Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 4 (2026) 918 - 928 e-ISSN: 2580-0760

A Hybrid VADER–IndoBERT Framework for Robust Sentiment Analysis of Long and Ambiguous Indonesian Texts

Margareta Valencia Suci Handayani[1], Ruri Suko Basuki[2], Muljono[3*], Raden Arief Nugroho[4], Dhendra Marutho[5], Yo Ceng Giap[6], Deshinta Arrova Dewi[7]
[1,2,3,6]Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
[4]Faculty of Humanities, Universitas Dian Nuswantoro, Semarang, Indonesia
[5]Department of Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia
[7]Faculty of Data Science and Information Technology, INTI International University, Putra Nilai, Malaysia

[*]*muljono@dsn.dinus.ac.id*

Abstract

The rapid expansion of digital learning platforms has increased the reliance on user-generated reviews for service evaluation and quality monitoring. However, sentiment analysis of Indonesian reviews remains challenging due to the prevalence of long sentences, mixed sentiments, and ambiguous linguistic expressions. This study introduces a Hybrid VADER–IndoBERT framework designed to improve sentiment classification robustness on complex Indonesian texts. A dataset of 4,904 Ruangguru application reviews was collected through web scraping and processed using a hybrid pipeline consisting of preprocessing, translation-based silver-standard sentiment labeling with VADER, and class balancing via Random Oversampling (ROS). The IndoBERT classifier was evaluated against a Bidirectional Long Short-Term Memory (BiLSTM) baseline. Experimental results show that IndoBERT achieved 90.9% accuracy, outperforming BiLSTM at 86.4%, demonstrating the superiority of Transformer-based architectures in capturing long-range dependencies and handling ambiguous sentiment cues. These findings highlight the effectiveness of integrating lexicon-based and Transformer-based approaches to achieve more robust sentiment analysis on linguistically complex Indonesian texts.

Keywords: BiLSTM; indobert; natural language processing; process innovation; ruangguru; sentiment analysis

How to Cite: M. V. S. Handayani, "A Hybrid VADER–IndoBERT Framework for Robust Sentiment Analysis of Long and Ambiguous Indonesian Texts", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 4, pp. 918 - 928, Jul. 2026.
*Permalink/*DOI: https://doi.org/10.29207/resti.v10i4.7377

Received: December 10, 2025
Accepted: June 21, 2026
Available Online: July 29, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The rapid expansion of digital learning technologies has accelerated the adoption of e-learning systems, enabling flexible online interaction between teachers and students [1], [2]. In Indonesia alone, more than 185 million citizens representing 66.5% of the population are active internet users, while over 139 million regularly engage with social media platforms [3]. This large digital population has increased the demand for technology-enabled education services. Among the major platforms supporting this transformation is Ruangguru, which serves over 38 million active learners and more than 300,000 private tutors.

User-submitted feedback on platforms like the Google Play Store play a crucial role in evaluating service quality and informing continuous improvement efforts [4], [5]. Recent studies highlight that AI-driven sentiment analysis is an instrument for deriving valuable insights from extensive textual datasets. to enhance service quality and understand consumer preferences [6], [7]. However, these reviews present several natural language challenges. Many texts contain long sentences, mixed sentiments, and ambiguous expressions including sarcasm, figurative language, and colloquial Indonesian which significantly complicate automated sentiment interpretation [8], [9]. Such linguistic complexity makes manual inspection impractical and often exceeds the capability of traditional computational models.

Sentiment analysis provides a systematic approach for extracting subjective opinions from textual data and grouping it into positive, negative, or neutral polarity classes [10], [11]. Nevertheless, classic machine learning algorithms like Naïve Bayes and Support

Vector Machines frequently struggle to capture contextual meaning, resolve ambiguity, and detect sentiment polarity shifts within long and compound sentences [4], [12-14]. Similarly, deep neural architectures based on Recurrent Neural Network (RNN) architectures, including Bidirectional Long Short-Term Memory (BiLSTM) networks, have been widely used for sequential data [15], they often fail to capture long-range dependencies and struggle to generalize when processing long, ambiguous reviews. These limitations highlight the need for more advanced language representation techniques.

Recent studies have explored various approaches to address these linguistic challenges. For instance, Anam and Kusnawi [16] implemented a VADER-IndoBERT hybrid approach for analyzing public opinion on political discourse, while Christian et al. [17] utilized IndoBERT with data augmentation for emotion classification in e-commerce reviews. Furthermore, Setiawan [18] noted that advanced Deep Learning architectures like IndoBERT are increasingly becoming the standard for handling complex Indonesian text. However, despite these advancements, there remains a limitation in effectively integrating lexicon-based and contextual approaches to handle highly ambiguous, informal, and long user-generated texts, particularly in domains where labeled data are scarce. While prior studies have explored hybrid VADER–IndoBERT approaches, most focus on improving classification performance. In contrast, this study emphasizes the integration of lexicon-based labeling and contextual learning as a unified weakly supervised framework that is more suitable for real-world scenarios with limited annotated data.

To solve these limitations, this paper proposes a hybrid VADER–IndoBERT model combining lexicon-based labeling and Transformer contextual learning. The motivation for combining these two approaches lies in their complementary strengths. VADER, as a rule-based lexicon model, provides fast, consistent, and interpretable sentiment scoring, making it suitable for generating silver-standard labels in the absence of manually annotated data [19], [20]. Such weakly supervised labeling strategies have been widely adopted to reduce the cost of manual annotation while maintaining reasonable labeling consistency [21]. However, lexicon-based approaches are inherently limited in capturing contextual nuances, such as sarcasm, implicit sentiment shifts, and long-range dependencies in complex sentences [10], [11].

In contrast, Transformer-based models such as IndoBERT excel at learning deep contextual representations and modeling complex linguistic structures through self-attention mechanisms [8], [22]. These models consistently outperform traditional baselines in diverse sentiment classification tasks, especially in low-resource language scenarios with rich morphological and informal variations like Indonesian [17], [18]. Nevertheless, their performance relies heavily on the availability of well-labeled datasets, which poses a significant bottleneck in real-world deployments.

By using labels generated from VADER as a weak supervision signal and utilizing contextual learning in IndoBERT, this framework can provide more stable sentiment classifications when data are sparse, noisy, or ambiguous linguistically. In particular, VADER generates an initial polarity signal based on lexical features, whereas IndoBERT learns how to improve this signal through the pattern in context, thus decreasing noise from ambiguous expressions and improving generalization on long and mixed-sentiment reviews [22]. Although this study does not perform ablation analysis of each component directly, the experimental results show that IndoBERT can learn well from silver labels generated by VADER as this signal, confirming the hybrid framework's practical quality and robustness in sentiment analysis within real-world datasets.

