JOURNAL OF DATA SCIENCE | Vol.2022:17
eISSN:2805-5160

Classification of MTI Student Thesis Documents at Bina Darma University Palembang Using Naïve Bayes

Dhea Noranita Putri[1], Tri Basuki Kurniawan*[1], Edi Surya Negara[1], Yesi Novaria Kunang[1], Misinem Misinem[2]

[1]Jurusan Magister Teknik Informatika, Universitas Bina Darma, Palembang, Indonesia
[2]Vocational Program, Universitas Bina Darma, Palembang, Indonesia

***Email**: tribasukikurniawan@binadarma.ac.id

Abstract

One of the resources that students might use as a guide when conducting research is the university library. A research thesis written by former students serves as reference material. Students must arrange the thesis documents following the concept or topic of their research because they are typically organized by faculty and department. Researchers, therefore, attempt to classify student thesis documents according to themes or subjects so that students can be more precise in their search for references to themes or topics that relate to the research they will do. The title, abstract, and important keta from the thesis document will be used as the study's data, which will then be classified using the best classification technique, the Naive Bayes Classification (NBC) approach. The learning stage and the testing stage are the two steps used in the naive Bayes classifier method's classification process. After establishing the Category and the quantity of data learning documents, probability calculations were then carried out for each category.

Keywords

Data Mining, Classification, Naïve Bayes, Text Mining

Introduction

Students require a reference when completing their studies. Reference is crucial in research writing since it establishes the reliability of the data we get (N. W. Pollock, 2021). The library is one of the options students consider while seeking references since it offers a variety of collections of earlier student studies from year to year. The library would typically arrange the research material on the bookshelves depending on faculties and departments.

One of the outcomes of study or scientific activity produced by post-graduate students is a thesis (O. Zuber-Skerritt,N. Knight, 1986). Thesis subjects and topics at the Faculty of Computer Science include, among others, data mining, networks, auditing, and e-government. The thesis's themes and subjects may continue to grow as technology progresses. The theme or topic of a thesis is a summary or overview of what the thesis is about.

Submission: 10 November 2022; **Acceptance:** 12 December 2022

Students who plan to conduct research first choose the theme or topic of the study they will perform, then they look for references to support the chosen theme or topic. Classification of thesis documents will result in a classification of thesis themes or topics, whose findings can assist students in finding precise thesis references in accordance with the subject, which is anticipated to assist students in finding their references more quickly.

The words in each sentence of the study's document were converted into numerical data using the Term Frequency-Inverse Document Frequency (TF-IDF) method before classification analysis was conducted. This was followed by the development of a classification model using the Naive Bayes Classifier (NB), Random Forest Classifier (RF), and Support Vector Machine (SVM) (SVM).

Methodology

The MTI student thesis paper from Bina Darma University Palembang was utilized as the source of the data for this investigation. There are 118 documents in the data set, divided into 9 classifications in total. Before pre-processing, which includes case folding, tokenization, stop word removal, and stemming, the data will first be manually labeled. Model training will be carried out following the pre-processing, and that will be followed by research testing (testing) and reaching research results. Figure 1 depicts the study framework in detail.

Framework

The framework outlines the exact steps taken by the model employed in this study, from the data input procedure through pre-processing to the classification outcomes. Figure 1 illustrates the research's framework.

Pre-Process

Pre-processing, often known as pre-processing, is a procedure used to prepare raw data before it is put through additional processes. Pre-processing is crucial because it frequently influences how well the classification model performs (S. B. Kotsiantis, et.al, 2006). In practice, the data representation frequently contains numerous properties, but only a small number of them are meaningful (impact). such that certain traits can be eliminated because the data they provide is insufficient or meaningless. Pre-processing phases such as case folding, tokenization, stop word removal, and stemming will be used in this study.

Case Folding

At this point, all lowercase letters will be used in the document's sentences ((S. Mujilahwati, 2016).

