Comparing Machine Learning and IndoBERT for PPKM Twitter Sentiment Classification

Fathan Nabil Rahman[1], Muhammad Nabil Fikri[2], Cahya Din Ramadhan[3], Rashya Farellyndo Revanza[4], Much Aziz Muslim[5]

[1,2,3,4] Department of Computer Science, Universitas Negeri Semarang, Indonesia
[5] Department of Technology Management, Universiti Tun Hussein Onn Malaysia
Email: [1]nblrhmn5@students.unnes.ac.id, [2]nabilfikri2209@students.unnes.ac.id,
[3]cahyagans6@students.unnes.ac.id, [4]rashyafarellyndo@students.unnes.ac.id,
[5]a212muslim@mail.unnes.ac.id

Abstrak. *Kebijakan PPKM selama pandemi COVID-19 memicu diskusi luas di Twitter/X, sehingga analisis sentimen diperlukan untuk memahami respons masyarakat. Cuitan berbahasa Indonesia bersifat informal, memiliki variasi ejaan dan negasi, serta distribusi label tidak seimbang. Penelitian ini membandingkan Random Forest, XGBoost, dan IndoBERT yang di-fine-tuning untuk klasifikasi sentimen tiga kelas pada cuitan tentang PPKM. Pada model klasik, prapemrosesan mempertahankan kata negasi penting dan teks direpresentasikan dengan TF-IDF. Karena kelas Netral mendominasi, Macro F1-score digunakan sebagai ukuran utama selain akurasi. IndoBERT memberikan kinerja terbaik dengan akurasi 91,27% dan Macro F1-score 0,8429. Random Forest memperoleh Macro F1-score 0,6664, sedikit di atas XGBoost 0,6620. Waktu eksekusi dilaporkan secara deskriptif karena tahapan yang tercatat tidak identik.*

Kata Kunci: *Klasifikasi Sentimen, IndoBERT, Random Forest, XGBoost, PPKM*

Abstract. *During the COVID-19 pandemic, Indonesia's PPKM policy generated broad discussion on Twitter/X, making sentiment analysis useful for understanding public responses. Indonesian tweets are informal, contain spelling variation and negation, and have imbalanced labels. This study compares Random Forest, XGBoost, and fine-tuned IndoBERT for three-class sentiment classification of PPKM-related tweets. For the classical models, preprocessing preserves important negation markers and represents text with TF-IDF. Because Neutral tweets dominate, Macro F1-score is used as the main metric alongside accuracy. IndoBERT achieves the best performance, with 91.27% accuracy and a Macro F1-score of 0.8429. Random Forest obtains Macro F1-score 0.6664, slightly above XGBoost at 0.6620. Execution time is reported descriptively because the recorded stages are not identical.*

Keywords: *Sentiment Classification, IndoBERT, Random Forest, XGBoost, PPKM*

1. INTRODUCTION

Social media has become an important channel through which citizens respond to public decisions and everyday policy changes. Twitter, currently known as X, is especially useful for this purpose because discussions appear quickly and often capture reactions while an issue is still unfolding. During the COVID-19 pandemic in Indonesia, the Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM) policy generated extensive public conversation, including support, criticism, uncertainty, and neutral information sharing [1], [2]. An automatic sentiment analysis system can help transform these large volumes of posts into structured evidence, allowing policy actors and researchers to observe public perception without relying solely on manual reading [3].

Classifying sentiment in Indonesian tweets is nevertheless difficult. Posts on social media rarely follow formal writing conventions; they often contain slang, abbreviations, spelling variation, code mixing, typographical errors, and context-dependent expressions [4]. Such characteristics undermine the performance of methods that rely primarily on surface-level lexical patterns. Traditional classifiers using Term Frequency-Inverse Document Frequency (TF-IDF), for example, can identify frequently associated words but do not directly model word order, discourse context, or implied meaning. As a result, Random Forest (RF) and Extreme Gradient

Boosting (XGBoost) may perform well on clear lexical signals yet struggle when sentiment is indirect, culturally specific, or expressed through informal Indonesian phrasing [5].

Several studies have investigated Indonesian sentiment classification from different methodological perspectives. Geni et al. [6] reported that IndoBERT produced higher accuracy and F1-score than conventional baselines for tweets about the 2024 Indonesian election. In research on the PPKM context, Fransiscus and Girsang [7] applied a BERT-based approach and achieved an F1 Score of 84%, indicating the usefulness of contextual language models for COVID-19 policy discourse. Other work has relied on more classical pipelines; for example, Sutranggono and Imah [8] combined Support Vector Machine (SVM), TF-IDF, and Chi-Square feature selection for PPKM-related tweets and achieved 75.28% accuracy. In another Indonesian sentiment domain, Alfarros et al. [9] compared SVM, Random Forest, and XGBoost and showed that model behavior varies with the underlying class distribution.

Building on these findings, this study evaluates two different modeling paradigms under a single experimental setting. The first uses TF-IDF features with ensemble-based machine learning classifiers, namely Random Forest and XGBoost, which are commonly used for sparse and high-dimensional text representations [10]. The second uses IndoBERT, an Indonesian BERT checkpoint built on the bidirectional transformer architecture [11]. This architecture conditions representations on both left and right context and is therefore well suited to informal and meaning-dependent expressions. All models are assessed on the same labeled PPKM tweet dataset using accuracy, precision, recall, F1-score, and confusion-matrix analysis. This design enables a direct comparison between lightweight statistical representations and a more computationally intensive contextual model.

The contribution of this work is threefold. First, previous studies have generally evaluated these models across different datasets, domains, or experimental settings. This study contributes a controlled comparison of Random Forest, XGBoost, and IndoBERT specifically on PPKM-related Indonesian Twitter/X data under a common held-out test protocol. Second, it reports Macro F1-score, per-class F1-score, error distributions, feature importance for the classical models, and pairwise McNemar tests, thereby examining both minority-class behavior and the statistical support for differences between models. Third, it documents imbalance handling, hyperparameter selection, and the computational environment, so that the comparison can serve as a reproducible reference for informal Indonesian social media classification.

2. METHOD

2.1. Research Design

This study employs a quantitative experimental design to compare Random Forest (RF), XGBoost, and IndoBERT for sentiment classification of Indonesian tweets related to PPKM. The procedure consists of dataset preparation, exploratory data analysis (EDA), text preprocessing, feature construction, model training, and evaluation. Because classical machine learning and transformer models require different input representations, two preprocessing routes are used. The overall workflow is shown in Figure 1.

