JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Sentence-Level Sentiment Analysis of Indonesian App Reviews Using IndoBERTweet

Inge Najwa Aqiilah *, Ristu Saptono *, and Akhmad Syaifuddin

Faculty of Information Technology and Data Science, Sebelas Maret University, Surakarta 57126, Indonesia; e-mail : inge.najwaa@student.uns.ac.id; ristu.saptono@staff.uns.ac.id; akhmadsyaifuddin@staff.uns.ac.id

* Corresponding Author : Inge Najwa Aqiilah, Ristu Saptono

Abstract: Document-level sentiment analysis assigns a single polarity label to an entire review, often obscuring opinion diversity within multi-sentence submissions. This limitation is particularly evident in reviews of multi-service platforms, where users frequently express heterogeneous opinions toward different aspects of the platform in the same review. To address this challenge, this study proposes a sentence-level sentiment analysis framework for Indonesian Gojek app reviews collected from the Google Play Store. The proposed framework introduces a two-stage segmentation strategy that combines punctuation-aware rules with conjunction-aware splitting based on coordinating and adversative conjunctions (e.g., *tapi* [but], *padahal* [even though]) to identify opinion boundaries and decompose mixed-sentiment reviews into independently classifiable sentence units. A total of 14,730 raw reviews collected between May and July 2025 were subjected to data cleaning and quality filtering, resulting in 7,187 valid reviews that were further segmented into 14,187 sentence-level instances. Each instance was manually annotated by three annotators using a four-class labeling scheme consisting of app-positive, app-negative, app-neutral, and service categories. Sentiment-level inter-annotator agreement, computed on the subset of instances unanimously categorized as app-related by all three annotators (n = 4,384), achieved substantial agreement (Fleiss' κ = 0.636). Hyperparameter optimization was conducted using Optuna with the Tree-structured Parzen Estimator (TPE) sampler across four experimental scenarios. The best performance was achieved by IndoBERTweet under Stratified K-Fold evaluation, attaining an accuracy of 0.751 and a macro F1-score of 0.729, outperforming all IndoBERT configurations. The results demonstrate the effectiveness of domain-adaptive pre-training on informal Indonesian text and highlight the value of conjunction-aware segmentation for preserving fine-grained opinion structures in mixed-sentiment reviews. These findings suggest that domain-aligned language representations provide a practical and effective solution for sentence-level sentiment analysis of Indonesian app reviews.

Keywords: Sentence-level Sentiment Analysis; IndoBERTweet; IndoBERT; Indonesian App Reviews; Natural Language Processing; Transformer-based Classification; User-Generated Content Analytics; Sustainable Digital Services.

Received: May, 20th 2026
Revised: June, 6th 2026
Accepted: June, 20th 2026
Published: June, 21st 2026

1. Introduction

The rapid growth of super-app platforms in Indonesia has generated a large volume of user-generated feedback on digital distribution channels such as the Google Play Store [1]–[3]. As of early 2026, one of Indonesia's largest super-apps, Gojek, had surpassed 100 million downloads and accumulated more than 6.6 million reviews, making it one of the richest publicly available sources of informal user opinion in the Indonesian language [4]. As an integrated platform offering transportation, logistics, food delivery, and digital payment services, Gojek provides a valuable environment for studying user perceptions across multiple service domains. Consequently, Natural Language Processing (NLP)-based sentiment analysis has become an increasingly important approach for transforming large volumes of unstructured review text into actionable opinion insights [3], [5], [6].

The dominant paradigm in app review sentiment analysis assigns a single sentiment label—typically positive, negative, or neutral—to an entire review [3], [7], [8]. While this document-level approach is effective for short and topically consistent reviews, it becomes inadequate when multiple opinions are expressed within a single submission [7]. This limitation is particularly evident in multi-service platforms such as Gojek, where users often evaluate different aspects of the platform simultaneously. For example, a review may praise the application interface while criticizing service-related issues such as driver availability or delivery delays [1]. Under document-level classification, these distinct opinions are collapsed into a single label, obscuring valuable information that could otherwise support more targeted service improvements and product development decisions [7], [9]. This limitation motivates the need for sentence-level sentiment analysis, where individual opinion units can be analyzed independently [7].

Aspect-Based Sentiment Analysis (ABSA) addresses this granularity issue by identifying sentiments associated with specific aspects such as usability, performance, or pricing [7], [10]. However, ABSA relies on explicit aspect definitions and annotations, an assumption that is often difficult to satisfy in Indonesian user-generated content characterized by slang, abbreviations, code-mixing, informal grammar, and implicit references [10]–[12]. In addition, ABSA requires the annotation of both aspect spans and sentiment labels, increasing annotation complexity and limiting scalability [7]. A more practical alternative is sentence-level sentiment analysis, which preserves opinion granularity without requiring predefined aspect categories by treating each segmented sentence as an independent classification unit [5] [7]. The primary challenge in this setting is determining meaningful segmentation boundaries within informal review text. To address this issue, we employ a conjunction-aware segmentation strategy that leverages coordinating and adversative conjunctions—such as *tapi* (but), *padahal* (even though), and *karena* (because)—which frequently signal opinion transitions in informal Indonesian. Combined with punctuation-aware rules, this strategy enables mixed-sentiment reviews to be decomposed into semantically coherent sentence units prior to classification.

A second challenge concerns the suitability of existing Indonesian pre-trained language models for informal app review data. IndoBERT [11], [13], one of the most widely adopted Indonesian transformer models, was pre-trained primarily on formal sources including Indonesian Wikipedia, news articles, and web-crawled text. Although this corpus provides strong linguistic representations for standard Indonesian, it does not fully capture the informal language patterns, spelling variations, abbreviations, and code-mixed expressions commonly found in app reviews [6]. Previous studies have reported reduced effectiveness of IndoBERT on informal Indonesian text, particularly for minority sentiment categories [6] [14]. In contrast, IndoBERTweet [12] was pre-trained on approximately 26 million Indonesian tweets using a domain-adaptive vocabulary initialization strategy specifically designed for informal language. Prior research has consistently demonstrated its superiority over IndoBERT in informal text classification tasks [12], [14]. However, its effectiveness for sentence-level sentiment analysis of Indonesian app reviews remains unexplored.

This study makes three main contributions. First, we propose a sentence-level sentiment analysis framework for Indonesian app reviews based on a conjunction-aware segmentation strategy. The proposed approach identifies semantically meaningful opinion boundaries while incorporating regex-based exceptions to preserve financial expressions (e.g., Rp10.000), temporal expressions (e.g., 10.30), and common Indonesian abbreviations, thereby preventing erroneous sentence fragmentation. Second, we investigate the application of IndoBERTweet to sentence-level app review classification and establish a benchmark baseline using a four-class labeling scheme consisting of app-positive, app-negative, app-neutral, and service categories, a setting that has not previously been examined [2], [12]. Third, we provide a comprehensive comparative evaluation of IndoBERTweet and IndoBERT under both Stratified 5-Fold Cross-Validation and Time-Based evaluation schemes, enabling assessment of model performance under both conventional and temporally realistic deployment scenarios [15], [16].

The remainder of this paper is organized as follows. Section 2 reviews previous studies on document-level and sentence-level sentiment analysis, informal Indonesian language modeling, and the research gap addressed in this work. Section 3 presents the proposed methodology, including data collection, preprocessing, conjunction-aware segmentation, annotation, resampling strategies, model development, and evaluation metrics. Section 4 reports the experimental setup and results, including hyperparameter optimization, comparative

performance analyses, and error analysis. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Literature Review

2.1. Document-Level Sentiment Analysis

Early studies on Indonesian app review sentiment analysis primarily relied on conventional machine learning algorithms and treated each review as a single classification unit. Handani et al. [1] employed Naive Bayes for sentiment classification of Gojek reviews collected from the Google Play Store, demonstrating the feasibility of large-scale automated opinion mining. Similarly, Pribadi et al. [8] applied Random Forest to sentiment analysis of PeduliLindungi reviews. These studies established the practicality of sentiment classification for mobile application reviews; however, they share a common methodological limitation in that each review is assigned a single sentiment label regardless of the number of opinions expressed within the text [4], [17].

