Journal of Intelligent System and Telecommunications

Journal homepage: https://journal.unesa.ac.id/index.php/jistel/index

Classification of Indonesian University Entrance Tweets Using Machine Learning

Angga Zakariya[1], Fahmi Hasan Firdaus[2], Muhammad Fahril Syahputra[3], Muchammad Abdulloh 'Ubaid[4], Wiyli Yustanti[5], Cendra Devayana Putra[6], Monica Cinthya[7]*

[1234567]Information System, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia

[6]Information Management, Faculty of Management, National Cheng Kung University, Tainan, Taiwan

*Correspondence: E-mail: putracendra@unesa.ac.id, r78117012@gs.ncku.edu.tw

ARTICLE INFO

Article History:
Submitted/Received 29 November 2025
First Revised 28 April 2026
Accepted 28 April 2026
First Available Online 25 May 2026
Publication Date 25 May 2026

Keyword:
Data Mining, Scrapping, Text Classification, Twitter, Machine Learning, Neural Network.

ABSTRACT

Entrance selection for State Universities (PTN) in Indonesia, specifically SNBP, SNBT, and Independent Selection (Mandiri), is a trending topic on social media, particularly Twitter. However, the high volume of tweets creates noise, making it difficult for prospective students to find relevant information. This study aims to classify tweets into three categories (SNBP, SNBT, Mandiri) and compare the performance of four machine learning models: Naïve Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Long Short-Term Memory (LSTM). The data consists of 12,442 tweets collected using specific keywords. The methodology involves preprocessing (cleaning, normalization, stemming), feature extraction using TF-IDF, and model evaluation. The results show that SVM achieved the best performance with an accuracy of 89.41% and an F1-Score of 89.44%, outperforming the deep learning models (ANN and LSTM) for this specific dataset. These findings indicate that traditional machine learning models can be more effective for text classification with moderate dataset sizes compared to complex deep learning architectures.

1. INTRODUCTION

One of the educational issues that always becomes an interesting topic of discussion every year is the State University (PTN) entrance selection process [1]. Given the existence of various selection pathways with different rules, requirements, and schedules, prospective students are required to access information carefully and in a timely manner to succeed in the increasingly fierce academic competition. The Ministry of Education, Culture, Research, and Technology has updated the new student admissions selection system which will be implemented starting

in 2023. The new student admissions system in Indonesia is regulated by Permendikbudristek Number 48 of 2022. In this regulation, there are three selection pathways for state universities, namely: 1) Academic Merit-based Admission (locally known as SNBP), 2) National Standardized Test-based Admission (SNBT), and 3) Independent University-level Selection (Mandiri) [1]. This regulatory change has given rise to a new wave of discussion in the digital space, especially among high school students preparing to enter college. System changes and increasing information needs are encouraging prospective students to turn to faster and more dynamic digital information sources.

In this context, massive information exchange occurs through social media. Social media itself can be defined as an internet-based instrument that facilitates users to represent themselves, collaborate, and build social relationships virtually [2]. Currently, these platforms have become an integral part of people's communication patterns, both in interpersonal and mass contexts. The high intensity of social media use (such as Facebook, Twitter, and YouTube) not only strengthens interactions but also contributes significantly to accelerating the widespread dissemination of public information [3].

However, the massive flow of information on Twitter often becomes a challenge in itself. The phenomenon of information overload, namely a condition where users have difficulty selecting relevant information [4]. This condition causes users to receive excessive exposure to information in a short time, making it difficult to determine the priority of valid information [5]. In addition, the many tweets mixed with personal comments, repetitive questions, and complaints create noise that covers up core information regarding educational selection [6]. As a result, prospective students often have difficulty in sorting out specific and relevant information regarding the Academic Merit-based Admission, National Standardized Test-based Admission, or Independent University-level Selection pathways [7]. This condition indicates the need for an automated system that is able to classify this information in a more structured manner.