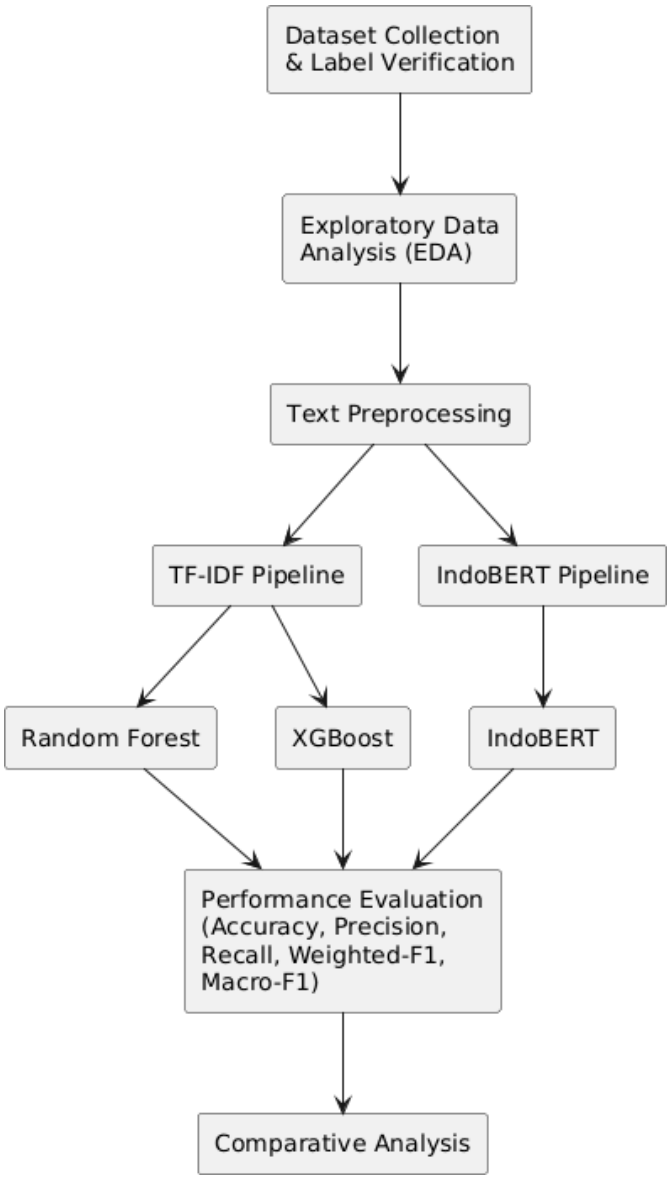


Figure 1. Research Workflow

Figure 1 summarizes the research procedure. The process starts with collecting the dataset and checking the sentiment labels to confirm their meaning. EDA is then performed to inspect the distribution of classes, tweet characteristics, and data quality. After this stage, the tweets are cleaned and transformed according to the model type. Random Forest and XGBoost receive TF-IDF vectors, whereas IndoBERT uses tokenization based on its corresponding tokenizer. A single 20% held-out test set is kept untouched; all model selection and validation are performed with five stratified folds on the remaining 80% of the data.

2.2. Data Collection

The data source used in this research is INA_TweetsPPKM_Labeled_Pure, a public tab-separated dataset obtained from Kaggle [12]. The original file contained 23,644 Indonesian Twitter/X posts collected with the #PPKM hashtag. The Kaggle release provides three integer sentiment labels: Positive (0), Neutral (1), and Negative (2). Representative examples from each label were manually inspected to verify this label mapping. The release does not provide an annotator protocol, annotator count, or inter-annotator agreement; accordingly, no independent annotation-quality statistic is claimed in this study.

2.3. Exploratory Data Analysis

EDA was conducted to describe the dataset prior to the modeling stage. The analysis included the number of instances in each class, tweet length measured by characters and words, and the most frequent vocabulary appearing in each sentiment group. Duplicate tweets were removed using the tweet text field to improve data quality. Class distribution was visualized with bar and pie charts, while word clouds were generated for Positive, Neutral, and Negative tweets using the WordCloud library. The EDA confirmed that Neutral tweets formed the largest portion of the dataset, which motivated the use of Macro F1-score as the main evaluation metric.

2.4. Text Preprocessing

Two preprocessing schemes were applied because the evaluated models use different text representations. For Random Forest and XGBoost, the TF-IDF preprocessing pipeline included lowercasing, URL and mention removal, hashtag conversion to ordinary words, deletion of non-alphabetic symbols, whitespace normalization, stopword removal with Sastrawi, removal of additional domain-specific terms, and stemming with the Sastrawi stemmer. Important Indonesian negation terms, including *tidak, tdk, ga, gak, nggak, bukan, belum*, and *jangan*, were deliberately kept outside the stopword list because they affect sentiment polarity.

For IndoBERT, a lighter cleaning strategy was used. Transformer models depend on contextual token sequences, so removing stopwords or applying stemming can reduce useful linguistic information. Consequently, this pipeline only removed URLs and user mentions and normalized whitespace, while keeping the original sentence structure and vocabulary as intact as possible for fine-tuning.

2.5. Train/Validation/Test Split

After duplicate removal, 158 records were removed. The two preprocessing steps then excluded 245 records with empty results, yielding a final modeling corpus of 23,241 posts. This corpus was split once into 18,592 development posts (80%) and 4,649 held-out test posts (20%) using stratification. Five-fold stratified cross-validation was performed on the development portion for model selection and validation. The held-out test set was not used in hyperparameter selection, early stopping, or cross-validation, and was evaluated only after the final model configuration was fixed.

2.6 TF-IDF Feature Extraction

Random Forest and XGBoost used Term Frequency-Inverse Document Frequency (TF-IDF) as the text representation. The initial TF-IDF inspection used a maximum of 15,000 features, an n-gram range of (1, 2), a minimum document frequency of 2, a maximum document frequency of 0.95, and sublinear TF scaling. For the model-selection pipelines, a model-specific TF-IDF vectorizer with 5,000 features was fitted inside each training fold, then applied to the corresponding validation fold and the held-out test set to avoid information leakage.

2.7 Model Training

The experiment included three main classifiers and one simple reference model. The reference model was a majority-class baseline that always predicted the most frequent label in the development data, Neutral. It has no learned parameters and is distinct from IndoBERT: IndoBERT is a fine-tuned contextual transformer classifier, whereas the baseline provides only a non-learning lower-bound comparison for the imbalanced dataset.

Random Forest combines many decision trees trained on bootstrap samples and random feature subsets, then determines the final class through majority voting. To address class imbalance, the Random Forest pipeline used class_weight = 'balanced', which assigns larger training weights to less frequent classes.

Hyperparameter selection for Random Forest used GridSearchCV with five stratified folds and Macro F1-score as the selection metric. The grid evaluated n_estimators in {100, 200, 300}, max_depth in {None, 10, 20}, and class_weight = 'balanced'. The selected configuration

was 300 estimators, unrestricted depth, and balanced class weights, with a mean cross-validation Macro F1-score of 0.6300.

