BAREKENG: Journal of Mathematics and Its Applications
June 2025 Volume 19 Issue 2 Page 0889–0902
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss2pp0889-0902

GOJEK DATA ANALYSIS THROUGH TEXT MINING USING SUPPORT VECTOR MACHINE (SVM) AND K-NEAREST NEIGHBOR (KNN)

Siti Hadijah Hasanah[1*], Muhamad Riyan Maulana[2], Dian Nurdiana[3]

[1]Statistics Study Program, Faculty of Science and Technology, Universitas Terbuka
[2,3]Information Systems Study Program, Faculty of Science and Technology, Universitas Terbuka
Jl. Pd. Cabe Raya, Tangerang Selatan, 15437, Indonesia

*Corresponding author's e-mail: * **sitihadijah@ecampus.ut.ac.id***

Article History:

Received: 31st July 2024
Revised: 24th January 2025
Accepted: 25th February 2025
Published: 1st April 2025

Keywords:

Gojek;
KNN;
SVM;
Text Mining.

ABSTRACT

The main focus of this research is to apply and test the effectiveness of SVM and KNN methods in Gojek data text analysis. This research will examine how the two methods can classify user comments and feedback and identify data sentiment analysis at the same time practically help Gojek understand user needs and improve service quality. The data obtained through scrapping is categorized into positive and negative sentiment. Data is taken from Gojek application user reviews throughout the year 2022 with a total of 1148 sentiment data with a percentage of 80% training data and 20% testing data. Evaluation of model performance using Confusion Matrix and AUC-ROC Curve shows that SVM is more effective than KNN, with accuracy on training data of 92.55% for SVM and 81.71% for KNN, as well as accuracy on testing data of 82.40% for SVM and 77,09% for KNN.

How to cite this article:

S. H. Hasanah, M. R. Maulana and D. Nurdiana., "GOJEK DATA ANALYSIS THROUGH TEXT MINING USING SUPPORT VECTOR MACHINE (SVM) AND K-NEAREST NEIGHBOR (KNN)," *BAREKENG: J. Math. & App.*, vol. 19, iss. 2, pp. 0889-0902, June, 2025.

Journal homepage:*https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article ·**Open Access**

1. INTRODUCTION

The development of information and communication technology has brought significant changes in various aspects of human life [1]. One of the most striking changes is the emergence of digital applications and platforms that make daily activities easier [2]. Gojek, as one of the largest technology companies in Indonesia, has been a pioneer in providing various on-demand services [3]. Starting from transportation services, food delivery, and shopping for daily necessities to financial services, Gojek has succeeded in reaching millions of active users who take advantage of the various services offered [4][5].

With the number of users continuing to increase, the data generated by the Gojek application has become very large and complex [6]. This data encompasses various types of information, such as transaction data, location data, review data, and user feedback [7]. However, the increasing volume and complexity of the data pose challenges in terms of storage, processing, and extracting meaningful insights in a timely manner [8]. To address this, Gojek has implemented several data management strategies, including the use of big data technologies and machine learning algorithms [9]. These approaches aim to optimize operations, improve service quality, and support decision-making processes [10].

Despite these efforts, challenges remain, especially in understanding and leveraging user feedback effectively [11]. Sentiment analysis, as one of the methods for analyzing textual data, can be an essential tool to interpret user reviews and feedback systematically [12][13]. By applying sentiment analysis, Gojek can identify patterns in user satisfaction, detect potential service issues, and understand customer expectations more accurately. Therefore, this study explores how sentiment analysis can contribute to overcoming the challenges of managing and analyzing large-scale user feedback data while aligning with the goal of enhancing user experience and operational efficiency.One method that can be used is Text Mining, which focuses on extracting valuable information from unstructured text [14].

Text Mining has experienced rapid development along with advances in machine learning and artificial intelligence technology [15]–[17]. This technique enables the automatic analysis of large amounts of text data, identifying patterns and trends that are useful for decision making [18][19]. Text mining can also be used to classify and group text, as well as to determine the sentiment or emotion contained in the text [20]–[23]. Various industries have utilized Text Mining for multiple purposes, such as sentiment analysis, document classification, and fraud detection [24][25]. In the context of customer service, Text Mining is used to analyze reviews and feedback from users, so that companies can better understand their customers' needs and preferences [26][27].

