KHATULISTIWA SMART :
JOURNAL OF ARTIFICIAL INTELLIGENCE

e-ISSN : 3089-9710
p-ISSN : XXXX-XXXX

Khatulistiwa Smart : Journal of Artificial Intelligence
Vol. 1 No. 2 (2025) : 49-56
Available online at : https://journal.literasikhatulistiwa.org/index.php/smart

Spam SMS Classification Analysis Using Naive Bayes with Python Language

Beny Yusman [1]
[1] Bisnis Digital, Fakultas Ilmu Pendidikan dan Ekonomi, Universitas Hafshawaty Zainul Hasan
Beny.univer@gmail.com[1]

Doi

Received: 20 Mei 2025 *Accepted: 11 Juni 2025* *Published: 30 Juni 2025*

Abstract Short Message Service (SMS) continues to be widely used in Indonesia, both by official institutions and private entities, despite the growing prevalence of internet-based communication technologies. This study aims to classify SMS messages into three categories—normal SMS, promotional SMS, and fraudulent (spam) SMS—using the Naïve Bayes algorithm. The dataset used in this study comprises 1,143 records, obtained from an open-source platform on GitHub. The research stages include dataset collection, text preprocessing (consisting of case folding, tokenization, filtering, normalization, and stemming), term weighting using two text representation techniques: Count Vectorizer and TF-IDF, and classification using the Multinomial Naïve Bayes algorithm. Classification performance was evaluated using a confusion matrix, along with accuracy, precision, recall, and F1-score metrics. The results show that both combinations—Multinomial Naïve Bayes with Count Vectorizer and with TF-IDF—performed well in classifying SMS messages. The Count Vectorizer model achieved an accuracy of 93%, while the TF-IDF model demonstrated competitive precision and recall values. These findings confirm that the Naïve Bayes algorithm, when paired with appropriate text representation techniques, can serve as an effective solution for automatic SMS classification systems, particularly for short messages in the Indonesian language. This research also opens opportunities for exploring more advanced classification algorithms in future studies.

Keywords *SMS fraud (spam); promotional SMS; normal SMS; Naïve Bayes; classification.*

1. INTRODUCTION

Short Message Service (SMS) was introduced in 1986 under the GSM (Global System for Mobile Communication) standard as a one-way service that enables the exchange of short alphanumeric text messages between two terminals, with a length of fewer than 140 characters (Battu, 2014). This feature is available on almost every mobile phone and allows users to exchange messages with both close contacts and strangers, such as for offering products, services, promotions, and more.

As technology continues to evolve, SMS is increasingly used by official parties such as government services, healthcare providers, and public agencies. In addition, private institutions such as product and service providers use SMS for purposes such as sending verification codes, billing notifications, promotional messages, and other essential information. According to CNBC Indonesia, Indonesia has approximately 25–28 million active SMS users and 37 million integrated users, indicating that SMS usage in the country remains relatively high (Reviantika et al., 2021).

However, the widespread use of SMS also creates opportunities for irresponsible parties to exploit it by sending spam SMS—messages that are unwanted by the recipient. These messages often mimic promotional content or disguise themselves as legitimate communication with the intent of phishing for users' personal information. Spam SMS messages cause discomfort, raise concerns both at the individual and societal levels, and pose significant privacy and security risks. Therefore, there is a need for an automated system that can classify incoming SMS into categories such as spam, genuine promotions, and normal messages. One of the widely used methods for text classification is the Naïve Bayes algorithm, a probability-based machine learning method used to predict the likelihood of future events based on historical data (Mozina et al., 2004).

Several previous studies have explored the effectiveness of the Naïve Bayes algorithm in SMS spam classification. Fitriana et al. (2020) compared the Naïve Bayes, Support Vector Machine (SVM), and Decision Tree algorithms and reported that Naïve Bayes achieved the best performance with a recall of 0.93, accuracy of 0.94, and an F1-score of 0.92. Chusna and Arif (2021) focused on Indonesian-language SMS spam classification using the Multinomial Naïve Bayes algorithm and obtained a precision of 93%, recall of 94%, F1-score of 94%, and accuracy of 95%. On the other hand, Pranata and Gunawan (2019) implemented the Naïve Bayes method using Java Programming and achieved a relatively low precision of 24%, a recall of 88%, and overall accuracy of 62%.

