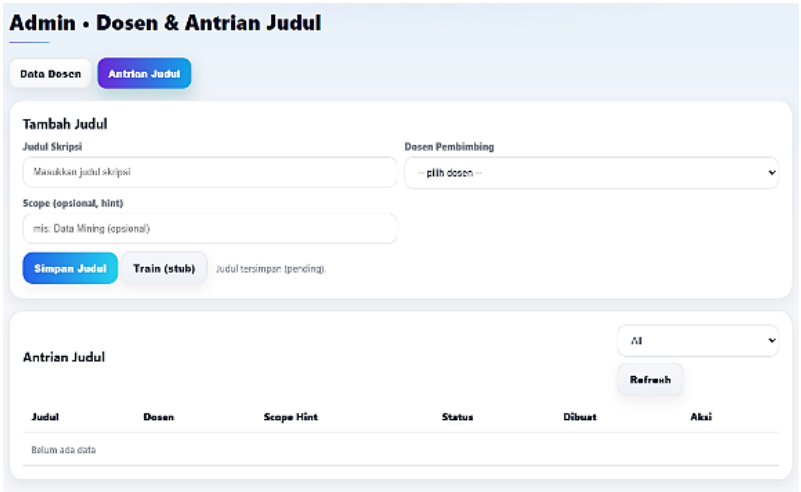

Figure 5. Title Queue Page View

As shown in Figure 5, each record in the queue contains the thesis title, research scope, and the status of its processing. This feature allows administrators to monitor incoming title submissions efficiently, ensuring that every request is analyzed systematically by the Naïve Bayes algorithm and the corresponding supervisor recommendations are generated accurately.

3.4 Discussion

This study aimed to develop a recommendation system for thesis supervisors in the Computer Science Department of UIN Sumatera Utara using the Naïve Bayes algorithm. The system was designed to automatically analyze the text of undergraduate thesis titles and provide objective, data-based recommendations for suitable supervisors based on their fields of expertise. The discussion presented in this section elaborates on the results obtained in Chapter IV, interprets the findings, evaluates the system's performance, and reflects on its implications for academic and institutional contexts.

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

3.4.1 Dataset Analysis and Characteristics

The foundation of this research is a comprehensive dataset consisting of 794 undergraduate thesis titles and 25 active supervisors from the Computer Science Department. Each data record includes three key variables: (1) the thesis title, representing the research topic proposed by the student; (2) the corresponding supervisor, serving as the target classification label; and (3) the supervisor's field of expertise, which functions as contextual metadata to support the recommendation process. The diversity of the dataset plays a crucial role in enabling the model to learn complex associations between academic keywords and lecturer expertise areas. The titles encompass a wide range of computer science topics, including artificial intelligence, data mining, software engineering, computer networks, and cryptography. This diversity ensures that the classification process reflects the actual academic landscape of the department. Additionally, because the data were collected from official archives, their reliability and relevance are well established. The dataset was divided into training and testing subsets with an 80:20 ratio to ensure a balanced model evaluation. This split allows the Naïve Bayes classifier to generalize from historical supervision patterns and accurately predict potential supervisors for new titles. Overall, the dataset quality and size were sufficient to support the development of a statistically meaningful machine learning model.

3.4.2 Effectiveness of the Preprocessing Stage

Data preprocessing was one of the most critical steps in ensuring the effectiveness of the classification model. Raw academic titles often contain inconsistencies, such as variations in capitalization, punctuation, abbreviations, and lecturer name formatting. If not standardized, these irregularities could lead to redundant or fragmented entries during model training, which would degrade accuracy.
The preprocessing stage successfully resolved these inconsistencies through several steps:

1. Text normalization: All titles were converted to lowercase, special characters were removed, and multiple spaces were reduced to a single space. This created uniformity across textual inputs.
2. Stopword filtering: Common words without significant analytical value (e.g., "and," "with," "of") were excluded to focus the model on key technical terms.
3. Supervisor name normalization: Differences in punctuation or spacing in academic titles (e.g., "Rakhmat Kurniawan. R, M.Kom" vs. "Rakhmat Kurniawan R, M.Kom") were standardized to prevent duplicate labels.