Therefore, the contributions of this study are threefold. First, it proposes a hybrid lexicon–Transformer framework that integrates sentiment labeling and contextual learning for linguistically complex Indonesian texts. Second, it constructs a sentiment-labeled dataset of Ruangguru reviews using VADER-based silver labeling and class balancing techniques. Third, it demonstrates the effectiveness of IndoBERT in learning from lexicon-based silver labels compared to a BiLSTM baseline. Overall, this study aims to advance sentiment analysis methods for Indonesian digital learning platforms by providing a scalable and robust approach for handling long, ambiguous, and user-generated texts.

2. Methods

In this study, a sentiment classification methodology is employed to examine user opinions toward the Ruangguru application using reviews from the Google Play Store.

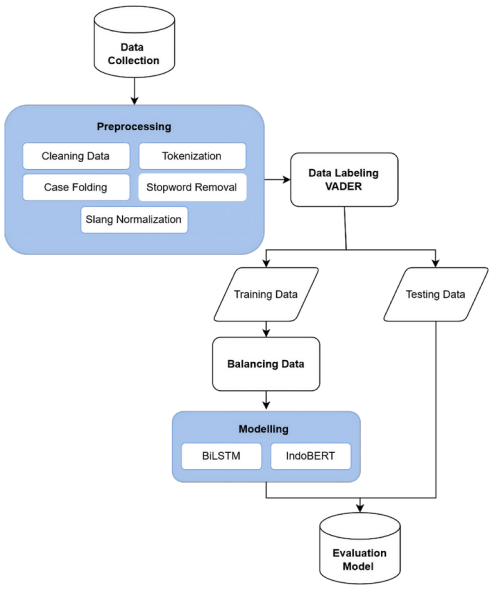


Figure 1. Research Workflow

As demonstrated in Figure 1, the methodology follows a structured workflow consisting of data collection, data labeling, preprocessing, oversampling, modeling, and evaluation stages. The processed data are used to train sentiment classification models, with class imbalance handled through a data balancing strategy and performance assessed using standard evaluation metrics.

2.1 Data Collection

The dataset used in this study comprises user reviews of the Ruangguru application collected 5,000 comments from the Google Play Store via web scraping. This platform was selected because it hosts a vast amount of user-generated feedback containing diverse linguistic patterns, ranging from short direct comments to long, complex narratives with mixed sentiments. The Python Google Play Scraper library was utilized to extract the review text along with its associated metadata. All data were saved in CSV format to facilitate the subsequent hybrid processing pipeline.

Table 1. Dataset Description

Attribute	Description
Rating	Numeric rating (1–5) reflecting the user's satisfaction level
Review Text	Comment text posted by the Google Play Store user

Table 1 provides a brief description of the attributes included in the dataset. The Rating attribute contains numeric values ranging from one to five, representing the user's level of satisfaction with the application. The Review Text attribute consists of written comments submitted by Google Play Store users. However, the initial data inspection revealed a non-linearity between the numerical ratings and sentiment expressed within the review text, where some users gave high ratings when voicing complaints, and others gave low ratings even when voicing positive or neutral feedback, indicating that the ratings may not be a reliable measure of sentiment in textual reviews [7]. Hence, it has been necessary to remove the Rating attribute in the later analysis and concentrate only on the Review Text to capture user sentiment in this study.

2.2 Preprocessing

The preprocessing stage improves the quality and readability of the data before it is used for model training. This procedure aligns with findings from international studies which emphasize that text cleaning and normalization are essential for enhancing sentiment-analysis performance, particularly for languages characterized by high linguistic variation such as Indonesian [23], [24]. Ensuring cleaner textual input helps the model learn more stable patterns and reduces potential disturbances caused by noisy user-generated content.

The process begins with the removal of URLs, excessive punctuation, emojis, numbers, non-alphabetic characters, and other elements that do not contribute to the semantic meaning of the text. This step helps reduce linguistic noise and strengthens the quality of the extracted features prior to modeling [13]. After cleaning, all text was converted to lowercase to ensure uniform token representation regardless of letter casing, thereby reducing sparsity and enhancing overall data consistency [24]. Slang normalization is then conducted to convert non-standard words, abbreviations, and informal slang into their standardized forms, reducing ambiguity in informal review texts. In this study, slang terms and their formal equivalents were sourced from a GitHub reference list and transformed into a dictionary using the pandas library, enabling accurate token-level normalization. This normalization step is particularly crucial in this study to maximize the accuracy of the subsequent machine translation process required for VADER labelling.

Once the text is normalized, tokenization is applied by splitting sentences into individual word units, allowing the model to analyze semantic patterns and word sequences more effectively. Consistent tokenization is essential for capturing the meaningful structure of user feedback and ensuring that sequential patterns are preserved [23]. Finally, stopwords that do not contribute to sentiment determination are removed to help the model focus on informative terms, improving modeling efficiency while reducing interference from irrelevant lexical elements [24].

2.3. Data Labelling

To overcome the scarcity of pre-labeled data and the unreliability of user-provided ratings, this study adopts a weakly supervised approach using VADER (Valence Aware Dictionary and Sentiment Reasoner) was employed to generate silver-standard sentiment labels. Since VADER is designed for English text, Google Translate API was used to initially translate the reviews into English [21]. Based on the computed compound polarity score, VADER classifies sentiments into positive (≥ 0.05), negative (≤ -0.05), while values in between as neutral [20], [25].

To guarantee the reliability of the generated labels, a manual validation procedure was performed on a sample of roughly 10% 10% of the dataset using stratified sampling to preserve class distribution. Two independent annotators with a background in Indonesian linguistics independently labeled the original (non-translated) review texts based on contextual meaning. The labeling criteria are summarized in Table 2.

Table 2. Manual Labeling Criteria

Sentiment Category	Criteria
Positive	Contains expresses satisfaction, praise, or positive experience
Neutral	Contains factual statements or mixed sentiment without a clear dominant polarity
Negative	Contains expresses complaints, dissatisfaction, or criticism toward application

The manually assigned labels were compared with VADER-generated labels using accuracy and Cohen's Kappa coefficient, where accuracy reflects overall agreement and Cohen's Kappa accounts for agreement by chance. Based on the criteria established by Jacob Cohen and adapted by Landis and Koch [26], Cohen's Kappa values are evaluated across five agreement levels ranging from slight (< 0.20) to almost perfect (0.81–1.00). This evaluation framework enables systematic comparison between human and automated labeling while identifying potential biases introduced by the translation-based annotation process.

2.4 Data Split

The processed review data were partitioned into training (80%) and test (20%) [12]. This approach helps maintain a balanced representation of sentiment labels across both subsets, ensuring that no class becomes disproportionately dominant. By preserving the original distribution, the model is exposed to data that accurately reflect real-world patterns during training.

This proportional splitting also contributes to a more reliable evaluation process. The test set remains unbiased because it mirrors the natural distribution of sentiments in the full dataset. Consequently, the model's performance metrics provide a more realistic indication of how well it generalizes to unseen data.

