Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

Sentiment Analysis of YouTube Comments on Indonesia-U.S. Trade Agreement: A Comparison of Machine Learning and Deep Learning

Dea Amanda*, Muhammad Iqbal, Mia Rosmiati

Faculty of Engineering and Computer Science, Computer Science, Universitas Bina Sarana Informatika Kampus Kota Pontianak, Pontianak
Jl. Abdul Rahman Saleh No. 18, Bangka Belitung Laut, Kecamatan Pontianak Tenggara, Kota Pontianak, Kalimantan Barat, Indonesia
Email: [1,*]15220479@bsi.ac.id, [2]Iqbal.mdq@bsi.ac.id, [3]mia.mrm@bsi.ac.id
Correspondence Author Email: 15220479@bsi.ac.id
Submitted: **29/06/2026**; Accepted: **18/07/2026**; Published: **20/07/2026**

Abstract−The Indonesia–United States Trade Agreement has sparked much public debate on YouTube, but due to the large volume of data and the informal nature of the text, manual analysis is ineffective. Through a comparative study of machine learning (Naïve Bayes and Support Vector Machine) and deep learning (Long Short-Term Memory and IndoBert), this research aims to identify the polarity of public opinion and evaluate the best computational model. This study contributes a comprehensive empirical comparison of the four algorithms within an identical experimental pipeline on an issue that has not been previously explored, thereby offering methodological insights for future sentiment analysis research on similarly informal, domain-specific social media text. The dataset consists of 2,284 comments labeled using the InSet Lexicon. The analysis results show that the neutral class dominates the sentiment distribution (52,0%), followed by the positive class (20,2%) and the negative class (27,8%). Performance evaluation reveals that IndoBERT significantly outperforms other models with an accuracy of 87.71% and a Macro F1-Score of 0.87. Support Vector Machine ranked second (79.00%), followed by Naïve Bayes (61.71%). In contrast, Long Short-Term Memory failed completely (accuracy of 52.00%) due to the "majority class collapse" phenomenon in small-scale datasets. This study concludes that the Transformer architecture (IndoBert) is the most robust approach for classifying sentiment in informal Indonesian-language text containing specific geopolitical and economic terms.

Keywords: Sentiment Analysis; Trade Agreements; Machine Learning; Deep Learning; IndoBert; YouTube

1. INTRODUCTION

International trade is a strategic tool in the global economy that enables countries to maximize their competitive advantages through the exchange of goods and services. Since 1949, Indonesia and the United States have maintained a strategic partnership, which will mark its seventy-fifth anniversary in 2024 [1]. However, when the United States imposed reciprocal tariffs of 32% on various Indonesian products in 2025, this raised concerns about the country's trade balance. Trade dynamics between the two countries became highly unstable [2]. In response to the 32% reciprocal tariffs, the two countries officially signed the Reciprocal Trade Agreement (ART) on February 19, 2026. Although initially positioned as a safeguard for domestic exporters, the ART contains controversial provisions such as exemptions for U.S. companies from Local Content Requirements (TKDN), restrictions on the operations of State-Owned Enterprises (SOEs), and a ban on barriers to mineral exports [3]. This policy has sparked intense public debate because it has a direct impact on economic sovereignty and the competitiveness of domestic industries.

In the digital age, the arena for public discourse has shifted to online platforms, where people actively support and oppose strategic policies. Unlike other platforms that impose character limits, YouTube comment sections tend to be more descriptive, contextual, and rich in discussion of policy issues [4]. The high volume of interactions on each video makes it a large-scale source of textual data that accurately reflects public opinion. However, the highly informal nature of YouTube comments, which contain slang, sarcasm, and non-standard structures, makes manual analysis ineffective [5]. Therefore, an automated computational approach using sentiment analysis is essential for systematically mapping the polarity of public opinion.

Conventional machine learning methods such as Naive Bayes and Support Vector Machines (SVMs) have long been used in the field of sentiment analysis. However, these methods often face challenges in capturing contextual semantics, handling non-standard structures in the Indonesian language, and understanding irony [6]. As Natural Language Processing (NLP) continues to evolve, deep learning models such as Long Short-Term Memory (LSTM) and transformers like IndoBert have emerged as solutions. LSTM excels at capturing long-term dependencies in sequential data, while IndoBert comprehensively understands the context of the Indonesian language through a bidirectional attention mechanism trained on a local corpus, making it more robust in handling informal text [7].

Several previous studies have evaluated the performance of machine learning and deep learning algorithms on various digital platforms. In the field of machine learning, SVM and Naïve Bayes have proven effective at mapping public sentiment in YouTube comments related to political issues [8],[9]. On the other hand, deep learning models such as LSTM demonstrate superior performance in capturing sequential context in informal text data on the X platform [10], [11]. Meanwhile, IndoBert has consistently outperformed traditional methods in various cases, such as sentiment analysis of the MBG program [12], a 12% increase in Value-Added Tax (VAT)

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

[13], and a review of the Tiket.com app [14]. IndoBert's ability to capture linguistic complexity was also confirmed in the Gojek review [15] and public opinion on electric cars [16].

However, the model's performance is highly dependent on the platform's feature characteristics. An interesting phenomenon was discovered by Fahrezi et al. [17] in the YouTube comment section related to the 2024 Presidential Debate, the Transformer model struggled to process slang that was too "noisy," so the LSTM model performed significantly better than IndoBERT. These findings indicate that there is no absolute consensus on the best model for YouTube data, and further evaluation is essential across various research contexts.

