Volume 11 Issue 1 Year 2026 | Page 150-159 ISSN: 2527-9866
Received: **28-10-2025** | *Revised:* **18-12-2025** | *Accepted:* **26-01-2026**

Comparative Evaluation of Preprocessing Techniques in Twitter Sentiment Analysis for Indonesia's 2024 Regional Elections

Asro[1], Solihin[2]
[1,2] PGRI Banten Polytechnic, Banten, Indonesia, 42161
e-mail: asrohaun6@gmail.com[1], tugasqu15@gmail.com[2]
*Correspondence: asroharun6@gmail.com

Abstract: *The rapid expansion of social media has positioned Twitter as a critical platform for capturing public opinion during political events, including Indonesia's 2024 Regional Elections. This study investigates the impact of preprocessing strategies and class balancing on the performance of sentiment analysis models applied to election-related tweets. An initial dataset of 9,096 tweets was collected and refined into 6,202 relevant entries from 2024–2025 through text cleaning, normalization, tokenization, and duplicate removal. Sentiment distribution analysis reveals a dominance of positive sentiment (58.4%), followed by negative (33.6%) and neutral (8.0%) expressions. Two classical machine learning classifiers—Naïve Bayes and Logistic Regression—were implemented using TF–IDF feature representation. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training data, and hyperparameter optimization was conducted using GridSearchCV. Model evaluation employed an 80/20 train–test split with accuracy, precision, recall, F1-score, and confusion matrices as performance metrics. Experimental results indicate that Logistic Regression combined with SMOTE and hyperparameter tuning achieved the highest accuracy of 93.08%, outperforming Naïve Bayes. The findings confirm that carefully designed preprocessing pipelines and class balancing significantly enhance the reliability of sentiment classification in political social media analysis.*

Keywords: *Sentiment Analysis, Twitter, Regional Elections 2024, Naïve Bayes, Logistic Regression*

1. Introduction

The rapid growth of social media has transformed platforms such as Twitter into primary channels for public expression, particularly during political events. In the Indonesian context, the 2024 Simultaneous Regional Elections (Pilkada Serentak 2024) generated extensive online discourse, positioning Twitter as a valuable data source for examining public opinion through sentiment analysis. Given the massive volume and unstructured nature of social media text, preprocessing stages play a crucial role in enabling machine learning models to effectively capture patterns in political communication.

2. Literature Review

Several recent studies have explored sentiment analysis of political events in Indonesia using Twitter data. Research by Ma'aly et al. [1] applied deep learning with multi-label classification to analyze online reviews related to Indonesian presidential elections, demonstrating the potential of neural approaches in capturing complex sentiment dimensions. Similarly, Sembiring and Dewa [2] evaluated various classifiers including Naïve Bayes, Support Vector Machine (SVM), LSTM, and GRU on election-related tweets and reported that Naïve Bayes achieved the highest accuracy among traditional models. Cahyanti et al. [3] compared multiple classification algorithms, such as Naïve Bayes, SVM, Random Forest, and neural networks, to analyze sentiment toward Indonesian presidential candidates, highlighting the strong performance of ensemble and neural-based approaches. In addition, Wibowo et al. [4]

employed a Long Short-Term Memory (LSTM) model to classify sentiment related to the 2024 General Election, achieving an accuracy of 84.3% after applying preprocessing and oversampling techniques. Firdaus et al. [5] further contributed by releasing a large-scale dataset for Indonesian presidential election sentiment analysis, emphasizing the importance of dataset quality in reliable opinion mining.

3. Research Gap and Contribution

Despite the growing body of research on election-related sentiment analysis, relatively few studies have systematically examined the combined impact of preprocessing strategies and class balancing techniques on model performance, particularly in the context of Indonesia’s 2024 Simultaneous Regional Elections. Most prior works focus primarily on algorithm comparison or deep learning performance, with limited attention to how preprocessing design choices influence classification reliability. This study addresses this gap by conducting a comparative evaluation of preprocessing techniques applied to Twitter sentiment analysis related to Pilkada Serentak 2024. An initial dataset of 9,096 tweets was refined into 6,202 relevant entries from 2024–2025 through normalization, tokenization, and duplicate removal. Sentiment distribution analysis shows a dominance of positive expressions (58.4%), followed by negative (33.6%) and neutral (8.0%) sentiments. Two classical classifiers Naïve Bayes (NB) and Logistic Regression (LR) were implemented using TF–IDF features. To mitigate class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training data, accompanied by hyperparameter optimization using GridSearch. The results demonstrate that the integration of well-designed preprocessing pipelines and balancing strategies significantly improves classification accuracy and model generalization.