More recent studies have adopted transformer-based architectures to improve classification performance. For example, Setiawan et al. [3] fine-tuned BERT on Tokopedia reviews and reported an accuracy of 89%. Although transformer models substantially improve predictive performance, they generally retain the document-level assumption and therefore inherit the same limitation of collapsing multiple opinions into a single sentiment label. Consequently, valuable opinion-level information may be lost when reviews contain heterogeneous evaluations of different services or application components. This limitation motivates the exploration of sentence-level sentiment analysis as a more fine-grained alternative.

2.2. Sentence-Level Sentiment Analysis

Sentence-level sentiment analysis provides an intermediate granularity between document-level classification and Aspect-Based Sentiment Analysis (ABSA), enabling individual opinion units to be analyzed without requiring explicit aspect annotations. Prior studies have shown that sentence-level modeling can improve both interpretability and sentiment resolution in texts containing multiple opinions.

Mifrah and Benlahmar [7] conducted a comparative study of sentence-level sentiment classification using LSTM, BiLSTM, GRU, and BERT on a dataset of 1.6 million tweets. Their results demonstrated that sentence-level analysis captures sentiment information more effectively than document-level aggregation, with BERT achieving the highest accuracy of 87.36%. Similarly, Du et al. [18] showed that sentence-level analysis of financial reports recovers sentiment signals that are often obscured by document-level representations. Li et al. [9] further demonstrated the value of sentence-level sentiment analysis for Amazon product reviews, where review decomposition enabled more detailed opinion extraction and improved explainability.

Recent research has also highlighted the importance of preprocessing strategies for informal user-generated content. Budianoor et al. [6] conducted a systematic preprocessing ablation study for IndoBERT-based sentiment classification of Indonesian culinary reviews and found that slang normalization contributed the largest performance improvement among all preprocessing components, yielding a macro F1-score gain of +0.0609. These findings suggest that sentence-level sentiment analysis in informal Indonesian text should be accompanied by preprocessing strategies that account for domain-specific linguistic characteristics.

2.3. Informal Indonesian Language Modeling

Indonesian user-generated content frequently contains slang, abbreviations, spelling variations, code-mixing, and informal grammatical constructions, creating challenges for language models trained primarily on formal text sources. Sebastian et al. [19] reported that BERT-based models exhibit performance degradation when applied to informal Indonesian text, highlighting the importance of domain alignment between pre-training and downstream tasks.

To address this limitation, Koto et al. [12] introduced IndoBERTweet, the first large-scale pre-trained language model specifically designed for informal Indonesian. The model was trained on approximately 26 million Indonesian tweets collected between December 2019 and December 2020 and incorporates a domain-adaptive vocabulary initialization strategy

that extends the IndoBERT vocabulary with social-media-specific tokens. This approach significantly reduces training cost while consistently outperforming IndoBERT, multilingual BERT (mBERT), and MalayBERT across multiple Indonesian Twitter classification benchmarks.

The importance of domain alignment has also been observed in downstream sentiment analysis studies. Hanafi et al. [4] applied Word2Vec and LSTM to Gojek review sentiment classification, while Singh et al. [5] employed BERT for Play Store review analysis. Their findings, together with evidence reported by Darnoto and Firmawan [20], indicate that alignment between the pre-training corpus and the target domain plays a critical role in classification performance for informal Indonesian text.

Additional evidence is provided by Supriyadi and Makatita [21], who successfully applied IndoBERT-based models to sentiment analysis of TikTok comments regarding QRIS adoption. Their work demonstrates that transformer models pre-trained on informal Indonesian data can be effectively transferred across different user-generated content domains. Collectively, these findings support the selection of IndoBERTweet as a suitable candidate for sentence-level sentiment analysis of Indonesian app reviews.

2.4. Research Gap

The existing literature reveals three important research gaps. First, most studies on Indonesian app review sentiment analysis continue to adopt document-level classification, assigning a single sentiment label to an entire review despite the presence of multiple opinions within the same submission [2], [3], [5], [8]. Second, studies that employ transformer-based models for Indonesian sentiment analysis often apply models pre-trained on formal corpora to informal user-generated content without explicitly addressing the resulting vocabulary and linguistic mismatch [2], [6]. Third, while sentence-level and aspect-level sentiment analysis have been explored in other domains and languages [9], [10], their application to Indonesian app reviews remains limited.

To the best of our knowledge, no previous study has investigated the use of IndoBERTweet [12] for sentence-level sentiment classification of Indonesian app reviews within a conjunction-aware segmentation framework. Furthermore, existing studies have not examined the effect of domain-adaptive pre-training under both conventional cross-validation and temporally realistic evaluation settings. This study addresses these gaps by introducing a conjunction-aware sentence segmentation strategy, a four-class annotation scheme that distinguishes app-directed and service-directed opinions, and a comprehensive comparison of IndoBERT [13] and IndoBERTweet [12] using both Stratified 5-Fold Cross-Validation and Time-Based evaluation protocols.

3. Proposed Method

This study follows a multi-stage methodology for sentence-level sentiment analysis of Gojek app reviews, encompassing data collection, text preprocessing, sentence segmentation, annotation, model development, and evaluation. The overall workflow of the proposed framework is illustrated in Fig. 1.

3.1. Data Collection

User reviews of the Gojek mobile application were collected from the Google Play Store using the google-play-scraper package [3]. This approach enables the automated acquisition of large-scale user-generated content and facilitates the construction of a sufficiently representative corpus for sentiment analysis. The data collection period spans from May to July 2025. This three-month window was selected to provide sufficient temporal coverage for the Time-Based evaluation scheme while maintaining a relatively stable application environment. Limiting the collection period to a post-feature-update interval helps reduce the influence of major product releases that could introduce abrupt shifts in user sentiment distributions.

A total of 14,730 raw reviews were initially collected. Several filtering procedures were subsequently applied to improve corpus quality. First, duplicate reviews containing identical text, user identifiers, and timestamps were removed, as these entries represent platform artifacts rather than distinct user opinions. Second, non-Indonesian reviews were excluded using language-based heuristic filtering to preserve linguistic consistency within the dataset. Finally, reviews containing fewer than three characters after lowercasing were discarded because they

provide insufficient semantic information for sentence-level sentiment analysis. After applying these filtering steps, 7,187 valid review entries remained and were forwarded to the sentence segmentation stage.

