JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.8 No.2 May 2026: 205-214
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v8i2.16081

Comprehensive Analysis of TF-IDF and Word2Vec-Based Text Feature Representation on Support Vector Machine Performance in Sentiment Analysis of Telemedicine Reviews

Citra Ayu Tri Lestari[1*], Ahsanun Naseh Khudori[2], Risqy Siwi Pradini[3]

[1-3]Department of Informatics, Faculty of Science and Technology,
ITSK RS dr. Soepraoen Malang,
Malang - Jawa Timur, Indonesia 65147
citraayutrilestari353@gmail.com; ahsanunnaseh@itsk-soepraoen.ac.id;
risqypradini@itsk-soepraoen.ac.id

[*]Correspondence: citraayutrilestari353@gmail.com

Abstract *– User reviews of telemedicine applications provide valuable insights into user satisfaction and service quality. Accurate sentiment analysis of these reviews depends greatly on the text representation method used before classification. Although TF-IDF and Word2Vec have been widely applied in sentiment analysis, their comparative performance on Indonesian-language telemedicine reviews with a highly imbalanced sentiment distribution has not been sufficiently investigated. So, in order to analyze the sentiment of Halodoc application evaluations using the Support Vector Machine (SVM) classifier, this study analyzes the performance of TF-IDF and Word2Vec, two text representation approaches. The Google Play Store was combined from January 2024 to October 2025 for a grand total of 10,443 reviews. After implementing user-rated automatic sentiment labeling, 6,574 valid reviews were obtained using Indonesian-specific preprocessing that included case folding, cleaning, stopword removal, and stemming. Stratified 5-fold cross-validation using recall, accuracy, precision, and F1-score metrics was used to test the models. Word2Vec only managed 75% accuracy, 78% precision, 79% recall, and 78% F1-score in the experiments, but TF-IDF achieved 90% accuracy, 89% precision, 90% recall, and 89% F1-score. TF-IDF outperformed Word2Vec because frequency-based features are more effective in representing short review texts, whereas Word2Vec generally requires a substantially larger corpus to learn robust semantic representations. These findings provide empirical evidence for selecting appropriate text representation methods in Indonesian telemedicine sentiment analysis and contribute practical guidance for developing more accurate digital healthcare opinion mining systems.*

Keywords: *Sentiment Analysis; Telemedicine; TF-IDF; Word2Vec; Support Vector Machine*

I. INTRODUCTION

Telemedicine applications are part of the digital health service transformation that has developed in recent years. This service allows people to access healthcare more quickly, conveniently, and efficiently. It allows patients to receive medical consultations without the need for face-to-face meetings. (Kichloo et al., 2020). Telemedicine has also been proven to increase access to health services, especially for people who have limited access to medical personnel. (Mahdavi et al., 2025). Since the COVID-19 pandemic, the high public interest in digital health services has led to an increase in the use of telemedicine. (Alexandra et al., 2021). One of the most popular telemedicine services in Indonesia is the Halodoc application with a user percentage of 46.5% (Farah et al., 2023). The high number of users of this application shows that telemedicine services have become an important part of the community's need for digital health services (Wijaya et al., 2022).

The increase in users of telemedicine applications has caused the number of user reviews on digital platforms to increase. The review contains user experiences, criticisms, and assessments of the quality of the application's services (Edi Sutoyo & Muhammad Cekas Permana, 2025). These reviews are a direct representation of user reviews and real-life problems experienced by users. If these reviews are not analyzed properly, application developers risk losing important information that can be used for system improvements. One approach that can be used to analyze these reviews is Natural Language Processing (NLP). NLP is able to analyze language patterns in text data to identify user opinions about a service, so it can be used to identify public opinion about the quality of that service. However, the effectiveness of NLP in performing sentiment analysis is greatly influenced by the text

Copyright © 2026

representation method used during the analysis process (Monika et al., 2022).

Text representation is an important stage in NLP to overcome problems in natural language processing, namely by converting text into numerical form so that it can be processed mathematically by an algorithm (Mandhasiya et al., 2024). Each text representation method has different capabilities in capturing the meaning and context of the text being analyzed (Cahyani & Patasik, 2021a). Therefore, selecting the right text representation method is an important factor in improving sentiment analysis performance. The methods frequently used in sentiment analysis cases are Term Frequency–Inverse Document Frequency (TF-IDF) and Word to Vector (Word2Vec) (Tan et al., 2023).

These two methods have different characteristics in presenting text, so it is necessary to compare the performance to find out which method is more effective in the sentiment analysis process (Zhan, 2025). According (Royyan & Setiawan, 2022) explains that the application of feature expansion in Word2Vec can deepen the understanding of language context. This research emphasizes the importance of selecting a text representation method that suits the characteristics of the language. While previous studies, such as Cahyani and Patasik (2021a) and Rifaldy et al. (2025), have compared TF-IDF and Word2Vec for emotion classification and investment-platform review analysis, their findings cannot be directly generalized to the telemedicine domain. Telemedicine reviews contain domain-specific vocabulary, including health-related terminology, consultation experiences, medication names, and service-quality evaluations, which differ from those found in general consumer-review datasets. Moreover, previous studies have not explicitly examined the effect of highly imbalanced sentiment distributions on the comparative performance of these text representation methods.