Based on a critical review of the latest literature, there are two significant research gaps. First, an "object gap." Previous studies have largely focused on domestic issues (such as the MBG Program, the Sandwich Generation, and electric vehicles) or reviews of commercial applications (such as Gojek and Tiket.com). No study has specifically investigated public perceptions of the Indonesia–United States Trade Agreement, a strategic geopolitical-economic issue discussed exclusively on YouTube. Second, the methodological gap. Most current studies only compare two or three algorithms separately. A comprehensive comparative study that tests all four methods (Naïve Bayes, SVM, LSTM, and IndoBert) simultaneously within an identical experimental pipeline on the same dataset has not yet been found.

Therefore, the objective of this study is to analyze public perceptions of ART in YouTube comments using machine learning methods (Naïve Bayes and SVM) and deep learning methods (LSTM and IndoBert). It is hoped that this study will provide comprehensive empirical evidence regarding which computational models are most effective and efficient for classifying sentiment in large-scale informal text data related to international trade issues. Additionally, this study will offer methodological recommendations for future researchers. The main contribution of this study is twofold: first, it provides the first empirical investigation of public sentiment toward the Indonesia–United States Trade Agreement, a geopolitical-economic issue that has not been examined in prior YouTube-based sentiment analysis research; second, it offers a comprehensive comparative evaluation of four classification algorithms within an identical experimental pipeline, thereby addressing the methodological gap left by previous studies that typically compare only two or three algorithms.

2. RESEARCH METHODOLOGY

2.1 Research Phases

This study employs a quantitative approach and a comparative experimental design. This method was used to evaluate and compare the performance of four sentiment classification algorithms Naive Bayes, Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and IndoBERT on an identical dataset. The research framework was designed through nine sequential stages: (1) data collection, (2) sentiment labeling, (3) text preprocessing, (4) class imbalance handling, (5) feature extraction, (6) data splitting, (7) modeling, (8) evaluation, and (9) comparative analysis. The research methodology for the entire research process is presented in Figure 1

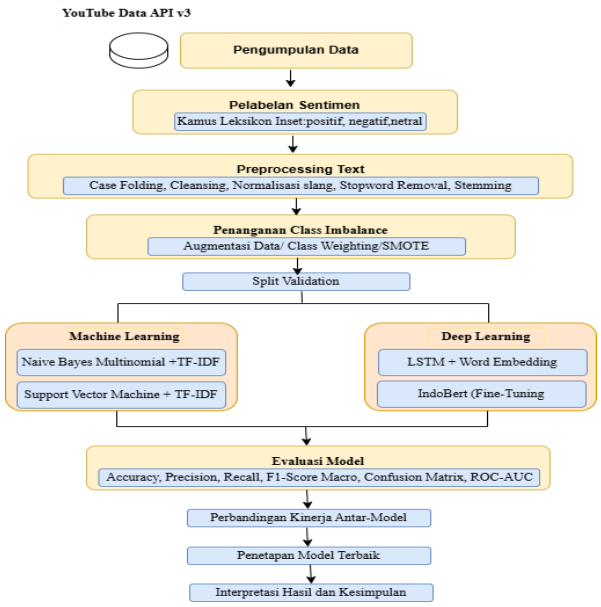

Figure 1. Research Methodology

Figure 1 shows the research methodology designed to ensure a fair comparison among the four algorithms. This study began by using the YouTube Data API v3 to collect YouTube comments. Next, automatic sentiment labeling was performed using the Inset Lexicon Dictionary (Positive, Negative, Neutral). Text processing was then applied to the data, including case normalization, correction, slang normalization, stopword removal, and

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

stemming. Subsequently, validation of the data split and handling of class imbalance were performed using data augmentation, class weighting, or SMOTE. The modeling phase encompasses two categories of approaches: (1) Machine Learning using Multinomial Naïve Bayes and Support Vector Machines, both of which utilize TF-IDF, and (2) Deep Learning using LSTM with Word Embedding and IndoBert through fine-tuning.

Each model was evaluated using the following metrics: Accuracy, Precision, Recall, Macro F1-Score, Confusion Matrix, and ROC-AUC. Next, a comparison of performance across models was conducted to identify the best model, followed by an interpretation of the results and the research conclusions.

2.2 Data Collection and Preprocessing

Research data in the form of user-generated content (UGC) comments was collected from 28 national news videos on YouTube discussing the Indonesia–United States Trade Agreement, using the YouTube Data API v3. After the crawling and initial cleaning process, 2,284 comments were obtained. Sentiment labeling at this stage was performed automatically using the InSet lexicon-based approach [18], which produces a class distribution of: Neutral (52.0%), Negative (27.8%), and Positive (20.2%). The text preprocessing stage was specifically designed to account for the informal nature of YouTube comments. As summarized in Table 1, there are different treatments for the Transformer-based Deep Learning model (IndoBert). The stopword removal and stemming stages were intentionally omitted in IndoBert because the WordPiece tokenization architecture is capable of capturing morphological context directly; thus, removing stopwords risks losing semantic information.

