Comparative Sentiment Analysis on Mobile JKN Application Using Logistic Regression with SMOTE Based Statistical Feature Selection

Rafika Farkhul Awaliyah [1]*, Aria Hendrawan [2]*
* Department of Information Technology and Communication, Universitas Semarang
farkhulfika@gmail.com [1], ariahendrawan@usm.ac.id [2]

Article Info

Article history:

Received 2025-07-28
Revised 2025-10-17
Accepted 2025-11-05

Keyword:

Sentiment Analysis,
Public Health Application,
Logistic Regression,
TF-IDF,
Word2Vec,
Feature Selection,
Mobile JKN.

ABSTRACT

This study investigates public sentiment on the Mobile JKN application using Logistic Regression enhanced with SMOTE-based statistical feature selection. Unlike prior works that relied solely on conventional feature combinations such as TF-IDF or Word2Vec, this research performs a comparative evaluation of three statistical feature selection techniques: Recursive Feature Elimination (RFE), Chi-Square, and Mutual Information, under both TF-IDF and Word2Vec representations in a low-resource Indonesian language setting. The dataset consists of 2,382 user reviews from the Google Play Store, balanced using SMOTE to mitigate class imbalance. The best configuration, TF-IDF combined with Mutual Information, achieved an accuracy of 73.38% and an F1-score of 50%, indicating a moderate yet consistent performance. A confusion matrix-based error analysis revealed that most misclassifications occurred between neutral and negative classes due to semantic overlap. The relatively low F1-score highlights challenges in sentiment separability, while the superior performance of Mutual Information demonstrates its ability to capture discriminative linguistic features. The superior performance of Mutual Information is attributed to its ability to capture non-linear dependencies between features and sentiment labels, yielding richer discriminative information compared to Chi-Square or RFE. This research establishes a comparative methodological framework that integrates feature selection and data balancing techniques, providing interpretable sentiment classification insights for under-resourced language settings.

I. INTRODUCTION

In the current era of digital transformation, mobile technology has become a vital component of daily life. Based on data from Statistics Indonesia (BPS), approximately 67.88% of the Indonesian population owned mobile phones by 2022 [1], indicating a substantial reliance on mobile devices for various services. In response to this trend, BPJS Health launched the Mobile JKN application as a digital health platform designed to facilitate public access to healthcare services. This application offers features such as health insurance management, online queue reservations, and teleconsultation services [2]. Since its inception, Mobile JKN has garnered thousands of user reviews on the Google Play Store, which reflect a wide spectrum of user sentiments and perceptions regarding service performance [3].

Sentiment analysis has emerged as a valuable tool to systematically extract and categorize opinions from textual data, particularly in evaluating public responses to digital services [4]. In this study, sentiment analysis is employed to classify user reviews of the Mobile JKN application into three categories: positive, neutral, and negative [5]. To represent the textual data effectively, two prominent feature representation techniques are utilized Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec aimed at capturing both statistical relevance and semantic context [6].

To enhance model performance and reduce dimensional noise, this study applies three statistical feature selection techniques: Recursive Feature Elimination (RFE), Chi-Square, and Mutual Information [7]. These methods are used to identify the most informative features that contribute to accurate sentiment classification. The dataset comprises 2,382 user reviews obtained through web scraping from the Google Play Store between November 2017 and May 2025. Sentiment labels were assigned based on user rating scores, and the Synthetic Minority Oversampling Technique (SMOTE) was employed to address the issue of class imbalance within the dataset [8].

Several previous studies have demonstrated the effectiveness of machine learning techniques such as Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression in sentiment classification tasks. However, most prior research focused only on implementing single algorithms or feature representations without addressing data imbalance and feature redundancy issues. Another study by [9] Employed TF-IDF and Chi-Square in conjunction with the Naïve Bayes algorithm to classify Mobile JKN reviews. Another study by [10] utilized the SVM algorithm optimized with Particle Swarm Optimization (PSO) to improve classification accuracy. However, limited research has explored the optimal combination of feature representations and feature selection methods for sentiment analysis in Indonesian-language datasets. To address this limitation, the present study proposes a comparative and methodological framework that integrates two types of textual feature representations (TF-IDF and Word2Vec) with three statistical feature selection techniques (RFE, Chi-Square, and Mutual Information), using Logistic Regression as the core classification algorithm [11]. Logistic Regression was selected over more complex algorithms such as SVM or BERT due to its interpretability, computational efficiency, and robustness for small and imbalanced datasets. The methodological pipeline includes data acquisition, text preprocessing, sentiment labeling, feature transformation, feature selection, data balancing, and model evaluation using performance metrics such as accuracy, precision, recall, and F1-score. This research is expected to support the development of intelligent feedback systems for digital health services and offer insights for sentiment analysis in low-resource language settings such as Bahasa Indonesia.