Data Mining and Natural Language Processing (NLP) techniques are relevant computational solutions for processing large amounts of text data [8]. In this process, Machine Learning approaches play a crucial role by enabling the system to automatically learn data patterns and improve prediction accuracy without the need for explicit programming for each rule. By utilizing search keywords such as SNBP, SNBT, and Mandiri, data obtained from Twitter can be processed using machine learning-based models to classify tweets according to the PTN selection pathway category. This approach is not only beneficial for prospective students, but also has the potential to support educational institutions in understanding the information needs of the community.

Previous research has shown that NLP and machine learning methods have been widely used in analyzing text data from social media. Previous studies revealed that the Naïve Bayes Classifier algorithm is able to effectively identify product review sentiment [9]. The Support Vector Machine model was also reported to provide high accuracy in classifying public sentiment on national policy issues [10]. In addition, other research shows that Big Data analysis through the Exploratory Data Analysis method produces important insights for data-driven decision-making [11]. These findings indicate that NLP, Data Mining, and machine learning techniques have great potential to be applied to the analysis of information regarding PTN admission selection.

Despite the widespread application of computational approaches to social media, there is a significant research gap regarding comparative studies between conventional Machine Learning algorithms (Naive Bayes and SVM) and Deep Learning approaches (ANN and LSTM), particularly in the domain of the latest PTN selection terminology (Academic Merit-based Admission and National Standardized Test-based Admission). The existing literature largely focuses on sentiment analysis of commercial products or political issues, so the

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

effectiveness of these models in handling the characteristics of educational text data that contains a variety of acronyms and specific noise has not been empirically tested. This absence of references creates an urgent need for research, especially considering the high risk of unclear information faced by prospective students due to dynamic regulatory changes. Therefore, this study aims to fill this gap by identifying the most adaptive and accurate classification model, which is expected to form the basis for the development of a more credible academic information filtering system.

Based on this urgency, this research was conducted to develop an automatic classification model for tweets related to state university admissions. This research involved Twitter data scraping, text pre-processing, feature extraction, and the application of several machine learning models such as Naïve Bayes, Support Vector Machine, Artificial Neural Network, and Long Short-Term Memory. The research results are expected to identify the most effective model for classifying tweets based on the SNBP, SNBT, and Mandiri categories, thus resulting in more structured and useful information access for prospective students and other stakeholders.

2. METHODS

This study adopts the Knowledge Discovery in Database (KDD) framework to classify the university entrance information. KDD is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [12]. To transform raw data into meaningful knowledge, this methodology involves five systematic stages: data selection, pre-processing, transformation, data mining, and evaluation.

2.1 Data Selection

The dataset used in this research was acquired from the social media platform X (formerly Twitter). Twitter was selected as the data source because it offers massive potential for understanding public sentiment and predicting trends through its unstructured and real-time data[13].

2.1.1 Scrapping Data

Data collection was performed using the scraping technique with the tweet-harvest library, utilizing a Twitter authentication token to access public tweets legally. The scraping process was categorized based on three specific university entrance paths: "SNBP," "SNBT," and "Mandiri" (Independent Selection).

For the SNBP category, the search query "snbp OR #SNBP" was filtered for the Indonesian language within the timeframe of January 1 to April 1, 2025. Similarly, for SNBT, the keywords "snbt OR #SNBT" were used for the period between March 20 and May 28, 2025. The Mandiri category covered a broader range (2020–2025) using specific keywords such as "#SIMAKUI," "#UTULUGM," and other university-specific terms. This process resulted in a total collection of 12,442 raw tweets, comprising 4,013 tweets for SNBP, 4,210 for SNBT, and 4,232 for Mandiri. The distribution of the acquired raw data for each category is illustrated in **Figure 1**.