Table 1. Text Preprocessing Steps

Step	Description	Implementation
Case Folding & Cleansing	Converting to lowercase and removing noise (URLs, mentions, symbols) using regular expressions.	Naïve Bayes, SVM, IndoBert, LSTM
Normalisasi slang	Replacing non-standard words/abbreviations with their standard forms using a slang dictionary	Naïve Bayes, SVM, IndoBert, LSTM
Stopword removal	Removal of high-frequency words that are not semantically significant	Naïve Bayes, SVM, LSTM
Stemming	Reducing affixed words to their base forms.	Naïve Bayes, SVM, LSTM

Based on Table 1, all models underwent case folding, data cleaning, and slang normalization, as these two stages are universal in addressing noise and inconsistencies in YouTube comment text. However, stemming and stopword removal were only applied to SVM, Naive Bayes, and LSTM, which rely on word-frequency-based feature representations. IndoBert, on the other hand, was not involved in these two stages because the Transformer architecture, based on subword tokenization (WordPiece), has the intrinsic ability to directly capture the morphological and contextual representations of words through a bidirectional self-attention mechanism (high bidirectional attention that has no significant meaning)[7].

Class imbalance—where the majority class is overrepresented—is addressed through strategies tailored to the algorithm to prevent model bias. Naïve Bayes uses the Synthetic Minority Oversampling Technique (SMOTE) [9]. In addition, SVM uses class weights (class_weight='balanced')[19], and IndoBert combines Focal Loss with class weights to highlight samples that are difficult to classify.

2.3 Feature Representation and Modeling

Although each of the four models receives the same input, they are processed using different feature representation schemes. Naïve Bayes and SVM use the Term Frequency-Inverse Document Frequency (TF-IDF) feature representation, a statistical method designed to evaluate how important a word (term) is in a document relative to a larger collection of documents [20]. SVM is implemented using a linear kernel [21] and the One-vs-Rest (OvR) strategy for multi-class classification

On the other hand, LSTM can capture long-term dependencies in sequential data by using end-to-end-trained word embeddings. The LSTM architecture consists of an embedding layer, an LSTM layer, a dense layer, and a dropout layer to prevent overfitting [22]. Meanwhile, IndoBert (indobert-base-p1) uses subword tokenization with a maximum length of 128 tokens. The classification process was performed via fine-tuning using the AdamW optimizer and a learning rate of 2×10^{-5}. All models were tested on test data that was not used during training. The results showed an 80:20 data split (stratified split) for NB, SVM, and LSTM, and a 70:15:15 split for IndoBert.

2.4 Evaluasi Model

The performance of the four models was comprehensively evaluated using the metrics Accuracy, Precision, Recall, and F1-Score based on the confusion matrix [23]. Given the class imbalance, the F1-Score Macro was used as the primary metric for determining the best model because it is more sensitive to performance on the minority class. The evaluation metric is calculated as follows:

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

$$\text{precission} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1} - \text{Score} = 2 \text{ x } \frac{\text{Precission x Recall}}{\text{Precission+Recall}} \quad (4)$$

where TP (True Positive) denotes the number of instances correctly predicted as belonging to a given sentiment class; TN (True Negative) denotes instances correctly predicted as not belonging to that class; FP (False Positive) denotes instances incorrectly predicted as belonging to the class when they do not; and FN (False Negative) denotes instances that actually belong to the class but were incorrectly predicted as not belonging to it. Since this study involves multi-class classification (Positive, Negative, and Neutral), each metric was calculated using a one-vs-rest approach for every class and subsequently averaged using the macro-averaging method, which treats all classes equally regardless of their sample size — an approach particularly suitable given the class imbalance observed in this dataset. In addition to basic metrics, the evaluation also includes a confusion matrix analysis to map the distribution of classification errors, as well as a Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) curve to measure the model's discriminatory power at various thresholds.

3. RESULT AND DISCUSSION

This section presents the results of a comparative implementation of the four sentiment classification algorithms, along with an in-depth discussion of the performance, limitations, and methodological implications of the research findings. The analysis focuses on the models' ability to handle the characteristics of informal YouTube text and class imbalance.

3.1 Model Implementation and Test Results

3.1.1 Distribution of Public Sentiment Toward the Indonesia-U.S. Trade Agreement

This study analyzes differences in public opinion based on the results of automatic tagging using the InSet lexicon on the initial dataset (before the augmentation stage), before discussing the algorithm's performance. The distribution of public sentiment toward the Indonesia–United States Trade Agreement reveals a clear hierarchy across the three sentiment categories, each reflecting a distinct facet of public engagement with the issue.

Neutral sentiment dominated the dataset, accounting for 52.0% (1,188 data points); this category is composed mainly of descriptive comments, factual questions, or direct quotations of news headlines that do not carry strong emotional undertones. This dominance is unsurprising given the nature of YouTube comment sections under policy-related videos, where a substantial share of users tend to restate or discuss factual developments rather than express explicit approval or disapproval. The negative class ranks second at 27.8% (635 data points), reflecting netizens' concerns over the erosion of economic sovereignty, the exemption of U.S. companies from Local Content Requirements (TKDN), and apprehension over the growing dominance of foreign corporations in strategic sectors; this pattern suggests that public anxiety is concentrated less on the trade agreement itself and more on its perceived structural asymmetry, namely provisions seen as disproportionately favoring U.S. economic interests. By contrast, the positive class stands at only 20.2% (461 data points), representing a smaller but notable segment of the public that expresses optimism regarding the potential for expanded exports and deeper bilateral economic cooperation; the comparatively low share of positive sentiment relative to the negative class indicates that public discourse around the ART is currently shaped more by skepticism than by enthusiasm, consistent with the controversial provisions highlighted in the policy background. The dominance of the Neutral class and the high proportion of the Negative class initially created a significant class imbalance in the training data. This imbalance was addressed at two stages of the pipeline: first, through minority-class oversampling prior to model training, which increased the representation of the Positive and Negative classes to match the size of the majority class; and second, through the use of a Focal Loss objective during the modeling phase, which further reduced the models' tendency to default to the statistically dominant Neutral class.

