AN EVALUATION OF MACHINE LEARNING MODELS FOR ANALYZING USER SENTIMENTS ON THE LIVIN' BY MANDIRI APPLICATION

RAMA ASHARI[1], ARIEF SETYANTO[2]

[1,2] Magister of Informatics Engineering, Universitas AMIKOM Yogyakarta, Indonesia
E-MAIL: ramaashari@students.amikom.ac.id, arief_s@amikom.ac.id

Abstract:

This study presents a sentiment analysis of user reviews on the Livin' by Mandiri mobile banking application, using a dataset of Indonesian-language reviews collected from the Google Play Store. Reviews were labeled based on their star ratings, with 5-star reviews classified as positive and 1-star reviews as negative. The data underwent extensive preprocessing, including tokenization, stopword removal, stemming, and lemmatization. A Term Frequency-Inverse Document Frequency (TF-IDF) approach was used to transform the text into feature vectors. Six machine learning algorithms Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM) were evaluated. Among these, the SVM achieved the best performance, with an accuracy of 93.4%, precision of 94.7%, recall of 91.6%, and an F1-score of 93.1%. Although ensemble models were also tested, they provided minimal performance gains. The results demonstrate that SVM offers a reliable and efficient solution for sentiment classification of Indonesian-language mobile app reviews, with strong potential for real-world financial technology applications.

Keywords:

sentiment analysis, machine learning, mobile banking, TF-IDF, Support Vector Machine, Livin' by Mandiri.

1. Introduction

Customer reviews have become a crucial component in influencing the decisions of new customers in the digital era, driven by significant advancements in technology and widespread connectivity. One effective way to make customers feel valued is by seeking their opinions and appreciating their feedback. As a result, users have become increasingly accustomed to sharing their experiences and opinions about products and services through various online platforms, such as discussion forums, e-commerce websites, and social media. Customer reviews not only serve as references for other customers but also assist service providers in understanding the market, improving products, and making better business decisions [1]. Customer satisfaction is one of the key factors in determining how well a product or service meets customer expectations [2]. Moreover, customer satisfaction is a vital tool that can assist businesses in various aspects of their operations, such as increasing revenue or reducing marketing costs.

On the other hand, the large and continuously growing number of reviews makes it increasingly difficult and time-consuming to analyze and interpret customer perspectives [3]. The increasing diversity of review data, which includes various writing styles, languages, and levels of detail, makes it more challenging to convey the emotions and meanings contained within them. Natural Language Processing (NLP) enables computer systems to automatically identify, categorize, and measure the sentiments expressed in customer review texts. Sentiment analysis is a method used to analyze information and determine opinions conveyed through text [4]. Computer systems can describe customer perspectives more accurately by using sentiment analysis to understand the meaning and specific features within the text.

By examining users' implicit attitudes and uncovering significant emotions hidden within their comments, sentiment analysis can draw conclusions about users' feelings toward various topics [5]. In practice, sentiment analysis can be classified into positive and negative sentiments. [6]. Reviews with "praise" comments can encourage new users to try the service, while "critical" reviews can serve as evaluations for service providers, particularly mobile applications, considering the intense competition in the app development industry. Millions of apps available on the Google Play Store [7] have already received reviews from their users. The sentiment in these user reviews correlates with the ratings or stars given by the users. In most positive reviews, the rating given is 5 stars, whereas negative reviews typically receive a 1-star rating. Reviews with ratings of 2, 3, and 4 stars tend to be more objective and less definitive.

Positive and negative sentiment analysis can be classified using various machine learning algorithms. In this study, several algorithms are used to classify the existing reviews, including DT, LR, K-NN, RF, and NB. The aim is to identify the most suitable algorithm for classifying negative

and positive reviews. By using multiple algorithms, this research aims to provide a comparison of the performance of each algorithm.

To address this challenge, this study implements and evaluates several supervised machine learning models, including DT, KNN, LR, RF, NB, and SVM, for binary sentiment classification. While the individual models employed are well-established in natural language processing tasks, this research contributes novelty through its practical application in the context of Indonesian-language reviews from the Livin' by Mandiri mobile banking application—a domain and language combination that remains underrepresented in the literature.

Furthermore, the study goes beyond basic classification by conducting additional ensemble model experiments (e.g., SVM combined with LR and RF) to explore potential performance improvements. The results confirm that SVM alone remains highly effective and efficient, highlighting its robustness for high-dimensional TF-IDF feature spaces. Thus, this research offers practical value in both implementation and validation of sentiment analysis techniques in a real-world financial technology context.

