Vol. 7, No. 4, December 2025

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Sentiment Analysis of User Reviews for the PLN Mobile Application Using Naïve Bayes and Long Short-Term Memory

Jose Mario Ayomi[1], Anik Vega Vitianingsih[2*], Yudi Kristyawan[3], Anastasia Lidya Maukar[4], Tjatursari Widiartin[5]

[1,2,3]Informatics Department, Universitas Dr. Soetomo, Surabaya, Indonesia

[4]Industrial Engineering Department, President University, Bekasi, Indonesia

[5]Informatics Department, Universitas Wijaya Kusuma, Surabaya, Indonesia

Email: joseayomi00@gmail.com, [2]vega@unitomo.ac.id*, [3]yudi.kristyawan@unitomo.ac.id, [4]almaukar@president.ac.id, [5]widiartin@gmail.com

Received: October 15, 2025

Revised: Nov 13, 2025

Accepted: Nov 21, 2025

Published: Dec 13, 2025

Corresponding Author:

Author Name*:

Anik Vega Vitianingsih

Email*:

vega@unitomo.ac.id

DOI:

10.63158/journalisi.v7i4.1342

Abstract. This study explores large-scale sentiment analysis of user reviews for the PLN Mobile application to better understand public perception and provide quantitative insights for improving digital electricity services in Indonesia. Addressing the lack of benchmarks for Indonesian public-service apps—where prior studies rely on smaller datasets and traditional machine learning—this research positions sentiment analysis as a tool for continuous user experience monitoring. A total of 50,000 Indonesian-language reviews from Google Play were collected and pre-processed using cleaning, case folding, tokenization, stopword removal, normalization, and stemming. Sentiments (positive, neutral, negative) were assigned using a domain-specific Indonesian sentiment lexicon, yielding approximately 40% positive, 35% neutral, and 25% negative labels. Two models were applied: Multinomial Naïve Bayes using TF-IDF features and a Long Short-Term Memory (LSTM) model with 100-dimensional word embeddings and a 128-unit LSTM layer. Naïve Bayes achieved 70.89% accuracy (F1-score: 0.6964), while LSTM outperformed it with 98.02% accuracy (F1-score: 0.9800). These results highlight the superiority of deep learning in sentiment monitoring and offer a scalable framework to help PLN and policymakers enhance digital public service delivery.

Keywords: Sentiment Analysis, Natural Language Processing, LSTM, PLN Mobile, Public Service Applications

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

The rapid expansion of digital technology has encouraged governments and utilities to deliver public services through mobile applications. In the electricity sector, PT Perusahaan Listrik Negara (PLN) launched the PLN Mobile application to centralize billing information, meter readings, outage reporting, and complaint submission in a single platform for customers across Indonesia [1]. As a flagship digital public-service channel, PLN Mobile is expected to enhance transparency and operational efficiency, yet increasing adoption has been accompanied by frequent complaints on the Google Play Store regarding login failures, slow service handling, payment errors, and unstable performance that may undermine user trust [2], [3]. Conventional complaint mechanisms are insufficient to capture this continuous stream of experience-based opinions, whereas user reviews provide a high-frequency source of evidence on perceived quality and usability of PLN Mobile as a socio-technical system [4]. Sentiment analysis offers a computational means to convert these unstructured reviews into structured negative, neutral, and positive categories, enabling PLN to monitor satisfaction trends, detect emerging service problems, and support data-driven interventions to improve digital electricity services at scale [5], [4], [6].

Several studies have investigated sentiment analysis of PLN Mobile user reviews using classical machine-learning algorithms. The work in [2] analyzed 3,000 reviews labeled into 2,099 positive (69.97%), 368 neutral (12.27%), and 533 negative (17.77%) and compared Naïve Bayes Classifier (NBC) with K-Nearest Neighbor (KNN); NBC achieved 77.69% accuracy, 53.14% recall, 59.84% precision, and 54.09% F1-score, outperforming KNN, which obtained 76.40% accuracy, 49.64% recall, 56.84% precision, and 50.67% F1-score. In [1], a Multinomial Naïve Bayes model trained on PLN Mobile reviews with an 80:20 split attained 76% accuracy, 76% precision, 100% recall, and 86% F1-score, indicating high sensitivity but only moderate correctness. The study in [3] applied Support Vector Machine (SVM) with TF–IDF features and reported accuracy above 90%, whereas a lexicon-based VADER labeling combined with Multinomial Naïve Bayes in [7] yielded 70% accuracy under a 90:10 scheme. More recently, [8] employed the Decision Tree algorithm to classify PLN Mobile reviews into three sentiment classes and achieved 96% accuracy, with precision of 91%, recall of 96%, and F1-score of 93%. Collectively, these PLN Mobile studies show that classical algorithms can attain accuracies between 70% and 96% depending on dataset size,

labeling strategy, and feature engineering, but they all rely on small corpora and sparse, order-insensitive representations.

Despite these encouraging results, several methodological and practical gaps remain in the existing PLN Mobile sentiment-analysis literature. First, all prior PLN Mobile studies operate on comparatively small datasets, typically involving only a few thousand reviews, which limits their ability to capture the full diversity of linguistic expressions, evolving user issues, and temporal dynamics present in the much larger pool of Google Play feedback; as PLN Mobile's user base grows, models calibrated on such restricted corpora may not generalize well [1], [2], [3], [7], [8] . Second, the employed algorithms—Naïve Bayes, KNN, SVM, and Decision Tree—are grounded in bag-of-words or TF–IDF feature spaces that treat tokens as independent, thereby ignoring word order, negation structures, and long-range contextual dependencies that are critical for accurately interpreting Indonesian-language sentiment, especially in the presence of affixation, informal spelling, and mixed registers [1], [2], [3], [7], [8]. Third, although [7] demonstrates lexicon-based automatic labeling with VADER combined with Naïve Bayes, no PLN Mobile study has yet integrated a large-scale, Indonesian-domain lexicon-based labeling approach with a deep-learning architecture such as Long Short-Term Memory, nor conducted a systematic, in-domain comparison between a probabilistic baseline and an LSTM model on the same multi-class dataset. Finally, existing works tend to emphasize classification accuracy rather than framing sentiment analysis as a comprehensive, scalable framework that links large-scale user opinion mining to the continuous monitoring and improvement of PLN Mobile as a critical digital public-service platform [1], [2], [3], [7], [8], [9], [10], [11], [12], [13].

In response to these gaps, the purposes of this research are:

1) To develop a large-scale sentiment-analysis framework for PLN Mobile that processes 50,000 Indonesian-language user reviews from the Google Play Store and automatically assigns negative, neutral, and positive labels using an extended Indonesian sentiment lexicon adapted to the PLN Mobile domain.
2) To rigorously evaluate and compare the performance of Multinomial Naïve Bayes and Long Short-Term Memory on the same three-class PLN Mobile dataset using accuracy, precision, recall, and macro F1-score, thereby quantifying the performance gains of sequence-aware deep learning over a probabilistic baseline.