3.1.2 Comparative Performance Evaluation of Four Algorithms

The four models were evaluated using the metrics Accuracy, Precision, Recall, Macro F1-Score, and AUC-ROC. Data splitting was performed consistently using class imbalance handling tailored to the characteristics of each algorithm. The comparative results are presented in Table 2.

Table 2. Comparison of the Performance of Four Sentiment Classification Algorithms

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	AUC-ROC (Macro)
Naïve Bayes TF-IDF	61,71%	0,61	0,64	0,62	0,80

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	AUC-ROC (Macro)
SVM TF-IDF	79,00%	0,78	0,78	0,78	0,91
LSTM (Random Embedding)	52,00%	0,51	0,34	0,23	0,50
IndoBert (Fine-Tuning)	87,71%	0,87	0,87	0,87	0,96

As shown in Table 2, with an accuracy of 87.71% and a Macro F1-Score of 0.87, IndoBERT clearly outperformed the other three models. With an AUC-ROC of 0.96, this model demonstrated excellent discriminative ability. In second place, SVM performed well, with an accuracy of 79% and an AUC of 0.91. SVM demonstrates that classical machine learning methods remain relevant when combined with bigram TF-IDF. While LSTM suffered a total failure with an accuracy of 52% and an AUC of 0.50, which is statistically equivalent to a random classifier, Naïve Bayes performed at a moderate level (61.71%).

To examine the distribution of classification errors across classes for each model, an in-depth analysis was conducted using confusion matrix visualizations. Figure 2 shows the confusion matrix visualizations for the four models tested.

