Eduvest – Journal of Universal Studies
Volume 6 Number 3, March, 2026
p- ISSN 2775-3735- e-ISSN 2775-3727

Comparison of F1-Score Naive Bayes, Logistic Regression, K-Nearest Neighbors, and SVM for Sentiment Classification X in Police Institutions

Robertos Hartanto Wijaya[1], Adi Nugroho[2]
Universitas Kristen Satya Wacana, Indonesia[1,2]
Email: 672021146@student.uksw.edu[1], adi.nugroho@uksw.edu[2]

KEYWORDS:
F1-Score, Machine Learning, SMOTE, Sentiment Analysis

ABSTRACT
Social media, especially platform X, is the main channel for the public to express their opinions on public institutions, including the police. Analysis of public sentiment on this platform can provide insight into police performance. This study aims to compare the performance of machine learning algorithms Naive Bayes, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) in classifying negative sentiments towards policing on social media X, as well as overcoming data imbalances using the SMOTE method. The dataset consisted of 1,274 Indonesian-language data collected by crawling, then processed using preprocessing techniques such as text cleaning, stopword removal, and TF-IDF feature extraction. Testing is conducted with and without the implementation of SMOTE for data balancing. Evaluate the model's performance using F1-Score. Without SMOTE, all algorithms fail to recognize neutral classes. After the implementation of SMOTE, Logistic Regression showed the best performance with an F1-Score of 80.85%, followed by SVM, Naive Bayes, and KNN. The implementation of SMOTE significantly improves the model's ability to classify negative sentiments. The combination of Logistic Regression and SMOTE is the best approach to classifying public sentiment towards policing, which can help police agencies understand public sentiment more accurately.

INTRODUCTION

Social media has become an essential platform that significantly influences how people communicate, access information, and interact. Social media is usually used by a person as a medium of communication, a means of information and as a medium of entertainment for its users (Nasution & Hayaty, 2019). X is a microblogging social media platform that facilitates user interaction. X has become one of the most well-known social media services in the world with over 200 million active users and more than 10.6 billion tweets generated.

In the ever-evolving digital era, data is an important component in various sectors, including government agencies such as the police. Social media acts as a means of public communication that allows the public to express opinions, criticisms, and suggestions on the performance and policies of government institutions. In this context, the public can actively participate in supporting, observing, and criticizing public policies (Al Mustaqim et al., 2024). Therefore, the data generated from interactions on social media is not only relevant for the sake

of communication, but also has value as a source of empirical data in analyzing public perceptions and sentiments on national strategic issues (Syahrohim et al., 2024).

Data generated from people's activities on social media has great potential to be analyzed systematically to understand the trends of public sentiment. Sentiment analysis is one of the widely used approaches to classify public opinion into positive, negative, or neutral categories (Permatasari et al., 2021). This approach is important because it can provide an objective picture of the level of trust, satisfaction, and public criticism of the performance of the police, which can ultimately be an input for policy evaluation and improvement of public services (Handika et al., 2024).

A number of previous studies have examined the use of machine learning algorithms in sentiment analysis. Research by Rangga Nasution shows that the K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms have a competitive level of accuracy in the classification of Twitter sentiment(Matarat et al., 2024). Meanwhile, a study by Y. Handika revealed that Naive Bayes is effective in analyzing sentiment on police performance, although it still has limitations in handling unbalanced data. Another study by A. Sabir showed that Logistic Regression was able to provide stable performance in social media-based text classification. In addition, a comparative study by K. Matarat emphasized that no one algorithm is always superior, so a comprehensive evaluation of various methods is needed to determine the best model (Julizar & Sulaeman, 2025).

Although various studies have been conducted, there are still some research gaps. First, most studies only compare algorithms without deeply considering the data imbalance problem that often occurs in social media datasets (Sabir et al., 2024). Second, research that specifically analyzes public sentiment towards police institutions in Indonesia is still limited, especially those that use a combination of resampling techniques such as the SMOTE. Third, model performance evaluations often focus only on accuracy, whereas metrics such as F1-Score are more relevant in unbalanced data conditions (Bahtiar et al., 2023).