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

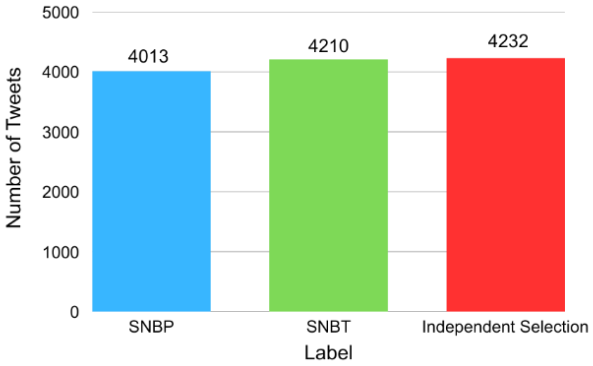

Figure 1. Distribution of Scraped Raw Data per Category

2.1.2 Data Labelling

In this study, the labelling was primarily determined by the source of the scraped data, where tweets retrieved via specific keywords for "SNBP," "SNBT," and "Mandiri" were automatically assigned their respective category labels. A manual verification was subsequently performed to ensure the relevance of the content to the assigned labels, resulting in a structured and labeled dataset ready for the pre-processing stage.

2.2 Pre-Processing

After the data labeling stage is complete, the next crucial step is text preprocessing. This step aims to transform the raw, unstructured, and noise-filled Twitter crawling results into clean and ready-to-process data. The process involves removing unnecessary characters and normalizing foreign terms into standard Indonesian vocabulary, ensuring the text is composed of proper base words. By preparing the data thoroughly, underlying patterns, trends, and valuable insights embedded in the text can be extracted more effectively for the subsequent analysis[14]. The text preprocessing techniques applied in this study follow the standard pipeline described by [15], comprising the following steps:

2.2.1 Data Cleaning

Data cleaning is the process of removing unnecessary characters, punctuation, and symbols from the text. This stage aims to eliminate noise from the dataset to ensure cleaner input for analysis.

2.2.2 Case Folding

Case folding involves converting all characters in the text into lowercase letters. This step ensures uniformity in the dataset, preventing the same word from being treated differently due to capitalization.

2.2.3 Normalization

Normalization transforms non-standard words, slang, or foreign terms into their standard Indonesian equivalents. This process ensures that the text is composed of formal vocabulary, making it easier to identify word patterns.

2.2.4 Stop Word Removal

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

Stop word removal eliminates common connecting words or conjunctions that frequently appear but carry little semantic meaning. Removing these words helps reduce the dataset size and allows the model to focus on more significant terms.

2.2.5 Tokenization

Tokenization is the process of breaking down complex sentences into smaller units known as tokens or individual words. This separation facilitates the analysis of word frequency and sentence structure during the mining process.

2.2.6 Stemming

Stemming transforms words with affixes back into their root or base forms. This technique reduces word variations to a common form, thereby improving the efficiency and accuracy of the classification algorithm.

2.3 Tranformation

In this phase, data transformation is performed to convert the pre-processed data into a format suitable for the data mining procedure. This step involves consolidating or transforming data to maximize the efficiency of the mining process and ensure the extracted patterns are easier to understand. Since this research utilizes unstructured text data, it is necessary to project the textual information into numerical feature spaces to be compatible with machine learning algorithms [16].

2.3.1 Term Frequency-Inverse Document Frequency (TF-IDF)

To transform the unstructured textual data into a numerical format, this study employs the Term Frequency-Inverse Document Frequency (TF-IDF) method with a limit of 5,000 features. As described by [17], TF-IDF is a prevalent technique used to evaluate the significance of a specific term within a document relative to its rarity across the entire corpus. By applying this method, each tweet is converted into a feature vector based on the resulting weight scores. The standard mathematical calculation for TF-IDF is defined as follows:

$$W_{t,d} = TF_{t,d} \times \log\left(\frac{N}{DF_t}\right)$$

Where $TF_{t,d}$ represents the frequency of term t in document d (the tweet), N is the total number of documents in the dataset, and DF_t is the number of documents containing term t.