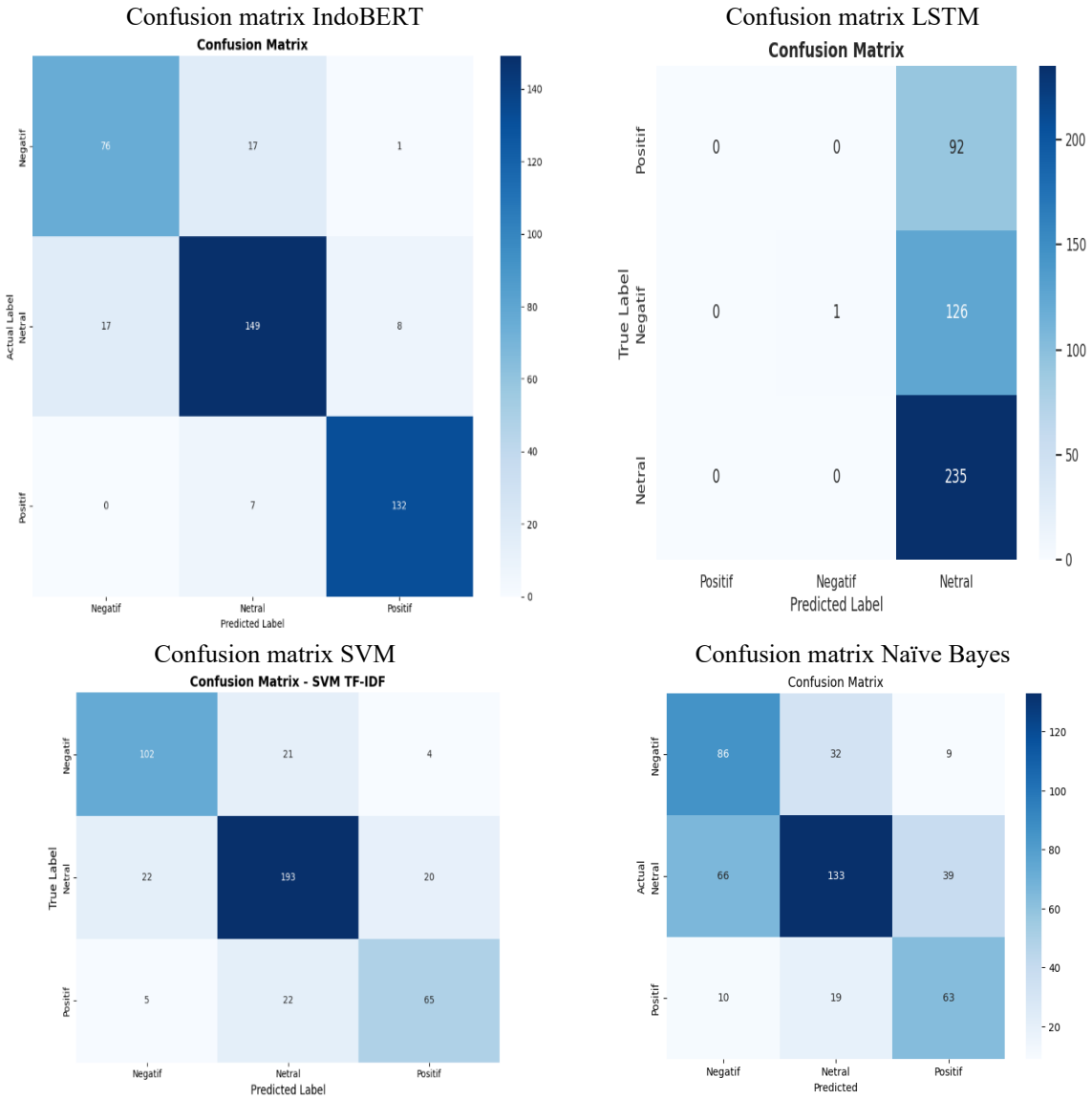

Figure 2. Confusion Matrix for Four Sentiment Classification Models

Based on Figure 2, the error patterns produced by each algorithm are clearly visible. IndoBERT exhibits a very strong diagonal matrix with minimal and evenly distributed classification errors, indicating that the model is capable of distinguishing between positive, negative, and neutral nuances. On the other hand, the LSTM Confusion Matrix shows a complete failure, with nearly all matrix cells filled only with the neutral class. The LSTM model fails to distinguish between the positive and negative classes, indicating that random embeddings on a small dataset are insufficient to capture the discriminative features of informal YouTube text. Meanwhile, Naive Bayes and

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

SVM exhibit significant classification errors between the Neutral and Negative classes due to semantic ambiguity in text containing words with negative connotations, but in a neutral context.

In addition to the confusion matrix, the Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) curve is also used to evaluate a model's ability to distinguish between classes at various thresholds. The comparative ROC-AUC curves for the four models are shown in Figure 3.

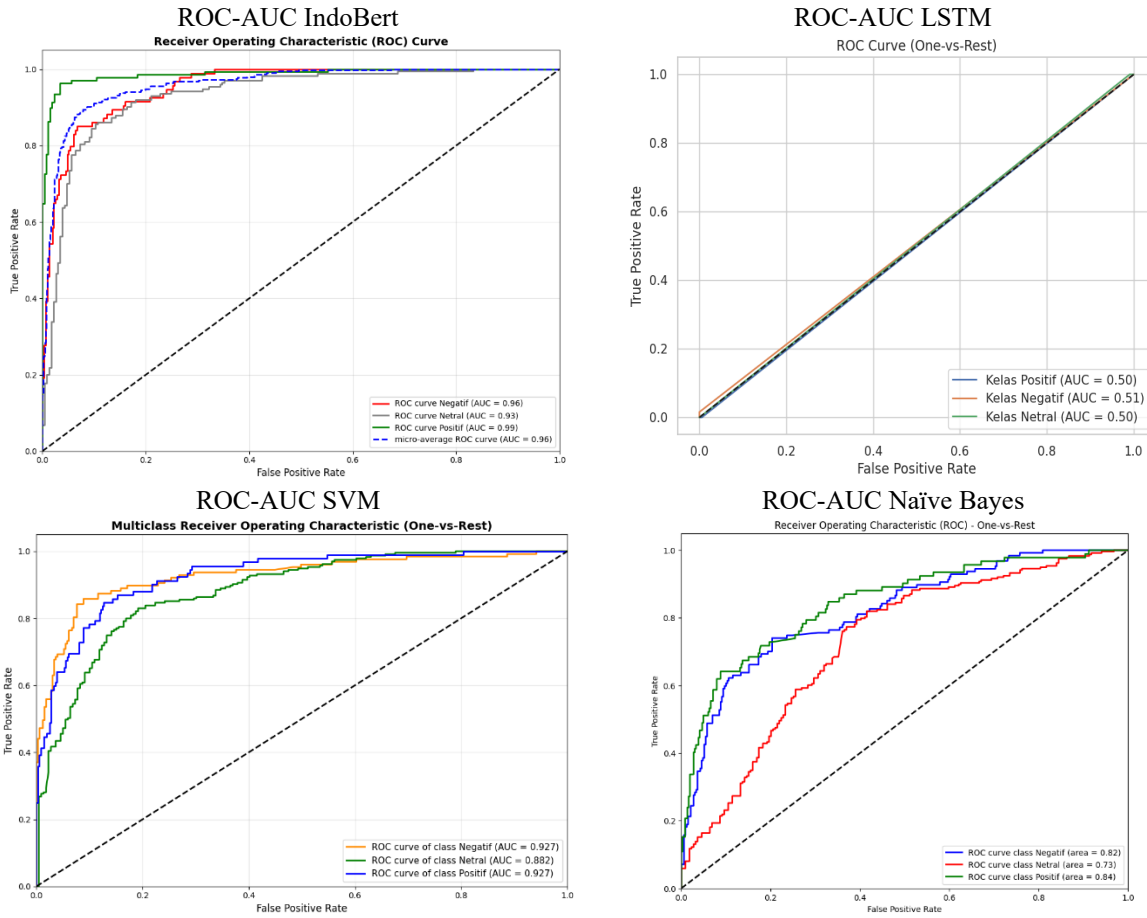

Figure 3. Comparative ROC-AUC Curves for Four Models

The graph in Figure 3 confirms that the IndoBert curve lies far above the diagonal line (random classifier) with an AUC of 0.96, meaning the model has a nearly perfect ability to distinguish sentiment classes. The SVM curve also shows excellent performance (AUC 0.91), placing it well above the baseline. However, the ROC curve for LSTM overlaps with the diagonal line (AUC 0.50), which mathematically proves that the LSTM model in this study does not have better predictive capabilities than random guessing. These findings empirically validate that, for informal Indonesian text data containing specific terms, the Transformer architecture (IndoBert) is far more robust than the sequential architecture (LSTM) that is not supported by pre-trained embeddings.