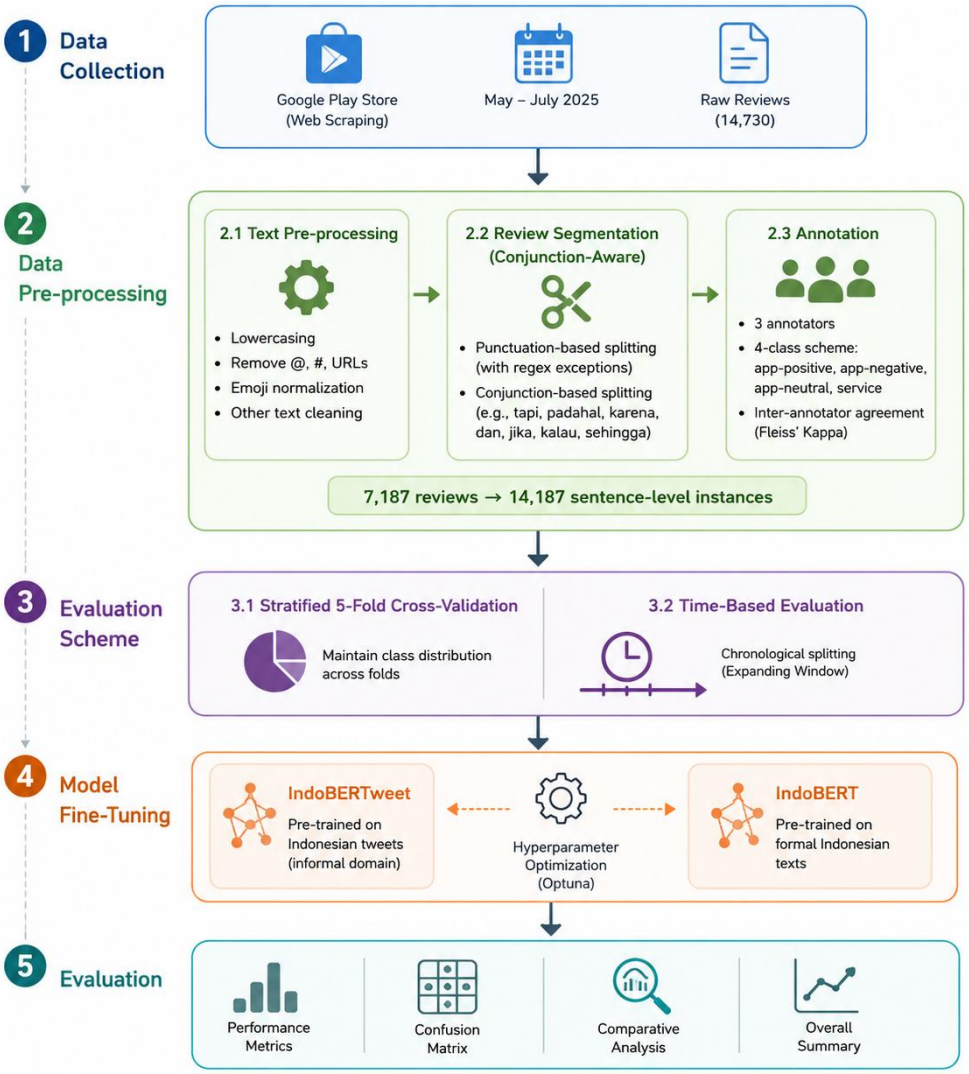

Figure 1. Proposed Method

3.2. Data Pre-processing

Data preprocessing plays a critical role in Natural Language Processing (NLP) and directly influences the quality of downstream classification models [7], [22]. The objective of this stage is to transform raw user-generated text into a standardized representation that can be effectively processed by transformer-based language models. The preprocessing pipeline consists of text normalization and sentence segmentation, as described in the following subsections.

3.2.1. Text Pre-processing

The 7,187 filtered reviews underwent a series of standard text normalization procedures, including lowercasing, removal of usernames (@), hashtags (#), and URLs. These operations reduce textual noise while preserving the semantic content required for sentiment classification. To retain non-verbal affective cues frequently found in informal user-generated content, emoji characters were converted into their textual descriptions using the demojize function provided by the Python emoji library. For example, the emoji 😊 was transformed into the textual token :smiling_face_with_smiling_eyes:. This conversion enables transformer models to exploit sentiment-related information conveyed through emoji usage rather than treating such symbols as opaque characters.

Table 1 presents an example of the preprocessing pipeline together with the resulting tokenization outputs generated by IndoBERTweet and IndoBERT. The example illustrates differences in vocabulary coverage and subword segmentation between the two models, particularly when processing informal Indonesian expressions such as bngettt and go-jek. As IndoBERTweet was pre-trained on large-scale informal Indonesian social media text, it generally produces tokenizations that are more aligned with colloquial language usage than IndoBERT [12], [13].

Table 1. Example of Text Pre-processing Pipeline

Stage	Output
Raw Text	*aku suka Bngettt, selama satu thn ini aku sering banget pake aplikasi GO-JEK untuk berpergian dan Memasan makanan 🥹 🤣 Mksih GO-JEK...*
After Lowercasing	*aku suka bngettt, selama satu thn ini aku sering banget pake aplikasi go-jek untuk berpergian dan memasan makanan 🥹 🤣 mksih go-jek...*
After Emoji Normalization	*aku suka bngettt, selama satu thn ini aku sering banget pake aplikasi go-jek untuk berpergian dan memasan makanan :smiling_face_with_3_hearts: :face_with_hand_over_mouth: mksih go-jek...*
After IndoBERTweet Tokenization	['[CLS]', '*aku*', '*suka*', '*bnget*', '##tt', ',', '*selama*', '*satu*', '*thn*', '*ini*', '*aku*', '*sering*', '*banget*', '*pake*', '*aplikasi*', 'go', '-', '*jek*', '*untuk*', '*berpergian*', '*dan*', '*memas*', '##an', '*makanan*', ':', 'smiling', '_', 'face', '_', 'with', '_', '3', '_', 'hearts', ':', ':', 'face', '_', 'with', '_', 'hand', '_', 'over', '_', 'mouth', ':', 'mks', '##ih', 'go', '-', 'jek', '.', '.', '.', '[SEP]']
After IndoBERT Tokenization	['[CLS]', '*aku*', '*suka*', 'bn', '##get', '##tt', ',', '*selama*', '*satu*', '*thn*', '*ini*', '*aku*', '*sering*', '*banget*', '*pake*', '*aplikasi*', 'go', '-', '*je*', '##k', '*untuk*', '*berper*', '##gian', 'dan', '*memas*', '##an', '*makanan*', ':', 'sm', '##iling', '_', 'face', '_', 'with', '_', '3', '_', 'heart', '##s', ':', ':', 'face', '_', 'with', '_', 'hand', '_', 'over', '_', 'mou', '##th', ':', 'mk', '##sih', 'go', '-', 'je', '##k', '.', '.', '.', '[SEP]']
IndoBERTweet Tokenization ID	[3, 2254, 4346, 20492, 4045, 16, 2015, 1713, 24374, 1540, 2254, 2633, 10218, 17420, 5952, 3686, 17, 13512, 1559, 20753, 1501, 2834, 1476, 3005, 30, 5989, 41, 22753, 41, 7002, 41, 23, 41, 7615, 30, 30, 22753, 41, 7002, 41, 6449, 41, 9316, 41, 7352, 30, 15722, 1567, 3686, 17, 13512, 18, 18, 18, 4]
IndoBERT Tokenization ID	[3, 2254, 4346, 10044, 4607, 4045, 16, 2015, 1713, 24374, 1540, 2254, 2633, 10218, 17420, 5952, 3686, 17, 10088, 942, 1559, 4248, 8399, 1501, 2834, 1476, 3005, 30, 2829, 4559, 41, 22753, 41, 7002, 41, 23, 41, 13452, 944, 30, 30, 22753, 41, 7002, 41, 6449, 41, 9316, 41, 18138, 3260, 30, 10259, 5541, 3686, 17, 10088, 942, 18, 18, 18, 4]

3.2.2. Review Segmentation

The primary methodological contribution of the preprocessing stage is the proposed sentence segmentation strategy. Each review, originally treated as a document-level unit, was decomposed into smaller sentence-level units using a two-stage segmentation mechanism designed to preserve opinion boundaries in informal Indonesian text. The first stage employs punctuation-based segmentation, where reviews are split using terminal punctuation marks (., !, and ?). To prevent erroneous segmentation, several regex-based exceptions were implemented to preserve common non-sentiment punctuation patterns, including (a) decimal numbers and currency expressions (e.g., Rp10.000), (b) time expressions (e.g., 10.30), and (c) frequently used Indonesian abbreviations such as dkk. and dll.. The second stage applies conjunction-aware segmentation based on coordinating and adversative conjunctions that frequently indicate opinion transitions in informal Indonesian reviews, including *tapi* (but), *padahal* (even though), *karena* (because), *kalau* (if), *jika* (if), *sehingga* (so that), and *dan* (and). Conjunction-based splitting was applied only when the conjunction connected clauses containing independent predicates or opinion-bearing expressions, thereby reducing the risk of over-segmentation in simple coordinated phrases.

This design directly addresses the mixed-sentiment nature of app reviews, where a single submission may contain multiple opinions directed toward different aspects of the platform. For example, the review "*Aplikasi bagus, tapi driver sering telat*" ("The application is good, but the driver is often late") contains both a positive application-related opinion and a service-

related complaint. Under the proposed framework, these opinions are separated into independent sentence units that can subsequently be classified individually. An illustrative example of the segmentation process and its corresponding split triggers is presented in Table 2. Through this procedure, the dataset expanded from 7,187 review entries to 14,187 sentence-level instances, substantially increasing the granularity of the sentiment analysis process.

Table 2. Example of Review Segmentation Results

Original Review	Segmented Sentence Unit	Split Trigger
Sebel banget, sekarang gojek nggak bisa pake promo, padahal udah langganan banget sama gojek. Dimasukin promo kodenya, error mulu jadi agak males sekarang T_T	*Sebel banget, sekarang gojek nggak bisa pake promo, padahal udah langganan banget sama gojek.*	Conjunction: *padahal* Punctuation: "."
	Dimasukin promo kodenya, error mulu jadi agak males sekarang T_T	End of review

As shown in Table 2, the proposed segmentation framework is capable of identifying semantically meaningful opinion boundaries that would otherwise remain embedded within a single review-level representation.