2.3.2 Encoding

Since machine learning algorithms are fundamentally mathematical, they require numerical input to perform calculations on the target variable. Consequently, the categorical labels in the dataset ("SNBP", "SNBT", and "Mandiri") were transformed into integer values using the Label Encoding technique. This transformation process is essential to map qualitative data into a machine-readable numeric format [18]. In this study, each unique class label was assigned a specific integer (e.g., 0, 1, and 2), ensuring compatibility with the loss functions of the classification models.

2.3.3 Split Data

The final step of the transformation phase is dividing the dataset into training and testing subsets to measure how well the model performs on unseen data. This study applies an 80:20 split, allocating most of the data for training while reserving the rest for performance

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

evaluation. The 80:20 proportion is widely used and commonly justified in practice as a balanced and practical rule for achieving reliable model assessment [19].

2.4 Data Mining

Data mining is a comprehensive analytical process that integrates techniques from multiple disciplines, including statistics, machine learning, artificial intelligence, and pattern recognition, to extract valuable information and uncover hidden relationships within large databases. This term is frequently used interchangeably with or regarded as a pivotal step within the KDD framework, as both share the fundamental objective of processing raw data collections into novel patterns or knowledge that are understandable and beneficial for data stakeholders [19]. To execute this phase, this study utilizes a comparative approach involving two non-neural network algorithms namely Naïve Bayes and Support Vector Machine as well as two neural network architectures specifically Artificial Neural Network and Long Short-Term Memory

2.4.1 Naïve Bayes

Naïve Bayes (NB) is widely regarded as a popular classifier, primarily due to its simplicity and reasonably good performance [16]. The algorithm operates on the fundamental assumption that attributes or features are independent of one another. While real-world problems do not always strictly follow this rule, the method is often preferred as it frequently produces accurate and reliable results. Additionally, it is characterized as a fast classifier that performs effectively with high-dimensional datasets. The mathematical formulation of the Bayes theorem used in this classifier is denoted as:

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

Where $P(C|X)$ describes the posterior probability of class C given input features X, $P(C|X)$ represents the conditional probability (likelihood) of observing features X given class C, $P(C)$ denotes the prior probability of the class, and $P(X)$ serves as the predictor probability.

2.4.2 Support Vector Machine

Support Vector Machine (SVM) is a contemporary supervised machine learning approach known for its high accuracy in classification tasks. As described by [20], the fundamental objective of this model is to determine linear separators within an n-dimensional vector space to facilitate the separation of different data categories. The algorithm endeavors to locate an optimal hyperplane that effectively partitions the data instances by maximizing the margin, which is defined as the distance between the hyperplane and the nearest support vectors. Once this optimal boundary is established, extracted text features can be classified into the appropriate categories. The decision function for the linear classifier is mathematically expressed as:

$$f(x) = \text{sign}(w \,.\, x \, + b)$$

Where w represents the weight vector normal to the hyperplane, x is the input feature vector, and b is the bias term. For multi-class classification, the algorithm constructs a set of hyperplanes to distinguish between the SNBP, SNBT, and Mandiri categories.

2.4.3 Artificial Neural Network

Artificial Neural Networks (ANNs) are computational models designed to simulate the human brain's working mechanism, consisting of interconnected processing units where weights are adjusted during training to learn specific tasks [21]. In this study, the ANN

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

architecture was constructed using the Keras framework, comprising a dense input layer of 128 neurons and a hidden layer of 64 neurons. To mitigate overfitting, a dropout regularization rate of 0.5 was applied. The fundamental mathematical operation for each layer in the network is defined by the general equation:

$$y = \sigma(W.x + b)$$

Where y represents the output vector, W denotes the weight matrix, x is the input vector, b is the bias vector, and σ represents the non-linear activation function. Specifically, the Rectified Linear Unit (ReLU) is utilized for the hidden layers, while the Softmax function is applied at the output layer to generate the final probability distribution for the three classes.