Consequently, it remains unclear whether TF-IDF or Word2Vec is more suitable for sentiment analysis of Indonesian-language telemedicine reviews. To address this research gap, the present study evaluates both methods using a large real-world dataset of 10,443 Halodoc reviews and investigates their performance under a naturally imbalanced sentiment distribution. The findings provide empirical evidence to support the selection of appropriate text representation methods for healthcare-related sentiment analysis.

Although recent studies have introduced more advanced language representation models, TF-IDF and Word2Vec remain two of the most widely used text representation methods in sentiment analysis because they provide competitive performance with relatively low computational complexity (Zhang, 2025). In addition to selecting an appropriate text representation method, sentiment analysis also requires an effective classification algorithm. This study employs Support Vector Machine (SVM) because it has consistently demonstrated strong performance in text classification tasks involving high-dimensional feature spaces. SVM is particularly suitable for TF-IDF and Word2Vec representations, as it can effectively identify optimal decision boundaries while maintaining good generalization performance. Furthermore, using the same classifier for both text representation methods ensures that any observed performance differences are primarily attributable to the representation techniques rather than the classification algorithm itself (Zhou et al., 2024).

This work intends to fill the highlighted research gap by comparing the performance of two text representation approaches, TF-IDF and Word2Vec, in sentiment analysis of reviews of telemedicine applications written in Indonesian using Support Vector Machine (SVM) as the classifier. This study addresses the identified research gap by evaluating TF-IDF and Word2Vec under identical experimental settings using the same SVM classifier. The findings are expected to provide empirical evidence regarding the most suitable text representation method for sentiment analysis of Indonesian-language telemedicine reviews. Furthermore, this study contributes to the evaluation of text representation methods in the telemedicine domain and provides practical insights for understanding user opinions in digital healthcare services. By conducting a systematic comparison of both approaches, the results are expected to serve as a reference for researchers, system developers, and digital healthcare practitioners in developing more accurate sentiment analysis systems.

II. METHODS

Finding out how well TF-IDF and Word2Vec do at analyzing the tone of reviews written in Indonesian for telemedicine apps is the goal of this quantitative study that employs a comparative experimental design. The research stages were carried out systematically, starting from collecting review data to evaluating and comparing model performance.

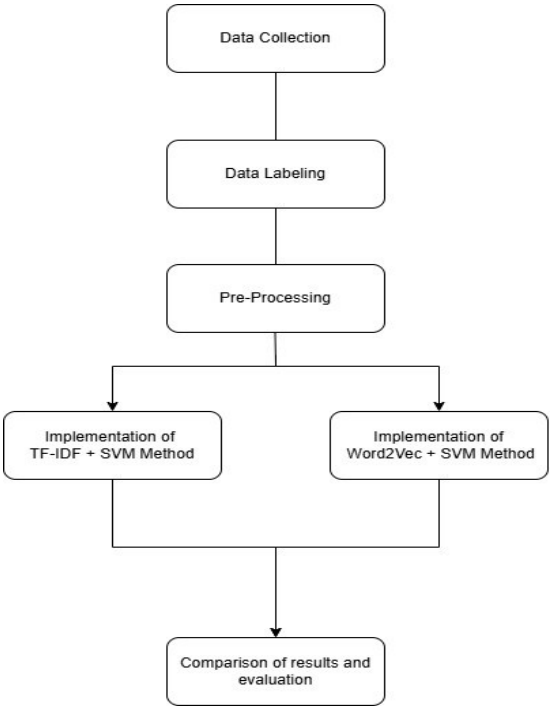


Figure 1. Research Flow Stages

The research stages are as shown in Figure 1. The research stages start from the data collection process, data labeling, pre-processing, implementation of text representation and comparison of evaluation results. This process uses Python 3.10 with Google Colab.

2.1 Data collection

Users' evaluations of the Halodoc app on the Google Play Store provided the secondary data used in this research. The reviews were gathered between January 1, 2024 and October 31, 2025 to ensure that the dataset reflected recent user experiences with the application. This period was selected to obtain recent and relevant user reviews. Review data were automatically extracted using a Python-based web scraping program that employed the Google Play Scraper library to retrieve publicly available user reviews in a structured format. Enabling the extraction of review data in a structured format (Imaduddin et al., n.d.). The collected dataset consists of four attributes: User Name, Review Content, Rating, and Review Time (at).