2.5 Balancing Data

After the data splitting, the class distribution is balanced using Random Oversampling (ROS) to address potential imbalance in the sentiment labels. While the Synthetic Minority Oversampling Technique (SMOTE) is commonly used in classical machine learning to generate synthetic feature vectors [27], it is not directly applicable to the discrete token inputs required by Transformer-based models like IndoBERT. This step ensures that the model learns to recognize ambiguous or critical sentiments that often found in minority classes as effectively as it recognizes positive praise.

This method is particularly effective because it maintains the original characteristics of the minority samples, allowing the model to learn from authentic data points [28]. ROS is also widely adopted in text-classification research due to its ability to stabilize the learning process. As a result, the model becomes less biased toward majority classes and is better equipped to recognize patterns across all sentiment categories.

2.6 Bidirectional Long Short-Term Memory (BiLSTM)

The BiLSTM architecture is a deep learning model tailored for processing sequential data from both forward and backward contexts. Its main strength lies in its ability to learn long-term dependencies in text from both directions [15]. This capability makes BiLSTM effective for text-classification tasks, especially when analyzing informal and unstructured text found in social media reviews.

The BiLSTM modeling workflow used in this study is illustrated in Figure 2. The labeled text data are first converted into indexed tokens and then mapped into an embedding layer to produce low-dimensional word vectors that capture semantic information and the relational meaning between words. These representations serve as the input for the BiLSTM layer. At this stage, two LSTM units operate in parallel, the forward LSTM reads from beginning to end, while the backward LSTM reads in reverse. The representations generated from both directions are subsequently concatenated to produce a comprehensive contextual vector at each sequence step [29].

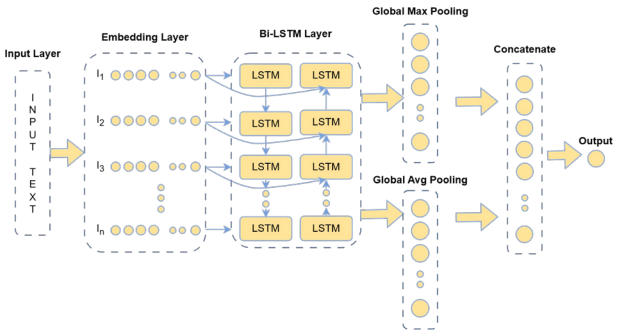


Figure 2. BiLSTM Modeling Workflow

The BiLSTM output is then summarized using two global pooling mechanisms Global Max Pooling, which captures the most dominant features in each dimension, and Global Average Pooling, which extracts the average distribution of information across the sequence. The results of both pooling operations are combined to form the final feature vector, which is subsequently passed to a softmax-activated output layer to perform sentiment classification.

The BiLSTM model in this study is configured using a set of hyperparameters optimized to achieve the best performance. The embedding layer employs 128-dimensional vectors to represent each token. The BiLSTM layer consists of 64 forward units and 64 backward units, allowing it to capture bidirectional context in a balanced manner [15], [29]. To prevent overfitting, dropout rates of 0.3 and 0.3 are applied to the internal BiLSTM connections. The model was optimized using the Adam algorithm, initialized with a learning rate of 0.001, selected for its stability in updating weights in sequential models. The training process was configured with a batch size of 32 and capped at a maximum threshold of 15 epochs, utilizing early stopping to halt the process when validation performance no longer showed improvement. Additionally, input sequences are truncated or padded to a maximum length of 120 tokens to maintain uniform input dimensions while preserving computational efficiency. This baseline serves specifically to test whether Deep Learning architectures based on RNNs are sufficient to handle the long-range dependencies found in complex Ruangguru reviews.

2.7 IndoBERT

This study employs IndoBERT (Indonesian BERT), a state-of-the-art Indonesian language model pre-trained on over 4 billion words sourced from diverse digital corpora, including news platforms and social media [22]. Inspired by the demonstrated effectiveness of the BERT model, IndoBERT was introduced as a specialized architecture trained on an extensive monolingual Indonesian dataset. This training approach makes IndoBERT more sensitive to local linguistic characteristics, including vocabulary variations and informal sentence structures, allowing it to perform more effectively on Indonesian text-understanding tasks such as sentiment analysis [22].

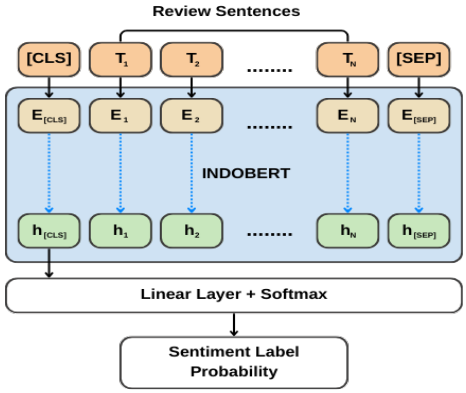


Figure 3. Fine-Tuning Architecture of IndoBERT for Sentiment Classification

This study employs the IndoBERT-Base variant, which has an architectural depth equivalent to BERT-Base, integrates 12 Transformer encoder layers, a hidden size of 768, and 12 parallel attention heads, representing a large-scale architecture with hundreds of millions of parameters [8], [29]. As illustrated in Figure 3, the IndoBERT-Base model is adapted for sentiment classification through a fine-tuning process. In this stage, a simple output layer is added on top of the encoder architecture. Specifically, the vector representation of the classification token ([CLS]), which aggregates the meaning of the entire sequence, is extracted and passed to a fully connected layer with the softmax function to translate the output into a ternary sentiment classification of Positive, Negative, and Neutral [16]. All parameters of the IndoBERT-Base model are updated during fine-tuning using the labeled review dataset. For the fine-tuning phase, the training parameters were set to a learning rate 2e-5 over 5 epochs, with batch sizes specified as 16 for training and 32 for validation. The model is optimized using the AdamW optimizer with a weight decay of 0.01. To stabilize training, warmup steps were allocated to 10% of the overall training steps, while the maximum sequence length was restricted to 128 tokens to achieve an optimal tradeoff between classification accuracy and computational overhead.

2.8 Evaluation

Model performance is evaluated using several metrics commonly applied in text classification research. Model evaluation relied on primary metrics including accuracy, precision, recall, and F1-score to deliver a comprehensive analysis of class prediction performance [10], [11], [30]. These metrics offer a thorough assessment of the model's effectiveness in distinguishing all sentiment categories.

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

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

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

$$F_1 - \text{Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (4)$$

In Equations 1-4, True Positives (TP) and True Negatives (TN) correspond to accurate predictions for positive and negative classes, while False Positives (FP) and False Negatives (FN) designate misclassified samples. As reflected in these equations, accuracy quantifies the ratio of correctly classified samples to the total evaluated instances, while precision evaluates how accurately the model identifies positive predictions, and recall evaluates the proportion of true positive instances successfully identified by the model. While the F1-score represents the harmonic mean of precision and recall offers a balanced assessment especially when the class distribution is uneven. Complementing these quantitative metrics, a Confusion Matrix provides a detailed breakdown of correct and misclassified predictions per class, enabling a deeper examination of the model's error dynamics.

