Journal of Deep Learning, Computer Vision and Digital Image Processing
E-ISSN: 2986-8939 P-ISSN: 2986-8920
Vol. 4, No. 2, June 2026
DOI. https://doi.org/10.61255/decoding.v4i2.1144

Comparative Analysis of IndoBERT and BiLSTM For Public Sentiment Classification Toward The Indonesian National Police on Youtube

Hardeva Satria Hazz✉
(Institut Sains dan Bisnis Atma Luhur, Pangkalpinang, Indoensia)

Delpiah Wahyuningsih
(Institut Sains dan Bisnis Atma Luhur, Pangkalpinang, Indoensia)

OPEN ACCESS

ARTICLE HISTORY
Received: April, 29 2026
Revised: May, 25 2026
Accepted: June, 19 2026

KEYWORDS
BiLSTM;
IndoBERT;
Indonesian national police;
Sentiment classification;
YouTube comments.

ABSTRACT

Purpose – This study aimed to compare the performance of IndoBERT and Bidirectional Long Short-Term Memory (BiLSTM) in classifying public sentiment toward the Indonesian National Police (INP) based on YouTube comments. This study sought to identify a robust sentiment classification model to support text-based public perception monitoring, particularly under a highly imbalanced sentiment distribution.

Method – YouTube comments were collected using the YouTube Data API. A total of 8,268 raw comments were obtained, and 7,197 comments were retained as the final dataset after preprocessing, automatic pseudo-labeling, and confidence filtering using a 0.5 threshold. To address concerns regarding threshold selection, an additional sensitivity analysis was conducted using confidence thresholds of 0.65 and 0.75. The experiment applied a dual-track preprocessing pipeline, cost-sensitive learning through class-weighted loss, bootstrap confidence interval analysis, and BiLSTM preprocessing ablation.

Findings – The results show that IndoBERT achieved stronger performance than BiLSTM. IndoBERT obtained an accuracy of 92.92% and a Macro-F1 Score of 0.8548, whereas BiLSTM achieved an accuracy of 76.11% and a Macro-F1 Score of 0.6124. Bootstrap analysis showed a Macro-F1 difference of 0.2424, with a 95% confidence interval of 0.1870 to 0.2959, indicating that IndoBERT's advantage was statistically significant. Sensitivity analysis also confirmed that IndoBERT consistently outperformed BiLSTM across all the tested thresholds.

Research Implications – The findings indicate that IndoBERT is more suitable for Indonesian sentiment classification in public perception monitoring than other models. However, because the dataset labels were generated using a BERT-based classifier, the evaluation may contain architectural circularity that favors the IndoBERT model. Future studies should use human-annotated gold-standard data and broader cross-platform validations.

Originality – This study provides a comparative evaluation of transformer-based and recurrent models using sensitivity analysis, bootstrap testing, cost-sensitive learning, and pre-processing ablation under imbalanced sentiment conditions.

Correspondence Author: ✉hardeva23@gmail.com
To cite this article : Hazz, S. H. & Wahyuningsih, D. (2026). Comparative Analysis of IndoBERT and BiLSTM For Public Sentiment Classification Toward The Indonesian National Police on Youtube. *Journal of Deep Learning, Computer Vision and Digital Image Processing, 4*(2), 58-82. https://doi.org/10.61255/decoding.v4i2.1144

INTRODUCTION

The Indonesian National Police, hereafter referred to as Polri or the Indonesian National Police (INP), has a central responsibility to maintain public security, enforce the law, and provide protection and services to citizens [1]. The development of digital communication has shifted public interaction with state institutions into more open online spaces, where institutional policies, public services, and law enforcement actions can be directly evaluated by the public. This transformation has created a dynamic discussion space that reflects how citizens perceive the performance and credibility of public institutions.

Public perception of the police is closely related to public trust, legitimacy, and institutional accountability. Public cooperation with law enforcement institutions is influenced not only by legal authority but also by how citizens evaluate fairness, responsiveness, and institutional credibility [2]. Therefore, monitoring public responses in digital spaces can help institutions identify emerging issues, evaluate communication strategies, and understand how their performance is perceived.

Public participation in monitoring the performance of state apparatuses is no longer limited to conventional complaint channels; it has broadly expanded into various social media platforms. In the context of police-related services, previous sentiment studies have shown that public responses to digital police services can be analyzed computationally to identify dominant public attitudes [3]. This indicates that digital opinion data can support institutional evaluation when processed using appropriate Natural Language Processing methods.

YouTube has emerged as a multimedia platform where citizens respond to news videos, public statements, interviews, and public controversies related to law enforcement institutions. Unlike short-text platforms such as Twitter or X, YouTube comments are often connected to audiovisual narratives and may contain longer, linked, and context-dependent responses. Prior YouTube-based sentiment studies have also shown that comment sections can provide rich public opinion data for classification tasks [4].

The utilization of public comments from digital platforms can complement structured survey methods to capture spontaneous public responses. Users often express criticism, support, disappointment, and emotional reactions more directly in the comments section. These digital traces can be a strategic source of information for evaluating public perceptions and supporting crisis communication strategies.

Although YouTube comment sections provide abundant opinion data, their linguistic characteristics are irregular. Indonesian internet users frequently combine formal Indonesian, slang, local dialects, abbreviations, emoticons, and nonstandard spelling in a single comment. This complicates sentiment classification because polarity is not always expressed through explicit positive or negative words [5], [6].

The process of extracting sentiment patterns from thousands of public comments is inefficient when performed manually. Manual classification is also vulnerable to subjectivity, fatigue, and inconsistency among the annotators. Therefore, Natural Language Processing (NLP) is needed to support a faster and more consistent classification process for large-scale social media texts.

Traditional machine learning methods, such as naïve Bayes, support vector machines, and random forests, have been widely used for sentiment classification. These methods are useful as baseline classifiers but rely heavily on manually engineered features, such as term frequency, word occurrence, and vector weighting schemes. Consequently, their ability to capture contextual relationships among words in informal sentence structures is limited [7].

Deep learning approaches have been introduced to overcome the limitations of manually engineered features. Studies comparing LSTM and BERT-based models in Indonesian social media sentiment analysis show that neural architectures can capture deeper textual patterns than traditional feature-based methods [8]. However, the performance of each architecture still depends on the characteristics of the dataset, preprocessing strategies, and class distribution.

Long Short-Term Memory (LSTM) is a recurrent neural network architecture designed to address the vanishing gradient problem and preserve information across longer sequences [9]. Its bidirectional extension, Bidirectional LSTM (BiLSTM), processes text in both forward and backward directions, enabling the model to capture broader contextual information than a unidirectional LSTM [10]. This capability makes the BiLSTM relevant for sentiment classification tasks that require sequential understanding.

In sentiment analysis studies, LSTM- and BiLSTM-based models have been applied to various textual domains, including news, social media, and public issue discussions. These studies indicate that recurrent models can achieve competitive performance when sentiment cues are relatively explicit and the input text is properly represented [11], [12]. However, sequential models may still face difficulties when dealing with noisy comments, implicit sentiments, long-distance dependencies, and informal expressions.

The weaknesses of sequential memory-based architectures have encouraged the adoption of transformer-based models. BERT introduced a bidirectional representation learning mechanism that conditions both the left and right contexts across all layers [13]. Through the self-attention mechanism, BERT-based models can assign different importance weights to tokens that contribute to the sentence meaning.

IndoBERT has become one of the most important transformer-based models for Indonesian NLP tasks. The indobenchmark/indobert-base-p1 model used in this study is associated with the IndoNLU benchmark, which provides Indonesian natural language understanding resources and pretrained models [14]. Therefore, the IndoNLU reference is essential for explaining the IndoBERT architecture used in this study.

Several studies have applied IndoBERT to Indonesian sentiment classification tasks in political discourse, public service reviews, and misinformation detection. These studies generally indicate that IndoBERT performs well in capturing contextual patterns in Indonesian text because it is built on a transformer-based architecture [15], [16]. Nevertheless, comparative studies between IndoBERT and BiLSTM on YouTube comments related to the public perception of law enforcement institutions remain limited.

YouTube-based sentiment studies in the Indonesian context have begun to grow, including studies on entertainment content, political debates, and public policy. A previous study compared LSTM and IndoBERT for sentiment analysis of YouTube comments related to the 2024 Indonesian presidential debate [17]. However, similar comparative evaluations remain limited in the context of public perception of law enforcement institutions, particularly the Indonesian National Police.

The complexity of sentiment classification is influenced not only by the model architecture but also by the class distribution in the dataset. Public discussions concerning law enforcement institutions often contain a high proportion of critical or negative comments because users tend to respond more actively when they feel dissatisfied with the police. If this imbalance is not handled properly, a model may achieve high overall accuracy while still failing to recognize minority classes, such as positive or neutral sentiments [18].

Cost-sensitive learning is a strategy for reducing the prediction bias caused by imbalanced data. This approach modifies the learning objective by assigning higher penalties to prediction errors in minority classes, encouraging the model to learn patterns beyond the dominant class [19]. Other imbalance-handling strategies, such as resampling, have also been used in sentiment analysis, but they may alter the original distribution of public opinion data [20].

A comprehensive evaluation metric is required to assess the model performance under imbalanced class conditions. The Macro-F1 Score is more appropriate than accuracy because it gives equal importance to each sentiment class, regardless of its frequency in the dataset [21]. Therefore, this metric is suitable for evaluating whether a model can recognize minority sentiment classes as well as the dominant class.

A review of prior studies shows two main gaps addressed by this study. First, many Indonesian sentiment studies still evaluate a single model architecture or compare models in contexts outside the perception of public institutions. Second, although comparative studies involving IndoBERT have been conducted in Indonesian text classification, the specific comparison between IndoBERT and BiLSTM on YouTube comments related to the Indonesian National Police remains unexplored [22].