2. Related Work

Research [8] on sentiment analysis of customer reviews in a restaurant has demonstrated the importance of this information in improving service quality and understanding customer perceptions. On the other hand, sentiment analysis is also applied across various platforms, including mobile banking applications, which classify user reviews into positive and negative categories using Support Vector Machine (SVM) [9]. Furthermore, sentiment analysis is used to detect reviews in e-commerce platforms [10] and online transportation services [11]. In study [12], sentiment analysis of customer feedback and airline service reviews was conducted using BERT, achieving the highest accuracy of 83%. Sentiment analysis research has also been conducted to analyze customer reviews of the Uber application [13]. Meanwhile, in study [14], reviews left by viewers on YouTube content related to Indian and Pakistani songs were analyzed. Comments written in Latin script Urdu were classified using several algorithms, resulting in the highest accuracy of 92.25% using LR.

In previous studies, several researchers have compared different models. Each model used yielded different accuracy results. The K-NN algorithm outperformed NB in sentiment analysis classification of Digital Payment [15]. On the other hand, the DT algorithm showed the best classification performance compared to Neural Network, Naïve Bayes, and SVM [11]. However, these studies demonstrated relatively low performance results. Therefore, data balancing is needed in machine learning algorithms for sentiment analysis classification in customer reviews.

Sentiment analysis of a sentence requires the removal of affixes that do not carry meaning. This process is known as stemming, where words are processed, affixes are discarded, and the sentence is normalized for better understanding [16]. Additionally, an important step is the removal of stopwords. Stopwords are words that frequently appear but have little meaning. The main benefit of stopword removal is to improve system performance by reducing memory usage and saving processing time during indexing and searching [17].

3. Method

3.1. Data Collection

The data collection process was carried out by scraping reviews for the Livin' by Mandiri app from the Google Play Store website. The scraping was done using the Python library, google_play_scraper, by inputting the token "id.bmri.livin." The Livin' by Mandiri mobile application has a total of 10 million active users since its launch until 2022 [18]. The collected reviews cover the period from September 30, 2021, to December 17, 2023, with a specific filter for reviews written in Indonesian by users located in Indonesia.

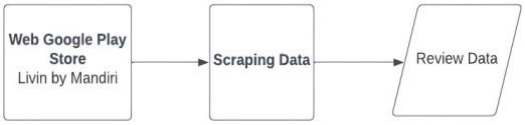

Figure 1. Data Scraping Flow

3.2. Preprocessing Data

The data preprocessing stage was designed to clean and standardize the user reviews before sentiment classification. First, sentiment labeling was performed by assigning a value of 1 (positive) to reviews with 5-star ratings and 0 (negative) to those with 1-star ratings. Reviews with ratings of 2, 3, or 4 stars were excluded due to their subjective and ambiguous nature. To address the issue of class imbalance, the dataset was balanced by randomly selecting 10,000 positive and 10,000 negative reviews, resulting in a total of 20,000 balanced samples.

Subsequent text cleaning involved converting all characters to lowercase, removing punctuation, numbers, emojis, and special symbols, and eliminating extra white spaces to standardize the review content. Tokenization was then applied to split each review into individual words.

Stopword removal followed, eliminating commonly used but semantically weak words (e.g., yang, dan, di) that do not contribute meaningfully to sentiment classification. Finally, stemming and lemmatization were carried out to reduce each token to its root or base form. This step was specifically tailored using Indonesian language processing rules to ensure proper morphological normalization, enhance textual consistency, and reduce dimensionality. The preprocessed data was then stored in a structured format for subsequent feature extraction.

3.3. Feature Extraction

After the data preprocessing, the next step is to determine the independent and dependent variables. These variables are named X and Y, respectively. Then, word weighting is performed using TF-IDF. TF-IDF is a numerical technique that allows the determination of weights for each term or word in each document [19]. This method is used to calculate the TF and IDF values for each token (word) in each document within the corpus [20]. In simple terms, the TF-IDF method is used to determine how frequently a word appears in a document. The formula used to calculate the weight of each token in the document is as follows [21]:

$$IDF_j = log\left(\frac{N}{DF_j}\right) \quad (1)$$

$$TF - IDF_j = TF_j \times IDF_j \quad (2)$$

Where IDF_j is the inverse document frequency of term j, N is the total number of documents, $\boldsymbol{DF_j}$ is the number of documents containing the term, and $\boldsymbol{TF_j}$ is the term frequency, representing the number of occurrences of the term in a document relative to the total number of words in that document.

3.4. Sentiment Classification

Six classification algorithms were evaluated: DT, K-NN, LR, RF, NB, and SVM. The dataset was split into 80% training and 20% testing subsets. Models were implemented using the scikit-learn library in Python. For SVM, a linear kernel was selected due to the high-dimensional sparse nature of the TF-IDF feature space.