The primary contribution of this study is not only in implementing traditional statistical models but in providing an in-depth comparative analysis of feature selection methods under different feature representations, coupled with SMOTE-based data balancing.

However, sentiment analysis in the Indonesian language remains a challenging task due to limited linguistic resources, high morphological variations, and semantic ambiguity, which often reduce the generalization ability of machine learning models. Previous studies primarily focused on model implementation without addressing data imbalance and feature redundancy that may degrade classification performance. Therefore, this study emphasizes a methodological contribution by integrating SMOTE-based data balancing with comparative statistical feature selection techniques to systematically evaluate their effects on sentiment classification performance. This framework aims to provide a more interpretable and adaptable approach to sentiment analysis in low-resource environments, particularly for Indonesian-language health applications.

II. Methods

The methodological pipeline includes data acquisition, text preprocessing, sentiment labeling, feature transformation, feature selection, data balancing, and model evaluation using performance metrics such as accuracy, precision, recall, and F1-score. This research is expected to support the development of intelligent feedback systems for digital health services and offer insights for sentiment analysis in low-resource language settings such as Bahasa Indonesia. [12].

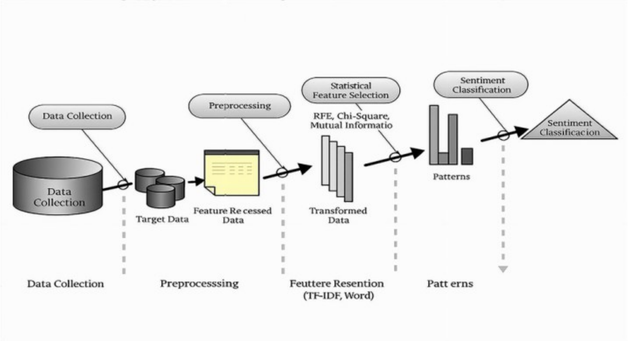


Figure 1. KDD process steps.

The presented figure illustrates the workflow of the Knowledge Discovery in Databases (KDD) process as applied to sentiment classification. The process begins with the collection of user review data sourced from the Google Play Store platform, followed by a series of text pre-processing stages, including cleaning, tokenization, stop word removal, and stemming.

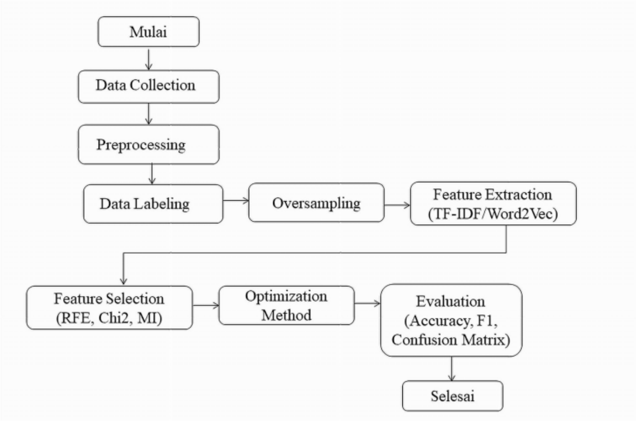

Mulai
Data Collection
Preprocessing
Data Labeling
Oversampling
Feature Extraction (TF-IDF/Word2Vec)
Feature Selection (RFE, Chi2, MI)
Optimization Method
Evaluation (Accuracy, F1, Confusion Matrix)
Selesai

Figure 2. Research steps.

The pre-processing data is then transformed using feature representation techniques such as TF-IDF and Word2Vec [13]. Subsequently, feature selection is carried out using statistical approaches such as Recursive Feature Elimination (RFE), Chi-Square, and Mutual Information to identify the most relevant attributes. Sentiment patterns are then extracted using classification algorithms such as Logistic Regression

[14]. Finally, the sentiment polarity is determined by categorizing the reviews into positive, neutral, or negative classes. This pipeline demonstrates a structured approach to converting unstructured textual data into meaningful information [15]. Figure 2 illustrates the sentiment classification process, as detailed in the following subsections.

A. Data Collection

The dataset used in this study was collected through web scraping of user reviews from the Google Play Store for the Mobile JKN application. A total of 2,382 user reviews were gathered, ranging from November 2017 to May 2025. Each review consists of a text comment and an associated user rating ranging from 1 to 5 stars. The star ratings were used to label sentiments: 1–2 stars as negative, 3 stars as neutral, and 4–5 stars as positive

B. Preprocessing

The raw review texts underwent several preprocessing steps, including lowercasing, removal of punctuation, stop words, and special characters, as well as tokenization and stemming [16]. This step aims to normalize the text and reduce dimensionality. A label encoding step was performed to transform sentiment labels into numerical categories: 0 for negative, 1 for neutral, and 2 for positive.