XGBoost, or Extreme Gradient Boosting, constructs decision trees sequentially, with each new tree designed to reduce errors made by the previous ensemble using gradient-based optimization [13]. To apply an imbalance strategy comparable in purpose to Random Forest, balanced sample weights computed from the development labels were supplied during XGBoost fitting. The model used the multiclass softmax objective and mlogloss as the evaluation metric, with TF-IDF vectors as input features.

XGBoost hyperparameters were selected with GridSearchCV and the same five stratified folds. The grid evaluated learning_rate in {0.01, 0.1, 0.2}, max_depth in {3, 5, 7}, and n_estimators in {100, 200}; Macro F1-score was optimized. The selected configuration was learning rate 0.2, maximum depth 7, and 200 estimators, with mean cross-validation Macro F1-score of 0.6572.

IndoBERT is a BERT-based model pre-trained on large Indonesian text resources, specifically indobenchmark/indobert-base-p1 [14]. In contrast to TF-IDF, IndoBERT generates contextual token representations, meaning that a word's representation changes based on its surrounding words. To address class imbalance, IndoBERT fine-tuning used a weighted cross-entropy loss, with class weights computed from the development labels. Text was tokenized using the matching tokenizer with a maximum sequence length of 128 tokens, with padding and truncation applied as needed.

The imbalance-handling mechanisms differ because the three models expose different training interfaces: Random Forest applies class_weight at the tree-fitting stage, XGBoost accepts a weight for each training observation so its boosting objective emphasizes selected examples, and IndoBERT incorporates class weights directly into the weighted cross-entropy loss used during gradient-based fine-tuning. Although the implementations are not identical, their shared purpose is to penalize errors on minority classes more heavily than errors on the majority class. Therefore, the comparison reflects both the models' representations and the particular imbalance treatment used by each model.

IndoBERT was evaluated with the same five stratified folds over four epochs, with batch sizes of 16 for training and 32 for evaluation, a learning rate of 2 × 10^-5, 100 warmup steps, weight decay of 0.01, and early stopping patience of two epochs based on validation Macro F1-score. This configuration was selected as a commonly used fine-tuning starting point for short Indonesian texts while limiting computational cost and training variability; the maximum sequence length of 128 was likewise chosen to cover the short tweet sequences without excessive padding. Unlike Random Forest and XGBoost, no exhaustive IndoBERT hyperparameter grid was conducted. Consequently, the comparison should not be interpreted as a fully equivalent optimization comparison, although all models used the same development/test split and five-fold validation protocol. The selected fixed configuration produced fold Macro F1-scores of 0.8416, 0.8552, 0.8587, 0.8604, and 0.8585, with a mean of 0.8549.

2.8. Experimental Environment

All experiments were executed locally on Windows using Python 3.10, PyTorch 2.6.0 with CUDA 12.4, and Transformers 5.15.1. The machine used an Intel Core i7-14700HX CPU with 20 physical cores and 28 logical processors, 16 GB system memory, and an NVIDIA GeForce RTX 4060 Laptop GPU with 8,188 MiB VRAM. Random Forest GridSearchCV used all available CPU cores (n_jobs = -1), XGBoost GridSearchCV used two parallel jobs, and IndoBERT fine-tuning used the NVIDIA GPU.

2.9. Evaluation and Analysis

The held-out test set was used to evaluate each model on accuracy, Macro Precision, Macro Recall, Macro F1-score, Weighted Precision, Weighted Recall, Weighted F1-score, and execution time. Macro F1-score was selected as the primary metric because it averages class-level

F1-scores without weighting them by frequency, making it more sensitive to minority-class performance. Weighted F1-score was also reported to show performance while considering class support. Confusion matrices were reported to assess correct and incorrect predictions for each class, and feature importance was analyzed for the Random Forest and XGBoost models to identify influential TF-IDF terms. Pairwise McNemar tests were used to assess whether differences in prediction correctness were statistically significant.

3. RESULTS AND DISCUSSION

3.1. Dataset Properties and Label Distribution

Before comparing the models, the label distribution in the Indonesian Twitter Sentiment Analysis Dataset-PPKM was examined to assess class imbalance. After duplicate removal, the EDA corpus contained 23,486 posts: 17,598 Neutral (74.9%), 3,945 Negative (16.8%), and 1,943 Positive (8.3%). The two preprocessing routes then excluded 245 empty records. The final modeling corpus contained 23,241 posts: 17,370 Neutral (74.7%), 3,929 Negative (16.9%), and 1,942 Positive (8.4%). Figure 2 presents the distribution after duplicate removal, while all model results in subsequent sections use the final modeling corpus.

This distribution shows a strong imbalance in favor of the Neutral class. A classifier could therefore report high accuracy simply by predicting the majority class, while still failing to detect Positive and Negative tweets. For this reason, the following evaluation places particular emphasis on Macro F1-score, which assigns equal weight to each sentiment class regardless of frequency.

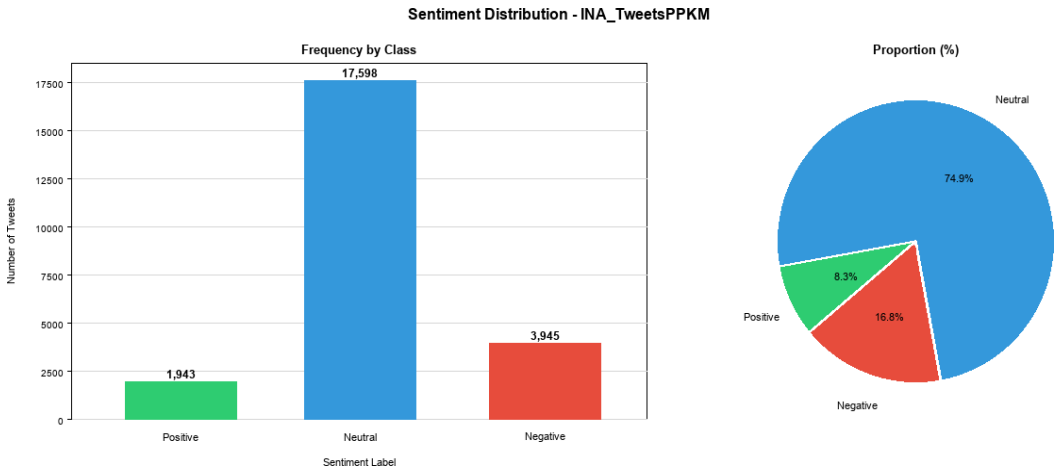


Figure 2. Sentiment Class Distribution of the PPKM Twitter Dataset

Figure 2 displays the class distribution in both counts and percentages. The bar chart reports the number of tweets in each class, whereas the pie chart emphasizes the dominance of the Neutral category. This pattern confirms that model evaluation must account for minority-class recognition rather than relying only on overall accuracy.

