Research Horizon

ISSN: 2808-0696 (p), 2807-9531 (e)

Research Horizon
Volume: 06
Issue: 05
Year: 2026
Page: 2409-2422

Citation:
Hidayatullah, A., & Abidin, Z. (2026). Evaluating text preprocessing effects on Indonesian sentiment classification of Islamic boarding school messages. *Research Horizon*, *6*(5), 2409-2422.

Article History:
Received: September 23, 2026
Revised: October 12, 2026
Accepted: October 16, 2026
Online since: October 25, 2026

Evaluating Text Preprocessing Effects on Indonesian Sentiment Classification of Islamic Boarding School Messages

Arfan Hidayatullah[1*], Zainal Abidin[1]

[1] Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia

[*] *Corresponding author: Arfan Hidayatullah (arfannj10@gmail.com)*

Abstract

Islamic boarding schools manage large volumes of digital messages from students' guardians and communities, making it difficult to identify stakeholder concerns. This study evaluated six text-preprocessing steps in Indonesian three-class sentiment classification and compared Multinomial Naive Bayes (NB) with Linear Support Vector Machine (SVM). The experiment used 2,257 messages from a pesantren communication channel. After identifiers and incomplete records were removed, three annotators labeled messages as positive, neutral, or negative; a fourth resolved cases without a majority. Agreement was measured using Fleiss' kappa. Steps were examined through leave-one-step-out ablation using unigram TF-IDF, five-fold stratified cross-validation, held-out testing, learning curves, and Wilcoxon tests. Cross-validation accuracy ranged from 0.8155 to 0.8238 for NB and 0.8100 to 0.8133 for SVM. On full-pipeline test data, accuracy was 0.8540 for NB and 0.8208 for SVM. Removing individual steps produced no substantial gains. No statistically significant difference between classifiers was detected. NB performed better overall but tended to overpredict neutral messages; both models struggled with this class. The findings support the full pipeline as a baseline and identify annotation consistency and neutral-class handling as priorities for further improvement.

Keywords

Ablation Study, Indonesian Language, Naive Bayes, Sentiment Analysis, Support Vector Machine, Text Preprocessing.

1. Introduction

Islamic boarding schools (*pondok pesantren*) are religious educational institutions that interact continuously with students' guardians and surrounding communities. These interactions generate administrative inquiries, information requests, service complaints, suggestions, and appreciation for educational programs. Digital communication has increased the volume of written messages administrators must manage. Research on public perceptions and negative issues within Islamic boarding school communities has discussed similar stakeholder concerns (Kamal et al., 2024). Manually reading and prioritizing every message is time-consuming alongside other educational and administrative responsibilities. As a result, administrators may overlook or delay complaints that need attention. An automated approach could organize messages and flag issues that require follow-up, helping administrators manage stakeholder communication more efficiently.

Sentiment analysis identifies opinions, emotions, and attitudes in text and classifies them by polarity (Liu, 2012). In *pesantren* communication, it can distinguish positive, neutral, and negative messages and identify dissatisfaction and service-related concerns. Previous studies by Geni et al. (2023), Kamal et al. (2024), and Ahmadian et al. (2024) have applied sentiment analysis to Indonesian public discussions, institutional issues, and stakeholder opinions. However, institutional messages differ from informal social media posts. They may contain formal expressions, greetings, religious terminology, and polite language that shape how opinions are conveyed. Therefore, sentiment classification in this setting needs to account for the linguistic characteristics and communicative purposes of institutional messages.

Text preprocessing is an important factor in classification because it transforms raw text into a structured representation. Common procedures include case folding, punctuation removal, tokenization, stopword removal, stemming, and normalization of non-standard words (Vijayarani et al., 2015). Preprocessing can improve or reduce classification performance depending on language, dataset, and algorithm (Haddi et al., 2013; Uysal & Gunal, 2014). This matters in Indonesian, where prefixes, suffixes, and confixes produce morphological variations. Stemming is commonly used to reduce these variations and consolidate words into root forms (Tala, 2003; Adriani et al., 2007). However, removing affixes may also eliminate information relevant to meaning or sentiment polarity. Likewise, normalization of non-standard language may be less necessary for formal institutional messages than for informal social media data. Each procedure therefore requires empirical testing in its target domain.