Although these studies demonstrate the effectiveness of Naïve Bayes in classifying short text messages such as SMS, there has been no systematic comparison of how different text representation techniques influence classification performance. Specifically, no previous studies have explicitly evaluated the comparison between using Count Vectorizer and Term Frequency-Inverse Document Frequency (TF-IDF) in supporting the Naïve Bayes algorithm for SMS classification.

Addressing this gap, the present study offers a novel approach by comparing the effectiveness of two different text representation techniques—Count Vectorizer and TF-IDF—in enhancing the performance of the Naïve Bayes algorithm for classifying SMS into three categories: spam, promotional, and normal. This study is expected to contribute to the development of more optimal and adaptive SMS classification systems for the Indonesian language context, particularly in relation to text preprocessing strategies in machine learning.

2. RESEARCH METHODS

This study implements several stages to reach its research conclusions. The steps undertaken include SMS dataset collection (mining), dataset preprocessing, dataset weighting, dataset classification using the Naïve Bayes algorithm, and finally, classification results analysis. These stages are illustrated in the flowchart shown in Figure 1.

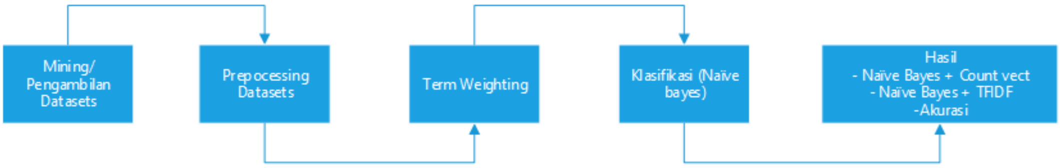

Mining/ Pengambilan Datasets
Prepocessing Datasets
Term Weighting
Klasifikasi (Naïve bayes)
Hasil
- Naïve Bayes + Count vect
- Naïve Bayes + TFIDF
-Akurasi

Figure 1. Flowchart of the Research Methodology

2.1. Dataset Mining

The dataset used in this study was sourced from https://github.com, an open-source platform. It was uploaded by Kuncahyo Setyo Nugroho under the title "Klasifikasi Spam SMS" on November 15, 2019. The dataset contains 1,143 data records and consists of two attributes (columns): Text and Label. The data is categorized into three classes: label 0 includes 596 records of normal SMS messages, label 1 includes 335 records of spam SMS messages (fraud/scam), and label 2 includes 239 records of promotional SMS messages.

2.2. Preprocessing

Preprocessing is the stage of preparing the data specifically using Natural Language Processing (NLP) techniques so that the data becomes suitable and usable for the next stages (Sutojo & Andono, 2017). This step is intended to ensure higher data precision and facilitate the execution of the classification process. The sequence of preprocessing steps is illustrated in Figure 2.

.

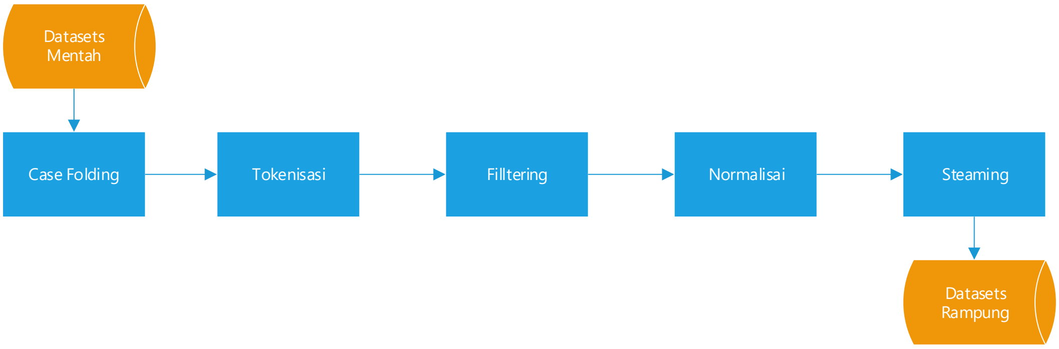


Figure 2. Preprocessing flow/stages

2.2.1. Case Folding

As part of the text folding process, all characters in the dataset (corpus) are converted to lowercase. This technique is useful for systems related to information retrieval, such as search engines (Kedia & Rasu, 2020).