3) To provide PLN with a practical analytical instrument for continuously monitoring user satisfaction and recurrent service issues reflected in user sentiment, so that evidence from large-scale reviews can inform decisions to improve the quality, reliability, and responsiveness of its digital electricity services.

The novelty of this research lies in its integration of methodological, data-scale, and application-level contributions within a single coherent framework for PLN Mobile. First, to the best of the authors' knowledge, this is the first study to apply a Long Short-Term Memory architecture to PLN Mobile user reviews and to benchmark it directly against a Multinomial Naïve Bayes classifier on the *same* three-class (negative, neutral, positive) dataset, thereby providing a rigorous, in-domain comparison between a probabilistic baseline and a sequence-aware deep-learning model. Second, the study operates on a lexicon-labeled corpus of 50,000 Indonesian-language reviews, substantially larger and more heterogeneous than the datasets used in previous PLN Mobile works, and employs an extended Indonesian sentiment lexicon that is explicitly adapted to the PLN Mobile domain, enabling scalable, consistent labeling without sacrificing domain relevance. Third, the proposed framework does not treat sentiment analysis as a purely academic classification task but explicitly positions it as a decision-support instrument for a national digital public-service platform, translating model outputs into interpretable patterns of user satisfaction and recurrent service issues that can inform PLN's strategic and operational efforts to monitor, evaluate, and continuously improve the quality of its digital electricity services.

2. METHODS

This research implements a comparative sentiment analysis framework that combines classical and deep-learning approaches to evaluate user perceptions of the PLN Mobile application.

2.1. Research Framework

As illustrated in Figure 1, the workflow comprises multiple stages: data collection, text preprocessing, lexicon-based sentiment labeling, feature representation, model training, and performance evaluation.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

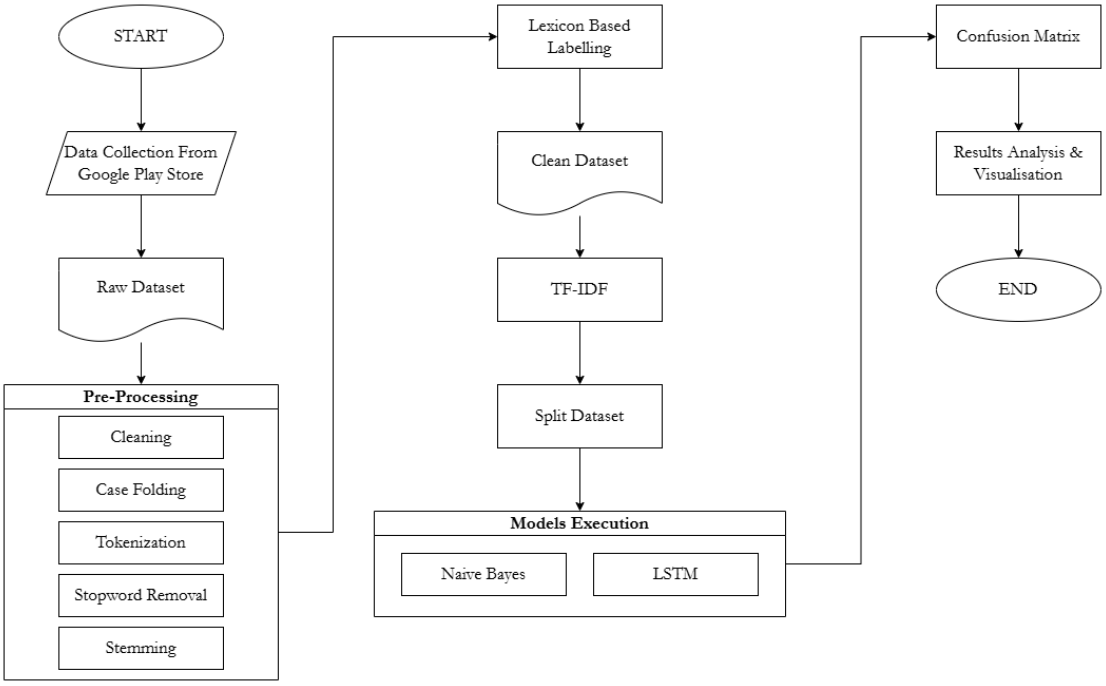

Figure 1. Research Flow Diagram

The framework consists of six sequential stages: data acquisition, text pre-processing, lexicon-based labelling, feature representation, model training, and performance evaluation, each grounded in natural language processing and text-mining theory for sentiment analysis [5], [4], [6]. User feedback were collected from the Google Play Store and processed through normalization, tokenization, stop-word elimination, slang handling, and stemming using the Sastrawi library to reduce lexical diversity, following established preprocessing conventions for Indonesian text analysis in application and product-review datasets [2], [1], [5], [14]. Sentiment labels were generated automatically using a lexicon adapted to PLN-specific terminology to produce three polarity classes, consistent with lexicon-based sentiment studies on Indonesian reviews and opinion data [7], [15], [16]. Text representation was then divided into two paradigms: TF–IDF weighting for statistical modeling and word embeddings for contextual modeling, in line with prior work that contrasts sparse term-weighting with dense vector representations for sentiment tasks [9], [10], [17], [6]. The Naïve Bayes classifier was used as the probabilistic baseline, reflecting its extensive use and documented performance in Indonesian sentiment analysis across various domains [2], [1], [5], [4], [18], whereas the Long Short-Term Memory model was used as the sequential deep-learning model designed to capture temporal dependencies within text, as demonstrated in several sentiment-analysis applications [9], [10], [11], [12]. Both

models were developed and evaluated using an 80:20 data partition, assessed through accuracy, precision, recall, and F1-score metrics commonly reported in sentiment-analysis evaluations [5], [9], [10], [19], [20], [21].

2.2. Data Collection

User-generated reviews concerning the PLN Mobile application were sourced from the Google Play Store using a Python-based web-scraping script that extracted review text, star ratings, and basic metadata in accordance with publicly available data-access policies [2], [1], [5], [22]. The scraper retrieved reviews in reverse chronological order, starting from the most recent entries at the time of data collection, until a total of 50,000 reviews was obtained; this procedure ensures that the corpus reflects the latest publicly available user feedback on PLN Mobile without manual date filtering [2], [3], [8], [22]. To improve data quality, exact duplicate reviews were removed, and entries dominated by non-linguistic or spam-like content—such as pure advertisements, phone numbers, or isolated URLs—were filtered out; reviews that contained no meaningful alphabetic tokens after pre-processing were discarded as extremely short or uninformative, following common practices in sentiment-analysis data preparation [5], [4], [6]. The Google Play Store was selected as the primary data source because it provides structured and large-scale review data that align with best practices in opinion-mining research on mobile applications [5], [22], [6], [20], [23]. Star ratings (1–5) were retained for descriptive purposes but were not used directly as sentiment labels; instead, the construction of the three sentiment classes relied on the lexicon-based procedure described in Section 2.4 [7], [15], [16]