Previous studies have demonstrated variations in the performance of sentiment analysis approaches. Lexicon-based methods such as VADER have been shown to perform well in handling informal language commonly found in social media. However, machine learning approaches, particularly SVM with TF-IDF or Bag-of-Words features, tend to achieve better classification accuracy in sentiment analysis tasks(Qi & Shabrina, 2023). In addition, it was found that a combination of lexicon and machine learning approaches, accompanied by good preprocessing, enabled SVM and MLP to achieve an accuracy of 0.89 in Twitter financial sentiment analysis. These findings confirm that data balancing techniques, feature selection, and text processing are crucial for model performance, with an even greater impact than the type of algorithm itself(Cam et al., 2024).

Despite these advancements, the issue of data imbalance remains a significant challenge in sentiment analysis, especially when dealing with real-world social media data. Imbalanced datasets can lead to biased model performance, particularly in minority classes. Previous studies have shown that SMOTE is effective in handling imbalanced data, this study proposes combining the comparison of three machine learning algorithms Random Forest, Logistic Regression, and SVM with a data balancing approach using the SMOTE method, as well as evaluating model performance using F1-Score metrics(Suandi et al., 2024).

To address these gaps, this study conducts a comparative analysis of four machine learning algorithms Naive Bayes, Logistic Regression, K-Nearest Neighbors, and Support Vector Machine in classifying public sentiment toward the police using data sourced from X. SMOTE is applied prior to model training to handle class imbalance, enabling a fair and rigorous comparison across all four algorithms. Rather than relying on accuracy alone, this study adopts F1-Score as the primary evaluation metric, given its superior sensitivity to class distribution disparities(Brownlee, 2016). Through this combination of multi-algorithm comparison, SMOTE-based resampling, and F1-Score evaluation, this study aims to identify the most effective approach for police sentiment classification in the Indonesian context. The findings are expected to contribute to the development of more reliable sentiment analysis models and serve as a practical reference for researchers and practitioners in selecting appropriate machine learning strategies for social media-based classification tasks.

METHOD

This research involves several stages, namely data collection, data preparation, and sentiment classification process using Naive Bayes algorithms, Logistic Regression, K-Nearest Neighbors, and Support Vector Machine to classify three classes of sentiment, namely positive, negative, and neutral. Furthermore, a comparison of prediction performance and evaluation of results were carried out to determine the algorithm with the most optimal performance, as shown in Figure 1. This research is focused on the classification of negative sentiment. This study uses Python version 3.11.4 and Jupyter Notebook as the development environment, as Python provides various libraries and frameworks that support data analysis and machine learning, such as scikit-learn, pandas, NumPy, and Matplotlib.