Classification performance may also depend on the algorithm used. Multinomial Naive Bayes applies a probabilistic approach to word occurrence and is widely used for text classification (McCallum & Nigam, 1998). Linear Support Vector Machine (SVM), by contrast, constructs decision boundaries to distinguish categories in high-dimensional feature spaces (Cortes & Vapnik, 1995; Joachims, 1998). Both algorithms have been used in Indonesian sentiment analysis, although their performance can vary across datasets and experimental settings (Supian et al., 2024; Putri et al., 2025; Febrilliani & Wibowo, 2025). Their comparison can reveal whether preprocessing effects differ by classifier.

The individual contribution of preprocessing steps in Indonesian religious-institutional communication remains underexplored. Their effects on three-class classification of institutional messages are especially unclear. This study addresses this gap by evaluating six preprocessing steps through a leave-one-step-out ablation design using Multinomial Naive Bayes and Linear SVM. Its novelty lies in systematically assessing each preprocessing step's contribution in the context of pesantren communication.

This study aims to evaluate the effects of individual preprocessing techniques on Indonesian-language sentiment classification in Islamic boarding school communication, compare Multinomial Naive Bayes and Linear SVM performance, identify effective preprocessing configurations, and determine whether performance differences are statistically meaningful. The study contributes a three-class dataset produced through a four-annotator annotation protocol, with inter-annotator agreement measured using Fleiss' kappa. It also compares six preprocessing steps using cross-validation, split test data, learning curves, and Wilcoxon statistical tests.

2. Literature Review

2.1. Indonesian Text Pre-processing

Text preprocessing converts raw documents into structured forms suitable for analysis. Common procedures include removing irrelevant characters and symbols, standardizing letter cases, tokenizing text into individual words, removing stopwords, normalizing non-standard expressions, and stemming (Vijayarani et al., 2015). These operations reduce textual noise and improve consistency, but their effects are not universal. Previous studies indicate that preprocessing choices can either improve or reduce classification performance depending on the language, dataset, and classifier (Haddi et al., 2013; Uysal & Gunal, 2014). Therefore, preprocessing procedures should be evaluated in relation to the target domain rather than treated as a mandatory sequence applied identically to every dataset within this communication domain.

Indonesian presents challenges because its agglutinative morphology produces word variations through prefixes, suffixes, and confixes. Stemming can reduce these variations by returning derived words to their root forms, potentially consolidating terms with related meanings. Indonesian stemming research has examined confix-stripping approaches to address this morphological structure (Adriani et al., 2007), while Tala (2003) investigated the effects of stemming on Indonesian information retrieval. However, its usefulness in sentiment classification cannot be assumed, since affixes and word forms may contribute to contextual meaning or polarity. In institutional messages, formal expressions and polite language may also make normalization needs different from those in informal social media. Accordingly, this study examines six preprocessing steps individually through a leave-one-step-out ablation design to determine their contribution to classification performance.

2.2. TF-IDF Representation and Classification Algorithms

After preprocessing, machine-learning algorithms require numerical representations of textual data. Term Frequency–Inverse Document Frequency (TF-IDF) is a weighting scheme that reflects a term's frequency in an individual document while reducing the weight of terms that occur widely across the corpus (Salton & Buckley, 1988; Robertson, 2004). This representation emphasizes terms that are relatively distinctive to particular messages and is commonly used for document classification (Martinčić-Ipšić et al., 2019). In sentiment analysis, TF-IDF provides a sparse feature representation based on word importance rather than contextual embeddings. Its use therefore offers a transparent basis for examining how preprocessing changes affect the terms and weights supplied to a classifier (Zhou et al., 2020).

Multinomial Naive Bayes estimates class probabilities from word-frequency information and assumes conditional independence, making it a common baseline for text classification (McCallum & Nigam, 1998). Linear SVM, by contrast, identifies a decision boundary that maximizes the margin between classes and is effective in high-dimensional feature spaces (Cortes & Vapnik, 1995; Joachims, 1998). Both algorithms have been applied to Indonesian sentiment classification, with comparative findings showing that performance depends on the dataset and

experimental setting (Supian et al., 2024; Putri et al., 2025). This study uses TF-IDF with Multinomial Naive Bayes and Linear SVM to compare preprocessing configurations under two traditional classifiers and assess whether individual preprocessing steps affect the models differently. This setup supports controlled, interpretable comparisons across experiments.

2.3. Transformer-Based Models and Research Positioning

Pre-trained transformer models have expanded Indonesian natural language processing by providing contextual representations learned from large text corpora. IndoBERT was introduced alongside the IndoLEM benchmark for evaluating Indonesian language understanding (Koto et al., 2020), while IndoNLU provides resources for assessing Indonesian natural language understanding systems (Wilie et al., 2020). Subsequent studies have applied transformer-based models to sentiment analysis. Geni et al. (2023), for example, investigated BERT-based sentiment analysis of Indonesian election-related tweets. Ahmadian et al. (2024) explored hybrid models for Indonesian emotion classification and sentiment analysis, while Jiwandono et al. (2026) examined fine-tuned IndoBERT for stock-market news sentiment. These studies demonstrate the relevance of contextual language models for Indonesian text analysis.

