Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16242
e-ISSN : 2541-2019
p-ISSN : 2541-044X

IndoBERT-Based Sentiment Analysis of Indonesian Social Media Discourse on AI-Generated Images

Halvino Iqbal Nataprawira[1)*], Ida Nurhaida[2)]
[1,2)]Department of Informatics, Universitas Pembangunan Jaya, Tangerang Selatan Indonesia,
[2)]Center of Urban Studies, Universitas Pembangunan Jaya, Tangerang Selatan, Indonesia
[1)]halvino.iqbalnataprawira@student.upj.ac.id, [2)]ida.nurhaida@upj.ac.id

Submitted :May 18, 2026 | **Accepted** : Jun 8, 2026 | **Published** : July 5, 2026

Abstract: The rapid emergence of generative artificial intelligence has disrupted creative ecosystems, prompting widespread discourse across Indonesian social media. However, the exact sentiment structure of this public reaction remains empirically unmapped due to the contextual complexities of informal language. The objective of this research is to evaluate the efficacy of contextual language models by fine-tuning IndoBERT and benchmarking it against classical machine learning classifiers—including Complement Naive Bayes, Logistic Regression, and Support Vector Machine—for classifying social media sentiment. A multi-platform dataset comprising 2,981 Indonesian-language posts from X, Reddit, and YouTube was collected and manually annotated into positive, neutral, and negative classes. To address inherent class imbalance, Synthetic Minority Oversampling Technique was applied to classical models, while class-weighted loss and Masked Language Modeling augmentation were utilized for IndoBERT. Performance was evaluated using macro-averaged F1-score across five repeated stratified random splits. IndoBERT achieved a mean macro-F1 of 0.7131 ± 0.0180, outperforming the best classical baseline by approximately 0.12, demonstrating a pronounced advantage in resolving ambiguous neutral discourse. Negative sentiment heavily dominated the corpus at 61.8%, reflecting a prevailing critical stance toward AI-generated imagery concerning ethical and copyright issues. Furthermore, evaluation variance across random seeds exceeded variance from augmentation strategies, indicating test set composition is a major performance determinant. In conclusion, this study establishes a robust empirical baseline for Indonesian sentiment analysis, proving transformer architectures superior for nuanced public opinion mining.

Keywords: Artificial Intelligence; IndoBERT; Machine Learning; Sentiment Analysis; Social Media

INTRODUCTION

The rapid advancement of Generative AI has fundamentally restructured the digital creative ecosystem. Text-to-image models such as DALL-E, Midjourney, and Stable Diffusion have democratized visual creation by producing high-fidelity imagery from text prompts within seconds. Yet this accessibility has simultaneously ignited polarized socio-economic and ethical debates over artistic legitimacy, intellectual property, and copyright infringement, as the technology systematically disrupts established value-creation pipelines within creative economies (Brewer et al., 2024; Kanbach et al., 2023; Laba, 2024).

In Indonesia, this tension is compounded by a regulatory vacuum: existing copyright law lacks the statutory framework to attribute ownership to algorithmic outputs (Gede Ari Rama et al., 2023). Public discourse has consequently concentrated on X (formerly Twitter), yet its volume and complexity resist manual interpretation, necessitating computational approaches to reveal the polarized stance of visual artists toward AI-driven ethical infringements (Miyazaki et al., 2024). Accurately classifying these sentiments is therefore not merely a computational exercise; it provides structured and scalable insights that can inform policymakers in designing regulatory interventions, assist platform operators in refining content governance, and enable creative industry stakeholders to evaluate the societal legitimacy of AI integration.

*Halvino Iqbal Nataprawira

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16242
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The most directly comparable prior work analyzed 1,958 tweets regarding Indonesian public opinion on AI-generated images on X. (Saputra et al., 2025). To classify this discourse, their study applied Complement Naïve Bayes (yielding a macro-F1 score of 0.64), Support Vector Machine (SVM) (0.68), and IndoBERT (0.57). While traditional feature-based approaches like Naïve Bayes and SVM remain widely adopted baselines in Indonesian sentiment analysis (Asokere et al., 2025; Fadilah Arfat et al., 2022; Syahputra et al., 2022), and deep learning architectures such as Long Short-Term Memory (LSTM) networks are frequently applied across diverse domains (Alfat et al., 2023; Mariam & Nurhaida, 2025), recent literature demonstrates a decisive shift toward transformer-based models (Khaqqi et al., 2025). Traditional approaches inherently possess limited capacity for capturing the deep contextual meaning and complex linguistic noise present in social media compared to transformer-based architectures (Bello et al., 2023). However, applying Transformers as plug-and-play models without addressing this linguistic noise often leads to deceptive underperformance; recent research demonstrates that when IndoBERT is paired with systematic text normalization to handle the informal abbreviations and slang pervasive on X, it definitively outperforms traditional baselines like SVM and Naïve Bayes (Adhim & Cahyono, 2025). Although IndoBERT possesses an inherent architectural advantage in learning these complex patterns, comparing it against non-DL models remains methodologically essential; evaluating both under strictly identical preprocessing and environmental constraints ensures a fair comparison, establishing a necessary quantitative baseline to measure exactly how much value Transformer architectures add to this specific domain.

This study therefore employs IndoBERT, a BERT model pre-trained specifically on Indonesian as the primary analytical framework. Its bidirectional self-attention mechanism and subword tokenization render it inherently robust to the linguistic noise of social media text (Nashrullah et al., 2025), equipping it to resolve the contextual nuances and complex discourse structures that prior methods have demonstrably failed to address.

This study makes the following contributions. First, it provides a multi-platform sentiment analysis by integrating data from X, Reddit, and YouTube, extending prior work that primarily focuses on a single platform. Second, it introduces a repeated stratified random split evaluation strategy to better capture model performance variability. Third, it presents a controlled comparison between classical machine learning models and IndoBERT under consistent preprocessing and evaluation settings. Fourth, it offers a detailed per-class analysis highlighting the challenges of neutral and positive sentiment classification in Indonesian social media discourse.