3.2. Exploratory Text Analysis

Exploratory text analysis was used to observe vocabulary tendencies across Positive, Neutral, and Negative Twitter/X posts. Word clouds offered a qualitative view of frequently appearing terms after duplicate removal. Although this visualization does not replace quantitative metrics, it helps identify general lexical patterns and possible overlaps among sentiment categories.



Figure 3. Word Cloud Visualization for Positive, Neutral, and Negative Classes

Figure 3 indicates that the sentiment classes share many contextual terms related to PPKM, COVID-19, government decisions, mobility, and public activity. The overlap suggests that sentiment labels cannot be separated only by topic words, because tweets with different polarities may discuss the same policy objects.

Neutral tweets are mostly characterized by informational terms associated with announcements, locations, and policy updates. Positive and Negative tweets contain more evaluative expressions, but their vocabulary still overlaps with the Neutral class. This overlap supports the need for models that can distinguish sentiment from context rather than relying only on isolated keywords.

These observations justify the use of two modeling strategies. TF-IDF-based classifiers are expected to capture discriminative lexical signals, while IndoBERT is expected to capture contextual meaning that may not be visible from word frequency alone.

3.3. Quantitative Model Evaluation

This section compares Random Forest, XGBoost, and IndoBERT with a Majority Baseline. The baseline always predicts Neutral, the most frequent development-set class, for every test post. It is not a variant of IndoBERT and contains no learned model parameters; it is included only to show the lower-bound performance of a trivial classifier on this imbalanced dataset.

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C} \text{F1}_i, \quad \text{F1}_i = \frac{2(\text{Precision}_i)(\text{Recall}_i)}{\text{Precision}_i + \text{Recall}_i} \quad (1)$$

In Equation (1), C denotes the number of sentiment classes, while Precision_i and Recall_i denote precision and recall for class i, respectively. Table 1 reports the complete results on the held-out test set.

Table 1. Comparative Model Evaluation on the Test Set (M-Prec = Macro Precision; M-Rec = Macro Recall; M-F1 = Macro F1-score; W-Prec = Weighted Precision; W-Rec = Weighted Recall; W-F1 = Weighted F1-score)

Model	Accuracy	M-Prec	M-Rec	M-F1	W-Prec	W-Rec	W-F1	Recorded time (s)
Majority Baseline	0.7475	0.2492	0.3333	0.2852	0.5587	0.7475	0.6395	0.0
Random Forest	0.8264	0.7287	0.6331	0.6664	0.8157	0.8264	0.8167	51.7
XGBoost	0.7686	0.6306	0.7300	0.6620	0.8286	0.7686	0.7859	14.9
IndoBERT	0.9127	0.8321	0.8546	0.8429	0.9150	0.9127	0.9137	15.6

Table 1 summarizes the test performance of all evaluated models. The Majority Baseline reached 0.7475 accuracy because it always selected the dominant Neutral label. However, its Macro F1-score was only 0.2852, showing that it did not recognize the Positive and Negative classes. This result demonstrates why accuracy alone is inadequate for this dataset.

For the classical machine learning models, Random Forest achieved 0.8264 accuracy and 0.8167 Weighted F1-score, while XGBoost achieved 0.7686 accuracy and 0.7859 Weighted F1-score. Random Forest also achieved a marginally higher Macro F1-score, 0.6664, compared with 0.6620 for XGBoost. The small absolute Macro F1 difference should therefore be interpreted cautiously.

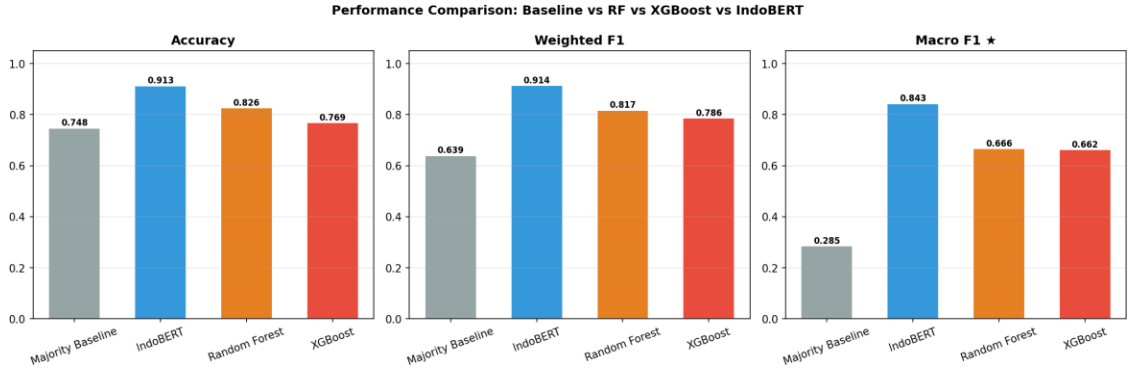


Figure 4. Comparison of Accuracy, Weighted F1-score, and Macro F1-score

Figure 4 compares Accuracy, Weighted F1-score, and Macro F1-score for all models. IndoBERT achieved the highest value on each metric, with 0.9127 accuracy, 0.9137 Weighted F1-score, and 0.8429 Macro F1-score. These results indicate that IndoBERT produced the strongest overall predictive performance.

The figure shows that Random Forest and XGBoost have very similar Macro F1-scores, whereas IndoBERT achieves a clearly higher score. Random Forest's slight Macro F1 advantage over XGBoost is consistent with its stronger precision for Positive and Negative posts. XGBoost's balanced sample weighting increased minority-class recall, but this was accompanied by lower precision and more false positives. IndoBERT's stronger result may be associated with its contextual token representations and its ability to process token sequences, but the present design does not isolate that contribution from weighted loss or the other model-specific imbalance treatments. These mechanisms should therefore be regarded as plausible explanations rather than experimentally established causes.

3.4. Per-Class Performance Analysis

Aggregate metrics provide an overview of model quality, but they do not fully reveal how each classifier handles majority and minority classes. A per-class analysis is therefore necessary because the dataset contains far more Neutral tweets than Positive or Negative tweets. Table 2 presents the F1-score for each sentiment category.

Table 2. Per-Class F1-score Comparison on the Test Set

Model	Positive F1	Neutral F1	Negative F1	Macro-F1
Majority Baseline	0.00	0.86	0.00	0.2852
Random Forest	0.47	0.90	0.63	0.6664
XGBoost	0.52	0.85	0.61	0.6620
IndoBERT	0.73	0.95	0.85	0.8429

Table 2 shows that the Majority Baseline failed to detect the Positive and Negative classes, producing an F1-score of 0.00 for both. Although its accuracy appeared relatively high

due to the large number of Neutral tweets, the low Macro F1-score confirms that it is not a reliable classifier for imbalanced sentiment data.