Despite their capabilities, transformer-based approaches generally involve more complex architectures and computational requirements than conventional TF-IDF-based classifiers (Kowsari et al., 2019; Cunha et al., 2021). More importantly, their contextual representations make it less direct to isolate the contribution of individual preprocessing operations, the central concern of this study. The objective is not to identify the highest-performing model across all available approaches, but to measure how six preprocessing steps influence classification performance under controlled feature representation and classifier settings. Therefore, this study focuses on TF-IDF, Multinomial Naive Bayes, and Linear SVM, using leave-one-step-out ablation experiments. This approach is consistent with feature-level ablation studies that examine the individual contributions of features to classification performance (Hidayat et al., 2021). Transformer-based models, including IndoBERT, are outside the experimental scope and are considered a direction for future research on formal, domain-specific messages exchanged between pesantren administrators and students' guardians, particularly in institutional communication contexts (Kobayashi et al., 2018).

3. Methods

This study employed a quantitative experimental design with a comparative computational approach to evaluate the effects of text preprocessing on Indonesian sentiment classification. The population comprised messages received through an Islamic boarding school's official communication channel. The initial dataset included 2,571 CSV records with sender identities, message categories, content, and sentiment annotations. Sender identities were removed for privacy and excluded as classification features. Records with missing email addresses (290) or message categories (30) were excluded through listwise deletion. After accounting for overlapping missing values, 314 records were discarded, leaving 2,257 messages analyzed using total sampling. Three annotators initially classified messages as positive, neutral, or negative, and majority voting determined the final labels. For 129 cases (5.72%) with differing annotations, a fourth annotator determined the classification (Latief et al., 2023). Inter-annotator agreement was assessed using Fleiss' kappa and Landis and Koch's (1977) criteria, yielding 0.5204, indicating moderate agreement.

The preprocessing pipeline consisted of six sequential steps: cleaning, including URL, number, and special-character removal; case folding; tokenization using

Natural Language Toolkit's (NLTK) word_tokenize; lexical normalization using a 22-entry dictionary; stopword removal using a library-based list; and stemming using a stemming library. The contribution of each step was examined through a leave-one-step-out ablation design. The complete pipeline was compared with six scenarios, each excluding one preprocessing step, as presented in Table 1. In the No Tokenizing scenario, tokenization and all subsequent token-dependent operations were disabled because those operations required tokenized input.

Table 1. Pre-processing Ablation Scenarios

Scenario	Description
Full Pipeline	All six steps enabled
No Cleaning	Without cleaning
No Case Folding	Without lowercase standardization
No Tokenizing	Only cleaning and case folding; all token-dependent steps disabled
No Normalization	Without lexical normalization
No Stopword	Without stopword removal
No Stemming	Without stemming

Note: Tokenization-dependent operations are disabled together in the No Tokenizing scenario.

The preprocessed messages were represented numerically using unigram Term Frequency–Inverse Document Frequency (TF-IDF) through Scikit-learn's TfidfVectorizer, with lowercase=False, ngram_range=(1,1), smooth inverse document frequency, and L2 normalization (Pedregosa et al., 2011). Two classifiers were evaluated: Multinomial Naive Bayes (NB) with a smoothing parameter of α = 1.0 and Linear SVM implemented using LinearSVC with C = 1.0 and random_state=42. The dataset was divided into training data (n = 1,805) and held-out test data (n = 452). No hyperparameter tuning was conducted.

Model evaluation involved five-fold stratified cross-validation, held-out test evaluation, learning curves, and Wilcoxon signed-rank tests. Cross-validation assessed performance across folds, while the Wilcoxon test compared fold-level accuracy between NB and SVM for each preprocessing scenario at α = 0.05 (Wilcoxon, 1945; Demšar, 2006). Test performance was evaluated using accuracy, precision, recall, class-specific F1-scores, macro-F1, and confusion matrices. Macro-F1 was calculated as the unweighted average of F1-scores across sentiment classes. Learning curves used training proportions of 10–100%. Overfitting was indicated by a training-validation accuracy gap exceeding 0.10, whereas underfitting occurred when both accuracies were below 0.65. Learning-curve analysis used 1,805 training samples for NB and the full dataset (2,257 samples) for SVM.