Various methods and algorithms have been developed to support the application of Text Mining. Two of the most popular are Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) [28]–[30]. SVM is a robust machine learning algorithm in classification and regression, especially in handling extensive data with many features [31]. Meanwhile, KNN is a simple but effective algorithm in classifying data based on proximity or similarity to other data whose categories are already known [32][33].

Previous research has shown that the SVM and KNN methods can provide good results in text analysis, particularly in terms of classification accuracy and computational efficiency. For example, studies comparing multiple machine learning methods, including Naive Bayes, Decision Trees, ANN and Random Forest, have demonstrated that SVM and KNN often outperform others in specific contexts, such as sentiment analysis and document classification [34]–[39]. For example, previous studies have successfully applied SVM and KNN in various domains such as social media sentiment analysis, scientific document classification, and spam detection. e-mail [40]–[42]. One of the studies by Muslim (2020) shows that using unigrams for feature extraction and grid search for parameter optimization in the SVM algorithm can increase the accuracy of customer review classification on e-commerce platforms. In this study, two datasets of reviews from Amazon (English) and Lazada (Indonesian) were used. The experimental results showed that the accuracy of Amazon reviews increased by 26.4%, while Lazada reviews increased by 4.26%. Meanwhile, research conducted by [41] showed findings that the proposed KNN-based Lao news text classification method, through preprocessing and feature extraction, KNN parameter adjustment, data normalization and dimension reduction, effectively improved text classification accuracy. Based on the research results above, it shows the fact that these two methods have their respective advantages in processing and analyzing text data.

However, research that specifically examines the application of these two methods in the context of Gojek data is still minimal. Most previous research focused more on analyzing data from different domains, such as social media or e-commerce [44][45]. Therefore, this research aims to fill this gap by comparing

SVM and KNN algorithms in analyzing Gojek user reviews and feedback texts. The choice of SVM and KNN is based on their proven performance in handling text classification tasks in previous studies, where both methods demonstrated strengths in accuracy and computational efficiency. By comparing these algorithms, this study seeks to identify which approach is more suitable for analyzing Gojek's large and complex textual data, specifically in the context of sentiment analysis.While sentiment analysis is a prominent application of text mining, the two are not identical. Text mining encompasses a broader scope, involving various techniques to extract meaningful patterns and insights from unstructured text data. Sentiment analysis, as a subset of text mining, focuses on determining the sentiment or emotional tone conveyed in the text. In this study, sentiment analysis serves as the primary focus to understand user satisfaction and feedback, contributing to improved service quality and user experience.Thus, this research not only expands the application of existing methods but also covers the gaps in previous research in the context of on-demand services such as Gojek.

The research gap identified is the limited number of studies that simultaneously examine the application of SVM and KNN methods in text analysis specific to Gojek data. While numerous studies have investigated each technique independently, very few have directly compared their performance within the same study context, particularly in the sentiment analysis of service-based platforms like Gojek. Previous research often focuses on general text analysis or alternative domains, such as social media or e-commerce, with limited emphasis on multi-service applications. This study addresses this gap by conducting a comprehensive comparative analysis of SVM and KNN for Gojek data, ensuring the findings are well-positioned to solve the stated problems. By evaluating the strengths and limitations of both methods, this research is expected to provide significant contributions to the advancement of Text Mining and machine learning in service-based sentiment analysis.

The main focus of this research is to apply and test the effectiveness of the SVM and KNN methods in text analysis of Gojek data. This research will examine how these two methods can classify user reviews and feedback and identify the sentiment. Researchers will experiment with various parameter configurations to find the most optimal combination. Apart from that, this research will also compare the performance of the two methods to determine the most effective and efficient method for text analysis of Gojek data.

.The urgency of this research is supported by data indicating the critical role of customer feedback in improving service quality for on-demand platforms like Gojek. Studies have shown that user satisfaction significantly influences the retention and growth of on-demand service platforms **[46][47]**. With the increasing number of Gojek users, analyzing user reviews and feedback becomes a necessity to maintain competitive advantage and ensure service quality aligns with user expectations. Furthermore, the volume and complexity of textual feedback generated by Gojek users present an opportunity to advance the application of Text Mining and machine learning techniques. This research contributes to filling the gap in the literature by demonstrating the effectiveness of sentiment analysis using SVM and KNN algorithms in a real-world context, specifically for on-demand services. By grounding the urgency in empirical evidence and its contribution to both industry and academia, this study emphasizes its relevance and significance.