METHODS

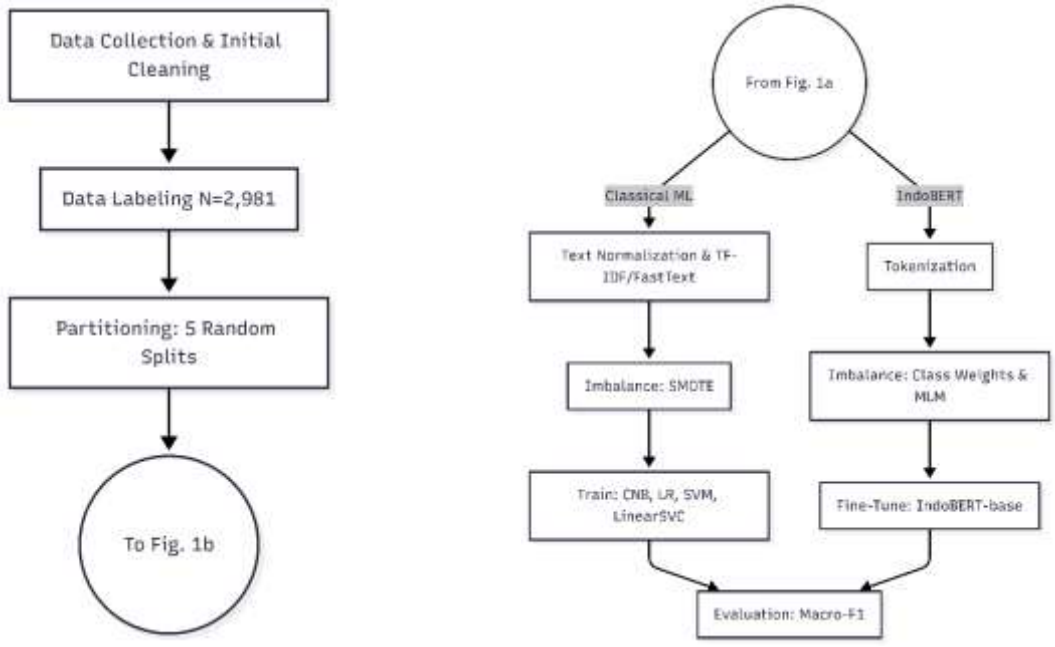

Fig 1. Research methodology. (a) Data preparation. (b) Model training and evaluation.

Research Methodology

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) as its overarching methodological framework, encompassing the phases of business understanding, data understanding, data preparation, modeling, and evaluation (Schröer et al., 2021). CRISP-DM was selected over alternatives, such as

*Halvino Iqbal Nataprawira

e-ISSN : 2541-2019
p-ISSN : 2541-044X

the IBM 10-stage model, due to its widespread adoption in academic NLP research and its flexibility in accommodating iterative experimental workflows. Within this framework, the study employs a quantitative experimental approach to design, train, and evaluate sentiment classification models on Indonesian social media text.

As illustrated in Figure 1, the research pipeline covers six phases: data acquisition via web scraping; two-level preprocessing tailored separately for IndoBERT and classical ML baselines; manual annotation of 2,981 samples into positive, neutral, and negative classes with inter-annotator validation; model development across IndoBERT and four classical classifiers; class imbalance handling via SMOTE (Chawla et al., 2002), for classical models and class-weighted loss with MLM augmentation for IndoBERT; and comparative evaluation using macro-averaged F1-score across five repeated random splits.

Data Sources & Collection

Indonesian-language social media data were collected from three platforms: X (formerly Twitter) as the primary source, supplemented by Reddit and YouTube. Data were gathered through automated web scraping using platform-specific APIs, targeting posts and comments containing keywords related to AI-generated images, including terms such as "gambar AI", "AI art", "Midjourney", and "DALL-E".

Collection was conducted iteratively across sessions spanning 2023 to 2026, yielding a raw corpus of approximately 11,000 samples. A random subset was drawn from the 11,000-sample corpus for annotation, with the final labeled dataset totaling 2,981 samples after filtering for relevance and language, deduplication, and removal of structurally incoherent entries.

It is important to note that the data spans multiple years (2023–2026), and temporal variation in public sentiment was not explicitly modelled in this study. As such, the results should be interpreted as an aggregated representation of sentiment across this period.

To support computational reproducibility, the final anonymized and annotated dataset (N=2,981), along with the specific random split partitions used for model evaluation, is publicly available at: GitHub repository. The raw scraped corpus is omitted from the repository to comply with the data sharing restrictions and privacy policies outlined in the platform developer agreements.

Data Preprocessing

Raw collected data underwent a two-stage preprocessing pipeline. The first stage applied universal cleaning across all samples, encompassing URL and mention removal, non-alphanumeric noise stripping, and case folding to produce annotation-ready text. The second stage was architecture-specific: classical ML baselines underwent additional slang normalization, stopword removal, and TF-IDF vectorization, whereas IndoBERT bypassed these destructive transformations to preserve linguistic context, relying solely on subword tokenization (indobenchmark/indobert-base-p2, max_len=256). All routines were implemented in Python 3.13 using PySastrawi, scikit-learn, and Hugging Face Transformers, with model training executed on Google Colab with an NVIDIA T4 GPU.

Prior to annotation, all collected samples were subjected to a uniform cleaning procedure to remove noise irrelevant to sentiment analysis. This included the removal of URLs, user mentions, and hashtags, normalization of whitespace by eliminating newlines and redundant spacing, and stripping of emojis, HTML entities, and malformed Unicode artifacts. Special characters were replaced with whitespace, and any samples rendered empty following these operations were discarded. A final deduplication pass using exact-match comparison was performed to eliminate redundant entries. This stage was applied identically across all samples and served as the preprocessing basis for IndoBERT, which requires no further transformation beyond tokenization.

Samples designated for classical machine learning models, specifically Complement Naive Bayes (CNB), Support Vector Machine (SVM), Logistic Regression, and Linear Support Vector Classification (LinearSVC), underwent an extended preprocessing pipeline building upon the initial cleaning stage.

