J-KOMA Journal of Computer Science and Applications

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Open Access

***E-ISSN**: 2620 - 4872*
***Vol**.09, **No**.01*
***Doi**: 10.21009/JKOMA.091.05*

***Recivied**: 30 June 2026*
***Accepted**: 22 July 2026*
***Published**: 31 July 2026*

Keywords:
sentiment analysis; Telemedicine; lexicon-based; machine learning; cross-validation.

***Correspondence Email:**
ahsanunnaseh@itsk-soepraoen.ac.id

Lexicon-Based and Machine Learning Approaches for Sentiment Classification of Telemedicine Reviews in Indonesia

Ahsanun Naseh Khudori[1], Fauziah[2]

[1] Institut Teknologi, Sains dan Kesehatan RS dr. Soepraoen Kesdam V/Brw, Indonesia
[2] Politeknik Negeri Malang, Indonesia

Abstract

The rapid growth of telemedicine services in Indonesia has generated a substantial volume of user reviews, an important source for evaluating service quality. This study compares the effectiveness of lexicon-based and machine learning approaches in classifying review sentiment, using the three largest telemedicine applications in Indonesia, Alodokter, Halodoc, and KlikDokter as a case study. A publicly available dataset of 19,389 reviews, manually labelled by two annotators under the guidance of a psychologist, served as the gold standard and was preprocessed through case folding, normalisation, stopword removal, and stemming. The lexicon-based approach employs the InSet dictionary, while the machine learning approach applies Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF) with TF-IDF features, evaluated via stratified 10-fold cross-validation. Given the highly imbalanced class distribution (positive 74.77%, negative 21.69%, neutral 3.54%), macro-F1 is adopted as the primary metric alongside accuracy. Machine learning approaches substantially outperformed the lexicon-based approach: SVM achieved the highest macro-F1 of 70.43% (88.68% accuracy), far exceeding the InSet dictionary's macro-F1 of 27.67% (34.92% accuracy). A further key finding is an evaluation paradox: although Naïve Bayes attained the highest accuracy (89.91%), its macro-F1 was comparatively low (60.11%) due to bias toward the majority class, demonstrating how accuracy alone can mislead on imbalanced data. The InSet dictionary also performed poorly in the telemedicine domain, frequently misclassifying positive reviews as negative. Finally, comparing three imbalance-handling strategies (no handling, class-weight, and random oversampling) showed that cost-sensitive learning (class-weight) was most effective, raising the neutral-class F1-score from 26.56% to 35.8% without compromising majority-class performance.

INTRODUCTION

The digital transformation of the health sector has driven the rapid growth of telemedicine services in Indonesia, especially since the COVID-19 pandemic. Apps like Alodokter, Halodoc, and KlikDokter allow users to consult medical professionals, order medications, and access laboratory services without a physical visit. Previous studies have shown that telemedicine applications in Indonesia have experienced a surge in users and have generally gained positive public sentiment [1], [2]. As adoption increases, the number of reviews on application distribution platforms is growing rapidly, providing valuable feedback that developers can use to evaluate and improve service quality [1], [2].

The large volume of reviews makes manual analysis inefficient, so automated sentiment analysis methods are needed to classify reviews as positive, negative, or neutral. Sentiment analysis is a branch of natural language processing that is widely applied to evaluate user opinions about products or services [3]. In general, there are two main approaches. The lexicon-based approach determines the polarity of the text by matching words to a weighted sentiment dictionary, thereby eliminating the need

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

for labelled, domain-independent training data [4], [5]. For Indonesian, the InSet (Indonesian Sentiment Lexicon) dictionary, developed by Koto and Rahmaningtyas, is the most widely used resource, containing 3,609 positive and 6,609 negative words [6]. In contrast, the machine learning approach requires labelled data to train the model, but generally achieves higher accuracy because it can learn contextual language patterns [3], [5]. The Support Vector Machine (SVM) algorithm [7], Naïve Bayes [8], and Random Forest [9] are among the most commonly used algorithms for text classification, and all three are often compared in sentiment studies of digital applications in Indonesia [10], [11].

A number of studies have examined sentiment analysis on telemedicine applications. Afifah et al. applied XGBoost to Halodoc reviews and achieved over 96% accuracy [1]; Hendra and Fitriyani used the Naive Bayes Classifier to classify Halodoc reviews [11]; and Sidabutar and Juardi analysed public sentiment towards Halodoc as a telemedicine service [2]. In the broader context of Indonesian, advances in pre-trained models such as IndoBERT have also improved performance on various text classification tasks [12]. Cross-domain comparative studies show trade-offs between approaches: machine learning excels in accuracy but relies on trained data and is domain-dependent, while lexicon-based approaches are more efficient but susceptible to context variation, sarcasm, and negation [3], [13], [14]. Studies on health applications also highlight the importance of addressing class imbalances, for example, through the Synthetic Minority Oversampling Technique (SMOTE) [15].