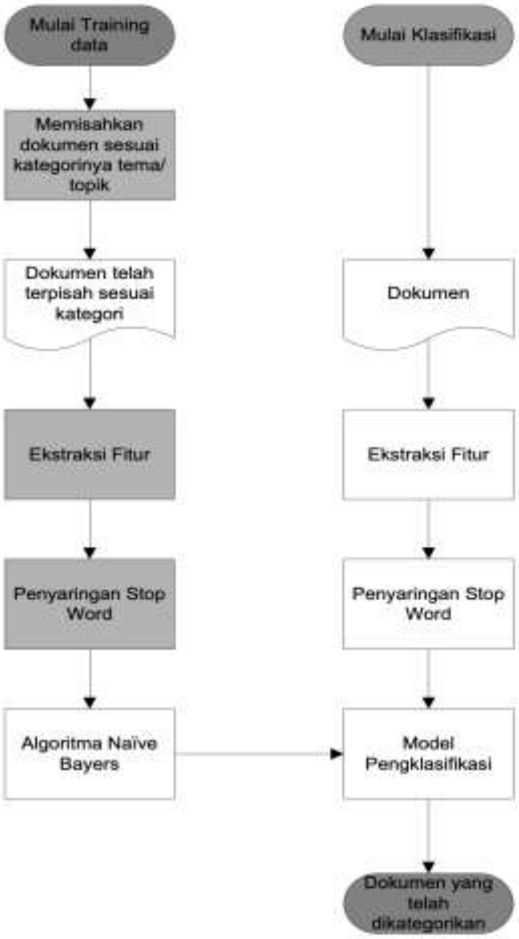

Figure 1. Framework for Document Classification

Tokenization

Tokenization is the act of converting a body of text into a collection of words, phrases, or symbols (J. A. Septian, et.al, 2019), (A. N. Rohman, et.al, 2019). Tokenization is used to investigate terms in a statement. Tokenization can then be used as a starting point for further pre-processing phases such as stop word removal and stemming. This step is frequently seen as less significant because the content is typically already saved in a computer-readable format. However, there are still issues like as deleting punctuation marks, other characters such as brackets and hyphens, and other issues such as shortening sentences. The identification of significant words is the core of tokenization (S. Kannan, et. Al., 2014).

Stop Word Removal

Many words in the document are often repeated (recurring words) but have no meaning. For example, (R. Ferdiana, et.al, 2019) is used to link two words in a phrase. Because of the high frequency of occurrence of terms, their existence in text mining is seen as an impediment to interpreting the content of a document.

The stop word in this study is taken from the Indonesian language; examples of stop words are 'yang,' 'and,' 'di,' and 'from.' Because these words have less relevance in document categorization, they will be removed before training models. This stop-word method also minimizes the amount of text to be processed, improving the overall performance of the system that will be constructed.

Stemming

Stemming is the process of converting all the words in a text or phrase into fundamental terms by eliminating the prefix, insertion (infix), suffix (suffix), and prefix combination (confix) (M. S. Saputri, et.al., 2018). For example, the term 'imitating' will be replaced with the root word 'imitate.' The stemming procedure was carried out in this investigation utilizing the Sastrawi library.

Term Frequency-Inverse Document Frequency (TF-IDF)

Inverse Frequency Term Document Frequency (TF-IDF) is the frequency with which words appear in each document. It is utilized in word weighting to locate meaningful judgments. Document Frequency (DF) is the frequency with which words appear in a document or the number of documents that include the term.

Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) performs modeling by representing words and document components as matrices. Singular Value Decomposition (SVD) is a technique for compressing the occurrence of big words into smaller regions known as semantic spaces (I. F. Rozi, et.al, 2018).

Naïve Bayes

Naïve Bayes was first proposed by the British scientist, Thomas Bayes. Based on prior knowledge, naive Bayes uses probability theory to forecast future opportunities (I. F. Rozi, et.al, 2018). Naive Bayes, therefore, sees the highest likelihood of a data set against all currently recognized classes when classifying data.