2.2.2. Tokenization

Tokenization is one of the simplest forms of text processing. It is the process of taking a string of characters and breaking it down into smaller parts or tokens typically words that occur most frequently (Thomas, 2020).

2.2.3. Filtering

In this step, sentence pairs that are not suitable for training the system are removed from the dataset. Several factors may determine whether a sentence pair is deemed inappropriate for example, if one of the sentences is too long (e.g., more than 1,000 words), it is likely unusable, since most machine translation (MT) models cannot process text that long (Hagiwara, 2021). In this context, meaningless or informal words such as "sih", "hehe", "ny", "gk", "dg", and so on are also removed.

2.2.4. Normalization

Text normalization is the process of converting text into a single canonical form. This typically involves standardizing various word forms that have the same meaning (Arumugam & Shanmugamani, 2018).

2.2.5. Stemming

Stemming is a text preprocessing task aimed at reducing related or derived word forms (such as "running") to their base form ("run"), since they carry the same meaning (Campesato, 2020).

2.3. Term Weighting

This study applies the term weighting process using the TF-IDF (Term Frequency–Inverse Document Frequency) method. TF (Term Frequency) is used to determine the frequency value of each term's occurrence (Tawalbeh et al., 2022). The TF-IDF formula is as follows:

$$tf - idf\ (t,d) = tf(t,d)\ X\ idf\ (t) \qquad (1)$$

$$idf_t = log\frac{N}{df(t)} \qquad (2)$$

TF(t,d) = frequency of term t in document d
IDF(t) = log(N / df(t))
N = total number of documents
df(t) = number of documents containing the term t

2.4. Classification (Naïve Bayes)

Naïve Bayes classification is a method based on Bayes' Theorem. This classification approach uses statistical and probabilistic principles introduced by the scientist Thomas Bayes (Saleh, 2015). The general formula for Bayes' Theorem is as follows:

$$\mathrm{P(A|B)} = \frac{\mathrm{P\ (B|A)P(A)}}{\mathrm{P(B)}} \qquad (3)$$

Where:

B = The class of data that is unknown

A = The data hypothesis

B is the specific class

P(A | B) = The probability of hypothesis A given condition B (posterior probability)

P(A) = The probability of hypothesis A (prior probability)

P(B | A) = The probability of observing B given A is true

P(B) = The overall probability of B (evidence)

2.5. Evaluation

Through the classification process, results are obtained and presented using several evaluation metrics, including Confusion Matrix, Precision, Accuracy, Recall, and F1-Score. The Confusion Matrix is a commonly used tool to simplify performance measurement, especially in classification problems involving two or more classes. It consists of four components: TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) (Tawalbeh et al., 2022). The following are the basic formulas derived from the Confusion Matrix:

$$Accuracy\ Formula = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision\ Formula = \frac{TP}{TP + FP} \quad (5)$$

$$Recall\ Formula = \frac{TP}{TP + FN} \quad (6)$$

$$F1 - Score\ Formula = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (7)$$

3. RESULTS AND DISCUSSION

As previously stated, this study utilized a total of 1,143 data records, which were then categorized into 569 records of normal SMS, 335 records of fraudulent (spam) SMS, and 239 records of promotional SMS. The distribution of the dataset can be seen in Figure 3.