2.3. Text Pre-Processing

Text pre-processing was conducted to standardize and clean the raw review data before feature extraction, because noise reduction, lexical normalization, and morphological simplification directly influence sentiment-classification accuracy and interpretability [5], [4], [6]. The pipeline consisted of case folding (converting all characters to lowercase), tokenization (segmenting each review into lexical units), stopword removal (eliminating high-frequency function words with minimal semantic contribution), and normalization. The normalization step corrected common slang and informal spellings using an Indonesian slang dictionary compiled and adapted from previous sentiment-analysis studies and PLN-related expressions, and collapsed repeated characters (for example,

"*baguuus*" → "*bagus*") while standardizing diacritics to maintain orthographic consistency [5], [17], [14], [6]. Non-alphabetic symbols, hyperlinks, digits, emojis, and redundant punctuation were removed rather than mapped to sentiment categories to reduce non-textual noise [2], [1], [23]. Finally, stemming was applied using the Indonesian Sastrawi library to convert derived and inflected terms into their base morphological forms, in line with established practices in Indonesian text-mining research [1], [5], [14].

2.4. Lexicon-Based Labeling

Lexicon-based labeling was employed to automatically assign sentiment polarity to each review based on the semantic orientation of its constituent words. This approach utilizes a predefined sentiment lexicon, in which each token carries a polarity weight that contributes either positively or negatively to the overall sentiment score [7], [15], [16]. In this study, the Indonesian InSet Lexicon was adopted and expanded with domain-specific terminology associated with electricity services, such as "*tagihan*" (billing), "*pemadaman*" (outage), and "*pelayanan*" (service), to enhance contextual precision [15], [16]. This modification ensured that sentiment orientation reflected user-specific experiences within the PLN Mobile application domain. Each tokenized review was processed by mapping its words to corresponding lexicon entries, aggregating their polarity scores to determine the overall sentiment strength and direction.

$$\text{Sentiment\_Score } (S^i) = \sum_{i=1}^{n} S_i \quad (1)$$

Equation (1) illustrates that a review's overall sentiment value was calculated by aggregating the polarity scores of each detected term. Positive words contributed to higher cumulative scores, while negative words decreased the total; neutral or unrecognized words had no effect. Reviews yielding positive totals were classified as positive, those with zero totals as neutral, and those with negative totals as negative. The resulting lexicon-labeled corpus contained approximately 40% positive, 35% neutral, and 25% negative reviews, reflecting a moderately imbalanced three-class distribution that later motivated the use of macro-averaged and weighted evaluation metrics in the model-comparison stage [5], [19], [18].

2.5. Feature Representation

Different feature-representation methods were applied to enhance model effectiveness. For the Multinomial Naïve Bayes classifier, text data were represented using the TF–IDF

scheme, where each document is transformed into a weighted vector whose term weights reflect their importance within a review relative to their occurrence across the entire corpus [5], [4], [17], [6]. This statistical representation reduces the dominance of very high-frequency but semantically weak words while emphasizing discriminative tokens that contribute more strongly to sentiment separation [2], [1], [5], [14]. In contrast, the Long Short-Term Memory model employed integer encoding followed by a 100-dimensional word-embedding layer to capture both syntactic and semantic dependencies between words in a dense, low-dimensional space [9], [10], [17], [6]. This context-aware representation enables the model to preserve sequential information and relational meaning across varying sentence structures, as recommended in prior LSTM-based sentiment-analysis studies [9], [10], [11], [12]. Together, these two paradigms—TF–IDF for frequency-based weighting and word embeddings for contextual modeling—facilitate a rigorous comparison between traditional probabilistic learning and deep-learning architectures under equivalent data conditions.

2.6. Naïve Bayes

The Naïve Bayes algorithm represents a probability-based classification approach derived from Bayes' theorem. It determines the likelihood that a document *d* falls within a specific sentiment category *c* by integrating prior class likelihoods with the conditional likelihoods of the observed terms. This relationship is formally described in Equation (2). The model assumes feature independence, allowing the joint probability of all features to be simplified into the product of individual probabilities, as formulated in Equation (3). This assumption, while simplistic, significantly reduces computational complexity and enhances scalability for large text datasets [5], [4], [6].

$$P(C|X)=\frac{P(X|C).P(C)}{P(X)} \tag{2}$$

$$P(X|C)=\prod_{i=1}^{n} P(xi|C) \tag{3}$$

Equation (2) defines the posterior probability as a proportional relationship between the likelihood of observing the document given the class and the prior probability of that class. Equation (3) extends this by decomposing the likelihood term into the product of word-level probabilities, normalized by the frequency of class occurrences within the

corpus. Laplace smoothing is applied to avoid zero-probability cases when unseen words occur during testing [5], [4], [6]. During classification, each review is assigned to the sentiment class that maximizes this posterior probability, effectively performing a Maximum A Posteriori (MAP) estimation. In this study, the Multinomial Naïve Bayes variant was implemented, using TF–IDF vectors as input features to capture both term importance and class distribution across the dataset. The model was trained on 80% of the 50,000-review corpus and validated on the remaining 20%. This formulation provides efficient learning and robust baseline performance for text-based sentiment classification, especially in Indonesian-language corpora [2], [1], [5], [4], [18], [24], [25], [26].

2.7. Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) framework is an advanced form of recurrent architecture designed to capture and preserve extended contextual relationships within time-ordered data.

$$ft = \sigma(Wf \times [xt + ht{-}1] + bf) \quad (4)$$

$$it = \sigma(Wi \times [xt + ht\text{-}1] + bi) \quad (5)$$

$$C\ t = \tanh(WC \times [xt + ht\text{-}1] + bC) \quad (6)$$

$$Ot = \sigma(Wo \,.\, [ht\text{-}1, Xt] + bo \quad (7)$$

$$ht = Ot \,.\, \tanh(Ct) \quad (8)$$

Its computational process, expressed in Equations (4)–(8), regulates the flow of information through three gating mechanisms: forget, input, and output. Equation (4) represents the forgetting mechanism, which identifies and discards outdated elements from the prior memory state to avoid retaining nonessential features [9], [11]. Equation (5) denotes the input mechanism, which controls how fresh data are integrated into the existing memory representation [9], [10]. Equation (6) computes the candidate memory update, integrating the newly processed input with the prior hidden state to form potential contextual information [10], [11]. Equation (7) defines the output gate, which

produces the hidden state used in the following sequence step, effectively representing the semantic context at each time step [9], [12]. Finally, Equation (8) updates the overall cell state by combining retained and newly added information, maintaining temporal and contextual coherence across sequences [9], [11], [12]. In this study, the LSTM model processes tokenized and padded review sequences using 100-dimensional word embeddings and an LSTM layer of 128 units. The network learns temporal and semantic patterns in text, enabling more accurate sentiment classification by capturing nuanced contextual dependencies [9], [10], [11], [12].