Table 1. Dataset Characteristics

No	Dataset Information	Description
1	Data Source	Google Play Store
2	Application	Halodoc
3	Collection Period	January 2024 – October 2025
4	Total Raw Reviews	10,443
5	Positive Reviews	9,353
6	Neutral Reviews	149
7	Negative Reviews	941
8	Reviews After Preprocessing	6,574

Dataset attributes utilized in this investigation are detailed in Table 1. We collected 10,443 evaluations written by actual Google Play Store customers throughout our data collecting period. A disproportionate number of positive evaluations dominate the sample, indicating a very unbalanced sentiment distribution. As a result, stratified 5-Fold Cross-Validation was used to provide a more trustworthy assessment of the classification models while maintaining the class distribution in each fold.

2.2 Data Labeling

Sentiment labels were assigned to each review to indicate whether the expressed opinion was positive, neutral, or negative before the classification stage. In this study, labeling was done automatically based on rating values in telemedicine review data. Positive, neutral, and negative are the three main ways that sentiments are classified. Within these groups, ratings of 1-2 indicate negative sentiment, ratings of 3 indicate neutral sentiment, and ratings of 4-5 indicate good sentiment (Fitriati et al., 2025). Using rating-based labeling enables a faster and more consistent annotation process than manually assigning sentiment labels to thousands of reviews. The labeling process is carried out using a rule-based approach, based on user rating values implemented using the Python programming language. The results of this automatic labeling are structured and labeled review data, so they can be used in the next process.

2.3 Pre-Processing

Pre-processing is the initial stage in data processing, which aims to change raw data into data that is ready to be used as input in the analysis process. This stage is carried out to ensure that the model used can recognize patterns more optimally through a pre-processing process using Python, as well as eliminating irrelevant elements in the data. Pre-processing was applied to the entire collected dataset. In this study, pre-processing was carried out in the following stages:

1) *Case Folding*
   At this stage, every alphabetical character was converted to lowercase so that identical words

with different capitalization would be represented consistently throughout the dataset (Faruqui et al., 2024). Case folding aims to ensure that the system does not differentiate words based on the use of capital and lowercase letters, thus being more consistent and preventing duplication in word representation.

2) *Cleaning*
Before feature extraction, each review was cleaned by removing textual elements that do not provide meaningful information for sentiment analysis, including URLs, special symbols, repeated punctuation, unnecessary numbers, and emoticons. Eliminating these non-informative components reduces textual noise and produces more consistent input data, allowing the subsequent feature extraction stage to focus on words that better represent the sentiment expressed in each review (Faruqui et al., 2024).

3) *Stemming*
Stemming is the process of removing common words that frequently appear but contribute little to the semantic meaning of a sentence, such as "yang," "dan," "atau," "di," and "dari" (Sarica & Luo, 2021). The goal of stopword removal is to free up model resources for more relevant phrases by removing words that don't contribute much to sentiment categorisation.
In this study, stemming was performed using the Sastrawi Stemmer library, which applies Indonesian morphological rules to convert inflected words into their root forms. This process reduces variations of words while preserving their semantic meaning, thereby improving feature consistency for the TF-IDF and Word2Vec representations.

4) *Stopword Removal*
The term "stopword removal" refers to the practice of eliminating overused words that do not advance the content of a phrase, like “yang”, “dan”, “atau”, “di”, “dari” (Sarica & Luo, 2021). The purpose of stopword removal is to eliminate words that provide limited information for sentiment classification, allowing the model to focus on more meaningful terms.

In this study, stopword removal was performed using the Indonesian stopword dictionary provided by the Sastrawi library. The Sastrawi stopword list was selected because it is specifically designed for Indonesian text processing and helps reduce noise while preserving important sentiment-related information for TF-IDF and Word2Vec feature extraction.

2.4 Implementation of Text Representation Method

In order to train a sentiment classification model to identify positive and negative evaluations of telemedicine applications, this research employs two text representation methods: TF-IDF and Word2vec. Each method has different goals, working methods, parameters and levels of representation depth. The application of the two text representation methods includes:

1) TF-IDF (*Term frequency- Inverse Document Frequency*)
One way to rank terms in a corpus is using TF-IDF, which takes into account both how often they appear in individual papers and how rare they are overall (Palanivinayagam et al., 2023). In addition, TF-IDF is a simple, stable model and has been proven to be effective on conventional machine learning models such as SVM, especially on high-dimensional data (Rianto et al., 2021) In this study, TF-IDF is used as a basic representation method that serves as a comparison for Word2Vec and is applied to a telemedicine review dataset that has gone through the pre-processing stage. The TF-IDF implementation process is carried out through the tokenization and word weighting stages based on TF-IDF similarities (Gretya Sinaga & Pribadi, 2026). In addition, model evaluation was carried out using the 5-Fold Cross Validation method to test the stability of model performance and reduce the risk of overfitting. The TF-IDF representation was configured using max_features = 5000 and ngram_range = (1,2), which are commonly adopted parameter settings in sentiment analysis studies. These parameters were kept fixed throughout the experiments to ensure a fair comparison with the Word2Vec representation under identical experimental conditions. The implementation process that will be carried out includes tokenization with TF-IDF equations:

$$TF = \frac{number\ of\ terms\ in\ the\ docks}{total\ number\ of\ terms\ in\ the\ docks} \quad (1)$$

To find out how often a word (term) appears in a text, one might use the Term Frequency (TF) approach, which is defined by Equation (1). The Inverse Document Frequency (IDF) and Term Frequency (TF) numbers are combined to make this approach function. While IDF determines how unique a word is in a set of texts, TF counts how often a word appears in a document. The combination of these two methods produces word weights that can represent important information in a document.