2.4.4 Long Short-Term Memory

Long Short-Term Memory (LSTM) is a specialized Recurrent Neural Network (RNN) architecture explicitly designed to address the vanishing gradient problem and learn long-term dependencies in sequential data [22]. In this study, the model is constructed with an Embedding layer (output dimension 128), followed by an LSTM layer comprising 64 units, and a Dropout regularization layer with a rate of 0.3. Subsequently, a dense layer with 64 neurons utilizing the ReLU activation function processes the features before the final classification. The network's ability to regulate information flow relies on gating mechanisms, particularly the Forget Gate, which determines the retention of information via the following equation:

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$$

Where f_t is the forget gate vector with values between 0 and 1, σ represents the sigmoid function, W_f denotes the weight matrix, h_{t-1} is the hidden state from the previous time step, x_t is the current input, and b_f is the bias term.

2.5 Evaluation

To measure the performance of the classification models, this study utilizes the Confusion Matrix as the fundamental basis for calculation. As described by [23], the confusion matrix provides a detailed breakdown of correct and incorrect predictions through four parameters: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

2.5.1 Accuracy

Accuracy is the most intuitive metric, representing the ratio of correctly predicted observations to the total observations. It measures the overall effectiveness of the model across all classes. The formula is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2.5.2 Precision

Precision quantifies the number of positive class predictions that actually belong to the positive class. It indicates the reliability of the model when it predicts a specific category. The equation is given by:

$$\text{Precision} = \frac{TP}{TP + FP}$$

2.5.3 Recall

Recall measures the ability of the model to identify all relevant instances within a dataset. It calculates the proportion of actual positives that were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

2.5.4 F1-Score

The F1-Score is the harmonic mean of Precision and Recall. This metric is particularly useful when the dataset has an uneven class distribution, as it seeks a balance between precision and recall rather than just optimizing for one.

$$\text{F-1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

3. RESULTS AND DISCUSSION

3.1 Word Cloud

Figure 2. Word Cloud Visualization.

The Word Cloud visualization displays the most frequently occurring terms in the analyzed tweet dataset. Keywords such as "SNBP," "SNBT," and "Mandiri" dominate the display, indicating that the scraping process successfully captured conversations relevant to the topic of university entrance selection.

In addition to these key words, several other terms emerged that reflect the dynamics of public discourse. Words such as "lolos," "takut," "semangat," and "deg-degan" indicate strong emotional involvement among prospective students in the selection process. This demonstrates that discourse regarding state university admissions is not only informative but also emotional.

Procedural terms such as "daftar," "bayar," "akun," and "barcode" indicate that the majority of conversations on Twitter focused on the technical and administrative processes of registration. Interestingly, the emergence of terms such as "joki," "privilege," and "biaya", particularly in the context of the Mandiri Pathway signifies growing public concerns about transparency, fairness, and affordability.

Overall, these initial findings not only demonstrate the language usage patterns typical of social media users, but also support the need to use NLP methods to perform more structured classifications given the high variety of contexts, emotions, and types of information involved in public conversations about PTN selection.

3.2 Comparation Model

Table 1. Performance Evaluation of Classification Algorithms

Model	Accuracy	Precision	Recall	F1-Score
SVM (Non-Neural)	89.41%	89.67%	89.41%	89.44%
LSTM (Keras)	87.12%	87.15%	87.12%	87.12%
Custom ANN (Keras)	86.31%	86.35%	86.30%	86.31%
Naive Bayes (Non-Neural)	82.01%	82.67%	82.01%	82.21%

Experimental results show that the Support Vector Machine (SVM) model outperformed all other models, with the highest accuracy of 89.41% and an F1-score of 89.44%. This superiority can be attributed to the SVM's ability to handle the high-dimensional feature space resulting from the TF-IDF representation, where optimal margin-based splitting provides excellent performance in sparse text classification tasks. This finding is consistent with several

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

previous studies reporting that SVM is often a strong baseline for TF-IDF-based textclassification on medium-sized datasets.

Interestingly, deep learning-based models such as LSTM and ANN demonstrated competitive performance with accuracies of 87.12% and 86.31%, respectively, although still below SVM. Although LSTM is theoretically designed to understand sequential dependencies between words, these results indicate that TF-IDF-based representations do not provide significant structural advantages for sequential models. This suggests that neural network model complexity does not always translate directly to improved performance, especially ondatasets of approximately 12,000 tweets.