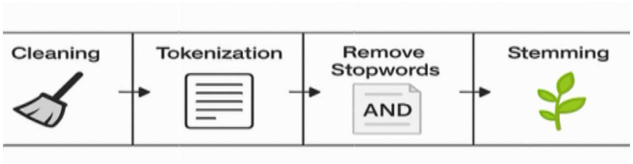


Figure 3. Data preprocessing steps

Data pre-processing in this study was carried out through four primary stages. The first stage, cleaning, involved purifying the textual data by removing irrelevant elements such as HTML tags, URLs, symbols, special characters, and punctuation marks that do not contribute meaningfully to sentiment interpretation. The subsequent stage was tokenization, which entailed segmenting documents or sentences into smaller word units (tokens) to facilitate linguistic manipulation and analysis. This was followed by stop word removal, namely the elimination of common words (e.g., "and," "or," "which") considered to have negligible informational value for the classification process. The final stage was stemming, which was carried out using the Sastrawi library to reduce each word to its root form (lemma). For instance, words such as "membaca", "pembacaan", and "dibaca" were normalized to the base form "baca". This process helps to unify variations of a word into a single representation, thereby improving the consistency and efficiency of feature extraction in the subsequent modelling stages [17]. All preprocessing steps were applied sequentially to generate clean, consistent, and standardized textual data before transforming it into feature representations with TF-IDF and Word2Vec, followed by feature selection and classification using the Logistic Regression algorithm [18].

TABLE I.
RESULTS OF TEXT PROCESSING STEPS

Steps	Result
Original Text	Difficult to get SMS verification code, always fails... please fix it
	At first, it was difficult to register, but I downloaded again, then registered again, and Alhamdulillah it worked... very helpful for online registration to avoid long queues
	The latest version of the JKN application has been installed on the device.
Cleaning	difficult to get sms verification code always fails please fix it
	at first it was difficult to register but i downloaded again then registered again and alhamdulillah it worked very helpful for online registration to avoid long queues
	jkn application latest version has been installed on the device
Tokenization	['difficult', 'to', 'get', 'sms', 'verification', 'always', 'fails', 'please', 'fix']
	['at', 'first', 'difficult', 'to', 'register', 'but', 'i', 'downloaded', 'again', 'then', 'registered', 'again', 'and', 'alhamdulillah', 'it', 'worked', 'very', 'helpful', 'for', 'online', 'registration', 'to', 'avoid', 'long', 'queues']
	['application', 'jkn', 'latest', 'version', 'has', 'been', 'installed', 'on', 'device']
Remove Stopwords	['difficult', 'get', 'sms', 'verification', 'fails', 'please', 'fix']
	['first', 'difficult', 'register', 'downloaded', 'again', 'registered', 'alhamdulillah', 'worked', 'helpful', 'online', 'registration', 'avoid', 'queues']
	['application', 'jkn', 'latest', 'version', 'installed', 'device']
Stemming	['difficult', 'get', 'sms', 'verify', 'fail', 'please', 'fix']
	['first', 'difficult', 'register', 'download', 'again', 'register', 'alhamdulillah', 'become', 'help', 'register', 'online', 'avoid', 'queue']
	['application', 'jkn', 'version', 'new', 'install', 'device']

C. Data Labelling

In this study, the sentiment labeling process for reviews of the Mobile JKN application was conducted by referencing the user rating scores available on the Google Play Store platform. Reviews with ratings ranging from 1 to 2 were classified as negative sentiment, a rating of 3 was designated as neutral sentiment, and ratings from 4 to 5 were categorized as positive sentiment [19].

This labeling approach was based on the general assumption that low scores reflect user dissatisfaction, high scores indicate a good level of satisfaction, and medium scores are considered neutral. From a total of 2,382 reviews successfully collected and processed, the sentiment distribution was as follows: 1,611 reviews (67.6%) were identified as neutral, 386 reviews (16.2%) as positive, and 385 reviews (16.2%) as negative. These findings suggest that the majority of users provided neutral evaluations regarding the services offered by the Mobile JKN application.

D. Data Splitting

In this study, the dataset was divided into two parts with a proportion of 80% for training data and 20% for testing data. The training data was used to build and train the model, while the testing data was used to evaluate its performance. To address class imbalance, the training data was balanced using the SMOTE technique.

TABLE II.
DATA LABELLING

Text	Negative Score	Positive Score	Total Score	Sentiment
[′susah′,′mendapatkan′,′kode′,′sms′,′verifikasi′,′gagal′,′terus′,′tolong′,′diperbaikin′]	-2	0	-2	negatif
[′awalnya′,′susah′,′daftar′,′download′,′ulang′,′daftar′,′alhamdulillah′,′jadi′,′membantu′,′mendaftar′,′online′,′antri′]	-1	2	1	positif
[′aplikasi′,′jkn′,′versi′,′terbaru′,′terpasang′,′perangkat′]	0	0	0	netral

This approach aligns with previous studies [8], ensuring a proportional evaluation of the model’s accuracy and generalization capability.