3.2 Discussion

3.2.1 Analysis of the Strengths and Weaknesses of the Model

The test results revealed an interesting phenomenon closely related to the architectural characteristics of the model and the nature of YouTube data.

Empirically, IndoBERT's superior performance (87.71%) demonstrates that Transformer-based models are the most suitable choice for informal Indonesian text data. This success can be attributed to two key factors. First, IndoBERT is able to capture long-range context and semantic ambiguities such as sarcasm or irony that sequential models struggle to represent, owing to its bidirectional self-attention mechanism, which allows every token in a sentence to be weighted against every other token simultaneously rather than being processed strictly in sequence. Second, the use of Focal Loss ($\gamma = 2.0$) during training played a meaningful role in mitigating the class imbalance problem. Focal Loss modifies the standard cross-entropy loss by introducing a modulating factor,$(1 - P_t)^{\gamma}$, which automatically down-weights the loss contribution of samples the model already classifies with high confidence, typically majority-class (Neutral) examples, while proportionally amplifying the loss for harder, misclassified minority-class (Positive and Negative) examples. This down-weighting effect is achieved solely through the model's own prediction confidence at each training step. This mechanism encouraged the model to allocate more

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

learning capacity toward distinguishing Positive and Negative sentiment, rather than defaulting to the statistically dominant Neutral class, and is reflected in the balanced diagonal pattern observed in IndoBERT's confusion matrix (Figure 2).

A key finding of this study is the extremely poor performance of LSTM, which achieved a Macro F1-Score of only 0.23. Analysis of the confusion matrix shows that the model predicted the neutral class for nearly 98% of the test data, a phenomenon referred to in the machine learning literature as the "collapse of the majority class" or the "trivial classifier problem." Three interrelated factors can explain this failure. With only 2,284 data points, an LSTM relying on randomly initialized embeddings lacks sufficient data capacity to learn meaningful word-vector representations, resulting in severe underfitting. Unlike IndoBERT, which benefits from embeddings pre-trained on billions of tokens of Indonesian-language corpora, the embedding layer in the LSTM configuration used in this study was initialized randomly and therefore had to learn semantic relationships between words from scratch — a task that a dataset of this size cannot adequately support. Compounding this limitation, class weighting proved far less effective in the LSTM architecture than in SVM: while SVM's class_weight parameter directly rescales the margin-based optimization objective and reliably shifts the decision boundary, the equivalent weighting applied to LSTM's loss function was insufficient to counteract the 52% dominance of the Neutral class once the model began converging toward a degenerate, single-class prediction pattern from the early training epochs onward.

SVM (79%) and Naïve Bayes (61.71%) demonstrate that TF-IDF representation remains a reasonably effective baseline, though it carries inherent limitations. Compared to Naïve Bayes, SVM using bigram n-grams is more accurate at capturing contextual phrases such as "foreign stooge" or "huge loss," which likely explains its considerably higher performance margin. However, both models struggle to distinguish neutral comments containing negatively toned vocabulary from genuinely negative sentiment — for instance, "waiting for the next policy" versus "this policy is harmful" — because TF-IDF represents text purely through word-frequency counts without capturing sentence order, syntactic structure, or contextual dependency, making it inherently prone to misclassification in semantically ambiguous text.

3.2.2 Methodological implications and comparison with previous research

The findings of this study make a significant methodological contribution and, at the same time, correct several assumptions in the existing literature on sentiment analysis.

These results also help resolve an apparent contradiction in prior literature regarding the relative strength of LSTM- and Transformer-based models on YouTube data. Previous research by Fahrezi et al. [17] on YouTube comments regarding the presidential debate found that LSTM (97.77%) significantly outperformed IndoBERT (46.6%) because the Transformer model struggled with slang that was too noisy. However, this study found the opposite result: IndoBERT (87.71%) outperformed LSTM (52%) by a wide margin. This fundamental difference stems from the linguistic characteristics of the topic. Comments made during the presidential election debate were dominated by political slang and extreme sarcasm, which may have reduced the value of WordPiece. In contrast, comments on the Indonesia-U.S. Trade Agreement, although informal, contained many technical economic terms such as "TKDN," "reciprocal tariffs," "BUMN," and "exports," which IndoBERT had extensively learned during the pre-training phase. This demonstrates the advantage of Transformers over LSTMs on the platform.

In terms of fine-tuning consistency, the accuracy of 87.71% and F1-score of 0.87 achieved in this study are highly consistent and competitive when compared to other recent studies. Agetia et al. [15] reported an IndoBERT accuracy of 92.46% for Gojek reviews, while Daniati et al. [16] achieved 91% for electric vehicle opinions. Although the results of this study are slightly lower, this is entirely understandable because this study uses multi-class classification (3 classes: Positive, Negative, and Neutral), which is mathematically more complex and more susceptible to noise than the binary classification (2 classes) used in previous studies.

