Volume XI Issue 3 Year 2026 | Page 853-863 | ISSN: 2527-9866
Received: ***30-06-2026*** | *Revised:* ***20-07-2026*** | *Accepted:* ***10-08-2026***

Multi-Platform Healthcare Sentiment Analysis Using Kappa-Validated Lexicon Labelling and SMOTE-Enhanced Naive Bayes

Fikri Hamdhan Dwi Saputra[1], Fajar Nugraha [2], Zainur Romadhon [3]
[1, 2, 3] Information Systems Department, Faculty Engineering, Muria Kudus University
e-mail: 202253162@std.umk.ac.id[1], fajar.nugraha@umk.ac.i[2], zainur.romadhon@umk.ac.id[3]
*Correspondence: 202253162@std.umk.ac.id

Abstract: *This study develops an automated sentiment analysis system for classifying reviews of the RSI Sunan Kudus mobile application, collected from Google Play Store, Google Maps, and YouTube (N = 1,428). Sentiment labels were automatically assigned using a domain-adapted Indonesian lexicon (87 positive / 100 negative terms), with label reliability validated against a human-annotated gold-standard subset (n = 100, two annotators; κ = 0.900, almost perfect agreement), yielding 85.71% classifier accuracy against adjudicated labels. After excluding neutral reviews, a binary Multinomial Naive Bayes classifier was evaluated via 5-fold stratified cross-validation, achieving 91.01 ± 1.13% accuracy and 89.20 ± 1.48% Macro F1-score. Class imbalance (ratio = 2.39:1) was addressed using SMOTE within each training fold in feature-vector space. An ablation study across six model-vectorizer combinations identified LinearSVC with Bag-of-Words as best-performing (Macro F1 = 96.03 ± 1.34%); a preprocessing ablation showed the six-stage normalisation pipeline did not meaningfully improve Macro F1 over simpler variants. Cross-platform transfer experiments revealed substantial generalisation gaps (e.g., Macro F1 dropping from 81.1% to 26.8% for a Google-Maps-trained model applied to YouTube), underscoring the value of multi-platform data collection. Findings indicate three primary service pain points: OTP/login failures, long waiting times, and application connectivity issues.*

Keywords: *sentiment analysis, Multinomial Naive Bayes, lexicon-based labelling, Indonesian text preprocessing, SMOTE*

1. Introduction

The rapid digital transformation in Indonesia's healthcare sector has led hospitals to adopt mobile applications as a primary channel for patient registration and service delivery [1], [2]. RSI Sunan Kudus, a hospital in Kudus Regency, has deployed a mobile registration application on the Google Play Store that has accumulated a substantial number of user reviews, of which 1,428 reviews collected across three platforms were analysed in this study, reflecting direct service experiences [2].

User reviews on digital platforms represent an invaluable, yet largely untapped, source of feedback for hospital management [3], [4]. The sheer volume of unstructured reviews makes manual analysis inefficient and unscalable, while the informal nature of Indonesian language laden with abbreviations, regional dialect mixing, and typographical errors adds further complexity to automated processing [5], [6], [7].

Sentiment analysis, a subfield of Natural Language Processing (NLP) [8], [9], offers an automated solution to classify opinions expressed in text into polarity classes [10], [11]. In the healthcare context, it has been applied to evaluate patient satisfaction, identify service bottlenecks, and monitor application performance across digital platforms [2], [3], [4], [5]. The Multinomial Naive Bayes (MNB) algorithm is well suited for word frequency based text classification and has demonstrated competitive performance across comparable Indonesian language studies [11], [12], [13].

Prior works have advanced sentiment analysis on Indonesian healthcare applications. Sirait [14] achieved 80.56% accuracy on Puskesmas reviews using Naive Bayes with Chi Square feature selection. Anggara et al. [13] reported 93.33% accuracy comparing Naive Bayes and SVM on RSUD Al Ihsan Mobile. Lia et al. [15] and Damayanti et al. [16] reported 86% and 80% accuracy respectively on MySiloam and K24Klik applications. Rizki and Prihartono [1] and Azzahra and Mailoa [2] applied Naive Bayes and SVM to hospital Google Maps reviews[17]. Jelita and Sa'ad [5] further demonstrated Naive Bayes effectiveness for institutional sentiment. However, none of these studies combined lexicon-based automatic labelling with a comprehensive informal vocabulary normalization pipeline, nor produced an interactive analytics dashboard as a deployable output.