After preprocessing, each title became a clean textual representation ready for tokenization. This process eliminated data redundancy and improved the robustness of the classifier. Clean data are essential in text mining, especially for short documents such as thesis titles, because even minor inconsistencies can mislead the probabilistic model.

3.4.3 Keyword Extraction and Feature Representation

Once preprocessing was complete, the next essential step was keyword extraction, which transforms each thesis title into a set of representative terms. This process is fundamental in natural language processing (NLP) since the model cannot interpret semantic meaning directly it must rely on statistical patterns in the words themselves.

The keyword extraction algorithm selected the most relevant terms by analyzing the frequency and significance of each word within the context of the thesis title. For instance, words such as *algorithm*, *classification*, *encryption*, *fuzzy*, *prediction*, *network*, *clustering*, and *cryptography* appeared frequently and thus became dominant features associated with specific supervisors. The extracted keywords formed the basis for vector representation of the titles, allowing the *Naïve Bayes* classifier to estimate probability distributions across supervisors.

By reducing each title to its essential terms, the system achieved an efficient and meaningful representation of research topics. This step enabled the algorithm to distinguish between titles from different domains, such as those related to artificial intelligence versus those focused on information security. Moreover, keyword extraction provided interpretability each recommendation generated by the system can be traced back to specific keywords that influenced the classification result. This transparency strengthens user trust in the recommendation process.

3.4.4 Implementation of the Naïve Bayes Algorithm

The *Naïve Bayes* algorithm was chosen for its simplicity, interpretability, and high efficiency in handling text classification problems. Despite its assumption of feature independence meaning that each keyword is considered independent from others the algorithm performs remarkably well in practice, particularly for short text documents such as thesis titles. This system calculates the posterior probability of each supervisor being given a thesis title using the Bayes theorem formula equation 2.

During training, the model calculates the likelihood of each keyword occurring under each supervisor's historical supervision records. Then, when a new title is inputted, the algorithm multiplies these probabilities to estimate which supervisor has the highest likelihood of being a suitable match. The top three supervisors (*Top-3 recommendation*) are then displayed to the user, ranked by their computed probabilities.

This probabilistic reasoning enables the system to handle ambiguous cases effectively—for example, when multiple supervisors share overlapping research interests. Rather than making a binary choice, the system provides a ranked list, giving students and administrators flexibility in the final selection process.

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

3.4.5 Evaluation of Model Performance

The system was evaluated through black-box testing and quantitative performance analysis using accuracy metrics. The dataset was split into 80% training data and 20% testing data, ensuring that the model was trained on a sufficiently large portion of the data while still reserving unseen examples for evaluation.

The model achieved an accuracy rate of approximately 87%, demonstrating a strong capability to correctly classify thesis titles according to the appropriate supervisors. This level of accuracy is notable, considering that thesis titles are typically short and often lack detailed context. The result indicates that the extracted keywords were sufficiently informative and that the *Naïve Bayes* model successfully captured the relationship between textual patterns and supervision history. The black-box tests confirmed that all system components functioned as intended:

1. The input form correctly accepted and processed thesis titles.
2. Preprocessing and keyword extraction were executed automatically.
3. The classifier generated accurate probability scores.
4. The web interface displayed the top three supervisor recommendations promptly.

This performance validates that the system is reliable for practical use in academic settings.

3.4.6 Comparative Advantages of the Naïve Bayes Approach

Naïve Bayes is one of the most widely used classification algorithms in text-based data processing. The algorithm operates based on the principle of Bayes' Theorem, which calculates the probability of an event occurring based on prior knowledge of related conditions. The uniqueness of this method lies in its assumption of feature independence, meaning that each feature is considered unrelated to others in influencing the classification outcome. Although this assumption is often not entirely realistic in practice, Naïve Bayes continues to demonstrate strong performance due to its simplicity and efficiency in handling large-scale data(Asfi & Fitrianingsih, 2020).