Nave Bayes performs permutation from a single hypothesis (data that is input) by assuming that every attribute is neither very consequential nor independent (O. Somantri, 2017). In the theory of Bayes, if there are two separate events, such as X and H, the theory may be stated as follows (O. Somantri, 2017), (Bustami, 2014):

$$P(H|X) = \frac{P(H)P(X)}{P(X)}$$

(Equation 1)

It is then created from equation (1) for classification requirements by converting X into a feature set F1, ..., Fn. The equation contains this equation (2).

$$P\ (C|F_1, \ldots, F_n)\ = \frac{P(C)P(F_1, \ldots, F_n|C)}{P(F_1, \ldots, F_n)}$$

(Equation 2)

In equation (2), C is a class whose classification probability will be shown, whereas F is a characteristic utilized for classification. P(C), often known as a prior or the probability of the class, is the likelihood that the class will be chosen in equation (2). According to Bayes' Theorem, the likelihood of each feature in each class and the probability of each feature are both referred to as evidence. The probability of a feature entering a class is also known as the posterior. Equation (3) is a different kind of equation and illustrates the idea from the preceding explanation (2).

$$Posterior\ = \frac{Prior\ x\ likelihood}{evidence}$$

(Equation 3)

Since the value of the evidence will remain constant in the naive Bayes classification, it may be ignored, leaving just the priors and likelihoods. The equation can be used to describe the probability value (4).

$$\begin{aligned}\mathrm{P}(\mathrm{C}|F_1,\dots,F_n) &= \mathrm{P}(\mathrm{C})\mathrm{P}(F_1,\dots,F_n|\mathrm{C}) = \mathrm{P}(\mathrm{C})\mathrm{P}(F_1|\mathrm{C})\mathrm{P}(F_2,\dots,F_n|\mathrm{C},F_1)\\ &= \mathrm{P}(\mathrm{C})\mathrm{P}((F_1|\mathrm{C})\mathrm{P}(F_2|\mathrm{C},\mathrm{F1})\mathrm{P}(F_3,\dots,F_n|\mathrm{C},F_1,F_2)\\ &= \mathrm{P}(\mathrm{C})\mathrm{P}(F_1|\mathrm{C})\mathrm{P}(F_2|\mathrm{C},F_1)\mathrm{P}(F_3|\mathrm{C},F_1,F_2),\mathrm{P}(F_4,\dots,F_n|\mathrm{C},F_1,F_2)\\ &= \mathrm{P}(\mathrm{C})\mathrm{P}(F_1|\mathrm{C})\mathrm{P}(F_2|\mathrm{C},F_1)\mathrm{P}(F_3|\mathrm{C},F_1,F_2)\dots\mathrm{P}(F_n|\mathrm{C},F_1,F_2,F_3,\dots,F_{n-1})\end{aligned}$$

(Equation 4)

Equation (4)'s solution will be taken to have independence (nave), and F1 - till Fn will be taken to be independent of one another (independent). These presumptions allow for the creation of equation (5)

$$P(P_i|F_j) = \frac{P(F_i\ \cap\ F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i)$$

for i =/ j

$$P(F_i|C,F_j) = P(F_i|C)$$

(Equation 5)

The probability in equation (4) may be reduced to $P(F_i|C)$ according to equation (5). Equation (4) may thus be changed into equation (6), which is the product of all probability features Fi in class C.

$$P\ (C|F_1,F_2,F_3,\dots,F_n)\ = P(C)\prod_{i=1}^{n}P(F_i|C)$$

(Equation 6)

From equation (6), it can be formed a naïve Bayes classification model by multiplying the overall probability of features Fi in class C according to equation (7).

$$\mathrm{P}(\mathrm{C}\mid\mathrm{F})\ =\ \mathrm{P}(F_1\mid\mathrm{C})\ \mathrm{P}(\ F_2\mid\mathrm{C})\ \mathrm{P}(\ F_3\mid\mathrm{C})\dots\mathrm{P}(\ F_n\mid\mathrm{C})\mathrm{P}(\mathrm{C})$$

(Equation 7)

Random Forest

A classification model known as a "random forest" is created by making judgments based on a random selection of data and factors. A random forest is generated by numerous trees describing a forest, which is then examined using the collection of trees. Random forest was used to analyze the data cluster with n observations and p explanatory factors. (Breiman 2001; Breiman & Cutler 2003)

Super Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning (supervised learning) classification technique that makes predictions about six classes based on models or patterns discovered during the training phase. To classify anything, one must find a decision boundary or hyperplane that divides it into distinct categories. In this example, the line helps to distinguish between tweets with positive feelings (labeled +) and tweets with negative sentiments (labeled -) (Han et al., 2006). According to the previous determination, the first stage of SVM is training, which entails examining text patterns to create models. The second stage is prediction (testing), which entails applying a model to categorize text and the outcomes are predictions from classification.