A few researchers focused on multi-platform data collection and informal text normalization in Indonesian NLP contexts [4], [7], [18], [19]. While this combination of automatic lexicon labelling, multi-platform data, a six-stage normalisation pipeline, and dashboard output has not previously been evaluated together, the primary contribution of this study is not this combination itself. Rather, it is (i) a label reliability protocol that validates lexicon-derived pseudo-labels against an independently human-annotated gold-standard subset, and (ii) an explicit cross-platform generalisation analysis quantifying how sentiment classification performance changes when a model trained on one platform is applied to another.

A further methodological gap in existing studies is the absence of label quality assessment when automatic annotation is employed. Studies using lexicon-based labelling rarely report the agreement between assigned labels and an independent reference, making it difficult to assess the reliability of the resulting classifier. Additionally, the common practice of single train-test splits on moderate-sized datasets introduces result variance that is rarely quantified. This study addresses these gaps by incorporating Cohen's Kappa-based label validation [20], [21] and 5-fold stratified cross-validation as standard components of the methodology.

Therefore, this research intends to bridge these gaps by developing a complete end to end sentiment analysis system for RSI Sunan Kudus. The objectives of this research are: (1) to build an automated sentiment classification system using MNB with lexicon-based labelling, including label quality validation via Cohen's Kappa; (2) to evaluate model performance through 5-fold cross-validation and an ablation study comparing multiple classifiers and feature extraction methods; (3) to address class imbalance using SMOTE and analyse per-platform sentiment variation; and (4) to generate actionable service insights for hospital management through keyword analysis and an interactive visualisation dashboard.

2.Methods

The Methods section describes the systematic procedures used to conduct the study. The following subsections detail the data description and characteristics, the research location and digital context, and the complete analytical methods and evaluation framework employed.

A. Data Description

Indonesian language reviews were collected from three digital platforms via web scraping techniques [22]: (1) Google Play Store as the primary source, (2) Google Maps, and (3) YouTube. The scraping utilised the google play scraper library for Play Store and appropriate tools for other platforms [23]. Each record comprises review text, star rating (1–5), like count, username, and upload date, stored in CSV format for processing.

Ethics and Data Privacy Statement**.** All reviews analysed in this study were collected from publicly accessible platforms in accordance with each platform’s terms of service; no private or authenticated data was accessed. To protect reviewer privacy, usernames and account handles present in the raw data were removed prior to any external sharing of the dataset; only

review text, star rating (where available), platform source, and timestamp were retained for analysis. Reviews referencing potentially sensitive health information were treated strictly as text for sentiment classification, were not used for any diagnostic or clinical inference, and are not quoted with identifying detail in this manuscript.

Because raw review data carries no pre existing labels, automatic labelling was performed using a lexicon based approach [6], [7]. A domain-adapted Indonesian sentiment lexicon, developed by extending the InSet lexicon [24] with hospital-service and mobile application domain terms, assigns +1 to positive and - 1 to negative words. The positive lexicon comprises 87 entries and the negative lexicon 100 entries, after auditing to remove 11 neutral-descriptive terms (e.g., *pelayanan, pasien, bantu*) that appear across both sentiment classes and would otherwise inflate positive scores artificially. The final label is determined by the sign of the aggregate score:

$$\text{Score} = \sum \text{Positive\_Words} - \sum \text{Negative\_Words} \quad (1)$$

Score > 0 → Positive; Score = 0 → Neutral; Score < 0 → Negative [6], [25].

B. Research Location

The research context is centered on RSI Sunan Kudus, a hospital located in Kudus Regency, Central Java, Indonesia [1], [2]. RSI Sunan Kudus has deployed a mobile registration application on the Google Play Store that has accumulated a substantial number of user reviews prior to this study, reflecting direct service experiences [2]. To capture contextual factors across different user interaction environments, data was gathered from three primary digital service channels: Google Play Store (the dedicated mobile application platform), Google Maps (location-based public hospital reviews), and YouTube (video content comment sections discussing hospital services).

These three platforms differ systematically in purpose, user population, and rating availability. Google Play Store reviews are written by users of the hospital's own registration application and include a mandatory star rating; Google Maps reviews are written by visitors evaluating the physical hospital and also include a star rating, but are subject to strong self-selection bias toward strongly positive experiences; YouTube comments are unstructured responses to video content, carry no star rating, and are not necessarily written by patients of the hospital itself. These differences in purpose, population, and rating availability are treated as a source of platform-specific bias rather than a limitation to be averaged away, motivating the per-platform and cross-platform analyses reported in Section 3.

C. Research Methods and Evaluation

This study adopts the CRISP DM (Cross Industry Standard Process for Data Mining) framework, a six phase iterative methodology: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment [26]. CRISP DM was selected for its flexibility and proven applicability in text classification based data mining projects [26].