In recent years, Twitter has become a key medium for political discourse, enabling large-scale and real-time analysis of public opinion beyond conventional survey methods [6]. While several studies have examined election sentiment analysis using machine learning and deep learning techniques [7], comprehensive evaluations focusing on preprocessing and class balancing remain limited. Therefore, the main contributions of this study are as follows:
(i) a systematic evaluation of preprocessing techniques combined with SMOTE,
(ii) benchmarking Naïve Bayes and Logistic Regression under multiple experimental configurations, and (iii) providing methodological insights to enhance sentiment analysis pipelines for socio-political contexts in Indonesia. Unlike prior studies that primarily focus on algorithmic comparison, this study introduces a controlled experimental design that systematically evaluates the combined impact of preprocessing strategies, class balancing using SMOTE, and hyperparameter optimization, while explicitly preventing information leakage in sentiment analysis of Indonesia’s 2024 Simultaneous Regional Elections.

4. Methods

This study employs a quantitative experimental approach to evaluate the impact of preprocessing strategies and class balancing techniques on Twitter sentiment analysis related to Indonesia’s 2024 Simultaneous Regional Elections. The overall research workflow follows the CRISP-DM framework, encompassing data acquisition, preprocessing, feature extraction, modeling, and evaluation, as illustrated in Figure 1.

A. Requirement Analysis

All experiments in this study were conducted on a Lenovo laptop running Windows 11 x64 (build 26100), equipped with an Intel® Core™ i7 processor and 16 GB of RAM. The development environment utilized Visual Studio Code version 1.103.2, with Node.js 22.17.0 employed for Twitter data crawling and Python (≥ 3.10) used for the entire sentiment analysis pipeline and model evaluation. The software stack was designed to support medium- to large-scale text processing, including data manipulation, NLP preprocessing, TF–IDF feature extraction, sentiment classification, class balancing, and result visualization. The hardware configuration provides sufficient memory to execute TF–IDF, SMOTE (applied only to the training set), and hyperparameter optimization using GridSearchCV without requiring GPU acceleration.

Tables I and II summarize the hardware and software requirements and indicate their usage across the research workflow shown in Figure 1.

Table I. Hardware Requirement

No	Component	Specification
1	CPU	Intel® Core™ i7 (64-bit)
2	RAM	16 GB
3	Storage	≥ 500 GB SSD
4	Device	Lenovo laptop

Table II. Software Stack

No	Category	Tools / Versions
1	IDE & Runtime	VS Code 1.103.2; Python ≥ 3.10; Node.js 22.17.0
2	Data & NLP	pandas; NumPy; NLTK
3	Feature Engineering	TF–IDF; train_test_split
4	Modeling & Balancing	Naïve Bayes; Logistic Regression; GridSearchCV; SMOTE
5	Visualization	matplotlib; seaborn; wordcloud; openpyxl

Note: SMOTE is applied only to the training set to prevent information leakage.