3. Results and Discussions

The results of this study are presented systematically, beginning with the stages of data collection and preprocessing, followed by the evaluation of the classification models used. The analysis evaluates the effectiveness of the proposed Hybrid VADER–IndoBERT approach in handling long and ambiguous user reviews and compares its performance with the BiLSTM baseline.

3.1 Data Collection

Utilizing a web scraping technique on the Google Play Store, the data acquisition phase successfully compiled 5,000 raw user feedback comments.

These reviews exhibit a wide range of user expressions, often involving informal language, emoticons, and various abbreviations. To illustrate the characteristics of the collected data, Table 3 presents several examples of the raw reviews used in this study.

Table 3. Sample of Collected Google Play Store Review

Rating	Review Text
5	saya suka belajar di sini karena bisa memperluas ilmu belajar anak saya terima kasih ruang guru🙏
1	ribet,banyak paketan,dah ada akun malah tdk bisa di gunakan,anehni apk
2	Ga bisa mabar virtual,kecewa berat 😔😔😔😔
4	Aku mau semuanya gratis tapi ga papa yang penting bisa belajar itu aja yang ku sampaikan bintang 🌟🌟🌟🌟
3	aplikasi bisa digunain, tapi kadang-kadang suka ngebug (force closed) pas buka aplikasi loadnya lama padahal internet lancar

3.2 Preprocessing

The preprocessing stage significantly reduced linguistic noise commonly found in long and informal user reviews, enabling both BiLSTM and IndoBERT models to focus on semantically relevant patterns rather than surface-level variations. This preprocessing strategy is crucial for mitigating ambiguity and sentiment dilution in long review texts, which directly supports the study's objective of robust sentiment classification.

During the data-cleaning stage, the number of comments was reduced from 5,000 to 4,904 usable entries. A total of 96 comments were removed because they contained empty text, duplicates, or irrelevant components including web URLs, user mentions, hashtags, numeric digits, and non-alphabetic symbols. An overview of the text quality before and after the cleaning process is presented in Table 4.

Table 4. Comparison Before and After Data Cleaning

Before	After
ALHAMDULILLAHH KEBANTU BANGET BUAT BELAJAR ULANGG KALAU ADA MATERI YANG GA PAHAM DI SEKOLAHH❤️❤️❤️	ALHAMDULILLAHH KEBANTU BANGET BUAT BELAJAR ULANGG KALAU ADA MATERI YANG GA PAHAM DI SEKOLAHH
jelekkkk bgt. ngelag terus, buat belajar susah!!!!!!!	jelekkkk bgt ngelag terus buat belajar susah
Ruangguru ko lebih sering errornya yaa sekarang?🙏	Ruangguru ko lebih sering errornya yaa sekarang

After the data were cleaned, case folding was performed to convert all characters into lowercase, thereby eliminating casing inconsistencies and ensuring standardized word forms. This step reduces unnecessary variation, preventing the model from treating the same word differently due to capitalization. The impact of case folding is illustrated in Table 5 through a comparison of sentence excerpts before and after lowercase normalization.

The tokenization stage breaks each review into individual word units (tokens) that are easier for the model to process systematically. By separating the text into tokens, the pipeline preserves the original word order while representing the sentence as a structured sequence that can be directly consumed by sequence-based models such as BiLSTM and by subsequent feature extraction steps. Table 6 presents the tokenization results and shows how the original sentence structure is transformed into a segmented list of tokens.

Table 5. Comparison Before and After Case Folding

Before	After
ALHAMDULILLAHH KEBANTU BANGET BUAT BELAJAR ULANGG KALAU ADA MATERI YANG GA PAHAM DI SEKOLAHH	alhamdulillahh kebantu banget buat belajar ulangg kalau ada materi yang ga paham di sekolahh
jelekkkk bgt ngelag terus buat belajar susah	jelekkkk bgt ngelag terus buat belajar susah
Ruangguru ko lebih sering errornya yaa sekarang	ruangguru ko lebih sering errornya yaa sekarang

Table 6. Comparison Before and After Tokenization

Before	After
alhamdulillahh kebantu banget buat belajar ulangg kalau ada materi yang ga paham di sekolahh	['alhamdulillahh', 'kebantu', 'banget', 'buat', 'belajar', 'ulangg', 'kalau', 'ada', 'materi', 'yang', 'ga', 'paham', 'di', 'sekolahh']
jelekkkk bgt ngelag terus buat belajar susah	['jelekkkk', 'bgt', 'ngelag', 'terus', 'buat', 'belajar', 'susah', 'sinyal', 'lancar', 'tapi', 'muat', 'ulang', 'terus', 'huhhh']
ruangguru ko lebih sering errornya yaa sekarang	['ruangguru', 'ko', 'lebih', 'sering', 'errornya', 'yaa', 'sekarang']

High-frequency words that lack explicit sentiment orientation, such as "yang," "dan," or "dengan," are removed from the text. Eliminating stopwords produces more concise sentences and highlights the meaningful terms that play an important role in sentiment analysis. A comparison before and after stopword removal is presented in Table 7.

Table 7. Comparison Before and After Stopwords Removal

Before	After
['alhamdulillahh', 'kebantu', 'banget', 'buat', 'belajar', 'ulangg', 'kalau', 'ada', 'materi', 'yang', 'ga', 'paham', 'di', 'sekolahh']	['alhamdulillahh', 'kebantu', 'banget', 'belajar', 'ulangg', 'materi', 'paham', 'sekolahh']
['jelekkkk', 'bgt', 'ngelag', 'terus', 'buat', 'belajar', 'susah', 'sinyal', 'lancar', 'tapi', 'muat', 'ulang', 'terus', 'huhhh']	['jelekkkk', 'bgt', 'ngelag', 'belajar', 'susah', 'sinyal', 'lancar', 'muat', 'ulang', 'huhhh']
['ruangguru', 'ko', 'lebih', 'sering', 'errornya', 'yaa', 'sekarang']	['ruangguru', 'ko', 'errornya']

Table 8. Comparison Before and After Slang Normalization

Before	After
['alhamdulillahh', 'kebantu', 'banget', 'belajar', 'ulangg', 'materi', 'paham', 'sekolahh']	alhamdulillah kebantu banget belajar ulang materi paham sekolah
['jelekkkk', 'bgt', 'ngelag', 'belajar', 'susah', 'sinyal', 'lancar', 'muat', 'ulang', 'huhhh']	jelek banget ngelag belajar susah sinyal lancar muat ulang huhhh
['ruangguru', 'ko', 'errornya']	ruangguru ko errornya

Slang normalization is performed to align non-standard words with their standard Indonesian forms. This process uses a slang dictionary sourced from GitHub, which is mapped into a dictionary structure. Several

terms, such as “bgt” and “ngga,” are converted to their formal equivalents to make the sentence's meaning clearer for the system. Examples of slang terms and their normalized are shown in Table 8.