4. Results

The final sentiment labels were examined to identify the distribution of positive, negative, and neutral messages across the entire dataset and the respective training and testing subsets. The distribution of each sentiment class is presented in Table 2 and Figure 1.

Table 2. Final Label Distribution

Class	All data (n = 2,257)	Training Data (n = 1,805)	Test Data (n = 452)
Positive	1,094 (48.47%)	879 (48.70%)	215 (47.57%)
Negative	837 (37.08%)	668 (37.01%)	169 (37.39%)
Neutral	326 (14.44%)	258 (14.29%)	68 (15.04%)

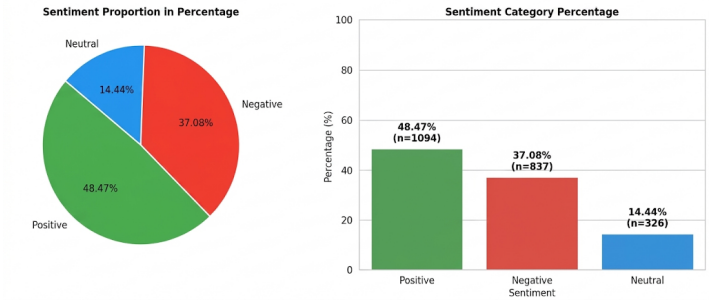


Figure 1. Final Sentiment Distribution

Of the 2,257 messages, positive sentiment predominated (48.47%), followed by negative (37.08%) and neutral sentiment (14.44%) in Table 2 and Figure 1. The class distributions in the training data (48.70% positive, 37.01% negative, and 14.29% neutral) and testing data (47.57% positive, 37.39% negative, and 15.04% neutral) were highly similar, indicating a relatively balanced distribution across the two datasets. Fleiss' kappa coefficient was 0.5204, indicating moderate inter-annotator agreement according to the classification proposed by Landis and Koch (1977). This moderate agreement suggests that sentiment boundaries were not always clearly distinguishable among human annotators, which may limit the maximum achievable performance of the classification models.

Table 3. Results of 5-fold Cross-Validation

Scenario	NB: Accuracy	NB: F1 Macro	SVM: Accuracy	SVM: F1 Macro
Full Pipeline	0.8155 ± 0.0189	0.7775	0.8133 ± 0.0235	0.7659
No Cleaning	0.8155 ± 0.0189	0.7775	0.8133 ± 0.0235	0.7659
No Case Folding	0.8183 ± 0.0159	0.7818	0.8105 ± 0.0246	0.7633
No Tokenizing	0.8238 ± 0.0214	0.7928	0.8122 ± 0.0226	0.7655
No Normalization	0.8155 ± 0.0178	0.7772	0.8133 ± 0.0235	0.7660
No Stopword	0.8211 ± 0.0193	0.7863	0.8100 ± 0.0225	0.7623
No Stemming	0.8211 ± 0.0233	0.7893	0.8105 ± 0.0237	0.7636

Table 3 presents the results of 5-fold cross-validation on the training data for Multinomial NB and Linear SVM. Average accuracy ranged from 0.8155 to 0.8238 for NB and from 0.8100 to 0.8133 for SVM, while macro F1 scores ranged from 0.7772 to 0.7928 and from 0.7623 to 0.7660, respectively. NB achieved its highest accuracy and macro F1 score under the No Tokenizing scenario (0.8238 and 0.7928, respectively). Meanwhile, SVM achieved its highest accuracy (0.8133) and macro-F1 score (0.7659) under both the Full Pipeline and No Cleaning scenarios. Notably, these two scenarios yielded identical results for both NB and SVM. The No Normalization scenario produced the same accuracy as the Full Pipeline for both models, while its macro-F1 scores were marginally lower for NB (0.7772) and higher for SVM (0.7660). Although some preprocessing ablation scenarios produced slightly higher scores, the differences were small relative to the standard deviations across folds (0.0159–0.0233 for NB and 0.0225–0.0246 for SVM). These findings suggest that removing individual preprocessing steps did not substantially or

consistently improve classification performance. Therefore, the full preprocessing pipeline was retained as a reasonable approach for subsequent model evaluation.