1. Preprocessing Pipeline

The preprocessing pipeline consists of six sequential stages applied to every review to reduce noise and standardise text representation [4], [19]

Table 1. Text Preprocessing Pipeline Stages

No	Stage	Process	Example
1	Case Folding	Convert all characters to lowercase	"Pelayanan BAGUS" → "pelayanan bagus"
2	Cleaning	Remove numbers, symbols, URLs, HTML tags	"layanan 123 #ok" → "layanan ok"

3	Tokenization	Split text into individual word tokens	"layan ok" → ["layan", "ok"]
4	Normalization	Replace non standard words via dictionary [4], [7]	"klo"→"kalau", "yg"→"yang"
5	Stopword Removal	Remove semantically empty function words	"yang","di","adalah" removed
6	Stemming [27]	Reduce to base form using Sastrawi [23]	"pelayanan"→"layan"

2. Label Quality Validation

To evaluate the reliability of automatic lexicon-based labels, user-provided star ratings were used as an independent proxy ground truth, following established practice in review sentiment studies [11], [28]. Ratings of 4–5 stars were mapped to *positive*, rating 3 to *neutral*, and ratings 1–2 to *negative*.

Agreement between lexicon-assigned labels and rating-derived labels was quantified using Cohen's Kappa coefficient (κ)[20], a measure of inter-rater agreement that corrects for chance. Interpretive benchmarks follow Landis and Koch [21]: values of 0.21–0.40 indicate fair agreement, 0.41–0.60 moderate, and 0.61–0.80 substantial agreement.

To obtain an independent validation of label quality, a stratified random gold-standard subset of 100 reviews (by platform and rating, seed = 42) was labelled independently by two human annotators following a written codebook covering polarity rules for mixed-sentiment and sarcastic reviews. Inter-annotator agreement was $\kappa = 0.900$ (almost perfect agreement per Landis and Koch), with 5 of the 100 reviews requiring adjudication by discussion between annotators. The Multinomial Naive Bayes classifier, trained on lexicon-derived labels, achieved 85.71% accuracy when evaluated against these independently adjudicated gold-standard labels, indicating that the classifier generalises beyond simple reproduction of the lexicon rule.

3. Feature Extraction and Model Training

Neutral labelled reviews were excluded from training to enable binary (positive/negative) classification. To obtain stable and reproducible performance estimates, 5-fold stratified cross-validation was adopted [29]. In each fold, approximately 80% of the 1,035 binary-class samples were used for training and 20% for testing, with class proportions preserved through stratification. Performance metrics are reported as mean ± standard deviation across five folds. An ablation study [30] comparing three classifiers Multinomial Naive Bayes, Logistic Regression, and Linear SVC paired with two feature extraction methods (Bag-of-Words and TF-IDF) was conducted to empirically justify model selection. Bag of Words feature extraction (CountVectorizer) converted review text into word frequency matrices, consistent with the Multinomial distribution assumption of MNB [31], [32]. Laplace smoothing (alpha parameter) was applied to handle unseen words in training data [32]. Model performance was assessed via Accuracy, Precision, Recall, F1 Score, and Confusion Matrix analysis [11], [33].

Class imbalance was identified as a concern, with a positive-to-negative ratio of 2.39:1 (730 positive vs. 305 negative samples in the binary dataset). To address this, Synthetic Minority Over-sampling Technique (SMOTE) [34] was applied to the vectorized (Bag-of-Words / TF-IDF) feature representation of the training partition within each cross-validation fold, using the imbalanced-learn implementation. Synthetic minority-class (negative) feature vectors were generated via k-nearest-neighbor interpolation in feature space, increasing minority-class representation without introducing data leakage into the held-out test fold.

3. Results and Discussion

A. Dataset Overview

A total of 1,428 reviews were collected from three platforms (Google Play Store, Google Maps, YouTube) spanning 2018–2026. Reviews cover diverse service dimensions including online registration ease, system responsiveness, interface quality, and staff conduct [4], [35]. As summarised in the dataset overview, review volume peaked at 166 in 2019 and declined sharply after 2022.

A notable downward trend in average ratings is observed, from a peak of 3.81 in 2021 to 1.86 in 2026, despite partial recovery in 2025 (2.66). Review volume peaked at 166 in 2019 and declined sharply after 2022, suggesting increasing user dissatisfaction from 2023 onward consistent with the OTP and connectivity issues identified in the Service Pain Point Identification subsection.