E. Oversampling

To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training data. This technique generates synthetic samples in the minority classes, thereby balancing the dataset and improving model generalization.

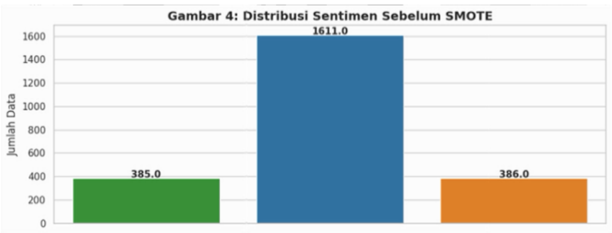


Figure 4. Data before SMOTE.

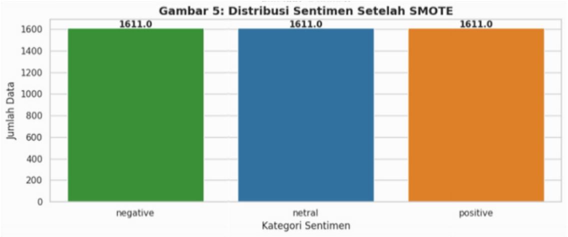


Figure 5. Data after SMOTE.

F. Feature Extraction

Following the data splitting process, feature extraction was carried out utilizing two primary methods, namely Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec. The TF-IDF approach assigns weights to each word based on its frequency within an individual document and its rarity across the entire corpus, thereby attributing higher significance to more informative terms. In contrast, the Word2Vec method transforms words into high-dimensional vectors while considering their surrounding context, enabling the model to capture the semantic relationships among words [20]. To further enhance the accuracy and efficiency of the Logistic Regression classification model, feature selection was implemented using three statistical techniques: Recursive Feature Elimination (RFE), Chi-Square, and Mutual Information. These methods were employed to identify the most relevant features and reduce the presence of less significant attributes in the sentiment classification of Mobile JKN application reviews [21].

G. Feature Selection

To reduce noise and improve classification performance, three statistical feature selection techniques were applied:

1. Recursive Feature Elimination (RFE): Selects the most significant features by recursively removing the least important ones based on model weight coefficients [22].
2. Chi-Square Test: Measures the dependence between features and sentiment labels [23].
3. Mutual Information: Quantifies the amount of information obtained about sentiment classes from the presence of particular features [24].

Each feature representation was tested with each selection method, resulting in six experimental configurations.

H. Optimization Method

In this study, Logistic Regression was employed as the primary classification algorithm to analyze sentiment in user reviews of the Mobile JKN application. Feature representation was conducted using two approaches: Term Frequency-Inverse Document Frequency (TF-IDF), which assigns weights to words based on their relative frequency and significance within the corpus, and Word2Vec, which transforms words into fixed-dimensional numerical vectors while capturing their semantic relationships [25]. To enhance model performance and eliminate less relevant features, feature selection was carried out by applying three statistical techniques, namely Recursive Feature Elimination (RFE), Chi-Square, and Mutual Information [26]. The entire classification workflow, including data transformation and model evaluation, was efficiently implemented utilizing the Scikit-Learn library.

TABLE III.
FUNCTION EXPRESSION OF METHODS USED IN RESEARCH

Method	Function Expression	Description
TF-IDF	w(i,j) = tf(i,j) · log(N/θ)	TF indicates frefk. i db. of doc.; Indicates how important an individual word is throughout the corpus
Word2Vec (CBOW/Skip-gram)	Hano explicit (model ba sebasis neural); output; vk∈R∘	Represents each word *k* as a fixed dimerise vector based on its contex
Logistic Regression	P(y=1∣X)=1+e(B1x1 +…+Bnxn)1	Binary classification model based on probabilities, iterably limited to multi-class class classification problems
Chi-Square (X^2)	χ2=x,y+∑Ej(Oi−Ei)2	Measures the association between eand class
Mutual Information	I(X;Y) = Σx,y p(x,y) · log(Py/p(x))	Measures how much information feature *X* has about target *Y*
Recursive Feature Elim. (Ekimmination)	No explicit mathematical expression; coefficient rank based	Iteratively remove feature with lowest contribution in model

To ensure reproducibility and transparency, the Word2Vec representation adopted the skip-gram architecture with a 100-dimensional vector space and a window size of 5 to capture broader semantic context. The Logistic Regression classifier was configured using the lbfgs solver with L2 regularization and a maximum iteration limit of 1000 to prevent overfitting. The dataset was split into 80% training and 20% testing portions, and the SMOTE technique was applied exclusively to the training data to avoid data leakage during evaluation. This configuration ensures that oversampling does not bias the testing set, thereby producing a more reliable performance estimation. All experiments were conducted using the Scikit-Learn and Gensim libraries under Python 3.10.

