# IMPLEMENTATION OF TEXT CLASSIFICATION ON USER REVIEWS IN DANA APPLICATION USING SUPPORT VECTOR MACHINE (SVM) AND GAUSSIAN NAÏVE BAYES (GNB)

**[1]Alfatha Fitrah Insan, [2]Detty Purnamasari**
*[1,2]Magister of Information System Management, Gunadarma University*
*[1,2]Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat*
[1]elfathfitrah@gmail.com, [2]dettygunadarma@gmail.com

**Abstrak**

*Metode atau perangkat konvensional tidak dapat secara efisien memproses volume besar dan beragam kategori informasi, yang secara kolektif disebut sebagai big data. Text mining adalah teknik yang umum digunakan untuk menganalisis big data. Penelitian ini mengevaluasi efektivitas Support Vector Machines (SVM) dan Gaussian Naïve Bayes (GNB) dalam klasifikasi ulasan pengguna dari aplikasi DANA, yang diperoleh dari Google Play Store. Empat tahap dasar pada penelitian ini adalah pengumpulan data, persiapan data, pemodelan data, dan evaluasi. Studi ini menggunakan data dari 15.451 ulasan pengguna, membaginya menjadi tiga subdata dengan ukuran data yang bervariasi dan masing-masing subdata memiliki rasio pelatihan dan pengujian yang berbeda-beda. Evaluasi menghitung empat pengukuran, yaitu akurasi, presisi, recall, F1-Score, dan ROC Curve. Hasil penelitian menunjukkan bahwa SVM dan GNB mencapai tingkat akurasi terkecil 75%. SVM mencapai akurasi rata-rata 84%, 88%, dan 91%, sedangkan GNB mencapai akurasi rata-rata 71%, 81%, dan 85%. Berdasarkan hasil implementasi, analisis sentimen lebih efektif ketika dilakukan dengan SVM dibandingkan dengan GNB.*

***Kata kunci:*** *big data, klasifikasi teks, naïve bayes classifier, support vector machine*

**Abstract**

*Conventional methods or devices are unable to efficiently process large volumes and diverse categories of information, which are collectively referred to as big data. Text mining is a commonly used technique for analyzing big data. This study evaluates the effectiveness of Support Vector Machines (SVM) and Gaussian Naïve Bayes (GNB) in the classification of user reviews from the DANA application, obtained from the Google Play Store. The four fundamental phases of the investigation are data collection, data preparation, data modelling, and evaluation. This study utilized a dataset of 15.451 user reviews, dividing it into three subsets with varying data sizes and each subset having varying training-to-testing ratios. The evaluation will calculate four measurements, which are accuracy, precision, recall, F1-Score, and ROC Curve. The results illustrate that SVM and GNB achieved accuracy rates of at least 75%. SVM achieves an average accuracy of 84%, 88%, and 91%, while GNB achieves an average accuracy of 71%, 81%, and 85%. Based on the implementation results, sentiment analysis is more effective when performed with SVM than with GNB.*

***Keywords:*** *big data, naïve bayes classifier, support vector machine, text classification*

## INTRODUCTION

Every transaction involving social media and several applications on the Google Play Store always involves feedback from users who are involved in transactions. The existence of feedback for every transaction provides advantages for both the developer and the user. Developers can have a deeper understanding of the limitations of applications

they have created in order to enhance user satisfaction [1]. As the amount of feedback increases, these user reviews are included in a larger collection of data called big data [2].

The problem with big data is divided into three characteristics, namely Volume, Velocity, and Variety (3Vs) [3]. Big data contains an expansive volume and assortment of information, thus it is hard to be processed physically or utilizing conventional devices. To overcome this, the most common approach to processing big data is text mining. One application of text mining is text classification, where text categorization is carried out. A text classification model, namely machine learning, is needed to categorize text into organized categories [4], [5]. After the text is analyzed, the model applies the appropriate identifier based on the content. The machine learning models that will be implemented in this study are Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB).