3.3 Data Labelling

Sentiment labeling was carried out using the VADER method after all reviews were translated into English. This approach is particularly effective for handling mixed and ambiguous sentiments, as VADER converts nuanced expressions into a numerical compound score. The resulting labels are consistent and objective, providing a silver-standard supervision that enables IndoBERT to learn deeper contextual representations beyond surface-level polarity during training.

From the 4,904 valid reviews, VADER identified 3,375 positive, 428 negative, and 701 neutral instances. The proportions of these sentiment categories are illustrated in Figure 4, which provides an overview of the distribution of annotations. A detailed breakdown of class frequencies, along with examples of how polarity scores are converted into sentiment labels, is presented in Table 9.

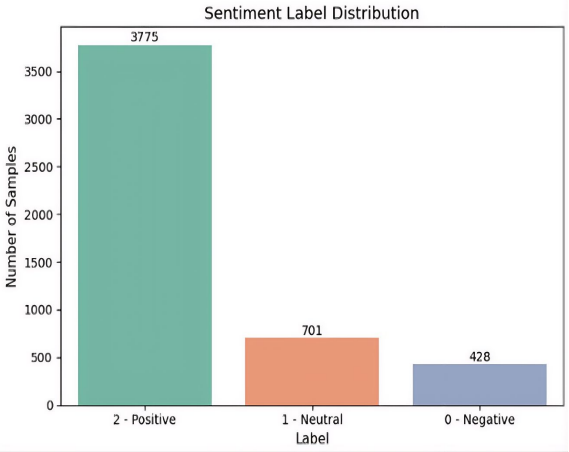


Figure 4. Visualization of Sentiment Label Distribution

Table 9. Labelling Data Results

Text	Translated	Score	Sentiment
aplikas terbaik untuk saat ini	*the best app for the moment*	0.6369	Positive
aplikasi rusak error nggak login	*The application is damaged, the error is not logging in*	-0.6808	Negative
bimbel sd smp sama sekolah smk	*tutoring for elementary and middle school and vocational school*	0.0	Neutral
ruangguru membantu kita mengerti	*Ruangguru helps us understand*	0.3818	Positive

To evaluate the reliability of the VADER-generated labels, a stratified subset of 463 reviews (10% of the dataset) was manually annotated by two independent annotators. The comparison between manual annotations and VADER labels resulted in an accuracy of 83.80% and a Cohen’s Kappa score of 0.7656, demonstrating a substantial level of agreement according to the criteria proposed by Landis and Koch. These results demonstrate a high level of agreement between human judgment and the automatically generated labels, suggesting that the VADER-based labeling process produces consistent sentiment classifications across the dataset, as illustrated in Figure 5.

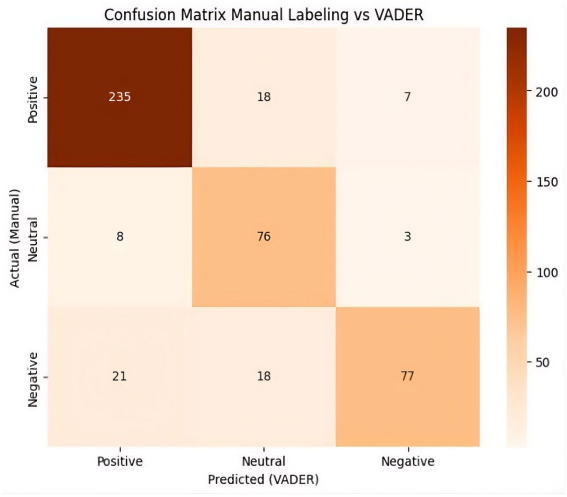


Figure 5. Confussion Matrix Manual Labels vs VADER

3.4 Data Split

The sentiment-labeled dataset was partitioned into training and test subsets through an 80:20 stratified sampling split to ensure that the proportions of the three sentiment classes remained balanced across both sets. This method preserves the original class distribution, allowing the training and evaluation processes to reflect the true characteristics of the dataset. By maintaining proportional representation, the model is exposed to realistic sentiment patterns during training while still being evaluated fairly on unseen data.

Table 10. Class Distribution in Training and Testing Data

Code	Sentiment	Training Data	Testing Data
0	Negative	342	86
1	Neutral	561	140
2	Positive	3020	755

Table 10 presents the class distribution resulting from this stratified split. The training set contains 342 negative, 561 neutral, and 3,020 positive reviews, while the test set consists of 86 negative, 140 neutral, and 755 positive reviews. These values show that the relative proportions of positive, neutral, and negative classes remain consistent between the training and test sets, confirming that the stratified splitting process successfully preserved the original label distribution.

3.5 Balancing Data

After the train test split, the training data still showed an imbalance among the three sentiment classes, with positive reviews appearing far more frequently than neutral or negative ones. To handle the observed disparity across sentiment classes, Random Oversampling (ROS) was implemented solely on the

training partition so that the evaluation on the test set remained fair and untouched. This method works by duplicating samples from the minority classes, allowing the model to learn from a more balanced set of examples without altering the original nature of the data.

The effect of this oversampling step can be seen in Figure 6, which compares the class distribution before and after ROS. The previously underrepresented neutral and negative classes become more proportionate once additional copies are added, giving the model a better chance to recognize their patterns during training. By implementing this technique, the model avoids over-fitting to the dominant class, encouraging more balanced feature learning across every sentiment label.

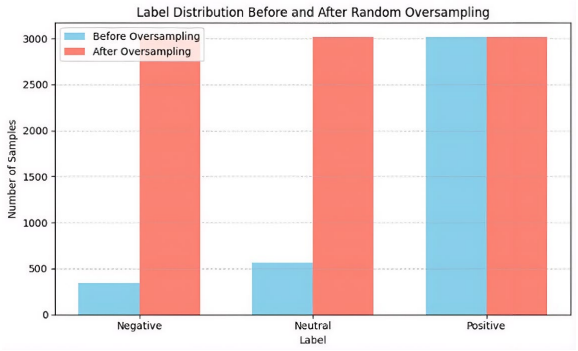


Figure 6. Label Distribution Before and After Oversampling

3.6 Bidirectional Long Short-Term Memory (BiLSTM)

The BiLSTM model was trained using data that had been fully preprocessed and balanced with ROS to ensure equal proportions across sentiment classes. Training employed 128-dimensional embeddings, 128 bidirectional LSTM units, two dense layers of 128 and 64 units equipped with a ReLU activation function alongside a 0.2 dropout rate to suppress overfitting and promote model generalization. Model optimization was executed using the Adam optimizer with a learning rate of 3e-4 and categorical cross-entropy loss, with training performed using a consistent batch size and validation at each epoch to monitor performance.