I. Evalations

The performance evaluation of the model in this study was carried out using a confusion matrix, which visually represents the classification outcomes to facilitate a more comprehensive analysis of model performance. This matrix comprises four fundamental components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) [26]. Based on these components, several evaluation metrics are computed, including accuracy, precision, recall, and F1-score, to assess the effectiveness of the model in classifying sentiments from user reviews of the Mobile JKN application [27]. Each model configuration was evaluated using the following metrics:

The respective formulations of these evaluation metrics are presented in Equations (1) to (4) [28][29].

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP} + FN \quad (2)$$

$$\text{Presisi} = \frac{TP}{TP} + FP \quad (3)$$

$$\text{F1-Score} = \frac{2 \text{ x Presisi X Recall}}{\text{Presisi + Recall}} \quad (4)$$

III. RESULTS AND DISCUSSION

This section presents the results of sentiment classification on Mobile JKN user reviews based on the six model configurations derived from combinations of feature representation techniques (TF-IDF and Word2Vec) and feature selection methods (RFE, Chi-Square, and Mutual Information). Each model was evaluated using accuracy, precision, recall, and F1-score.

1. TF-IDF Experiment + Feature Selection

In the sentiment classification experiment using TF-IDF feature representation and the Logistic Regression algorithm, three feature selection techniques were evaluated: RFE, Chi-Square, and Mutual Information. The model with RFE yielded the lowest performance, with an accuracy of 59.33%, precision of 51%, recall of 56%, and an f1-score of 52%. The Chi-Square configuration showed improved performance, achieving an accuracy of 71.7%, precision of 75%, but a relatively low recall of 44%, and an f1-score of 46%. Meanwhile, the Mutual Information method produced the best overall results, with an accuracy of 73.38%, precision of 75%, recall of 47%, and an f1-score of 50%.

These results indicate that the choice of feature selection method significantly affects the model's classification performance. Although the precision values are consistently high, the lower recall and f1-scores suggest challenges in accurately identifying positive and negative sentiments. This highlights the need for further enhancement in feature representation and selection to achieve better class balance in sentiment detection.

TABLE IV.
TF-IDF EXPERIMENT + FEATURE SELECTION

Rep. (%)	Sel. (%)	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
TF-IDF	RFE	59.33	51	56	52
TF-IDF	Chi-Square	71.7	75	44	46
TF-IDF	MI	73.38	75	47	50
Word2Vec	RFE	50.73	47	53	46
Word2Vec	Chi-Square	43.4	41	45	39
Word2Vec	MI	46.54	44	48	42

2. Word2Vec Experiment + Feature Selection

The sentiment classification experiment using the Logistic Regression algorithm with Word2Vec feature representation and three feature selection techniques—RFE, Chi-Square, and Mutual Information—yielded relatively lower performance compared to the TF-IDF-based approach. Among the configurations, RFE achieved the highest performance, with an accuracy of 50.73%, precision of 47%, recall of 53%, and an f1-score of 46%. The Chi-Square method produced an accuracy of 43.4%, precision of 41%, recall of 45%, and f1-score of 39%. Meanwhile, Mutual Information reached an accuracy of 46.54%, precision of 44%, recall of 48%, and f1-score of 42%.

Overall, these findings indicate that the Word2Vec feature representation did not significantly enhance classification performance in this context. Although minor improvements in recall were observed in some feature selection techniques, the overall f1-scores remained low. This suggests that the combination of Word2Vec and the applied statistical selection methods has not been sufficiently effective in capturing sentiment characteristics, particularly for the positive and negative classes. Further investigation into more suitable feature representation and selection strategies is warranted to improve classification outcomes.

TABLE V.
WORD2VEC EXPERIMENT + FEATURE SELECTION

Rep. (%)	Sel. (%)	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
Word2Vec	RFE	50.73	47	53	46
Word2Vec	Chi-Square	43.4	41	45	39
Word2Vec	MI	46.54	44	48	42

3. Comparison of Accuracy and F1 Score of Various Methods

Table VI presents a comparison of the performance of the Logistic Regression algorithm in classifying sentiment in reviews of the Mobile JKN application, utilizing two feature representation approaches, namely TF-IDF and Word2Vec, each combined with three feature selection techniques: Recursive Feature Elimination (RFE), Chi-Square, and Mutual Information.