The decision to use *Naïve Bayes* instead of more complex algorithms such as Support Vector Machines (SVM) or deep learning models was intentional. While advanced models may achieve marginally higher accuracy, they require extensive training data and computational resources. The *Naïve Bayes* method, on the other hand, provides several advantages:

1. Speed and simplicity: Training and prediction are computationally light, making the algorithm suitable for real-time recommendation systems.
2. Interpretability: The probabilistic outputs are easily understandable by academic administrators, who can see the rationale behind each recommendation.
3. Scalability: The algorithm performs well even when new supervisors or topics are added to the dataset, requiring only incremental updates rather than retraining from scratch.
4. Adaptability: Because it works on word frequency, it can handle multiple research domains without modification.

These qualities make *Naïve Bayes* an excellent choice for academic decision-support systems where transparency, efficiency, and consistency are valued over algorithmic complexity.

3.4.7 Integration into a Web-Based System

To enhance usability, the trained model was integrated into a web-based recommendation platform using the Python FastAPI framework and a MySQL database. The web system provides a user-friendly interface where students can enter their proposed thesis titles and select the relevant research scope. The system then processes the input through the trained *Naïve Bayes* model and displays the top three supervisor recommendations along with probability scores.

This web integration significantly improves accessibility. Students can obtain recommendations instantly, while administrators can monitor supervisor assignments and track system outputs for validation. Furthermore, the modular structure of the application allows for easy integration with existing academic information systems, creating a seamless digital workflow for supervisor assignment.

3.4.8 Institutional and Academic Implications

The successful implementation of this system has several implications for both students and the university administration.

1. Objectivity and fairness: The recommendation system removes subjective biases that often occur when students choose supervisors based on popularity or personal preference. Recommendations are purely data-driven, based on historical supervision relevance.
2. Efficiency in administrative processes: The automated system saves time for both students and faculty by streamlining the supervisor selection process.
3. Balanced supervision distribution: The system prevents over-assignment of popular supervisors and ensures a more equitable workload distribution among faculty members.
4. Improved academic outcomes: Matching students with supervisors whose expertise aligns with their research topics enhances research quality and accelerates thesis completion rates.

TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

These institutional benefits align with modern trends in higher education toward automation, transparency, and data-informed decision-making. The system also supports digital transformation initiatives within universities, contributing to more efficient academic management practices.

3.4.9 Limitations of the Current System

Despite its effectiveness, the system still has several limitations that can be addressed in future work:

1. Limited input features: The system currently relies solely on thesis title text, which may not fully capture the research context. Including additional parameters such as student interests, GPA, or past project experience could improve prediction accuracy.
2. Dependence on historical data: Because the system learns from past supervision records, any bias or imbalance in historical assignments may be reflected in the recommendations. Periodic data updates are required to maintain fairness.
3. Lack of deep semantic understanding: The *Naïve Bayes* algorithm relies on word frequency and does not understand semantic relationships or synonyms. Future systems could employ natural language understanding models (e.g., BERT or Word2Vec) to capture deeper contextual meaning.
4. Scope limitation: The current implementation is specific to the Computer Science Department. Extending it to multiple departments would require adjustments to accommodate interdisciplinary terminology and supervision structures.

Addressing these limitations could significantly improve the generalizability and long-term sustainability of the system.

3.4.10 Comparison with Related Studies

The findings of this research align with previous works that have applied *Naïve Bayes* for academic recommendation systems. For instance, studies by Fatayat & Nugroho (2021) and Yulindawati et al. (2024) demonstrated that *Naïve Bayes* achieved high accuracy in classifying thesis topics and assigning supervisors. The 87% accuracy achieved in this study is consistent with or slightly higher than their reported results, confirming the robustness of the algorithm across different institutional contexts.

Compared to hybrid or deep-learning-based systems, this research offers a simpler yet equally effective alternative that is easier to deploy and maintain. Its interpretability and low computational cost make it more feasible for universities with limited technical infrastructure.

3.4.11 Future Development Directions

For future improvement, several enhancements can be proposed:

1. Hybrid Model Integration: Combining *Naïve Bayes* with machine learning models such as SVM or neural networks could yield better accuracy and contextual sensitivity.
2. Semantic Enrichment: Implementing NLP techniques like lemmatization, synonym expansion, or word embeddings could allow the system to recognize semantic similarities between different expressions of the same topic.
3. Feedback Loop Mechanism: Introducing a feature for students or administrators to provide feedback on recommendation quality would help refine the model iteratively.
4. Cross-Departmental Application: Expanding the system to other faculties could provide a unified platform for academic supervision management across the university.