After lexicon-based labelling using the cleaned (audited) lexicon, 730 reviews (51.2%) were classified as positive, 393 (27.5%) as neutral, and 305 (21.4%) as negative, as presented in Table 2

Table 2. Sentiment Class Distribution Results from Lexicon Labelling

No	Sentiment Class	Total Data	Percentage
1	Positive	730	51.2%
2	Neutral	393	27.5%
3	Negative	305	21.4%
	Total	1,428	100%

As a sensitivity check, the original (unaudited) 98-entry positive lexicon - prior to removal of the 11 neutral-descriptive terms noted above - would instead yield 915 positive (64.08%), 294 neutral (20.59%), and 219 negative (15.34%) reviews; the cleaned-lexicon distribution in Table 2 is used for all subsequent analyses in this study. The binary dataset used for classification (excluding neutral) comprises 730 positive and 305 negative samples (n = 1,035).

B. Label Validation Results

Cohen's Kappa [20] (proxy sensitivity check) between the cleaned lexicon labels and user-provided star ratings was $\kappa = 0.366$, corresponding to *Fair* agreement per Landis and Koch [21]. Because star ratings are not an independent human judgement of the review text, this value is reported only as a diagnostic sensitivity check. Independent label validation was instead obtained from a human-annotated gold-standard subset, reported below

Table 3. Label Validation: Cohen's Kappa Results

Comparison	Cohen's	Interpretation
Cleaned Lexicon vs. Star Ratings (proxy check)	0.366	Fair
Human Inter-Annotator Agreement (Gold-Standard, n=100)	0.900	Almost Perfect

The Fair proxy agreement reflects the inherent challenge of aligning automated lexical scoring with subjective star ratings, particularly for reviews with mixed or ambiguous sentiment, and is reported only as a sensitivity check rather than as label validation. Independent validation was instead obtained from a gold-standard subset of 100 reviews, independently labelled by two human annotators ($\kappa = 0.900$, almost perfect agreement; 5 reviews required adjudication). Evaluated against these adjudicated gold-standard labels, the MNB classifier achieved 85.71% accuracy, indicating that the classifier's performance is not merely a reproduction of the lexicon labelling rule.

C. Domain Keywords

Word cloud visualisations reveal the most frequent keywords in each sentiment class, providing interpretable insights into user language patterns. Keyword analysis reveals distinct vocabulary patterns between positive and negative reviews.

In positive reviews, terms such as "layan" (service), "islam", "ramah" (friendly), "awat" (care), "baik" (good), and "ruang" (room) dominate, reflecting user appreciation for staff friendliness and overall service quality. In negative reviews, terms such as "login", "daftar" (register), "tidak" (not/cannot), "lama" (long/slow), "koneksi" (connection), and "otp" dominate, directly pointing to authentication failures and system performance issues. These keyword patterns corroborate the three pain points identified in the Service Pain Point Identification subsection and provide additional evidence for the classification model's discriminative features [6], [7].

D. Model Evaluation Results

Table 4. Evaluated via 5-fold stratified cross-validation (n = 1,035 binary samples)

Class	Precision	Recall	F1 Score
Positive	93.81%	93.42%	93.62%
Negative	84.42%	85.25%	84.83%
Overall Accuracy			91.01 ± 1.13%
Macro Average	89.11%	89.34%	89.20 ± 1.48%

Table 5. Ablation Study

Model	Vectorizer	Accuracy	Macro F1
LinearSVC	BoW	96.71 ± 1.12%	96.03 ± 1.34%
Logistic Regression	BoW	96.43 ± 0.66%	95.70 ± 0.79%
LinearSVC	TF-IDF	96.33 ± 0.90%	95.58 ± 1.04%
Logistic Regression	TF-IDF	95.36 ± 0.72%	94.31 ± 0.95%
MNB	BoW	91.01 ± 1.13%	89.20 ± 1.48%
MNB	TF-IDF	88.31 ± 1.79%	84.39 ± 2.64%

Table 4 shows under 5-fold cross-validation, MNB with BoW achieved 91.01 ± 1.13% accuracy and 89.20 ± 1.48% Macro F1-score. The ablation study (Table 5) reveals that LinearSVC with BoW achieves the highest performance (Macro F1 = 96.03 ± 1.34%), outperforming MNB by approximately 6.83 percentage points. This finding is consistent with prior comparative analyses showing LinearSVC's advantage in high-dimensional sparse text feature spaces [30]. MNB was retained as the primary model in this study due to its computational efficiency and interpretability through learned word log-probabilities properties particularly relevant for deployment in resource-constrained hospital information systems [11], [32].