3.2.3. Annotation

Following the segmentation stage, each sentence-level instance was manually annotated to construct the final classification dataset. Three annotators were recruited for this task. To ensure annotation quality and domain familiarity, all annotators were required to (1) possess strong proficiency in the Indonesian language, (2) have prior experience in data annotation activities, (3) be active users of the Gojek mobile application, and (4) hold or be currently pursuing an undergraduate degree.

The annotation scheme distinguishes between two broad categories of review content: app-related reviews and service-related reviews. App-related reviews refer to opinions concerning the application itself, including user interface design, usability, functionality, and technical performance. These instances were further assigned one of three sentiment labels: app-positive, app-negative, or app-neutral. In contrast, service-related reviews describe experiences associated with external service components, such as drivers, couriers, merchants, or delivery operations. Since the primary objective of this study is application-focused sentiment classification, service-related statements were assigned a dedicated service category rather than being associated with sentiment polarity labels. Because the three annotators did not always agree on the category assignment (app vs. service) for a given instance, the category and sentiment labels for each instance were subsequently resolved through majority voting and consensus discussion. Following annotation and majority-voting aggregation, a total of 14,187 sentence-level instances were retained for subsequent experiments. The final class distribution is illustrated in Fig. 2.

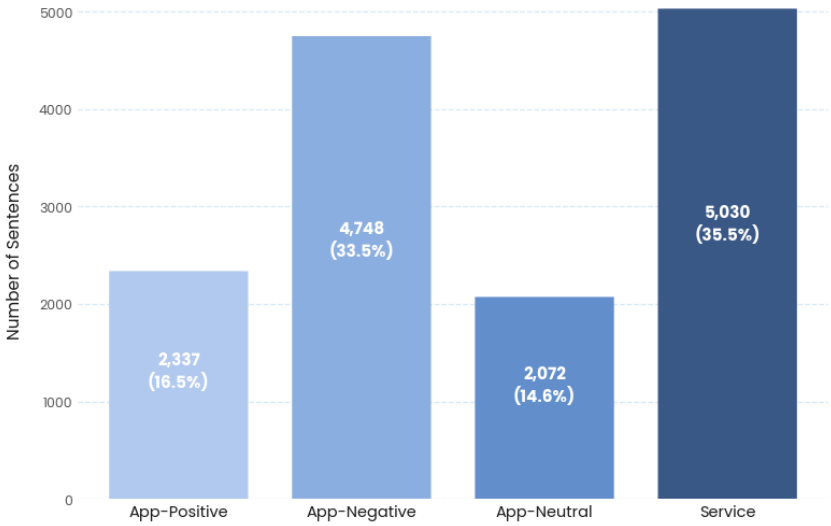

Figure 2. Final dataset distribution.

To assess annotation reliability, inter-rater agreement was measured using Fleiss' Kappa, a widely adopted extension of Cohen's Kappa for evaluating agreement among more than two annotators [23], [24]. Unlike raw agreement measures, Fleiss' Kappa accounts for agreement occurring by chance and therefore provides a more reliable assessment of annotation consistency [24]. The coefficient is computed as:

$$\kappa = \frac{\bar{P}_o - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

where $\bar{P}_o$ denotes the mean observed agreement and $\bar{P}_e$ represents the expected agreement by chance. The resulting coefficient ranges from values below 0 (systematic disagreement) to 1 (perfect agreement). The interpretation adopted in this study follows the benchmark proposed by Landis and Koch [25], as summarized in Table 3.

Table 3. Interpretation of Fleiss' Kappa Values

Kappa Value (κ)	Strength of Agreement
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Sentiment-level inter-rater reliability was assessed on the subset of 4,384 instances unanimously categorized as app-related by all three annotators. Within this subset, the Fleiss' Kappa analysis yielded a coefficient of $\kappa = 0.636$ ($\bar{P}_o = 0.797$, $\bar{P}_e = 0.445$), indicating substantial agreement according to the Landis and Koch interpretation scale [25]. A total of 3,111 instances (70.96%) received identical sentiment labels from all three annotators, 1,161 instances (26.48%) achieved majority agreement (2 out of 3 annotators), and 112 instances (2.55%) exhibited complete disagreement; these unresolved cases were reviewed through a consensus discussion process.

3.3. Evaluation Scheme

After annotation, the sentence-level dataset was partitioned according to two evaluation schemes designed to assess model performance under both conventional machine learning and temporally realistic deployment settings. The first scheme employs Stratified 5-Fold Cross-Validation to evaluate generalization performance while preserving class distributions across folds. The second scheme adopts a Time-Based evaluation protocol that simulates real-world deployment by training models on historical data and testing them on future observations. The overall design of both evaluation schemes is presented in Fig. 3 and Fig. 4, respectively.

3.3.1. Stratified 5-Fold Cross-Validation

Stratified K-Fold Cross-Validation was employed to provide a robust and statistically reliable assessment of model performance. The primary objective of stratification is to preserve the original class distribution of the dataset within each fold, thereby reducing evaluation bias caused by class imbalance. The final dataset consists of four classes: app-positive (16.5%), app-negative (33.5%), app-neutral (14.6%), and service (35.5%). Maintaining these proportions across folds ensures that each training and testing cycle reflects the overall data distribution.

In this study, the number of folds was set to k=5. During each iteration, one fold was used as the testing set, while the remaining folds were utilized for model development. An internal partitioning procedure was subsequently applied to generate validation data for hyperparameter optimization and model selection. By repeating this process across all five folds, every instance contributes to both training and testing exactly once, providing a comprehensive estimate of model generalization performance. The detailed workflow and data allocation strategy for each fold are illustrated in Fig. 3.

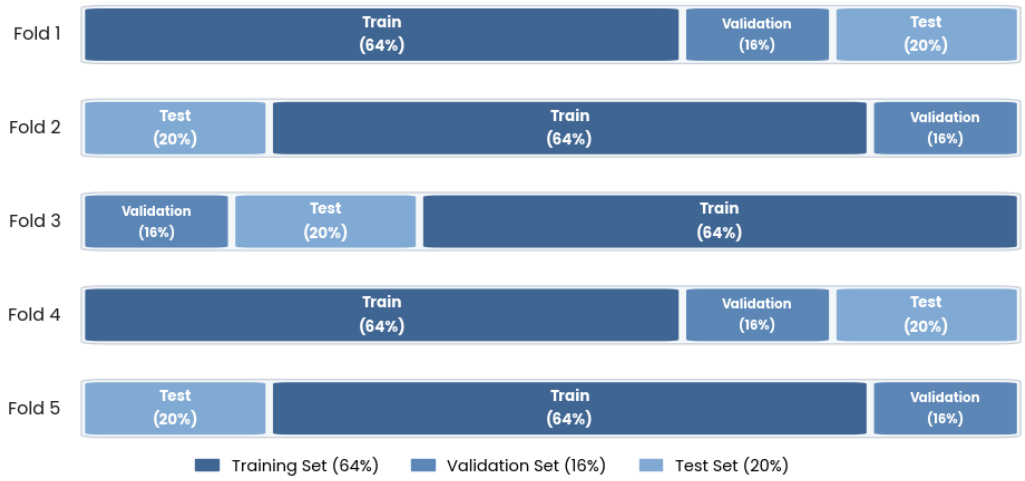

Figure 3. Visualization of the stratified 5-fold cross-validation

3.3.2. Time-Based Evaluation

In addition to conventional cross-validation, this study employs a Time-Based evaluation scheme to better reflect practical deployment conditions. Unlike random partitioning approaches, Time-Based evaluation preserves the chronological ordering of reviews, ensuring that models are trained exclusively on historical data and evaluated on future observations. This design prevents information leakage from future samples and provides a more realistic assessment of model robustness under temporal distribution shifts.

An expanding-window strategy was adopted across four evaluation iterations. In each iteration, the training set grows cumulatively by incorporating additional historical data, while the testing set advances to the subsequent time period. This setup enables the analysis of how increasing temporal coverage influences model performance. To support hyperparameter optimization and model selection, each training partition was further divided using a stratified 80:20 split while preserving the original class proportions. The detailed temporal allocation, sample distribution, and expanding-window structure are presented in Fig. 4.