The TF value is calculated from the number of occurrences of a term and compared to the total of all terms in the document. So, the more frequently a word appears in a document, the greater the TF value will be. The TF calculation can be expressed in equation (1).

$$IDF = log_2 \frac{total\ number\ of\ docks}{number\ of\ docks\ in\ term} \quad (2)$$

Then, to lessen the impact of frequently occurring terms in the text, the IDF value is employed. Words that appear will have a small IDF value, while words that rarely appear will have a large IDF value because they are considered more specific in representing the contents of the document. The IDF calculation is shown in equation (2).

$$TF\ IDF = \text{TF} \times \text{IDF} \quad (3)$$

The TF-IDF weight is determined by multiplying the TF and IDF values, which are obtained in the previous step. A term's significance in describing the document's contents is proportional to its TF-IDF value. It is seen in the equation that the TF-IDF equation (3).

2) Word2Vecv

Depending on their context, Word2Vec can represent words as fixed-dimensional vectors. This method captures the semantic relationships between words that TF-IDF cannot handle. Word2Vec was chosen because it remains a strong baseline in modern embedding studies, especially in resource-constrained languages such as Indonesian. As a skip-gram architecture is successful in learning representations of uncommon words, this study's Word2Vec model was trained using gensim with the following parameters: vector_size = 200, window = 5, min_count = 1, using this technique. Each review was then converted into a vector using the average embedding method (Rahmanda & Setiawan, 2022).

To ensure that the Support Vector Machine model consistently performs well, we utilize the representation results as input and evaluate it using 5-fold cross-validation. According to the Skip-Gram architecture, the Word2Vec model was set up with the following parameters: vector_size = 200, window = 5, min_count = 1, and random_seed = 42. All studies used these same settings so that we could compare them fairly to the TF-IDF representation under the same experimental circumstances.

2.5 Classification Algorithm

The classification algorithm used in this study is Support Vector Machine (SVM). This algorithm is used because it is able to work effectively on high-dimensional text data, with TF-IDF and Word2Vec representation results, and has been proven to have stable performance on text classification tasks. (Suharman & Kamayani Sulaeman, 2025) By using SVM in the classification process, this study is consistent in seeing the performance differences between TF-IDF and Word2Vec. This approach ensures evaluation results and shows how effective each method is in representing text.

In this study, SVM receives input in the form of vectors from TF-IDF and Word2Vec. The use of a linear kernel was chosen because it is most suitable for text modeling and has been widely used in sentiment analysis research, including in research that combines TF-IDF and SVM (Rafly Gusmansyah & Hendrawan, 2025).

The SVM classifier was implemented using a linear kernel with the default regularization parameter (C = 1.0). The linear kernel was selected because it is computationally efficient and has consistently demonstrated strong performance for high-dimensional text classification tasks. The linear kernel is used in this study because of its simplicity: it can calculate the similarity level between features in two vectors using the dot product. This kernel is formulated as:

$$K(x_i, x_j) = x_i^T\ x_j \quad (4)$$

This approach is considered effective in the process of handling large-dimensional data such as TF-IDF and Word2Vec representations, so that the model is able to recognize patterns effectively without having to carry out a complicated feature transformation process.

2.6 Evaluation

Evaluation and comparison of sentiment analysis models were conducted to measure how effective the two text representation methods, namely TF-IDF and Word2vec, were in the task of sentiment classification of Indonesian-language telemedicine application reviews. This evaluation employed the same classification algorithm, namely support vector machine (SVM), with a 5-fold cross-validation scheme. This method is used to produce a more stable evaluation because all data is used alternately as training data and test data, so that the model is able to recognize data patterns better and reduce the risk of overfitting. Considering the substantial class imbalance in the dataset (89.56% positive, 9.01% negative, and 1.43% neutral), the 5-fold split was performed using a stratified scheme so that the proportion of each sentiment class remained consistent across folds. This study does not yet apply

explicit class-imbalance mitigation techniques (e.g., SMOTE or class-weighting); this limitation is acknowledged and is addressed further in the Error Analysis and Conclusion sections. To ensure the reproducibility of the experiments, all data splitting processes were performed using a fixed random seed (random_state = 42). Using a fixed random seed ensures that the same data partitions are generated across repeated experiments, thereby enabling fair and reproducible performance comparisons between the TF-IDF and Word2Vec representations.