This research has significant implications both from an academic and practical perspective. Academically, this research will add to the existing literature regarding the application of Text Mining with the SVM and KNN methods and provide new insights into how these two methods can be optimized in Gojek data analysis. This research can also be a reference for other researchers interested in further studying the Text Mining application in the context of on-demand services. Practically, this research's results can help Gojek better understand user needs and preferences, improve service quality, and identify problems proactively. By analyzing user reviews and feedback, Gojek can identify areas that need improvement and take appropriate steps to increase user satisfaction. In addition, the results of this research can also be used by Gojek to design more effective marketing and promotional strategies, based on patterns and trends identified from user review data.

The main objective of this research is to explore and evaluate the application of SVM and KNN methods in text analysis of Gojek data. This research aims to identify the most effective methods for classifying user reviews and feedback and determining their sentiment. Thus, it is hoped that this research can significantly contribute to the development of knowledge in the fields of Text Mining and machine learning, as well as providing practical benefits for the development of Gojek services. This research is also expected to provide new insights for Gojek in understanding user needs and preferences to improve service quality and overall user satisfaction.

2. RESEARCH METHODS

The research data comes from scrapping results on the Gojek application throughout 2022, with values 1 and 2 categorized as negative sentiment analysis, while values 3, 4, and 5 are categorized as positive sentiment analysis. A detailed explanation of the research flow diagram is summarized in **Figure 1** above as described below.

2.1 Start

The initial stages of this research are preparation and planning. At this stage, the researcher determines the research objectives, designs the methodology, and identifies the data for use. The main objective of this research is to evaluate and compare the effectiveness of the Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) methods in text analysis of Gojek data.

2.2 Data

The next stage is data collection. The data used in this research consists of reviews and feedback from 1178 Gojek application users from January 2022 to December 2023. This data is obtained from various sources such as Gojek's internal database or scraping reviews from online platforms. After the data is collected, the next step is data preprocessing, including text cleaning, tokenization, and feature extraction to ensure the data is ready for analysis.

2.3 Data Sharing: Training Data (80%) and Testing Data (20%)

Once the data is ready, the next step is to divide the data into two sets: training data and testing data. This distribution is usually carried out randomly with a proportion of 80% for training data and 20% for testing data from data cleaning . Training data will be used to train the classification model, while testing data will be used to test and evaluate the performance of the model that has been trained.

2.4 Feature Selection

At this stage, the features used in the classification model are selected. Feature selection selects a subset of the most relevant features to improve model performance. The selected features can be words, phrases, or other attributes that are considered important for the classification of user reviews and feedback **[48]**.

2.5 Classification Method

Once the features are selected, the next step is to apply the classification method. In this research, the two classification methods that will be applied are Support Vector Machine (SVM) and K-Nearest Neighbor (KNN).

1. SVM (Support Vector Machine)

   SVM is a powerful machine learning algorithm for classification tasks **[31]**, **[49]**. This algorithm finds the best hyperplane that separates data into different classes. SVM is effective in handling high-dimensional data and is often used in text analysis due to its ability to overcome overfitting **[50]**–**[52]**.

2. KNN (K-Nearest Neighbor)

   KNN is a simple but effective classification algorithm. This algorithm works by classifying new data based on its proximity to data whose categories are already known **[53]**–**[55]**. KNN classifies data based on the feature space's majority class of the K nearest neighbors **[39]**, **[56]**. KNN is suitable for small to medium datasets and does not require strict assumptions about data distribution**[57]**.

2.6 Evaluation: Confusion Matrix and AUC-ROC Curve

Once the SVM and KNN models are trained and tested, the performance of both models is evaluated using two key metrics:

1. Confusion Matrix

Confusion matrix is an evaluation tool that shows the performance of a classification model by presenting the number of correct and incorrect predictions in detail. This matrix consists of four elements: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). From this matrix, other evaluation metrics can be calculated such as accuracy, precision, recall, and F1- score **[58]**–**[60]**.