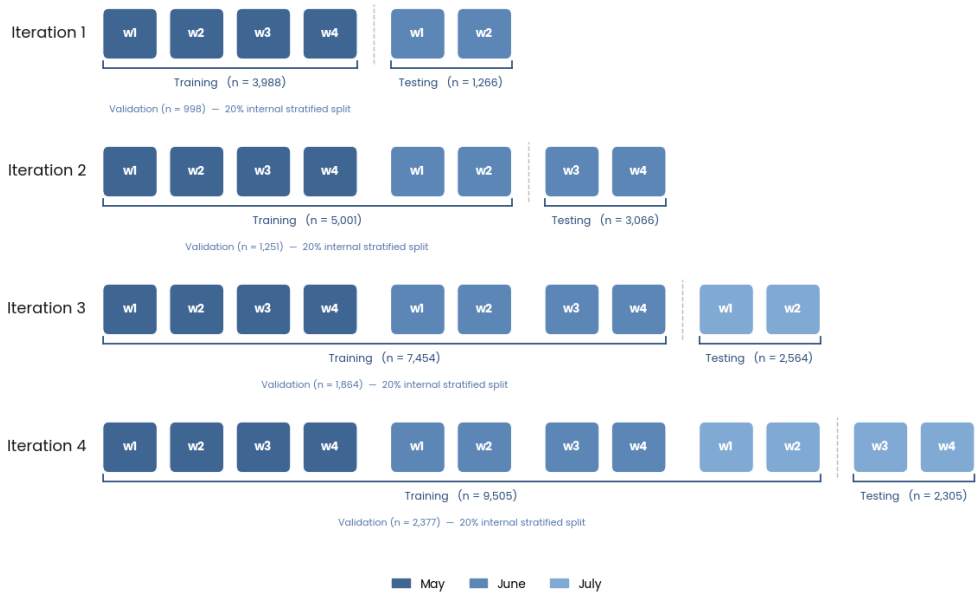

Figure 4. Visualization of the Time-Based Evaluation Scheme

3.4. Model Fine-Tuning

This study evaluates two Indonesian pre-trained transformer models: IndoBERTweet and IndoBERT. Both models are based on the BERT architecture and have been widely

adopted in Indonesian Natural Language Processing (NLP) tasks. However, they differ substantially in terms of pre-training data and domain coverage, making them suitable candidates for investigating the effect of domain-adaptive pre-training on sentence-level sentiment classification of app reviews.

The comparative design serves two purposes. First, it allows the effectiveness of IndoBERTweet on sentence-level app review classification to be systematically evaluated. Second, it enables an assessment of whether pre-training on informal Indonesian text provides measurable advantages over models trained predominantly on formal language resources. The overall model comparison framework is illustrated in Fig. 1.

3.4.1. IndoBERTweet

IndoBERTweet was selected as the primary model because it was specifically designed for informal Indonesian language processing. The model was pre-trained on approximately 26 million Indonesian-language tweets containing slang, abbreviations, emoticons, spelling variations, and other characteristics commonly found in user-generated content [12]. Through a domain-adaptive vocabulary initialization strategy, IndoBERTweet extends the lexical coverage of standard IndoBERT and provides stronger representations for informal language phenomena frequently encountered in social media and online reviews.

Given that Google Play Store reviews often contain colloquial expressions, acronyms, creative spellings, and informal writing styles, IndoBERTweet is expected to provide more effective contextual representations than models trained primarily on formal text sources [12]. Its suitability for informal Indonesian sentiment analysis has been demonstrated in several previous studies, motivating its selection as the primary model in this work.

3.4.2. IndoBERT

IndoBERT serves as the baseline model in this study. It is a monolingual Indonesian language model based on the BERT architecture and was pre-trained on a large corpus comprising Indonesian Wikipedia articles, major news portals such as Kompas, Tempo, and Liputan6, as well as the Indonesian Web Corpus [11], [13] In total, the pre-training corpus contains more than 220 million Indonesian words, enabling the model to learn rich contextual representations across a broad range of formal language domains.

Although IndoBERT has achieved strong performance across numerous Indonesian NLP tasks, its training data predominantly reflects formal language usage. Consequently, the model may be less effective when processing highly informal user-generated content containing slang, abbreviations, code-mixing, and non-standard orthographic variations. This characteristic makes IndoBERT an appropriate baseline for evaluating the contribution of domain-adaptive pre-training in the context of sentence-level sentiment analysis for Indonesian app reviews.

3.5. Evaluation Metrics

The performance of IndoBERTweet and IndoBERT was evaluated on the four-class sentence-level sentiment classification task consisting of app-positive, app-negative, app-neutral, and service categories. To provide a comprehensive assessment of model effectiveness, multiple evaluation metrics were employed based on the confusion matrix. The confusion matrix summarizes the relationship between actual and predicted class labels and enables detailed analysis of class-specific prediction behavior, particularly in the presence of class imbalance.

Let $T_{i,j}$ denote the number of instances whose true class is C_i and that are predicted as class C_j. Accordingly, diagonal elements ($T_{i,i}$) represent correctly classified instances, while off-diagonal elements correspond to misclassifications. Based on this notation, Accuracy, Precision, Recall, F1-Score, and Macro F1-Score were calculated to evaluate model performance from both overall and class-specific perspectives. Accuracy measures the proportion of correctly classified instances among all observations and provides an overall indication of classification performance:

$$Accuracy = \frac{\sum_{i=1}^{C} T_{i,i}}{\sum_{i=1}^{C} \sum_{j=1}^{C} T_{\mathrm{j,i}}} \tag{2}$$

where C denotes the number of classes. In this study, $C = 4$, corresponding to the app-positive, app-negative, app-neutral, and service categories.

Precision evaluates the reliability of positive predictions by measuring the proportion of correctly predicted instances within a given class:

$$Precision\ (C_i) = \frac{T_{i,i}}{\sum_{j=1}^{C} T_{\mathrm{j,i}}} \tag{3}$$

A high precision value indicates that most instances predicted as class C_i are correctly classified.

Recall measures the ability of the model to identify all instances belonging to a particular class:

$$Recall\ (C_i) = \frac{T_{i,i}}{\sum_{j=1}^{C} T_{\mathrm{i,j}}} \tag{3}$$

A high recall value indicates that the model successfully captures the majority of instances belonging to class C_i.

F1-Score represents the harmonic mean of Precision and Recall and is particularly useful when both false positives and false negatives are important:

$$F1\text{-}Score(C_i) = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Finally, Macro F1-Score is computed as the unweighted average of the F1-scores across all classes:

$$Macro\ F1\text{-}Score(C_i) = \frac{1}{C}\sum_{i=1}^{C} \mathrm{F1\text{-}Score}(C_i) \tag{5}$$

Macro F1-Score was selected as the primary evaluation metric in this study because it assigns equal importance to all classes regardless of their frequency in the dataset. This characteristic is particularly important given the imbalanced class distribution observed in the annotated corpus. Unlike accuracy, which may be dominated by majority classes, Macro F1-Score explicitly penalizes poor performance on minority classes and therefore provides a more balanced assessment of classification effectiveness. Consequently, it serves as the principal metric for comparing IndoBERTweet and IndoBERT across all experimental scenarios.

4. Results and Discussion

This section addresses two primary research questions. The first concerns the distribution of sentence-level sentiments in Gojek app reviews after conjunction-aware segmentation and annotation. The second examines the comparative performance of IndoBERTweet and IndoBERT under different evaluation schemes using Accuracy, Precision, Recall, and Macro F1-Score as performance indicators.

4.1. Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA RTX A4000 GPU with 16.1 GB of VRAM and 67.23 GB of system memory. Model training and inference were implemented using PyTorch 2.10.0 with CUDA acceleration, while the Hugging Face Transformers library was used for model fine-tuning. Hyperparameter optimization was performed using Optuna 4.8.0 with the Tree-structured Parzen Estimator (TPE) sampler [26], [27]. To ensure a fair comparison between models, both IndoBERTweet and IndoBERT were trained under a common experimental framework. The model configurations adopted throughout all experimental scenarios are summarized in Table 4.

The hyperparameter search space explored during optimization is presented in Table 5. These ranges were selected based on commonly adopted configurations for transformer fine-tuning and were applied consistently across all experimental scenarios to ensure comparability. For each experimental scenario, the final model was trained using the optimal hyperparameter configuration identified by Optuna. Across all runs, the training process completed the predefined maximum of 10 epochs without triggering the early-stopping criterion, indicating stable convergence behavior under the selected search space and optimization settings.