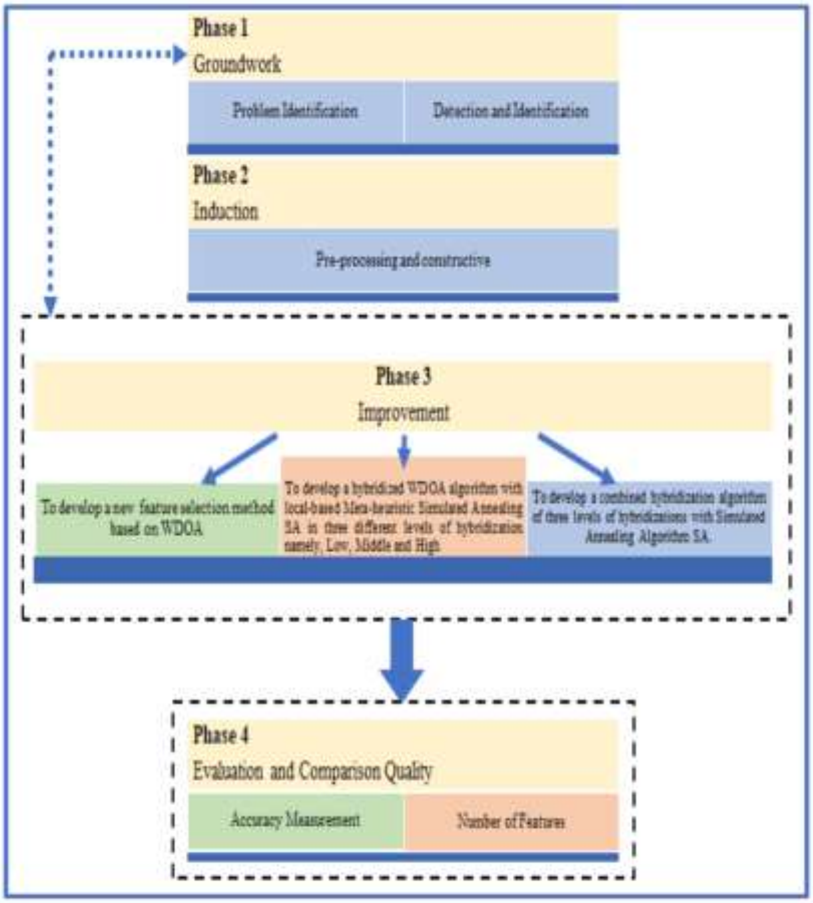
Phase 1
Groundwork
Problem Identification
Detection and Identification
Phase 2
Induction
Pre-processing and constructive
Phase 3
Improvement
To develop a new feature selection method based on WDOA
To develop a hybridized WDOA algorithm with local-based Meta-heuristic Simulated Annealing SA in three different levels of hybridization namely, Low, Middle and High
To develop a combined hybridization algorithm of three levels of hybridizations with Simulated Annealing Algorithm SA
Phase 4
Evaluation and Comparison Quality
Accuracy Measurement
Number of Features

Figure 1. The research framework of Text classification

Problem Identification

The goal of this task is to identify the relevant related studies. We are mainly focusing on understanding the issues facing the development of an effective text classifier. This task was realized by reviewing the most recent text classifiers and their details to identify the existing techniques' strengths and drawbacks.

Two major text classifiers- related problems were identified in this phase. The identified problems are, first, the ambiguity performance of three classifiers to detect the class of new documents. Second, feature reduction while maintaining high performance in feature selection problems.