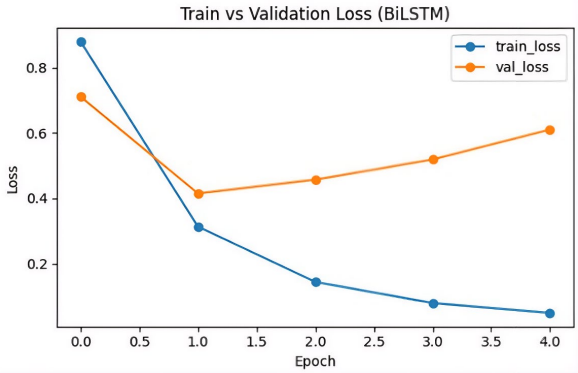


Figure 7. BILSTM Model Training and Test Loss Curve

The training dynamics of the BiLSTM model are visualized in Figure 7. The blue line represents the training loss, which shows a consistent and steep decline from around 0.9 in the initial epoch to approximately 0.05 by the fourth epoch. This learning trend indicates that the model rapidly absorbs complex textual features and contextual patterns from the training data within the initial epochs. However, the validation loss decreases only until the first epoch and then exhibits an upward rebound trend through the final epoch. The diverging directions of the two curves indicate the presence of overfitting, where the model continues to improve its performance on the training data but fails to maintain accuracy on the validation set. This overfitting behavior indicates that the BiLSTM architecture struggles to generalize when processing long and ambiguous reviews, confirming the limitations of RNN-based models highlighted in the Introduction.

3.7 IndoBERT

The IndoBERT model was trained on a dataset that underwent automatic labeling with the VADER method and was balanced using Random Oversampling (ROS). This approach ensures the model learns sentiment patterns from lexicon-based labels generated by VADER while maintaining proportional class distribution. Subsequent to the architectural configuration, a fine-tuning procedure was executed on the pre-trained IndoBERT-base model to adapt its representations to the target review dataset, aiming to enhance VADER's label representation with IndoBERT's ability to capture deeper Indonesian linguistic context, providing greater expressiveness than static embedding methods.

Tokenization was executed via the dedicated IndoBERT tokenizer, configured with a predefined maximum sequence length to meet the specific requirements of the architecture. The IndoBERT model was then trained with batch sizes of 16 for training and 32 for validation, a learning rate of 2e-5, a weight decay coefficient of 0.01, and an early-stopping strategy with a patience threshold of 2 epochs based on the F1-score. During training, evaluation was carried out at every epoch, and the best model was automatically selected using the *load_best_model_at_end* function based on the highest F1-score.

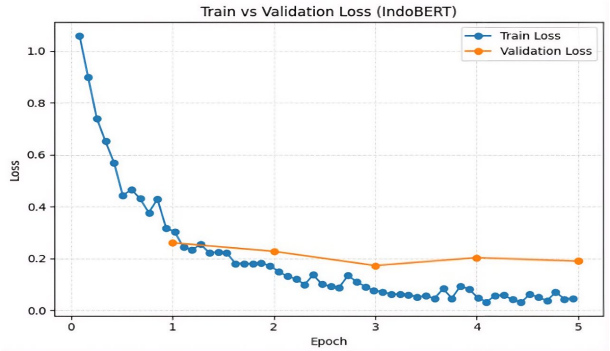


Figure 8. IndoBERT Model Training and Test Loss Curve

The training performance of the IndoBERT model is shown in Figure 8. Unlike the BiLSTM baseline, IndoBERT demonstrates stable learning behavior when handling long and ambiguous user reviews. The

training loss decreases sharply from values above 1.0 to around 0.05 by the end of training, reflecting the effectiveness of leveraging pre-trained weights and the model's ability to quickly adapt to the Ruangguru review dataset. Meanwhile, the validation loss remains relatively smooth and stable at around 0.2, with the best point achieved at the third epoch. Furthermore, the absence of a substantial gap between training and validation loss confirms that the model avoids overfitting, demonstrating a more robust generalization capacity than the BiLSTM benchmark. This demonstrates that IndoBERT can capture contextual patterns more accurately and apply them effectively to unseen data.

3.8 Evaluation

Both BiLSTM and IndoBERT models were evaluated on classifying user reviews into three sentiment categories (negative, neutral, and positive). The evaluation employed accuracy, precision, recall, and F1-score for each class to provide a more detailed understanding of the strengths and weaknesses of both models.

Table 11. Evaluation of BiLSTM and IndoBERT Model

Model	Class	Acc (%)	Prec (%)	Rec (%)	F1 (%)
BiLSTM	Negative	86.4	62.5	58.1	60.2
	Neutral		65.5	84.2	73.7
	Positive		94.3	90.0	92.1
IndoBERT	Negative	90.9	75.2	77.9	76.5
	Neutral		83.8	74.9	78.7
	Positive		93.8	94.6	94.6

The evaluation results in Table 11 reveal a clear performance difference between BiLSTM and IndoBERT. Overall, IndoBERT achieves an accuracy of 90.9 percent, which is higher than the 86.4 percent obtained by BiLSTM. At the per-class level, BiLSTM performs well in recognizing Positive sentiment, but its performance declines for the Neutral class and is lowest for the Negative class, where it achieves an F1-score of 60.2 percent. This indicates that the RNN-based architecture of BiLSTM struggles to capture complex linguistic patterns, such as negation and implicit criticism, which frequently appear in negative reviews. IndoBERT demonstrates a far more stable and consistent performance across all classes. The most substantial improvement appears in the Negative class, where IndoBERT achieves an F1-score of 76.5 percent, an increase of approximately 16 percent compared to BiLSTM. In addition, the F1-scores for the Neutral and Positive classes also improve, reaching 78.7 percent and 94.6 percent, respectively. These advantages highlight the effectiveness of the Transformer architecture and its Attention mechanism in IndoBERT. This mechanism enables bidirectional context understanding and captures long-range dependencies, improving classification of ambiguous reviews compared to conventional models.

A deeper consideration of the confusion matrices in Figures 9 and 10 illustrates still more that the performance is different on ambiguous sentiments between BiLSTM and IndoBERT. In the Positive class, IndoBERT has 721 correct predictions, the rate outperformed BiLSTM with 680 predictions, and significantly reduces the misclassification into Neutral from 53 cases to 18 cases. This finding implies that BiLSTM fails to detect implicit positive sentiment, which is particularly challenging in long or complex sentences.