This study was conducted to identify a sentiment classification model that is robust for noisy and imbalanced Indonesian YouTube comments. The comparison focuses on IndoBERT as an attention-based transformer model and BiLSTM as a recurrent sequence-based model. The findings are expected to provide empirical insights for developing text-based public perception monitoring systems in institutional communication contexts.

The primary objective of this study was to compare the performance of IndoBERT and BiLSTM in classifying public sentiment toward the Indonesian National Police based on YouTube comments. This experimental study evaluated both models using the same dataset, dual-track preprocessing scheme, stratified data splitting, and cost-sensitive learning. The hypothesis proposed in this study is that IndoBERT will achieve a higher Macro-F1 Score than BiLSTM because its self-attention mechanism is better suited to preserving contextual relationships in noisy and imbalanced Indonesian social media text.

METHOD

This study employs a comparative quantitative method with an experimental approach to evaluate the performance of two deep learning architectures for Indonesian sentiment classification. The models compared in this study were IndoBERT as a Transformer-based model and Bidirectional Long Short-Term Memory (BiLSTM) as a recurrent sequence-based model. The comparison was designed to assess how both architectures process noisy, informal, and imbalanced YouTube comment data related to public perceptions of the Indonesian National Police.

Data collection was performed automatically using the YouTube Data API v3 to retrieve public comments from videos related to the Indonesian National Police [23]. The search process used eight keywords representing institutional performance, traffic enforcement, public service, internal supervision, and police-related issues. The search was limited to videos published from early 2025 to April 2026, with a maximum target of 30 videos per keyword and a minimum threshold of 50 comments per video per keyword.

The extraction process retrieved top-level comments from each selected video using pagination via the nextPageToken parameter. Duplicate video links were removed using video_id to prevent the same video from being collected using different search keywords. The modeling process used only comment text and sentiment labels, while metadata such as usernames, publication time, and video title were not used as predictive features.

Ethical considerations were applied because the dataset originated from public, user-generated content on YouTube. The data collection process was limited to publicly available comments accessed through the YouTube Data API and was used only for academic text analysis in accordance with the platform's API service framework [24]. Usernames were not used in the modeling, and any examples presented in the manuscript were anonymized or masked to reduce privacy risks.

The collected comments underwent text pre-processing to remove irrelevant elements and standardize the input format. The preprocessing stage included null-value removal, duplicate comment removal, case folding, URL removal, mention removal, hashtag removal, punctuation removal, number removal, emoji removal, repeated character normalization, and whitespace normalization. These cleaning steps were applied to reduce noise while preserving the main semantic content of comments.

To handle informal Indonesian expressions, the comments were normalized using an Indonesian colloquial lexicon. This normalization process converted slang, abbreviations, and non-standard

spellings into more formal, Indonesian equivalents. This step was necessary because YouTube comments often contain informal writing patterns that may reduce the consistency of textual representations.

In this study, a dual-track preprocessing scheme was applied to adapt the text format to the input characteristics of each model. The IndoBERT track retained stop words and conjunctions after cleaning and normalization because BERT-based models rely on contextual token relationships. The BiLSTM track applied additional Indonesian stopword removal using the Sastrawi Python library to produce a denser word sequence for recurrent modeling [26].

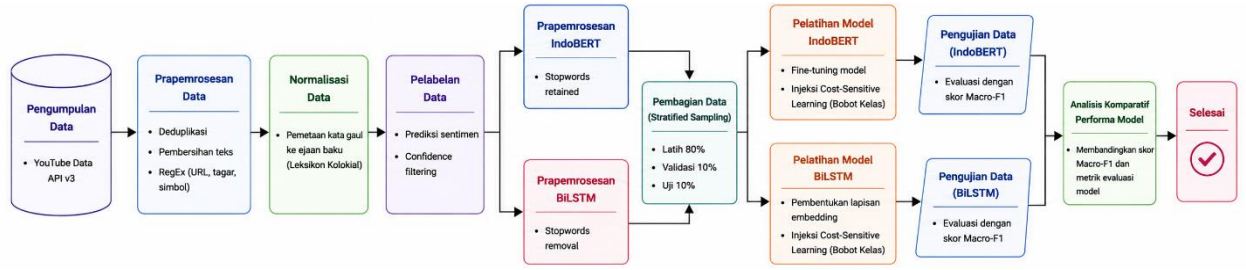

Figure 1. Comparative Stages of IndoBERT and BiLSTM Sentiment Classification

After preprocessing, comments containing fewer than three words were removed because very short comments generally provide a limited semantic context for sentiment classification. The filtered text was then prepared for automatic annotation and training of the model. This filtering was applied to maintain the minimum textual information required for the three-class sentiment classification.

Sentiment labeling was conducted through automatic pseudo-labeling using the pre-trained fine-tuned Indonesian sentiment classifier mdhugol/indonesia-bert-sentiment-classification from the Hugging Face library [27]. The model assigned each comment to one of three sentiment classes: positive, neutral, or negative. The labeling process was performed on the IndoBERT preprocessing track because this track preserved a more complete sentence context for the BERT-based classifier.

The pseudo-labeling process generated sentiment labels and confidence scores for each comment. A confidence threshold of 0.5 was used as the main filtering criterion to remove predictions with lower classifier confidence while retaining sufficient data for training and evaluating the model. Comments with confidence scores below this threshold were excluded from the main experimental data set.

To address the methodological sensitivity of the confidence threshold, we conducted a threshold sensitivity analysis using alternative thresholds of 0.65 and 0.75. This analysis examined whether stricter confidence filtering substantially changed the label distribution and model performance. The sensitivity analysis is reported in the Results section to evaluate whether the main conclusion remained stable under more conservative filtering settings.

The class distribution after pseudo-labeling showed a severe imbalance, with negative sentiments being the dominant class. To reduce the prediction bias toward the majority class, this study applied cost-sensitive learning using class-weighted cross-entropy loss. The class weights were computed using the balanced class-weighting approach implemented in Scikit-learn [28].

The class weight for each sentiment class was calculated using the following formula:

$$\boldsymbol{w_c} = \frac{\boldsymbol{N}}{\boldsymbol{Kxn_c}} \qquad (1)$$

In this formula, $\boldsymbol{w_c}$ represents the weight of class c, N represents the total number of samples, K represents the number of sentiment classes, and $\boldsymbol{n_c}$ represents the number of samples in class c. The computed class weights were converted into tensor format and injected into the cross-entropy loss function during model training.

The dataset was divided into training, validation, and testing subsets using stratified sampling methods. The split proportions were set at 80% for training, 10% for validation, and 10% for testing. This strategy was used to preserve the original sentiment class ratio across all the data subsets.

The IndoBERT model was developed using indobenchmark/indobert-base-p1 as the pretrained base model [14]. The model was fine-tuned for three-class sentiment classification using the Hugging Face Transformers Framework [29]. Tokenization was performed with a maximum input length of 128 tokens, with padding and truncation applied to standardize the sequence length.

The IndoBERT fine-tuning process used five training epochs, a learning rate of 2e-5, a training batch size of 16, and an evaluation batch size of 32. The training configuration also applied a weight decay of 0.01 and a warmup ratio of 0.1. The best model was selected based on the validation Macro-F1 Score using early stopping with a patience value of 2.

The BiLSTM model was built from scratch using word-index sequences generated from the BiLSTM pre-processing track. The vocabulary was constructed from the training data, and each input sequence was padded or truncated to a maximum of 100 tokens. This configuration ensured that the recurrent model processed a consistent sequence length during the training and evaluation.

The BiLSTM architecture consisted of an embedding layer with an embedding dimension of 256, two bidirectional LSTM layers with 128 hidden units, a dropout layer with a dropout rate of 0.5, and a fully connected output layer for the three sentiment classes. The model was trained using Adam optimization with a learning rate of 5e-4, weight decay of 1e-4, and batch size of 64. A ReduceLROnPlateau scheduler was applied with a reduction factor of 0.5 and patience of 1, while early stopping used a patience value of 5 based on the validation Macro-F1 Score.

Both IndoBERT and BiLSTM were trained using class-weighted cross-entropy losses. This ensured that both architectures used the same imbalance handling strategy during optimization. Therefore, the comparison focused on architectural differences rather than differences in the loss function treatment.

To empirically examine the preprocessing design, additional BiLSTM preprocessing ablation was conducted. The first BiLSTM configuration used the BiLSTM preprocessing track with stopword removal, whereas the second configuration used the IndoBERT preprocessing track without stopword removal as the BiLSTM input. Both configurations used the same data split and training configuration so that the effect of stop word removal on BiLSTM performance could be evaluated more directly.

The classification performance was evaluated using accuracy, macro precision, macro recall, Macro-F1 Score, Weighted-F1 Score, classification report, and confusion matrix. The Macro-F1 Score was used as the primary evaluation metric because it gives equal importance to each sentiment class, regardless of its frequency in the dataset. This metric was considered more appropriate than accuracy because the dataset contained a highly imbalanced sentiment distribution [31].

To assess the statistical reliability of the performance gap between the two models, we computed a bootstrap confidence interval for the Macro-F1 difference. Bootstrap analysis was conducted using 5,000 resampling iterations on the same test set. The confidence interval was used to examine whether the observed Macro-F1 advantage remained stable under repeated resampling.

This study also conducted a token truncation analysis for the IndoBERT input sequence. The analysis measured the number and proportion of comments exceeding the 128-token limit used for fine-tuning. This step was included to evaluate whether truncation could become a meaningful source of information loss in the IndoBERT evaluation process.

A qualitative error analysis was also performed by comparing the prediction outcomes of IndoBERT and BiLSTM on the same test set. The analysis grouped errors into three categories: IndoBERT correct and BiLSTM wrong, IndoBERT wrong and BiLSTM correct, and both models were incorrect. Selected examples from these categories were examined to identify recurring linguistic challenges, such as implicit sentiment, sarcasm, short comments, and ambiguous expressions.

This study focuses exclusively on text-based sentiment classifications. Other variables, such as view counts, like counts, channel identity, video duration, upload time, and engagement statistics, were not included in the modeling process. This limitation was applied to keep the comparison focused on the ability of IndoBERT and BiLSTM to process the comment text.