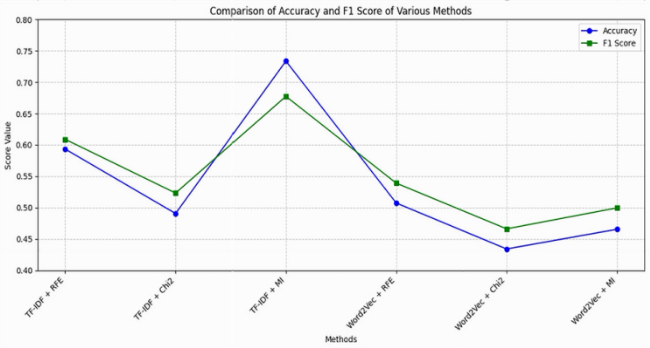
TABLE VI.
COMPARISON OF ACCURACY AND F1 SCORE OF VARIOUS METHODS

Rep. (%)	Sel. (%)	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
TF-IDF	RFE	59.33	51	56	52
TF-IDF	Chi-Square	71.7	75	44	46
TF-IDF	MI	73.38	75	47	50

This study proposed a sentiment classification framework for analyzing Mobile JKN user reviews using Logistic Regression combined with statistical feature selection techniques and SMOTE-based data balancing. A comparative assessment of two feature representation approaches (TF-IDF and Word2Vec) with three selection strategies (RFE, Chi-Square, and Mutual Information) revealed that the TF-IDF + Mutual Information configuration achieved the best results with 73.38% accuracy and a 50% F1-score.

Although this configuration outperformed other approaches, the overall performance should be considered moderate rather than state-of-the-art, particularly in capturing the nuances of Indonesian sentiment. This finding emphasizes that while traditional statistical models remain useful, they are limited in addressing semantic complexity in low-resource languages.

The main contribution of this research lies in providing a methodological comparison framework that integrates feature representation, feature selection, and data balancing to support more interpretable sentiment classification. For future work, more advanced techniques such as deep learning or transformer-based architectures (e.g., BERT or IndoBERT) are recommended to achieve higher performance and broader generalization.



Fiure 6. Comparison of Accuracy and F1 Score of Various Methods.

Figure 6 presents the comparison of Accuracy and F1-Score across all methods. The visualization highlights the consistent superiority of TF-IDF with Mutual Information and clearly illustrates the performance gaps among feature representation methods.

4. *Erorr Analysis Based on Confusion Matrix*

The confusion matrix analysis provided further insight into the performance of the sentiment classification model. The majority of misclassifications occurred between the neutral and negative sentiment classes, primarily due to overlapping lexical patterns such as the frequent use of polite expressions (e.g., "please fix," "hope better") that mask negative sentiment in formal reviews. Although the application of SMOTE successfully balanced the dataset distribution, it did not entirely resolve semantic ambiguity within the data.

These findings indicate that while the TF-IDF + Mutual Information configuration achieved the best overall performance (73.38% accuracy and an F1-score of 50%), the relatively low recall for the negative class shows that subtle linguistic cues in Indonesian reviews remain challenging for traditional statistical models. Future studies are encouraged to incorporate context-aware embedding models such as IndoBERT or multilingual BERT, which are better suited for capturing contextual nuances and improving sentiment separation, particularly between neutral and negative classes.

5. *Comparison with Previous Studies*

To provide a broader comparison, Table VII summarizes the performance of our proposed framework against results from previous studies employing alternative algorithms. While our Logistic Regression model with TF-IDF and Mutual Information achieved a moderate performance (73.38% accuracy and an F1-score of 50%), SVM-based methods generally produced higher accuracy in similar tasks, particularly when optimized with metaheuristics such as PSO [10]. Furthermore, transformer-based models such as IndoBERT substantially outperformed traditional statistical models, achieving above 80% accuracy and F1-scores close to 80%.

These findings reinforce the notion that, although our framework offers interpretability and methodological contributions, more advanced architectures are needed to achieve state-of-the-art performance in Indonesian sentiment analysis. Nonetheless, the superior performance of Mutual Information compared to Chi-Square and RFE can be attributed to its ability to capture non-linear dependencies between features and sentiment labels, providing richer discriminative information. While Chi-Square primarily measures linear associations, Mutual Information quantifies how much information a feature contributes to predicting sentiment classes regardless of dependency type. This enables the model to detect subtle linguistic cues within Indonesian-language reviews, especially when feature distributions overlap semantically. Consequently, the TF-IDF + Mutual Information configuration effectively balances statistical relevance and semantic expressiveness, leading to consistently better performance across evaluation metrics.

TABLE VII.
COMPARISON WITH PREVIOUS METHODS (FROM LITERATURE)

Method	Dataset Size	Best Acc. (%)	F1 (%)	Notes
This study (LR + TF-IDF + MI)	2, 382 Reviews	73.38	50	Moderate, balanced with SMOTE
SVM + PSO [10]	2,000 reviews (JKN)	76.2	-	Optimized with PSO
SVM + Word2Vec [4]	3,000 reviews (vaccines)	78.5	-	Word2Vec enhanced SVM
IndoBERT [19]	5,000 Shopee reviews	84.3	79	Transformer-based model

IV. CONCLUSION

This study proposed a sentiment classification framework to analyze user reviews of the Mobile JKN health application using Logistic Regression enhanced with statistical feature selection techniques. Through a comparative assessment of two feature representation methods (TF-IDF and Word2Vec) combined with three selection strategies (RFE, Chi-Square, and Mutual Information), the most effective configuration for categorizing sentiment into positive, neutral, and negative classes was identified.