Several studies have been a basis for this study. In a previous study on sentiment analysis of Indonesian-language tweets about the 2014 presidential election [6], researchers also utilized the GNB and SVM algorithms. The classification process uses term frequency and TF-IDF features producing the results that the SVM algorithm was more accurate than the GNB method. Another study that discusses sentiment analysis is the implementation of the Lexicon approach to analyze the sentiment of 2017 DKI Jakarta Gubernatorial Candidates on Twitter. The approach uses three attribute classifications (positive, negative, and neutral) to evaluate sentiment [7]. By using the 10-fold cross-validation technique, the result showed that the GNB proved to be the most accurate in classifying with an average accuracy, precision, and recall of 95%. It achieved a true positive (TP) rate of 96.8% and a true negative (TN) rate of 84.6%. The findings of this study revealed that the GNB classification strategy outperformed the SVM method in accurately assessing the sentiment of Indonesian tweets. However, additional studies, such as those conducted by [8], and [9] have shown that SVM is more accurate than other techniques.

This study aims to compare the performance of the SVM and GNB algorithms in classifying user reviews from the DANA application. DANA is one of the popular digital wallets in Indonesia and widely used for its convenient financial services, thus relies heavily on user feedback to maintain and improve its offerings. Classification performance then will be evaluated by accuracy, precision, recall, F1-Score, and ROC curve.

## RESEARCH METHODOLOGY

The detailed approach that was used in this study to implement text classification on user reviews in the DANA application from the Google Play Store consists of four methodology stages, as shown in Figure 1.

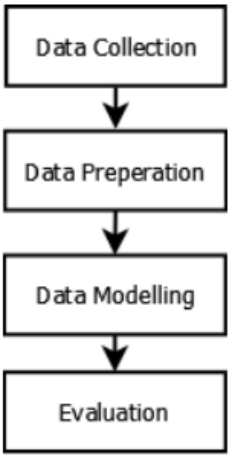


Figure 1. Research Methodology

**Data Collection**

The process starts with collecting data of DANA's user reviews from the Google Play Store. The raw data has been collected by using a third-party application, namely appfollow.io. This method allows efficient conversion of reviews into CSV format making it easily manageable for analysis.

**Data Preparation**

The second stage is data preparation. An essential part of data analysis is to ensure that the data is easy for machines to understand because they can comprehend only 1's and 0's. A technique to make it happen is on this stage, where the raw data will be cleaned and transformed into a basic form, a data format that can be understood by the machine [10]. Since this is the large text dataset, it is necessary to clean it to remove certain differences to avoid inconsistent data. The approach to cleaning the data is quite simple. Data preparation consists of text cleaning, stopword removal, and stemming. All errors and unnecessary component in the data will be removed [11].

**Data Modelling**

Once the data is clean and ready, the data will proceed to the Data modelling stage that consists of labelling, term weighting, and data splitting. Labelling is done to measure the accuracy of the SVM or GNB at the evaluation stage. Term weighting is a mechanism for scoring the occurrence frequency of a word in a text document. The method used for term weighting is TF-IDF (Term Frequency-Inverse Document Frequency), which combines two weight calculation concepts. In TF-IDF, the document is converted into numerical values [12], [13]. Term Frequency (TF) is a method used to determine how often terms appear in a text. In contrast, Inverse text Frequency (IDF) is calculated by dividing the total number of documents by the number of documents that

contain a certain word (Document Frequency, DF), and the results are then converted into logarithmic value [14]. TF-IDF uses two key equations.

$$idf(t) = \log\left(\frac{n}{df(t)}\right) \quad (1)$$

Equation (1) represents the inverse document frequency (IDF) of a term $t$, where $n$ is the total number of documents and $df(t)$ is the number of documents that contain term $t$.

$$w(t,d) = tf\ (t,d) * idf(\text{t}) \quad (2)$$

Equation (2) shows the calculation of TF-IDF weight for term $t$ in document $d$. The term frequency of $t$ in document $d$ is expressed as $tf\ (t,d)$, and $idf(t)$ represents the inverse document frequency of term $t$. Once TF-IDF has been calculated, the next step is to separate the data into training and testing sets.

**Evaluation**

The last phase is evaluation. This study uses SVM and GNB algorithms that are supervised learning methods used for classification text. GNB is a probabilistic classifier that implements Bayes' theorem, while SVM uses statistical learning theory [15]. Accuracy refers to the proportion of accurately predicted cases [16]. Precision measures the number of predicted instances for a class, recall measures the number of correctly predicted items among all relevant items, and F1-Score is the sum of the calculated precision and recall values. The ROC curve is an analytical method represented as a graph to evaluate the ability of the model to differentiate between positive and negative labels [17]. After obtaining the classification report contains those measurements from both algorithms, a comparison will conclude which one performed better.