Case folding was first applied to ensure lexical uniformity. Punctuation was normalized, including the collapsing of laughter expressions (e.g., 'wkwk', 'haha') and elongated character repetitions into their base forms. Standalone punctuation marks and isolated numeric tokens were subsequently stripped.

Given the informal register of social media text, a dictionary-based slang normalization step was applied to map non-standard Indonesian terms to their formal equivalents, improving lexical consistency for TF-IDF vectorization (Sindu et al., 2024). The normalization dictionary was manually constructed from common Indonesian social media expressions and abbreviations, consisting of approximately 70 slang-to-formal lexical mappings, with particular emphasis on negation forms and sentiment-bearing terms to preserve polarity information. This procedure was applied only to the classical machine learning pipeline, whereas IndoBERT operated on minimally processed text to retain contextual information.

Stopword removal was subsequently performed using a lightweight Indonesian stopword list adapted for social media text. To preserve sentiment semantics, negation words, specifically *tidak*, *bukan*, *jangan*, and *belum*, were

*Halvino Iqbal Nataprawira

e-ISSN : 2541-2019
p-ISSN : 2541-044X

explicitly excluded from removal through a whitelist mechanism. Finally, whitespace was normalized, and any samples rendered empty or duplicated following the full preprocessing pipeline were discarded. Table 1 illustrates representative examples of these text transformations.

Table 1. Examples of text transformations across preprocessing stages

Stage	Unit
Raw	@inggridient_ 30% kemungkinan di generate AI. 70% by hand. Gue sampe sampe skrg belum nemu AI yg gambarnya bisa pake tema dengan modulasi yg detail banget. Gambar ini termasuk detail.
Initial Cleaning (IndoBERT)	30% kemungkinan di generate AI. 70% by hand. Gue sampe sampe skrg belum nemu AI yg gambarnya bisa pake tema dengan modulasi yg detail banget. Gambar ini termasuk detail.
Classical ML	kemungkinan generate ai by hand saya sampe sampe skrg belum nemu ai gambarnya bisa pakai tema modulasi detail sangat gambar termasuk detail

Prior to model-specific preprocessing, the cleaned dataset was partitioned into training, validation, and test sets using stratified random splitting across five split seeds (42, 7, 123, 99, 2026), ensuring consistent class proportions across partitions. The use of repeated random splits rather than a single fixed partition follows evidence that standard splits produce unstable system rankings, though it is acknowledged that random repartitioning of a single corpus remains a pragmatic approximation of truly independent test set evaluation (Søgaard et al., 2021). Model-specific preprocessing, specifically, slang normalization, stopword removal, and TF-IDF vectorization for classical ML baselines and subword tokenization for IndoBERT, was then applied to each partition independently to prevent data leakage.

Data Labeling

The cleaned corpus was manually annotated by the primary author into three sentiment classes: positive (appreciation, support, or favorable reception), negative (criticism, concern, or rejection), and neutral (factual, ambiguous, or non-opinionated expressions). To validate consistency and mitigate subjective bias, a 100-sample subset was co-annotated by two volunteer annotators with IT/CS backgrounds, provided solely with class definitions and labeling guidelines, and declaring no competing interests. This subset-based validation approach via Cohen's Kappa prior to full-scale single-author annotation is established practice in Indonesian sentiment analysis (Abdul Chamid & Kusumaningrum, 2024), with agreement scores presented in Table 2.

Table 2. Inter-annotator agreement scores (Cohen's Kappa)

Annotator Pair	Cohen's Kappa	Interpretation
Author vs. Annotator 1	0.6543	Substantial
Author vs. Annotator 2	0.5455	Moderate
Annotator 1 vs. Annotator 2	0.5337	Moderate

By established Kappa interpretation benchmarks, the obtained values correspond to moderate-to-substantial agreement, considered acceptable for informal social media text where sentiment boundaries are inherently ambiguous. Disagreements were resolved through majority voting. The final annotated dataset comprises 2,981 samples: 1,841 negative (61.8%), 731 neutral (24.5%), and 409 positive (13.7%), with the pronounced class imbalance directly motivating the imbalance-handling strategies described in subsequent sections.

Although the full dataset was primarily annotated by a single annotator, the subset-based inter-annotator agreement provides an estimate of labeling consistency that is acceptable for informal social media text. This approach is commonly adopted in sentiment analysis studies where large-scale multi-annotator labeling is constrained. In addition, transformer-based models such as IndoBERT are known to be relatively robust to moderate levels of label noise.

Feature Extraction and Classification Models

For classical machine learning classifiers, this study utilized TF-IDF (Term Frequency-Inverse Document Frequency), a statistical weighting scheme that computes a combined score by multiplying a term's frequency within a document by its inverse frequency across the broader corpus to highlight discriminative tokens. For the SVM (RBF) configuration, these high-dimensional sparse vectors were supplemented with dense FastText word embeddings to better capture morphological variations common in informal Indonesian text (Bojanowski et al., 2017). To address class imbalance in these classical models, SMOTE-based oversampling was applied to the TF-

*Halvino Iqbal Nataprawira

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16242
e-ISSN : 2541-2019
p-ISSN : 2541-044X

IDF features. While this approach effectively mitigates imbalance and improves F1-scores in Indonesian text classification (Dwi Cahya et al., 2023), it may introduce synthetic samples that do not perfectly reflect realistic linguistic patterns in high-dimensional sparse spaces.

Among the classical classifiers evaluated, this study explicitly highlights Complement Naïve Bayes (CNB). While standard Naïve Bayes computes the posterior probability of a class based on Bayes' theorem assuming conditional feature independence, CNB estimates the probability that a sample belongs to the *complement* of a class. This formulation reduces the predictive bias introduced by majority class dominance, making it significantly more robust for skewed datasets.

To capture deeper contextual semantics, this study also leverages IndoBERT (specifically indobenchmark/indobert-base-p2), a BERT-based model pretrained on large-scale, domain-relevant Indonesian corpora (Wilie et al., 2020). Rather than relying on static feature extraction, BERT architectures are pretrained using a Masked Language Modeling (MLM) objective, which trains the model to reconstruct masked input tokens using bidirectional context (Devlin et al., 2019). For the sentiment classification task, IndoBERT was fine-tuned by appending a classification head to the [CLS] token representation. Class imbalance during this fine-tuning stage was handled by applying a class-weighted cross-entropy loss, which assigns a higher penalty to minority class prediction errors (Korsakpaisarn et al., 2026).