Table 4. Training configuration

Parameter	IndoBERTweet	IndoBERT
Model checkpoint	indolem/indobertweet-base-uncased	indobenchmark/indobert-base-p1
Transformers version	v5.6.2	v5.2.0
Optimizer	AdamW	AdamW
Gradient clipping	1.0	1.0
Maximum sequence length	128 tokens	128 tokens
Hyperparameter optimization	Optuna TPE, 10 trials	Optuna TPE, 10 trials

Table 5. Hyperparameter search space

Hyperparameter	Search Space
Learning rate	[2e-5, 5e-5]
Batch size	{8, 16, 32}
Dropout rate	[0.1, 0.3]
Weight decay	[0.005, 0.1]
Warmup ratio	[0.0, 0.2]
Maximum epochs	10
Early stopping patience	5

4.2. Experimental Results

4.2.1. Hyperparameter Optimization Results

To ensure a fair comparison across models and evaluation schemes, hyperparameter optimization was conducted independently for each experimental scenario. Four separate Optuna studies were performed corresponding to: (1) IndoBERTweet with Stratified 5-Fold Cross-Validation, (2) IndoBERT with Stratified 5-Fold Cross-Validation, (3) IndoBERTweet with Time-Based evaluation, and (4) IndoBERT with Time-Based evaluation. Independent optimization was necessary because the optimal configuration may depend on both the underlying model architecture and the data partitioning strategy. In particular, Time-Based evaluation introduces temporal distribution shifts between training and testing data, potentially requiring different optimization dynamics compared with stratified evaluation, where data distributions remain relatively stable across folds [28].

The search process explored the hyperparameter ranges presented in Table 5, and the best configuration identified for each scenario is summarized in Table 6. The results indicate that different evaluation schemes led to distinct optimal configurations, particularly with respect to batch size, dropout rate, and weight decay, highlighting the importance of scenario-specific hyperparameter tuning.

Table 6. Best Hyperparameter configuration for each experimental scenario

Scenario	Model	Evaluation Scheme	Learning Rate	Batch Size	Dropout Rate	Weight Decay
1	IndoBERTweet	Stratified 5-Fold	4.5e-5	32	0.188	0.012
2	IndoBERT	Stratified 5-Fold	3.3e-5	32	0.186	0.029
3	IndoBERTweet	Time-Based	4.7e-5	16	0.293	0.083
4	IndoBERT	Time-Based	3.1e-5	8	0.162	0.052

4.2.2. Performance under Stratified 5-Fold Cross-Validation

Tables 7 and 8 present the fold-level performance of IndoBERTweet and IndoBERT under Stratified 5-Fold Cross-Validation. Across all folds, IndoBERTweet consistently achieved higher performance than IndoBERT in terms of Accuracy, Macro F1-Score,

Precision, and Recall. The average Macro F1-Score of IndoBERTweet reached 0.729, exceeding IndoBERT by 0.015 points, while the average Accuracy improved from 0.740 to 0.751.

Table 7. Stratified 5-fold performance of IndoBERTweet (Scenario 1)

Fold	Accuracy	Macro F1-Score	Precision	Recall
Fold 1	0.750	0.730	0.748	0.721
Fold 2	0.763	0.746	0.749	0.746
Fold 3	0.744	0.719	0.723	0.718
Fold 4	0.751	0.725	0.733	0.722
Fold 5	0.749	0.723	0.726	0.724
Average	0.751	0.729	0.736	0.726
Std. Dev.	0.007	0.010	0.012	0.011

Table 8. Stratified 5-fold performance of IndoBERT (Scenario 2)

Fold	Accuracy	Macro F1-Score	Precision	Recall
Fold 1	0.735	0.717	0.717	0.717
Fold 2	0.745	0.716	0.738	0.709
Fold 3	0.740	0.711	0.711	0.713
Fold 4	0.744	0.719	0.721	0.721
Fold 5	0.736	0.709	0.713	0.708
Average	0.740	0.714	0.720	0.714
Std. Dev.	0.004	0.004	0.010	0.005

To further investigate model behavior, Table 9 reports class-wise F1-scores averaged across all folds. The largest performance difference was observed in the app-neutral category, where IndoBERTweet achieved an F1-score of 0.533 compared with 0.496 for IndoBERT. In contrast, the gap was relatively small for app-positive, app-negative, and service classes.

Table 9. Per-class F1-score comparison

Class	IndoBERTweet	IndoBERT
App-Positive	0.839	0.837
App-Negative	0.772	0.764
App-Neutral	0.533	0.496
Service	0.772	0.760
Average	0.729	0.714

The observed advantage of IndoBERTweet is consistent with its pre-training on large-scale informal Indonesian text [12]. Informal user reviews frequently contain colloquial expressions, abbreviations, spelling variations, and conversational constructions that are less prevalent in the formal corpora used to train IndoBERT [13]. As a result, domain-adaptive pre-training appears particularly beneficial for categories that rely on subtle contextual interpretation rather than explicit sentiment-bearing words. Nevertheless, both models achieved comparable performance in the app-positive class, suggesting that strongly positive expressions are sufficiently represented across both formal and informal pre-training corpora.

Another notable observation concerns model stability. IndoBERT exhibited lower inter-fold variance (Macro F1 standard deviation: ±0.004) than IndoBERTweet (±0.010), indicating more consistent performance across folds. However, this stability came at the cost of a lower overall performance ceiling. Collectively, the results demonstrate that IndoBERTweet provides a consistent performance advantage under stratified evaluation while maintaining acceptable robustness across folds.

4.2.3. Performance under Time-Based Evaluation

The results of the Time-Based evaluation are presented in Tables 10 and 11. Unlike the stratified setting, the temporal evaluation protocol introduces chronological separation

between training and testing data, thereby providing a more realistic assessment of deployment performance.

Table 10. Time-Based performance of IndoBERTweet (Scenario 3)

Iteration	Accuracy	Macro F1-Score	Precision	Recall
Iteration 1	0.689	0.670	0.680	0.704
Iteration 2	0.711	0.685	0.693	0.711
Iteration 3	0.738	0.710	0.714	0.724
Iteration 4	0.773	0.739	0.748	0.736
Average	0.728	0.701	0.709	0.719
Std. Dev.	0.036	0.030	0.029	0.014

Table 11. Time-Based performance of IndoBERT (Scenario 4)

Iteration	Accuracy	Macro F1-Score	Precision	Recall
Iteration 1	0.707	0.682	0.693	0.708
Iteration 2	0.697	0.674	0.683	0.691
Iteration 3	0.730	0.709	0.703	0.722
Iteration 4	0.766	0.737	0.744	0.734
Average	0.725	0.701	0.706	0.714
Std. Dev.	0.030	0.028	0.026	0.018

Both models exhibited a clear upward trend as the training window expanded. IndoBERTweet improved from a Macro F1-Score of 0.670 in Iteration 1 to 0.739 in Iteration 4, while IndoBERT improved from 0.682 to 0.737. This pattern suggests that increased temporal coverage and training volume contribute substantially to model performance, enabling better adaptation to evolving linguistic patterns and user expressions.

A notable exception occurred in Iteration 1, where IndoBERT slightly outperformed IndoBERTweet (0.682 vs. 0.670). This result suggests that under limited training data conditions, the broader and more formal linguistic representations learned by IndoBERT may generalize more effectively than the highly domain-specific representations of IndoBERTweet. However, as additional historical data became available, IndoBERTweet consistently recovered and ultimately achieved the highest performance in the final iteration. By Iteration 4, both models achieved nearly identical Macro F1-Scores (0.739 vs. 0.737), indicating that the performance gap narrows as training data volume increases. Nevertheless, IndoBERTweet maintained a small but consistent advantage, suggesting that domain-adaptive pre-training continues to provide benefits even under larger training regimes.

4.2.4. Overall Comparison

To provide a consolidated view of model performance across all experimental settings, Table 12 summarizes the average results obtained from the four scenarios. For Stratified 5-Fold Cross-Validation, the reported values correspond to the mean performance across five folds. For Time-Based evaluation, the reported values represent the unweighted average across four temporal iterations.