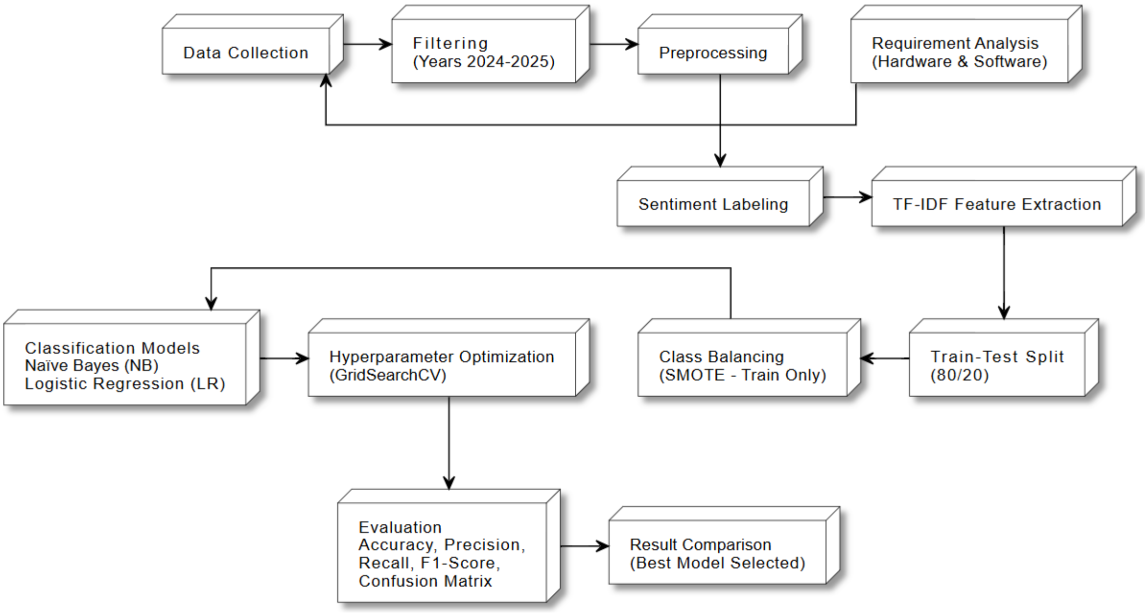


Figure 1. Research Methodology Workflow

Figure 1 illustrates the overall research workflow, starting from requirement analysis, followed by data collection, preprocessing, feature extraction using TF–IDF, model training under multiple experimental configurations (with and without SMOTE and GridSearch), and performance evaluation on a held-out test set. The workflow is designed to ensure data integrity, prevent information leakage, and enable a fair comparison between classification models.

B. Data Collection

Twitter data were collected through a crawling process using election-related keywords associated with the 2024 Regional Elections (Pilkada 2024). The initial corpus consisted of

9,096 tweets. To ensure temporal relevance and contextual consistency, only tweets published during the period 2024–2025 were retained. After filtering, a total of 6,202 tweets were deemed suitable for subsequent analysis [10]. The collected dataset represents spontaneous public expressions related to electoral processes, candidates, and election institutions, making it a reliable source for political sentiment analysis.

C. Data Preprocessing

Preprocessing [13} was conducted to transform noisy and unstructured social media text into a consistent representation suitable for machine learning models. The preprocessing pipeline consisted of the following stages:

1. **Text Cleaning:** Removal of URLs, user mentions, hashtags (while preserving the lexical content), emojis, special characters, excessive whitespace, and non-linguistic artifacts.
2. **Case Folding:** Conversion of all characters to lowercase to reduce lexical redundancy.
3. **Tokenization:** Segmentation of text into individual word tokens.
4. **Normalization:** Standardization of informal and non-standard spellings (e.g., slang and abbreviations) into their canonical forms. Approximately 2.2% of tokens were affected by this process.
5. **Duplicate Removal:** Elimination of duplicated tweets to minimize bias.
6. **Stopword Removal and Stemming:** Removal of function words and reduction of tokens to their root forms using an Indonesian stemming algorithm.

These steps were essential to improve linguistic consistency and feature quality, thereby enhancing model performance.

D. Sentiment Labeling

Each tweet was categorized into one of three sentiment classes: **Positive**, **Negative**, or **Neutral**. After preprocessing and filtering, the sentiment distribution showed a predominance of positive sentiment (58.4%), followed by negative (33.6%) and neutral (8.0%). This class imbalance reflects common characteristics of political discourse on social media and motivated the application of class balancing techniques in the modeling stage.

E. Feature Extraction

Textual features were extracted using the Term Frequency–Inverse Document Frequency (TF–IDF) method. Unigrams and bigrams (1–2 grams) were employed to capture both individual terms and short contextual patterns. The feature space was limited to a maximum of 90,000 features to balance representational richness and computational efficiency. TF–IDF was selected due to its effectiveness in representing short texts and its proven performance in sentiment classification tasks involving high-dimensional sparse data [11].