## RESULTS AND DISCUSSION

This study obtained the dataset from user reviews of the DANA application on the Google Play Store. A third-party software, appfollow.io, processed the reviews and subsequently transformed them into CSV format. Due to limitations in computing power and the need for precise and methodical examination, we only examined 15.451 reviews from the original dataset of 29.887 reviews. In order to evaluate the model's performance across different dataset sizes, the dataset was divided into three subsets, with each subset having 5.237, 11.045, and 15.451 reviews.

**Data Preparation Results**

The dataset is still unrefined, therefore it is necessary to prepare data until it is ready to be analyzed by eliminating irrelevant elements.

Table 1. Cleaning Text Result

| Before | After |
|---|---|
| Mantap pokoknya Dana! Terus berkembang ya.. | mantap pokoknya dana terus berkembang ya |
| Sejauh ini memudahkan buatku :) | sejauh ini memudahkan buatku |
| Sukaaaaa | sukaaaaa |
| CASHBACKNYA TIDAK SESUAI, PAYAH | cashbacknya tidak sesuai payah |

Table 2. Stopword Removal Result

| Before | After |
|---|---|
| mantap pokoknya dana terus berkembang ya | mantap pokoknya dana terus berkembang |
| sejauh ini memudahkan buatku | sejauh memudahkan buat |
| cashbacknya tidak sesuai payah | cashbacknya sesuai payah |

Table 3. Stemming Result

| Before | After |
|---|---|
| mantap pokoknya dana terus berkembang | mantap pokok dana terus kembang |
| sejauh memudahkan buat | jauh mudah buat |
| banyak membantu | banyak bantu |

1. Text Cleaning

   The procedures for text cleaning in this study are eliminating punctuation marks and URLs, converting the uppercase letters to lowercase, and leaving only the words remaining. Some results of text cleaning can be seen in Table 1.

2. Stopword Removal

   The stopwords are words with no meaning and usually appear in large numbers in a text document. The purpose of removing stopwords is to reduce the number of features in text mining to optimize model formation in the classification process [18]. Some results of stopword removal can be seen in Table 2.

3. Stemming

   In this stage, all words are transformed into their base forms by removing prefixes, suffixes, possessive pronouns, and additional particles. Some results of stemming can be seen in Table 3.

**Data Modelling Results**

After the data is cleaned through the data preparation process, the next step is data modelling which consists of labeling, term weighting, and data splitting. In this study, the sentences are given two labels namely positive and negative.

Table 4. Labelling

| Review | Label |
|---|---|
| Mantap pokoknya Dana! Terus berkembang ya | 1 |
| Lazis | 1 |
| Sangat disayangkan applikasi yang menang 2 kategori sebagai best user choice dan best essential app sekarang promonya sudah tidak seperti dulu lagi | -1 |
| Bagus mempercepat proses | 1 |
| Good | 1 |

1. Labelling

   The labeling stage was done by giving a label for each review sentence a label of either 1 or -1. Positive sentences are marked with the number 1 while negative sentences are marked with the number -1. Some results of the labelling process can be seen in Table 4.

2. Term Weighting

   After labelling all text features, the TF-IDF method is used to parse each data point into a vector format. In this process, each word will be weighted, and then TF-IDF will be calculated for each word that appears in the dataset using Tfidf-Vectorizer.

3. Data Splitting

   The data will be split for testing purposes using the SVM and GNB algorithms into various training-test ratios, such as 90:10, 80:20, 70:30, 60:40, and 50:50. This approach of trying various ratios aims to find out the impact of size variations on model performance. In other words, it seeks to assess the level of consistency in the model's performance.

**Evaluation**

In evaluation, a Python-based machine learning library called sci-kit-learn is used to fit the SVM algorithm into the dataset. The metrics are calculated using the LinearSVC function to implement and gain accuracy, precision, recall, and F1-score for each class, i.e., positive and negative. Tables 5, 6, and 7 present the SVM's evaluation results for datasets of different sizes and varying training-testing ratios.