Random Forest produced a substantially better result than the baseline. It achieved F1-scores of 0.47 for Positive, 0.90 for Neutral, and 0.63 for Negative posts. The balanced class weighting improved recognition of the two minority classes relative to the non-learning baseline. XGBoost achieved F1-scores of 0.52 for Positive, 0.85 for Neutral, and 0.61 for Negative posts. Its higher minority-class recall was accompanied by lower precision, which explains why its Macro F1-score was slightly below Random Forest despite the use of balanced sample weights.

IndoBERT delivered the strongest per-class performance. It obtained F1-scores of 0.73 for Positive, 0.95 for Neutral, and 0.85 for Negative posts. Its stronger per-class results may be associated with its ability to process contextual token sequences, although the present experiment does not isolate the contribution of contextual representations from that of weighted loss and other model-specific differences. The per-class comparison confirms that IndoBERT is the most dependable model for three-class Indonesian sentiment classification in this experiment. Random Forest and XGBoost should be compared primarily by their closely matched Macro F1-scores rather than by a claim of a definitive computational-efficiency ranking, because the notebook records different execution stages for the three approaches.

3.5. Impact of Class Imbalance on Model Performance

The label distribution in Section 3.1 indicates that the dataset is dominated by Neutral tweets. This imbalance affected all models, particularly in their ability to recognize the smaller Positive and Negative groups. The Majority Baseline illustrates the issue clearly: it reached 0.7475 accuracy but only 0.2852 Macro F1-score because it did not identify any minority-class instances.

The Random Forest and XGBoost results further show the influence of class imbalance. Random Forest reached a macro F1-score of 0.6664, and XGBoost reached 0.6620. Random Forest's bagging mechanism aggregates diverse trees and, with class weighting, preserved higher minority-class precision. XGBoost used balanced sample weights and obtained higher minority-class recall, but its lower precision reduced its Macro F1-score. The difference is small and should be interpreted as a model behavior within this dataset rather than a general superiority claim.

IndoBERT was the most robust model under this imbalanced distribution. Its higher Macro F1-score and per-class F1-scores may be associated, among other factors, with its contextual token representations. However, because the imbalance-handling mechanisms differ across approaches, the current design cannot attribute the performance gain to contextual representations alone.

These findings confirm that class imbalance has a major effect on sentiment model evaluation. Accuracy and Weighted F1-score can favor models that perform well on the majority class, whereas Macro F1-score better reflects balanced recognition across all classes. Similar evaluation concerns have also been observed in Twitter-based public health sentiment studies [15].

3.6. Computational Efficiency Trade-off

Recorded execution time was retained as descriptive implementation information. It should not be interpreted as a directly comparable training-time benchmark because the notebook records the maximum mean fit time from GridSearchCV for Random Forest and XGBoost, but test-set inference time for IndoBERT.

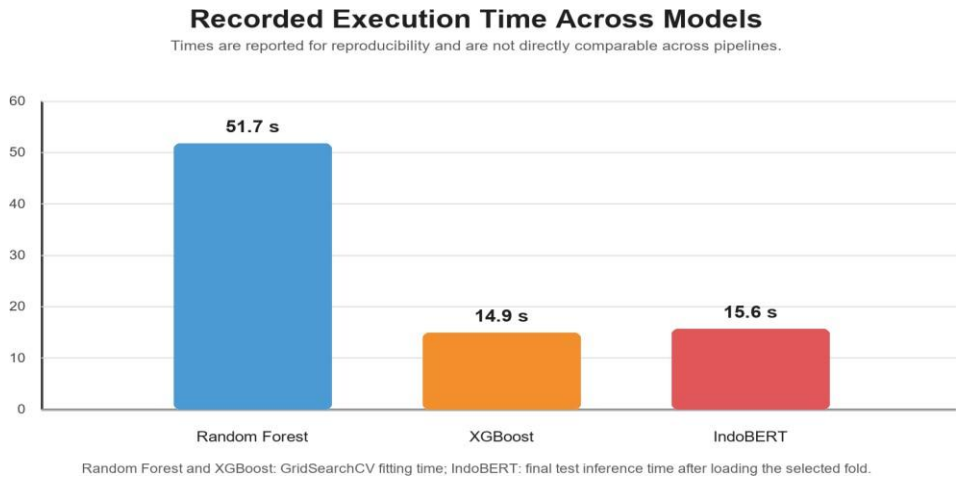


Figure 5. Recorded Execution Time Across Models

Figure 5 reports the recorded values: 51.7 seconds for Random Forest, 14.9 seconds for XGBoost, and 15.6 seconds for IndoBERT. The Random Forest and XGBoost values are GridSearchCV mean-fit records, whereas the IndoBERT value is test-set inference time; the values therefore do not measure identical pipeline stages.

IndoBERT achieved the strongest predictive results. A valid computational-efficiency comparison would require each model's end-to-end fit time to be measured under the same timing protocol; this was not available in the latest notebook output. Figure 5 is therefore included only as a record of the reported execution values.

3.7. Error Analysis and Ambiguous Cases

A qualitative review of test-set predictions was conducted to identify recurring error patterns. The analysis found 181 instances that were misclassified by all three models. These shared mistakes suggest that some posts contain ambiguity that remains difficult for both TF-IDF-based classifiers and transformer-based models.

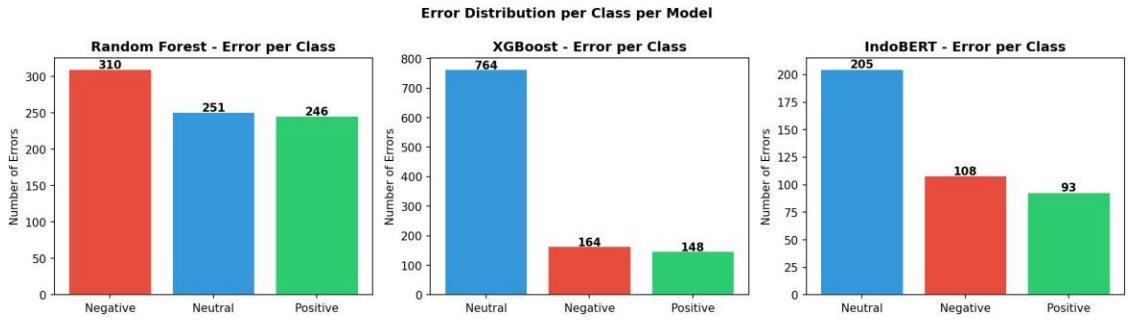


Figure 6. Error Distribution per Class Across Models

Figure 6 shows prediction errors by class for Random Forest, XGBoost, and IndoBERT. Random Forest made 310 errors on Negative posts, 251 on Neutral posts, and 246 on Positive posts. XGBoost made 764 Neutral errors, 164 Negative errors, and 148 Positive errors. IndoBERT produced fewer errors overall, consisting of 205 Neutral errors, 108 Negative errors, and 93 Positive errors.