F. Experimental Design and Modeling

To prevent information leakage, the dataset was split into training and testing sets using an 80/20 ratio. Two classical machine learning classifiers were evaluated:

- **Naïve Bayes (NB):** Implemented using the Multinomial Naïve Bayes algorithm, which is widely used in text classification due to its simplicity and computational efficiency.
- **Logistic Regression (LR):** A linear classification model that maps TF–IDF features to sentiment class probabilities using a logistic (sigmoid) function.

To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied only to the training data, ensuring that the test set remained untouched [14].

G. Hyperparameter Optimization

Hyperparameter tuning was performed using GridSearchCV with 3-fold cross-validation. The parameter search space was defined as follows:

- **Naïve Bayes:** α ∈ {0.1, 0.5, 1.0, 2.0}, fit_prior ∈ {True, False}
- **Logistic Regression:** C ∈ {0.01, 0.1, 1, 10}, solver ∈ {liblinear, saga}, max_iter ∈ {1000, 5000}

These configurations reflect commonly adopted best practices for text classification, balancing model flexibility and computational feasibility on CPU-based systems [12], [15].

H. Evaluation Metrics

Model performance was evaluated on an untouched test set using the following metrics:

- Accuracy
- Precision
- Recall
- Macro-averaged F1-score
- Confusion Matrix for class-level error analysis

In addition, four experimental configurations were compared to assess the contribution of SMOTE and hyperparameter tuning to overall model performance.

5. Results and Discussion

A. Research Questions and Discussion Framework

This section is structured to address three main research questions:

RQ1. What are the characteristics of the Twitter corpus from 2024–2025 after cleaning and normalization?

RQ2. To what extent do class balancing (SMOTE) and hyperparameter tuning (GridSearch) affect sentiment classification performance?

RQ3. Between two classical algorithms Naïve Bayes (NB) and Logistic Regression (LR) which model demonstrates the best generalization on a held-out test set?

The discussion proceeds systematically from corpus profiling (RQ1), to comparative model performance (RQ2), and finally to error analysis of the best model (RQ3). Visual evidence (figures and tables) is provided to support the interpretation.

B. Corpus Profile and Lexical Signals (RQ1)

The initial crawling process yielded **9,096 tweets** related to the 2024 Simultaneous Regional Elections. After year-based filtering (**2024–2025**), the dataset was reduced to **6,202 relevant tweets**, ensuring temporal alignment with the electoral context. This filtering strategy is consistent with prior studies that employ Twitter as a primary source for political discourse analysis in Indonesia. Following preprocessing, the sentiment distribution shows a dominance of **positive sentiment (58.4%)**, followed by **negative (33.6%)** and **neutral (8.0%)**, as illustrated in **Figure 2**. This pattern suggests generally optimistic public sentiment toward the electoral process, while the substantial negative share reflects critical opinions inherent to democratic competition.

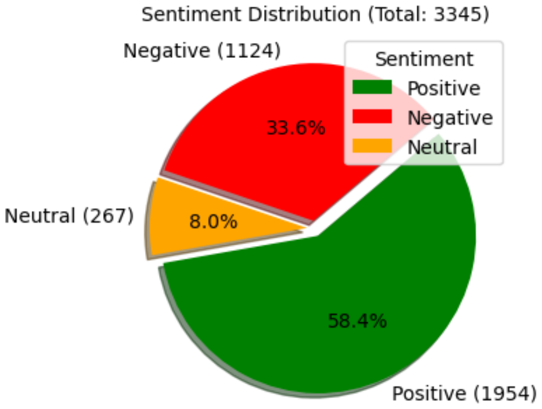


Figure 2. Sentiment Distribution (Positive, Negative, Neutral)

Lexical analysis using word clouds and top-frequency terms (**Figure 3**) reveals dominant keywords such as *pilkada*, *serentak*, *kpu*, *jakarta*, *candidate*, and *governor*. Their consistent appearance across sentiment classes indicates strong topical coherence centered on the election process, institutions, and candidates.