A methodological limitation of this study is the use of automatic pseudo-labeling instead of human-annotated gold-standard labels. Because the pseudo-labeling model is based on a BERT-like architecture, there is potential architectural circularity that may favor IndoBERT during comparative evaluation. Therefore, the results should be interpreted as performance on silver-standard labels and require further validation using human annotations.

RESULTS

Data Collection

The initial stage of this study focused on collecting public comment data from the YouTube platform using the YouTube Data API, v3. The data collection process was executed automatically using the YouTubeDataExtractor class to retrieve the video metadata and top-level comment texts. This approach enabled the collection of a large-scale comment corpus while preserving the relationship between each comment and its corresponding video.

The search parameters were configured to focus on recent public discussions regarding the Indonesian National Police. The researcher restricted the search to videos published between January 2025 and April 2026 and applied a minimum threshold of 50 comments per video. This limitation was used to ensure that each selected video contained sufficient public interaction for the sentiment analysis.

Data collection used eight keywords reflecting various public perspectives on police institutions. These keywords cover aspects of institutional performance, traffic enforcement, internal supervision, case handling, and field operational issues. The categories and search keywords used in the data extraction process are listed in Table 1.

Table 1. Categories and Video Search Keywords

Query Category	Search Keywords
Performance and Institutional	*"kinerja polri", "prestasi polri"*
Traffic Enforcement	*"polisi lalu lintas", "razia polisi"*
Cases and Supervision	*"oknum polisi viral", "propam polri", "bareskrim polri", "kasus polisi"*

Based on these eight keywords, the system searched for a maximum of 30 videos for each keyword. The collected video list was then filtered using video_id to remove duplicate videos that appeared under different keywords. This final filtering process identified 84 unique videos that met the criteria as comment-data sources.

Comments were extracted from the selected videos using the fetch_comments function with a maximum quota of 120 top-level comments per video. The system used the nextPageToken parameter to continue downloading comments from the YouTube server until the specified quota was reached, or no further comments were available. The final raw dataset used in this study consisted of 8,268 rows of comments.

The collected raw dataset contained five main attributes: video_id, author, comment_text, published_at, and title. These attributes were retained to support data tracing prior to the preprocessing stage. However, only the comment text was used as the primary input for sentiment

classification, whereas metadata such as usernames and publication time were not used as predictive features.

The collected comments represent authentic social media writing patterns, including informal spelling, abbreviations, emoticons, punctuation noise, and inconsistent sentence structures. These characteristics indicate that the raw dataset requires further preprocessing before being used for model training. A sample of the raw comment data is presented in Table 2.

Table 2. Raw Comment Data

video_id	author	comment_text	published_at	title
v0JJL-b-GGc	User_001	*Jujur razia helm dadakan itu salah satu masyarakat tidak suka polisi*	2025-11-24T13:23:45Z	*Warga Lebih Percaya Damkar Dibanding Polisi, Kapolri Janji Perbaiki Kinerja*
rMu_o_pkEUw	User_002	*Polisi hadir cuma untuk perusak dinegeri ini*	2026-03-17T02:36:23Z	*DPR RI Kritik Kinerja Polri: Kenapa Polisi Suka Sekali Menjadikan Tersangka Orang yang Jadi Korban!*
qDZfghD-8B8	User_003	*Polri untuk Rakyat dan pejabat yg utama.*	2025-12-30T13:45:46Z	*Akuntabilitas Kinerja Polri 2025: Penegakan Hukum, Keamanan, dan Pelayanan Masyarakat #beritasatu*
su97rc1wErA	User_004	*Minta maaf doang dan minta maaf Mulu gada peningkatan jad lebih baik*	2025-12-31T06:24:25Z	*Kapolri Minta Maaf atas Kinerja Polri Sepanjang 2025*
Ja6gRFMufwQ	User_005	*Harapan saya selaku masyarakat Indonesia. Bintang ⭐⭐⭐ Dan ⭐⭐⭐⭐ Dengan harapan sesegera mungkin di pensiunkan. Dan Bintang ⭐⭐ Dibawah Gubernur dan KOMPOLNAS. Optimis masyarakat dapat merasakan pelayanan Kepolisian. TERIMAKASIH 🙏.*	2025-11-25T06:52:03Z	*(LENGKAP) WAKAPOLRI KOMJEN DEDI PRASETYO BLAK BLAKAN SOAL RAPOR MERAH POLRI*

The author values in Table 2 have been anonymized to avoid displaying identifiable YouTube usernames in the manuscript. This adjustment aligns with ethical considerations, as user identifiers were not analyzed as model features. The raw dataset was then exported to raw_sentimen_polri_dataset.csv for the preprocessing stage.

Data Preprocessing

The preprocessing stage commenced with the removal of duplicate and invalid entries from the raw dataset. From 8,268 raw comments, the deduplication process removed 128 repeated comments identified as exact duplicates or copied comments. This process reduced the dataset to 8,140 unique comments for further text cleaning and analysis.

Advanced text cleaning was then applied to standardize the comment structure. This stage included case folding, URL removal, mention removal, hashtag removal, punctuation removal, number removal, emoji removal, repeated character normalization, and whitespace normalization. These steps were used to reduce textual noise while retaining the main semantic content of each comment.

To handle informal Indonesian expressions, cleaned comments were normalized using a colloquial Indonesian lexicon containing 4,330 entries. This process converted slang, abbreviations, and

nonstandard spellings into more formal Indonesian equivalents. Text normalization is important because inconsistent writing patterns can affect the quality of token representation in sentiment classification models [32].

In this study, we applied a dual-track preprocessing pipeline to match the input characteristics of each model architecture. The IndoBERT track retained stop words and conjunctions after cleaning and normalization to preserve the sentence context. The BiLSTM track applied additional stopword removal to produce a more compact word sequence for recurrent modeling.

The application of this dual-track mechanism produced two clean-text versions from the same raw comment. The IndoBERT model received a more complete sentence structure, whereas the BiLSTM model received a shorter sequence after stopword removal. The differences between the two preprocessing outputs are listed in Table 3.

Table 3. Dual-Track Preprocessing Results

Raw Comment	IndoBERT Preprocessing	BiLSTM Preprocessing
Jujur razia helm dadakan itu salah satu masyarakat tidak suka polisi	*jujur razia helm dadakan itu salah satu masyarakat tidak suka polisi*	*jujur razia helm dadakan salah satu masyarakat suka polisi*
Polisi hadir cuma untuk perusak dinegeri ini	*polisi hadir cuma untuk perusak dinegeri ini*	*polisi hadir cuma perusak dinegeri*
Polri untuk Rakyat dan pejabat yg utama.	*polri untuk rakyat dan pejabat yang utama*	*polri rakyat pejabat utama*
Minta maaf doang dan minta maaf Mulu gada peningkatan jad lebih baik	*meminta maaf doang dan meminta maaf mulu enggak ada peningkatan jadi lebih baik*	*meminta maaf doang meminta maaf mulu enggak peningkatan jadi lebih baik*
Harapan saya selaku masyarakat Indonesia. Bintang ☆☆☆ *Dan* ☆☆☆☆ *Dengan harapan sesegera mungkin di pensiunkan. Dan Bintang* ☆☆ *Dibawah Gubernur dan KOMPOLNAS.*	*harapan saya selaku masyarakat indonesia bintang dan dengan harapan sesegera mungkin di pensiunkan dan bintang dibawah gubernur dan kompolnas optimis masyarakat dapat merasakan pelayanan kepolisian polri terimakasih*	*harapan selaku masyarakat indonesia bintang harapan sesegera mungkin pensiunkan bintang dibawah gubernur kompolnas optimis masyarakat merasakan pelayanan kepolisian polri terimakasih*

After all comments were normalized, those with fewer than three words were removed from the dataset. Very short comments were excluded because they generally provide limited semantic context for the three-class sentiment classification. This filtering stage removed 865 comments, reducing the dataset from 8,140 to 7,275 valid comments prior to automatic labeling.

Exploratory Data Analysis was conducted to observe the linguistic characteristics of the preprocessed dataset. The average comment length in the IndoBERT track was 13.2 words, whereas the median length was nine words. This distribution indicates that public responses in the dataset were dominated by short, direct comments.

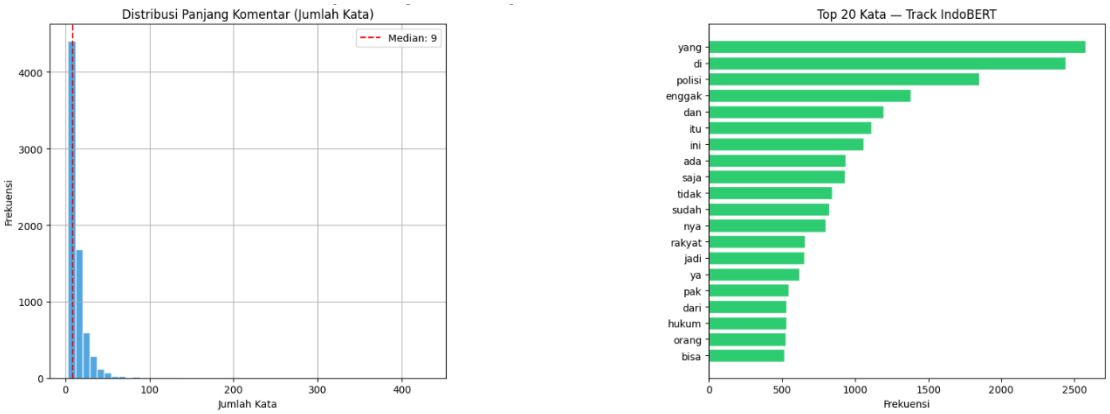

Figure 2. Distribution of Comment Lengths and Top 20 Words

Vocabulary exploration showed that the IndoBERT track contained 13,394 unique words, whereas the BiLSTM track contained 13,275 unique words. The difference between the two tracks was relatively small, indicating that stop word removal did not substantially reduce the vocabulary size. Therefore, the use of stopword removal in the BiLSTM track should be understood mainly as a sequence simplification strategy rather than as strong evidence of major vocabulary reduction.