The experimental results revealed that the combination of TF-IDF and Mutual Information achieved the highest overall performance, with an accuracy of 73.38% and an F1-score of 50%. These findings demonstrate that well-optimized traditional statistical techniques remain effective for sentiment classification tasks, particularly in low-resource language environments such as Bahasa Indonesia.

The methodological contribution of this study lies in providing a comparative framework that integrates feature representation, statistical feature selection, and data balancing through SMOTE to enhance sentiment classification robustness. The application of SMOTE effectively mitigated class imbalance and improved the stability of classification across sentiment categories. Furthermore, the framework demonstrates that even with a moderate-sized dataset, systematic feature selection and balancing strategies can produce consistent and interpretable results comparable to more complex models.

The findings can serve as a practical foundation for future research on sentiment analysis in other low-resource linguistic contexts. Future studies are encouraged to extend this framework using transformer-based architectures such as IndoBERT or multilingual BERT to improve semantic comprehension, accuracy, and model generalizability. Future studies may incorporate ROC-AUC and precision-recall metrics to provide a more comprehensive model evaluation.

REFERENCES

[1] Y. D. et al. Mai, Thanh; M, Shahbaz; Tong, "Pr ep rin t n ot pe er r iew Pr ep rin t n ot pe er ed," *Fusion*, pp. 1–8, 2023, doi: 10.2139/ssrn.5277059.

[2] Renny, Harmendo, and D. Kusmadeni, "Analisis Transformasi Digital BPJS Kesehatan Dalam Mendukung Mutu Layanan Jaminan Kesehatan Nasional," *J. Penelit. Perawat Prof.*, vol. 6, pp. 2075–2091, 2024, doi: 10.37287/jppp.v6i5.3142.

[3] N. Z. B. Jannah and K. Kusnawi, "Comparison of Naïve Bayes and SVM in Sentiment Analysis of Product Reviews on Marketplaces," *Sinkron*, vol. 8, no. 2, pp. 727–733, 2024, doi: 10.33395/sinkron.v8i2.13559.

[4] C. A. Nurhaliza Agustina, R. Novita, Mustakim, and N. E. Rozanda, "The Implementation of TF-IDF and Word2Vec on Booster Vaccine Sentiment Analysis Using Support Vector Machine Algorithm," *Procedia Comput. Sci.*, vol. 234, pp. 156–163, 2024, doi: 10.1016/j.procs.2024.02.162.

[5] A. Madasu and S. Elango, "Efficient feature selection techniques for sentiment analysis," *Multimed. Tools Appl.*, vol. 79, no. 9–10, pp. 6313–6335, 2020, doi: 10.1007/s11042-019-08409-z.

[6] F. Rifaldy, Y. Sibaroni, and S. S. Prasetiyowati, "Effectiveness of word2vec and tf-idf in sentiment classification on online investment platforms using support vector machine 1.," vol. 10, no. 2, pp. 863–874, 2025, doi: https://doi.org/10.29100/jipi.v10i2.6055.

[7] R. D. Kurniawan, "GoPay App Review Sentiment Classification Optimization Using a Combination of Text Representation and Machine Learning," vol. 6, no. 2, pp. 31–36, 2024, doi: 10.24246/ijiteb.622024.31-36.

[8] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Appl. Sci.*, vol. 13, no. 6, 2023, doi: 10.3390/app13064006.

[9] T. Ramadhani, P. Hermawan, and A. R. Dzikrillah, "Penerapan Metode Naïve Bayes untuk Analisis Sentimen pada Ulasan Pengguna Aplikasi ChatGPT di Google Play Store," *Technol. Sci.*, vol. 6, no. 1, pp. 430–439, 2024, doi: 10.47065/bits.v6i1.5400.

[10] N. Maulida, N. Suarna, and W. Prihartono, "Analisis Ulasan Sentimen Aplikasi Mobile Jkn Dengan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1651–1658, 2024, doi: 10.36040/jati.v8i2.9105.

[11] B. Setiawan, "A Review of Sentiment Analysis Applications in Indonesia Between 2023-2024," vol. 08, pp. 71–83, 2024, doi: 10.1016/j.procs.2021.01.190.

[12] T. Husain and N. Hidayati, "The Optimize Of Association Rule Method For The Best Book Placement Patterns In Library : A Monthly Trial," vol. 4, no. 2, pp. 53–59, 2021, doi: 10.31943/teknokom.v4i2.63.