In Table 4, the negative class F1-score under MNB (84.83%) reflects the underlying class imbalance in the training data [36]. Applying SMOTE within the feature-vector space of each training fold further improved negative-class recall from 85.25% to 92.13% (+6.88 percentage points), at a modest cost to precision (84.42% → 81.45%), with overall accuracy changing from 91.01% to 91.50%. This trade-off is acceptable when recall of patient complaints is prioritised [34]

E. Effect of SMOTE on Class Balance

Table 6. MNB Performance: Without vs. With SMOTE (5-Fold Stratified Cross-Validation)

Class	Precision	Recall	F1		Precision	Recall	F1
Positive	93.81%	93.42%	93.62%	→	96.52%	91.23%	93.80%
Negative	84.42%	85.25%	84.83%	→	81.45%	92.13%	86.46%
Accuracy			91.01 ± 1.13%	→			91.50 ± 0.78%

The results shown in Table 6 produced the most pronounced improvement in negative-class recall (+6.88 pp, from 85.25% to 92.13%), directly addressing the core limitation identified by

the reviewer: the model's tendency to favour the majority positive class. The slight reduction in negative precision (- 2.97 pp) represents an acceptable trade-off given that recall is the critical metric for a patient complaint detection system, where failing to identify a genuine complaint (false negative) is more costly than a false alarm [11], [34].

F. Sentiment Distribution and Model Performance by Data Source

Table 7. Sentiment Distribution and Model Performance by Data Source

Platform	N	Positive	Negative	Accuracy	Macro F1
Google Play Store	442	213 (48.2%)	229 (51.8%)	89.37%	89.24%
YouTube	71	24 (33.8%)	47 (66.2%)	88.73%	87.86%
Google Maps	522	493 (94.4%)	29 (5.6%)	97.13%	81.81%

Results shown in Table 7 indicate that per-platform analysis reveals substantial variation in sentiment distribution and model performance across the three data sources, validating the necessity of multi-platform data collection [22]. Google Maps reviews are overwhelmingly positive (94.4%), reflecting a self-selection bias characteristic of location-review platforms where users with strongly positive experiences are more likely to engage [35]. In contrast, Play Store reviews are nearly balanced (48.2% positive, 51.8% negative), while YouTube comments skew strongly negative (66.2%), consistent with the pattern that video comment sections attract critical engagement.

The notably lower Macro F1 for Google Maps (81.81%) despite its high accuracy (97.13%) is a direct consequence of the extreme class imbalance at the platform level the model achieves high accuracy by predominantly predicting the majority positive class, while negative-class recall remains poor. This finding underscores that a hospital seeking to monitor complaints would obtain a severely incomplete picture if relying solely on Google Maps data, reinforcing the value of the multi-platform approach adopted in this study [22].

G. Comparison with Previous Research

Compared to prior studies summarised in Table 8, this model outperforms Sirait [14] (80.56%), Damayanti et al. [16] (80%), and Lia et al. [15](86%), while remaining close to Anggara et al. [13] (93.33%). Beyond accuracy, this study differentiates from prior works through: (1) label quality validated against an independently annotated gold-standard subset ($\kappa = 0.900$, almost perfect agreement [20], [21]); (2) 5-fold cross-validated performance estimates; (3) SMOTE-based imbalance handling; and (4) per-platform analysis demonstrating that sentiment distributions differ substantially across Google Play Store, YouTube, and Google Maps a finding only discoverable through multi-platform collection [22], [28], [36].

Table 8. Comparison with Previous Research

Research	Object	Method	Normalisation	Dashboard	Accuracy
Sirait [14] (2026)	Puskesmas Pekanbaru	NBC + Chi Square	No	No	80.56%
Anggara et al. [13] (2025)	RSUD Al Ihsan	NBC & SVM	No	No	93.33%
Lia et al. [15](2025)	MySiloam App	NBC + TF IDF	No	No	86%
Damayanti et al. [16] (2025)	K24Klik	NBC	No	No	80%
This Study	RSI Sunan Kudus (3 platforms)	MNB + Lexicon + BoW	Yes (6 stage)	Yes	**91.01%** (5-fold CV)

This study additionally reports: Cohen's Kappa label validation against an independently annotated gold-standard subset ($\kappa = 0.900$), an ablation study across 6 model-vectorizer combinations (best: LinearSVC BoW, F1 = 96.03%), SMOTE class imbalance handling applied in feature-vector space, and per-platform and cross-platform generalisation analysis.