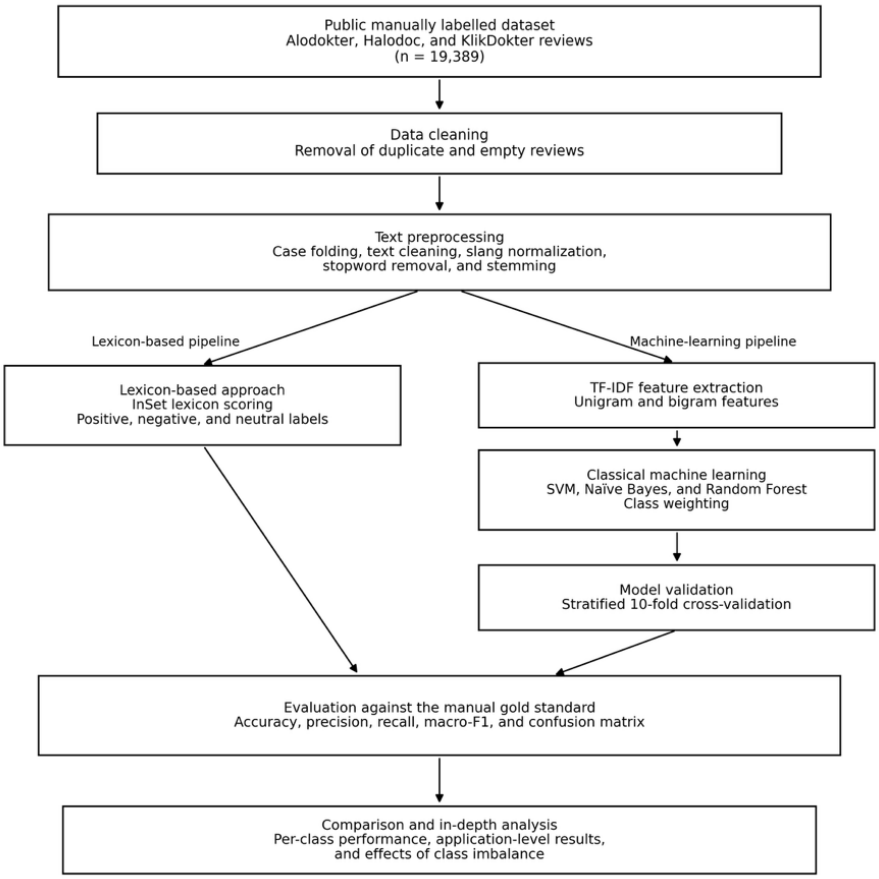
Although sentiment analysis in telemedicine applications has been extensively studied, several gaps remain inadequately addressed. First, the majority of studies focused on a single application, generally Halodoc, making it difficult to generalise the findings across service providers [1], [2], [11]. Second, many studies apply only one approach without directly comparing lexicon-based and machine learning on the same dataset, making it difficult to fairly assess the relative advantages of the two. Third, in studies based on lexicon labelling, machine learning models are commonly evaluated against the labels produced by the lexicon itself. This practice creates an optimistic evaluation bias because there is no reference to a truth independent of the method. Fourth, many studies ignore class imbalances and report only accuracy, which can be misleading when one class is highly dominant.

Departing from this gap, this research offers several interrelated novelties. This study conducted cross-platform comparisons of the three largest telemedicine applications in Indonesia, Alodokter, Halodoc, and KlikDokter, simultaneously, so that the findings are more generalizable than those from studies limited to one application. In addition, this study compares the lexicon-based approach (InSet) with three machine learning algorithms (SVM, Naïve Bayes, and Random Forest) head-to-head on identical datasets, ensuring a controlled, comparable comparison. Methodologically, this study evaluates against the gold standard of manual annotation by two annotators under the guidance of a psychologist, not against the lexical label, thereby eliminating the evaluation bias common in similar studies. Finally, the study explicitly addressed and reported class imbalances through class weighting and macro-F1, applied stratified 10-fold cross-validation to obtain more robust performance estimates, and compared three imbalance management strategies to assess their impact. Thus, the contribution of this study is to provide fair, comparative empirical evidence on the advantages and limitations of the two approaches in the domain of Indonesian-language telemedicine, as well as methodological recommendations for researchers and practitioners in choosing methods based on data availability and accuracy needs.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

METHODS

This study uses a quantitative experimental approach with a flow as shown in Figure 1.



Note. Both pipelines are evaluated using the same manually labelled gold-standardd dataset.