On the other hand, Naïve Bayes produced the lowest performance with an accuracy of 82.01%. This can be attributed to the feature independence assumption in Naïve Bayes, which is less appropriate for natural language because the context between words is interdependent. As a result, this model produces a higher misclassification rate than the discriminative SVMapproach.

Overall, these results indicate that conventional vector- and margin-based models like SVM are still very competitive and even superior to Deep Learning approaches for text classification tasks with medium-sized datasets and TF-IDF based features. These findings formed the basis for selecting SVM as the primary model for the subsequent analysis phase.

3.3 Model Prediction Results

Table 2. Sample of Model Predictions with Actual Labels

Tweet	Model Prediction	Actual
rata rata nilai rapor bagus, alhamdulillah dapet warna biru	SNBP	SNBP
open joki buat ujian masuk ptn, hubungi wa0987652798019	Jalur Mandiri	Jalur Mandiri
mau jj dulu soalnya dapet qr code (happy)	SNBT	SNBT
pm dapet 1000 itu makan apa nder?	SNBT	SNBT
simak ui ngambil untungnya gede banget dari pendaftaran woilah	Jalur Mandiri	Jalur Mandiri
ipi yang murah ada univ apa aja ya?	Jalur Mandiri	Jalur Mandiri
dapet bar code nih derrr, mana pilihan 1 lagi wkwkkw	SNBT	SNBT
sekolah masuk top 1000 chance nya gede gak kk?	SNBP	SNBP
soal pk sama pm susah banget kemaren hari keduaaa	SNBP	SNBT
tahun in simak ui bakaln online lagi, takut bgt kalo banyak yang curang...	Jalur Mandiri	Jalur Mandiri

The model generally demonstrated good linguistic pattern recognition and associated contextual keywords with the appropriate classes. For example, a tweet like "rata rata nilai rapor bagus, alhamdulillah dapet warna biru" which are indicators commonly associated with the merit-based selection process.

The model's ability to handle informal language was also evident in the prediction of a tweet with the theme "open joki," which was correctly classified as the Independent Selection Process. This indicates that the model learned not only from formal keywords but also from slang vocabulary and social context commonly used in online conversations about state university entrance selection.

However, some misclassifications were still found. For example, the tweet "soal pk sama pm sulit banget kemaran hari keduaaa" was predicted as SNBP, when the actual label was SNBT. This error likely occurred because the tweet did not contain keywords explicitly referring to the selection process but instead simply reflected an emotional experience or general complaint. Thus, the feature vectors generated from the TF-IDF representation may overlap with similar terms appearing in the SNBP category, particularly tweets with an anxious or reflective tone.

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X

Overall, this analysis shows that the model performs well when tweets contain discriminatory keywords, but performance declines in short text contexts that are ambiguous or do not directly mention the selection pathway. This finding suggests that improvements based on contextual representations such as word embeddings, n-grams, or transformer based models can improve the model's sensitivity to semantic context, particularly in tweets with informal language structures and few keywords.

4. CONCLUSION

This study successfully classified public discourse related to State University (PTN) admission pathways SNBP, SNBT, and Mandiri using Twitter data. Employing the KDD framework on 12,442 tweets, four machine learning models were evaluated: Naïve Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Long Short-Term Memory (LSTM).

The experimental results demonstrate that the Support Vector Machine (SVM) achieved the best performance among all evaluated models, achieving an accuracy of 89.41% and an F1-Score of 89.44%. This outcome indicates that SVM is highly effective for classifying Indonesian short-text social media data, particularly due to its robustness in handling highdimensional sparse TF-IDF features.

Qualitative inspection of prediction samples also revealed the model's capability in recognizing both formal terminology and informal expressions commonly found in Twitter discourse. However, misclassifications were observed in tweets lacking explicit keywords or containing ambiguous context, indicating limitations in TF-IDFb ased representations forsemantic understanding.