RESULTS

Experimental Setup and Dataset Distribution

Prior to model evaluation, an exploratory analysis of the filtered dataset (N=2,981) revealed a pronounced class imbalance. As illustrated in Figure 2, negative sentiment dominates the corpus at 61.8%, compared to neutral (24.5%) and positive (13.7%) samples. This natural skew, which is typical of social media discourse regarding AI-generated imagery, directly necessitated the implementation of targeted imbalance-handling strategies for each model architecture.

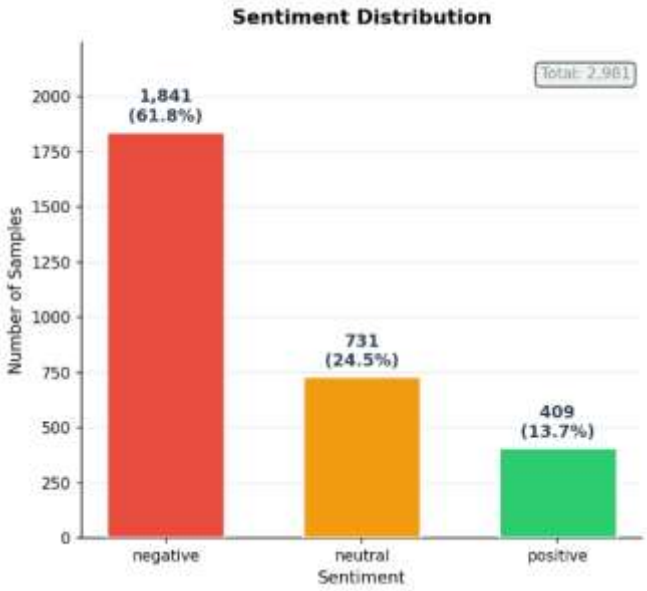

Fig 2. Sentiment class distribution of the filtered dataset

To ensure robust and reproducible evaluation across this imbalanced corpus, all models were evaluated under a strictly unified 70/15/15 stratified split. To account for the optimization instability and performance variance attributable to random initialization in transformer fine-tuning, the experiments were repeated across five distinct random seeds (seeds: 42, 7, 123, 99, 2026) (Mosbach et al., 2021), maintaining consistent class proportions across all partitions.

Classical machine learning baselines utilized cross-validation on the training partition for hyperparameter tuning, whereas the IndoBERT architecture leveraged the validation partition for early stopping and optimal checkpoint selection. The final model configurations, alongside their respective data partitions and balancing techniques, are summarized in Table 3. Crucially, the 'Total Real' column explicitly denotes the count of original, unmodified dataset instances; all test partitions remained strictly static and consisted exclusively of unaugmented samples to prevent data leakage.

*Halvino Iqbal Nataprawira

Table 3. Dataset partitioning and imbalance handling strategies per model configuration

Model	Features	Imbalance Handling	Train	Val	Test	Total Real
CNB	TF-IDF	None	2,086	447	448	2,981
CNB	TF-IDF	SMOTE	3,864*	447	448	2,981
LinearSVC	TF-IDF	SMOTE	3,864*	447	448	2,981
Logistic Regression	TF-IDF	SMOTE	3,864*	447	448	2,981
SVM (RBF)	TF-IDF + FastText	SMOTE	3,864*	447	448	2,981
IndoBERT	—	Class Weighting	2,086	447	448	2,981
IndoBERT + Augmentation	—	Class Weighting + MLM Augmentation	~2,560**	447	448	2,981

*Training set expanded via SMOTE synthetic oversampling on TF-IDF vectors; original sample count unchanged.
**Training set expanded via MLM-based augmentation of minority classes (neutral and positive) in the training partition only; augmented training size varied across split seeds (range: approximately 2,531–2,560 samples) due to differing minority class compositions per split.

Table 4. IndoBERT hyperparameter configuration

Parameter	Value
Model	indobenchmark/indobert-base-p2
Max Sequence Length	256
Batch Size	8
Epochs	5
Learning Rate	1e-5
Warmup Ratio	0.1
Weight Decay	0.05
Label Smoothing	0.1
Optimizer	AdamW

Table 5. Classical ML hyperparameter configuration

Parameter	Value
TF-IDF Max Features	30,000
TF-IDF N-gram Range	(1,2)
TF-IDF Min DF	2
SMOTE k-Neighbors	3
CNB Alpha	1.0
LR / LinearSVC C	1.0
*GridSearchCV Folds	*10

*GridSearchCV with 10-fold cross-validation was applied exclusively to the SVM (RBF) configuration for kernel and regularization parameter selection.

Hyperparameter configurations for IndoBERT and classical ML models are summarized in Tables 4 and 5 respectively. For SVM, GridSearchCV with 10-fold cross-validation was applied over a search space comprising two kernels (RBF and linear), four regularization values ($C \in \{1, 10, 100, 1000\}$), two gamma settings ({scale, auto}, applicable to RBF only), and two class weighting schemes ({None, balanced}), yielding 24 total candidate configurations. The epoch count for IndoBERT was set to five, informed by validation loss monitoring indicating overfitting onset beyond this threshold, computational constraints of the T4 GPU environment, and consistency with BERT fine-tuning ranges reported in comparable classification tasks (Perwira et al., 2025). All configurations were held constant across split seeds to ensure observed variance reflects data partitioning rather than model configuration differences. Models were evaluated across five repeated stratified random splits rather than a single fixed partition, following evidence that standard splits produce unreliable system rankings, while acknowledging that random repartitioning of a single corpus remains an approximation of truly independent evaluation (Søgaard et al., 2021).

*Halvino Iqbal Nataprawira

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16242
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Overall Model Performance

Table 6 presents the macro-F1 scores across all model configurations, evaluated over five repeated random splits. Results are reported as mean ± standard deviation to account for evaluation variance attributable to test set composition.