2.8. Model Evaluation

The performance of the Multinomial Naïve Bayes and LSTM models was evaluated on the same three-class sentiment dataset (negative, neutral, positive) using a confusion-matrix-based assessment. For each sentiment class, the confusion matrix records true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), which provide the basis for deriving standard classification metrics in multi-class sentiment analysis [5], [9], [10], [19], [18]. Model effectiveness was evaluated using four principal indicators—accuracy, precision, recall, and F1-score—that together provide a comprehensive measure of classification quality. These evaluation metrics were derived using the formulas presented in Equation 9 10 12.

$$\text{Accuracy } \frac{(TP+TN)}{(TP+TN+FP+FN)} \text{ x } 100\% \tag{9}$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \text{x} 100\% \tag{10}$$

$$Precision = \frac{TP}{TP+FP} x\ 100\% \tag{11}$$

$$\text{F1 score} = 2\text{x } \frac{\text{RecallxPrecesion}}{\text{Recall+Precesion}} \text{ x } 100\% \tag{12}$$

Overall accuracy is defined in Equation (9) as the proportion of correctly classified reviews (sum of all true positives across classes) to the total number of reviews. Precision for a given class ccc is defined in Equation (10) as the ratio between true positives and the sum of true and false positives for that class, while recall for class ccc is defined in Equation (11) as the ratio between true positives and the sum of true positives and false negatives. The F1-score for class ccc, shown in Equation (12), is the harmonic mean of

precision and recall, providing a balanced measure of correctness and completeness for each sentiment class [5], [9], [10], [6], [20]. Because the lexicon-labeled dataset exhibits a moderately imbalanced distribution across the three sentiment classes, macro-averaged precision, recall, and F1-score were computed by averaging the per-class values, giving equal weight to each class regardless of its frequency [5], [19], [18]. In addition, weighted averages were also examined, in which each class metric is weighted by its relative support; these weighted scores followed the same qualitative trends as the macro-averaged metrics. Therefore, macro F1 is emphasized in the reported results as the primary indicator of overall model performance in this imbalanced multi-class sentiment-classification setting.

3. RESULTS & DISCUSSION

3.1. Dataset Summary

The experimental corpus consists of 50,000 user-generated reviews of the PLN Mobile application collected from the Google Play Store, representing a large-scale sample of authentic user feedback on billing, outages, payments, and application usability [2], [3], [5], [22]. As summarized in Table 1, the reviews have an average length of 7.61 tokens (median 5, with a range from 1 to 87 tokens) and a vocabulary of 24,084 unique lowercased tokens, indicating short and highly varied user comments that are typical of mobile-application review datasets in Indonesian [2], [1], [5], [14], [23]. This combination of large dataset size and linguistically noisy, informal text underscores the need for robust pre-processing and feature-representation strategies. The corpus is subsequently transformed into a three-class sentiment dataset (negative, neutral, positive) through the lexicon-based labeling procedure described in Section 2.4, and its detailed sentiment distribution is analyzed in Section 3.3 as the basis for model training and evaluation. The detailed statistic Is shown In Table 1.

Table 1. Summary Statistics of the PLN Mobile Review Dataset

Statistic	Values
Total number of reviews	50,000
Average review length (tokens)	7.61
Median review length (tokens)	5
Minimum review length (tokens)	1

Statistic	Values
Maximum review length (tokens)	87
Vocabulary size (unique tokens, lowercased)	24,084

3.2. Pre-Processing Results

Text pre-processing was conducted to prepare the raw user reviews for the feature extraction and modeling stages, ensuring linguistic consistency, noise reduction, and a standardized Indonesian text structure before classification [5], [4], [6]. To illustrate the impact of these operations, a representative PLN Mobile review containing a mixture of positive and negative expressions, emojis, and informal language was selected as an example. The sentence was sequentially transformed through case folding (conversion to lowercase), removal of punctuation, digits, emojis, and hyperlinks, tokenization into individual lexical units, stopword elimination to discard high-frequency function words, normalization of slang and repeated characters, and stemming using the Indonesian Sastrawi library to map derived forms to their base lemmas [1], [5], [14]. As shown in Table 2, each stage progressively simplifies and regularizes the original text, reducing orthographic variation while preserving the core semantic content. The final pre-processed output consists of standardized tokens that are suitable for conversion into numerical features for both the Multinomial Naïve Bayes and LSTM models [5], [4], [6].

Table 2. Example of Pre-Processing Steps

Pre-processing	Pre-processing (Indonesia)
Original Review	*"Aplikasi PLN Mobile ini cukup membantu untuk bayar tagihan listrik 👍, tapi layanan pengaduan ke rumah terlalu lama dan error mulu."*
Cleaning	*"aplikasi pln mobile ini cukup membantu untuk bayar tagihan listrik, tapi layanan pengaduan ke rumah terlalu lama dan error mulu"*
Case Folding	*"aplikasi pln mobile ini cukup membantu untuk bayar tagihan listrik, tapi layanan pengaduan ke rumah terlalu lama dan error mulu"*
Tokenization	*["aplikasi", "pln", "mobile", "ini", "cukup", "membantu", "untuk", "bayar", "tagihan", "listrik", "tapi", "layanan", "pengaduan", "ke", "rumah", "terlalu", "lama", "dan", "error", "mulu"]*
Stopword Removal	*["aplikasi", "pln", "mobile", "membantu", "bayar", "tagihan", "listrik", "layanan", "pengaduan", "rumah", "lama", "error", "mulu"]*

Pre-processing	Pre-processing (Indonesia)
Normalization	*["aplikasi", "pln", "mobile", "membantu", "bayar", "tagihan", "listrik", "layanan", "pengaduan", "rumah", "lama", "sering", "error"]*
Stemming	*["aplikasi", "pln", "mobile", "bantu", "bayar", "tagih", "listrik", "layan", "adu", "rumah", "lama", "sering", "error"]*

3.3. Lexicon-Based Sentiment Distribution

After the lexicon-based labeling procedure described in Section 2.4 was applied to the 50,000 cleaned PLN Mobile reviews, each review was assigned to one of three sentiment classes: negative, neutral, or positive. The resulting distribution is shown in Table 3: 11,924 negative reviews (23.85%), 18,222 neutral reviews (36.44%), and 19,854 positive reviews (39.71%). This pattern is consistent with findings from other Indonesian application-review corpora, where positive feedback is slightly dominant but a substantial proportion of neutral and negative opinions still reflects persistent usability and service issues [2], [1], [5], [22], [23]. Overall, PLN Mobile is viewed somewhat favorably, yet more than 60% of reviews are non-positive, indicating meaningful scope for improving user experience and service reliability. From a modeling perspective, the 39.71/36.44/23.85 split constitutes a moderately imbalanced three-class dataset, motivating the use of macro-averaged precision, recall, and F1-score—as emphasized in Section 2.8—and providing a basis for later analysis of how classical classifiers such as Naïve Bayes may overpredict the majority (positive) class in Indonesian sentiment-analysis tasks [5], [4], [14], [6].