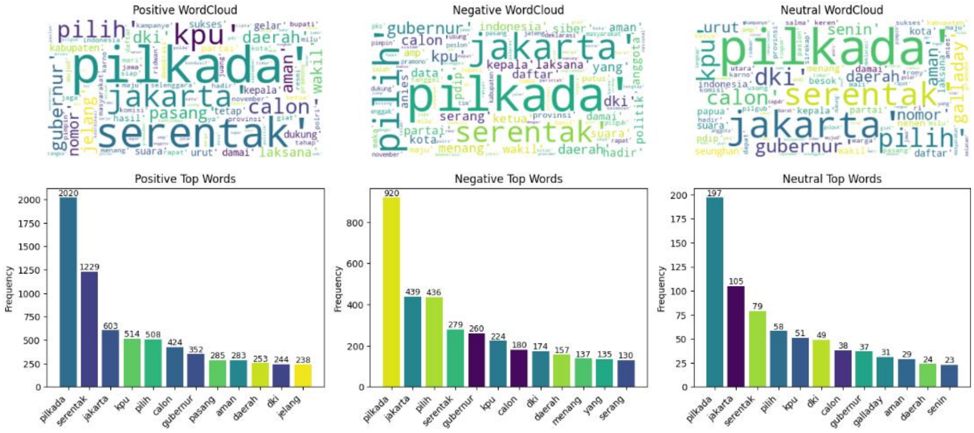


Figure 3. Word Clouds and Top Terms per Sentiment Class

Text normalization plays an important role in handling spelling variation and informal language. Approximately **2.2% of tokens** were modified through normalization (**Figure 4**). Although relatively small, this step improves linguistic consistency and feature quality, in line with findings reported in prior Indonesian and Malay text processing studies [16].

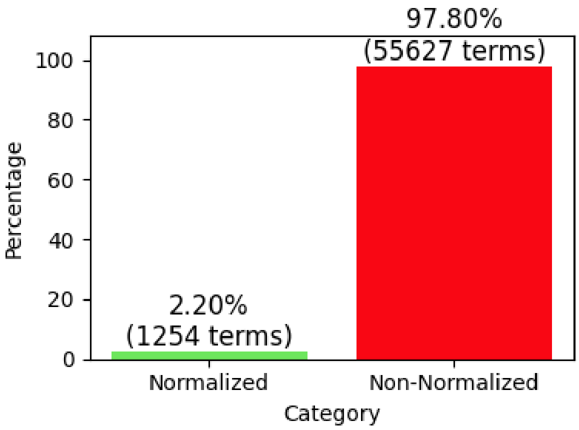


Figure 4. Ratio of Normalized (~2.2%) vs. Non-Normalized Tokens

In addition, **Figure 5** illustrates the distribution of the initially collected 9,096 tweets across different temporal categories. Tweets published outside the 2024–2025 period (1,835 entries) were excluded, while 6,202 tweets from 2024 and 1,059 tweets from 2025 were retained for further analysis.

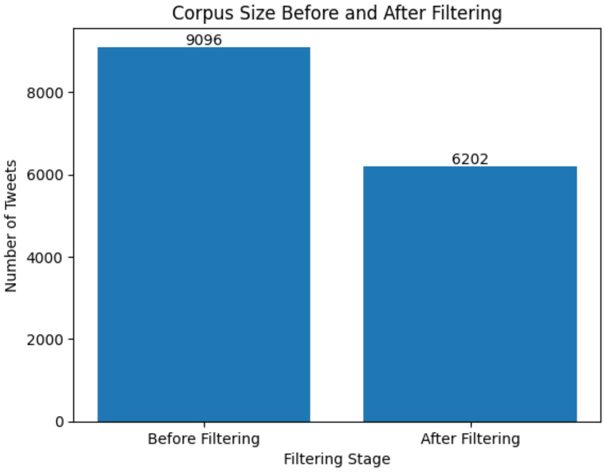


Figure 5. Corpus Size Before (9,096) and After (6,202) Filtering

C. Comparative Accuracy Across Experimental Settings (RQ2 & RQ3)

Model evaluation was conducted using NB and LR under four experimental settings:

- S_1: No SMOTE, no GridSearch
- S_2: No SMOTE, with GridSearch
- S_3: SMOTE applied (training set only), no GridSearch
- S_4: SMOTE applied (training set only), with GridSearch