Overall, this research demonstrates that automated classification of PTN admission discussions on social media is feasible and can help structure unorganized public information. Future work may incorporate context-aware embeddings such as BERT, IndoBERT, or n-gram extensions to improve semantic sensitivity and reduce ambiguity-driven errors.

5. ACKNOWLEDGMENT

The authors would like to praise Allah SWT for His blessings, which allowed the completion of this research. We express our deepest gratitude to the Department of Information Systems, Faculty of Engineering, Universitas Negeri Surabaya, and the Department of Information Management, National Cheng Kung University, for the support and resources provided. We also wish to acknowledge all parties who provided intellectual contributions and technical assistance during the preparation of this article.

6. AUTHORS' NOTE

The authors declare that this research was conducted solely for academic and research purposes. All data utilized in this study were obtained from public sources and have been anonymized to uphold privacy standards and adhere to data usage ethics. No personal orsensitive information has been stored, disseminated, or used outside the scope of this research.

The authors further declare that there is no conflict of interest regarding the publication of this article and that no external funding was received that could influence the results or direction of this study. The authors welcome academic feedback and discussion for further research development.

7. AUTHORS' CONTRIBUTION/ROLE

All authors contributed equally to this research in accordance with the CRediT (Contributor Roles Taxonomy). The team collaborated in the conceptualization of the research problem and objectives, data curation including extraction and cleaning and the design of the methodology, comprising the preprocessing pipeline and machine learning implementation. Software execution, formal analysis, and result visualization were also carried out jointly. In addition, all authors participated in the writing process, including the original draft preparation and subsequent review and editing to ensure clarity, accuracy, and alignment with the journal standards. All authors have read and approved the final version of this manuscript.

8. AI USE AND DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used Grammarly in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

8. REFERENCES

[1] J. Pengabdian Magister Pendidikan IPA *et al.*, "Pelatihan Tes Penalaran Matematika Bagi Siswa SMA Untuk Menghadapi Seleksi Nasional Berbasis Tes (SNBT)," *Jurnal Pengabdian Magister Pendidikan IPA*, vol. 6, no. 2, 2023, doi: 10.29303/jpmpi.v6i2.4128.

[2] F. Yusuf, H. Rahman, S. Rahmi, A. Lismayani, and P. Guru Sekolah Dasar Universitas Negeri Makassar, "JHP2M: Jurnal Hasil-Hasil Pengabdian dan Pemberdayaan Masyarakat PEMANFAATAN MEDIA SOSIAL SEBAGAI SARANA KOMUNIKASI, INFORMASI, DAN DOKUMENTASI: PENDIDIKAN DI MAJELIS TAKLIM ANNUR SEJAHTERA", [Online]. Available: https://journal.unm.ac.id/index.php/JHP2M

[3] N. A. Azmi, A. T. Fathani, D. P. Sadayi, I. Fitriani, and M. R. Adiyaksa, "Social Media Network Analysis (SNA): Identifikasi Komunikasi dan Penyebaran Informasi Melalui Media Sosial Twitter," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 4, p. 1422, Oct. 2021, doi: 10.30865/mib.v5i4.3257.

[4] A. A. Nuralmi, M. N. Abdurrazaq, and I. Maulana, "Kondisi Information Overload pada Mahasiswa IAI AL-AZIS Akibat Penggunaan Media Sosial yang Tidak Sehat," 2024.

[5] E. Charles and S. Limanto, "JURNAL EDUPEDIA Universitas Muhammadiyah Ponorogo http://studentjournal.umpo.ac.id/index.php/edupedia DAMPAK PENGGUNAAN SOCIAL MEDIA OVERLOAD TERHADAP PERFORMA AKADEMIK DI KOTA BATAM," 2021, [Online]. Available: http://studentjournal.umpo.ac.id/index.php/edupedia

[6] N. Permatasari, R. Yosral, C. Fitri Annisa, P. S. Stis, B. Pusat, and S. Ri, "Analisis Media Sosial Twitter Tentang Pendidikan Daring Pada Masa Pandemi COVID-19 di Indonesia (Permatasari, dkk) ANALISIS MEDIA SOSIAL TWITTER TENTANG PENDIDIKAN DARING PADA MASA PANDEMI COVID-19 DI INDONESIA (Twitter Analysis About Online Education During COVID-19 Pandemic In Indonesia)."