FIGURE 1. Research stage flow chart

Data Sources

Data are derived from the public dataset, A Weighted Sentiment Dataset for Indonesian Telemedicine Text Classification, available on Mendeley Data [16]. The dataset contains Indonesian-language user reviews from three telemedicine applications, Alodokter, Halodoc, and KlikDokter, collected from the Google Play Store. Each review is manually annotated by two annotators with guidance compiled by psychologists, including platform attributes, app_name, rating, review, sentiment (positive/negative/neutral), and weight (review complexity). This study uses review attributes as input and sentiment as the gold-standard label.

Cleaning and Preprocessing

From the initial dataset of 19,680 rows, blank and duplicate reviews were removed, yielding 19,527 unique reviews. After preprocessing, reviews that become empty (containing only stopwords or non-alphabetic characters) are issued, leaving 19,389 reviews for modelling with an average of 6.95 tokens per review. The preprocessing stage includes case folding (converting all letters into lowercase letters); cleaning (removing URLs, numbers, punctuation, emojis, and nonalphabetic characters, and normalizing the elongation of letters such as “baaguss” to “bagus”); normalization of non-standard words/slang into standard forms using normalization dictionaries; stopword removal to remove words without significant sentimental meaning; and stemming to change words to basic forms using a confix-stripping approach that is in accordance with the morphology of the Indonesian language [17].

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Lexicon-Based (InSet) Approach

The sentiment score of each review is calculated as the number of word weights that match the InSet dictionary [5]:

$$Score = \sum_{i=1}^{n} w_i \quad (1)$$

where *wi* is the polarity weight of the i-th word. A review is categorised as positive if $Score > 0$, negative if $Score < 0$, and neutral if $Score = 0$. This approach has been widely used in studies of Indonesian-language sentiment [18].

Machine Learning Feature Extraction and Classification

The text is converted into a numerical representation using the Term Frequency–Inverse Document Frequency (TF-IDF) with a combination of unigrams and bigrams, which effectively gives weight to discriminatory words while suppressing common words [19]:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log \frac{N}{DF(t)} \quad (2)$$

Three algorithms were used as comparators: SVM [7], which is effective in high-dimensional feature space, Naïve Bayes [8] as a probabilistic baseline, and Random Forest [9] as a tree-based ensemble. To address class imbalances, class-weighting parameters are applied to SVM and RF. The implementation uses the scikit-learn library [20].

Evaluation with Cross-Validation

The machine learning model will be evaluated using stratified 10-fold cross-validation to ensure that performance estimates do not depend on a single random data split [21]. Performance was measured by accuracy, precision, recall, and F1-score based on the confusion matrix [22]. Because the class distribution is highly imbalanced, the macro-averaged F1-score is used as the primary metric to ensure the performance of the minority (neutral) class is equally represented [22]. In addition, to explicitly assess the impact of imbalances, three strategies were compared in the best model: (a) no treatment, (b) class-weight (cost-sensitive learning), and random oversampling of minority classes. To avoid data leakage, the entire process of oversampling and TF-IDF feature extraction was performed only on the training data for each cross-validation fold, rather than on the entire dataset before partitioning [15].

RESULTS AND DISCUSSION

Descriptive Statistics and Sentiment Distribution

Of the 19,389 reviews, the distribution per application is: Alodokter 7,483, Halodoc 7,200, and KlikDokter 4,706. The distribution of sentiment labels is very unbalanced: positive 14,497 (74.77%), negative 4,205 (21.69%), and neutral only 687 (3.54%). Figure 2 shows the distribution per application.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

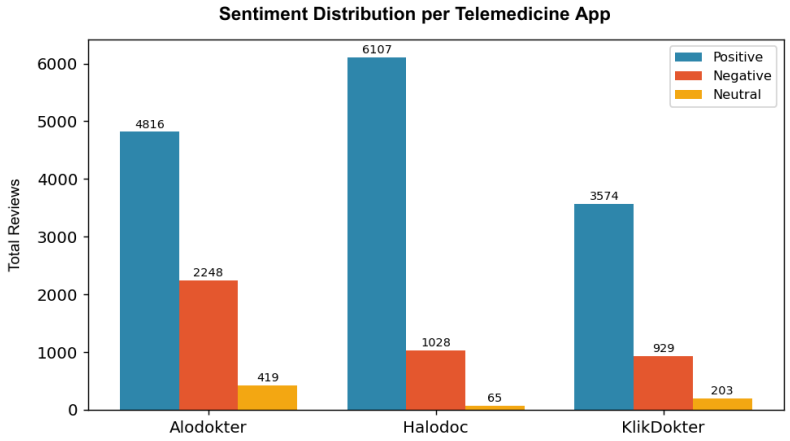


FIGURE 2. Sentiment distribution for each telemedicine app

The dominance of positive sentiment is consistent across all three applications, indicating relatively high user satisfaction with core services and aligning with findings from previous studies on telemedicine applications in Indonesia [1], [2]. However, the very small proportion of neutral classes, especially on Halodoc, which has only 65 neutral reviews, poses a classification challenge in itself. This extreme imbalance is the main reason accuracy alone is inadequate as a performance measure and justifies using macro-F1 as a key metric in this study.