2. AUC-ROC Curve

Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) curve are evaluation tools used to measure the performance of binary classification models. ROC curve is a graph that shows the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) at various thresholds. AUC is the area under the ROC curve and indicates how well the model can differentiate between classes. The greater the AUC value, the better the model performance **[61][62]**.

2.7 End

The final stage of the research is to summarize the results obtained. Researchers will compare the performance of SVM and KNN based on the calculated evaluation metrics. The results of this research will be analyzed to determine the most effective and efficient method for classifying Gojek user reviews and feedback. In addition, researchers will also identify potential improvements and next steps for future research. It is hoped that the results of this research can significantly contribute to the development of Gojek services and provide new insights in the fields of Text Mining and machine learning.

In this research, researchers have adopted a systematic approach to analyzing the sentiment of Gojek application users using two classification methods: Support Vector Machine (SVM) and K-nearest neighbor (KNN). The process begins with data collection and cleaning, dividing the data into proportional training and testing sets. By applying appropriate feature selection techniques, researchers can ensure that the trained model has an optimal data representation.

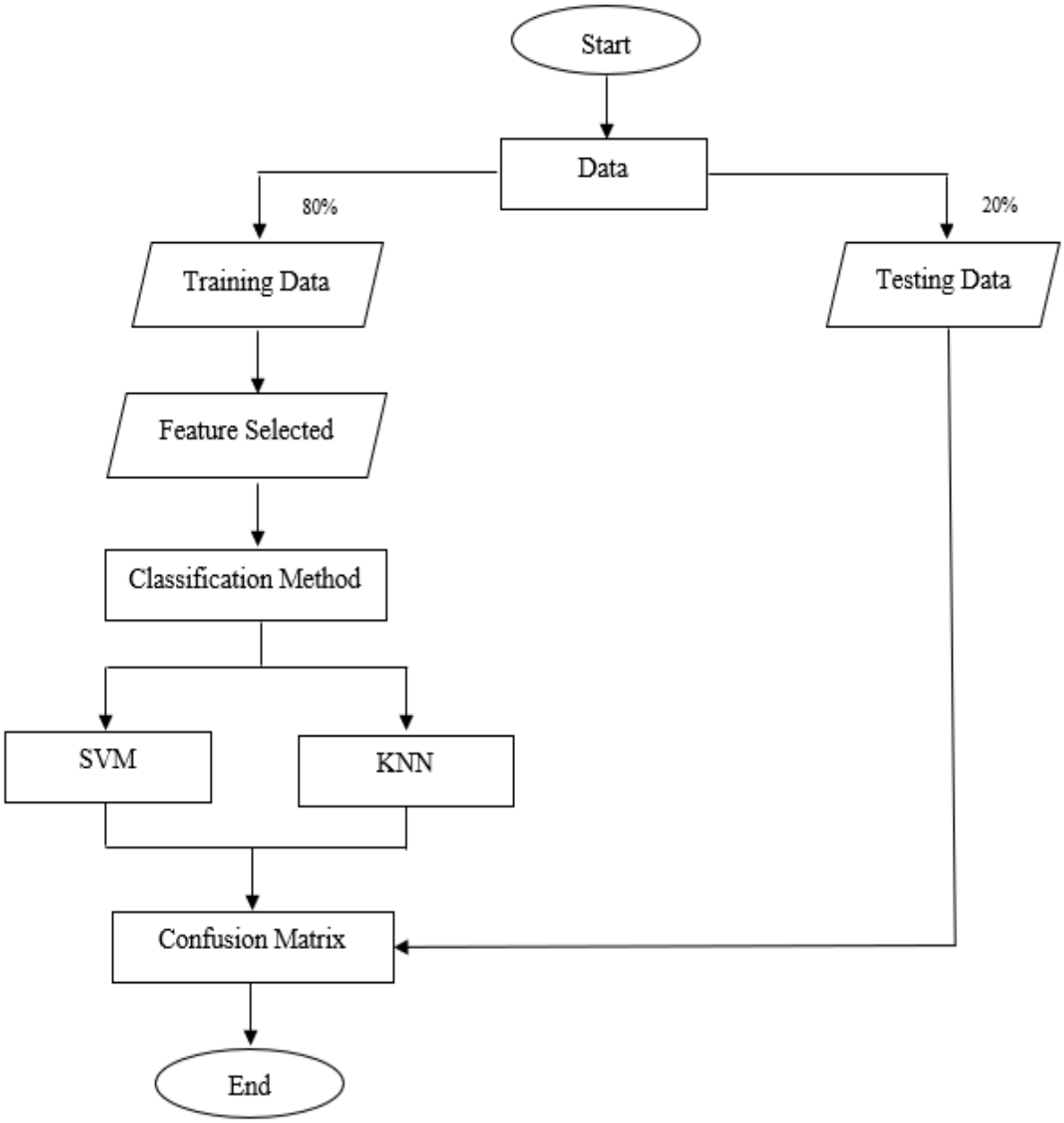

Figure 1. Research Flow Diagram

Figure 1 is a flowchart that illustrates the workflow process in a classification project using two methods, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The first step begins with dividing the dataset into two parts, training data (80%) and test data (20%). Training data is used for the feature selection process to identify relevant and important features for the classification model. After selecting the features, the SVM and KNN classification methods are applied to the training data to build a prediction model.

Next, the model that has been trained using the training data will be evaluated with the test data. The evaluation is carried out using the Confusion Matrix, which functions to measure the performance of the classification model. The Confusion Matrix provides information about the number of correct predictions (true positive and true negative) and incorrect predictions (false positive and false negative). This flow ends after the model performance is evaluated and the analysis results from the Confusion Matrix are obtained.

3. RESULTS AND DISCUSSION

The following are the results of the descriptive analysis carried out.