To validate the robustness of the results and explore the potential of ensemble approaches, additional experiments were conducted by combining classifiers through soft voting ensembles. These included: SVM + RF, SVM + LR, and SVM + RF + LR. While some of these hybrid models showed minor performance gains, the standalone SVM classifier consistently achieved the highest balance of accuracy and efficiency, confirming its suitability for this task.

3.5. Performance Evaluation

The evaluation stage was carried out after the processed data, using several algorithms, produced positive or negative sentiment classifications. The purpose of this stage is to determine how accurate the classification results of each algorithm are. The technique used to evaluate the algorithms in this study is the confusion matrix. The confusion matrix provides information regarding Accuracy, F1-Score, Precision, and Recall.

4. Result

The flow of the dataset collection is shown in Figure 1. Unnecessary features are removed, leaving the data with the features "at," "content," "score," "thumbUpCount," and "appVersion." The data is then sorted based on "at" and saved in CSV format. The top five rows of the data are shown in Table 1.

Tabel 1. Data hasil *scraping*

No	at	content	score	thumbsUpCount	appVersion
1	12/17/2023 6:58	Sangat membantu	5	0	
2	12/17/2023 6:54	Susah dipakenya, setiap masuk selalu ada di halaman Selamat Datang, dan tdk ada tombol buat login or skip.. Caranya gmn kl mau login? Pas perlu mo pake, jd dipersulit kek gini ðŸ˜¢	2	0	1.7.0
3	12/17/2023 6:53	kenapa kalo tiap mau buka aplikasinya,harus di update terus kalo ga di update aplikasi ga bisa kebuka walau cuma intip saldo doang,,,dr pertama menggunakan livin mandiri setahun yg lalu,,,entah sdh brp kali diupdate terus utk bisa dipake,memakan memory yg tdk sedikit jgðŸ˜”	3	0	
4	12/17/2023 6:48	Kenapa sudah beberapa hari ini saya mau upgrate livin saya...tpi malah status tertunda...jdinya pengen lakukan transaksi jdi terkendali.	3	0	
5	12/17/2023 6:47	Aplikasi ini kok sulit diupdate yaa,,	1	0	1.6.0
…	…	…	…	…	…

The data, initially in CSV format, was then subjected to a preprocessing stage, resulting in a total of 19,530 reviews. This number decreased from the initial dataset due to the removal of several empty reviews identified during the data cleaning process. The final dataset consisted of 9,557 positive sentiment reviews and 9,973 negative sentiment reviews.

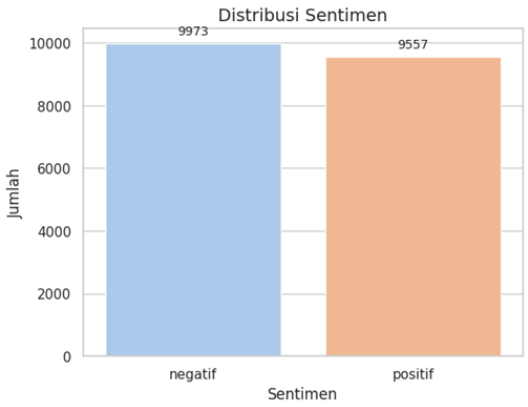

Figure 2. Distribution of Data by Sentiment Category

The difference in the number of reviews resulting from the preprocessing stage is 416, as shown in Figure 2. The relatively insignificant difference in data makes the dataset suitable for inclusion in a machine learning model and is expected to yield high performance. The average review length after preprocessing is 7 words, with the longest review consisting of 90 words and the shortest containing only 1 word.

Tabel 2. Sample of Data After Preprocessing

Original Review	Preprocessed Review	Sentiment
“Meringankan urusan, bagus aku suka”	“ringan urus bagus suka”	1 (Positive)
“Sangat bagus dan mudah digunakan”	“bagus mudah”	1 (Positive)
“Lebih mudah”	“mudah”	1 (Positive)
“Sering gangguan payah”	“ganggu payah”	0 (Negative)
“Kadang susah login”	“kadang susah login”	0 (Negative)

The preprocessing results produced more organized and relevant data, improving the efficiency of sentiment analysis as shown in table 2. This clean and structured dataset facilitates machine learning models in identifying significant sentiment patterns, as noise such as irrelevant words has been removed. With preprocessed data, the models are better equipped to deliver accurate sentiment classification outcomes, thus supporting the overall objectives of the research.