The most frequent words in both preprocessing tracks were terms related to police institutions and law enforcement issues. Words such as “polisi,” “rakyat,” “hukum,” and “enggak” appeared frequently, reflecting the central focus of public discourse in the dataset. This pattern indicates that the collected comments were relevant to the research topic and contained recurring vocabulary related to the public evaluation of the police.

Figure 3. Word Cloud of IndoBERT Track and BiLSTM Track

The word cloud visualization reinforced that public discussion was concentrated on police performance, law enforcement, and the relationship between the authorities and citizens. However, the similarity in vocabulary size between the two tracks suggests that the dual-track design should not be interpreted as proof that stopword removal strongly reduces the computational memory load. The preprocessed dataset was then converted into a CSV format and prepared for automatic sentiment labeling.

Data Labeling

The data labeling stage was conducted automatically using the pretrained fine-tuned Indonesian sentiment classifier mdhugol/indonesia-bert-sentiment-classification. This automatic pseudo-labeling process was applied to 7,275 clean comments from the IndoBERT preprocessing track. The model classified each comment into three sentiment classes: positive, neutral, and negative sentiments.

The pseudo-labeling process produced sentiment labels and confidence scores for each prediction. The overall average confidence score was 0.9299, with a minimum value of 0.3566 and a maximum value of 0.9985. These results indicate that most predictions were produced with relatively high confidence, although several comments received low-confidence scores.

To improve the reliability of the generated labels, confidence filtering was applied with a threshold of 0.5. Predictions with confidence scores below the threshold were removed because they were considered less reliable for model training and evaluation purposes. This filtering process removed 78 comments and reduced the dataset from 7,275 to 7,197 validated comments.

The final label distribution shows a highly imbalanced sentiment composition. Negative sentiment dominated the dataset with 5,664 comments, followed by positive and neutral sentiments with 792 and 741 comments, respectively. The detailed distribution of sentiment classes after confidence filtering is shown in Table 4.

Table 4. Sentiment Class Distribution After Confidence Filtering

Sentiment Class	Number of Comments	Percentage
Negative	5,664	78.7%
Positive	792	11.0%
Neutral	741	10.3%
Total Valid	**7,197**	**100%**

This distribution indicates that public comments related to the Indonesian National Police were dominated by negative sentiments. The high proportion of negative comments reflects the strong presence of criticism in the YouTube discussions. This imbalance also created a modeling challenge because the classifier could become biased toward the majority class if no imbalance-handling strategy was applied.

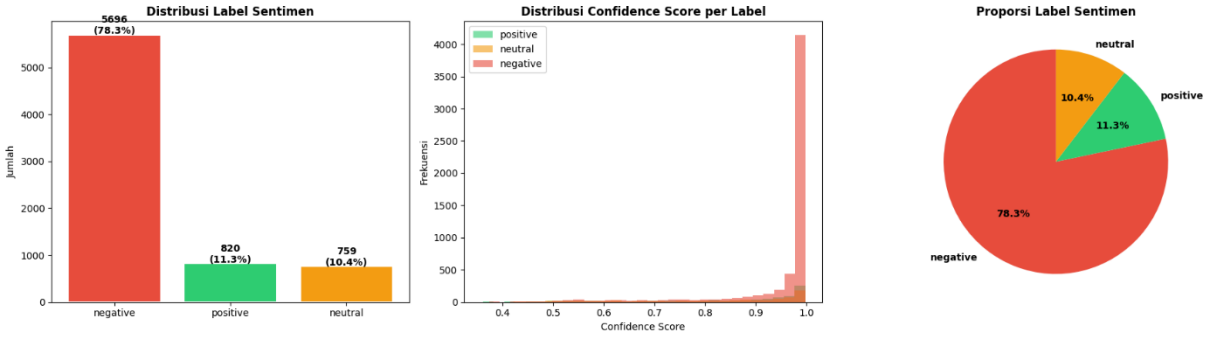

Figure 4. Sentiment Label Distribution, Confidence Score Distribution, and Label Proportion After Confidence Filtering

In addition to the overall confidence score, this study examined the confidence statistics for each sentiment class. This analysis was conducted to identify whether the pseudo-labeling model produced different confidence levels for negative, neutral, and positive comments. A summary of the class-stratified confidence scores is presented in Table 5.

Table 5. Class-Stratified Confidence Score Summary

Sentiment Class	Number of Comments	Mean Confidence	Minimum Confidence	Maximum Confidence
Negative	5,664	0.9570	0.5014	0.9985
Neutral	741	0.8364	0.5005	0.9979
Positive	792	0.8711	0.5016	0.9975

The class-stratified confidence analysis showed that the negative class had the highest mean confidence score of 0.9570. The positive class recorded a mean confidence score of 0.8711, whereas the neutral class had the lowest mean confidence score of 0.8364. This pattern indicates that the pseudo-labeling model assigned more confident predictions to negative comments than to the minority classes.

The minimum confidence scores across all classes were slightly above the filtering threshold of 0.5. Negative sentiment had a minimum confidence score of 0.5014, neutral sentiment had a score of 0.5005, and positive sentiment had a score of 0.5016. These values show that the confidence filtering mechanism successfully removed predictions below the specified threshold, although some retained labels were close to the minimum acceptance boundary.

The lower mean confidence in neutral sentiment suggests that neutral comments were more difficult for the pseudo-labeling model to distinguish between positive and negative comments. This condition is reasonable because neutral comments in social media discussions often contain factual statements, indirect criticisms, or ambiguous expressions. Therefore, the neutral class should be interpreted cautiously in subsequent comparative evaluations.

To address the methodological concern regarding the selection of the 0.5 confidence threshold, this study conducted a sensitivity analysis using two stricter alternative thresholds: 0.65 and 0.75. This analysis was used to examine whether higher filtering thresholds substantially changed the label distribution in the dataset. The label distributions for each confidence threshold are presented in Table 6.

Table 6. Confidence Threshold Sensitivity Analysis on Label Distribution

Threshold	Total Comments	Removed Comments	Negative	Positive	Neutral
0.50	7,197	78	78.70%	11.00%	10.30%
0.65	6,768	507	80.91%	10.27%	8.82%
0.75	6,492	783	82.38%	9.55%	8.07%

Sensitivity analysis showed that stricter confidence thresholds reduced the number of retained comments. At the 0.65 threshold, the dataset decreased to 6,768 comments, whereas at the 0.75 threshold, it decreased to 6,492 comments. Despite this reduction, the general class pattern remained consistent, with negative sentiment still dominating the dataset and positive and neutral sentiments remaining minority classes.

The proportion of negative sentiment increased from 78.70% at the 0.50 threshold to 82.38% at the 0.75 threshold level. Meanwhile, the positive and neutral classes gradually decreased as the filtering thresholds became stricter. This pattern indicates that higher-confidence predictions were more concentrated in the negative class, whereas minority classes contained more predictions with moderate confidence scores.

These findings support the use of 0.5 as the main operational threshold because it retained a larger amount of training data while still removing predictions that were below the minimum confidence boundary. The alternative thresholds did not change the central characteristics of the dataset, namely, its highly imbalanced distribution with negative sentiment as the majority class. Therefore, threshold selection did not substantially alter the main label distribution used in the comparative experiment.

The labels produced in this stage should be interpreted as silver standard labels rather than human-annotated gold standard labels. This distinction is important because the annotation process relies on a pre-trained classifier instead of manual validation by expert annotators. Consequently, labeling errors may still exist, especially in comments containing sarcasm, implicit sentiments, or ambiguous language.

Another methodological concern is the architectural similarity between the pseudo-labeling and IndoBERT models evaluated in this study. Because the pseudo-labeling model is based on a BERT-like architecture, the resulting labels may align more closely with transformer-based representations than with recurrent models such as BiLSTM. This potential circularity should be considered when interpreting IndoBERT's performance advantage in comparative evaluations.

Modelling

The modeling stage began by dividing the validated dataset into training, validation, and testing subsets using stratified sampling. The split proportions were set at 80% for training, 10% for validation, and 10% for testing. This process produced 5,757 comments for training, 720 for validation, and 720 for testing.

The class distribution in each subset was preserved to maintain the same sentiment ratio as that of the final dataset. This strategy was important because the data contained a dominant negative class and two minority classes. The same data indices were applied to both the IndoBERT and BiLSTM datasets to ensure a fair comparison of the models.

To address the class imbalance, both models were trained using class-weighted cross-entropy loss. The computed class weights were 0.4236 for negative sentiment, 3.0290 for positive sentiment, and 3.2375 for neutral sentiment. These weights assigned larger penalties to the prediction errors in minority classes during training.

The first model was developed using the indobenchmark/indobert-base-p1 pretrained model. IndoBERT was fine-tuned for three-class sentiment classification, with a maximum input length of 128 tokens. The training process used five epochs, a learning rate of 2e-5, a training batch size of 16, an evaluation batch size of 32, a weight decay of 0.01, and a warmup ratio of 0.1.

The IndoBERT fine-tuning process applied early stopping with a patience value of 2 based on the validation Macro-F1 Score. The training process showed stable convergence, and the validation performance improved during the training phase. IndoBERT completed the full training process in 462.6 s.