Table 6. Overall model performance across five random splits (Mean ± Standard Deviation)

Model	Features	Imbalance Handling	Mean Accuracy	Mean Macro-F1	Macro-F1 Std	Mean MCC
CNB	TF-IDF	None	0.6424	0.5754	±0.0130	0.3890
CNB	TF-IDF	SMOTE	0.6594	0.5799	±0.0039	0.3846
LinearSVC	TF-IDF	SMOTE	0.6857	0.5894	±0.0265	0.4068
Logistic Regression	TF-IDF	SMOTE	0.6848	0.5975	±0.0280	0.4045
SVM (RBF)	TF-IDF + FastText	SMOTE	0.7156	0.5822	±0.0340	0.4221
IndoBERT	**—**	**Class Weighting**	**0.7638**	**0.7131**	**±0.0180**	**0.5727**
IndoBERT + Augmentation	—	Class Weighting + MLM	0.7656	0.7132	±0.0258	0.5522

IndoBERT consistently outperformed all classical ML baselines across all five split seeds, achieving a mean macro-F1 of 0.7131 and a Matthews Correlation Coefficient (MCC) of 0.5727. MCC was also reported because it has been widely advocated as a robust metric for imbalanced classification problems (Chicco & Jurman, 2020). This represents an improvement of approximately 0.12 in F1 and 0.17 in MCC over the best classical baseline (Logistic Regression). This performance differential is attributable to IndoBERT's bidirectional self-attention mechanism, which captures deep contextual dependencies and resolves linguistic ambiguity in social media text that TF-IDF-based feature representations inherently cannot model. Among classical models, SMOTE yielded marginal improvement for discriminative classifiers (LinearSVC, Logistic Regression) but provided negligible benefit for CNB ($\Delta = 0.005$), consistent with known sensitivity of complement-based probability estimation to synthetic sample distributions. MLM-based augmentation yielded no meaningful improvement over the baseline IndoBERT configuration (with MCC actually decreasing slightly from 0.5727 to 0.5522), consistent with findings that BERT augmentation benefits diminish significantly at dataset sizes exceeding 1,000 samples and imbalance ratios below 9:1(Hu et al., 2022).

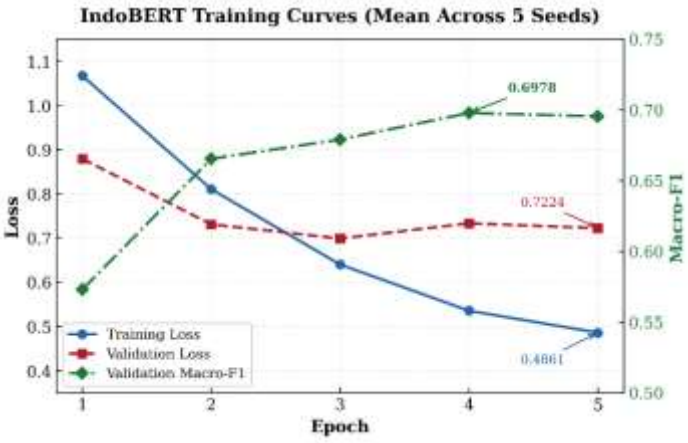

Fig 3. Mean IndoBERT training and validation curves across five random seeds

Figure 3 illustrates the training dynamics of the IndoBERT model, displaying the mean training loss, validation loss, and validation macro-F1 scores across all five random initializations over five epochs. The learning curves demonstrate stable convergence: training loss steadily decreases, while the validation macro-F1 score climbs and peaks at an average of 0.6978 by the fourth epoch. Although average validation loss reaches its minimum at epoch three before exhibiting a marginal increase, the concurrent stability in the validation F1 score indicates mild,

*Halvino Iqbal Nataprawira

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16242
e-ISSN : 2541-2019
p-ISSN : 2541-044X

expected overfitting rather than severe model degradation. This trajectory empirically validates our use of optimal checkpoint selection, ensuring that the model weights evaluated on the test set were saved at peak generalizability.

The slightly higher mean test macro-F1 (0.7131) compared to the validation peak (0.6978) is not paradoxical; under repeated random splits, independently resampled test sets can, by chance, yield more favorable class compositions, especially for the positive class leading to marginally higher scores. To further characterize IndoBERT's evaluation stability, Figure 4 presents per-seed macro-F1 scores across all five split seeds.

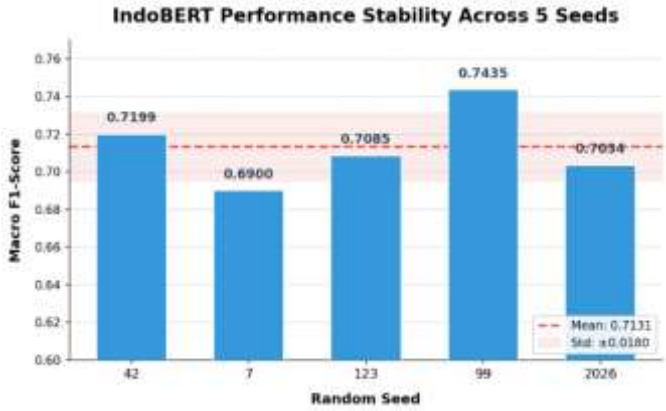

Fig 4. IndoBERT Macro-F1 performance variance across five random initializations

Performance varied across split seeds, ranging from 0.6900 (seed 7) to 0.7435 (seed 99), yielding a standard deviation of ±0.0180. This variance notably exceeds the improvement yielded by augmentation or oversampling strategies, suggesting that test set composition is a more significant source of performance variation than minority class handling at this dataset scale, consistent with documented BERT fine-tuning instability attributable to optimization variance across random initializations (Mosbach et al., 2021).

Relative to the most directly comparable prior work on the same task (Saputra et al., 2025), this study's IndoBERT configuration improved macro-F1 by +0.14, and exceeded their best overall model (SVM + FastText, 0.68) by +0.033. Classical ML scores in this study are comparatively lower, attributable to the repeated random splits evaluation protocol, which produces more conservative estimates than single-run reporting, combined with a noisier dataset spanning a broader collection period.