All experiments used identical **TF–IDF** features [21]. The results indicate consistent performance improvements as model configurations become more advanced. First, SMOTE yields substantial gains. For Naïve Bayes, accuracy increases from 62.68% to 83.65% (+20.97 percentage points). For Logistic Regression, combining SMOTE with GridSearch achieves the highest accuracy of 93.08%, the best result among all configurations. Second, hyperparameter tuning further amplifies the effect of SMOTE. Naïve Bayes improves from 83.65% to 87.00% (+3.35 points), while Logistic Regression increases from 77.76% to 93.08% (+15.32 points relative to S_2) [23]. This highlights the importance of parameters such as regularization strength (*C*) and solver choice for linear models. Third, Logistic Regression demonstrates superior generalization. Under S_4, LR outperforms NB by 6.08 percentage points, consistent with prior evidence that TF–IDF-based linear models are stable and effective for high-dimensional text data [17].

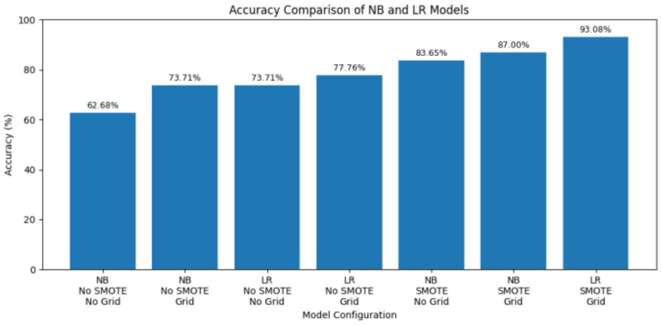


Figure 6. Accuracy Comparison of NB and LR Across S_1–S_4

Table 3. Accuracy by Model and Experimental Setting

No	Model	Experimental Setting	Accuracy (%)
1	Naïve Bayes (NB)	S_1 – No SMOTE, No GridSearch	62.68
2	Naïve Bayes (NB)	S_2 – No SMOTE, GridSearch	73.71
3	Logistic Regression (LR)	S_1 – No SMOTE, No GridSearch	73.71
4	Logistic Regression (LR)	S_2 – No SMOTE, GridSearch	77.76
5	Naïve Bayes (NB)	S_3 – SMOTE, No GridSearch	83.65
6	Naïve Bayes (NB)	S_4 – SMOTE, GridSearch	87.00
7	Logistic Regression (LR)	S_4 – SMOTE, GridSearch	93.08

As summarized in Table I, Logistic Regression with SMOTE and GridSearch (S_4) achieved the highest accuracy of 93.08%, outperforming all other experimental configurations. This result indicates that the combination of class balancing and hyperparameter optimization significantly enhances the generalization capability of linear classifiers when applied to high-dimensional TF–IDF feature spaces.

D. Accuracy Trends Across Configurations (RQ2)

To visualize performance progression, **Figure 7** presents accuracy trends from S_1 to S_4. Accuracy increases steadily with configuration complexity, with the most pronounced improvement occurring when SMOTE is introduced (S_3) and peaking when combined with GridSearch (S_4). This trend confirms that balanced training data and optimized hyperparameters are crucial for achieving optimal performance in short-text sentiment analysis on Twitter.

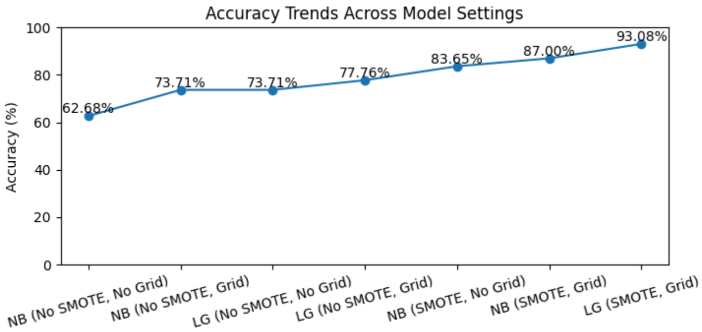
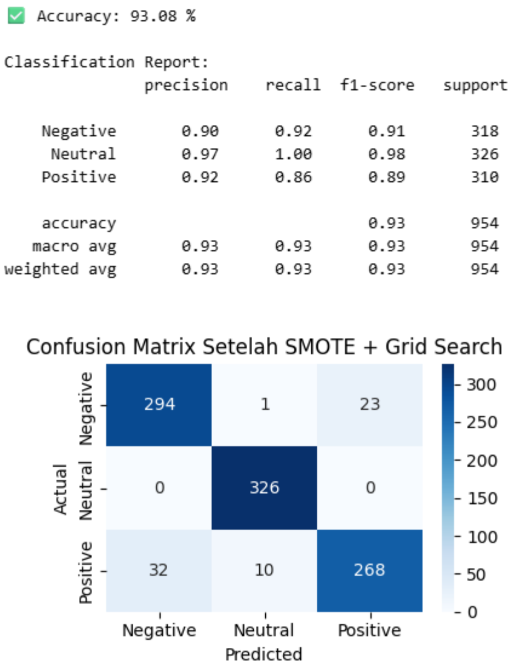