Table 3. Sentiment Distribution

Sentiment Category	Count	Percentage (%)
Positive	19,854	39.71
Neutral	18,222	36.44
Negative	11,924	23.85%

3.4. Sentiment Classification Results

Figure 2 shows the distribution of model predictions on the 10,000-review test set. The Naïve Bayes model classified 4,467 reviews as positive (44.67%), 3,554 as neutral (35.54%), and 1,979 as negative (19.79%), whereas the LSTM model predicted 3,951 positive (39.51%), 3,745 neutral (37.45%), and 2,304 negative (23.04%) reviews. When compared with the

lexicon-based ground-truth distribution reported in Section 3.3 (39.71% positive, 36.44% neutral, 23.85% negative), Naïve Bayes clearly shifts probability mass toward the positive class and underestimates neutral and especially negative sentiment. This behaviour is consistent with previous Indonesian sentiment-analysis studies, where probabilistic bag-of-words models tend to overpredict the majority class due to high-frequency generic positive tokens (e.g., "*bagus*", "*mantap*", "*keren*", "*mudah*") and skewed class priors [5], [4], [14], [6]. In contrast, the LSTM predictions are much closer to the ground-truth proportions, reflecting a more balanced allocation of reviews across the three classes and reinforcing the quantitative performance differences discussed in the subsequent Naïve Bayes and LSTM evaluation subsections.

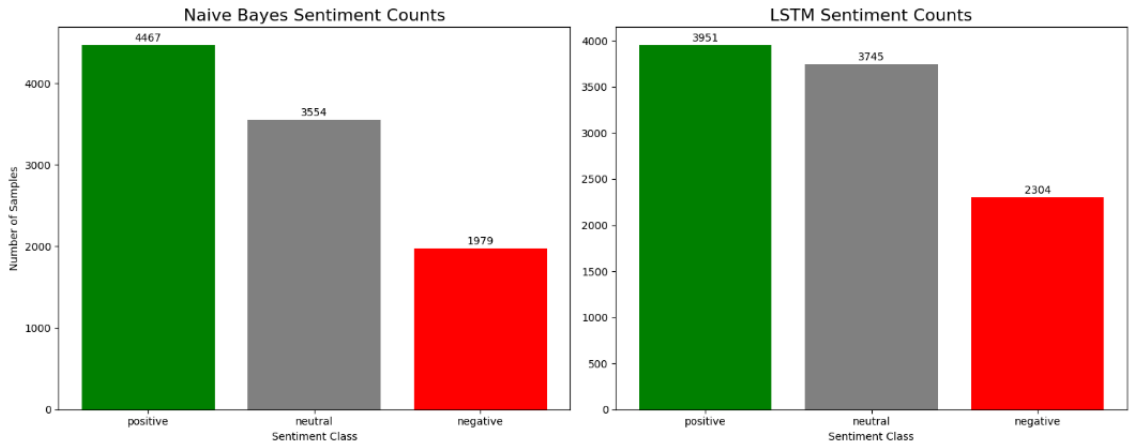

Figure 2. Sentiment Distribution

To further examine the linguistic characteristics of each sentiment class, word cloud visualizations were generated, as shown in Figure 3. The positive word cloud is dominated by terms such as "*mudah*", "*cepat*", "*bantu*", "*baik*", and "*terima kasih*", indicating that users particularly appreciate the ease of use, transaction speed, and helpfulness of PLN Mobile for paying bills and managing electricity services. Neutral reviews are characterized by more factual and descriptive tokens, including "*informasi*", "*tagihan*", *token*, and *update*, which suggest routine usage of the application as an informational tool without strong emotional evaluation. The negative word cloud, on the other hand, highlights words such as "*sangat*", "*lama*", "*error*", "*tidak bisa*", "*gangguan*", "*login*", and "*aplikasi*", pointing to recurring pain points related to failed logins, delayed or unsuccessful transactions, application errors, and perceived instability of the service. These lexical patterns indicate that, while overall sentiment toward PLN Mobile is generally favorable, user

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

dissatisfaction is concentrated around reliability and responsiveness issues, providing concrete directions for PLN to prioritize improvements in system stability, transaction handling, and complaint resolution.

Figure 3. Word Cloud Visualisation

3.5. Naïve Bayes Model Evaluation

The Naïve Bayes model was trained on 40,000 reviews and tested on 10,000 reviews using TF–IDF vectorization to represent term-weight importance across the corpus. Model effectiveness was quantified using accuracy, precision, recall, and F1-score to evaluate It's ability to classify PLN Mobile user sentiments into positive, neutral, or negative categories. The classifier achieved an overall test accuracy of 0.7089 (≈71%), with macro precision of 0.7055, macro recall of 0.6923, and a macro F1-score of 0.6964. At the class level, the F1-scores were 0.6523 for negative reviews, 0.6388 for neutral reviews, and 0.7981 for positive reviews, with corresponding recalls of 0.6019, 0.6272, and 0.8479. These values indicate that the model is most sensitive to positive sentiment but less effective at recovering neutral and especially negative comments, consistent with its tendency to overpredict the positive class relative to the lexicon-based distribution described in Section 3.3. Figure 4 presents the confusion matrix, illustrating the model's classification outcomes. Of the 10,000 test reviews, 3,367 positive, 2,313 neutral, and 1,409 negative samples were correctly predicted on the main diagonal, while misclassifications primarily occurred between neutral and negative categories (e.g., 829 negative reviews predicted as neutral and 997 neutral reviews predicted as positive). Many of these errors arise in reviews that combine mixed or mitigating expressions—such as "*lambat tapi membantu*" ("slow but helpful")—which are difficult for a bag-of-words model with a conditional-independence assumption to interpret. Despite these limitations in capturing contextual nuance, the Naïve Bayes classifier remains computationally efficient, easy to

implement, and provides a transparent probabilistic baseline against which the performance gains of the LSTM model can be rigorously compared.