[900/900 07:38, Epoch 5/5]

Epoch	Training Loss	Validation Loss	Accuracy	Precision Macro	Recall Macro	F1 Macro
1	0.659354	0.323268	0.818056	0.681406	0.870435	0.738726
2	0.226868	0.343075	0.879167	0.746568	0.884837	0.798758
3	0.109769	0.712785	0.916667	0.837652	0.816255	0.825575
4	0.034154	0.769847	0.926389	0.857207	0.843870	0.850023
5	0.013685	0.732509	0.927778	0.857771	0.852006	0.854159

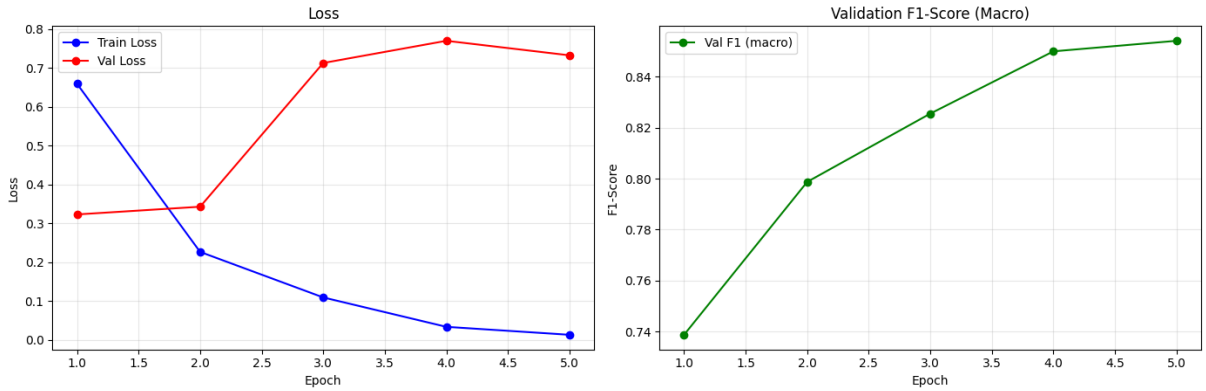

Figure 5. IndoBERT Training Process and Learning Curves

The IndoBERT learning curves showed that the validation Macro-F1 score continued to improve across epochs, although the validation loss increased after the early training phase. Because the model selection was based on the validation Macro-F1, the final checkpoint still reflected the strongest classification performance on the validation set. This pattern indicates that IndoBERT can adapt to the YouTube comment dataset, although the validation loss should be interpreted cautiously.

The second model was developed using a Bidirectional LSTM architecture trained from scratch on the BiLSTM preprocessing track. The model used a vocabulary constructed from the training data with an actual vocabulary size of 11,367 tokens. Each input sequence was padded or truncated to a maximum of 100 tokens.

The BiLSTM architecture consisted of an embedding layer with 256 dimensions, two bidirectional LSTM layers with 128 hidden units, a dropout of 0.5, and a fully connected output layer for the three sentiment classes. The model was trained using the Adam optimizer with a learning rate of 5e-4, weight decay of 1e-4, and batch size of 64. A ReduceLROnPlateau scheduler was applied with a reduction factor of 0.5 and a patience of 1.

The BiLSTM training was configured for a maximum of 30 epochs and early stopping with a patience value of 5. The validation performance began to fluctuate after several epochs, indicating that the recurrent model had difficulty maintaining stable generalization on the imbalanced dataset. Training was stopped at the 16th epoch, with a total training time of 22.4 s.

```
Training BiLSTM...
Epoch  1/30 | Train Loss: 1.0906 | Val Loss: 1.0674 | Val F1: 0.4414 ← best
Epoch  2/30 | Train Loss: 1.0370 | Val Loss: 0.9452 | Val F1: 0.4777 ← best
Epoch  3/30 | Train Loss: 0.9075 | Val Loss: 0.8899 | Val F1: 0.5293 ← best
Epoch  4/30 | Train Loss: 0.8165 | Val Loss: 0.8543 | Val F1: 0.5297 ← best
Epoch  5/30 | Train Loss: 0.7657 | Val Loss: 0.8688 | Val F1: 0.5410 ← best
Epoch  6/30 | Train Loss: 0.7054 | Val Loss: 0.8532 | Val F1: 0.5607 ← best
Epoch  7/30 | Train Loss: 0.6587 | Val Loss: 0.8948 | Val F1: 0.5496
Epoch  8/30 | Train Loss: 0.6293 | Val Loss: 0.9435 | Val F1: 0.5793 ← best
Epoch  9/30 | Train Loss: 0.5632 | Val Loss: 1.0320 | Val F1: 0.5586
Epoch 10/30 | Train Loss: 0.5231 | Val Loss: 0.9648 | Val F1: 0.5756
Epoch 11/30 | Train Loss: 0.4993 | Val Loss: 1.0084 | Val F1: 0.5918 ← best
Epoch 12/30 | Train Loss: 0.4892 | Val Loss: 1.0397 | Val F1: 0.5839
Epoch 13/30 | Train Loss: 0.4813 | Val Loss: 1.0417 | Val F1: 0.5785
Epoch 14/30 | Train Loss: 0.4562 | Val Loss: 1.0978 | Val F1: 0.5527
Epoch 15/30 | Train Loss: 0.4498 | Val Loss: 1.0928 | Val F1: 0.5580
Epoch 16/30 | Train Loss: 0.4566 | Val Loss: 1.1112 | Val F1: 0.5556

Early stopping at epoch 16 (patience=5)

BiLSTM training complete! (22.4s)
Best Val F1 (macro): 0.5918
```

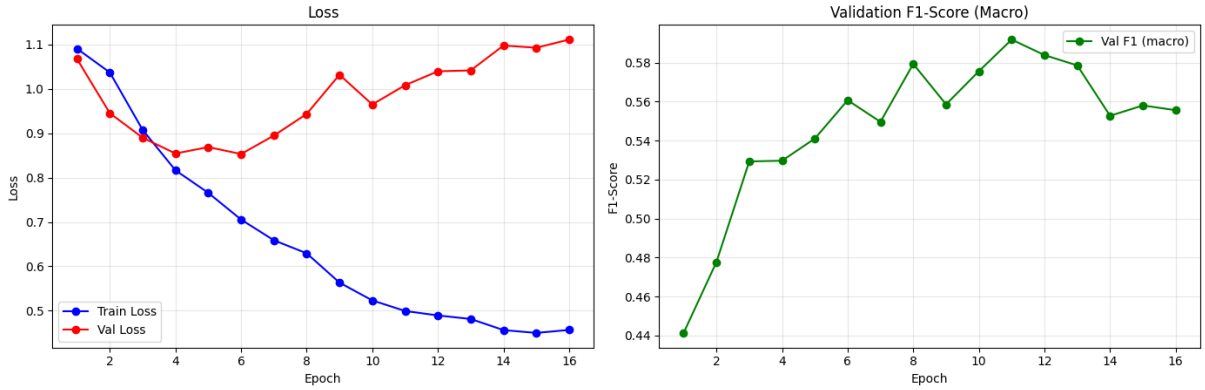

Figure 6. BiLSTM Training Process and Learning Curves

The BiLSTM learning curves showed greater fluctuations than the IndoBERT curves. Although the model was trained much faster, its validation performance was less stable across epochs. This result indicates that the BiLSTM architecture is more computationally efficient but less robust in learning sentiment patterns from noisy and imbalanced YouTube comments.

Comparative Evaluation

The comparative evaluation was conducted using a blind test set containing 720 comments. Both IndoBERT and BiLSTM were evaluated on the same test partition to ensure a fair comparison between the two architectures. The evaluation focused on accuracy, macro precision, macro recall, Macro-F1 Score, Weighted-F1 Score, confusion matrix, class-wise F1-Score, and training time.

Macro-F1 Score was used as the primary evaluation metric because the dataset contained a highly imbalanced class distribution. This metric assigns equal importance to the positive, neutral, and negative classes, regardless of the number of samples in each class. Therefore, Macro-F1 provided a more balanced assessment than accuracy in this study.

Based on the main test results at the 0.50 confidence threshold, IndoBERT outperformed BiLSTM in all classification metrics. IndoBERT achieved an accuracy of 0.9292 and a Macro-F1 Score of 0.8548, whereas BiLSTM achieved an accuracy of 0.7611 and a Macro-F1 Score of 0.6124. A detailed comparison of the evaluation metrics between the two models is presented in Table 7.

Table 7. Comparative Evaluation Metrics

Classification Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	F1-Score (Weighted)	Training Time
IndoBERT	0.9292	0.8641	0.8462	0.8548	0.9286	462.6
BiLSTM	0.7611	0.5983	0.6531	0.6124	0.7765	22.4

The Macro-F1 gap between IndoBERT and BiLSTM reached 0.2424 or 24.24 pp. This gap indicates that IndoBERT produced a more balanced classification performance across sentiment classes than BiLSTM. However, this result should be interpreted together with the bootstrap confidence interval analysis to assess the reliability of the observed performance differences.

To evaluate the statistical reliability of the Macro-F1 performance gap, a bootstrap confidence interval analysis was conducted using 5,000 resampling iterations on the same test set. The 95% confidence interval was computed for IndoBERT Macro-F1, BiLSTM Macro-F1, and the Macro-F1 difference between the two models. The bootstrap confidence interval summary is shown in Table 8.

Table 8. Bootstrap Confidence Interval for Macro-F1 Score

Metric	Observed Value	95% CI Lower	95% CI Upper
IndoBERT Macro-F1	0.8548	0.8145	0.8908
BiLSTM Macro-F1	0.6124	0.5654	0.6576
Macro-F1 Difference (IndoBERT - BiLSTM)	0.2424	0.1870	0.2959

Bootstrap analysis showed that the 95% confidence interval for the Macro-F1 difference ranged from 0.1870 to 0.2959. Because this interval did not include zero, IndoBERT's Macro-F1 advantage over BiLSTM can be considered statistically reliable for the test set. This result supports the interpretation that the observed performance gap was not merely caused by random variations in the test samples.

The confidence intervals also show that the two models occupied clearly different performance ranges. IndoBERT's Macro-F1 confidence interval ranged from 0.8145 to 0.8908, whereas BiLSTM's ranged from 0.5654 to 0.6576. This separation further supports the conclusion that IndoBERT achieved a stronger and more stable sentiment classification performance than BiLSTM.