Performance of the Lexicon-Based Approach (InSet)

The InSet Dictionary performs poorly against the gold standard: accuracy is only 34.92%, with macro-F1 at 27.67%. The confusion matrix in Figure 5 reveals the root of the problem.

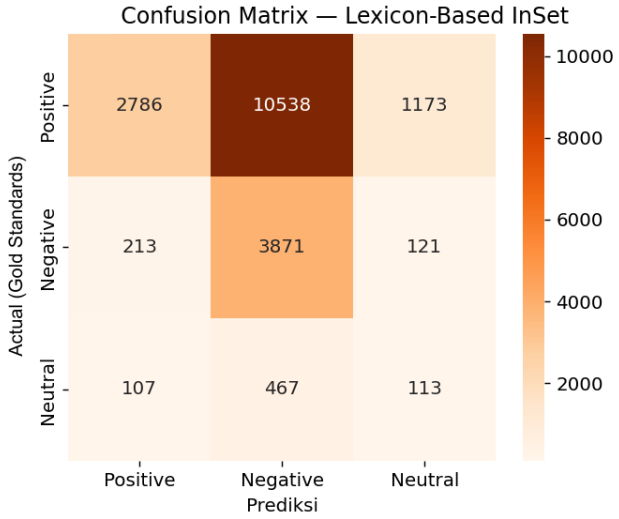


FIGURE 3. Confusion matrix in the set lexicon-based approach

The most striking finding was severe negative bias. Of the 14,497 reviews that were actually positive, InSet labelled 10,538 of them (72.7%) as negative. Overall, InSet classified 76.7% of all reviews as negative, when the actual proportion of negative reviews was only 21.69%. As a result, the F1 score per class is low (positive: 31.7%, negative: 40.6%, neutral: 10.8%).

At least three factors explain this failure. First, domain mismatch: InSet is built from a common tweet corpus [6], so many healthcare terms and app-specific phrases are underrepresented or have undue weight in the context of telemedicine. Second, the inequality of the dictionary: the number of negative words of the InSet (6,609) is almost twice as many as positive words (3,609), so the accumulated score tends to lean towards the negative, especially in long reviews that contain a lot of words. Third, the absence of context handling and negation: the word weight addition approach does not recognise

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

polarity reversals, so expressions such as “not disappointing” or “the queue is not long” are calculated negatively because they contain negatively weighted words. These findings are important because they contradict the common perception, also reported in some studies, that ready-made sentiment dictionaries can be applied directly across domains. The contrast between InSet’s poor performance here and the high accuracy reported in machine learning-based studies for Halodoc [1], [11] confirms that the advantages of a method depend heavily on domain suitability and evaluation design; general dictionaries require domain adaptations before they are suitable for use in telemedicine.

Machine Learning Approach Performance (10-Fold CV)

The results of the stratified 10-fold cross-validation for the three algorithms are presented in Table 1 (mean ± standard deviation).

TABLE 1. Machine learning model performance (10-fold CV)

Algorithm	Accuracy (%)	Macro Accuracy (%)	Recall macro (%)	Macro - F1 (%)	Weighted - F1 (%)
Naive Bayes	89.91 ± 0.63	73.51	59.53	60.11 ± 1.23	88.28
Random Forest	86.23 ± 0.83	65.85	68.05	66.33 ± 1.72	86.84
SVM	88.68 ± 0.65	69.51	71.67	70.43 ± 1.12	88.91

The results of the stratified 10-fold cross-validation for the three algorithms are presented in Table 1 (mean ± standard deviation). The small inter-fold standard deviation (≤ 1.72%) indicated consistent performance and was insensitive to specific data sharing, a key advantage of cross-validation over the single-split schemes widely used in previous studies. Based on macro-F1, SVM is the best model (70.43%), followed by Random Forest (66.33%) and Naïve Bayes (60.11%). The advantages of SVM align with its effectiveness in the high-dimensional, sparse TF-IDF feature space and are consistent with findings from comparative studies of NB and SVM in other digital applications in Indonesia [11], [15].