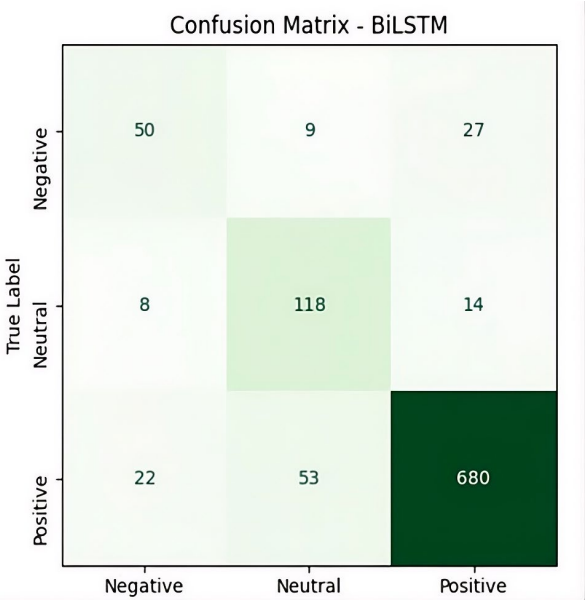


Figure 9. Confusion Matrix of the BiLSTM Model

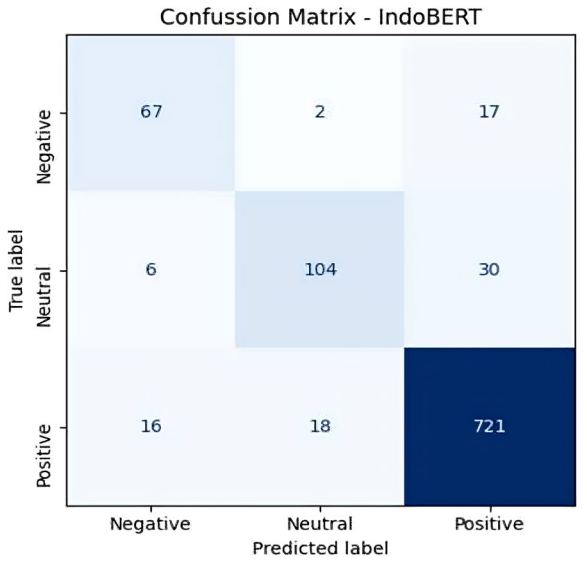


Figure 10. Confusion Matrix of the IndoBERT Model

For example, this applies to the Negative class, where IndoBERT correctly identifies 67 instances over 50 by BiLSTM and significantly reduces critical errors that could lead to misclassifying negative reviews into positive 17 cases for IndoBERT compared to 27 for BiLSTM. These results indicate that the self-attention mechanism in IndoBERT is more effective at recognizing contextual signals such as negation, sarcasm, and implicit criticism.

Overall, BiLSTM in general overpredicts the Neutral class with 118 correct predictions, but that also results in a large number of misclassified samples collected from other classes. Such a behaviour implies that the model defaults to neutral labels under uncertainty. IndoBERT can demonstrate more consistent classification error distribution overall showing better generalization and robustness in handling long and ambiguous reviews. Extending from these results, the proposed Hybrid VADER–IndoBERT framework has been developed to better describe complex sentiment patterns. To further position this approach within the broader research context, a comparison with related studies is presented in Table 11.

Table 12. Performance Comparison with Related Studies

Method	Acc (%)	Prec (%)	Rec (%)	F1 (%)
SVM-PSO [13]	82.27	-	-	-
Flair + IndoBERT [16]	81.29	81.30	81.29	81.26
VADER + GBM [31]	84.00	66.00	64.00	65.00
Proposed Method	90.93	90.82	90.93	90.83

According to Table 12, the Hybrid VADER–IndoBERT approach demonstrated superior performance across all evaluation metrics, reaching an accuracy of 90.93% and an F1-score of 90.83%. This result significantly outperforms SVM-PSO and VADER+GBM, suggesting that traditional machine learning and standalone lexicon-based approaches are less effective in handling complex and ambiguous review texts. Compared to the Flair + IndoBERT model, the proposed method also demonstrates superior performance, indicating that the integration of lexicon-based sentiment signals provides additional support for interpreting implicit and mixed sentiments.

Although differences in datasets limit direct comparability, the evaluation results indicate that the proposed hybrid framework offers enhanced robustness in processing long and ambiguous Indonesian user reviews compared to standalone baselines. This is particularly valuable for real-world scenarios with informal language and context-dependent expressions, highlighting the benefit of combining lexicon-based and Transformer-based approaches.

4. Conclusions

The findings of this research confirm the effectiveness of a Hybrid VADER–IndoBERT framework for sentiment analysis on long and ambiguous Indonesian user reviews of the Ruangguru application. By integrating lexicon-based silver-standard labeling using VADER with the contextual representation capabilities of IndoBERT, the proposed approach addresses key challenges related to mixed sentiments, sentence length, and linguistic ambiguity. Experimental results on 4,904 reviews show that the IndoBERT model achieves an overall accuracy of 90.9%, outperforming the BiLSTM baseline at 86.4%. These findings confirm that Transformer-based architectures provide more robust sentiment classification for complex Indonesian review texts. From a practical perspective, the proposed framework enables more reliable extraction of user sentiment insights to support data-driven service improvement. Nevertheless, this study is constrained by its reliance upon on lexicon-based labeling and a single application domain, and future work should incorporate human-annotated data, data augmentation strategies, and domain-specific models to further enhance robustness.

Acknowledgements

This research was made possible through the financial support provided by the Ministry of Education, Culture, Research, and Technology Indonesia (Kemdiktisaintek) under Fundamental Research Regular BIMA 2026 (Grant number: 195/DST/C3/DT.05.00/2026). Furthermore, the authors would also like to thank Universitas Dian Nuswantoro for the continuous support in completing this study

References

[1] A. Haleem, M. Javaid, M. A. Qadri, and R. Suman, "Understanding The Role of Digital Technologies in Education: A review," *Sustainable Operations and Computers*, vol. 3, pp. 275–285, 2022, doi: 10.1016/j.susoc.2022.05.004.

[2] I. Y. Alyoussef, "Acceptance of E-Learning in Higher Education: The Role of Task-Technology Fit with The Information Systems Success Model," *Heliyon*, vol. 9, no. 3, p. e13751, Mar. 2023, doi: 10.1016/j.heliyon.2023.e13751.

[3] We Are Social, "Meltwater," *Digital 2024 Indonesia*, 2024. Accessed: Sep. 28, 2025. [Online]. Available: https://wearesocial.com/id/wp-content/uploads/sites/19/2024/02/Digital_2024_Indonesia.pdf

[4] Mussalimun, E. H. Khasby, G. I. Dzikrillah, and Muljono, "Comparison of K- N earest Neighbor (K -NN) and Naïve Bayes Algorithm for Sentiment Analysis on Google Play Store Textual Reviews," in *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, Semarang, Indonesia: IEEE, Sep. 2021, pp. 180–184. doi: 10.1109/ICITACEE53184.2021.9617217.

[5] R. Kurniawan, H. O. L. Wijaya, and R. P. Aprisusanti, "Sentiment Analysis of Google Play Store User Reviews on Digital Population Identity App Using K-Nearest Neighbors," *SISFOKOM*, vol. 13, no. 2, pp. 170–178, Jun. 2024, doi: 10.32736/sisfokom.v13i2.2071.