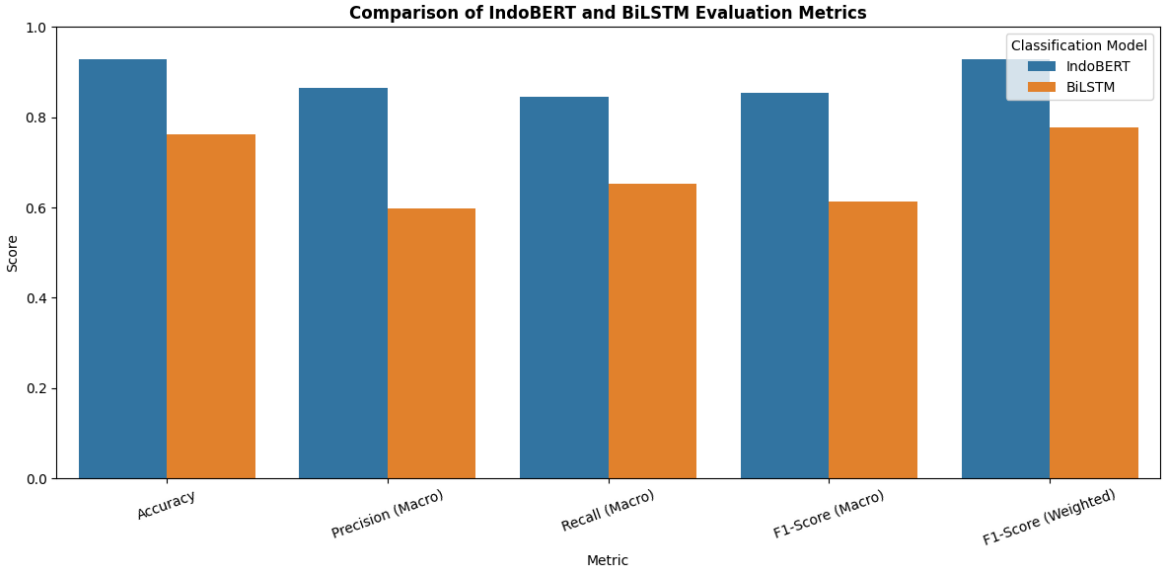

Figure 7. Comparison of IndoBERT and BiLSTM Evaluation Metrics

The largest practical difference between the two models appeared in the macro precision and Macro-F1 Score. IndoBERT achieved a macro precision of 0.8641, whereas BiLSTM achieved only 0.5983. This indicates that BiLSTM produced more incorrect predictions across sentiment classes, particularly when identifying positive and neutral comments.
An additional sensitivity analysis was conducted to examine whether the main conclusions remained stable under stricter confidence thresholds. The evaluation compared the main threshold of 0.50 with two alternative thresholds: 0.65 and 0.75. The model performance for each threshold is presented in Table 9.

Table 9. Confidence Threshold Sensitivity Analysis on Model Performance

Threshold	Final Data	IndoBERT Macro-F1	BiLSTM Macro-F1	Difference
0.50	7,197	0.8548	0.6124	0.2424
0.65	6,768	0.8456	0.5095	0.3360
0.75	6,492	0.9261	0.6106	0.3155

The sensitivity analysis showed that IndoBERT consistently outperformed BiLSTM across all tested confidence thresholds. At the 0.65 threshold, IndoBERT achieved a Macro-F1 Score of 0.8456, whereas BiLSTM decreased to 0.5095. At the stricter 0.75 threshold, IndoBERT achieved a Macro-F1 Score of 0.9261, whereas BiLSTM achieved 0.6106.

These results indicate that the main conclusion was not dependent on the selection of a 0.50 confidence threshold. Although stricter thresholds reduced the number of retained comments, IndoBERT maintained a substantial Macro-F1 advantage over BiLSTM. Therefore, the use of 0.50 as the main threshold can be justified as a practical choice that preserves more training data while producing the same comparative conclusion as stricter filtering settings.

The higher performance observed at the 0.75 threshold should be interpreted with caution. This threshold retained fewer comments and removed more moderate-confidence samples, which may have produced a cleaner and easier-to-classify subset. Therefore, the threshold sensitivity results should not be interpreted as evidence that 0.75 is universally better, but rather as evidence that IndoBERT's advantage remained stable under alternative filtering conditions.

An in-depth evaluation was conducted using confusion matrix analysis. A confusion matrix was used to examine how each model classified positive, neutral, and negative comments on the same test set. A comparison of the IndoBERT and BiLSTM confusion matrices is shown in Figure 8.

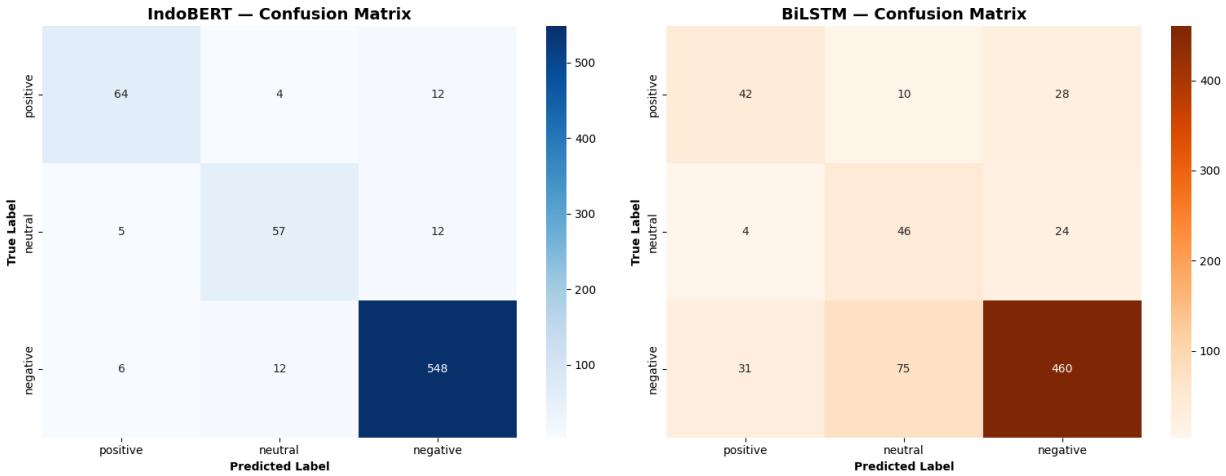

Figure 8. Comparison of IndoBERT and BiLSTM Confusion Matrices

The IndoBERT confusion matrix showed a more consistent classification across the sentiment classes. The model correctly classified 64 of 80 positive comments, 57 of 74 neutral comments, and 548 of 566 negative comments. This pattern indicates that IndoBERT maintained a strong performance in the majority negative class while still preserving relatively better recognition of positive and neutral comments.

The BiLSTM confusion matrix showed a stronger tendency to confuse minority classes with negative classes. It correctly classified 42 of 80 positive comments and 46 of 74 neutral comments, whereas 28 positive comments and 24 neutral comments were shifted into the negative class. This suggests that the class-weighted loss reduced the effect of imbalance but did not fully resolve the classification difficulty for BiLSTM.

Class-wise F1-Score analysis further confirmed the performance gap between the two models. IndoBERT achieved more stable F1-Scores across the positive, neutral, and negative classes, whereas BiLSTM showed weaker performance in the minority sentiment categories. The class-wise F1-Score comparison is shown in Figure 9.

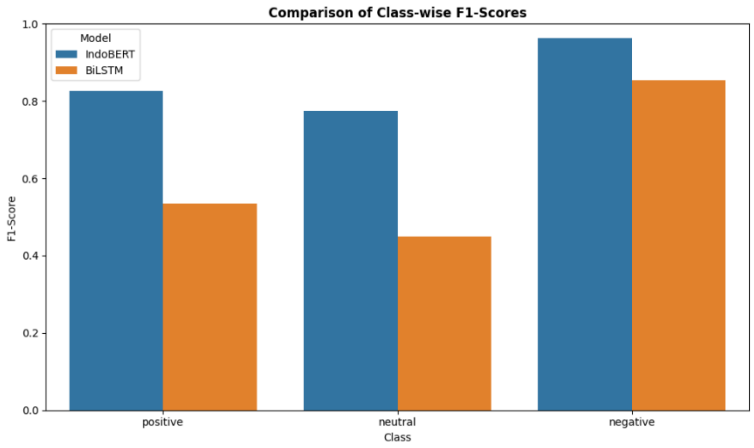

Figure 9. Comparison of Class-wise F1-Scores

The class-wise F1-Score comparison shows that IndoBERT outperformed BiLSTM in all sentiment classes. IndoBERT achieved F1-Scores of 0.8258, 0.7755, and 0.9631 for positive, neutral, and negative sentiments, respectively. In comparison, BiLSTM achieved F1-Scores of 0.5350, 0.4488, and 0.8534 for positive, neutral, and negative sentiments, respectively.

The largest performance gap appeared in the neutral class, where IndoBERT achieved an F1-Score of 0.7755, while BiLSTM reached only 0.4488. This result indicates that BiLSTM had greater difficulty in distinguishing neutral comments from positive and negative comments. This result is consistent with the earlier confidence analysis, where neutral sentiment also had the lowest mean confidence score.

Both models achieved their highest F1-Score in the negative class because negative comments dominated the dataset. However, IndoBERT still showed a stronger performance than BiLSTM in this majority class, with an F1-Score of 0.9631 compared with 0.8534. This pattern confirms that IndoBERT was not only better at recognizing the dominant class but also more stable in handling minority sentiment classes.

DISCUSSION

The experimental results show that IndoBERT achieved a stronger classification performance than BiLSTM across all evaluation metrics. This finding indicates that the self-attention mechanism in IndoBERT was more effective in capturing the contextual relationships in noisy Indonesian YouTube comments. This result is also supported by the bootstrap confidence interval analysis, which showed that the Macro-F1 difference between IndoBERT and BiLSTM remained stable under repeated resampling.