Figure 7. Accuracy Trend from S_1 to S_4

E. Error Analysis of the Best Model (RQ3)

On a held-out test set of 954 tweets, Logistic Regression with SMOTE and GridSearch achieves an overall accuracy of 93%, with Precision = 0.93, Recall = 0.93, and Macro F1-score = 0.93 (Figure 8). The Neutral class is the easiest to identify (Precision 0.97; Recall 1.00), indicating reliable detection of non-opinionated tweets. Most misclassifications occur between Positive and Negative classes, typically in tweets containing sarcasm, irony, or emotionally ambiguous expressions well-known challenges in social media sentiment analysis [18], [19].



```
✅ Accuracy: 93.08 %

Classification Report:
              precision    recall  f1-score   support

    Negative       0.90      0.92      0.91       318
     Neutral       0.97      1.00      0.98       326
    Positive       0.92      0.86      0.89       310

    accuracy                           0.93       954
   macro avg       0.93      0.93      0.93       954
weighted avg       0.93      0.93      0.93       954
```

Figure 8. Confusion Matrix and Class-wise Metrics for LR (SMOTE + GridSearch)

Table 4. Class-wise Performance Metrics on the Test Set

Class	Precision	Recall	F1-Score	Support
Negative	0.90	0.92	0.91	318
Neutral	0.97	1.00	0.98	326
Positive	0.92	0.86	0.89	310
Overall Accuracy	—	—	0.93	954

4. Conclusions

This study investigated the impact of preprocessing strategies and class balancing techniques on Twitter sentiment analysis related to Indonesia’s 2024 Simultaneous Regional Elections. Using a refined dataset of 6,202 tweets collected during the 2024–2025 period, the performance of two classical machine learning models Naïve Bayes and Logistic Regression was evaluated based on TF–IDF feature representation. The experimental results demonstrate that preprocessing plays a critical role in improving sentiment classification performance. Text cleaning, normalization, tokenization, duplicate removal, and stemming significantly enhanced data quality and reduced noise inherent in social media content. Furthermore, addressing class imbalance through the application of the SMOTE applied exclusively to the training data resulted in substantial performance improvements for both classifiers. Among the evaluated models, Logistic Regression combined with SMOTE and hyperparameter optimization achieved the highest accuracy of 93.08%, consistently outperforming Naïve Bayes across all experimental configurations. These findings confirm that classical machine learning models remain highly competitive for sentiment analysis tasks when supported by well-designed preprocessing pipelines and balanced training data. Despite these promising results, this study has several limitations. First, nested cross-validation and formal statistical significance testing were not applied, which may affect the robustness of comparative conclusions. Second, the analysis was limited to Twitter data and did not address complex linguistic phenomena such as sarcasm and irony, which remain challenging in sentiment classification. Future research may extend this work by incorporating transformer-based models such as IndoBERT to capture deeper contextual semantics, expanding data sources to other social media platforms such as YouTube and TikTok, and integrating sarcasm detection techniques. These extensions are expected to further enhance the reliability and generalizability of sentiment analysis in political and socio-digital contexts in Indonesia. The novelty of this work lies in its systematic experimental design that isolates the contribution of preprocessing, SMOTE, and hyperparameter tuning on classical sentiment classification models within a real-world electoral context.

References

[1] A. N. Ma'aly, D. Pramesti, A. D. Fathurahman, and H. Fakhrurroja, "Exploring Sentiment Analysis for the Indonesian Presidential Election Through Online Reviews Using Multi-Label Classification with a Deep Learning Algorithm," *Information (Switzerland)*, vol. 15, no. 11, Nov. 2024, doi: 10.3390/info15110705.