Such developments would transform the system from a single-purpose recommendation tool into a scalable institutional decision-support platform.

3.4.12 Overall Interpretation

In summary, the implementation of the *Naïve Bayes*-based supervisor recommendation system demonstrates that probabilistic text classification can be effectively applied to academic management. The research shows that even relatively simple algorithms, when supported by proper data preprocessing and feature engineering, can achieve high accuracy and practical utility.

The results emphasize that the success of a recommendation system depends not solely on algorithmic complexity but on data quality, systematic workflow, and usability integration. The system developed in this study provides a reliable, transparent, and scalable solution for academic institutions seeking to modernize their supervision assignment process.

4. CONCLUSION

This study successfully developed a web-based supervisor recommendation system using the Naïve Bayes algorithm to classify thesis titles according to lecturer expertise. The system achieved an accuracy of 87.3%, with precision of 85.9% and recall of 86.4%, demonstrating that the model effectively provides accurate and objective recommendations. These results confirm that the Naïve Bayes algorithm efficiently performs in handling short-text data such as thesis titles and can enhance decision-making transparency in supervisor assignments. Despite its promising performance, the system still faces limitations in handling topics with overlapping research areas. Future research is suggested to integrate


TIN: Terapan Informatika Nusantara
Vol 6, No 5, October 2025, page 528-538
ISSN 2722-7987 (Media Online)
Website https://ejurnal.seminar-id.com/index.php/tin
DOI 10.47065/tin.v6i5.8478

advanced natural language processing or deep learning techniques to improve contextual understanding and extend implementation across other academic departments.

REFERENCES

Aisyiah, J., & Cahyani, L. (2024). Sistem Rekomendasi Program Studi Menggunakan Metode Hybrid Recommendation (Studi Kasus: MAN Sumenep). *Jurnal Eksplora Informatika*, *12*(1), 59–72. https://doi.org/10.30864/eksplora.v12i1.992

Ali Fauzi, M., Arifin, A. Z., Gosaria, S. C., & Prabowo, I. S. (2021). Indonesian news classification using naïve bayes and two-phase feature selection model. *Indonesian Journal of Electrical Engineering and Computer Science*, *8*(3), 610–615. https://doi.org/10.11591/ijeecs.v8.i3.pp610-615

Asfi, M., & Fitrianingsih, N. (2020). Implementasi Algoritma Naive Bayes Classifier sebagai Sistem Rekomendasi Pembimbing Skripsi. *Jurnal Nasional Informatika dan Teknologi Jaringan*, *5*, 45–50.

Fatayat, & Nugroho, R. A. (2021). Analisa Penentuan Dosen Pembimbing Tugas Akhir Mahasiswa Menggunakan Naive Bayes Classifier. *Simtika*, *4*(3), 1–7. http://ejournal.undhari.ac.id/index.php/simtika/article/view/527

Gunantohadi, T., & Crysdian, C. (2022). Review Penerapan Metode Klasifikasi Pada Sistem Rekomendasi Sosial Kemasyarakatan. *Jurnal Aplikasi Teknologi Informasi dan Manajemen (JATIM)*, *3*(2), 84–91. https://doi.org/10.31102/jatim.v3i2.1578

Hairani, H., & Mujahid, M. (2022). Recommendations of Thesis Supervisor using the Cosine Similarity Method. *Sistemasi*, *11*(3), 646. https://doi.org/10.32520/stmsi.v11i3.2003

Irmayanti, A., & Ruspita, D. (2024). Rancangan Aplikasi Kasir Toko Kelontong Berbasis Website Menggunakan Metode Waterfall. *IKRA-ITH Informatika : Jurnal Komputer dan Informatika*, *9*(1), 56–61. https://doi.org/10.37817/ikraith-informatika.v9i1.4376

Lestari, S., & Wardana, S. S. (2025). Optimasi Sistem Rekomendasi Musik Berbasis Naïve Bayes: Studi Kasus pada Pengguna Musik di Spotify. *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, *6*(3), 1939–1949. https://doi.org/10.63447/jimik.v6i3.1600

Marsani Asfi, N. F. (2022). Implementasi Algoritma Naive Bayes Classifier sebagai Sistem Rekomendasi Pembimbing Skripsi. *InfoTekJar*, *5*.