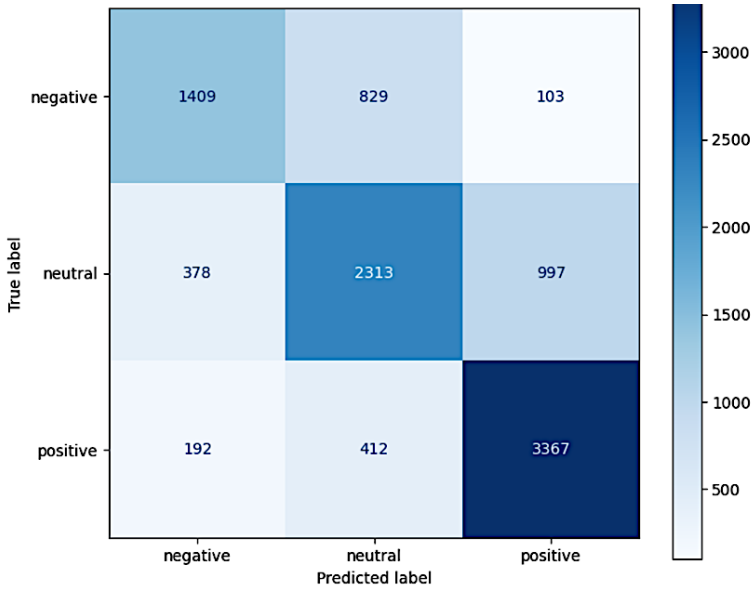

Figure 4. Naïve Bayes Confusion Matrix

3.6. Long Short-Term Memory Model Evaluation

The LSTM model was trained on 40,000 user reviews and tested on 10,000 reviews, using an embedding layer of 100 dimensions to capture contextual and sequential dependencies within the text. The model obtained a test accuracy of 0.9802, representing a substantial enhancement compared to the Naïve Bayes benchmark described in Section 3.5 (0.7089). Based on the confusion-matrix–derived metrics defined in Section 2.8, the LSTM achieved a macro precision of 0.9810, macro recall of 0.9791, and a macro F1-score of 0.9800, indicating highly consistent and near-optimal performance across the three sentiment categories. At the class level, the model recorded F1-scores of 0.9800 for negative reviews, 0.9740 for neutral reviews, and 0.9861 for positive reviews, with precision and recall values exceeding 0.96 for every class. These results demonstrate robust stability and balanced generalization across diverse linguistic expressions in Indonesian PLN Mobile reviews. As shown in Figure 5, the confusion matrix illustrates that misclassifications are minimal, with the vast majority of test samples correctly predicted on the main diagonal and only small amounts of confusion between neighboring sentiment categories.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

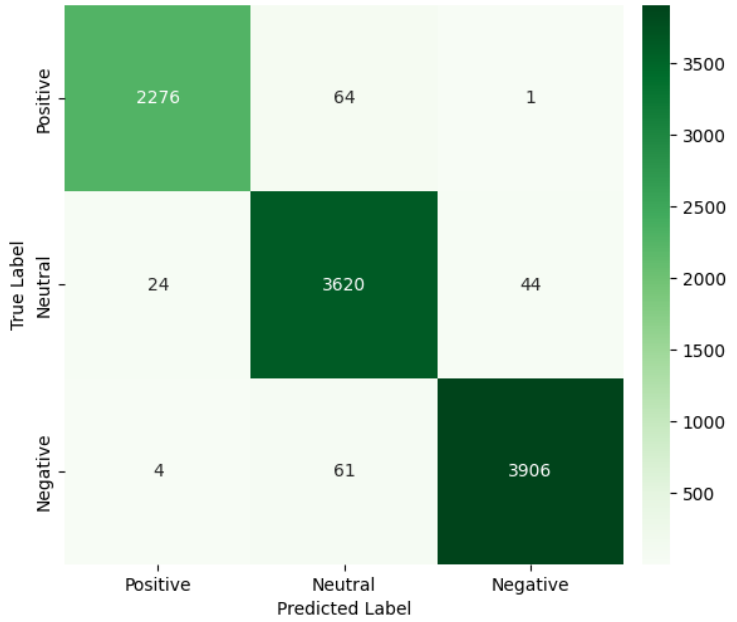

Figure 5. LSTM Confusion Matrix

This pattern reflects the model's ability to learn contextual word dependencies through memory-gating mechanisms that effectively manage long-term relationships within text sequences. During training, the recorded training and validation loss curves decreased smoothly and converged with only a small gap, indicating stable learning dynamics and no evidence of severe overfitting. Unlike the probabilistic approach of Naïve Bayes, the LSTM integrates semantic and syntactic cues, allowing it to interpret nuanced sentiment variations arising from negations, intensifiers, or mixed-polarity sentences [9], [11]–[13]. The high accuracy, balanced class-wise performance, and stable training behaviour validate the model's robustness in large-scale sentiment analysis of Indonesian-language data. Thus, the LSTM model not only surpasses traditional classifiers in quantitative performance but also establishes a scalable and reliable framework for sentiment detection in digital public-service applications such as PLN Mobile.

3.7. Discussion

The empirical findings of this study provide compelling evidence that combining large-scale, lexicon-labeled data with sequence-aware deep learning architectures significantly enhances sentiment classification performance for Indonesian-language reviews of the PLN Mobile application. The comparative evaluation between the Multinomial Naïve Bayes (MNB) and Long Short-Term Memory (LSTM) models reveals clear performance disparities

that have both methodological and operational implications for digital service monitoring.

The lexicon-labeled corpus of 50,000 user reviews represents the largest known dataset in PLN Mobile sentiment-analysis research to date. The sentiment distribution—39.71% positive, 36.44% neutral, and 23.85% negative—suggests that while public perception is slightly skewed toward satisfaction, a substantial 60.29% of reviews express non-positive sentiment. This indicates that many users still experience service-related friction, such as delays, login failures, billing confusion, or system instability. These concerns mirror earlier qualitative findings in [1], [2], and [3], which highlighted recurring usability issues, but the present study quantifies them at scale using sentiment analytics.

When benchmarked, the MNB model achieves 70.89% accuracy and a macro F1-score of 0.6964. However, its class-wise performance reveals imbalances: a high recall of 0.8479 for positive reviews, but only 0.6272 and 0.6019 for neutral and negative reviews, respectively. These figures confirm a known tendency in classical bag-of-words classifiers to overpredict the majority class (in this case, positive) due to their frequency-based assumptions and lack of sequence awareness. This behavior is consistent with prior Indonesian sentiment studies [2], [5], [14], [18], which showed that probabilistic models struggle to accurately capture neutral or negative feedback, especially in short, noisy, or emotionally nuanced texts.

In contrast, the LSTM model demonstrates superior balance and precision across all sentiment categories, achieving 98.02% accuracy and a macro F1-score of 0.9800. Its performance exceeds MNB by 27.13% in accuracy and 28.36% in macro F1, representing a significant leap in classification quality. More importantly, precision and recall exceed 0.96 for all sentiment classes, highlighting LSTM's ability to detect subtle differences in emotional tone and interpret complex linguistic structures, including negations, intensifiers, and mixed-polarity expressions. These gains align with findings in deep-learning sentiment analysis literature [9]–[13], where recurrent models like LSTM have consistently outperformed classical baselines, especially in morphologically rich languages such as Indonesian.

Beyond raw performance metrics, qualitative insights from word cloud visualizations further support the models' interpretations. Positive reviews are frequently associated with terms like “mudah” (easy), “cepat” (fast), and “bantu” (helpful), indicating user appreciation for usability and speed—features aligned with PLN Mobile’s core service goals. Neutral reviews center around transactional and informational terms such as “tagihan” (billing), “token”, and “update”, suggesting routine app use without strong affective valence. Meanwhile, negative reviews prominently feature “error”, “lama” (slow), “gangguan” (disruption), and “tidak bisa” (can’t)—indicating consistent technical grievances, especially around app stability, login failures, and complaint resolution delays. These themes resonate with user concerns previously captured anecdotally in [1], [2], and [4], but are now evidenced through robust NLP methods.