Table 4. Performance on Test Data (n = 452)

Scenario	NB: Accuracy	NB: F1 Macro	SVM: Accuracy	SVM: F1 Macro
Full Pipeline	0.8540	0.82	0.8208	0.77
No Cleaning	0.8540	0.82	0.8230	0.77
No Case Folding	0.8584	0.83	0.8252	0.77
No Tokenizing	0.8540	0.82	0.8208	0.76
No Normalization	0.8540	0.82	0.8208	0.77
No Stopword	0.8496	0.82	0.8230	0.77
No Stemming	0.8584	0.83	0.8230	0.77

On the test data in Table 4, NB achieved accuracy values ranging from 0.8496 to 0.8584 and macro F1 scores ranging from 0.82 to 0.83, while SVM achieved accuracy values ranging from 0.8208 to 0.8252 and macro F1 scores ranging from 0.76 to 0.77. NB achieved its highest accuracy under the No Case Folding and No Stemming scenarios (0.8584), whereas SVM achieved its highest accuracy under the No Case Folding scenario (0.8252). Compared with the Full Pipeline scenario, these improvements corresponded to only two additional correctly classified messages out of 452 for each model. Performance differences across preprocessing scenarios were therefore relatively small. A similar pattern was observed in the cross-validation results, suggesting that removing individual preprocessing steps did not consistently improve classification performance on the test data. These findings further support using the full preprocessing pipeline for the final model evaluation.

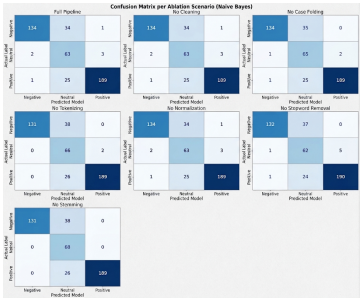


Figure 2. Naive Bayes Confusion Matrix on Seven Ablation Scenarios

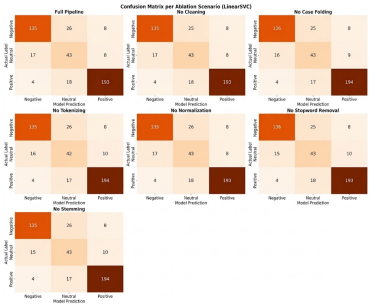


Figure 3. Linear SVM Confusion Matrix on Seven Ablation Scenarios

Figures 2 and 3 present the confusion matrices of NB (blue) and SVM (orange), respectively, across seven preprocessing ablation scenarios. Both figures show the distribution of correctly and incorrectly classified messages for each sentiment class, illustrating the models' classification performance under different preprocessing configurations.

Table 5. Performance per Class on Full Pipeline (Test data)

Class	NB: Precision	NB: Recall	NB: F1-score	SVM: Precision	SVM: Recall	SVM: F1-score	Support
Negative	0.98	0.79	0.88	0.87	0.80	0.83	169
Neutral	0.52	0.93	0.66	0.49	0.63	0.55	68
Positive	0.98	0.88	0.93	0.92	0.90	0.91	215
Macro average	0.82	0.87	0.82	0.76	0.78	0.77	452

Table 5 shows the classification performance of both models across sentiment classes using the full preprocessing pipeline. Both models performed best in the positive class, with F1 scores of 0.93 for NB and 0.91 for SVM, followed by the negative class, with scores of 0.88 and 0.83, respectively. However, both models performed worse at classifying neutral messages, with F1 scores of 0.66 for NB and 0.55 for SVM. The confusion matrices (Figures 2 and 3) further reveal differences in their misclassification patterns. NB classified 122 messages as neutral, although only 68 messages actually belonged to this class. Specifically, 34 negative messages and 25 positive messages were misclassified as neutral. Consequently, NB achieved high neutral recall (0.93) but relatively low precision (0.52), while its precision for both positive and negative classes reached 0.98. In contrast, SVM predicted 87 messages as neutral, resulting in a lower neutral recall (0.63) and precision (0.49). Although NB achieved higher accuracy (0.8540) and macro F1 (0.82) than SVM (0.8208 and 0.77, respectively), the differences in their classification patterns highlight the importance of considering class-specific errors in model selection. In practical deployment, the choice of model should also account for the consequences of misclassification, particularly whether missing negative messages is more detrimental than incorrectly prioritizing neutral messages.

Table 6. Wilcoxon Test of 5-fold Cross-Validation (NB vs SVM)

Scenario	Accuracy NB	Accuracy SVM	Difference (NB − SVM)	W	p-value	Significant
Full Pipeline	0.8155	0.8133	+0.0022	6.50	0.8750	No
No Cleaning	0.8155	0.8133	+0.0022	6.50	0.8750	No
No Case Folding	0.8183	0.8105	+0.0078	1.00	0.2500	No
No Tokenizing	0.8238	0.8122	+0.0116	3.00	0.3125	No
No Normalization	0.8155	0.8133	+0.0022	6.00	0.8125	No
No Stopword	0.8211	0.8100	+0.0111	2.50	0.2500	No
No Stemming	0.8211	0.8105	+0.0105	0.00	0.1250	No

Table 6 shows that NB consistently achieved higher accuracy than SVM across all seven scenarios, with differences ranging from 0.0022 to 0.0116. However, none of the differences were statistically significant ($p = 0.1250–0.8750$). These results should be interpreted cautiously because, with only five folds, the minimum attainable p-value for an exact two-sided Wilcoxon test is 0.0625, exceeding the 0.05 significance threshold. Therefore, the results do not establish model equivalence but indicate insufficient statistical evidence of a difference.