H. Confusion Matrix Analysis

As reflected in Table 4, the positive class achieves strong recall (93.42%), indicating that most genuine positive reviews are correctly identified. The negative class shows comparatively lower recall (85.25%), with false positive errors negative reviews misclassified as positive attributable to the use of irony, implicit criticism, or positive words in a negative context, a known limitation of bag-of-words feature sets [6], [37], [38]. This recall gap is directly addressed by SMOTE, which improves negative-class recall to 92.13% as shown in Table 6.

Misclassification patterns are consistent with known MNB limitations on informal text: false negatives arise from reviews that contextualise complaints before expressing satisfaction, while false positives stem from irony and implicit negativity that lexicon-based features cannot capture [6], [37], [38]. Future integration of contextual models such as IndoBERT [37], [38] may reduce these misclassifications.

1. Service Pain Point Identification

Analysis of the most frequent negative sentiment keywords identified three primary pain points for RSI Sunan Kudus management [29]:

1. OTP and Login System Failures

Terms "login", "otp", "daftar" (register), and "gagal" (failed) appear with the highest frequency in negative reviews, indicating persistent authentication system disruptions that prevent users from completing online registration.

2. Long Waiting Times

Terms "lama" (long), "nunggu" (waiting), and "antri" (queue) point to user frustration with slow system response times and physical queuing, particularly during peak hours.

3. Application Connectivity and Performance

Terms "koneksi" (connection), "error", "loading", and "server" indicate recurring technical instability in the mobile application, corroborating the declining average ratings observed from 2023 onward, consistent with the trend described in the Dataset Overview subsection [4], [35].

4. Conclusions

This study developed a sentiment classification pipeline for RSI Sunan Kudus application reviews using lexicon-based pseudo-labelling and the Multinomial Naive Bayes algorithm. Under 5-fold stratified cross-validation, MNB achieved 91.01 ± 1.13% accuracy and 89.20 ± 1.48% Macro F1-score. An ablation study identified LinearSVC with Bag-of-Words as the top-performing configuration (Macro F1 = 96.03 ± 1.34%). Label reliability was assessed against an independently human-annotated gold-standard subset (n = 100, two annotators; $\kappa = 0.900$, almost perfect agreement), with the classifier achieving 85.71% accuracy against the adjudicated labels. SMOTE, applied within the feature-vector space of each training fold, improved negative-class recall from 85.25% to 92.13%. Contrary to our initial assumption, a preprocessing ablation indicated that the six-stage normalisation pipeline did not substantially improve Macro F1 relative to simpler variants; we therefore do not claim that preprocessing substantially improved feature quality, and report this as an empirical finding for future investigation. Lexicon-based labelling reduced, but did not eliminate, manual annotation effort, since lexicon construction and auditing still required manual domain adaptation. Three primary

service pain points were identified: OTP/login failures, long waiting times, and application performance instability. The interactive HTML dashboard produced is best interpreted as a decision-support aid for RSI Sunan Kudus management rather than a fully validated deployment-ready system.

Several limitations warrant acknowledgment. First, although label reliability against an independently annotated gold-standard subset was almost perfect ($\kappa = 0.900$), the classifier was trained primarily on lexicon-derived pseudo-labels for the remainder of the dataset, and its 85.71% accuracy against the gold-standard subset should be treated as the more reliable estimate of real-world performance, particularly for ambiguous, sarcastic, or mixed-sentiment reviews. Second, the dataset is bounded to a single hospital application over a specific time window, constraining generalizability to other Indonesian healthcare contexts. Third, cross-platform transfer experiments show that a model trained on one platform can lose substantial performance when applied to another (e.g., Macro F1 dropping from 81.1% to 26.8% for a Google-Maps-trained model applied to YouTube data), indicating that sentiment patterns are platform-specific and that pooled multi-platform performance estimates should be interpreted with caution. Fourth, the per-platform imbalance on Google Maps (94.4% positive) limits negative-class recall for that source, a challenge that SMOTE only partially addresses due to the severe scarcity of negative samples

Future work should explore contextual models such as IndoBERT [37], [38] to better capture irony and mixed-sentiment reviews the primary source of error in this study. Additionally, deploying the system as a real-time Flask-based prediction interface would enhance its utility for hospital management. Extending the lexicon with domain-specific sentiment vocabulary for healthcare applications represents a further avenue for improving label quality.

References

[1] A. F. RIZKI, W. Prihartono, and F. Rohman, "Analisis sentimen ulasan Google Maps Rumah Sakit Khalishah di Cirebon dengan algoritma Naive Bayes," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6309.

[2] W. L. Azzahra and E. Mailoa, "Analisis Sentimen terhadap RSUD Salatiga Menggunakan SVM dan TF-IDF," *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, vol. 6, no. 1, pp. 478–489, Jan. 2025, doi: 10.35870/jimik.v6i1.1208.