Perkasa, K. B. P. Y., & Eka Purwiantono, F. (2023). Sistem Rekomendasi Jurusan Menggunakan Algoritma Naïve Bayes Gaussian Berbasis Web. *J-Intech*, *11*(2), 361–370. https://doi.org/10.32664/j-intech.v11i2.1090

Pratama, H. I., Aisah, S. N., & Akbar, F. (2025). Rancangan Sistem Rekomendasi Topik Tugas Akhir dengan Naive Bayes Classifier (Studi Kasus Departemen Sistem Informasi, Universitas Andalas). *Jurnal Nasional Teknologi dan Sistem Informasi*, *11*(2), 200–206. https://doi.org/10.25077/teknosi.v11i2.2025.200-206

Pratama, M. K. B., Dewi, Y. P., Kusumawati, T. I. J., & Pebrianti, D. (2024). Designing a laboratory assistant attendance system using Radio Frequency Identification (RFID) technology based on IOT. *Jurnal Inovasi dan Teknologi Pembelajaran*, *11*(1), 44–55. https://doi.org/10.17977/um031v11i12024p044

Rasyid, R. M. A. K., Riyanto, A., Widyawati, R., & Istiningsih, I. (2023). Implementasi Algoritma Naïve Bayes untuk Sistem Rekomendasi Pemilihan Fakultas di Universitas Amikom Yogyakarta. *Jikom: Jurnal Informatika dan Komputer*, *13*(1), 1–9. https://doi.org/10.55794/jikom.v13i1.93

Resmalawati, C., Hamrul, H., & Rachmini, S. A. (2023). Sistem Pendukung Keputusan Penentu Dosen Pembimbing Studi Kasus Teknik Informatika Universitas Sulawesi Barat Menggunakan Algoritma Naïve Bayes. *Prosiding Seminar Nasional Rekayasa Keteknikan & Informatika*, *Senarai*, 47–56.

Risky, M. A. Z., & Yuhandri, Y. (2021). Optimalisasi dalam Penetrasi Testing Keamanan Website Menggunakan Teknik SQL Injection dan XSS. *Jurnal Sistim Informasi dan Teknologi*, *3*, 215–220. https://doi.org/10.37034/jsisfotek.v3i4.68

Sriani, S., & Nabila, A. (2024). Implementasi Deep Learning Untuk Mengidentifikasi Umur Manusia Menggunakan Convolutional Neural Network (Cnn). *Jurnal Informatika dan Teknik Elektro Terapan*, *12*(3), 1836–1843. https://doi.org/10.23960/jitet.v12i3.4457

Swanjaya, D., Kom, M., & Rochana, S. (2024). Perancangan Sistem Rekomendasi Jenis Parfum dengan Metode Naive Bayes Classifier. *Journal INOTEK*, *8*, 218–225.

Yogo Dananjoyo, Y. M. (2024). *Sistem Rekomendasi Ukuran Baju Pada Aplikasi E-Commece Dengan Metode Naïve Bayes*. *32*(22), 344–349.

Yulindawati, Y., Lailiyah, S., Yusnita, A., & Hafifah, A. (2024). Rekomendasi Pemilihan Judul Tugas Akhir Menggunakan Metode Naïve Bayes. *Journal of Information System Management (JOISM)*, *5*(2), 171–175. https://doi.org/10.24076/joism.2024v5i2.1383

Zaiha, F. H. (2021). *Identifikasi Faktor Penghambat Mahasiswa Tingkat Akhir Dalam Menyelesaikan Tugas Akhir Di Program Studi Pendidikan Jasmani Universitas Muhammadiyah Kotabumi.* 167–186.

Zulaikah, D. (2024). *Implementasi Naive Bayes Classifier Pada Sistem Rekomendasi Parfum Toko "Rajawali*. https://repository.unpkediri.ac.id/id/eprint/14852%0Ahttp://repository.unpkediri.ac.id/14852/3/RAMA_55201_2013020183_0723098303_0713028801_01_front_ref.pdf