Factors like accuracy, precision, recall, and F1-Score were used to measure the model's performance. Accuracy is used to measure the model's accuracy in classifying all data. Precision is used to measure how accurately the model predicts data in a particular class. Recall can measure the model's ability to find all data that belongs to a particular class. By the same token, F1-Score is helpful for gauging the accuracy and recall of a classification model, which is an important metric for balancing its performance. To further understand the model's sentiment class categorization findings, this study leverages both assessment metrics and a confusion matrix. At this stage, the SVM model is tested using both types of TF-IDF representation, as well as Word2vec. Next, the metric results will be compared to determine the performance differences between the two approaches. The performance of each model will be explained in the results and discussion section. This methodology section focuses solely on explaining the evaluation procedures used.

III. RESULTS AND DISCUSSION

3.1 Data Scraping and Labeling Results

Based on the data collection process using scraping techniques from the Google Play Store, 10,443 telemedicine user review data were obtained between January 1, 2024, and October 31, 2025. The data collection process was carried out using the Google Play Scraper library. Furthermore, an automatic labeling process was performed based on review ratings, resulting in the sentiment classification presented in Table 2. The sentiment classification is also visualized in the following diagram.

Table 2. Sentiment Classification Results

No	Sentiment Classificatior	Amount of Data	Percentage (%)
1.	Positive	9.353	89,56
2.	Negative	941	9,01
3.	Neutral	149	1,43
	Amount	10.443	100.00

Based on Table 2, positive sentiment dominated the classification results, accounting for 9,353 items, or 89.56% of all reviews. Meanwhile, negative sentiment accounted for 941 items (9.01%), while neutral sentiment accounted for only 149 items (1.43%). These findings indicate that most users gave a positive response to the Halodoc application. In addition, the large differences in the amount of data between categories indicate that the data distribution tends to be unbalanced. This imbalance has the potential to affect the performance of classification models, especially in identifying sentiments with relatively small amounts of data, such as neutral sentiment.

3.2 Pre-processing Results

In the pre-processing section, raw text data is transformed into cleaner and more structured data so that it can be used as input for machine learning models. The goal of this procedure is to make the text simpler and less noisy so that the classification model can handle it more easily. Based on the data processing results, the dataset after the pre-processing stage contained 6,574 reviews. This number has decreased compared to the initial data due to the process of deleting invalid data, such as empty reviews, duplicate data, and text that has no meaning.

Because the model's efficacy in sentiment categorization is dependent on the quality of the output data, this part is vital. Folding cases, cleaning, removing stop words, and stemming are all parts of the preprocessing procedure used in this research. In order to keep things uniform and prevent different representations of the same word, the case folding procedure changes all the letters to lowercase. The cleaning process also gets rid of symbols, URLs, punctuation, and numerals. The next step is to eliminate frequent terms like conjunctions and demonstratives that don't contribute much to the process of detecting mood. The final step is stemming, which involves breaking words down into their most fundamental forms in order to simplify the model's ability to detect text patterns and decrease word variability. To provide a clearer picture of the results of each pre-processing stage, an example of data transformation is shown in Table 3.

Table 3. Results of the Review Data Pre-Processing Stage

Original Text	Case Folding	Cleaning	Stop Word Removal	Steaming
A really useful app, Thank you, Halodoc	A really useful app – thank you, Halodoc	A really helpful app – thank you, Halodoc	The app is helpful – thank you, Halodoc	helpful app – thank you, Halodoc

Table 3 shows that each preprocessing stage gradually changes the review text. Case folding and cleaning are used to normalize the text, while stopword removal and stemming significantly simplify the sentence structure without losing the main meaning. This indicates that the preprocessing stages were effective in producing more representative data for sentiment classification.

3.3 Comparison of TF-IDF and Word2Vec using SVM

At the text representation stage, the pre-processing data is converted into numerical form so that it can be processed by the SVM classification algorithm. Word2Vec and TF-IDF are the two text representation algorithms used in this study. To capture both unigrams and bigrams, the TF-IDF representation is executed with the settings max_features = 5000 and ngram_range = (1,2). The use of bigrams allows the model to recognize simple contexts between words, such as "not helpful" or "very good", which plays an important role in determining sentiment. TF-IDF produces high-dimensional vectors that can increase the importance of each word in a document, making it very effective for use in SVM classification algorithms.

Meanwhile, Word2vec is used with Skip-Gram architecture with parameters vector_size = 200, window = 5, and min_count=1. Each review is represented as a vector using the average embedding method, averaging the vectors across all data within a single document. This approach allows the model to capture semantic relationships among words within a document. However, unlike TF-IDF, Word2Vec focuses more on the meaning of words than their frequency of occurrence.