Figure 3. Word cloud negative sentiment

Figure 4. Word cloud positive sentiment

Word cloud visualizations show that users associate negative sentiment with issues such as login errors and update failures, while positive sentiment is linked to ease of use and satisfaction with words like “bagus,” “mudah,” and “mantap.”

Figure 5. Word Cloud Visualization Based on TF-IDF Scores

TF-IDF analysis highlights the most influential words in sentiment classification, with “bagus,” “bantu,” and “mantap” holding the highest weights. These terms reflect the most relevant expressions influencing model predictions.

With the application of the TF-IDF transformation, the review data was successfully represented in the form of vectors that emphasize contextually important words while

reducing the weight of overly common words with low informative value. This representation was then used as input in the sentiment classification stage, which is discussed in the following section.

After obtaining the numerical representation of customer reviews through the TF-IDF process, the next step was to build a classification model to predict the sentiment labels of each review. The cleaned and transformed dataset was divided into two subsets, namely training data (80%) and testing data (20%), with the aim of training the model and objectively evaluating its performance on previously unseen data.

Table 3. Model Evaluation Results

Model Name	Accuracy	F1 Score	Precision	Recall
DT	0.884	0.877	0.906	0.850
KNN	0.773	0.790	0.721	0.873
LR	0.930	0.927	0.943	0.913
RF	0.914	0.907	0.963	0.858
NB	0.860	0.867	0.810	0.933
SVM	0.933	0.930	0.944	0.916
SVM+RF	0.931	0.927	0.949	0.907
SVM+LR	0.933	0.931	0.942	0.920
SVM+RF+LR	0.933	0.930	0.949	0.911

Table 3 summarizes the performance of all tested classification models using four standard evaluation metrics: accuracy, F1 score, precision, and recall. Among the evaluated models, the SVM achieved the best overall performance as a single classifier, with an accuracy of 93.3%, F1 score of 0.930, precision of 0.944, and recall of 0.916. This confirms SVM's robustness and consistency in handling high-dimensional, sparse features generated by TF-IDF for sentiment classification tasks.

LR followed closely with an accuracy of 93.0%, F1 score of 0.927, and slightly higher recall at 0.913, though with a marginally lower precision than SVM. RF performed moderately well accuracy 91.4% and achieved the highest precision of 0.963, but at the expense of lower recall 0.858, indicating that it may be more conservative in identifying positive sentiment cases.

KNN recorded the lowest accuracy 77.3% and F1 score 0.790, revealing its limited effectiveness in this context. While it achieved relatively high recall 0.873, its low precision 0.721 indicates a high rate of false positives. Similarly, NB had the lowest precision 0.810 among all models, although its recall was relatively high 0.933, suggesting it overpredicts the positive class.

Several ensemble experiments were also conducted to evaluate whether combining classifiers could improve performance. The ensemble models SVM+RF, SVM+LR, and SVM+RF+LR all achieved the same accuracy 93.3% as the standalone SVM. Although the ensemble models showed slightly different balances in precision and recall, none of them significantly outperformed the standalone SVM model. For example, SVM+RF achieved higher precision 0.949 but lower recall 0.907, while SVM+LR showed a slight gain in recall 0.920 but slightly lower precision 0.942. These results confirm that while ensemble techniques provide robustness, the SVM model alone offers the best trade-off between simplicity, performance, and computational efficiency.

In conclusion, SVM was selected as the primary model for this study due to its superior and stable performance across all metrics, validating its suitability for sentiment classification of Indonesian-language mobile banking reviews.

5. Conclutions

This study implemented and evaluated several machine learning models to perform sentiment analysis on user reviews of the Livin' by Mandiri mobile banking application, using a real-world Indonesian-language dataset collected from the Google Play Store. After a systematic preprocessing pipeline and feature extraction via TF-IDF with unigram and bigram representations, six classifiers were trained and tested: DT, KNN, LR, RF, NB, and SVM.

The experimental results revealed that the SVM model consistently outperformed other models, achieving the highest F1 score and balanced performance across precision and recall. Although several ensemble models were also evaluated (SVM combined with RF and LR), their performance gains were negligible. Therefore, the standalone SVM model was selected as the most suitable and efficient model for the sentiment classification task in this context.

The findings confirm the effectiveness of traditional machine learning methods particularly SVM for high-dimensional textual data, even without deep learning or neural architectures. Furthermore, the research contributes practically by addressing a real-world application involving sentiment analysis of Indonesian-language reviews in the financial technology domain.

Future work may explore deep learning models such as BERT or aspect-based sentiment analysis to extract more granular insights.