The small inter-fold standard deviation (≤ 1.72%) indicated consistent performance and was insensitive to specific data sharing—a key advantage of cross-validation over the single-split schemes widely used in previous studies. Based on macro-F1, SVM is the best model (70.43%), followed by Random Forest (66.33%) and Naïve Bayes (60.11%). The advantages of SVM align with its effectiveness in the high-dimensional, sparse TF-IDF feature space and are consistent with findings from comparative studies of NB and SVM in other digital applications in Indonesia [10], [15].

Thorough Comparison and Accuracy Paradox vs Macro-F1

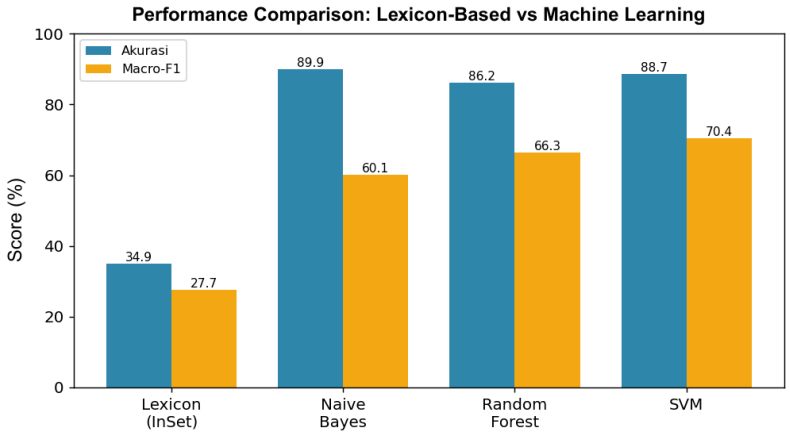


FIGURE 4. Performance comparison: lexicon-based vs. machine learning

Two important findings emerged. First, the machine learning approach has a landslide advantage over the lexicon-based approach: the difference between the macro-F1 scores of the SVM and InSet

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

reaches 42.8 percentage points. This reinforces a consistent pattern reported in the cross-domain literature that models trained on domain-specific data outperform static dictionaries [3], [14]. Second, and more methodologically instructive, is the paradox of accuracy vs macro-F1. Naïve Bayes recorded the highest accuracy (89.91%) but the lowest macro-F1 among ML models (60.11%). This happens because NB, without class weight adjustments, tends to predict the majority class (positive), so that the macro recall falls (59.53%). The model "plays it safe" by guessing the dominant class and still obtaining high accuracy. On the other hand, the SVM with class weighting achieved the best class balance (macro recall: 71.67%). These findings serve as a methodological warning: in unbalanced data, accuracy can be misleading, and macro-F1 reporting becomes crucial—a practice that is still often overlooked in similar sentiment studies.

Compared with previous studies, the SVM accuracy in this study (88.68%) was within a reasonable range. Some studies reported higher accuracy for example, XGBoost was above 96% in Halodoc reviews [1] and Logistic Regression/SVM was around 92% in healthcare applications [15] but direct comparisons need to be cautious due to design differences: most of these studies used binary (positive/negative) classifications on a single application, while this study dealt with three class classifications (including difficult neutrals) across three applications at once. The addition of neutral classes and cross-platform diversity inherently lowers accuracy while making results more realistic and generalizable. In other words, the slightly lower accuracy figures here reflect more difficult tasks and more honest evaluations.

Per-Class Analysis and SVM Matrix Confusion

TABLE 2. Performance each class: SVM vs Lexicon-Based (InSet)

Classes	Support	SVM Precision (%)	SVM Recall (%)	SVM F1 (%)	InSet F1 (%)
Positive	14,497	94.1	92.8	93.4	31.7
Negatives	4,205	81.5	82.7	82.1	40.6
Neutral	687	32.8	39.6	35.9	10.8

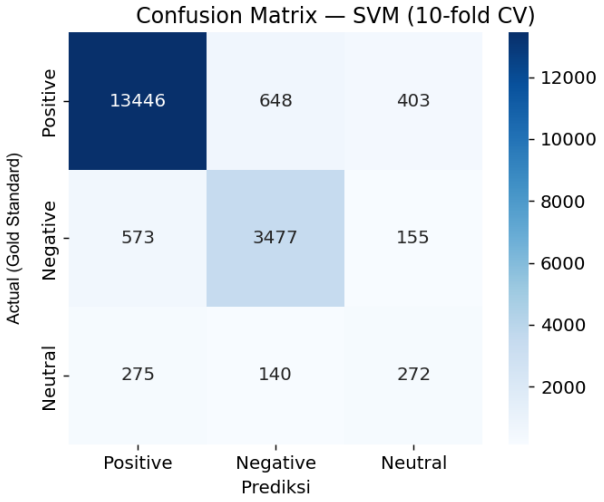


FIGURE 5. Best model confusion matrix (SVM, 10-fold CV)