4. CONCLUSION

This study successfully mapped public perceptions of the Indonesia–United States Trade Agreement (ART) through sentiment analysis of YouTube comments, addressing the research problem of identifying opinion polarity and determining the most effective computational model. The sentiment distribution was dominated by the Neutral class (52.0%), followed by Negative (27.8%) and Positive (20.2%) sentiment, indicating that public discourse tends to be descriptive and informative, though marked by notable skepticism toward the agreement's impact on domestic economic sovereignty. Among the four models compared, IndoBERT optimized with Focal Loss achieved the best performance, with an accuracy of 87.71% and a Macro F1-Score of 0.87, substantially outperforming SVM (79.00%) and Naïve Bayes (61.71%). In contrast, LSTM failed completely, with an accuracy of 52.00% and an AUC-ROC of 0.50—statistically equivalent to random prediction—a failure attributed to randomly initialized embeddings on a small-scale dataset, which triggered a majority-class collapse. These findings confirm that for informal social media text containing domain-specific terminology, Transformer architectures built on pre-trained language models substantially outperform sequential architectures trained from scratch. Despite these methodological contributions, several limitations remain. Manual validation by linguistic experts is recommended to improve ground-truth quality, given that lexicon-based labeling may introduce noise

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

in ironic or sarcastic comments. Reliance on YouTube as the sole data source may also introduce demographic selection bias, suggesting the value of future cross-platform comparisons with X or TikTok. Finally, the simple augmentation techniques applied here could be enhanced through LLM-based paraphrasing to generate richer syntactic variation and improve generalization to minority classes.