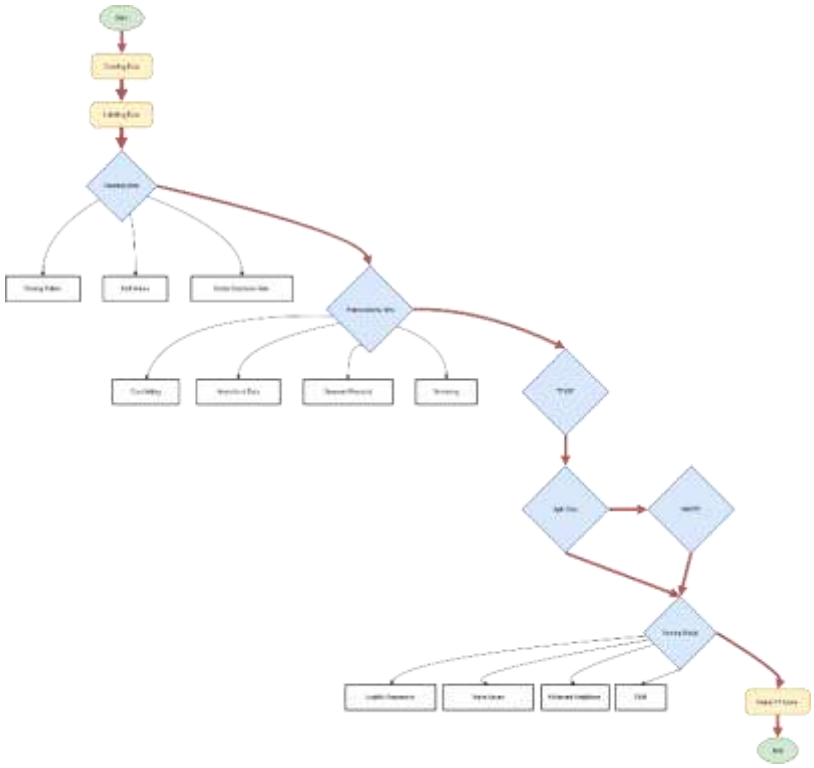

Figure 1. Research Method
Source: Illustration of research flow by researcher (2025)

Data Crawling

Data *crawling* is a technique in *machine learning* that is used to collect data contained in a website. This method works by retrieving information with *keywords* that have been entered by the *user* (Dikiyanti et al., 2021).

Data Cleaning

Data *cleaning* is an important technique for identifying and correcting data errors (Iqbal et al., 2024). In this study, whose data was collected through *crawling* techniques, a lot of data may contain *noise*, therefore data cleaning was carried out to search for missing values, null values and duplicate values.

Synthetic Minority Over-Sampling Technique (SMOTE)

In classification data, class imbalance occurs when the amount of data in the majority class is much greater than that in the minority class(Wang et al., 2021). This condition can cause the model to favor the majority class, thereby reducing the performance of minority class detection. One commonly used approach to address this problem is data-level resampling, specifically oversampling methods such as SMOTE *(Chen et al., 2024)*.

SMOTE works by creating synthetic samples in minority classes, rather than simply duplicating existing data. In general, for each minority sample, $\mathbf{x}_i$ *SMOTE* will look for k*the nearest neighbors* of the minority class. Then *SMOTE* selects one of the neighbors at random and generates new data with linear interpolation as follows$\mathbf{x}_{zi}$ (Taskiran et al., 2025)

$$\mathbf{x}_{new} = \mathbf{x}_i + \lambda(\mathbf{x}_i{}^{(k)} - \mathbf{x}_i)$$

with the following descriptions:

a. $\mathbf{x}_{new}$: is a new synthetic sample generated from the SMOTE process.
b. $\mathbf{x}_i$: is data selected from the minority class as the starting point for forming new samples.
c. $\mathbf{x}_i{}^{(k)}$: is one of the closest neighbors of x_i based on the k-nearest neighbors value.
d. λ : is a random number in the range of 0 to 1 that regulates the distance between x_i and its neighbors in the interpolation process.

Performance Metrics and Evaluation

This study will use F1-Score as the main performance metric to evaluate the performance of the Naive Bayes, Logistic Regression, and K-Nearest Neighbors algorithms. The evaluation will include an analysis of the strengths and weaknesses of each algorithm in the context of sentiment analysis.

The F1-Score is calculated using the following equation:

$$F1\ SCORE = 2 \text{ x } \frac{\text{Precission x Recall}}{\text{Precission + Recall}}$$

With precision and recall defined as follows(Hardjita et al., 2022):

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

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

Remarks:

a. *True Positives (TP)* is the amount of data that is predicted to be positive and correct according to the actual class.
b. *False Positives (FP)* are the amount of data that is predicted to be positive but actually belongs to the negative class.
c. *False Negatives (FN)* are the amount of data that is predicted to be negative but actually belongs to the positive class.