Per-Class Analysis

To provide a more granular evaluation beyond the aggregate macro-F1, per-class performance was analyzed across all model configurations. Quantitative per-class F1 scores, as reported in Figure 5, represent mean values across all five random split seeds, consistent with the aggregate evaluation protocol applied throughout this study. This ensures that reported per-class performance reflects stable central tendency estimates rather than the outcome of any single initialization, and that observed performance degradation on minority classes is attributable to inherent dataset characteristics rather than optimization variance.

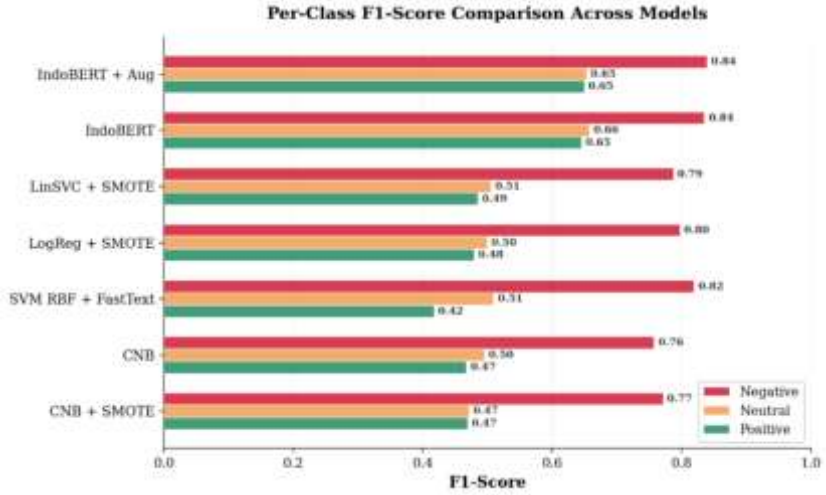

Fig 5. Per-class F1-scores across all model configurations

*Halvino Iqbal Nataprawira

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16242
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Negative sentiment classification achieved the highest F1 scores across all models (0.76–0.84), reflecting its 61.8% majority representation. The narrow performance gap between classical ML and IndoBERT on this class suggests that sparse bag-of-words features suffice when a target class is abundantly represented.

The most pronounced gap emerged in neutral classification: classical models plateaued at 0.42–0.51 F1, while IndoBERT achieved 0.66 This margin suggests that contextual representations in IndoBERT are more effective in handling semantic ambiguity compared to traditional feature-based approaches, particularly for neutral Indonesian discourse where TF-IDF features cannot model discourse-level context or conditional meaning. This difficulty is not unique to the Indonesian domain; neutral sentiment introduces ambiguity that makes classification more difficult (Ceh-Varela et al., 2025).

Positive sentiment yielded the weakest results (0.42–0.65 F1) due to severe underrepresentation (13.8%) and inherent boundary ambiguity with neutral discourse. Notably, applying SMOTE to CNB degraded neutral class performance compared to the baseline CNB (0.47 vs. 0.50). This reinforces that synthetic oversampling introduces distributional noise rather than genuine minority class signals during complement probability estimation.

Further qualitative inspection of misclassified samples reveals that sentences containing mixed sentiment or implicit contrast are particularly challenging. For example, statements such as 'AI ini keren, tapi tetap bermasalah secara etika' tend to be misclassified as positive or negative depending on dominant lexical cues. This indicates that ambiguity and contrastive structures remain difficult even for contextual models.

To illustrate these misclassification patterns, Figure 6 presents the IndoBERT confusion matrix for Seed 123. With a macro-F1 of 0.7085, it most closely approximates the five-seed mean (0.7131), serving as a representative single-run reference.

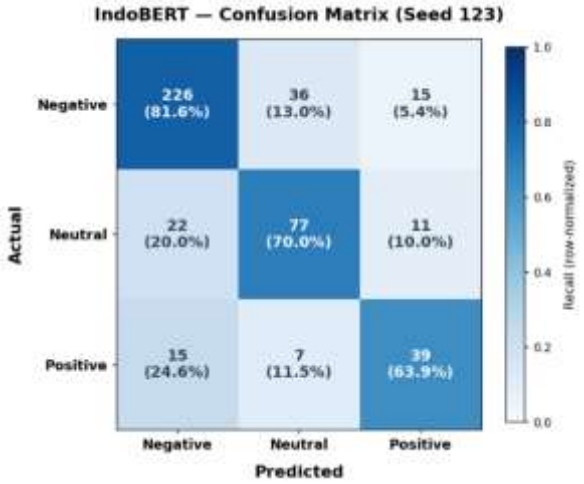

Fig 6. IndoBERT confusion matrix (Seed 123)

The matrix reveals that neutral and positive misclassifications predominantly collapse toward the negative class, consistent with the majority class bias inherent to imbalanced training distributions. This pattern corroborates the per-class F1 trends observed across all models and confirms that residual classification error is concentrated at the boundary between minority and majority classes rather than distributed uniformly.

DISCUSSION

The results confirm that transformer-based contextual representations substantially outperform classical feature-based approaches for Indonesian sentiment classification on informal social media text, particularly on semantically ambiguous classes. The most pronounced gap emerged in neutral classification (+0.16 F1), consistent with prior evidence that TF-IDF features cannot capture discourse-level conditionality in informal Indonesian registers. The overall IndoBERT macro-F1 of 0.7131 represents a +0.14 improvement over the reference paper's IndoBERT configuration (Saputra et al., 2025), and a +0.033 margin over their strongest model (SVM + FastText, 0.68). This improvement is attributable to a larger and more linguistically diverse dataset, multi-platform data integration, and the repeated random split evaluation protocol, which produces more conservative and reliable estimates than single-run reporting. Notably, MLM-based augmentation yielded negligible gain (Δ = 0.0001), suggesting that at this dataset scale and imbalance ratio, augmentation offers diminishing returns, consistent with findings by (Hu et al., 2022).