Detection and Identification Cases

The advancement of the Internet and the increased amount of online information has significantly impacted the ability to detect and identify huge documents. The classification approach is capable of simulating human thinking, and it has been successfully applied in various areas due to its salient features in mining. However, text classification to detect and identify new events, documents, or sentiments is complicated. This is because the large volume of data degrades classifiers' performance due to the high dimensionality of feature space.

Results and Discussion

On 118 data manually collected at the Bina Darma University Palembang library, research testing was done. The thesis's title, abstract, and abstract keywords make up the utilized data. Then, each piece of data will be individually cleansed to check for errors. Following a final check for writing mistakes, the pre-processing step and creation of a classification model are carried out.

The Naive Bayes Classifier (NBC) classification model is utilized, and then the Random Forest Classifier (RFC) method and the Support Vector Machine are compared (SVM). A confusion matrix will be used for research testing, and additional metrics will be determined using the confusion matrix's value. This study will employ an f-score as well as accuracy, precision, recall, and other metrics. Figures 2 to 4 show the performance outcomes for each categorization model.

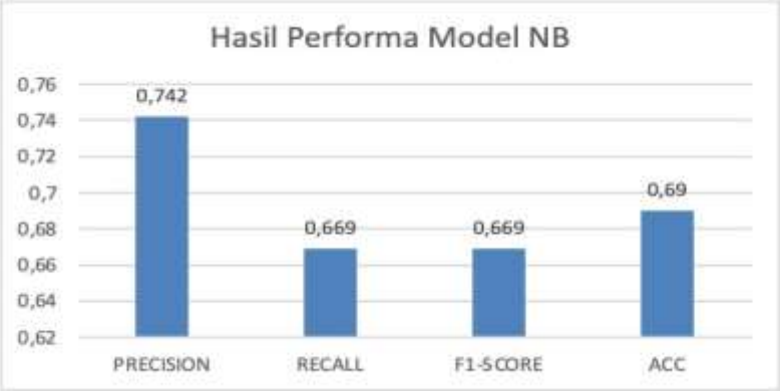

Figure 2. Result Metric Naïve Bayes Classifier

The outcome of the nave Bayes classifier classification model is shown in Figure 2. Gaussian nave Bayes is the type of naive Bayes model utilized. The accuracy of the NB model is 69%, as are the precision, recall, and f-score percentage values. The precision percentage value is 74%.

The performance of the random forest classification model is shown in Figure 3. In RF, the number of n estimators employed is 1000, indicating that 1000 trees were utilized. The accuracy, precision, recall, and f-score percentage values for the RF model are each 78%, 85%, 73%, and 77%, respectively.

The performance of the classification model using support vector machines is shown in Figure 4. The RBF kernel is the parameter for the kernel that is utilized in SVM. The accuracy of the SVM model is 64%, as are the precision, recall, and f-score percentage values. The precision, recall, and f-score percentage values are all above average.