Based on the test results using the SVM algorithm, the TF-IDF method produced better classification performance compared to Word2Vec. This is indicated by higher accuracy, precision, recall, and F1-score values in TF-IDF. Word frequency-based representations were deemed more effective in identifying sentiment patterns in telemedicine app reviews than word embedding-based representations. A likely contributing factor to Word2Vec's lower performance is the relatively small size of the training corpus. After pre-processing, the dataset consisted of only 6,574 reviews, which is considerably smaller than the corpus sizes typically required to train Word2Vec embeddings from scratch and obtain semantically robust word vectors. Word embeddings trained on limited data tend to produce noisier, less generalizable representations, particularly for infrequent words, which may explain why the embedding-based approach struggled to capture sentiment polarity as effectively as the simpler, frequency-based TF-IDF representation in this study.

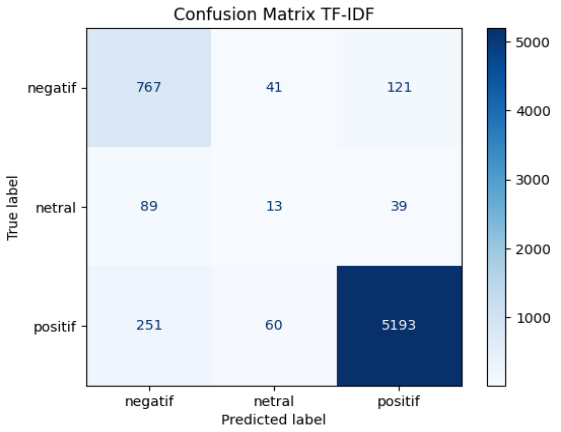


Figure 3. TF-IDF Confusion Matrix Results

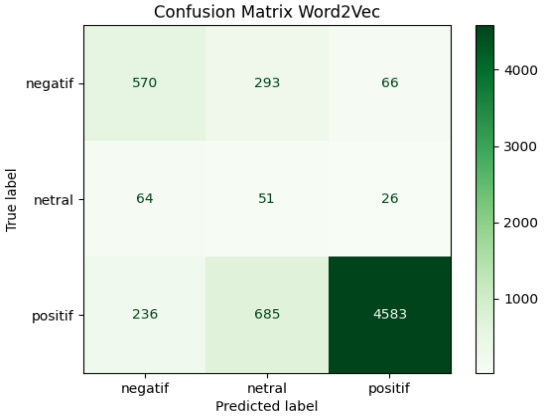


Figure 4. Word2Vec Confusion Matrix Results

Based on the confusion matrix in Figure 3 and Figure 4, the TF-IDF model shows a lower misclassification rate compared to Word2Vec. In TF-IDF, most of the data was successfully classified correctly, especially in the positive class with a total of 5193 correct predictions. Additionally, 767 negative data points were correctly predicted as negative. However, there were still some misclassifications, such as 121 negative data points being predicted as positive.

In contrast, in Word2Vec, there are more frequent misclassifications in the neutral class. This is evident from the 293 negative data points

predicted as neutral and the 685 positive data points also predicted as neutral. Although the Word2Vec model successfully classified 4,583 positive data points correctly, its misclassification rate was still higher than that of TF-IDF. This shows that Word2Vec has difficulty in distinguishing sentiment polarity.

3.4 Model Evaluation Results

To get less reliability on any one data distribution and more consistent findings, the stratified 5-fold cross-validation approach was used to conduct the assessment. Accuracy, precision, recall, and F1-score were some of the assessment measures used to quantify the model's performance. The categorization findings were further examined using a confusion matrix. The stratified scheme further strengthens the evaluation by maintaining the same proportion of positive, negative, and neutral reviews in each fold. Consequently, the reported performance metrics better reflect the models' generalization ability and are less affected by bias resulting from the highly imbalanced dataset.

1) TF-IDF Evaluation
   Based on the test results using the 5-Fold Cross Validation method, the SVM model with TF-IDF representation shows very good performance. The accuracy value obtained was around 0.90, with high and balanced precision, recall, and F1-score values. These results indicate that TF-IDF is capable of representing text effectively in sentiment classification tasks.

2) Word2vec Evaluation
   The SVM model with Word2Vec representation produces lower performance compared to TF-IDF, with an accuracy value of around 0.75. Although the precision value is relatively high, the recall value is lower, indicating that the model has not been able to optimally detect all data. This shows that Word2Vec is less effective for sentiment classification on short texts.

Table 4. Model Evaluation Results

Method	Accuracy	Precision	Recall	F1 Score
TF-IDF + SVM	0,90	0,89	0,90	0,89
Word2Vec + SVM	0,75	0,78	0,79	0,78

3.5 Comparative Discussion

Based on this comparison, the TF-IDF method demonstrated superior performance compared to Word2Vec across all evaluation metrics. This indicates that the word frequency-based approach is more effective for sentiment analysis of short telemedicine reviews than the embedding-based representation used in this study.