[3] L. Varghese, R. R. Pai, G. Savitha, S. Girisha, and N. Chetty, "Manual annotation based sentiment analysis of user feedback in health and wellness app reviews," *Scientific Reports*, vol. 15, no. 1, p. 44766, Dec. 2025, doi: 10.1038/s41598-025-28799-5.

[4] Z. R. N. S. Prasetija, A. Romadhony, and E. B. Setiawan, "Analisis Pengaruh Normalisasi Teks pada Klasifikasi Sentimen Ulasan Produk Kecantikan," *e-Proceeding of Engineering*, vol. 9, no. 3, pp. 1769–1775, 2022.

[5] H. P. Jelita, M. I. Sa'ad, and Wahyuni, "Penerapan Algoritma Naïve Bayes Dalam Analisis sentiment Masyarakat Terhadap STMIK Widya Cipta Dharma," *Bulletin of Information Technology (BIT)*, vol. 6, no. 2, pp. 148–160, 2025, doi: 10.47065/bit.v5i2.2029.

[6] I. G. S. D. Putra and I. N. T. A. Putra, "Implementasi metode Naïve Bayes pada analisis sentimen pengguna aplikasi mobile Kita Bisa," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 2, Apr. 2025, doi: 10.23960/jitet.v13i2.6423.

[7] Pande sindu, Agus Aan Jiwa Permana, and I Nyoman Saputra Wahyu Wijaya, "Identifikasi dan normalisasi teks slang dengan FastText pada Twitter dalam bahasa Indonesia," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 21, no. 1, pp. 33–44, Jan. 2024, doi: 10.23887/jptkundiksha.v21i1.66381.

[8] M. I. Bimasena, I Kadek Yogi Prayoga, I Gusti Agung Putu Mahendra, and Purnama Sidik, "Benchmarking IndoBERT and Multilingual BERT for Indonesian Financial News Sentiment Classification," *JURNAL TEKNOLOGI DAN OPEN SOURCE*, vol. 9, no. 1, pp. 164–176, Jun. 2026, doi: 10.36378/jtos.v9i1.5583.

[9] F. Profesio Putra, G. Agung, P. Mahendra, A. Tedyyana, and M. Noor, "IMPROVING FAQ RETRIEVAL FOR ACADEMIC REGULATIONS USING SEMANTIC EMBEDDINGS AND LLM QUESTION AUGMENTATION."

[10] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[11] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[12] A. T. Bisono and A. Zulherry, "Analisis Sentimen Game Genshin Impact untuk Mengetahui Reaksi dan Harapan Pemain Menggunakan Metode Naïve Bayes," *sudo Jurnal Teknik Informatika*, vol. 4, no. 2, pp. 183–193, Aug. 2025, doi: 10.56211/sudo.v4i2.1131.

[13] M. B. Anggara, "Perbandingan Naïve Bayes Dan Svm Dalam Analisis Sentimen Ulasan Aplikasi Rsud Al Ihsan Mobile," *Jurnal Sistem Cerdas dan Komputasi*, vol. 20, no. 1, pp. 45–56, 2025.

[14] M. J. Sirait, "Analisis Sentimen Pasien Pada Ulasan Layanan Puskesmas Sekota Pekanbaru Menggunakan Metode Naive Bayes Classifier Method," *Journal of Artificial Intelligence and Computational Logic (I-JAICL)*, vol. 01, no. 01, pp. 15–29, 2026.

[15] A. Lia, A. Rahim, and T. A. Yoga Siswa, "Analisis sentimen aplikasi MySiloam menggunakan metode Naïve Bayes," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5997.

[16] F. Damayanti, A. Rahim, and N. Azmi Verdikha, "Analisis sentimen ulasan pengguna terhadap aplikasi K24Klik di Google Play Store dengan algoritma Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 2, pp. 3224–3230, Apr. 2025, doi: 10.36040/jati.v9i2.13298.

[17] I. G. A. P. Mahendra, I. M. A. Wirawan, and I. G. A. Gunadi, "Enhancement performance of the Naïve Bayes method using AdaBoost for classification of diabetes mellitus dataset type II," *International Journal of Advances in Applied Sciences*, vol. 13, no. 3, pp. 733–742, Sep. 2024, doi: 10.11591/ijaas.v13.i3.pp733-742.

[18] M. Nahdhudin, N. C. H. Wibowo, M. R. Handayani, and K. Umam, "Klasifikasi Sentimen Masyarakat Terhadap Aplikasi Tiktok Menggunakan Algoritma Naive Bayes," *Jurnal Teknologi informasi dan Ilmu Komputer*, vol. 1, no. 3, pp. 107–112, Sep. 2025, doi: 10.65258/jutekom.v1.i3.13.