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Meanwhile, recall is the ratio of correctly predicted positive observations to all actual observations belonging to the positive class.

RESULT AND DISCUSSION

Dataset Class Distribution

The data used in this study was obtained through a crawling process on platform X using keywords "Polisi OR Kepolisian OR Kinerja Polisi OR Kinerja Kepolisian OR Polri". Data collection was conducted between March 13, 2024, and September 7, 2025, followed by a filtering process to select data in Indonesian. The crawling process yielded 1,319 data points, as shown in Table 1. The amount of data obtained was relatively limited due to the policies and access restrictions imposed by platform X on data collection activities. Next, a data cleaning process was carried out to remove duplicate data, resulting in a final number of 1,274 data used in this study.

Tabel 1 Data Hasil Crawling

No	Text	Sentimen
1	@HerumawanPA @merapi_uncover @FXHarminanto icj itu lebih kuat dari akun X karena anggotanya lebih kritis dari adminnya. pernah inget kasus kapolres bantul dibully di ICJ? besoknya polda bikin himbauan/ acara2 cari simpati tapi tiap kali polisi salah tetep dibully	Negatif
2	Gini kah lemes tuh? Udah terlalu malas buat ladenin ya Allah i wish that this country had good cops yg do their work as they should. Masalah gini harusnya ditangani polisi.. tapi.. polisi di sini tuh..	Negatif
3	@persebayaupdate Kalo yang melanggar polisi seperti menembak gas air mata berani nuntut ga?	Negatif
4	Fix polisi sekarang kurang duit. Sampe kek gini. Mending ga usah pake polisiÃ‚Â²an.	Negatif
5	@mardigu024 Tidak Ada Berita Baik Dari Polisi.	Negatif

The results of the analysis of data collected on platform X are presented in Figure 1. Based on the visualization of the diagram, the percentage distribution is divided into three categories. The largest distribution shows negative sentiment with a proportion of 70.5%. Positive sentiment follows in second place with a percentage of 23.7%. Meanwhile, neutral sentiment has the lowest percentage at 5.8%.

Figure 2. Percentage of Positive, Negative, and Neutral Sentiment

Figure 1. Wordcloud Sentiment
Source: Data visualization using Python (WordCloud) by researchers (2025)

Figure 2 shows a visualization of frequently occurring words presented in the form of a word cloud. The frequency of keywords is represented by the proportional font size, where larger words indicate higher occurrence. In this visualization, the most frequently appearing keywords are “polisi”, “polri”, and “masyarakat”. The dominance of the word “polisi” indicates that it is the primary focus of the dataset. In addition, other frequently occurring words include “kerja”, “hukum”, “oknum”, “anggota”, and “rakyat”. The presence of the words “oknum” and “salah” suggests that the discussions are largely related to controversial issues, violations, or criticism of institutional performance.

Data Preprocessing Results

The preprocessing stage includes cleaning mentions and URLs, normalizing words, removing punctuation, case folding, and removing stopwords and stemming. The preprocessing results show that sentence structure becomes simpler and focuses on meaningful words, thus improving the performance of the classification algorithm as seen in Table 2.