It should be noted that the comparison was conducted using fixed parameter settings for both TF-IDF and Word2Vec. Therefore, the reported performance reflects the effectiveness of the two text representation methods under identical experimental conditions rather than the optimal performance achievable through extensive hyperparameter optimization. Future studies are encouraged to investigate the influence of different parameter configurations using systematic optimization techniques, such as Grid Search or Bayesian Optimization, to further improve classification performance. Furthermore, this conclusion is based on a direct comparison of the average accuracy, precision, recall, and F1-score values obtained from stratified 5-fold cross-validation, without applying a formal statistical significance test (e.g., a paired t-test or McNemar's test). Future studies are encouraged to incorporate such statistical analyses to determine whether the observed performance differences are statistically significant. In addition, the metrics reported in Table 4 represent overall (weighted/macro-averaged) performance and are therefore influenced by the dominant positive class. A more thorough assessment of each model's capacity to categorize minority-class sentiment might be achieved by reporting per-class precision, recall, and F1-score, given that the neutral class comprises only 1.43% of the sample.

3.6 Model Error Analysis

Some classification errors occurred because the review texts contained ambiguous expressions, non-standard Indonesian words, abbreviations, and informal language that were difficult for the model to interpret correctly. In addition, the substantial class imbalance in the dataset also affected the model's performance, particularly in detecting the neutral class, which represented only 1.43% of the total data. As a result, the classifier tended to predict the majority positive class more frequently, leading to lower recall for minority classes. These findings indicate that both linguistic variability and class imbalance remain important challenges for sentiment classification of Indonesian telemedicine reviews. Future studies may reduce these errors by applying data balancing techniques and employing contextual embedding models, such as IndoBERT, which are better suited to capturing semantic information in informal Indonesian text.

IV. CONCLUSION

Based on the results of the research that has been conducted, it can be concluded that the text representation method can have a significant influence on the performance of the sentiment analysis model. This is indicated by the TF-IDF accuracy value, which reaches around 0.90, as well as the precision, recall, and F1-score values , which are more stable compared to Word2Vec, which only achieves an accuracy of around 0.75.

TF-IDF is considered more suitable for capturing the direct occurrence of words in the text, making it effective for use on review data that tends to be short. Meanwhile, Word2Vec, which relies on the context of word usage, hasn't yielded optimal results on this type of data. Furthermore, the imbalanced data distribution also impacts classification results, particularly for the neutral class, which is smaller in number than the other classes. This process causes the model to tend to be better at recognizing the dominant class.

For further research, it is recommended to consider data balancing techniques or the use of more complex representation methods to improve model performance, especially in classes with limited data amounts.

Before drawing any conclusions from this study, it is important to keep in mind that it has a number of limitations. First, the dataset was drawn exclusively from a single telemedicine application (Halodoc), so the results may not generalize directly to other telemedicine platforms. Second, sentiment labels were generated automatically from user rating scores rather than through manual annotation, which may introduce label noise since a numerical rating does not always align precisely with the sentiment expressed in the review text. Third, the dataset exhibits substantial class imbalance, with the neutral class comprising only 1.43% of the data; although a stratified 5-fold cross-validation scheme was used, no explicit imbalance-mitigation technique (e.g., SMOTE or class-weighting) was applied, so performance on the minority class should be interpreted with caution. Future studies are encouraged to address these limitations through manual or hybrid labeling validation, multi-platform datasets, and imbalance-handling strategies.

AVAILABILITY DATA AND MATERIALS

The dataset used can be accessed via the link: *https://github.com/CITRALESTARI28/digital-health-sentiment-dataset*.

REFERENCES

Alexandra, S., Handayani, P. W., & Azzahro, F. (2021). Indonesian hospital telemedicine acceptance model: the influence of user behavior and technological dimensions. *Heliyon*, *7*(12). https://doi.org/10.1016/j.heliyon.2021.e08599

Cahyani, D. E., & Patasik, I. (2021a). Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, *10*(5), 2780–2788. https://doi.org/10.11591/eei.v10i5.3157

Edi Sutoyo, & Muhammad Cekas Permana. (2025). work contributes to enhancing telemedicine service quality and advancing Indonesian natural language processing research. *Elsevier Inc*. https://doi.org/10.17632/j8yjvsyr9p.1

Farah, F. R., Putri, A. W., & Rani, T. B. (2023). *Analysis of the Factors Influencing the Decision to Use Halodoc Telemedicine in the City of Bogor* [Analisis Pengaruh Keputusan Penggunaan Telemedicine Halodoc di Kota Bogor]. *Manajemen Kesehatan Indonesia*. https://doi.org/10.14710/jmki.11.2.2023.190-202

Faruqui, N., Thatoi, P., Choudhary, R., Roncevic, I., Alqahtani, H., Sarker, I. H., & Khanam, S. (2024). AI-Analyst: An AI-Assisted SDLC Analysis Framework for Business Cost Optimization. *IEEE Access*, *12*, 195188–195203. https://doi.org/10.1109/ACCESS.2024.3519423