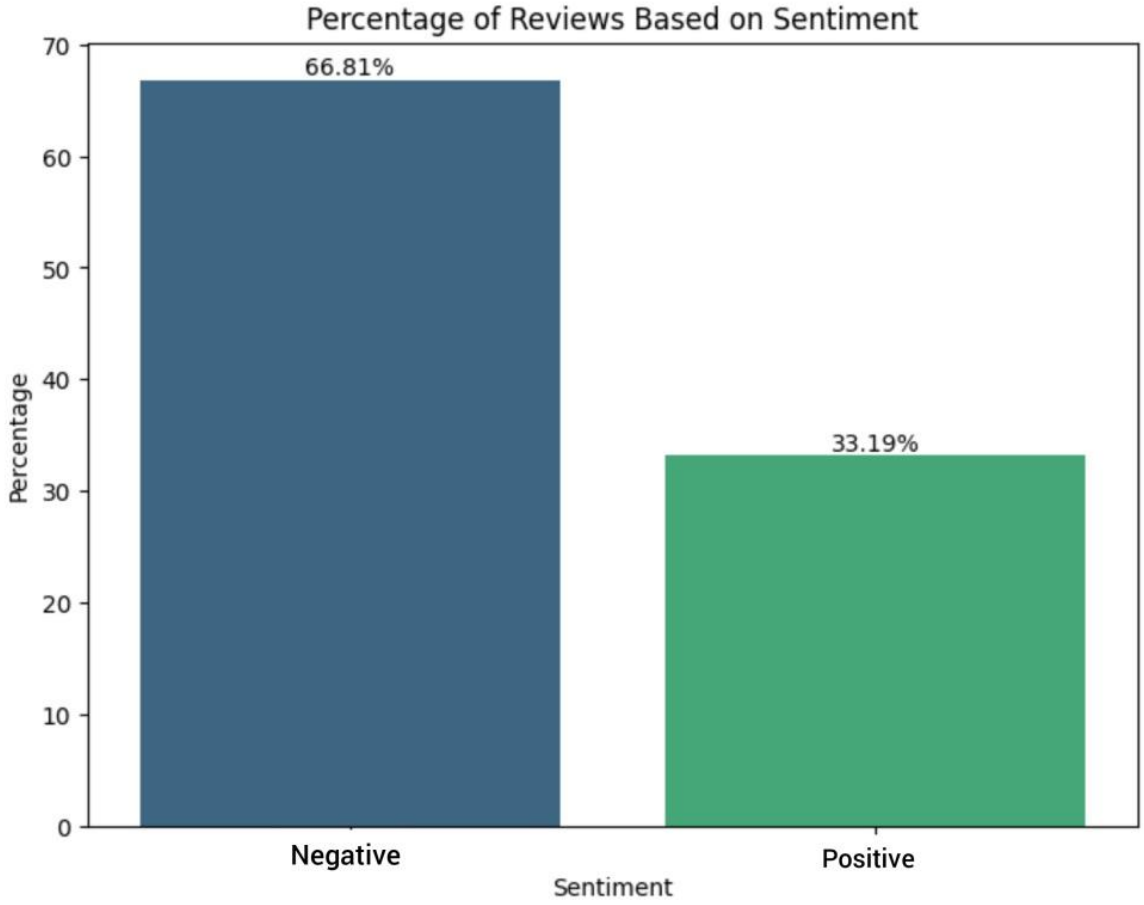

Figure 1. Sentiment Analysis Results

Based on the analysis results shown in **Figure 2**, most Gojek application users positively respond to the services provided. Specifically, 66.81% of all reviews analyzed showed positive sentiment. This indicates that most users are satisfied with the services they receive, both in terms of service quality, ease of use of the application, and responsiveness of the service provider.

On the other hand, several reviews also showed negative sentiment, reaching 33.19% of the total reviews analyzed. This percentage shows that although most users are satisfied, several users still experience dissatisfaction or problems using Gojek services. Further analysis of these negative reviews is important to identify areas Gojek needs to improve to improve the overall user experience.

Next, **Figure 3** below displays the percentage of the top 20 words most frequently mentioned in user reviews of the Gojek application. From this analysis, the words " driver " and "application" emerged as the most frequently mentioned words, each with a percentage of 0.66%. This shows that interaction with the driver and the experience of using the application are the two aspects most often discussed by users. The high frequency of the word " driver " can illustrate the importance of the driver's role in the overall user experience. In contrast, the high frequency of "application" emphasizes the importance of the features and user interface in the Gojek application.