Table 2 Preprocessing Results Data,

No	Original	Preprocessing
1	@HerumawanPA @merapi_uncover @FXHarminanto icj itu lebih kuat dari akun X karena anggotanya lebih kritis dari adminnya. pernah inget kasus kapolres bantul dibully di ICJ? besoknya polda bikin himbauan/ acara2 cari simpati tapi tiap kali polisi salah tetep dibully	icj itu lebih kuat dari akun kali karena anggota lebih kritis dari adminnya pernah ingat kasus kapolres bantul dibully di icj besok polda bikin himbauan acara cari simpati tapi tiap kali polisi salah tetap dibully
2	Gini kah lemes tuh? Udah terlalu malas buat ladenin ya Allah i wish that this country had good cops yg do their work as they should. Masalah gini harusnya ditangani polisi.. tapi.. polisi di sini tuh..	begini kah lemes tuh sudah terlalu malas buat laden ya allah i wish that this country had good cops yang di their work as they should masalah begini harus tangan polisi tapi polisi di sini tuh
3	@persebayaupdate Kalo yang melanggar polisi seperti menembak gas air mata berani nuntut ga?	kalau yang langgar polisi seperti tembak gas air berani tuntut tidak
4	Fix polisi sekarang kurang duit. Sampe kek gini. Mending ga usah pake polisiÃ‚Â²an.	fix polisi sekarang kurang duit sampai kayak begini mending tidak usah pakai polisi
5	@mardigu024 Tidak Ada Berita Baik Dari Polisi.	begini kah lemes tuh sudah terlalu malas buat laden ya allah i wish that this country had good cops yang di their work as they should masalah begini harus tangan polisi tapi polisi di sini tuh

Source: Results of data preprocessing by researchers (2025)

Model Results without Using Smote.

Initial testing was performed on the original dataset without class imbalance handling. The results of the performance evaluation of each algorithm are shown in Table 3. Based on the F1-Score score, Logistic Regression showed the best performance with a score of 76.93%, followed by SVM at 75.59%, KNN at 75.48%, and Naive Bayes at 75.05%.

Tabel 1 Akurasi Algoritma tanpa SMOTE

	Naive Bayes	Logistic Regression	*KNN*	SVM
Accuracy	0.8039	0.8157	0.7961	0.8078
Precision	0.7633	0.7564	0.7480	0.7657
Recall	0.8039	0.8157	0.7961	0.8078
F1-Score	0.7505	**0.7693**	0.7548	0.7559

Source: Data processing results using Python (2025)

Based on Figure 3, the initial dataset shows a significant class imbalance, with 894 negative sentiment data, 225 positive sentiment, and 155 neutral sentiment data. After applying the SMOTE technique with sampling_strategy = 'auto' to the data train, the amount of data in each class became balanced, which was 709 data each. This shows that the auto strategy succeeds in oversampling the minority class while cleaning up data that has the potential to cause classification ambiguity.

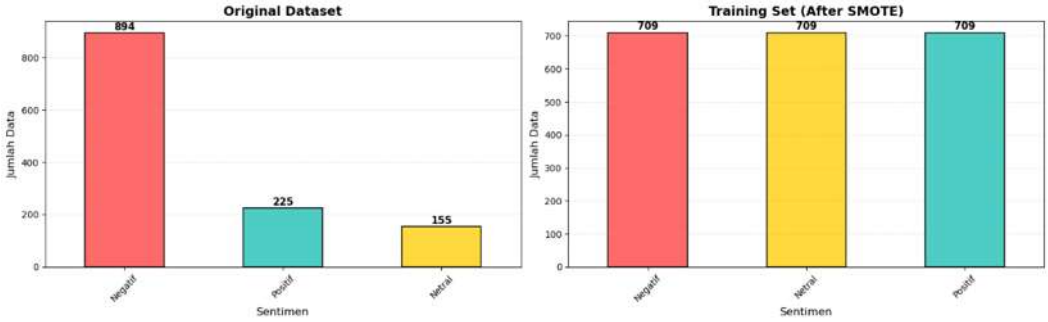

Figure 2 Data Distribution After Balancing with SMOTE
Source: Visualization of data distribution after SMOTE using Python by researchers (2025)

Tabel 2 Akurasi SMOTE

	Naive Bayes	Logistic Regression	*KNN*	SVM
Accuracy	0.7255	0.8078	0.6863	0.8118
Precision	0.8181	0.8106	0.7758	0.7929
Recall	0.7255	0.8078	0.6863	0.8118
F1-Score	0.7627	**0.8085**	0.7240	0.7752

Source: Data processing results after the application of the SMOTE method by the researcher (2025)