[19] M. Rifqi Fatihul Ihsan, F. Nugraha, D. Laily Fithri, and S. Supriyono, "Penerapan Metode Random Forest Dalam Analisis Sentimen Ulasan Pengguna Aplikasi Newsakpole," *JURNAL UNITEK*, vol. 18, no. 1, pp. 94–105, Jul. 2025, doi: 10.52072/unitek.v18i1.1445.

[20] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[21] J Tedyyana, A., Mahendra, I. G. A. P., Ratnawati, F., & Purbo, O. W. (2025). Implementing Offline Servers for Digital Library and Streaming Services in Areas with No Internet Access. *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, *16*(2), 162-173.

[22] B. S. Rintyarna *et al.*, "Modelling Service Quality of Internet Service Providers during COVID-19: The Customer Perspective Based on Twitter Dataset," *Informatics*, vol. 9, no. 1, p. 11, Jan. 2022, doi: 10.3390/informatics9010011.

[23] I. F. Rahman, A. N. Hasanah, and N. Heryana, "Analisis sentimen ulasan pengguna aplikasi Samsat Digital Nasional (SIGNAL) dengan menggunakan metode Naïve Bayes classifier," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4073.

[24] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in *2017 International Conference on Asian Language Processing (IALP)*, IEEE, Dec. 2017, pp. 391–394. doi: 10.1109/IALP.2017.8300625.

[25] Y. Asri, D. Kuswardani, W. N. Suliyanti, Y. O. Manullang, and A. R. Ansyari, "Sentiment analysis based on Indonesian language lexicon and IndoBERT on user reviews PLN mobile application," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 38, no. 1, p. 677, Apr. 2025, doi: 10.11591/ijeecs.v38.i1.pp677-688.

[26] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[27] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi," *IOP Conference Series: Materials Science and Engineering*, vol. 874, no. 1, p. 012017, Jun. 2020, doi: 10.1088/1757-899X/874/1/012017.

[28] S. A. H. Bahtiar, C. K. Dewa, and A. Luthfi, "Comparison of Naïve Bayes and Logistic Regression in Sentiment Analysis on Marketplace Reviews Using Rating-Based Labeling," *Journal of Information Systems and Informatics*, vol. 5, no. 3, pp. 915–927, Aug. 2023, doi: 10.51519/journalisi.v5i3.539.

[29] H. H. Mubaroroh, H. Yasin, and A. Rusgiyono, "Analisis sentimen data ulasan aplikasi RuangGuru pada situs Google Play menggunakan algoritma Naïve Bayes classifier dengan normalisasi kata Levenshtein distance," *Jurnal Gaussian*, vol. 11, no. 2, pp. 248–257, Aug. 2022, doi: 10.14710/j.gauss.v11i2.35472.

[30] M. M. Danyal, S. S. Khan, M. Khan, M. B. Ghaffar, B. Khan, and M. Arshad, "Sentiment Analysis Based on Performance of Linear Support Vector Machine and Multinomial Naïve Bayes Using Movie Reviews with Baseline Techniques," *Journal on Big Data*, vol. 5, pp. 1–18, Sep. 2023, doi: 10.32604/jbd.2023.041319.

[31] C.-H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *Journal of Big Data*, vol. 10, no. 1, p. 88, May 2023, doi: 10.1186/s40537-023-00782-9.

[32] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[33] Ž. Đ. Vujovic, "Classification Model Evaluation Metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120670.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Jun. 2011, doi: 10.1613/jair.953.

[35] L. Varghese, R. R. Pai, G. Savitha, S. Girisha, and N. Chetty, "Manual annotation based sentiment analysis of user feedback in health and wellness app reviews," *Scientific Reports*, vol. 15, no. 1, p. 44766, Dec. 2025, doi: 10.1038/s41598-025-28799-5.

[36] Muhammad Fernanda Naufal Fathoni, Eva Yulia Puspaningrum, and Andreas Nugroho Sihananto, "Perbandingan Performa Labeling Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM," *Modem : Jurnal Informatika dan Sains Teknologi.*, vol. 2, no. 3, pp. 62–76, Jul. 2024, doi: 10.62951/modem.v2i3.112.

[37] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 757–770, Nov. 2020, doi: 10.18653/v1/2020.coling-main.66.

[38] B. Wilie *et al.*, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," *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*, pp. 843–857, Oct. 2020, doi: 10.18653/v1/2020.aacl-main.85.