The error distribution indicates that IndoBERT reduced misclassification across all sentiment categories compared with the classical models. Even so, the remaining errors show that

contextual models still struggle with ambiguous wording, mixed sentiment, sarcasm, irony, and informal abbreviations.

Sarcasm and irony were common sources of error. Some tweets conveyed criticism indirectly or used humorous expressions, making the intended sentiment difficult to infer from visible tokens alone. In several cases, tweets with neutral sentence structure were classified as Negative because they contained words commonly associated with complaints or pessimistic statements.

Mixed sentiment also contributed to incorrect predictions. Tweets that combined approval and criticism in one statement created competing sentiment signals. This problem was more apparent in TF-IDF-based models because word-frequency features do not always capture the overall intention of a sentence.

In summary, most of the remaining errors were due to ambiguous language, mixed sentiment, sarcasm, and informal expressions. Future training strategies such as oversampling or SMOTE [16] may help by increasing the model's exposure to underrepresented sentiment classes. These results also show that sentiment classification on Indonesian Twitter/X data still requires attention to pragmatic meaning, not only lexical patterns.

3.8. Feature Importance Analysis

Feature importance was extracted from the final Random Forest and XGBoost TF-IDF pipelines. Table 3 reports the highest-ranked terms. These rankings are descriptive: they indicate the terms used by the fitted models to split, rather than causal explanations of sentiment. Similar Indonesian sentiment studies and reviews also report model-dependent lexical patterns [17], [18].

Table 3. Most Influential TF-IDF Features

Model	Top TF-IDF terms	Note
Random Forest	level; semangat; lengkap; panjang; gak; ga; tidak	Topic and negation terms
XGBoost	jagajakarta; kuybisa; semangat; polres; lengkap	Corpus-specific lexical terms

3.9. Hyperparameter Sensitivity Analysis

Figure 7 reports the cross-validation sensitivity analysis used for Random Forest and XGBoost. It shows Macro F1-score across the tested Random Forest estimator counts and depths, and across the tested XGBoost learning rates and depths with 200 estimators. Application-focused sentiment research further shows that performance can vary across domains [19].

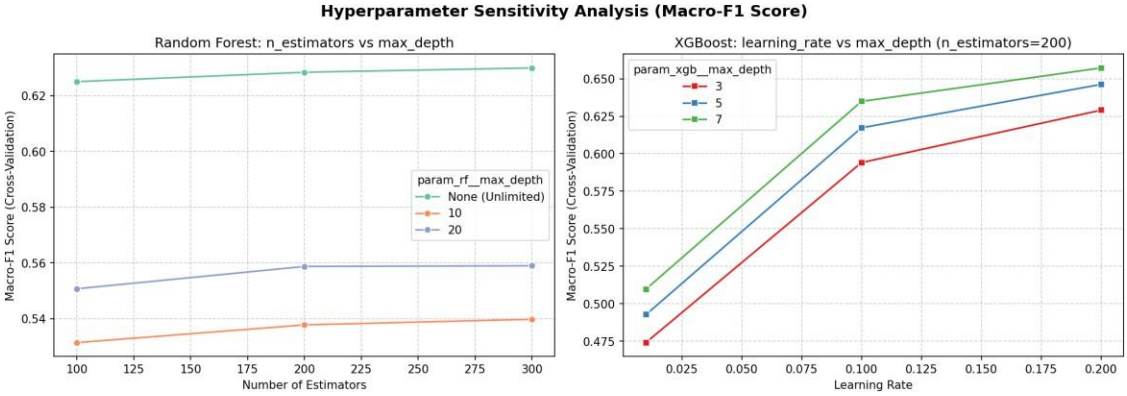


Figure 7. Hyperparameter Sensitivity Analysis

The selected Random Forest and XGBoost configurations are reported in Section 2.7. IndoBERT used a fixed configuration evaluated across five folds; its fold scores ranged from

0.8416 to 0.8604. The fixed IndoBERT settings were chosen as a commonly used, computationally manageable fine-tuning configuration for short Indonesian texts. Because the notebook did not perform a separate IndoBERT hyperparameter grid search, these fold results provide evidence of stability for the selected setting rather than a full sensitivity sweep; this asymmetry is an experimental limitation and prevents interpreting the comparison as an equivalent optimization across models.

3.10. Statistical Significance Testing

Pairwise McNemar tests were performed on the held-out test predictions. All three comparisons yielded $p < 0.05$, indicating statistically significant differences in prediction correctness at the 0.05 level. Table 4 reports the p-values. Reproducible evaluation also requires careful separation of model selection from final testing [20].

Table 4. Pairwise McNemar Test Results

Comparison	p-value	Result
RF vs XGBoost	1.6072 x 10^-20	Significant
RF vs IndoBERT	2.7684 x 10^-47	Significant
XGBoost vs IndoBERT	3.4799 x 10^-103	Significant

3.11. Limitations

This study has several limitations. It evaluates a single Kaggle dataset from the PPKM policy domain and only Twitter/X posts, so the results may not generalize to other Indonesian domains, time periods, or platforms. The source release does not provide inter-annotator agreement, and no external dataset validation was conducted. In addition, the available recorded-time values cover non-identical execution stages across models. These limitations define the scope of the empirical comparison. Comparative classifier studies, Indonesian BERT evaluations, and imbalance analyses likewise support cautious interpretation of model differences [21], [22], [23].

3.12. Confusion Matrix Analysis

Figure 8 presents the confusion matrix for Random Forest, Figure 9 presents the confusion matrix for XGBoost, and Figure 10 presents the confusion matrix for IndoBERT.