Table 5. SVM Result of 5.237 Data

| Data Ratio | Accuracy | Precision | | Recall | | F1-Score | | Support | |
|---|---|---|---|---|---|---|---|---|---|
| | | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| 90:10 | 85% | 87% | 84% | 81% | 89% | 84% | 87% | 244 | 280 |
| 80:20 | 84% | 87% | 82% | 80% | 89% | 83% | 85% | 513 | 535 |
| 70:30 | 84% | 87% | 81% | 78% | 89% | 82% | 85% | 761 | 1572 |
| 60:40 | 83% | 87% | 81% | 77% | 89% | 82% | 85% | 1019 | 1076 |
| 50:50 | 84% | 88% | 81% | 77% | 90% | 82% | 85% | 1258 | 1361 |
| Average | 84% | 87% | 82% | 79% | 89% | 83% | 85% | | |

Table 6. SVM Result of 11.045 Data

| Data Ratio | Accuracy | Precision | | Recall | | F1-Score | | Support | |
|---|---|---|---|---|---|---|---|---|---|
| | | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| 90:10 | 89% | 83% | 91% | 77% | 94% | 80% | 92% | 326 | 779 |
| 80:20 | 88% | 83% | 90% | 74% | 94% | 78% | 92% | 637 | 1572 |
| 70:30 | 88% | 82% | 90% | 73% | 94% | 77% | 92% | 932 | 2382 |
| 60:40 | 87% | 82% | 89% | 72% | 94% | 77% | 94% | 1274 | 3144 |
| 50:50 | 88% | 82% | 90% | 73% | 94% | 77% | 92% | 1581 | 3942 |
| Average | 88% | 82% | 90% | 74% | 94% | 78% | 92% | | |

Table 7. SVM Result of 15.451 Data

| Data Ratio | Accuracy | Precision | | Recall | | F1-Score | | Support | |
|---|---|---|---|---|---|---|---|---|---|
| | | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| 90:10 | 91% | 87% | 93% | 75% | 97% | 81% | 95% | 368 | 1178 |
| 80:20 | 92% | 86% | 93% | 77% | 96% | 81% | 95% | 746 | 2345 |
| 70:30 | 91% | 86% | 93% | 77% | 96% | 81% | 94% | 1117 | 3519 |
| 60:40 | 91% | 86% | 92% | 75% | 96% | 80% | 94% | 1507 | 4674 |
| 50:50 | 90% | 86% | 91% | 72% | 96% | 78% | 94% | 1886 | 5840 |
| Average | 91% | 86% | 92% | 75% | 96% | 80% | 94% | | |

Based on the percentages generated from various experiments conducted, namely dividing the dataset into 3 stages of size, and each trying various training-testing ratios, it can be observed that the lowest accuracy obtained by the SVM algorithm is 88% and the highest is 91%. As the size of the data increases, the accuracy rises.

Furthermore, the GNB algorithm also uses a Python-based machine learning library called sci-kit-learn, and the metrics are calculated using the GaussianNB function to implement and gain the accuracy, precision, recall, and F1-score for each class, i.e., positive and negative. Tables 8, 9, and 10 present the GNB's evaluation results for datasets of different sizes and varying training-testing ratios.

Table 8. GNB Result of 5.237 Data

| Data Ratio | Accuracy | Precision | | Recall | | F1-Score | | Support | |
|---|---|---|---|---|---|---|---|---|---|
| | | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| 90:10 | 79% | 84% | 76% | 68% | 89% | 75% | 82% | 244 | 280 |
| 80:20 | 79% | 84% | 76% | 71% | 87% | 77% | 87% | 513 | 535 |
| 70:30 | 78% | 83% | 75% | 69% | 87% | 75% | 80% | 761 | 811 |
| 60:40 | 78% | 83% | 75% | 69% | 87% | 75% | 80% | 1019 | 1076 |
| 50:50 | 79% | 83% | 77% | 72% | 86% | 77% | 81% | 1258 | 1361 |
| Average | 79% | 83% | 76% | 70% | 87% | 76% | 82% | | |

Table 9. GNB Result of 11.045 Data

| Data Ratio | Accuracy | Precision | | Recall | | F1-Score | | Support | |
|---|---|---|---|---|---|---|---|---|---|
| | | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| 90:10 | 80% | 78% | 81% | 46% | 94% | 58% | 87% | 326 | 779 |
| 80:20 | 80% | 75% | 81% | 45% | 94% | 56% | 87% | 637 | 1572 |
| 70:30 | 81% | 77% | 82% | 47% | 94% | 58% | 88% | 932 | 2382 |
| 60:40 | 81% | 77% | 82% | 49% | 94% | 59% | 88% | 1274 | 3144 |
| 50:50 | 81% | 75% | 82% | 50% | 94% | 60% | 87% | 1581 | 3942 |
| Average | 81% | 76% | 82% | 47% | 94% | 58% | 87% | | |