[7] F. A. Rahman, A. Hadiapurwa, and H. Nugraha, "Pengaruh Penggunaan Media Sosial Terhadap Perilaku Pencarian Informasi Akademis Siswa SMAN 2 Cimahi," *Al-Ma mun Jurnal Kajian Kepustakawanan dan Informasi*, vol. 4, no. 2, pp. 93–108, Dec. 2023, doi: 10.24090/jkki.v4i2.8489.

[8] A. Al Tawil, L. Almazaydeh, D. Qawasmeh, B. Qawasmeh, M. Alshinwan, and K. Elleithy, "Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT," *Computers, Materials and Continua*, vol. 81, no. 2, pp. 3395–3412, 2024, doi: 10.32604/cmc.2024.057279.

[9] A. Hanafiah *et al.*, "SENTIMEN ANALISIS TERHADAP CUSTOMER REVIEW PRODUK SHOPEE BERBASIS WORDCLOUD DENGAN ALGORITMA NAÏVE BAYES CLASSIFIER SENTIMENT ANALYSIS OF CUSTOMER REVIEWS OF SHOPEE PRODUCTS BASED ON WORDCLOUD USING NAÏVE BAYES CLASSIFIER ALGORITHM," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 6, no. 1, 2023.

[10] L. Rofiqi and M. Akbar, "Analisis Sentimen Terkait RUU Perampasan Aset dengan Support Vector Machine," *JEKIN - Jurnal Teknik Informatika*, vol. 4, no. 3, pp. 529–538, Aug. 2024, doi: 10.58794/jekin.v4i3.824.

[11] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Mag*, vol. 17, no. 3, p. 37, Mar. 1996, doi: 10.1609/aimag.v17i3.1230.

[12] T. Srikanth1 *et al.*, "Unveiling Insights with Twitter Data: Exploring Trends, Sentiments, and Predictions Through Social Media Mining," *Journal for Educators, Teachers and Trainers JETT*, vol. 15, no. 5, pp. 355–365, 2024, doi: 10.47750/jett.2024.15.05.35.

[13] M. I. Ghozali, W. H. Sugiharto, and A. Fajar Iskandar, "KLIK: Kajian Ilmiah Informatika dan Komputer Analisis Sentimen Pinjaman Online Di Media Sosial Twitter Menggunakan Metode Naive Bayes," *Media Online)*, vol. 3, no. 6, pp. 1340–1348, 2023, doi: 10.30865/klik.v3i6.936.

[14] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.

[15] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2011.

[16] W. Ahmed, M. Hammad, and K. M. Amin, "Sentiment Analysis on Twitter Using Machine Learning Techniques and TF-IDF Feature Extraction: A Comparative Study," 2023.

[17] Brett. Lantz, *Machine learning with R : learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Packt Publishing, 2013.

[18] V. R. Joseph, "Optimal ratio for data splitting," *Stat Anal Data Min*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.

[19] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, p. 437, Apr. 2020, doi: 10.30865/mib.v4i2.2080.

[20] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach," *Soc Netw Anal Min*, vol. 13, no. 1, Dec. 2023, doi: 10.1007/s13278-023-01030-x.

[21] E. Kariri, H. Louati, A. Louati, and F. Masmoudi, "Exploring the Advancements and Future Research Directions of Artificial Neural Networks: A Text Mining Approach," *Applied Sciences (Switzerland)*, vol. 13, no. 5, Mar. 2023, doi: 10.3390/app13053186.

[22] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

DOI: http://doi.org/10.26740/jistel.v2n1.p64-76
e- ISSN 3090-613X