SVM was excellent for the positive class (F1 93.4%, recall 92.8%) and the negative class (F1 82.1%, recall 82.7%), but performed poorly on the neutral class (F1 35.9%, recall 39.6%). Interestingly, although negative is a minority class (4,205 reviews), the model still detects it well. This can be explained by linguistic characteristics: negative reviews tend to have a distinctive and discriminatory vocabulary of complaints (e.g. "can't list", "old", "error", "expensive"), so that TF-IDF with bigrams is able to capture these strong markers. In contrast, the neutral class is the weakest point for two mutually reinforcing reasons: the sample size is very small (687 reviews, 3.54%), and semantically neutral reviews are ambiguous and often overlap with brief positive reviews (e.g. "okay", "not bad"). The confusion matrix shows that most errors in the neutral class are classified as positive rather than negative. However, SVM's neutral performance (35.9%) remained far outpacing InSet (10.8%),

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

suggesting that pattern-based learning remained superior to static dictionaries in even the most difficult classrooms.

Application Analytics (Cross-Platform)

The advantages of the multi-application dataset enable cross-platform analysis (Table 3) using SVM predictions.

TABLE 3. SVM and Lexicon performance for each application

Application	SVM Accuracy (%)	SVM Macro - F1 (%)	Inset Accuracy (%)
Alodokter	87.51	72.12	42.07
Halodoc	92.85	68.42	30.69
ClickDoctor	84.19	66.55	30.0

An interesting pattern emerged: Halodoc achieved the highest accuracy (92.85%), but its macro-F1 was lower (68.42%) than Alodokter's (72.12%). Again, this reflects the influence of class distribution: Halodoc reviews are mostly positive, with neutral classes almost absent (65 reviews), so accuracy is high, but macro-F1 is depressed by the weakness of minority classes. In contrast, Alodokter has a relatively more balanced class distribution (4,816 positive, 2,248 negative, 419 neutral), so its macro-F1 is at its best. These findings reinforce the argument that metric selection should account for the specific class distribution of each platform. Consistent across applications is SVM's landslide advantage over InSet, which reinforces the generalisation of cross-platform findings and shows that research conclusions do not depend on the characteristics of one particular application.

Handling Class Imbalance

To assess the impact of the imbalance directly, three strategies were tested on the best model (SVM) with the same cross-validation scheme (Table 4, Figure 6). All resampling is done leakage-free and only on the training data in each fold.

TABLE 4. Comparison of class imbalance management strategies (SVM, 10-fold CV)

Strategy	Accuracy (%)	Macro - F1 (%)	F1 Positive (%)	F1 Negative (%)	F1 Neutral (%)
No Handling	89.53	67.38	93.74	81.84	26.56
Class - Weight	88.68	70.43	93.4	82.09	35.8
Random Oversampling	87.25	67.16	92.4	80.51	28.57