What distinguishes the LSTM approach is not only its quantitative superiority but also its ability to provide interpretable, actionable insights. By reliably identifying and clustering negative and neutral sentiment, the LSTM model can serve as a real-time sentiment-monitoring tool for PLN. Its outputs can be integrated into dashboards or feedback systems that allow PLN operators to track shifts in user satisfaction, detect emerging complaint patterns, and prioritize system enhancements. This transforms sentiment analysis from a retrospective academic exercise into a practical decision-support instrument, aligning with the research objective of improving public-service responsiveness.

Furthermore, this study demonstrates that lexicon-labeled data, when scaled and adapted to the domain context, can produce reliable ground truth even in low-resource language environments. The integration of domain-specific sentiment lexicon expansion ensures that PLN-specific vocabulary (e.g., “pemadaman”, “tagihan”, “layanan”) is properly interpreted, improving labeling quality and reducing noise. This lexicon-based approach also circumvents the cost and time constraints of manual annotation, offering a sustainable model for other Indonesian public-service platforms that seek to analyze user feedback at scale.

Methodologically, this study also addresses and advances several critical limitations in existing PLN Mobile sentiment-analysis literature. Unlike prior works [1], [2], [3], [7], and [8] that relied on small datasets (typically <5,000 reviews) and feature-insensitive models

(e.g., KNN, Decision Trees, SVMs), this research operates on a 50,000-review dataset and explicitly contrasts a probabilistic baseline (MNB) with a sequence-aware deep-learning model (LSTM) under consistent data conditions. The findings demonstrate not only a marked performance advantage for deep learning but also the importance of contextual modeling in understanding sentiment in Indonesian-language reviews—where informality, slang, and morphological variation are prevalent.

From a policy and infrastructure standpoint, these findings carry important implications. As PLN Mobile continues to evolve as a national digital platform, ensuring that the voice of the user is continuously heard and responded to is critical for maintaining trust and improving service delivery. The proposed framework empowers PLN to transition from static feedback systems to dynamic, data-driven customer-experience strategies, leveraging real-time analytics to guide development priorities and operational refinements. Furthermore, this framework can be adapted and extended to other sectors of Indonesian public services—such as health, education, or transportation—where mobile applications are increasingly central to citizen engagement.

The combination of lexicon-based labeling, deep-learning modeling, and user-level text mining produces a scalable, interpretable, and high-performing sentiment-analysis pipeline. It represents not just a technical contribution but also a strategic enabler for public service innovation in Indonesia. The robust results of the LSTM model—both in metric performance and semantic understanding—establish it as a valuable tool for sentiment monitoring in Indonesian public-service ecosystems and offer a replicable model for similar use cases across the Global South.

4. CONCLUSION

This research proposed a comparative sentiment-analysis framework that combines lexicon-based labeling with Multinomial Naïve Bayes and LSTM architectures to analyze public perception of the PLN Mobile application. By processing 50,000 user reviews from the Google Play Store, the framework provided a large-scale, data-driven assessment of user sentiment toward Indonesia's national electricity service platform. The methodological pipeline—covering data collection, text pre-processing, lexicon-based polarity assignment using an extended InSet dictionary, and supervised model training—

transformed unstructured textual feedback into structured sentiment categories that characterize both user satisfaction and dissatisfaction.

Experimental results showed that the LSTM model delivered markedly superior predictive performance compared with the Naïve Bayes benchmark. On the 10,000-review test set, Naïve Bayes achieved an accuracy of 0.7089 and a macro F1-score of 0.6964, whereas the LSTM model attained 0.9802 accuracy and a macro F1-score of 0.9800. Thus, LSTM improved accuracy by 0.2713 and macro F1 by 0.2836 relative to Naïve Bayes, with precision and recall exceeding 0.96 for all sentiment classes. While Naïve Bayes effectively identified sentiment-relevant keywords, it lacked contextual sensitivity and tended to overpredict the positive class. In contrast, the sequential LSTM architecture successfully modeled temporal dependencies, negations, and mixed-polarity expressions (e.g., "*pelayanannya cepat tapi sering error*"), producing a more realistic picture of user experience and a more balanced recovery of neutral and negative sentiment.

The novelty and practical contribution of this study lie in integrating automatic lexicon-based annotation with scalable deep-learning on a 50,000-review corpus, and in benchmarking a domain-adapted LSTM directly against a Naïve Bayes baseline for PLN Mobile sentiment analysis. Unlike previous studies limited to smaller datasets or single classical algorithms, this work demonstrates that sequence-aware deep-learning architectures offer substantial and quantifiable gains for Indonesian public-service sentiment. For PLN and other public-sector stakeholders, the proposed framework can function as a decision-support tool to track sentiment trends, detect emerging complaint themes, and prioritize interventions related to reliability, transaction handling, and customer support.

This research has several limitations. Sentiment labels were generated automatically using a lexicon-based approach, which may introduce noise for sarcasm or highly informal slang. The dataset is restricted to Indonesian-language Google Play Store reviews of PLN Mobile, limiting generalizability to other platforms or services. Moreover, the comparative analysis focused on Naïve Bayes and a single LSTM configuration. Future work may incorporate partially manually validated labels, extend analysis to multi-platform or multilingual settings, and evaluate more advanced architectures such as

transformer-based models and aspect-level sentiment analysis to obtain finer-grained insights into specific service components.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universitas Dr. Soetomo, particularly the Faculty of Engineering and the Informatics Department, as well as the professors who provided guidance and support throughout this research. The authors also extend heartfelt thanks to their friends whose encouragement and assistance were invaluable during the more challenging stages of this work.

REFERENCES

[1] A. Carla, H. Soetanto, and Y. Yuliazmi, "Implementasi text mining untuk analisis sentimen pada pengguna PLN Mobile menggunakan metode Naïve Bayes," *Bit (Fakultas Teknologi Informasi Universitas Budi Luhur)*, vol. 21, no. 1, pp. 72–77, 2024.

[2] S. Syafrizal, M. Afdal, and R. Novita, "Analisis Sentimen Ulasan Aplikasi PLN Mobile Menggunakan Algoritma Naïve Bayes Classifier Dan K-Nearest Neighbor," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 10–19, Dec. 2023, doi: 10.57152/malcom.v4i1.983.

[3] H. Faisal, A. Febriandirza, and F. N. Hasan, "Analisis sentimen terkait ulasan pada aplikasi PLN Mobile menggunakan metode Support Vector Machine," *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, vol. 5, no. 1, pp. 303–312, 2024.