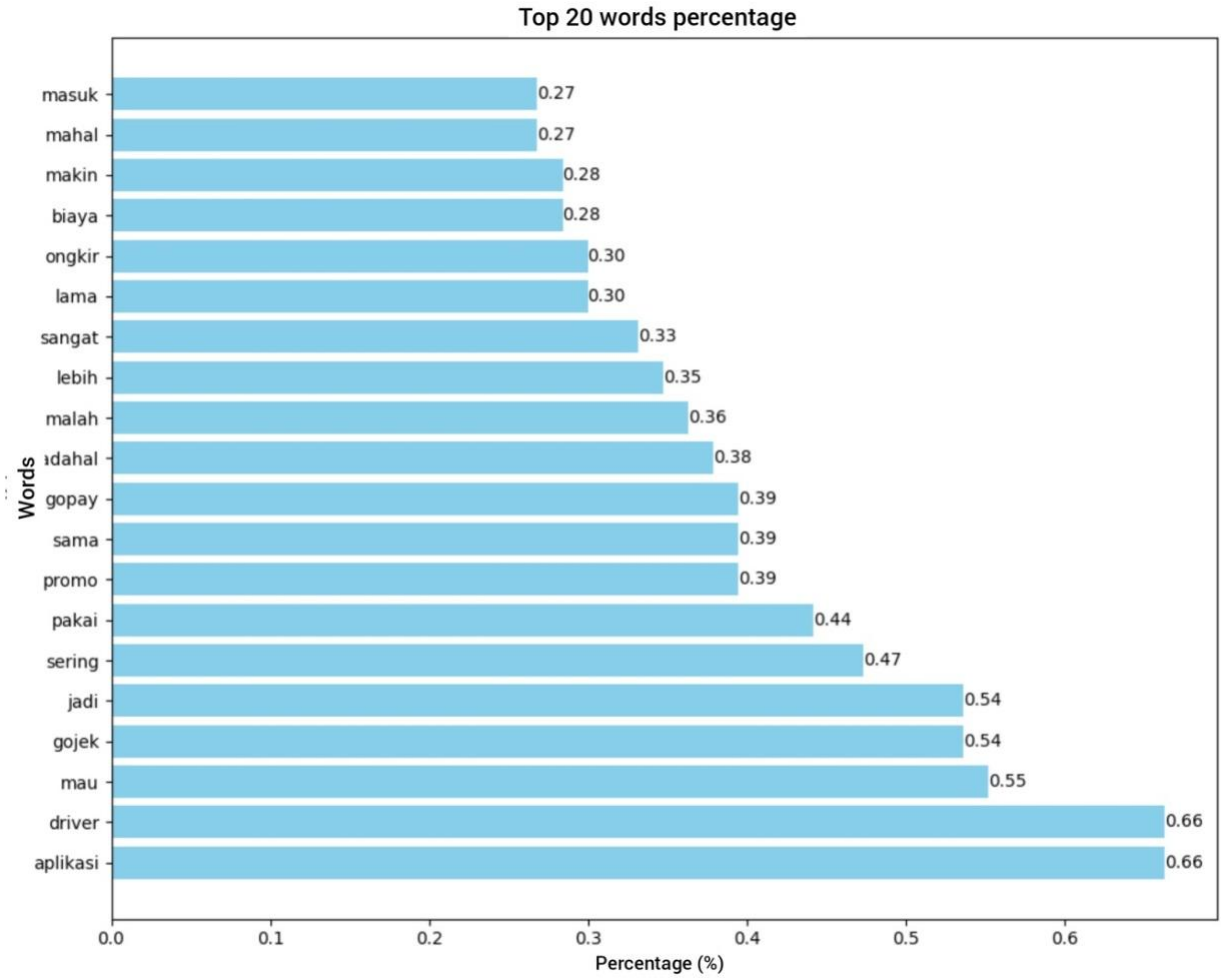

Figure 2. Top 20-Word Results

To provide a clearer visual picture of the distribution and frequency of frequently used words in user reviews, a wordcloud technique was used, the results of which are shown in **Figure 4** below. This wordcloud presents the words that appear most frequently in reviews with a font size proportional to their frequency of appearance. From this wordcloud, we can easily identify key words often used by users when providing reviews of the Gojek application.

Figure 4. Wordcloud Data Gojek

This descriptive analysis provides valuable insight into how users respond to the services provided by Gojek. By understanding user sentiment and words frequently appearing in reviews, Gojek can identify the strengths and weaknesses of its services. This information can be used as a basis for making strategic decisions to improve service quality, increase user satisfaction, and ultimately, strengthen Gojek's position in the market. This research also highlights the importance of text analysis in understanding user behavior and preferences. By using text analysis techniques such as sentiment analysis and wordcloud, companies can gain deep insights into user views and experiences that may not be revealed through traditional survey methods [20], [63]–[65]. Therefore, applying text analysis in customer service management can be a very effective tool to improve service quality and maintain customer satisfaction continuously [66]–[68].