The performance advantage of IndoBERT was particularly evident in the minority sentiment classes. IndoBERT maintained a stronger Macro-F1 performance than BiLSTM, even though the dataset was dominated by negative comments. This result suggests that the transformer-based architecture was more effective in preserving the contextual information needed to distinguish positive, neutral, and negative sentiments under severe class imbalance.

The threshold sensitivity analysis further supports this interpretation. IndoBERT consistently outperformed BiLSTM at 0.50, 0.65, and 0.75 confidence thresholds. This indicates that the main conclusion was not dependent on the use of the 0.50 threshold, although the 0.50 setting was retained as the main experiment because it preserved a larger amount of training data.

The confusion matrix analysis confirmed that BiLSTM still experienced a prediction bias toward the majority class. BiLSTM confused positive and neutral comments more frequently with negative comments than IndoBERT. This indicates that the class-weighted loss reduced the imbalance effect but did not fully eliminate the difficulty faced by the recurrent model.

Qualitative error analysis provides further insight into the differences between the two models. IndoBERT correctly classified 144 comments that were misclassified by BiLSTM, whereas BiLSTM correctly classified only 23 comments that were misclassified by IndoBERT. In addition, both models misclassified 28 comments, indicating that some comments remained difficult for both architectures

Table 10. Summary of Qualitative Error Analysis

Error Category	Number of Comments
IndoBERT correct, BiLSTM wrong	144
IndoBERT wrong, BiLSTM correct	23
Both models wrong	28

Several examples show that BiLSTM often loses important contextual cues after stopword removal. For example, the comment *"enak banget jadi penjahat bisa dibela polisi..."* was correctly classified as negative by IndoBERT but predicted as positive by BiLSTM. This error suggests that BiLSTM struggled to capture implicit criticism when the sentence contained words that may appear positive when considered in isolation.

Table 11. Selected Misclassified Examples

Cleaned Comment	True Label	IndoBERT Prediction	BiLSTM Prediction	Error Pattern
rusaklah sudah organisasi kepolisian ini	Negative	Negative	Neutral	BiLSTM missed negative context
baguslah bocah bocah tenggil begini di beri pelajaran	Positive	Positive	Negative	BiLSTM confused evaluative tone
mudah an besi jembatan nya enggak jadi tersangka	Negative	Negative	Positive	BiLSTM failed to capture implicit criticism
enak banget jadi penjahat bisa dibela polisi	Negative	Negative	Positive	BiLSTM misread ironic expression
kompensasi harus kasih minimal juta buat orang	Neutral	Neutral	Negative	BiLSTM shifted neutral statement to negative

The selected examples indicate that many errors were related to implicit meaning, sarcasm, and mixed expressions. Sarcasm can affect sentiment interpretation because the surface polarity of a sentence may differ from its intended meaning [33], [34]. Comments containing words such as *"baguslah"* or *"enak"* were not always positive because they could appear in sarcastic or critical contexts.

The results also show that neutral comments are difficult to classify. This is consistent with the class-stratified confidence analysis, in which the neutral class had the lowest mean confidence score among the three sentiment classes. Neutral comments often contain factual statements, short remarks, or indirect evaluations that are harder to separate from positive or negative sentiments.

The short-comment characteristic of the dataset also influenced the model behavior. The average comment length was 13.2 words, whereas the median length was only nine words. This indicates that many comments contained limited context, making sentiment classification dependent on subtle lexical and contextual cues.

Short comments can be challenging for both recurrent and transformer-based models. For BiLSTM, short and noisy sequences may reduce the amount of sequential information available for learning stable pattern. For IndoBERT, the bidirectional attention mechanism still provides an advantage because it can weigh the available tokens more flexibly within a sentence context.

The token length analysis showed that the maximum input length of 128 tokens did not substantially affect the IndoBERT evaluation. Only 12 of the 7,197 comments exceeded 128 tokens, resulting in a truncation rate of 0.17%. Therefore, token truncation is unlikely to be a major source of performance loss in this study.

Table 12. IndoBERT Token Length and Truncation Summary

Total Comments	Token Threshold	Comments Exceeding Threshold	Truncation Rate	Mean Token Length	Median Token Length	Maximum Token Length

7,197	128	12	0.17%	17.60	13	470

The low truncation rate indicates that the 128-token limit was sufficient for almost all the comments in the dataset. Although one comment reached a maximum length of 470 tokens, such cases were rare and did not represent general data distribution. This supports the use of 128 tokens as a practical input length for IndoBERT in the experiment.

Another point that requires careful interpretation is the difference in vocabulary between the two preprocessing tracks. The IndoBERT track contained 13,394 unique words, while the BiLSTM track contained 13,275 unique words. This difference is relatively small; therefore, stopword removal should not be interpreted as a major vocabulary reduction strategy in this dataset.

The benefit of stop word removal in the BiLSTM track should instead be understood as sequence simplification. Removing stopwords can reduce some non-informative tokens from the recurrent input sequence, but it may also remove functional words that carry contextual meanings. Therefore, the effect of stopword removal needs to be examined empirically rather than assumed only from vocabulary size.

To validate the preprocessing design, we conducted an additional BiLSTM ablation experiment. The first configuration used the BiLSTM preprocessing track with stop word removal, whereas the second used the IndoBERT preprocessing track without stop word removal as the input for BiLSTM. The results of the ablation experiment are presented in Table 13.

Table 13. BiLSTM Preprocessing Ablation

BiLSTM Input Track	Stopword Removal	Accuracy	Macro-F1	Weighted-F1
BiLSTM track	Yes	0.7611	0.6124	0.7765
IndoBERT track	No	0.6875	0.5398	0.7185

The ablation results show that BiLSTM performed better when trained on the preprocessing track with the removal of stopwords. The Macro-F1 Score decreased from 0.6124 to 0.5398 when stopword removal was not applied. This result provides empirical support for the dual-track preprocessing design, particularly for using a simplified input sequence in the BiLSTM model.

However, the ablation results should not be interpreted as proof that stop word removal is universally beneficial for all recurrent models. The benefits observed in this study may be influenced by the specific dataset, vocabulary composition, class imbalance, and model configuration. Further experiments on other Indonesian social media datasets are required to evaluate the generalizability of this preprocessing choice.

Computational Efficiency

The training time comparison shows a clear computational tradeoff between the two models. IndoBERT required 462.6 s to complete training, whereas BiLSTM required only 22.4 s. This means that BiLSTM was substantially more efficient in terms of training time, although its classification performance was lower.

The faster training time of BiLSTM makes it suitable for resource-limited environments. However, the lower Macro-F1 Score and weaker minority class performance indicate that speed should not be the only deployment consideration. For institutional monitoring systems, the choice of model should balance the computational cost, prediction quality, and importance of detecting minority sentiments.

IndoBERT is more suitable when the classification quality and minority-class detection are prioritized. BiLSTM may still be considered when computational resources are limited or rapid

retraining is required. However, any practical deployment should include latency testing, real-time inference evaluation, and robustness testing of new comments before institutional use.

Methodological Considerations

The results should also be interpreted in relation to the pseudo-labeling strategy used in this study. The dataset labels were generated using a pretrained fine-tuned BERT-based sentiment classifier rather than human annotation. Therefore, the labels should be considered silver-standard labels rather than gold-standard labels.

This labeling strategy introduces potential architectural circularity. Because the pseudo-labeling model is functionally closer to IndoBERT than BiLSTM, the generated labels may align more naturally with transformer-based representations. This issue may partially contribute to IndoBERT's observed performance advantage and should be acknowledged as a limitation of the comparative evaluation study.

Therefore, future validation using human-annotated data is necessary. A gold-standard subset annotated by multiple human evaluators would help to measure the reliability of the pseudo-labeling process. This would also allow future studies to examine whether IndoBERT's advantage remains stable when evaluated against independently verified sentiment labels.

Overall, the findings indicate that IndoBERT is more robust than BiLSTM in classifying imbalanced Indonesian YouTube comments in this dataset. The advantage is supported by aggregate metrics, class-wise F1-Scores, confusion matrix patterns, bootstrap confidence intervals, confidence-threshold sensitivity analysis, and qualitative error analysis. Nevertheless, these conclusions should be understood within the constraints of silver-standard labels, platform-specific data, and retrospective evaluations.

Theoretical and Practical Implications

The findings of this study have both theoretical and practical implications for Indonesian sentiment classification research. Theoretically, this study contributes empirical evidence that a transformer-based model can outperform a recurrent BiLSTM model when applied to noisy and imbalanced Indonesian YouTube comments. This finding supports the relevance of contextual attention mechanisms for processing informal social media text, particularly when sentiment classes are unevenly distributed in the text.

This study also shows that cost-sensitive learning can be used as a practical strategy to reduce the majority-class bias in imbalanced sentiment classification. However, these results should not be interpreted as proof that IndoBERT and class weighting are universally optimal for all Indonesian sentiment analysis tasks. Their effectiveness still depends on dataset characteristics, annotation quality, preprocessing design and target deployment context.

From a practical perspective, the stronger Macro-F1 Score and minority-class performance of IndoBERT indicate its potential use as a supporting tool for public perception monitoring. For institutions such as the Indonesian National Police, this type of model may help summarize large volumes of public comments and identify emerging patterns of criticism, support, and neutral public responses. When combined with human review, such information can support communication evaluation and early issue detection.

Nevertheless, practical deployment should be approached with caution. This study was conducted on a retrospective static dataset and did not evaluate the real-time inference speed, system latency, adversarial robustness, or human-in-the-loop validation. Therefore, further testing is required before the model can be integrated into the operational digital infrastructure.

CONCLUSION

This study compared IndoBERT and BiLSTM for classifying public sentiment toward the Indonesian National Police based on YouTube comments data. The results show that IndoBERT achieved a

stronger classification performance than BiLSTM, with an accuracy of 0.9292 and a Macro-F1 Score of 0.8548. In comparison, BiLSTM achieved an accuracy of 0.7611 and a Macro-F1 Score of 0.6124.