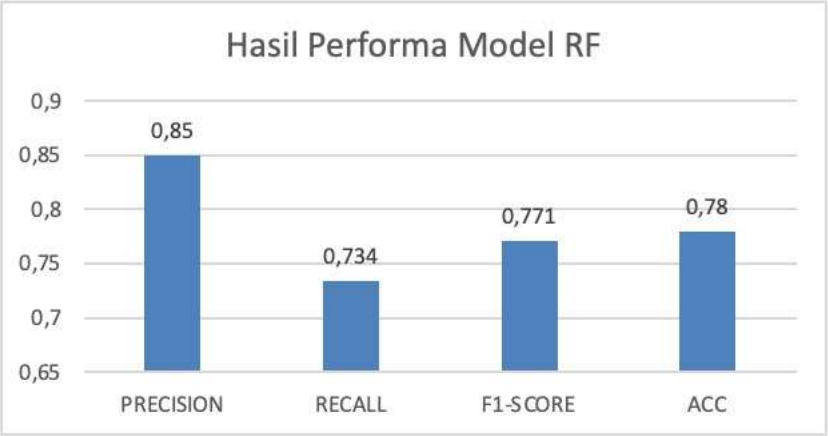

Figure 3. Result Metric Random Forest Classifier

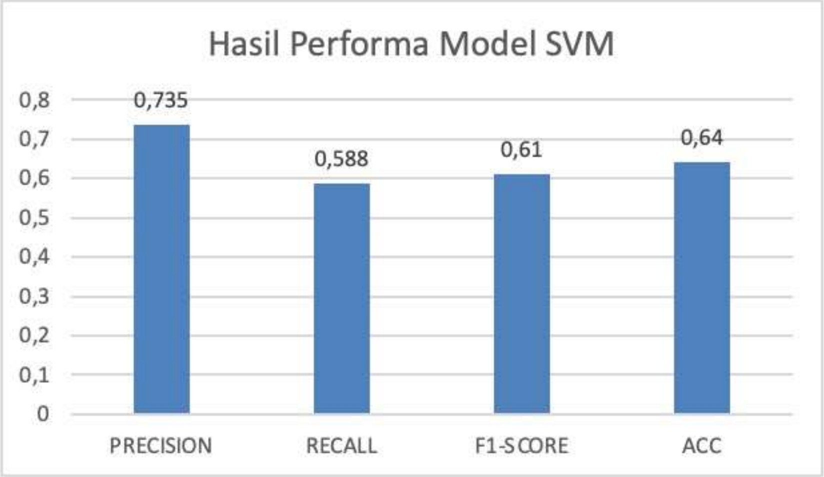

Figure 4. Result Metric Support Vector Machine

With an accuracy score of 78%, NB 69%, and SVM 64%, Table 1 shows that the random forest classifier (RFC) test results outperform the Naive Bayes (NB) and Support Vector Machine (SVM) models.

Table 1. Comparison of Classification Models

Classifier	Precision	Recall	F1-Score	Accuracy
RF	0,85	0,734	0,771	0,78
NB	0,742	0,669	0,669	0,69
SVM	0,735	0,588	0,61	0,64
Classifier	Precision	Recall	F1-Score	Accuracy

With an average accuracy of 85%, recall of 73%, and f-score of 77%, Random Forest also offers additional metrics with better results than the NB and SVM models.

The results of the random forest classification are superior to those of the nave Bayes model classification and the support vector machine, according to a comparison of research test findings for each classification model in Figure 5.

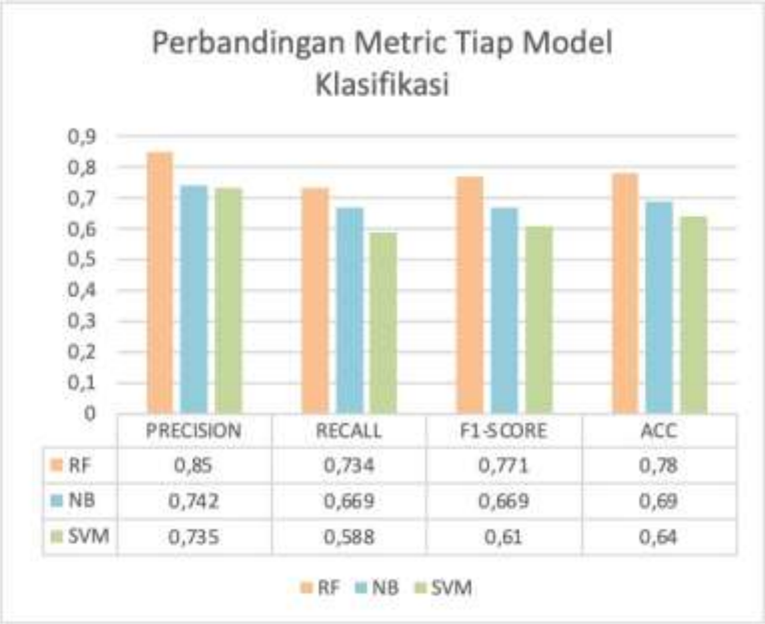

Figure 5. Model Comparison for Classification

According to the outcomes of the experiments that were run, it can be said:

To determine which model produced a superior classification for thesis document classification, the random forest classifier (RF), nave Bayes classifier (NB), and support vector machine (SVM) classification models were compared in this work.