[6] A. A. Mohammad *et al.*, "Customer Sentiment Analysis for Food and Beverage Development in Restaurants using AI in Jordan," *Data and Metadata*, vol. 4, p. 922, Apr. 2025, doi: 10.56294/dm2025922.

[7] S. Al-Natour and O. Turetken, "A Comparative Assessment of Sentiment Analysis and Star Ratings for Consumer Reviews," *International Journal of Information Management*, vol. 54, p. 102132, Oct. 2020, doi: 10.1016/j.ijinfomgt.2020.102132.

[8] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, "Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 30–40, Mar. 2025, doi: 10.20473/jisebi.11.1.30-40.

[9] A. S. Paramita, "Fine-Grained Sentiment Analysis Approach on Customer Reviews Based on Aspect-Level Emotion Detection," *J. Appl. Data Sci.*, vol. 6, no. 3, pp. 2235–2247, Sep. 2025, doi: 10.47738/jads.v6i3.964.

[10] A. Mabrouk, R. P. D. Redondo, and M. Kayed, "Deep Learning-Based Sentiment Classification: A Comparative

Survey," *IEEE Access*, vol. 8, pp. 85616–85638, 2020, doi: 10.1109/ACCESS.2020.2992013.
[11] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, p. 102048, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.
[12] Z. A. Diekson, M. R. B. Prakoso, M. S. Q. Putra, M. S. A. F. Syaputra, S. Achmad, and R. Sutoyo, "Sentiment analysis for customer review: Case study of Traveloka," *Procedia Computer Science*, vol. 216, pp. 682–690, 2023, doi: 10.1016/j.procs.2022.12.184.
[13] S. Fahmi, L. Purnamawati, G. F. Shidik, M. Muljono, and A. Z. Fanani, "Sentiment Analysis of Student Review in Learning Management System Based on Sastrawi Stemmer and SVM-PSO," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia: IEEE, Sep. 2020, pp. 643–648. doi: 10.1109/iSemantic50169.2020.9234291.
[14] Y. Handayani, A. R. Hakim, and Muljono, "Sentiment Analysis of Bank BNI User Comments Using the Support Vector Machine Method," in *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia: IEEE, Sep. 2020, pp. 202–207. doi: 10.1109/iSemantic50169.2020.9234230.
[15] Z. Hameed and B. Garcia-Zapirain, "Sentiment Classification Using a Single-Layered BiLSTM Model," *IEEE Access*, vol. 8, pp. 73992–74001, 2020, doi: 10.1109/ACCESS.2020.2988550.
[16] K. Kusnawi and K. Anam, "Comparison ff Sentiment Labeling Using Textblob, Vader, and Flair in Public Opinion Analysis Post-2024 Presidential Inauguration with IndoBERT," *J. Tek. Inform. (JUTIF)*, vol. 6, no. 2, pp. 803–818, Apr. 2025, doi: 10.52436/1.jutif.2025.6.2.4015.
[17] W. Christian, D. Adamlu, A. Yu, and D. Suhartono, "Leveraging IndoBERT and DistilBERT for Indonesian emotion classification in e-commerce reviews," *Procedia Computer Science*, vol. 269, pp. 321–330, 2025, doi: 10.1016/j.procs.2025.08.284.
[18] B. Setiawan, "A Review of Sentiment Analysis Applications in Indonesia Between 2023-2024," *Journal Information Engineering and Educational Technology*, vol. 08, 2024, doi: https://doi.org/10.26740/jieet.v8n2.p71-83.
[19] K. Barik and S. Misra, "Analysis of Customer Reviews with an Improved VADER Lexicon Classifier," *Journal of Big Data*, vol. 11, no. 1, p. 10, Jan. 2024, doi: 10.1186/s40537-023-00861-x.
[20] D. Marutho, Muljono, S. Rustad, and Purwanto, "Sentiment Analysis Optimization Using Vader Lexicon on Machine Learning Approach," in *2022 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Surabaya, Indonesia: IEEE, Jul. 2022, pp. 98–103. doi: 10.1109/ISITIA56226.2022.9855341.
[21] F. S. Lestari, H. Harliana, M. M. Huda, and T. Prabowo, "Sentiment Analysis of iPusnas Application Reviews on Google Play Using Support Vector Machine," *Proceedings of the International Seminar on Business, Education and Science*, vol. 1, pp. 178–188, Oct. 2022, doi: 10.29407/int.v1i1.2656.
[22] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Nov. 02, 2020, *arXiv*: arXiv:2011.00677. doi: 10.48550/arXiv.2011.00677.
[23] I. Gambo, R. Massenon, R. O. Ogundokun, S. Agarwal, and W. Pak, "Identifying and Resolving Conflict in Mobile Application Features Through Contradictory Feedback Analysis," *Heliyon*, vol. 10, no. 17, p. e36729, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36729.
[24] H.-T. Duong and T.-A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Computational Social Networks*, vol. 8, no. 1, p. 1, Dec. 2021, doi: 10.1186/s40649-020-00080-x.
[25] S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, "A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets," *IEEE Access*, vol. 11, pp. 11811–11826, 2023, doi: 10.1109/ACCESS.2023.3242290.
[26] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, 1977, doi: https://doi.org/10.2307/2529310.
[27] Hairani, A. Anggrawan, Muhammad Ridho Akbar, Khasnur Hidjah, and Muhammad Innuddin, "Optimizing Sentiment Analysis for Lombok Tourism Using SMOTE and Chi-Square with Machine Learning," *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 4, pp. 706–713, Jul. 2025, doi: 10.29207/resti.v9i4.6623.
[28] D. A. Sani, "A Random Oversampling and BERT-based Model Approach for Handling Imbalanced Data in Essay Answer Correction," *J.INFOTEL*, vol. 16, no. 4, pp. 729–739, Dec. 2024, doi: 10.20895/infotel.v16i4.1224.
[29] A. He and M. Abisado, "Text Sentiment Analysis of Douban Film Short Comments Based on BERT-CNN-BiLSTM-Att Model," *IEEE Access*, vol. 12, pp. 45229–45237, 2024, doi: 10.1109/ACCESS.2024.3381515.
[30] B. Rahman and Maryani, "Optimizing Customer Satisfaction Through Sentiment Analysis: A BERT-Based Machine Learning Approach to Extract Insights," *IEEE Access*, vol. 12, pp. 151476–151489, 2024, doi: 10.1109/ACCESS.2024.3478835.
[31] W. Aljedaani, F. Rustam, S. Ludi, A. Ouni, and M. W. Mkaouer, "Learning Sentiment Analysis for Accessibility User Reviews," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, Melbourne, Australia: IEEE, Nov. 2021, pp. 239–246. doi: 10.1109/ASEW52652.2021.00053.