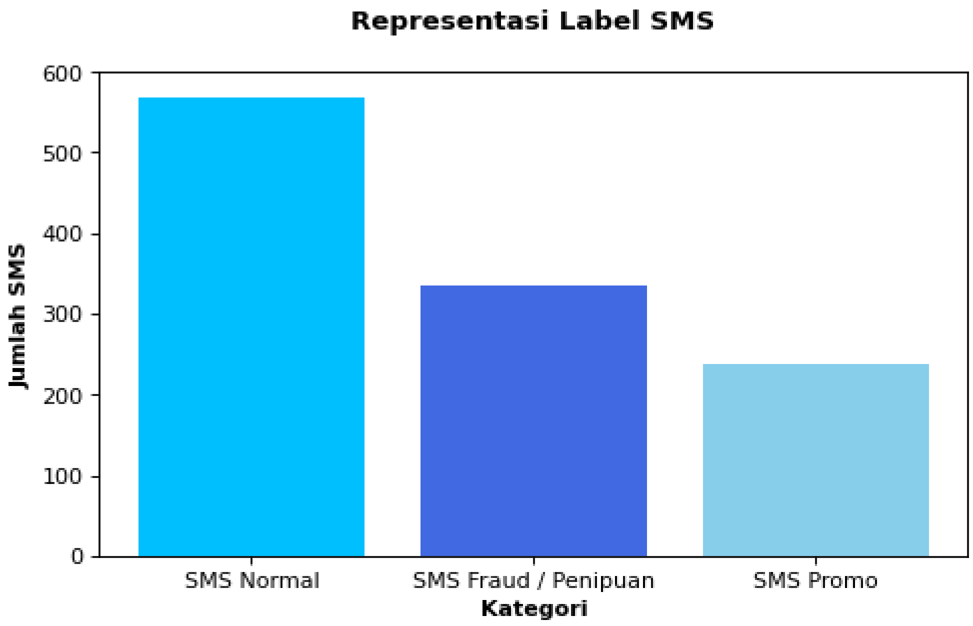
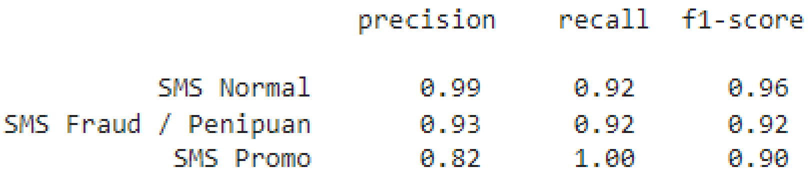


Figure 3. Visualizing Datasets with diagrams

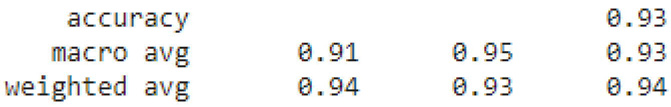
In this study, two comparative classification evaluation processes were conducted to enhance and validate the accuracy of the results. The first approach combined Multinomial Naïve Bayes with Count Vectorizer, and the precision, recall, and F1-score values for each SMS category were obtained, as shown in Figure 4, while the overall accuracy result is presented in Figure 5.



```
                      precision    recall  f1-score

          SMS Normal       0.99      0.92      0.96
SMS Fraud / Penipuan       0.93      0.92      0.92
           SMS Promo       0.82      1.00      0.90
```

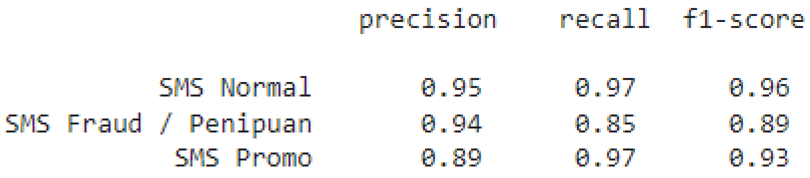
Figure 4. Precision, recall, f1-score values of Multinomial Naive Bayes with count vectorizer



```
    accuracy                           0.93
   macro avg       0.91      0.95      0.93
weighted avg       0.94      0.93      0.94
```