Bootstrap confidence interval analysis further supported the reliability of the performance gap. The observed Macro-F1 difference between IndoBERT and BiLSTM was 0.2424, with a 95% confidence interval ranging from 0.1870 to 0.2959. Because this interval did not include zero, the result indicates that IndoBERT's Macro-F1 advantage was statistically reliable in the test set.

The confidence threshold sensitivity analysis showed that the main conclusion remained stable under stricter filtering thresholds. IndoBERT consistently outperformed BiLSTM at 0.50, 0.65, and 0.75 confidence thresholds. This finding supports the use of 0.50 as the main operational threshold because it retained more training data while producing the same comparative conclusion as the more conservative thresholds.

The additional BiLSTM preprocessing ablation also strengthens the empirical basis of the dual-track preprocessing design. BiLSTM achieved better performance when trained on the preprocessing track with stop word removal, reaching a Macro-F1 Score of 0.6124 compared with 0.5398 when stop word removal was not applied. This result indicates that sequence simplification through stop word removal was beneficial for BiLSTM in this dataset.

The findings suggest that the self-attention mechanism in IndoBERT is more effective for processing noisy and imbalanced Indonesian YouTube comments than the recurrent sequence-based mechanism in BiLSTM. This advantage is particularly relevant for maintaining the classification performance across minority sentiment classes. Therefore, IndoBERT is more suitable when the classification quality and minority-class detection are prioritized.

From a practical perspective, IndoBERT has the potential to be used as a supporting model for text-based public perception monitoring. The model can help summarize large volumes of public comments and identify patterns of criticism, support, or neutral responses toward public institutions. However, actual deployment requires further testing for real-time inference, latency, computational cost, robustness, and human-in-the-loop validation.

This study had an important methodological limitation. The sentiment labels were generated through automatic pseudo-labeling using a pre-trained fine-tuned BERT-based classifier; therefore, the dataset should be considered a silver standard rather than a gold standard. This creates a potential architectural circularity that may partly favor IndoBERT over BiLSTM in comparative evaluations.

Future research should validate the dataset using human-annotated gold standard labels created by multiple annotators. Further studies should compare different pseudo-labeling models, test additional confidence thresholds, and conduct broader ablation experiments on preprocessing strategies. Cross-platform evaluation using data from TikTok, X, Instagram, or other social media platforms is also needed to examine whether the findings remain stable beyond the YouTube comments.

ACKNOWLEDGMENT

The authors would like to express their appreciation to all parties who provided support during the preparation of this study. Any institutional, technical, or administrative assistance that contributed to the completion of this study is gratefully acknowledged.

AUTHOR CONTRIBUTION STATEMENT

HSH conceptualized the study, collected the data, developed the model, and performed the data analysis. DW supervised the research process, reviewed the methodology and validated the findings. Both authors contributed to manuscript writing and approved the final version of the article.

AI DISCLOSURE STATEMENT

The authors used AI-assisted language tools to refine the language and for editorial revision. All research designs, data collection, model implementation, analysis, interpretation, and final manuscript decisions were reviewed and validated by the authors. The AI tool was not used to generate research data or replace the author's responsibility for the scientific content.

REFERENCES

[1] Republic of Indonesia, "Law of the Republic of Indonesia Number 2 of 2002 concerning the Indonesian National Police," 2002.

[2] T. R. Tyler, "Enhancing police legitimacy," *The ANNALS of the American Academy of Political and Social Science*, vol. 593, no. 1, pp. 84–99, 2004, doi: 10.1177/0002716203262627.

[3] D. A. Fauzi, P. R. Nurajijah, and Engkus, "Analisis sentimen masyarakat terhadap aplikasi Digital Korlantas Polri pada ulasan Google Play Store menggunakan model IndoBERT," *Jurnal Pendidikan Sosial dan Humaniora*, vol. 5, no. 1, pp. 1181–1204, 2025.

[4] S. Riyadi, L. K. Salsabila, C. Damarjati, and R. A. Karim, "Sentiment analysis of YouTube users on Blackpink Kpop group using IndoBERT," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 8, no. 2, pp. 233-245, 2024.

[5] A. Rustamaji, R. R. Huizen, and D. P. Hostiadi, "Sentiment analysis for hotel reviews using Snowball and VADER," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 14, no. 2, 2025.

[6] N. L. Kirana, L. Muflikhah, and Indriati, "Analisis sentimen ulasan aplikasi SP4N LAPOR! dengan IndoBERT dan koreksi ejaan berbasis Levenshtein Distance," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 4, 2026.

[7] R. Tangke, D. T. Salaki, W. W. Kalengkongan, and E. Ketaren, "Analisis sentimen aplikasi TikTok menggunakan algoritma Support Vector Machine (SVM) dan Random Forest," *Jurnal TIMES*, vol. 13, no. 2, pp. 53-62, 2024.

[8] M. Khadapi and V. M. Pakpahan, "Analisis sentimen berbasis jaringan LSTM dan BERT terhadap diskusi Twitter tentang Pemilu 2024," *JUKI: Jurnal Komputer dan Informatika*, vol. 6, no. 2, pp. 130–137, 2024.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005, doi: 10.1016/j.neunet.2005.06.042.

[11] I. K. Wijaya and R. Artana, "Analisis sentimen berbahasa Inggris dengan metode LSTM studi kasus berita online pariwisata Bali," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 6, pp. 1325–1334, 2024.

[12] W. Astuti, B. Irawan, and N. A. Ramdhan, "Analisis sentimen terhadap isu pemblokiran thrifting pada platform TikTok menggunakan Bidirectional Long Short-Term Memory," *ELKOM: Jurnal Elektronika dan Komputer*, vol. 18, no. 2, pp. 332–339, 2025.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[14] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *Proc. 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*, 2020, pp. 843–857.

[15] P. Sayarizki, Hasmawati, and H. Nurrahmi, "Implementation of IndoBERT for sentiment analysis of Indonesian presidential candidates," *IndoJC*, vol. 9, no. 2, pp. 61–72, 2024.
[16] I. G. N. L. Wijayakusuma, "Perbandingan kinerja IndoBERT dan mBERT untuk deteksi berita hoaks politik dalam bahasa Indonesia," *Jurnal Sains dan Teknologi (Undiksha)*, vol. 14, no. 1, pp. 114–123, 2025.
[17] M. Fahrezi, Y. B. Pratama, and A. Pramudiyantoro, "Analisis sentimen debat publik Pilpres 2024 menggunakan metode algoritma LSTM dan IndoBERT pada platform YouTube," *JPIM: Jurnal Penelitian Ilmiah Multidisipliner*, vol. 2, no. 3, 2025.
[18] F. Salsabilla and A. Witanti, "Analisis sentimen akhir masa jabatan Presiden Jokowi pada media sosial X menggunakan Naïve Bayes," *SKANIKA*, vol. 8, no. 1, pp. 106-115, 2025.
[19] M. Z. Sarwani, M. Khoiron, and M. Udin, "Optimization of the Naïve Bayes classifier algorithm using cost-sensitive learning to detect lung diseases with an imbalanced dataset," *Journal of Artificial Intelligence and Software Engineering*, vol. 5, no. 1, pp. 332–338, 2025.
[20] E. M. O. N. Haryanto, A. K. A. Estetikha, and R. A. Setiawan, "Implementasi SMOTE untuk mengatasi imbalanced data pada sentimen analisis hotel di Nusa Tenggara Barat," *Informasi Interaktif*, vol. 7, no. 1, 2022.
[21] L. N. Hayati, F. Y. Randana, and H. Darwis, "An in-depth exploration of sentiment analysis on Hasanuddin Airport using machine learning approaches," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 2, pp. 195-208, 2025.
[22] C. Ramadhan, V. Atina, and H. Permatasari, "Analisis perbandingan model CNN dan IndoBERT dalam sentimen berita politik Indonesia," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB)*, pp. 110-118, 2025.
[23] Google Developers, "YouTube Data API Overview," *Google for Developers*. Available: https://developers.google.com/youtube/v3/getting-started.
[24] Google Developers, "YouTube API Services Terms of Service," Google for Developers. Available: https://developers.google.com/youtube/terms/api-services-terms-of-service.
[25] M. Z. Rahman, Y. A. Sari, and N. Yudistira, "Analisis sentimen Tweet COVID-19 menggunakan Word Embedding dan metode Long Short-Term Memory (LSTM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 11, pp. 5120–5127, 2021.
[26] H. Rabbani, "PySastrawi: Indonesian stemmer. Python port of PHP Sastrawi project," GitHub repository. Available: https://github.com/har07/PySastrawi.
[27] M. Hugol, "indonesia-bert-sentiment-classification," Hugging Face model repository. Available: https://huggingface.co/mdhugol/indonesia-bert-sentiment-classification.
[28] F. Pedregosa *et al*., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[29] T. Wolf *et al*., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.
[30] D. R. Alghifari, M. Edi, and L. Firmansyah, "Implementasi Bidirectional LSTM untuk analisis sentimen terhadap layanan Grab Indonesia," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 12, no. 2, pp. 89–99, 2022.
[31] P. A. Riyantoko, T. M. Fahrudin, D. A. Prasetya, T. Trimono, and T. D. Timur, "Analisis sentimen sederhana menggunakan algoritma LSTM dan BERT untuk klasifikasi data spam dan non-spam," *Prosiding Seminar Nasional Sains Data*, vol. 2, no. 1, pp. 103–111, 2022.
[32] D. Khairani, A. P. N. S. Ginting, dan R. B. Syafi'i, "Pengaruh Tahapan Preprocessing Terhadap Model IndoBERT dan IndoBERTweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 11, no. 4, pp. 887-894, 2024.
[33] D. Maynard and M. Greenwood, "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.
[34] Y. Yunitasari, A. Musdholifah, and A. K. Sari, "Sarcasm detection for sentiment analysis in Indonesian tweets," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 1, pp. 53–62, 2019.