Table 12. Overall performance comparison across all scenarios

Scenario	Model	Evaluation Scheme	Accuracy	Macro F1-Score	Precision	Recall
1	IndoBERTweet	Stratified 5-Fold	0.751	0.729	0.736	0.726
2	IndoBERT	Stratified 5-Fold	0.740	0.714	0.720	0.714
3	IndoBERTweet	Time-Based	0.728	0.701	0.709	0.719
4	IndoBERT	Time-Based	0.725	0.701	0.706	0.714

Across all scenarios, IndoBERTweet achieved the highest overall performance, with the best results obtained under Stratified 5-Fold Cross-Validation (Accuracy = 0.751; Macro F1-

Score = 0.729). Although both models obtained the same average Macro F1-Score under Time-Based evaluation (0.701), IndoBERTweet consistently achieved slightly higher Accuracy, Precision, and Recall. Overall, these findings indicate that domain-adaptive pre-training on informal Indonesian text provides measurable benefits for sentence-level sentiment classification, particularly under evaluation settings where class discrimination is more challenging.

4.2.5. Confusion Matrix Analysis

To better understand model behavior beyond aggregate performance metrics, an error analysis was conducted using the aggregated confusion matrix of Scenario 1 (IndoBERTweet under Stratified 5-Fold Cross-Validation). The matrix was generated by accumulating predictions across all five folds, resulting in a total of 14,187 evaluated sentence-level instances. This aggregation covers the entire annotated dataset and provides a comprehensive view of class-specific prediction patterns. The resulting confusion matrix is shown in Fig. 5.

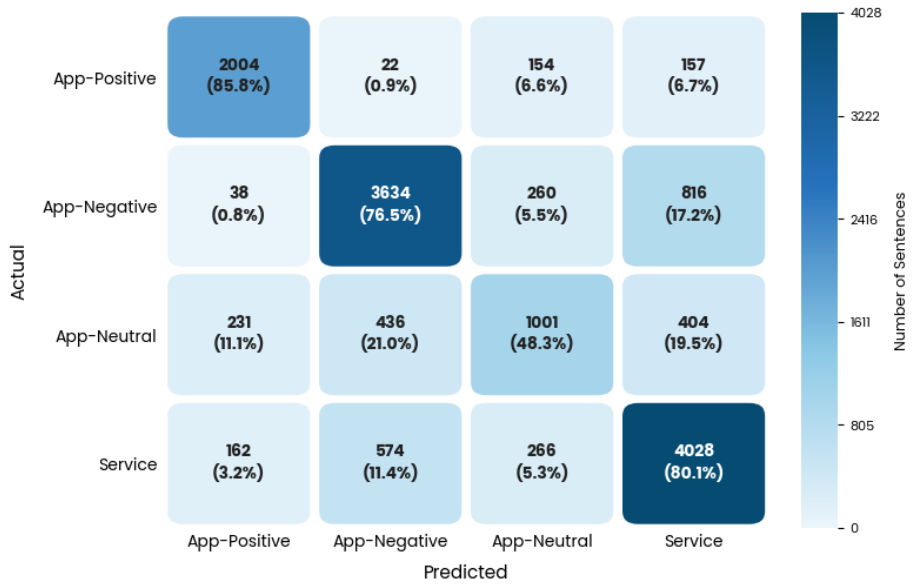

Figure 5. Aggregated confusion matrix for scenario 1

The most prominent source of classification error is associated with the app-neutral class. Although this category captures factual or descriptive statements about the application, many instances contain complaint-adjacent language that lacks explicit sentiment markers. As a result, a substantial proportion of app-neutral sentences are misclassified as app-negative or service-related statements. This observation is consistent with the relatively low recall obtained for the app-neutral category and highlights the inherent difficulty of identifying neutrality in informal user-generated content, where sentiment is often conveyed implicitly rather than through explicit affective expressions.

A second source of error arises from the bidirectional confusion between app-negative and service categories. In many reviews, users describe service-related failures using application-oriented language, thereby blurring the distinction between software functionality and operational service quality. For example, statements such as "*aplikasi nggak bisa nelpon driver*" ("the app cannot call the driver") simultaneously reference the application and a service interaction. Consequently, the model may struggle to determine whether the primary target of the complaint is the application itself or the associated service component.

In contrast, the app-positive category achieved the highest recall among all classes. Positive reviews frequently contain explicit sentiment indicators such as bagus (good), cepat (fast), mudah (easy), and mantap (excellent), which provide strong lexical cues for classification. The remaining misclassifications are primarily associated with reviews praising service personnel rather than the application, reflecting the conceptual overlap between positive application experiences and positive service experiences. Overall, the confusion matrix analysis suggests that the primary challenge in this task is not the identification of positive or negative sentiment itself, but rather the discrimination of neutral statements and the separation of app-directed

and service-directed opinions. These findings provide valuable guidance for future model improvements and annotation refinements.

4.2.6. Ablation Study: Impact of Emoji Normalization

To quantify the contribution of emoji normalization within the proposed preprocessing pipeline, an ablation study was conducted by removing the emoji normalization step while preserving all other preprocessing components, including lowercasing, URL and username removal, and conjunction-aware segmentation. To ensure a fair comparison, the optimal hyperparameter configurations identified in Section 4.2.1 were retained without further tuning. Consequently, any observed performance differences can be attributed directly to the presence or absence of emoji normalization.

A parallel set of experiments was performed for all four evaluation scenarios. Table 13 summarizes the resulting Macro F1-Scores with and without emoji normalization. Across all scenarios, the complete preprocessing pipeline consistently outperformed its ablated counterpart, demonstrating that emoji normalization provides a measurable contribution to classification performance.

Table 13. Macro F1-score comparison with and without emoji normalization

Model	Evaluation Scheme	With Emoji Normalization	Without Emoji Normalization
IndoBERTweet	Stratified 5-Fold	0.729	0.723
IndoBERTweet	Time-Based	0.701	0.699
IndoBERT	Stratified 5-Fold	0.714	0.708
IndoBERT	Time-Based	0.701	0.686

The observed performance gains range from +0.002 to +0.015 Macro F1 points, with the largest improvement occurring for IndoBERT under Time-Based evaluation. Although the magnitude of these improvements is modest, the consistency of the gains across all experimental settings indicates that emoji normalization contributes useful sentiment-related information that would otherwise be lost during preprocessing.

A class-level analysis further reveals that the app-neutral category benefits most from emoji normalization. Removing this step reduces the app-neutral F1-score from 0.533 to 0.514 for IndoBERTweet and from 0.496 to 0.487 for IndoBERT. This finding is consistent with the role of emojis as non-verbal sentiment indicators in informal Indonesian reviews. By converting emoji symbols into textual representations, the models gain access to semantic cues that help distinguish neutral expressions from mildly negative statements. Without normalization, these cues remain inaccessible or are only partially represented through subword tokenization.

Interestingly, the performance gain is more pronounced for IndoBERT than for IndoBERTweet. This result is consistent with the different pre-training corpora used by the two models. Because IndoBERTweet was pre-trained on large-scale Twitter data that naturally contains abundant emoji usage, it appears to possess a degree of inherent robustness to emoji-related language patterns. In contrast, IndoBERT relies primarily on formal text sources and therefore benefits more substantially from explicit emoji normalization. These findings experimentally justify the inclusion of emoji normalization in the proposed preprocessing pipeline and demonstrate that seemingly simple preprocessing choices can have measurable downstream effects, particularly for sentiment categories that rely on subtle contextual signals.

5. Conclusions

This study proposed a sentence-level sentiment analysis framework for Indonesian app reviews based on conjunction-aware review segmentation and transformer-based language models. By decomposing multi-opinion reviews into independently classifiable sentence units, the proposed framework addresses a key limitation of conventional document-level sentiment analysis, namely the loss of opinion granularity in reviews containing multiple sentiment targets.

Experimental results demonstrate that IndoBERTweet consistently outperformed IndoBERT under Stratified 5-Fold Cross-Validation, achieving an accuracy of 0.751 and a

Macro F1-Score of 0.729. The largest performance improvement was observed in the app-neutral category, suggesting that domain-adaptive pre-training on informal Indonesian text provides advantages when modeling subtle and implicit opinion expressions. Under Time-Based evaluation, both models achieved the same average Macro F1-Score of 0.701, while IndoBERTweet maintained a slight advantage in overall predictive performance. In addition, both models exhibited improved performance as the temporal training window expanded, indicating the importance of historical data coverage for robust sentiment classification.