After analyzing the results from the wordcloud, which provides an overview of the keywords that appear most frequently in user reviews, we can bridge these findings with the results obtained from the training and testing data analysis. Table 1 provides a clear picture of the effectiveness of each method in classifying training data. The SVM method shows a very high accuracy percentage of 92.55%. This shows that SVM effectively recognizes existing data patterns to minimize classification errors. In contrast, KNN shows a lower accuracy of 81.71%, which reflects that although KNN is a simple and easy to understand method, it is less effective in handling large and complex datasets such as the one used in this study.

Table 1. Training Data

Method/	SVM		KNN	
Response	Positive (%)	Negative (%)	Positive (%)	Negative (%)
Positive	89.81	4.71	82.19	18.78
Negative	10.19	95.29	17.81	81.22
Accuracy	92.55		81.71	

Based on Table 1, The percentage of positives successfully identified by SVM reached 89.81%, with only 4.71% misclassified as negative. Meanwhile, KNN could classify 82.19% of positive responses, but with 18.78% classification errors. In this case, SVM has a higher accuracy level and shows better consistency in predicting positive responses. In contrast, KNN has a larger proportion of errors, indicating that this model is more susceptible to noise in the data.

These results are very significant in the context of the Gojek application, where understanding user sentiment is key to improving services. With the high accuracy of SVM, companies can be more confident in making data-based decisions that rely on user sentiment analysis. SVM methods can direct service improvement efforts based on more accurate and relevant feedback, while KNN may require more setup and data preprocessing to achieve similar results.