Several limitations constrain the generalizability of these findings. The dataset was primarily labeled by a single annotator, with inter-annotator validation restricted to 100 samples yielding moderate agreement (κ = 0.53–0.65), introducing non-trivial label noise risk particularly at the neutral-positive boundary. Additionally, evaluation

*Halvino Iqbal Nataprawira

e-ISSN : 2541-2019
p-ISSN : 2541-044X

relied on random repartitioning of a single corpus, while more reliable than a fixed split, remains an approximation of truly independent test evaluation (Søgaard et al., 2021). The dataset's temporal span (2023–2026) and single-domain coverage further limit cross-domain and cross-temporal generalizability. These factors collectively suggest that the reported macro-F1 represents a dataset-bound performance ceiling rather than an absolute upper bound for the task.

This dataset-bound ceiling is most evident in the disproportionately low F1-score for the positive class (0.42–0.65), a persistent challenge across all evaluated models. Qualitative analysis of the misclassified samples reveals that this performance degradation is driven heavily by the prevalence of sarcasm and the semantic ambiguity of positive vocabulary within the specific discourse of AI-generated art. First, negative sentiment regarding AI disruptions is frequently disguised using positive lexical amplifiers. For instance, a misclassified sample stating, "Keren banget AI sekarang, modal ngetik doang udah bisa nyolong karya orang" (AI is so cool now, just by typing you can steal people's work) contains strong positive indicators (keren banget) but carries a fundamentally negative stance regarding copyright infringement. Models lacking deep pragmatic reasoning consistently misclassify these sarcastic critiques as positive. Second, genuine positive sentiment in this domain often lacks explicit emotional polarity, instead manifesting as objective technical praise. A statement such as, "Hasil render dari prompt Midjourney ini presisi dan teksturnya rapi" (The render from this Midjourney prompt is precise and the texture is neat) expresses positive appreciation but uses descriptive, analytical language that models frequently confuse with the neutral class. This limited and highly technical positive vocabulary blurs the boundary between objective observation and subjective praise, directly contributing to the aforementioned label noise and the suppressed positive F1-score.

Beyond dataset constraints and discourse complexities, IndoBERT's architectural limitations further bound these results. The 256-token truncation discards contextual information in longer threads, and fine-tuning instability on imbalanced data risks suboptimal convergence. Critically, IndoBERT's pretraining corpus is predominantly formal Indonesian text, creating domain mismatch with the informal, code-switched registers of social media. This constraint has been addressed in related work through domain-adaptive pretraining on Indonesian Twitter corpora, demonstrating improved downstream performance in comparable settings (Koto et al., 2021); applying architectures like IndoBERTweet to this domain represents a direct and necessary avenue for future improvement.

CONCLUSION

This study evaluated IndoBERT against classical ML baselines for Indonesian sentiment classification on AI-generated image discourse across five repeated random splits. Three key findings emerged.

First, class imbalance was the dominant performance determinant: the negative class (61.8% of data) achieved consistently high F1 scores (0.76–0.84), while neutral and positive classes exhibited substantially higher variance attributable to underrepresentation and semantic boundary ambiguity in informal Indonesian text.

Second, evaluation variance across split seeds consistently exceeded variance introduced by augmentation or oversampling, indicating that test set composition is a more significant performance factor than minority class handling at this dataset scale, underscoring repeated random splits as a more reliable evaluation protocol than single-run reporting.

Third, IndoBERT demonstrated clear superiority over classical ML approaches, most pronounced in neutral class F1 (~0.16 gap), confirming that bidirectional contextual representations resolve semantic ambiguity more effectively than sparse TF-IDF features. Overall, IndoBERT achieved a mean macro-F1 of 0.7131 ± 0.0180, outperforming the reference paper's IndoBERT baseline by +0.14.

A primary limitation of this study is that 2,981 samples were labeled by a single annotator, with IAA validation restricted to 100 samples yielding only moderate inter-annotator agreement ($\kappa = 0.53–0.65$); this introduces non-trivial label noise risk, particularly at the neutral-positive boundary. Another limitation is reliance on a single-domain dataset without cross-domain validation, which may limit the generalizability of the findings to other types of Indonesian text or platforms. Future work should prioritize model-level improvements: domain-adaptive pretraining of IndoBERT on Indonesian social media corpora, exploration of larger transformer architectures such as XLM-R or IndoBERT-large, and ensemble strategies combining contextual embeddings with sentiment lexicons to address persistent positive class underperformance.

REFERENCES

Abdul Chamid, A., & Kusumaningrum, R. (2024). *Labeling Consistency Test of Multi-Label Data for Aspect and Sentiment Classification Using the Cohen Kappa Method*. https://doi.org/10.18280/isi.290118

*Halvino Iqbal Nataprawira

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Adhim, N. F., & Cahyono, N. (2025). Optimization of IndoBERT for Sentiment Analysis of FOMO on Social Media Through Fine-Tuning and Hybrid Labeling. *Journal of Applied Informatics and Computing*, *9*(6), 3786–3797. https://doi.org/10.30871/JAIC.V9I6.11686

Alfat, L., Nasucha, M., Uddin, N., Salleh, K. A., & Baharun, N. (2023). Sentiment Classification of Indonesian Emotion Related to Vaccination Event using LSTM. *2023 IEEE World AI IoT Congress, AIIoT 2023*, 825–829. https://doi.org/10.1109/AIIOT58121.2023.10174581

Asokere, M., Wusu, A., & Olabanjo, O. (2025). Twitter (X) as an electoral barometer: systematic evidence from sentiment analysis of Twitter data. *International Journal of Information Technology 2025*, 1–24. https://doi.org/10.1007/s41870-025-03039-1

Bello, A., Ng, S. C., & Leung, M. F. (2023). A BERT Framework to Sentiment Analysis of Tweets. *Sensors 2023, Vol. 23, Page 506*, *23*(1), 506. https://doi.org/10.3390/s23010506