REFERENCES

[1] R. Qoni'ah, "Kemitraan Strategis Perdagangan Indonesia-Amerika Serikat : Menavigasi Tantangan Global dan Potensi Keunggulan Kompetitif," J. Transform. Glob., vol. 11, no. 2, 2024, doi: https://doi.org/10.21776/ub.jtg.011.02.5
[2] M. S. Mandalika and V. D. Muaja, "Analisis Hukum terhadap Dampak Pengenaan Tarif 32% oleh Amerika Serikat terhadap Perdagangan Indonesia: Tinjauan Perjanjian Perdagangan Internasional dan Kebijakan Ekonomi," J. Polit. Sos. Huk. dan Hum., vol. 3, no. 2, 2025, doi: https://doi.org/10.59246/aladalah.v3i2.1285
[3] S. D. Ariapramuda, "Transformasi Pembatasan Kedaulatan Regulasi dalam Hukum Investasi Indonesia Pasca Agreement on Reciprocal Trade (ART) Indonesia – Amerika Serikat 2026," J. Glob. Ilm., vol. 3, no. 8, pp. 1942–1953, 2026, doi: https://doi.org/10.55324/jgi.v3i8.370
[4] I. Sartika and I. Saputra, "Analisis Sentimen Masyarakat Terhadap Pembatasan Akses Media Sosial Bagi Anak Pada Platform YouTube Menggunakan Metode Support Vector Machine (SVM)," JIBEMA J. Ilmu Bisnis, Ekon. Manajemen, dan Akunt., vol. 3, no. 4, pp. 567–578, 2026, doi: https://doi.org/10.62421/jibema.v3i4.251
[5] A. Pangestu and F. Maulana, "Implementasi Analisis Sentimen Komentar YouTube Berbasis IndoBERT dengan Evaluasi Confidence Score," Semin. Nas. Teknol. Sains, vol. 5, no. 1, pp. 664–672, 2026, [Online]. Available: https://proceeding.unpkediri.ac.id/index.php/stains/article/view/9348
[6] A. A. Qolbu, N. Fitriyati, and N. Inayah, "Performa Naïve Bayes, SVM, dan IndoBERT pada Analisis Sentimen Twitter IndiHome dengan Strategi Penanganan Data Tidak Seimbang," J. Fourier, vol. 814, no. 1, pp. 29–44, 2025, [Online]. Available: https://fourier.or.id/index.php/FOURIER/article/view/252
[7] D. D. Purwanto, "Empirical Evaluation of IndoBERT and LSTM for Sentiment Analysis of Tourism Reviews : A Data-Driven Study on Kenjeran Park," J. Tek. Inform., vol. 7, no. 1, pp. 463–474, 2026, doi: https://doi.org/10.52436/1.jutif.2026.7.1.4901
[8] S. A. Rahmadhani, L. D. Rusanti, and H. Al Rosyid, "Klasifikasi Sentimen Komentar Youtube Demo DPR RI Menggunakan Support Vector Machine," Arcitech J. Comput. Sci. Artif. Intell., vol. 5, no. 2, pp. 356–375, 2025, doi: https://doi.org/10.29240/arcitech.v5i2.15316
[9] S. F. Huwaida, R. Kusumawati, and B. Isnaini, "Analisis sentimen komentar youtube terhadap pemindahan ibu kota negara menggunakan metode Naïve Bayes," Jambura J. Informatics, vol. 6, no. 1, pp. 26–39, 2024, doi: https://doi.org/10.37905/jji.v6i1.24718
[10] E. Darmawan, M. A. Hasan, N. Rahmawati, and V. Kurniawan, "PEMANFAATAN LSTM UNTUK MENGANALISIS SENTIMEN PENGGUNA TWITTER : STUDI KASUS PADA TWEET BERITA TERKINI," JATI (Jurnal Mhs. Tek. Inform., vol. 9, no. 2, pp. 2954–2963, 2025, doi: https://doi.org/10.36040/jati.v9i2.13194
[11] N. R. Ramadhan and N. Hendrastuty, "Perbandingan Algoritma Naïve Bayes dan LSTM untuk Analisis Sentimen Terhadap Opini Masyarakat Tentang Sandwich Generation," Build. Informatics, Technol. Sci., vol. 6, no. 3, pp. 1677–1687, 2024, doi: https://doi.org/10.47065/bits.v6i3.6385
[12] H. Abriananta et al., "Comparison of Naive Bayes, Support Vector Machine, and Indobert Methods for Classifying Public Sentiment towards the MBG Program on Platform X," J. Appl. Informatics Comput., vol. 10, no. 3, pp. 2806–2815, 2026, [Online]. Available: https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/12721
[13] M. R. Manoppo, I. C. Kolang, D. N. Fiat, R. Michelly, and C. Mawara, "ANALISIS SENTIMEN PUBLIK DI MEDIA SOSIAL TERHADAP KENAIKAN PPN 12% DI INDONESIA MENGGUNAKAN INDOBERT," J. Kecerdasan Buatan dan Teknol. Inf., vol. 4, no. 2, pp. 152–163, 2025, doi: https://doi.org/10.69916/jkbti.v4i2.322
[14] D. Nuryadi et al., "FINE TUNING INDOBERT UNTUK ANALISIS SENTIMEN PADA ULASAN PENGGUNA APLIKASI TIKET . COM DI GOOGLE PLAY STORE," JATI (Jurnal Mhs. Tek. Inform., vol. 9, no. 2, pp. 3577–3583, 2025, doi: https://doi.org/10.36040/jati.v9i2.13204
[15] I. W. A. Agetia, N. Luh, E. Armoni, I. P. Ari, and U. Irawan, "Evaluasi Kinerja Algoritma Naïve Bayes , SVM , dan IndoBERT pada Analisis Sentimen Ulasan Pengguna Gojek Berbasis Text Mining," TIN Terap. Inform. Nusant., vol. 7, no. 1, pp. 89–98, 2026, [Online]. Available: https://ejurnal.seminar-id.com/index.php/tin/article/view/9617
[16] E. Daniati et al., "Perbandingan Kinerja Algoritma SVM , LSTM , dan Fine-tuned IndoBERT dalam Analisis Sentimen Opini Masyarakat Indonesia terhadap Mobil Listrik," IJCSR, vol. 5, no. 1, pp. 1–13, Jan. 2026, doi: 10.59095/ijcsr.v5i1.245.
[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 J. Penelit. Ilm. Multidsipliner, vol. 02, no. 03, pp. 1936–1961, 2025, [Online]. Available: https://ojs.ruangpublikasi.com/index.php/jpim/article/view/1110
[18] D. Pratiwi, N. Khoerani, S. Sari, U. Trisakti, J. Barat, and P. Korespondensi, "ANALISIS SENTIMEN TERHADAP KEBIJAKAN SUBSIDIPEMBELIAN KENDARAAN BERTENAGA LISTRIK DI INDONESIA MENGGUNAKAN PENDEKATAN INSET LEXICON DAN METODE SUPPORT VECTOR MACHINE," J. Teknol. Inf. dan Ilmu Komput., vol. 12, no. 6, pp. 1303–1314, 2025, [Online]. Available: https://jtiik.ub.ac.id/index.php/jtiik/article/view/9548
[19] J. Siringoringo, S. P. Tanjung, H. Budi, S. Lukito, and J. Prancis, "Klasifikasi Sentimen Opini Publik pada Isu Anggaran DPR Menggunakan Support Vector Machine Berbasis Pembobotan Kelas," J. Algoritm., pp. 1780–1790, 2026, doi: https://doi.org/10.33364/algoritma/v.23-1.3539
[20] D. Septiani and I. Isabela, "Analisis Term Frequency Inverse Document Frequency (TF-IDF) Dalam Temu Kembali Informasi pada Dokumen Teks," SINTESIA J. Sist. dan Teknol. Inf. Indones., vol. 25, no. 2, pp. 81–88, 2022, [Online]. Available: https://journal.unj.ac.id/unj/index.php/SINTESIA/article/view/39364
[21] A. M. Yolanda and R. T. Mulya, "Implementasi Metode Support Vector Machine untuk Analisis Sentimen pada Ulasan

Journal of Information System Research (JOSH)
Volume 7, No. 4, Juli 2026, pp 1067−1075
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v7i4.10515

Aplikasi Sayurbox di Google Play Store," VARIANSI J. Stat. Its Appl. Teach. Res. Vol., vol. 6, no. 2, pp. 76–83, 2024, [Online]. Available: https://jurnalvariansi.unm.ac.id/index.php/variansi/article/view/258

[22] D. S. Hani and C. I. Ratnasari, "Klasifikasi Masalah Pada Komunitas Marah-Marah di Twitter Menggunakan Long Short-Term Memory," J. MEDIA Inform. BUDIDARMA, vol. 7, no. 4, pp. 1829–1837, 2023, [Online]. Available: https://paperity.org/p/341530804/klasifikasi-masalah-pada-komunitas-marah-marah-di-twitter-menggunakan-long-short-term

[23] T. Gori et al., "PREPROCESSING DATA DAN KLASIFIKASI UNTUK PREDIKSI KINERJA AKADEMIK SISWA," JTIIK, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: 10.25126/jtiik.20241118074.