Fitriati, L. R., Guntara, R. G., & Purwaamijaya, B. M. (2025). *Sentiment Analysis of GoPay User Reviews on the Google Play Store using the IndoELECTRA Model* [Analisis Sentimen Ulasan Pengguna GoPay di Google Play Store menggunakan Model IndoELECTRA]. *Jurnal Algoritma*, *22*(2). https://doi.org/10.33364/algoritma/v.22-2.3053

Gretya Sinaga, A., & Pribadi, O. (2026). Comprehensive Comparison of TF-IDF and Word2Vec in Product Sentiment Classification Using Machine Learning Models. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 10, Number 1). https://doi.org/10.30871/jaic.v10i1.11582

Imaduddin, H., Yusfida A'la, F., & Nugroho, Y. S. (n.d.). Sentiment Analysis in Indonesian

Healthcare Applications using IndoBERT Approach. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Number 8). Retrieved www.ijacsa.thesai.org. https://doi.org/10.14569/IJACSA.2023.0140813

Kichloo, A., Albosta, M., Dettloff, K., Wani, F., El-Amir, Z., Singh, J., Aljadah, M., Chakinala, R. C., Kanugula, A. K., Solanki, S., & Chugh, S. (2020). Telemedicine, the current COVID-19 pandemic and the future: a narrative review and perspectives moving forward in the USA. In *Family Medicine and Community Health* (Vol. 8, Number 3). BMJ Publishing Group. https://doi.org/10.1136/fmch-2020-000530

Mahdavi, S., Fekri, M., Mohammadi-Sarab, S., Mehmandoost, M., & Zarei, E. (2025). The use of telemedicine in family medicine: a scoping review. In *BMC Health Services Research* (Vol. 25, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s12913-025-12449-7

Mandhasiya, D. G., Murfi, H., & Bustamam, A. (2024). The hybrid of BERT and deep learning models for Indonesian sentiment analysis. *Indonesian Journal of Electrical Engineering and Computer Science*, *33*(1), 591–602. https://doi.org/10.11591/ijeecs.v33.i1.pp591-602

Monika, P., Kulkarni, C., Harish Kumar, N., Shruthi, S., & Vani, V. (2022). Machine learning approaches for sentiment analysis. *International Journal of Health Sciences*, 1286–1300. https://doi.org/10.53730/ijhs.v6ns4.6119

Palanivinayagam, A., El-Bayeh, C. Z., & Damaševičius, R. (2023). Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review. *Algorithms*, *16*(5). https://doi.org/10.3390/a16050236

Rahmanda, R., & Setiawan, E. B. (2022). Word2Vec on Sentiment Analysis with Synthetic Minority Oversampling Technique and Boosting Algorithm. *Jurnal RESTI*, *6*(4), 599–605. https://doi.org/10.29207/resti.v6i4.4186

Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00413-1

Rifaldy, F., Sibaroni, Y., & Prasetiyowati, S. S. (2025). Effectiveness of Word2Vec and TF-IDF in Sentiment Classification on Online Investment Platforms Using Support Vector Machine. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *10*(2), 863–874. https://doi.org/10.29100/jipi.v10i2.6055

Royyan, A. R., & Setiawan, E. B. (2022). Feature Expansion Word2Vec for Sentiment Analysis of Public Policy in Twitter. *Jurnal RESTI*, *6*(1), 78–84. https://doi.org/10.29207/resti.v6i1.3525

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLoS ONE*, *16*(8 August). https://doi.org/10.1371/journal.pone.0254937

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. In *Applied Sciences (Switzerland)* (Vol. 13, Number 7). MDPI. https://doi.org/10.3390/app13074550

Wijaya, J. H., Octavius, G. S., & Hwei, L. R. Y. (2022). A LITERATURE REVIEW OF TELEMEDICINE IN INDONESIA: PAST, PRESENT, AND FUTURE PROSPECTIVE. In *Indonesian Journal of Health Administration* (Vol. 10, Number 2, pp. 261–272). Airlangga University. https://doi.org/10.20473/jaki.v10i2.2022.261-272

Zhan, Z. (2025). Comparative Analysis of TF-IDF and Word2Vec in Sentiment Analysis: A Case of Food Reviews. *ITM Web of Conferences*, *70*, 02013. https://doi.org/10.1051/itmconf/20257002013

Zhang, L. (2025). Features extraction based on Naive Bayes algorithm and TF-IDF for news classification. *PLOS ONE*, *20*(7 July). https://doi.org/10.1371/journal.pone.0327347

Zhou, J., Ye, Z., Zhang, S., Geng, Z., Han, N., & Yang, T. (2024). Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data. *Heliyon*, *10*(16). https://doi.org/10.1016/j.heliyon.2024.e359