[4] A. A. Purnama and Y. R. Sipayung, "Sentiment Analysis of Public Service Using Naïve Bayes Classifier," *Journal of Information Systems and Informatics*, vol. 7, no. 3, pp. 2439–2457, Sep. 2025, doi: 10.51519/journalisi.v7i3.1207.

[5] K. Nurfebia and S. Sriani, "Sentiment Analysis of Skincare Products Using the Naive Bayes Method," *Journal of Information Systems and Informatics*, vol. 6, no. 3, pp. 1663–1676, Sep. 2024, doi: 10.51519/journalisi.v6i3.817.

[6] K. P. Gunasekaran, "Exploring sentiment analysis techniques in natural language processing: A comprehensive review," *arXiv preprint*, arXiv:2305.14842, 2023.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[7] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile," *PETIR*, vol. 15, no. 2, pp. 264–275, Nov. 2022, doi: 10.33322/petir.v15i2.1733.

[8] Ihsan Zulfahmi, "Analisis Sentimen Aplikasi PLN Mobile Menggunakan Metode Decission Tree," *Jurnal Penelitian Rumpun Ilmu Teknik*, vol. 3, no. 1, pp. 11–21, Dec. 2023, doi: 10.55606/juprit.v3i1.3096.

[9] R. Y. Ahmed, N. F. Yuosif, S. A. Ahmed, and A.-B. A. Mohammed, "Comparison of RNN and LSTM Classifiers for Sentiment Analysis of Airline Tweets," *Journal of Information Systems and Informatics*, vol. 7, no. 2, pp. 1893–1913, Jun. 2025, doi: 10.51519/journalisi.v7i2.1140.

[10] A. Alfarel, H. Hasmawati, and B. Bunyamin, "Sentiment Analysis of Telkom University using the Long Short-Term Memory and Word2Vec Feature Expansion," *Jurnal Teknologi Informasi dan Pendidikan*, vol. 17, no. 2, pp. 468–481, Dec. 2024, doi: 10.24036/jtip.v17i2.889.

[11] A. Azrul, A. I. Purnamasari, and I. Ali, "Analisis sentimen pengguna Twitter terhadap perkembangan artificial intelligence dengan penerapan algoritma Long Short-Term Memory (LSTM)," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 1, pp. 413–421, 2024.

[12] A. S. Widagdo, K. N. Qodri, F. E. N. Saputro, and N. A. R. Putri, "Analisis sentimen mobil listrik di Indonesia menggunakan Long-Short Term Memory (LSTM)," *JURNAL FASILKOM*, vol. 13, no. 3, pp. 416–423, 2023.

[13] L. Y. Wibowo, N. Annisa, P. A. Khairunnisa, V. H. Pranatawijaya, and R. Priskila, "Implementasi Long Short-Term Memory dalam analisis sentimen pengguna aplikasi Twitter yang mengandung ujaran kebencian," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 3, pp. 3170–3174, 2024.

[14] S. K. Wardani, Y. A. Sari, and I. Indriati, "Analisis sentimen menggunakan metode Naïve Bayes Classifier terhadap review produk perawatan kulit wajah menggunakan seleksi fitur N-gram dan Document Frequency Thresholding," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 12, pp. 5582–5590, 2021.

[15] A. D. Pramesti, K. Umam, and M. R. Handayani, "Identification of buzzers in skincare reviews using a lexicon-based sentiment analysis method," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2598–2606, 2025.

[16] H. Firda, P. Putra, N. R. Oktadini, P. E. Sevtiyuni, K. O. Ilir, and S. Selatan, "Perbandingan pelabelan rating-based dan Inset Lexicon-based dalam analisis sentimen menggunakan SVM (studi kasus: ulasan aplikasi GoBiz di Google Play Store)," *Jurnal Sistemasi*, vol. 14, no. 2, pp. 516–528, 2025.

[17] R. Ramadhan, Y. A. Sari, and P. P. Adikara, "Perbandingan pembobotan Term Frequency-Inverse Document Frequency dan Term Frequency-Relevance Frequency terhadap fitur N-Gram pada analisis sentimen," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 11, pp. 5075–5079, 2021.

[18] P. R. Sari, D. R. Indah, E. Rasywir, and G. Athalina, "Comparison of Naive Bayes and SVM algorithms for sentiment analysis of PUBG Mobile on Google Play Store," *SISTEMASI*, vol. 13, no. 6, pp. 2767–2779, 2024.

[19] S. A. Z. Kusuma, D. E. Ratnawati, and N. Y. Setiawan, "Analisis sentimen pengguna sosial media Twitter/X terhadap acara Clash of Champions menggunakan metode Multinomial Naïve Bayes," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 3, 2025.

[20] M. N. Arifin, Amir Hamzah, Moh. A. Huda, and N. Hasanah, "Analysis of Google Play Store User Sentiment Towards Application X Using the SVM Algorithm," *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 1, pp. 249–258, Jun. 2025, doi: 10.47709/brilliance.v5i1.6024.

[21] Z. Mahendra and A. Ridok, "Analisis sentimen opini masyarakat terhadap fenomena TikTokShop di Indonesia menggunakan metode K-Nearest Neighbor berbasis N-gram dengan seleksi fitur Information Gain," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 5, 2024.

[22] E. Eskiyaturrofikoh and R. R. Suryono, "Analisis Sentimen Aplikasi X Pada Google Play Store Menggunakan Algoritma Naïve Bayes Dan Support Vector Machine (SVM)," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 3, pp. 1408–1419, Aug. 2024, doi: 10.29100/jipi.v9i3.5392.

[23] Z. Rahman, P. Sakinah, Y. Hendra, B. Satria, F. Maulana, and A. Q. Ayun, "Sentiment analysis of Gojek app reviews on Google Play Store with natural language processing using Naive Bayes' algorithm," *Jurnal Multimedia dan Teknologi Informasi (Jatilima)*, vol. 6, no. 03, pp. 60–69, 2024, doi: 10.54209/jatilima.v6i03.1189.

[24] S. D. Prasetyo, S. S. Hilabi, and F. Nurapriani, "Analisis sentimen relokasi Ibukota Nusantara menggunakan algoritma Naïve Bayes dan KNN," *Jurnal KomtekInfo*, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[25] S. Wulandari, G. Atha, P. F. Alam, and M. Rendra, "Identification of Karleen Hijab Fashion SME Competitors Based on Sentiment Analysis Using Naïve Bayes Classifier Algorithm," *JTERA (Jurnal Teknologi Rekayasa)*, vol. 7, no. 2, p. 323, Dec. 2022, doi: 10.31544/jtera.v7.i2.2022.323-330.

[26] R. Zahrani, N. Rosa Damayanti, E. Yulianingsih, and M. Ariandi, "Analisis Sentimen Opini Terhadap Novel Pada Website Goodreads Menggunakan Metode Naive Bayes Classifier," *Jurnal Teknologi Rekayasa)*, vol. 9, no. 2, p. 25, 2024, doi: 10.31544/jtera.v9.i1.2024.77-84.