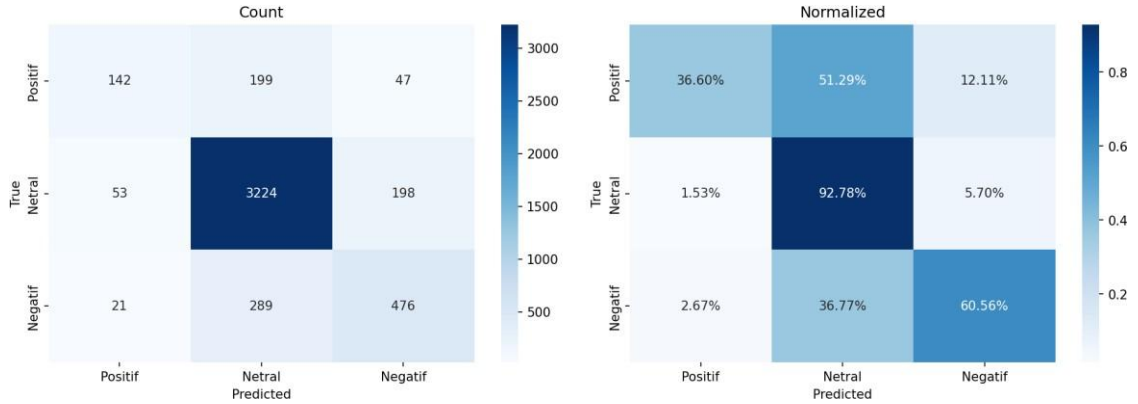


Figure 8. Confusion Matrix for Random Forest

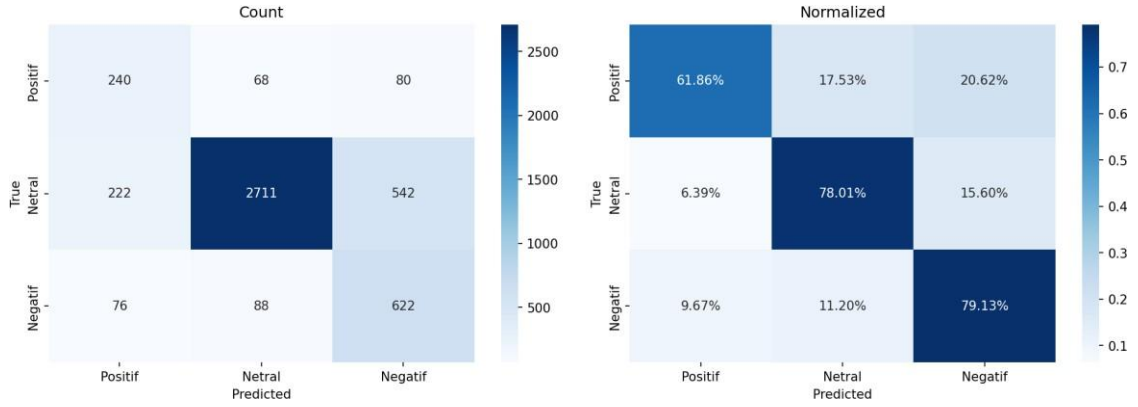


Figure 9. Confusion Matrix for XGBoost

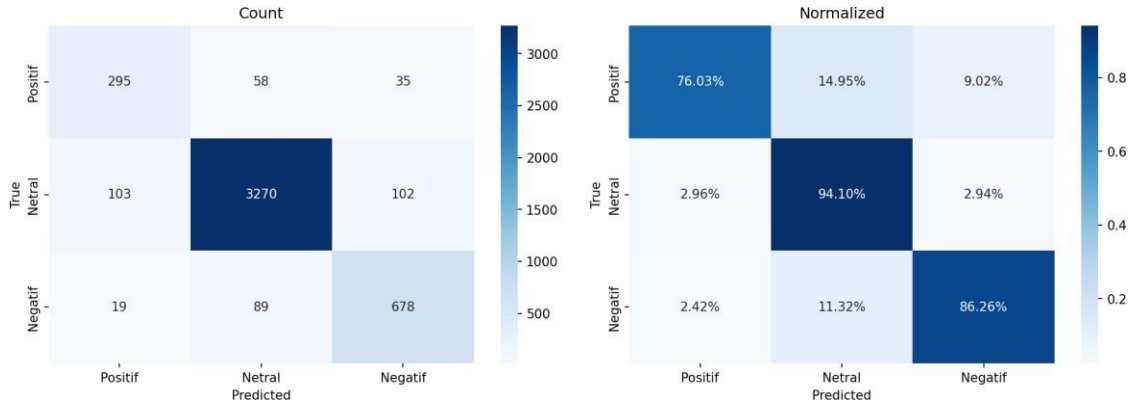


Figure 10. Confusion Matrix for IndoBERT

4. CONCLUSION

This study compared classical machine learning and transformer-based approaches for three-class sentiment classification of Indonesian Twitter/X posts about the PPKM policy. The evaluated models were Random Forest with TF-IDF features, XGBoost with TF-IDF features, and fine-tuned IndoBERT. Model selection used five-fold stratified cross-validation on the development data, and the final evaluation used a single untouched 20% test set. Because the dataset was highly skewed toward the Neutral class, Macro F1-score was used as the main metric.

The results indicate that IndoBERT achieved the strongest performance, with 91.27% accuracy and a Macro F1-score of 0.8429. This result is consistent with the use of contextual token representations for informal Indonesian posts, but it does not establish contextual representation as the sole cause of the improvement, as the imbalance treatments and optimization procedures differed across models. Among the classical classifiers, Random Forest achieved a macro F1-score of 0.6664, and XGBoost achieved 0.6620. The small difference between these classical models should be interpreted cautiously. A direct efficiency ranking cannot be drawn because the latest notebook records non-identical execution stages across the three models. Recent Indonesian BERT- and hybrid-sentiment studies provide related evidence for contextual modeling in Indonesian text classification [24], [25]. A related Indonesian study analyzed sentiment on Twitter using Bi-LSTM and Word2Vec and evaluated four data-splitting scenarios; its best-reported scenario was 80:10:10 [26]. Related Indonesian Twitter/X studies have also applied IndoBERT to presidential candidate sentiment [27] and to political discourse on Platform X [28].

The error analysis found that several tweets remained difficult to classify, particularly those containing ambiguity, sarcasm, irony, abbreviations, and mixed sentiment. These findings

indicate that informal Indonesian sentiment analysis remains challenging, even with transformer-based models [29]. Future research can examine imbalance-handling methods such as oversampling [30] or SMOTE to improve recall for minority classes. Additional work may also extend the framework to aspect-based sentiment analysis or other Indonesian public policy datasets.

5. ACKNOWLEDGEMENTS

The authors acknowledge the support of the Department of Computer Science at Universitas Negeri Semarang in completing this research. The authors also appreciate the public availability of the PPKM sentiment dataset on Kaggle and the open-source tools that supported the implementation. The authors thank lecturers, reviewers, and colleagues for their feedback and suggestions for improving this manuscript.

References

[1] Y. Wang, J. Guo, C. Yuan, and B. Li, "Sentiment Analysis of Twitter Data," Applied Sciences, vol. 12, no. 22, Art. no. 11775, 2022, doi: 10.3390/app122211775.

[2] S. Hansun, A. Suryadibrata, R. Nurhasanah, and J. Fitra, "Tweets Sentiment on PPKM Policy as a COVID-19 Response in Indonesia," Indian Journal of Computer Science and Engineering, vol. 13, no. 1, pp. 51-58, 2022, doi: 10.21817/indjcse/2022/v13i1/221301302.

[3] W. Budiharto and M. Meiliana, "Prediction and Analysis of Indonesia Presidential Election from Twitter Using Sentiment Analysis," Journal of Big Data, vol. 5, Art. no. 51, 2018, doi: 10.1186/s40537-018-0164-1.

[4] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp. 1320-1326, 2010.