[2] A. R. Sembiring and C. K. Dewa, "Sentiment Analysis On Indonesian Tweets about the 2024 Election," *Sinkron*, vol. 9, no. 1, pp. 413–422, Jan. 2025, doi: 10.33395/sinkron.v9i1.14481.

[3] R. Cahyanti, D. N. Maftuhah, A. B. Santoso, and I. Budi, "Twitter Sentiment Analysis Towards Candidates of the 2024 Indonesian Presidential Election," *Jurnal RESTI*, vol. 8, no. 4, pp. 516–524, Aug. 2024, doi: 10.29207/resti.v8i4.5839.

[4] A. W. W, H. H. Andana, J. Zeniarja, and A. Febriyanto, "Sentiment Analysis of the 2024 General Election Through Twitter using Long-Short-Term Memory Algorithm," *Journal of Informatics and Web Engineering*, vol. 4, no. 2, pp. 387–400, Jun. 2025, doi: 10.33093/jiwe.2025.4.2.25.

[5] H. L. A. Asro, A. Wicaksono, and P. Herwanto, "Strategi Dinamis Menggunakan MongoDB untuk Analisis Sentimen terhadap Komentar YouTube Pilkada Gubernur Indonesia 2024 Dynamic Strategy Using MongoDB for Sentiment Analysis of YouTube Comments on the 2024 Indonesian Gubernatorial Elections."

[6] Olayemi Mikail Olaniyi, SalaudeenTajudeen Muhamad, Emmanuel Daniya, Abdullahi Mohammed Ibrahim, and Taliha Abiodun Folorunso, "Development of maize plant dataset for intelligent recognition and weed control," 2023, doi: 10.17632/jjbfcckrsp.2.

[7] N. Sulistianingsih and I. N. Switrayana, "Enhancing Sentiment Analysis for the 2024 Indonesia Election Using SMOTE-Tomek Links and Binary Logistic Regression," *International Journal of Education and Management Engineering*, vol. 14, no. 3, pp. 22–32, Jun. 2024, doi: 10.5815/ijeme.2024.03.03.

[8] D. Ziaul, H. Iskandar, and Y. Ramdhani, "Optimasi Parameter Random Forest menggunakan Grid Search Untuk Analisis Time series," 2023.

[9] A. Robi Padri, A. Asro, and I. Indra, "Classification of Traffic Congestion in Indonesia Using the Naive Bayes Classification Method," *Journal of World Science*, vol. 2, no. 6, pp. 877–888, Jun. 2023, doi: 10.58344/jws.v2i6.285.

[10] H. L. A. Asro, A. Wicaksono, and P. Herwanto, "Strategi Dinamis Menggunakan MongoDB untuk Analisis Sentimen terhadap Komentar YouTube Pilkada Gubernur Indonesia 2024 Dynamic Strategy Using MongoDB for Sentiment Analysis of YouTube Comments on the 2024 Indonesian Gubernatorial Elections."

[11] D. Dey *et al.*, "The proper application of logistic regression model in complex survey data: a systematic review," *BMC Med Res Methodol*, vol. 25, no. 1, Dec. 2025, doi: 10.1186/s12874-024-02454-5.

[12] Tedyyana, A., Ratnawati, F., & Kurniati, R. (2019). Rancangan sistem informasi penelitian dan pengabdian masyarakat Politeknik Negeri Bengkalis menggunakan metode UML (Unified Modeling Language). *Sistemasi: Jurnal Sistem Informasi*, *8*(3), 413-423.

[13] A. Azizan *et al.*, "Normalization of Malay Noisy Text in Social Media using Levenshtein Distance and Rule-Based Techniques," 2024, doi: 10.47772/IJRISS.

[14] V. Sinaj and H. Curma, "Handling Data Imbalance in Text Classification," 2023. [Online]. Available: https://www.researchgate.net/publication/384767044

[15] B. Moores and V. Mago, "A Survey on Automated Sarcasm Detection on Twitter," Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.02516

[16] N. S. Jonnala *et al.*, "Leveraging hybrid model for accurate sentiment analysis of Twitter data," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-09794-2.