References

[1] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in

Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

[2] Z. K. Chepukaka and F. K. Kirugi, "Service Quality and Customer Satisfaction At Kenya National Archives and Documentation Service, Nairobi County: Servqual Model Revisited," *nternational J. Cust. Relations*, vol. 7, no. 1, pp. 1–14, 2019.

[3] Z. A. Diekson, M. R. B. Prakoso, M. S. Q. Putra, M. S. A. F. Syaputra, S. Achmad, and R. Sutoyo, "Sentiment analysis for customer review: Case study of Traveloka," *Procedia Comput. Sci.*, vol. 216, no. 2022, pp. 682–690, 2022, doi: 10.1016/j.procs.2022.12.184.

[4] P. Mehta and S. Pandya, "A review on sentiment analysis methodologies, practices and applications," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 601–609, 2020.

[5] E. Asani, H. Vahdat-Nejad, and J. Sadri, "Restaurant recommender system based on sentiment analysis," *Mach. Learn. with Appl.*, vol. 6, no. April, p. 100114, 2021, doi: 10.1016/j.mlwa.2021.100114.

[6] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, pp. 1–5, 2019, doi: 10.1109/ICIC47613.2019.8985884.

[7] "Google Play Store." [Online]. Available: https://play.google.com/store/apps

[8] R. Obiedat *et al.*, "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," *IEEE Access*, vol. 10, pp. 22260–22273, 2022, doi: 10.1109/ACCESS.2022.3149482.

[9] Y. Handayani, A. R. Hakim, and Muljono, "Sentiment analysis of Bank BNI user comments using the support vector machine method," *Proc. - 2020 Int. Semin. Appl. Technol. Inf. Commun. IT Challenges Sustain. Scalability, Secur. Age Digit. Disruption, iSemantic 2020*, pp. 202–207, 2020, doi: 10.1109/iSemantic50169.2020.9234230.

[10] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," *2018 IEEE Int. Conf. Innov. Res. Dev. ICIRD 2018*, pp. 1–6, 2018, doi: 10.1109/ICIRD.2018.8376299.

[11] A. R. Prananda and I. Thalib, "Sentiment Analysis for Customer Review: Case Study of GO-JEK Expansion," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 1, p. 1, 2020, doi: 10.20473/jisebi.6.1.1-8.

[12] A. Patel, P. Oza, and S. Agrawal, "Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model," *Procedia Comput. Sci.*, vol. 218, pp. 2459–2467, 2022, doi: 10.1016/j.procs.2023.01.221.

[13] S. Ali, G. Wang, and S. Riaz, "Aspect based sentiment analysis of ridesharing platform reviews for kansei engineering," *IEEE Access*, vol. 8, pp. 173186–173196, 2020, doi: 10.1109/ACCESS.2020.3025823.

[14] M. A. Qureshi *et al.*, "Sentiment Analysis of Reviews in Natural Language: Roman Urdu as a Case Study," *IEEE Access*, vol. 10, no. December 2021, pp. 24945–24954, 2022, doi: 10.1109/ACCESS.2022.3150172.

[15] H. Wisnu, M. Afif, and Y. Ruldevyani, "Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes," *J. Phys. Conf. Ser.*, vol. 1444, no. 1, 2020, doi: 10.1088/1742-6596/1444/1/012034.

[16] K. Divya, B. S. Siddhartha, N. M. Niveditha, and B. M. Divya, "An Interpretation of Lemmatization and Stemming in Natural Language Processing," *J. Univ. Shanghai Sci. Technol.* , vol. 22, no. 10, p. 351, 2020, [Online]. Available: https://www.researchgate.net/publication/348306833

[17] R. Rani and D. K. Lobiyal, "Performance evaluation of text-mining models with Hindi stopwords lists," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 2771–2786, 2022, doi: 10.1016/j.jksuci.2020.03.003.

[18] J. Soegiastuti and A. Review, "THE INFLUENCE OF INFORMATION SYSTEMS AND KNOWLEDGE ON THE IMPLEMENTATION EFFECTIVENESS OF LIVIN' BY MANDIRI APPLICATION," vol. 1, pp. 52–60, 2022.

[19] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Eng.*, vol. 69, pp. 1356–1364, 2014, doi: 10.1016/j.proeng.2014.03.129.

[20] M. A. Rahmat, Indrabayu, and I. S. Areni, "Hoax web detection for news in bahasa using support vector machine," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 332–336, 2019, doi: 10.1109/ICOIACT46704.2019.8938425.

[21] H. J. Kim, J. W. Baek, and K. Chung, "Optimization of associative knowledge graph using TF-IDF based ranking score," *Appl. Sci.*, vol. 10, no. 13, 2020, doi: 10.3390/app10134590.