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

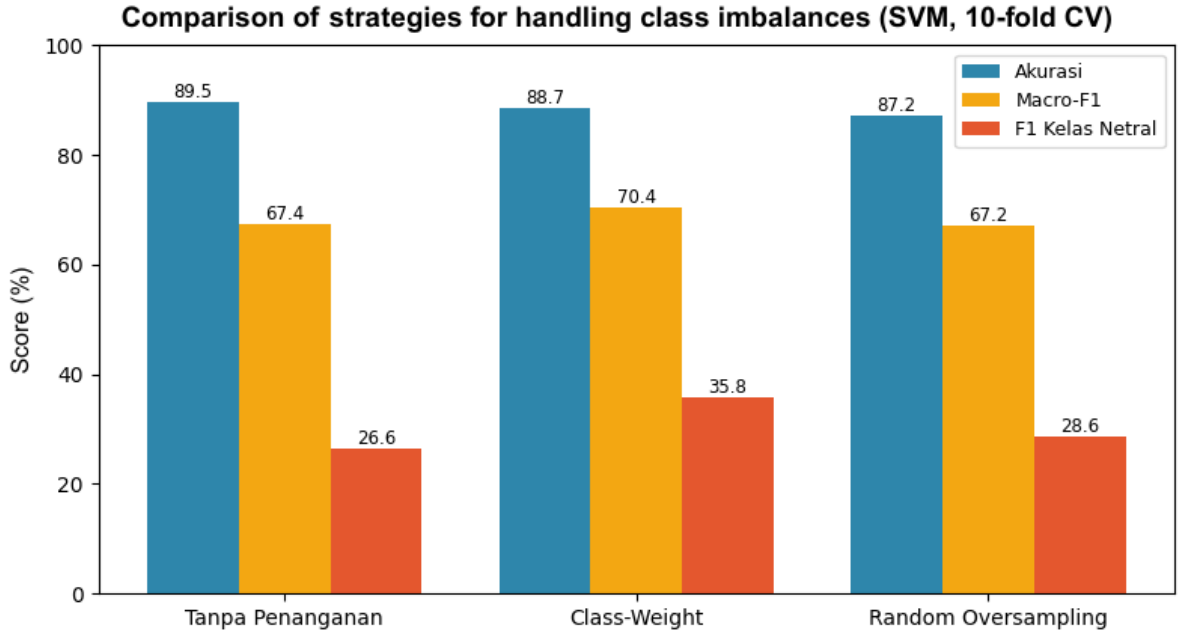


FIGURE 6. Comparison of strategies for handling class imbalances

These results provide three important lessons. Firstly, without handling, the model achieves the highest accuracy (89.53%) but at the expense of the minority class: the F1 is only 26.56%, which is neutral. Second, class-weights proved to be the most effective, raising neutral F1 substantially from 26.56% to 35.8% (up 9.2 points) and macro-F1 from 67.38% to 70.43%, with a negligible drop in accuracy (0.85 points). Third, random oversampling was less effective: the improvement for the neutral class was small (28.57%), while macro-F1 and accuracy decreased.

The finding that oversampling does not outperform cost-sensitive learning is in line with the characteristics of text representation. In sparse and high-dimensional TF-IDFs, duplicate minority samples (random oversampling) tend to cause overfitting to the same pattern without adding diversity to the information, so the model loses generalisation. It also adds nuance to studies that apply SMOTE to sentiment analysis in health applications [15]: although SMOTE produces synthetic samples (not just duplicates), its effectiveness in the text feature space still needs to be tested carefully and is not always superior to cost-sensitive learning. Therefore, the class-weight approach was chosen as the main configuration in this study and served as the basis for all results reported in the previous sub-chapter. It is also important to emphasise that resampling is applied only within the training fold; Applying it before data sharing will cause the synthetic sample to leak into the test data and artificially inflate accuracy, a mistake that is still common in the literature

Error Analysis and Limitations

Qualitative analysis of misclassification reveals several consistent patterns. Sarcasm and irony, such as “steady, list can’t be done”, are difficult to recognise in either the lexicon and the bag-of-words model, because the actual polarity is at odds with the words of its author. A mixture of languages (Indonesian–English–regional languages) and non-standard abbreviations not covered in the normalisation dictionary or InSet reduces accuracy, especially in the lexicon-based approach. In addition, the neutral class is the largest source of error due to its data scarcity and semantic ambiguity.

This research has several limitations that, at the same time, open up opportunities for further research. First, TF-IDF’s feature representation is a bag-of-words and does not capture word context or order; transformer-based contextual models such as IndoBERT [12] have the potential to address the sarcasm and ambiguity that are the main sources of error. Second, the neutral class imbalance has not been fully resolved, even with class weighting implemented; other techniques, such as SMOTE text variants or data augmentation, can be carefully explored [15]. Third, the InSet dictionary is used as is, without domain adaptation. Preparing an expanded lexicon could significantly improve the performance

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

of the lexicon-based approach. Fourth, the stemming in this study uses a mild confix-stripping approach; using a full Stemmer could slightly alter the outcome.

Implications

These findings provide clear practical guidance. When labelled data is available, and inter-class fairness is a priority, TF-IDF-based SVMs with class weighting are the strongest and most computationally efficient option. InSet's lexicon-based approach, as it is, is not recommended for the telemedicine domain without adaptation, given its severe negative bias. For telemedicine app developers, mapping complaints to negative classes that have proven reliably detectable can serve as a basis for prioritising service improvements, such as the registration process, doctor consultation queues, and payment features. For the research community, this study emphasises the importance of evaluating independent gold standards, macro-F1 reporting on unbalanced data, and implementing data leak-free resampling.

CONCLUSION

This study compared the lexicon-based (InSet) and machine learning (SVM, Naïve Bayes, Random Forest based on TF-IDF) approaches for sentiment classification of 19,389 user reviews of three telemedicine applications in Indonesia (Alodokter, Halodoc, KlikDokter), evaluated against the gold standard of manual annotation with stratified 10-fold cross-validation. The results showed that SVM performed best, with a macro-F1 of 70.43% (88.68% accuracy), a landslide victory over InSet's lexicon-based approach, which achieved only 27.67% (34.92% accuracy). The InSet Dictionary, as it is, has proven to perform poorly in the telemedicine domain due to its bias in labelling positive reviews as negative.