After data balancing using the SMOTE method, there was a change in the performance of each algorithm seen in Table 4. Logistic Regression showed an improvement in performance with an F1-Score of 80.85%, which is the highest value compared to other algorithms. SVM also experienced an increase with an F1-Score of 77.52%. Naive Bayes experienced a smaller increase, while KNN still showed the lowest performance among the four algorithms. This performance improvement shows that SMOTE is effective in improving minority class representation, so that the model becomes more balanced in studying positive and neutral sentiment patterns. Algorithms such as Logistic Regression and SVM based on decision boundary formation tend to benefit more from a balanced class distribution, thus being able to significantly increase recall values and F1-Scores.

Comparative Analysis of F1-Score by Class

Tabel 5 Perbandingan F1-Score per-kelas sebelum dan sesudah SMOTE

	Negatif	Netral	Positif
NB	0.88	0.00	0.58
LR	0.89	0.00	0.65
KNN	0.87	0.00	0.63
SVM	0.88	0.00	0.60
NB + SMOTE	0.83	0.22	0.73
LR + SMOTE	**0.89**	**0.27**	**0.74**
KNN + SMOTE	0.80	0.16	0.69
SVM + SMOTE	0.89	0.08	0.66

Source: Results of machine learning model analysis by researchers (2025)

The results of the performance evaluation per class before and after the implementation of SMOTE are presented in Table 5. This evaluation was conducted to look in more detail at the model's ability to classify each category of sentiment, especially minority classes. Prior to the implementation of SMOTE, all algorithms showed an F1-Score value of 0.00 in the neutral class. This suggests that the model is not able to recognize the neutral class at all and tends to classify the entire data into the majority class, i.e. the negative class.

This condition indicates model bias due to the unbalanced distribution of classes in the dataset. After data balancing using SMOTE, there was an increase in the model's ability to recognize minority classes, especially in neutral and positive classes. Logistic Regression showed the most significant increase in the neutral class with an F1-score of 0.27, followed by Naive Bayes of 0.22. However, the F1-Score in the neutral class is still lower than the negative and positive classes in the entire algorithm. This suggests that the classification of neutral sentiment is still a challenge, likely due to the characteristics of neutral texts that tend to be ambiguous and have similarities to both positive and negative classes.

CONCLUSION

From the results of data collection and cleaning, 1,274 data in Indonesian were obtained which were used as research datasets. The results of the evaluation before the implementation of SMOTE showed that all algorithms failed to recognize neutral classes with an F1-Score value of 0.00, which indicates a model bias towards the majority class. The results of the test of the model without SMOTE showed that Logistic Regression provided the best performance with an F1-score of 76.93%, followed by SVM of 75.59%, KNN of 75.48%, and Naive Bayes of 75.05%.

However, after applying SMOTE to the data train, there was an increase in the model's ability to recognize minority classes, especially in neutral and positive classes. Logistic Regression showed the most significant improvement with the neutral class F1-Score of 0.27 and the overall F1-Score of 80.85%, followed by Naive Bayes and KNN with moderate improvements, while SVM showed a relatively limited increase in the neutral class. These results suggest that data balancing plays an important role in improving the model's sensitivity to minority classes, although challenges in classifying neutral sentiment remain. Based on the overall results, the combination of Logistic Regression and SMOTE is the most optimal approach in this study to classify public sentiment towards policing on social media data, both in terms of overall performance and ability to recognize minority classes.

REFERENCE

Al Mustaqim, D., Hakim, F. A., Atfalina, H., & Fatakh, A. (2024). Peran media sosial sebagai sarana partisipasi warganet dalam mewujudukan keadilan dan akuntabilitas penegakan hukum di Indonesia. *Journal of Multidisciplinary Research and Development*, *1*(1), 53–66.