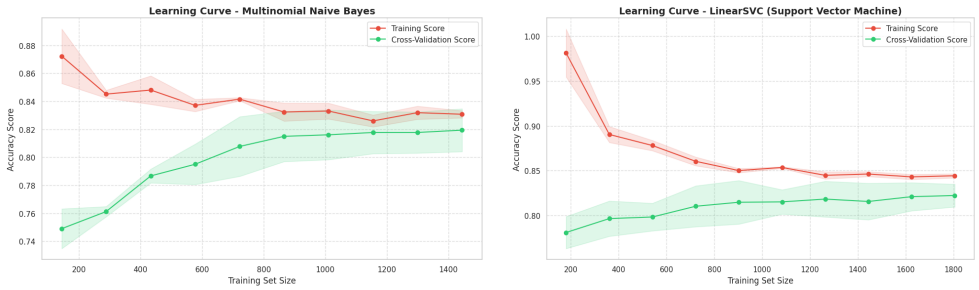


Figure 4. (a) Multinomial Naive Bayes Learning Curve and (b) The SVM Learning Curve

The learning curves in Figure 4 (a) and (b) indicate that both models exhibit good fitting performance. For NB, the final training accuracy was 0.8307, while the validation accuracy was 0.8194, resulting in a difference of 0.0113. For SVM, the final training accuracy was 0.8443, and the validation accuracy was 0.8223, with a difference of 0.0220. These relatively small differences suggest no substantial overfitting in either model under the default configuration. The curves were generated using raw text and built-in TF-IDF without additional text preprocessing. The NB curves were calculated using the training data (n = 1,805), whereas the SVM curves used the entire dataset (n = 2,257).

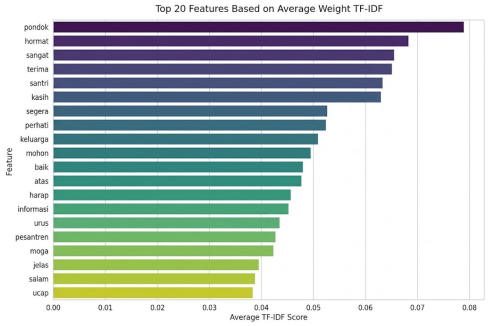


Figure 5. Twenty Features with the Highest TF-IDF Weighted Average

The twenty features with the highest mean TF-IDF weights in the training data under the full preprocessing pipeline are presented in Figure 5. The top features include "pondok" (0.0789), "sincerely" (0.0683), "very" (0.0655), "terima" (0.0651), "santri" (0.0634), and "kasih" (0.0630). Most of these features consist of domain-specific terms (e.g., *pondok*, *santri*, *pesantren*, and family-related terms) and expressions of politeness commonly used in formal correspondence (e.g., respect, gratitude, greetings, and requests). The mean TF-IDF weights reflect feature importance across the corpus rather than their ability to distinguish between sentiment classes. Therefore, this list primarily illustrates the linguistic characteristics of the messages rather than sentiment-specific features. Further analysis of class-specific feature contributions, such as SVM coefficients or NB log-probabilities, is needed to identify the features that distinguish between sentiment classes.

5. Discussion

The findings indicate that individual preprocessing steps contributed only marginally and inconsistently to sentiment classification performance. Although several ablation scenarios produced slightly higher accuracy and macro F1 scores than the full pipeline, the improvements were small and did not consistently occur

across both classifiers. This finding supports Uysal and Gunal (2014), who emphasized that the impact of preprocessing on text classification depends on data characteristics and the classification method. Similarly, Haddi et al. (2013) highlighted that preprocessing techniques do not necessarily improve sentiment analysis performance in every context. In this study, the similar results between the full pipeline and the No Cleaning scenario may also be associated with the built-in text processing functions of TF-IDF vectorization, which already handle certain textual features, such as punctuation and single-character tokens.