The study's main methodological contributions include cross-platform comparisons that reinforce the generalization of findings; evaluation of the Gold Standard manual that removes the evaluation bias based on lexical labels; affirmation that misleading accuracy in data is not balanced so macroF1 is mandatory to report; and a comparison of imbalance management strategies that showed cost-sensitive learning (class-weight) was more effective than random oversampling for the classification of the text in this data. Further research is suggested to apply contextual models such as IndoBERT to address sarcasm and ambiguity, explore advanced resampling techniques for neutral classes, and develop a lexicon specific to the telemedicine domain to improve the lexicon-based approach.

ACKNOWLEDGEMENTS

The author would like to thank for Institut Teknologi, Sains dan Kesehatan RS dr. Soepraoen Kesdam V/Brw and Politeknik Negeri Malang their support during the research, as well as the compilers of the A Weighted Sentiment Dataset for Indonesian Telemedicine Text Classification, who have made the data available openly.

REFERENCES

[1] K. Afifah, I. N. Yulita, I. Sarathan, B. Data, and U. Padjadjaran, "Sentiment Analysis on Telemedicine App Reviews using XGBoost Classifier," 2021.
[2] T. Grace, W. Margaretha, D. Juardi, U. S. Karawang, and T. Timur, "TELEMEDICINE DI INDONESIA TERHADAP," vol. 13, no. 1, 2025.
[3] S. J. Mahajani, S. Srivastava, and A. F. Smeaton, "A Comparison of Lexicon-Based and ML-Based Sentiment Analysis : Are There Outlier Words ?".
[4] V. Bonta, N. Kumaresh, and N. Janardhan, "A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," vol. 8, no. March, pp. 1–6, 2019.
[5] A. M. V. D. V. Id and E. Bleich, "The advantages of lexicon-based sentiment analysis in an age of machine learning," pp. 1–19, 2025, doi: 10.1371/journal.pone.0313092.
[6] F. Koto, "InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," pp. 391–394, 2017.
[7] L. Saitta, "Support-Vector Networks," vol. 297, pp. 273–297, 1995.
[8] A. Mccallum, "A Comparison of Event Models for Naive Bayes Text Classification," 1997.
[9] L. E. O. Breiman, "Random Forests," pp. 5–32, 2001.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

[10] D. A. Kristiyanti, D. A. Putri, and E. Indrayuni, “E-Wallet Sentiment Analysis Using Naïve Bayes and Support Vector Machine Algorithm E-Wallet Sentiment Analysis Using Na ¨ ıve Bayes and Support Vector Machine Algorithm,” 2020, doi: 10.1088/1742-6596/1641/1/012079.

[11] A. Hendra, “Analisis Sentimen Review Halodoc Menggunakan Nai¨ve Bayes Classifier,” vol. 6, no. 2, pp. 78–89, 2021.

[12] T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” 2020.

[13] E. Lunando and A. Purwarianti, “Indonesian Social Media Sentiment Analysis with Sarcasm Detection,” 2013.

[14] R. Catelli, S. Pelosi, C. Comito, C. Pizzuti, and M. Esposito, “Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy,” *Comput. Biol. Med.*, vol. 158, no. April, p. 106876, 2023, doi: 10.1016/j.compbiomed.2023.106876.

[15] R. Pasaribu and I. A. G. S. Putra, “Analisis sentimen ulasan aplikasi dengan Multinomial Naïve Bayes, Logistic Regression, dan SVM,” *J. Nas. Teknol. Inf. dan Apl.*, vol. 4, no. 1, pp. 63–72, 2025, doi: 10.24843/JNATIA.2025.v04.i01.p08.

[16] and E. W. A. F. W. Athallah, I. C. Ulumudin, R. Sutoyo, “A weighted sentiment dataset for Indonesian telemedicine text classification,” *Mendeley Data, V2*, 2026, doi: 10.17632/rmmmf2dp8d.2.

[17] M. Adriani, J. Asian, B. Nazief, and H. E. Williams, “Stemming Indonesian : A Confi x-Stripping Approach,” vol. 6, no. 4, pp. 1–33, 2007, doi: 10.1145/1316457.1316459.

[18] D. Musfiroh, U. Khaira, P. Eko, P. Utomo, and T. Suratno, “Sentiment Analysis of Online Lectures in Indonesia from Twitter Dataset Using InSet Lexicon Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon,” vol. 1, no. April, pp. 24–33, 2021.

[19] N. Arifin, U. Enri, and N. Sulistiyowati, “PENERAPAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DENGAN TF-IDF N-GRAM UNTUK TEXT CLASSIFICATION,” vol. 6, no. 2, 2021.

[20] F. Pedregosa, R. Weiss, and M. Brucher, “Scikit-learn : Machine Learning in Python,” vol. 12, pp. 2825–2830, 2011.

[21] R. Kohavi and S. Edu, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and M o d e l Selection,” pp. 1137–1143, 1993.

[22] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.