[5] G. Venkateshwarlu, S. Akhila, V. Kavyasree, S. Vishnu, and V. S. Prasad, "Enhanced Text Classification Using Random Forest: Comparative Analysis and Insights on Performance and Efficiency," International Journal of Computer Engineering in Research Trends, vol. 11, no. 1s, pp. 1-8, 2024, doi: 10.22362/ijcert/2024/v11/i1/v11i1s01.

[6] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI), vol. 9, no. 3, pp. 746-757, 2023, doi: 10.26555/jiteki.v9i3.26490.

[7] Fransiscus and A. S. Girsang, "Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact Using BERT Method," International Journal of Engineering Trends and Technology, vol. 70, no. 12, pp. 281-288, 2022, doi: 10.14445/22315381/IJETT-V70I12P226.

[8] A. N. Sutranggono and E. M. Imah, "Tweets Emotions Analysis of Community Activities Restriction as COVID-19 Policy in Indonesia Using Support Vector Machine," CommIT (Communication and Information Technology) Journal, vol. 17, no. 1, pp. 13-25, 2023, doi: 10.21512/commit.v17i1.8189.

[9] M. A. D. Naufalino, M. Al-Husaini, and Rianto, "Comparison of Support Vector Machine, Random Forest and XGBoost for Sentiment Analysis on Indodax," Journal of Computer Networks, Architecture and High Performance Computing, vol. 7, no. 2, pp. 562-572, 2025, doi: 10.47709/cnahpc.v7i2.5894.

[10] A. Samih, A. Ghadi, and A. Fennan, "Enhanced Sentiment Analysis Based on Improved Word Embeddings and XGBoost," International Journal of Electrical and Computer Engineering, vol. 13, no. 2, pp. 1827-1836, 2023, doi: 10.11591/ijece.v13i2.pp1827-1836.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, vol. 1, pp. 4171-4186, 2019, doi: 10.18653/v1/N19-1423.

[12] A. Widiarta, "Indonesian Twitter Sentiment Analysis Dataset-PPKM," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/anggapurnama/twitter-dataset-ppkm. [Accessed: Aug. 24, 2026].

[13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016, doi: 10.1145/2939672.2939785.

[14] B. Wilie et al., "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, 2020, doi: 10.18653/v1/2020.aacl-main.85.

[15] M. H. Tsai and Y. Wang, "Analyzing Twitter Data to Evaluate People's Attitudes towards Public Health Policies and Events in the Era of COVID-19," International Journal of Environmental Research and Public Health, vol. 18, no. 12, Art. no. 6272, 2021, doi: 10.3390/ijerph18126272.

[16] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A Comprehensive Evaluation of Oversampling Techniques for Enhancing Text Classification Performance," Scientific Reports, vol. 15, Art. no. 21631, 2025, doi: 10.1038/s41598-025-05791-7.

[17] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications Using IndoBERT Approach," International Journal of Advanced Computer Science and Applications, vol. 14, no. 8, pp. 113-117, 2023, doi: 10.14569/IJACSA.2023.0140813.

[18] B. Setiawan, "A Review of Sentiment Analysis Applications in Indonesia Between 2023-2024," Journal of Information Engineering and Educational Technology, vol. 8, no. 2, pp. 71-83, 2024, doi: 10.26740/jieet.v8n2.p71-83.

[19] E. Kurniawati, A. I. Purnamasari, I. Ali, R. Kurniawan, and O. Nurdiawan, "Analisis Sentimen Ulasan Pengguna Aplikasi FLO di Google Play Store dengan Menggunakan Algoritma Naive Bayes," Jurnal Informatika dan Teknik Elektro Terapan, vol. 14, no. 1, pp. 791-799, Jan. 2026, doi: 10.23960/jitet.v14i1.8776.

[20] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[21] D. Pradipto and I. M. A. Agastya, "Comparative Analysis of Random Forest and Support Vector Machine Algorithms for Predicting Student Retention at Ibnu Sina University," SISTEMASI, vol. 15, no. 4, pp. 1491-1503, 2026, doi: 10.32520/stmsi.v15i4.6298.

[22] F. A. I. Suhendra, A. Darmayantie, A. S. Suhendra, and Pa Pa Min, "Comparative Analysis of Indonesian Pre-trained BERT Models for the Extractive Question Answering Task on an Indonesian-Translated SQuAD Dataset," MATRIK, vol. 25, no. 2, pp. 311-322, Mar. 2026, doi: 10.30812/matrik.v25i2.5847.

[23] L. Dube and T. Verster, "Enhancing Classification Performance in Imbalanced Datasets: A Comparative Analysis of Machine Learning Models," Data Science in Finance and Economics, vol. 3, no. 4, pp. 354-379, 2023, doi: 10.3934/dsfe.2023021.

[24] M. F. Kono, I. N. Fajri, and Y. Pristyanto, "Public Sentiment Analysis on Corruption Issues in Indonesia Using IndoBERT Fine-Tuning, Logistic Regression, and Linear SVM," Journal of Applied Informatics and Computing, vol. 9, no. 5, pp. 2616-2628, 2025, doi: 10.30871/jaic.v9i5.10537.

[25] D. G. Mandhasiya, H. Murfi, and A. Bustamam, "The Hybrid of BERT and Deep Learning Models for Indonesian Sentiment Analysis," Indonesian Journal of Electrical Engineering and Computer Science, vol. 33, no. 1, pp. 591-602, 2024, doi: 10.11591/ijeecs.v33.i1.pp591-602.

[26] F. Bouchra, I. M. A. D. Suarjaya, and N. K. D. Rusjayanthi, "Analisis Sentimen Masyarakat terhadap Tayangan Televisi Nasional menggunakan Metode Deep Learning," Jurnal Buana Informatika, vol. 15, no. 2, pp. 89-99, Oct. 2024, doi: 10.24002/jbi.v15i2.9118.

[27] P. Sayarizki, H. Hasmawati, and H. Nurrahmi, "Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates," Indonesian Journal on Computing, vol. 9, no. 2, pp. 61-72, Aug. 2024, doi: 10.34818/indojc.2024.9.2.934.

[28] V. E. Sidauruk and W. Herowati, "IndoBERT-Based Sentiment Analysis of Political Discourse on Platform X: The Case of Prabowo-Gibran Administration," Journal of Applied Informatics and Computing, vol. 10, no. 1, pp. 673-683, Feb. 2026, doi: 10.30871/jaic.v10i1.11586.

[29] V. D. Setiawan, D. U. Iswavigra, and E. Anggiratih, "Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court's Decision Regarding the Minimum Age of Vice Presidential Candidates," Scientific Journal of Informatics, vol. 12, no. 3, pp. 397-406, Aug. 2025, doi: 10.15294/sji.v12i3.26320.

[30] A. W. Wahyu, 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.