The ablation study further confirmed the effectiveness of the proposed preprocessing pipeline. Emoji normalization consistently improved performance across all evaluation scenarios, with gains ranging from +0.002 to +0.015 Macro F1 points. The results highlight the importance of preserving non-verbal sentiment cues when processing informal user-generated content and demonstrate that preprocessing decisions can contribute meaningfully to downstream classification performance.

Despite these encouraging results, several challenges remain. First, the app-neutral category continues to be the most difficult class to identify, reflecting the inherent ambiguity of neutral sentiment expressions in informal Indonesian reviews. Second, confusion between app-directed and service-directed complaints indicates the presence of referential ambiguity that cannot always be resolved through flat multi-class classification. Third, the current study assumes that conjunction-aware segmentation produces semantically valid sentence boundaries; however, segmentation quality was not evaluated independently and therefore warrants further investigation. Future work may explore hierarchical classification architectures, targeted data augmentation strategies, contrastive pre-training approaches, and cross-platform validation on other Indonesian super-app ecosystems such as Grab, Tokopedia, and Shopee to assess the generalizability of the proposed framework.

Author Contributions: Conceptualization, I.N. Aqiilah and R. Saptono; Methodology, R. Saptono and A. Syaifuddin; Software, I.N. Aqiilah; Validation, I.N. Aqiilah and R. Saptono; Formal Analysis, I.N. Aqiilah; Investigation, I.N. Aqiilah; Resources, I.N. Aqiilah; Data Curation, I.N. Aqiilah; Writing – Original Draft Preparation, I.N. Aqiilah; Writing – Review and Editing, R. Saptono and A. Syaifuddin; Visualization, I.N. Aqiilah; Supervision, R. Saptono; Project Administration, I.N. Aqiilah. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Hibah Penguatan Kapasitas Grup Riset (PKGR-UNS) C, Universitas Sebelas Maret, under contract number 462/UN27.22/PT.01.03/2026.

Data Availability Statement: The raw dataset comprising user reviews of the Gojek application scraped from Google Play Store (May–July 2025) is publicly available at https://github.com/thisntinge/raw-gojek-reviews-dataset.git. The dataset includes review text, ratings, dates, and metadata. The annotated sentiment labels are not publicly shared due to the substantial manual annotation effort involved.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] S. Wahyu Handani, D. Intan Surya Saputra, Hasirun, R. Mega Arino, and G. Fiza Asyrofi Ramadhan, "Sentiment Analysis for Go-Jek on Google Play Store," *J. Phys. Conf. Ser.*, vol. 1196, p. 012032, Mar. 2019, doi: 10.1088/1742-6596/1196/1/012032.

[2] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken DW, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," in *6th International Conference on Sustainable Information Engineering and Technology 2021*, Sep. 2021, pp. 258–264. doi: 10.1145/3479645.3479679.

[3] I. H. Setiawan, M. Rahardi, A. Aminuddin, and F. F. Abdulloh, "Sentiment Analysis of Tokopedia Application Reviews on Google Play Store Using BERT," in *2024 International Conference on Information Technology Systems and Innovation (ICITSI)*, Dec. 2024, pp. 242–247. doi: 10.1109/ICITSI65188.2024.10929357.

[4] M. Hanafi, S. Adi, and A. Setiawan, "A Model of Sentiment Analysis on Gojek Application Review using Word Vector Representation and Long Short-Term Memory (LSTM)," in *2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*, Jan. 2025, pp. 944–949. doi: 10.1109/ICoCSETI63724.2025.11019757.

[5] A. P. Singh, A. Singh, A. Prakash, A. Kumar, and Vikas, "Sentiment Analysis on Play Store Application Reviews Using BERT Model," in *2025 International Conference on Networks and Cryptology (NETCRYPT)*, May 2025, pp. 495–499. doi: 10.1109/NETCRYPT65877.2025.11102489.

[6] R. Budianoor, S. W. Saputro, F. Abadi, R. A. Nugroho, and A. Farmadi, "Quantifying the Impact of Text Preprocessing on IndoBERT Fine-Tuning for Indonesian Informal Culinary Sentiment Analysis," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 564–581, May 2026, doi: 10.62411/jcta.15980.

[7] S. Mifrah and E. H. Benlahmar, "Sentence-Level Sentiment Classification A Comparative Study Between Deep Learning Models," *J. ICT Stand.*, May 2022, doi: 10.13052/jicts2245-800X.10213.

[8] M. R. Pribadi, D. Manongga, H. D. Purnomo, I. Setyawan, and Hendry, "Sentiment Analysis of the PeduliLindungi on Google Play using the Random Forest Algorithm with SMOTE," in *2022 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Jul. 2022, pp. 115–119. doi: 10.1109/ISITIA56226.2022.9855372.

[9] X. Li, X. Sun, Z. Xu, and Y. Zhou, "Explainable Sentence-Level Sentiment Analysis for Amazon Product Reviews," in *2021 5th International Conference on Imaging, Signal Processing and Communications (ICISPC)*, Jul. 2021, pp. 88–94. doi: 10.1109/ICISPC53419.2021.00024.

[10] D. R. I. M. Setiadi, D. Marutho, and N. A. Setiyanto, "Comprehensive Exploration of Machine and Deep Learning Classification Methods for Aspect-Based Sentiment Analysis with Latent Dirichlet Allocation Topic Modeling," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 12–22, May 2024, doi: 10.62411/faith.2024-3.

[11] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.

[12] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.

[13] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *arXiv*, Nov. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[14] A. F. Hidayatullah, K. Kalinaki, M. M. Aslam, R. Y. Zakari, and W. Shafik, "Fine-Tuning BERT-Based Models for Negative Content Identification on Indonesian Tweets," in *2023 8th International Conference on Information Technology and Digital Applications (ICITDA)*, Nov. 2023, pp. 1–6. doi: 10.1109/ICITDA60835.2023.10427046.

[15] A. P. Kirani, R. Saptono, and R. Anggrainingsih, "Which Features Matter Most? Evaluating Numerical and Textual Features for Helpfulness Classification in Imbalance Dataset using XGBoost," *Sci. J. Informatics*, vol. 12, no. 4, pp. 731–742, Nov. 2025, doi: 10.15294/sji.v12i4.33443.

[16] R. Saptono and T. Mine, "Time-based Sampling Methods for Detecting Helpful Reviews," in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Dec. 2020, pp. 508–513. doi: 10.1109/WIIAT50758.2020.00076.

[17] I. P. Windasari, F. N. Uzzi, and K. I. Satoto, "Sentiment analysis on Twitter posts: An analysis of positive or negative opinion on GoJek," in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Oct. 2017, pp. 266–269. doi: 10.1109/ICITACEE.2017.8257715.

[18] C.-H. Du, M.-F. Tsai, and C.-J. Wang, "Beyond Word-level to Sentence-level Sentiment Analysis for Financial Reports," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 1562–1566. doi: 10.1109/ICASSP.2019.8683085.

[19] D. Sebastian, H. D. Purnomo, and I. Sembiring, "BERT for Natural Language Processing in Bahasa Indonesia," in *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, Dec. 2022, pp. 204–209. doi: 10.1109/ICICyTA57421.2022.10038230.

[20] B. R. P. Darnoto and D. B. Firmawan, "Language-Similarity-Guided Transfer Fine-Tuning of Pre-trained Transformer Models for Sentiment Analysis Across 12 Indonesian Regional Languages," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 547–563, May 2026, doi: 10.62411/jcta.15975.

[21] E. Supriyadi and P. N. Makatita, "Sentiment Analysis of TikTok User Comments on QRIS Adoption in Indonesia Using IndoBERT," *Procedia Comput. Sci.*, vol. 269, pp. 121–130, 2025, doi: 10.1016/j.procs.2025.08.265.

[22] D. R. I. M. Setiadi, A. R. Muslikh, S. W. Irianandra, W. Warto, J. Gondohanindijo, and A. A. Ojugo, "Outlier Detection Using Gaussian Mixture Model Clustering to Optimize XGBoost for Credit Approval Prediction," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 244–255, Nov. 2024, doi: 10.62411/jcta.11638.

[23] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[24] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, Nov. 1971, doi: 10.1037/h0031619.

[25] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.

[26] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[27] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[28] P. Röttger and J. Pierrehumbert, "Temporal Adaptation of BERT and Performance on Downstream Document Classification: Insights from Social Media," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2400–2412. doi: 10.18653/v1/2021.findings-emnlp.206.