Bahtiar, S. A. H., Dewa, C. K., & Luthfi, A. (2023). Comparison of Naive Bayes and logistic regression in sentiment analysis on marketplace reviews using rating-based labeling. *Journal of Information Systems and Informatics*, *5*(3), 915–927.

Brownlee, J. (2016). *Machine Learning Mastery With Python Understand Your Data, Create Accurate Models and Work Projects End-To-End*.

Cam, H., Cam, A. V., Demirel, U., & Ahmed, S. (2024). Sentiment analysis of financial Twitter posts on Twitter with the machine learning classifiers. *Heliyon*, *10*(1). https://doi.org/10.1016/j.heliyon.2023.e23784

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, *57*(6). https://doi.org/10.1007/s10462-024-10759-6

Dikiyanti, T. D., Rukmi, A. M., & Irawan, M. I. (2021). Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm. *Journal of Physics: Conference Series*, *1821*(1), 12054.

Handika, Y., Hanif, I. F., & Hasan, F. N. (2024). Analysis of Public Sentiment Towards POLRI's Performance using Naive Bayes and K-Nearest Neighbors. *IJID (International Journal on Informatics for Development)*, *13*(1), 386–399.

Hardjita, P. W., Nurochman, & Hidayat, R. (2022). Sentiment Analysis of Tweets on Prakerja Card using Convolutional Neural Network and Naive Bayes. *IJID (International Journal on Informatics for Development)*, *10*(2), 82–91. https://doi.org/10.14421/ijid.2021.3007

Iqbal, M., Afdal, M., & Novita, R. (2024). Implementasi Algoritma Support Vector Machine Untuk Analisa Sentimen Data Ulasan Aplikasi Pinjaman Online di Google Play Store: Implementation of Support Vector Machine Algorithm for Sentiment Analysis of Online Loan Application Review Data on Google Play Store. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(4), 1244–1252.

Julizar, A., & Sulaeman, M. K. (2025). Evaluation of Logistic Regression and Random Forest Algorithms for Hate Speech Identification. In *International Journal of Scientific Engineering and Science* (Vol. 9, Number 5). http://ijses.com/

Matarat, K., Mingmuang, C., & Charoenrat, W. (2024). A Comprehensive Performance Analysis of Supervised Machine Learning Techniques for Sentiment Analysis. *International Journal of Computer Applications*, *186*(7), 975–8887. https://doi.org/10.5120/ijca2024923409

Nasution, M. R. A., & Hayaty, M. (2019). Perbandingan akurasi dan waktu proses algoritma K-NN dan SVM dalam analisis sentimen twitter. *Jurnal Informatika*, *6*(2), 226–235.

Permatasari, P. A., Linawati, L., & Jasa, L. (2021). Survei Tentang Analisis Sentimen Pada Media Sosial. *Majalah Ilmiah Teknologi Elektro*, *20*(2), 177.

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach. *Social Network Analysis and Mining*, *13*(1). https://doi.org/10.1007/s13278-023-01030-x

Sabir, A., Ali, H. A., & Aljabery, M. A. (2024). ChatGPT tweets sentiment analysis using machine learning and data classification. *Informatica*, *48*(7).

Suandi, F., Anam, M. K., Firdaus, M. B., Fadli, S., Lathifah, L., Yumami, E., Saleh, A., & Hasibuan, A. Z. (2024). *Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms* (pp. 126–138). https://doi.org/10.2991/978-94-6463-620-8_10

Syahrohim, I., Saputra, S. D., Saputra, R. W., Pranatawijaya, V. H., & Priskila, R. (2024). Perbandingan analisis sentimen setelah pilpres 2024 di Twitter menggunakan algoritma machine learning. *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(2). https://doi.org/10.23960/jitet.v12i2.4249

Taskiran, S. F., Turkoglu, B., Kaya, E., & Asuroglu, T. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Scientific Reports*, *15*(1), 21631.

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*, *9*, 64606–64628. https://doi.org/10.1109/ACCESS.2021.3074243