The comparison between NB and SVM showed that NB generally achieved higher accuracy and macro F1 scores than SVM across the evaluated scenarios. However, the Wilcoxon test did not identify statistically significant differences between the classifiers. Therefore, the results indicate a numerical advantage for NB but do not provide sufficient statistical evidence to establish its superiority over SVM. This finding differs from the assumption that more complex classification boundaries necessarily produce better performance. Although SVM is widely used for text categorization because of its effectiveness in high-dimensional feature spaces (Cortes & Vapnik, 1995; Joachims, 1998), its performance still depends on the dataset's characteristics. Differences in classifier performance have also been examined in Indonesian sentiment analysis studies, including those by Supian et al. (2024) and Putri et al. (2025). Thus, these findings reinforce the importance of evaluating classifiers empirically within specific language and domain contexts.

Regarding class-specific performance, both models performed better on positive and negative messages than on neutral messages. The neutral class showed the lowest F1 scores, particularly for SVM. This difficulty may be related to the moderate inter-annotator agreement ($\kappa = 0.5204$), indicating that sentiment boundaries were not always interpreted consistently by annotators. Landis and Koch (1977) classify this agreement level as moderate, while Latief et al. (2023) emphasize the importance of clear annotation guidelines for Indonesian text classification. In addition, messages in the *pesantren* context may combine appreciation, requests, and administrative information within a single message, making sentiment identification more challenging. This context is relevant to Kamal et al. (2024), which also examines sentiment in Islamic boarding school communities. These findings suggest that improving annotation consistency and clarifying neutral-class criteria may be as important as refining preprocessing techniques.

The learning curves further indicate relatively small differences between training and validation accuracy for both classifiers, suggesting no substantial overfitting under the evaluated configurations. Meanwhile, the TF-IDF feature analysis shows that the most prominent terms primarily represent *pesantren*-related vocabulary and expressions commonly used in formal communication. This finding aligns with Salton and Buckley (1988) on TF-IDF's role in representing term importance within a document corpus. However, high TF-IDF weights do not necessarily indicate strong discriminative power between sentiment classes. Consequently, the results describe the linguistic characteristics of the messages rather than directly identifying sentiment-specific features.

This study demonstrates that the full preprocessing pipeline provides a reasonable baseline for classifying *pesantren* messages, although removing individual steps does not substantially reduce performance. Future improvements may therefore focus on strengthening annotation guidelines, addressing neutral-class classification, and examining class-specific feature contributions. More advanced Indonesian language models, such as IndoBERT, may also provide opportunities for further comparison in this domain (Koto et al., 2020; Wilie et al., 2020).

6. Conclusion

This study evaluated six preprocessing stages for three-class sentiment classification of 2,257 boarding school messages using NB and SVM. Five-fold cross-validation produced accuracies of 0.8100–0.8238, with differences between scenarios smaller than fold-to-fold variation. On test data, the complete pipeline achieved 0.8540 accuracy for NB and 0.8208 for SVM. No preprocessing stage demonstrated a significant improvement, and the Wilcoxon test found no significant difference between classifiers. NB achieved higher macro-averaged performance and F1-score but tended to overpredict neutral messages, whereas SVM produced more balanced predictions but performed less effectively on the neutral class. These findings indicate that preprocessing effects should be evaluated empirically. The results provide a baseline for classifying *pesantren* communications and highlight neutral-message classification and annotation consistency as areas needing improvement.

These findings are limited by the use of one domain and data source, five folds for the Wilcoxon test, non-stratified test data, and the absence of class-imbalance handling. Hyperparameters were not tuned, one random seed was used, and the experiments relied on unigram TF-IDF rather than transformer comparisons. The no-tokenization scenario also disabled token-dependent steps, while the manual normalization dictionary was small. Moderate annotator agreement may have introduced label noise. Future research should test multiple institutions and sources, apply repeated cross-validation, stratified sampling and class-imbalance strategies, and tune hyperparameters across multiple seeds. It should also compare traditional classifiers with Indonesian pretrained models such as IndoBERT, examine class-specific feature contributions, expand normalization dictionaries, and strengthen annotation guidelines, particularly for neutral messages.

References

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M., & Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, *6*(4), 1-33. https://doi.org/10.1145/1316457.1316459.

Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2024). Hybrid models for emotion classification and sentiment analysis in Indonesian language. *Applied Computational Intelligence and Soft Computing*, *2024*(1), 282-297. https://doi.org/10.1155/2024/2826773.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., ... & Gonçalves, M. A. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*, *58*(3), 1024-1041. https://doi.org/10.1016/j.ipm.2020.102481.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*(1), 1–30.