[13] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bull. Electr. Eng. Informatics*, vol. 10, no. 5, pp. 2780–2788, 2021, doi: 10.11591/eei.v10i5.3157.

[14] S. Kumar, N. Kaur, Kavita, and A. Joshi, "Tweet Sentiment Analysis using Logistic Regression," *IET Conf. Proc.*, vol. 2023, no. 11, pp. 332–336, 2023, doi: 10.1049/icp.2023.1801.

[15] S. I. R. Adi, B. Bakkara, K. A. Zega, F. N. Vielita, and N. A. Rakhmawati, "Analisis Sentimen Masyarakat Terhadap Progress Ikn Menggunakan Model Decision Tree," *JIKA (Jurnal Inform.*, vol. 8, no. 1, p. 57, 2024, doi: 10.31000/jika.v8i1.9803.

[16] A. Mirugwe *et al.*, "Sentiment Analysis of Social Media Data on Ebola Outbreak Using Deep Learning Classifiers," *Life*, vol. 14, no. 6, pp. 8–14, 2024, doi: 10.3390/life14060708.

[17] H. T. Duong and T. A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput. Soc. Networks*, vol. 8, no. 1, pp. 1–16, 2021, doi: 10.1186/s40649-020-00080-x.

[18] S. N. Cahyani and G. W. Saraswati, "Implementation of Support Vector Machine Method in Classifying School Library Books With Combination of Tf-Idf and Word2Vec," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1555–1566, 2023, doi: 10.52436/1.jutif.2023.4.6.1536.

[19] K. Hasanah, "Comparison of Sentiment Analysis Model for Shopee Comments on Google Play Store," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 13, no. 1, pp. 21–30, 2024, doi: 10.32736/sisfokom.v13i1.1916.

[20] A. H. Dani, E. Y. Puspaningrum, and R. Mumpuni, "Studi Performa TF-IDF dan Word2Vec Pada Analisis Sentimen Cyberbullying," *Router J. Tek. Inform. dan Terap.*, vol. 2, no. 2, pp. 94–106, 2024, [Online]. Available: https://doi.org/10.62951/router.v2i2.76

[21] E. Edwar, I. G. A. N. R. Semadi, M. Samsudin, and I. K. Dharmendra, "Perbandingan Metode Seleksi Fitur Pada Analisis Sentimen (Studi Kasus Opini Pilkada DKI 2017)," *INFORMATICS Educ. Prof. J. Informatics*, vol. 8, no. 1, p. 11, 2023, doi: 10.51211/itbi.v8i1.2408.

[22] C. A. Ramezan, "Transferability of Recursive Feature Elimination (RFE)-Derived Feature Sets for Support Vector Machine Land Cover Classification," *Remote Sens.*, vol. 14, no. 24, 2022, doi: 10.3390/rs14246218.

[23] M. B. Hamzah, "Classification of Movie Review Sentiment Analysis Using Chi-Square and Multinomial Naïve Bayes with Adaptive Boosting," *J. Adv. Inf. Syst. Technol.*, vol. 3, no. 1, pp. 67–74, 2021, doi: 10.15294/jaist.v3i1.49098.

[24] W. Han, H. Chen, and S. Poria, "Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis," *EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc.*, no. Mi, pp. 9180–9192, 2021, doi: 10.18653/v1/2021.emnlp-main.723.

[25] I. Hendrawan Rifky, E. Utami, and A. Hartanto Dwi, "Analisis Perbandingan Metode Tf-Idf dan Word2vec pada Klasifikasi Teks Sentimen Masyarakat Terhadap Produk Lokal di Indonesia," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 11, no. 3, pp. 497–503, 2022, doi: 10.30591/smartcomp.v11i3.3902.

[26] M. Ali Kawo, G. Muhammad, D. Gabi, and M. Sule Argungu, "A Comparative Study of Some Selected Classifiers on an Imbalanced Dataset for Sentiment Analysis," *Int. J. Innov. Sci. Res. Technol.*, vol. 9, no. 5, pp. 2826–2832, 2024, doi: 10.38124/ijisrt/ijisrt24may1751.

[27] M. Janaah and A. Nugroho, "Performance of SVM Optimized with PSO as Classification Method for Sentiment Analysis UNNES ' s Social Media," pp. 68–80, 2025, doi: 10.20895/infotel.v17i1.1266.

[28] F. M. Anto, L. S. Abimanyu, and T. Herdi, "Penerapan Algoritma Naïve Bayes Dengan Feature Selection Pada Data Penjualan Konstruksi," *J. Ilm. FIFO*, vol. 15, no. 2, p. 102, 2024, doi: 10.22441/fifo.2023.v15i2.002.

[29] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, 2022, doi: 10.1038/s41598-022-09954-8.