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Brewer, P. R., Cuddy, L., Dawson, W., & Stise, R. (2024). Artists or art thieves? media use, media messages, and public opinion about artificial intelligence image generators. *AI & SOCIETY 2024 40:1*, *40*(1), 77–87. https://doi.org/10.1007/s00146-023-01854-3

Ceh-Varela, E., Shakya, S. R., & Imhmed, E. (2025). Challenges in Detecting Nuanced Sentiment with Advanced Models. *Cloud Computing and Data Science*, *6*, 115–135. https://doi.org/10.37256/ccds.6220256316

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics 2019 21:1*, *21*(1), 6-. https://doi.org/10.1186/S12864-019-6413-7

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/V1/N19-1423

Dwi Cahya, L., Luthfiarta, A., Imanuel, J., Krisna, T., Winarno, S., & Nugraha, A. (2023). Improving Multi-label Classification Performance on Imbalanced Datasets Through SMOTE Technique and Data Augmentation Using IndoBERT Model. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *9*(3), 290–298. https://doi.org/10.25077/teknosi.v9i3.2023.290-298

Fadilah Arfat, M., Nurkholis, A., Kurniawan, I., Pagari Alam No, J. Z., Ratu, L., Kedaton, K., & Bandari Lampung, K. (2022). Analisis Sentimen Masyarakat Indonesia Terkait Vaksin Covid-19 Pada Media Sosial Twitter Menggunakan Metode Support Vector Machine (Svm). *Jurnal Informatika: Jurnal Pengembangan IT*, *7*(2), 96–103. https://doi.org/10.30591/jpit.v7i2.3549

Gede Ari Rama, B., Julia Mahadewi, K., Kunci, K., Buatan, K., Cipta, H., & Hukum, S. (2023). Urgensi Pengaturan Artificial Intelligence (AI) Dalam Bidang Hukum Hak Cipta Di Indonesia. *JURNAL RECHTENS*, *12*(2), 209–224. https://doi.org/10.56013/RECHTENS.V12I2.2395

Hu, L., Li, C., Wang, W., Pang, B., & Shang, Y. (2022). Performance Evaluation of Text Augmentation Methods with BERT on Small-sized, Imbalanced Datasets. *Proceedings - 2022 IEEE 4th International Conference on Cognitive Machine Intelligence, CogMI 2022*, 125–133. https://doi.org/10.1109/CogMI56440.2022.00027

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2023). The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science 2023 18:4*, *18*(4), 1189–1220. https://doi.org/10.1007/s11846-023-00696-z

Khaqqi, F. N., Alfat, L., & Nurhaida, I. (2025). Sentiment Analysis of US-China Tariffs using IndoBERT and Economic Impact on Indonesia. *Journal of Applied Informatics and Computing*, *9*(6), 3000–3011. https://doi.org/10.30871/JAIC.V9I6.11544

Korsakpaisarn, O., Noraset, T., Lapamnuaypol, J., & Jin'No, K. (2026). *Optimizing BERT for Sentiment Classification of Amazon Product Reviews: A Study on Class Imbalance and Misclassification Analysis*. 1–6. https://doi.org/10.1109/isai-nlp66160.2025.11320736

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *EMNLP 2021 - 2021 Conference*

*Halvino Iqbal Nataprawira

on Empirical Methods in Natural Language Processing, Proceedings, 10660–10668. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.833

Laba, N. (2024). Engine for the imagination? Visual generative media and the issue of representation. *Media, Culture and Society*, *46*(8), 1599–1620. https://doi.org/10.1177/01634437241259950

Mariam, S., & Nurhaida, I. (2025). Analisis Sentimen berbasis Deep Learning Terhadap Kesetaraan Gender di Bidang STEM: Perspektif dan Implikasinya. *Edumatic: Jurnal Pendidikan Informatika*, *9*(1), 69–78. https://doi.org/10.29408/edumatic.v9i1.29071

Miyazaki, K., Murayama, T., Uchiba, T., An, J., & Kwak, H. (2024). Public perception of generative AI on Twitter: an empirical study based on occupation and usage. *EPJ Data Science 2024 13:1*, *13*(1), 2-. https://doi.org/10.1140/epjds/s13688-023-00445-y

Mosbach, M., Andriushchenko, M., & Klakow, D. (2021). *On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines*. https://openreview.net/forum?id=nzpLWnVAyah

Nashrullah, M. A., Puspitaningayu, P., Agustin Tjahyaningtijas, H. P., Fransisca, Y., & Fikril Alan, M. B. (2025). *BERT-based Transfer Learning for Two-Class Sentiment Detection in Indonesian X apps Comments*. 139–143. https://doi.org/10.1109/ICVEE66651.2025.11281404

Perwira, R. I., Permadi, V. A., Purnamasari, D. I., & Agusdin, R. P. (2025). Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content. *Journal of Information Systems Engineering and Business Intelligence*, *11*(1), 30–40. https://doi.org/10.20473/JISEBI.11.1.30-40

Saputra, R., Pristyanto, Y., & Fajri, I. N. (2025). Generative AI Image Sentiment Analysis on Social Media X using TF-IDF and FastText. *Journal of Applied Informatics and Computing*, *9*(5), 2509–2520. https://doi.org/10.30871/JAIC.V9I5.10627

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/J.PROCS.2021.01.199

Sindu, pande, Permana, A. A. J., & Wijaya, I. N. S. W. (2024). IDENTIFIKASI DAN NORMALISASI TEKS SLANG DENGAN FASTTEXT PADA TWITTER DALAM BAHASA INDONESIA. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *21*(1), 33–44. https://doi.org/10.23887/jptkundiksha.v21i1.66381

Søgaard, A., Ebert, S., Bastings, J., & Filippova, K. (2021). We Need To Talk About Random Splits. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 1823–1832. https://doi.org/10.18653/v1/2021.eacl-main.156

Syahputra, R., Yanris, G. J., & Irmayani, D. (2022). SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *6*(2), 671–678. https://doi.org/10.33395/sinkron.v7i2.11430

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. *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, AACL-IJCNLP 2020*, 843–857. https://doi.org/10.18653/v1/2020.aacl-main.85

*Halvino Iqbal Nataprawira