Table 10. GNB Result of 15.451 Data

| Data Ratio | Accuracy | Precision | | Recall | | F1-Score | | Support | |
|---|---|---|---|---|---|---|---|---|---|
| | | Positive | Negative | Positive | Negative | Positive | Negative | Positive | Negative |
| 90:10 | 85% | 83% | 86% | 49% | 97% | 62% | 91% | 368 | 1178 |
| 80:20 | 85% | 83% | 86% | 49% | 97% | 62% | 91% | 746 | 2345 |
| 70:30 | 85% | 81% | 86% | 49% | 96% | 61% | 91% | 1117 | 3519 |
| 60:40 | 85% | 82% | 86% | 49% | 96% | 62% | 91% | 1507 | 4674 |
| 50:50 | 85% | 80% | 85% | 49% | 96% | 61% | 90% | 1886 | 5840 |
| Average | 85% | 82% | 86% | 49% | 96% | 62% | 91% | | |

Based on the three various steps conducted, namely dividing the dataset into 3 stages of size, and each trying various training-testing ratios, it can be observed that the lowest accuracy obtained by the GNB algorithm is 79% and the highest is 85%. Similar to what happened in the SVM's evaluation, the accuracy is getting higher as the data size increases. This proves that the more data trained, the better the accuracy because the machine learns a lot. Additionally, the ROC Curve of SVM can be seen in Figure 2, and the ROC Curve of GNB can be seen in Figure 3.

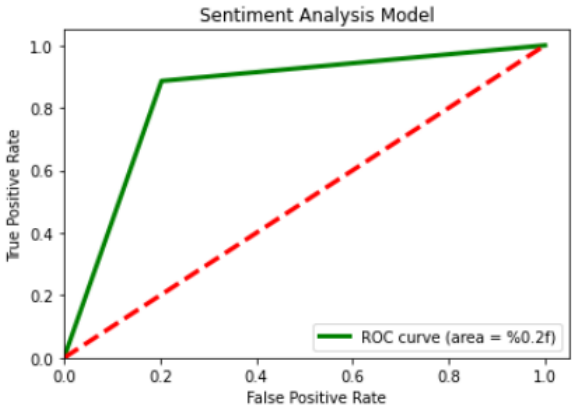


Figure 2. ROC Curve of SVM

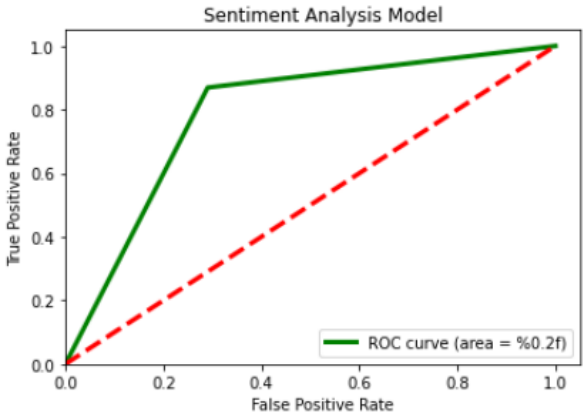


Figure 3. ROC Curve of GNB

Figure 2 and 3 show similar ROC graphs. It has a high AUC (Area Under the Curve) with the conclusion of True Positive Rate more than 80%, and the rest is False Positive Rate. This means that both the SVM and GNB algorithms for classification models have good performance in detecting positive cases because the error rate is relatively low.

## CONCLUSIONS AND SUGGESTIONS

This study has focused on the implementation of machine learning algorithms with SVM and GNB for sentiment analysis text classification. Both algorithms have proven excellent results through their evaluations involving accuracy, precision, recall, F1-score, and ROC curve. The dataset used is only 15.451 out of the total, then split into 3 types of sizes, namely 5.237, 11.045, and 15.451, with each stage having a variation in the training-testing ratios. In terms of accuracy, SVM achieves an average of 84%, 88%, and 91%, while GNB achieves an average accuracy of 79%, 81%, and 85%. The accuracy of SVM consistently surpasses GNB, indicating its superiority in sentiment analysis text classification.