Figure 5. Accuracy value of Multinomial Naive Bayes with count vectorizer

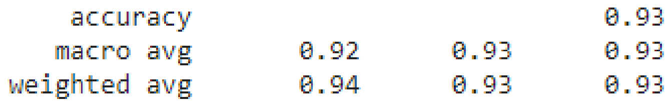
The second evaluation process involved the combination of Multinomial Naïve Bayes with TF-IDF. The resulting precision, recall, and F1-score values for each SMS category are shown in Figure 6, while the overall accuracy score is presented in Figure 7.



```
                      precision    recall  f1-score

          SMS Normal       0.95      0.97      0.96
SMS Fraud / Penipuan       0.94      0.85      0.89
           SMS Promo       0.89      0.97      0.93
```

Figure 6. Precision, recall, f1-score values of Multinomial Naive Bayes with tfidf



```
    accuracy                           0.93
   macro avg       0.92      0.93      0.93
weighted avg       0.94      0.93      0.93
```

Figure 7. Accuracy value of Multinomial Naive Bayes with count tfidf

4. CONCLUSIONS

This study successfully demonstrated the application of the Naïve Bayes classification algorithm in categorizing SMS messages into three classes: fraud (spam), promotional, and normal. Using a dataset of 1,143 records sourced from an open-access platform, the research applied two evaluation approaches: Multinomial Naïve Bayes with Count Vectorizer and with TF-IDF. Both approaches showed effective classification capabilities, with the highest accuracy reaching 93%. The findings imply that the Naïve Bayes algorithm—when combined with appropriate text representation techniques—can serve as a reliable and lightweight solution for spam detection in SMS-based communication systems, particularly within the Indonesian language context. This has practical value for improving user safety, minimizing digital fraud attempts, and enhancing automated message filtering in telecom systems. For future research, it is recommended to expand the dataset to include more diverse SMS samples from various regions and languages. Additionally, further comparative studies involving other machine learning algorithms (e.g., SVM, Random Forest, or deep learning models) and hybrid preprocessing techniques could provide deeper insights into optimizing text classification performance across different messaging platforms.

5. REFERENCES

Karo, I. M. Arumugam, R., & Shanmugamani, R. (2018). Hands-on natural language processing with Python: A practical guide to applying deep learning architectures to your NLP applications. Packt Publishing.

Battu, D. (2014). New telecom networks enterprises and security. Wiley.

Campesato, O. (2020). Artificial intelligence, machine learning, and deep learning. David Pallai.

Chusna, H. N. L., & Arif, M. S. (2021). Klasifikasi SMS spam berbahasa Indonesia menggunakan algoritma. Jurnal Media Informatika Budidarma, 5(4), 1316–1325.

Fitriana, D. N., Setifani, N. A., & Yusuf, A. (2020). Perbandingan algoritma Naïve Bayes, SVM, dan Decision Tree. JUSIM (Jurnal Sistem Informasi Musirawas), 5(2), 167–174.

Hagiwara, M. (2021). Real-world natural language processing: Practical applications with deep learning. Manning.

Kedia, A., & Rasu, M. (2020). Hands-on Python natural language processing: Explore tools and techniques to analyze and process text with a view to building real-world NLP applications. Packt Publishing.

Mozina, M., Demsar, J., Kattan, M., & Zupan, B. (2004). Nomograms for visualization of Naive Bayesian classifier. In Springer-Verlag Berlin Heidelberg (pp. 337–348).

Pranata, E. A., & Gunawan, G. F. (2019). Penerapan metode Naïve Bayes untuk klasifikasi SMS spam menggunakan Java Programming. J-INTECH, 7(2), 104–108.

Reviantika, F., Azhar, Y., & Marthasari, G. I. (2021). Analisis klasifikasi SMS spam menggunakan logistic regression. Asosiasi Prakarsa Indonesia Cerdas (APIC), 4, 155–160.

Saleh, A. (2015). Implementasi metode klasifikasi Naïve Bayes dalam memprediksi. Citec Journal, 2(3), 207–217.

Sutojo, T., & Andono, P. N. (2017). Pengolahan citra digital. ANDI Publisher.

Tawalbeh, L., Alazab, M., Maleh, Y., Baddi, Y., & Gahi, Y. (2022). Big data intelligence for smart applications. Springer Nature Switzerland AG.

Thomas, A. (2020). Natural language processing with Spark NLP: Learning to understand text at scale. O'Reilly Media., K