The RF classification model's performance was 78% accurate, 85% precise, 73.4% recall, and 77.1% f-score. The accuracy, precision, recall, and f-score for the NB classification model are 69%, 74.2%, 66.9%, and 66.9% respectively. The accuracy, precision, recall, and f-score for the SVM classification model are 64%, 73.5%, 58.8%, and 61% respectively.

Based on the criteria of accuracy, precision, recall, and f-score, RF produces the classification model with the best performance, while the SVM classification model produces the model with the worst performance.

The student academic repository contains existing documents that can be used in future studies (RAMA). For diploma to doctorate students, RAMA serves as a national archive of research papers (S3). The title and abstract of RAMA publications that have been published generally can be utilized as a data source for additional study. Additional research can employ additional document categories (various disciplines of study, such as computer science, medicine, and literature) and can also look into tertiary institutions to find out what subjects students are most interested in studying there.

Conclusion

The random forest classifier (RF), the nave Bayes classifier (NB), and the support vector machine (SVM) classification models were tested in this paper to ascertain whether the model gave a better classification for thesis document categorization. Performance metrics for the RF classification model were 73.4% recall, 85% precision, 78% accuracy, and 77.1% f-score. The NB classification model has a 69% accuracy rate, a 74.2% precision rate, a 66.9% recall rate, and an f-score of 66.9%. The SVM classification model has a 64% accuracy rate, a 73.5% precision rate, a 58.8% recall rate, and a 61% f-score. The classification model produced by RF has the best performance based on the accuracy, precision, recall, and f-score metrics, while the classification model produced by SVM has the poorest.

References

N. W. Pollock, “Referencing in Scientific Writing.,” Wilderness Environ. Med., 2021.

O. Zuber-Skerritt and N. Knight, “Problem definition and thesis writing,” High. Educ., vol. 15, no. 1, pp. 89–103, 1986.

S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, “Data preprocessing for supervised leaning,” Int. J. Comput. Sci., vol. 1, no. 2, pp. 111–117, 2006.

S. Mujilahwati, “Pre-Processing Text Mining Pada Data Twitter,” Semin. Nas. Teknol. Inf. dan Komun., vol. 2016, no. Sentika, pp. 2089–9815, 2016.

J. A. Septian, T. M. Fachrudin, and A. Nugroho, “Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor,” J. Intell. Syst. Comput., vol. 1, no. 1, pp. 43–49, 2019.

A. N. Rohman, E. Utami, and S. Raharjo, “Deteksi Kondisi Emosi pada Media Sosial Menggunakan Pendekatan Leksikon dan Natural Language Processing,” Eksplora Inform., vol. 9, no. 1, pp. 70–76, 2019.

S. Kannan, V. Gurusamy, S. Vijayarani, J. Ilamathi, M. Nithya, S. Kannan, and V. Gurusamy, “Preprocessing techniques for text mining,” Int. J. Comput. Sci. Commun. Networks, vol. 5, no. 1, pp. 7–16, 2014.

R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, “Dataset Indonesia untuk Analisis Sentimen,” J. Nas. Tek. Elektro dan Teknol. Inf., vol. 8, no. 4, pp. 334–339, 2019.

M. S. Saputri, R. Mahendra, and M. Adriani, “Emotion classification on indonesian twitter dataset,” in 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 90–95.

I. F. Rozi, E. N. Hamdana, and M. B. I. Alfahmi, “Pengembangan Aplikasi Analisis Sentimen Twitter Menggunakan Metode Na{\"\i}ve Bayes Classifier (Studi Kasus Samsat Kota Malang),” J. Inform. Polinema, vol. 4, no. 2, p. 149, 2018.

O. Somantri, “Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB),” J. Telemat., vol. 12, no. 01, 2017.

C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, “Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning,” in IISA 2013, 2013, pp. 1–6.

Bustami, “Penerapan Algoritma Naive Bayes,” J. Inform., vol. 8, no. 1, pp. 884–898, 2014.