For further research, it is expected to conduct more thorough testing because there are still words that are not included in the KBBI such as contemporary language, abbreviations, and a mix between Indonesian, English and regional languages. Thus, the results obtained will be more helpful in determining which classification algorithm has better capabilities and a better framework. Hopefully, in future research, the data or reviews used will have a better sentence and word structure. Thus, the accuracy of the data values will also be better and more valid.

## BIBLIOGRAPHY


[1] V. A. and S. S. Sonawane, "Sentiment analysis of twitter data: a survey of techniques," *Int. J. Comput. Appl.*, vol. 139, no. 11, pp. 5–15, 2016, doi: 10.5120/ijca2016908625.

[2] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014, doi: 10.1109/ACCESS.2014.2332453.

[3] A. M. Simanjuntak, S. Thamrin, and S. Sundari, "The influence of big data analytics on human resource management strategies for company sustainability", [Online]. Available: https://e-conf.usd.ac.id/index.php/icebmr/

[4] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[5] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text mining: predictive methods for analyzing unstructured*

*information*. 2004. doi: 10.1007/978-0-387-34555-0.

[6] A. F. Hidayatullah and S. N. Azhari, "Analisis sentimen dan klasifikasi kategori terhadap tokoh publik pada data twitter menggunakan naive bayes classifier," vol. 2016, no. semnasIF, pp. 1–8, 2016.

[7] G. A. Buntoro, "Analisis sentimen calon gubernur DKI Jakarta 2017 di twitter," *INTEGER J. Inf. Technol.*, vol. 2, no. 1, pp. 32–41, 2017, doi: 10.31284/j.integer.2017.v2i1.95.

[8] V. Chandani and R. S. Wahono, "Komparasi algoritma klasifikasi machine learning dan feature selection pada analisis sentimen review film," *J. Intell. Syst.*, vol. 1, no. 1, pp. 55–59, 2015.

[9] I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig, and N. A. Mahyoub, "Automatic Arabic text categorization: a comprehensive comparative study," *J. Inf. Sci.*, vol. 41, no. 1, pp. 114–124, 2015, doi: 10.1177/0165551514558172.

[10] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment analysis and classification of Indian farmers' protest using twitter data," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100019, 2021, doi: 10.1016/j.jjimei.2021.100019.

[11] T. N. Prakash and A. Aloysius, "Data preprocessing in sentiment analysis using twitter data," *Int. Educ. Appl. Res. J.*, vol. 3, no. 07, pp. 89–92, 2019, [Online]. Available: https://www.researchgate.net/publication/334670363

[12] S. Fahmi, L. Purnamawati, G. F. Shidik, M. Muljono, and A. Z. Fanani, "Sentiment analysis of student review in learning management system based on sastrawi stemmer and SVM-PSO," *Proc. - 2020 Int. Semin. Appl. Technol. Inf. Commun. IT Challenges Sustain. Scalability, Secur. Age Digit. Disruption, iSemantic 2020*, pp. 643–648, 2020, doi: 10.1109/iSemantic50169.2020.9234291.

[13] G. A. Dalaorao, A. M. Sison, and R. P. Medina, "Integrating collocation as TF-IDF enhancement to improve classification accuracy," *TSSA 2019 - 13th Int. Conf. Telecommun. Syst. Serv. Appl. Proc.*, pp. 282–285, 2019, doi: 10.1109/TSSA48701.2019.8985458.

[14] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, 2019, doi: 10.1186/s13673-019-0192-7.

[15] K. Zishumba, "Sentiment analysis based on social media data," *J. Inf. Telecommun.*, pp. 1–48, 2019, [Online]. Available: http://repository.aust.edu.ng/xmlui/bitstream/handle/123456789/4901/Kudzai

Zishumba.pdf?sequence=1&isAllowed=y

[16] Y. N. Kunang and W. P. Mentari, "Analysis of the impact of vectorization methods on machine learning-based sentiment analysis of tweets regarding readiness for offline learning," *JUITA J. Inform.*, vol. 11, no. 2, p. 271, 2023, doi: 10.30595/juita.v11i2.17568.

[17] F. S. Nahm, "ROC Curve: overview and practical use for clinicians," *Korean J. Anesthesiol.*, vol. 75, no. 1, pp. 25–36, 2022.

[18] D. Marutho, S. Handaka, E. Wijaya, and M. Muljono, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," *2018 Int. Semin. Appl. Technol. Inf. Commun.*, pp. 533–538, 2018, doi: 10.1109/ISEMANTIC.2018.8549751.