Febrilliani, J. S., & Wibowo, A. (2025). Twitter public sentiment towards vasectomy in Indonesia using SVM and Naïve Bayes. *Research Horizon*, *5*(4), 1415–1424. https://doi.org/10.54518/rh.5.4.2025.756.

Geni, L., Yulianti, E., & Sensuse, D. (2023). Sentiment analysis of tweets before the 2024 elections in Indonesia using BERT language models. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, 9(3), 746–757. https://doi.org/10.26555/jiteki.v9i3.26490.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, *17*(1), 26–32. https://doi.org/10.1016/j.procs.2013.05.005.

Hidayat, R. A., Khasanah, I. N., Putri, W. C., & Mahendra, R. (2021). Feature-rich classifiers for recognizing textual entailment in Indonesian. *Procedia Computer Science*, *189*(1), 148-155. https://doi.org/10.1016/j.procs.2021.05.094.

Jiwandono, T. A., Achmad, M. S. H., & Suhirman. (2026). Fine-tuned IndoBERT for stock market sentiment analysis: Evidence from CNBC Indonesia news. *Indonesian Journal of Electrical Engineering and Computer Science*, *42*(3), 774–785. http://doi.org/10.11591/ijeecs.v42.i3.pp774-785.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning* (Lecture Notes in Computer Science, Vol. 1398, pp. 137–142). Berlin: Springer.

Kamal, M. R., Noviyanto, K., Subhan, H., & Afiana, L. (2024). Sentiment analysis on social media and stakeholders about negative issues among Islamic boarding school community in Indonesia. *Jurnal Theologia*, *35*(2), 261–280. http://dx.doi.org/10.21580/teo.2024.35.2.23164.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text classification for organizational researchers: A tutorial. *Organizational Research Methods*, *21*(3), 766-799. https://doi.org/10.1177/1094428117719322.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). Barcelona: International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.66.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150-167. https://doi.org/10.3390/info10040150.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310.

Latief, A. D., Jarin, A., Uliniansyah, M. T., Nurfadhilah, E., & Afra, D. I. N. (2023). A proven sentiment annotation guideline for Indonesian Twitter data. In *2023 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)* (pp. 31–36). New York: IEEE. https://doi.org/10.1109/IC3INA60834.2023.10285807.

Liu, B. (2012). *Sentiment analysis and opinion mining*. California: Morgan & Claypool Publishers.

Martinčić-Ipšić, S., Miličić, T., & Todorovski, L. (2019). The influence of feature representation of text on the performance of document classification. *Applied Sciences*, *9*(4), 743-755. https://doi.org/10.3390/app9040743.

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization* (pp. 41–48). New York: AAAI Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(23), 2825–2830.

Putri, G. K., Sujaini, H., & Ulumi, D. I. (2025). Perbandingan algoritma Support Vector Machine (SVM) dan Naive Bayes pada analisis sentimen Bahasa Jawa dan Sunda. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, *13*(2), 200-211. https://doi.org/10.26418/justin.v13i2.88285.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503-520. https://doi.org/10.1108/00220410410560582.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0.

Supian, A., Tri Revaldo, B., Marhadi, N., Efrizoni, L., & Rahmaddeni, R. (2024). Perbandingan kinerja Naïve Bayes dan SVM pada analisis sentimen Twitter Ibu Kota Nusantara. *Jurnal Ilmiah Informatika*, *12*(1), 15–21. https://doi.org/10.33884/jif.v12i01.8721.

Tala, F. Z. (2003). *A study of stemming effects on information retrieval in Bahasa Indonesia*. Amsterdam: University of Amsterdam (Master's thesis).

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 843–857). Suzhou: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.aacl-main.85.

Zhou, X., Gururajan, R., Li, Y., Venkataraman, R., Tao, X., Bargshady, G., ... & Kondalsamy-Chennakesavan, S. (2020). A survey on text classification and its applications. In *Web intelligence* (Vol. 18, No. 3, pp. 205-216). Thousand Oaks: SAGE Publications. https://doi.org/10.3233/WEB-200442.

Acknowledgment
We gratefully acknowledge the contributions of individuals who supported the completion of this article.

Funding Information
This research did not receive any funding.

Conflict of Interest Statement
The authors declare that there is no conflict of interest.

Ethical Approval and Originality Statement
Ethical approval was obtained for this study. The manuscript represents original work and has not been previously published, nor is it under consideration by another journal.

Data Disclosure Statement
The data that support the findings of this study are available from the corresponding author upon reasonable request.



Copyright: © 2026 by the authors.
This work is licensed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-sa/4.0/).