Table 2. Percentage of Testing Data

Method/	SVM		KNN	
Response	Positive (%)	Negative (%)	Positive (%)	Negative (%)
Positive	78.76	13.95	74.61	32.56
Negative	21.24	86.05	18.65	97.67
Accuracy	82.40		77.09	

In addition, Table 2 provides more detailed information regarding each method's performance on testing data. SVM recorded an accuracy percentage of 82.40%, which shows that this method still shows solid performance even though it is in the testing phase. On the other hand, KNN has lower accuracy, namely 77.09%. This difference in accuracy confirms that SVM can better handle data complexity and provide more reliable results.

In terms of percentage of positive responses, SVM recorded 78.76%, with 13.95% of misclassifications being negative. KNN, on the other hand, recorded 74.61% for positive responses, but had a higher error rate at 32.56%. This suggests that KNN tends to misclassify large amounts of positive data as negative, which can be detrimental in sentiment analysis, where capturing the positive nuances of user feedback is important.

The results of this test provide significant insight for Gojek in understanding and responding to user sentiment. With SVM, which is more accurate, Gojek can be more effective in making strategic decisions

based on sentiment analysis. For example, if SVM shows high satisfaction, Gojek can maintain its existing marketing strategy. Conversely, if there is an increase in negative responses, the management team can immediately respond with corrective action.

4. CONCLUSIONS

This research analyzed and compared the performance of Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) in classifying user sentiment for the Gojek application. The findings demonstrate that SVM achieved superior accuracy in both training (92.55%) and testing (82.40%) phases, highlighting its capability to handle complex datasets effectively. KNN, while slightly less accurate (training: 81.71%, testing: 77.09%), still provided competitive results, affirming its potential as a simpler alternative for certain sentiment analysis tasks.

However, the study acknowledges the limitation of using these two methods independently, as combining their strengths could yield better performance. This insight emphasizes the need for future research to explore hybrid approaches or ensemble techniques, such as integrating SVM and KNN, to improve overall accuracy and robustness. Furthermore, while this study focused on SVM and KNN due to their reliability in sentiment analysis, subsequent research may benefit from re-evaluating the algorithm selection process and testing other robust methods such as Random Forests (RF) and Neural Networks (NN) to verify their suitability within the same context.Lastly, the importance of aligning the choice of algorithms with the dataset's characteristics and research objectives was reinforced. Future work should incorporate larger datasets and advanced NLP techniques, such as contextual processing, to enhance the understanding and utility of user sentiment analysis for practical applications like personalized recommendation systems.

ACKNOWLEDGMENT

The research team would like to express our deepest gratitude to the individuals and institutions that have made important contributions to the success of our research. First, we would like to express our deepest appreciation to the Institute for Research and Community Service of the Open University (LPPM-UT) for the financial support in the internal research grant with research contract number: B/592/UN31.LPPM/PT.01.03/2023. The support from LPPM-UT has enabled us to engage students and carry out this research with dedication. We would also like to thank the journal for giving us the opportunity to publish the results of this research. The support from the journal provides space for the dissemination of the results of this research to a wider scientific realm.

REFERENCES

[1] C. Zhang, I. Khan, V. Dagar, A. Saeed, and M. W. Zafar, "ENVIRONMENTAL IMPACT OF INFORMATION AND COMMUNICATION TECHNOLOGY: UNVEILING THE ROLE OF EDUCATION IN DEVELOPING COUNTRIES," *Technol. Forecast. Soc. Change*, vol. 178, no. January, p. 121570, 2022, doi: 10.1016/j.techfore.2022.121570.

[2] M. Jovanovic, D. Sjödin, and V. Parida, "CO-EVOLUTION OF PLATFORM ARCHITECTURE, PLATFORM SERVICES, AND PLATFORM GOVERNANCE: EXPANDING THE PLATFORM VALUE OF INDUSTRIAL DIGITAL PLATFORMS," *Technovation*, vol. 118, p. 102218, Dec. 2022, doi: 10.1016/j.technovation.2020.102218.

[3] L. Dessyanawaty and Y.-S. Yen, "AN OPTIMIZING OMNI-CHANNEL STRATEGY FOR RIDE-HAILING COMPANIES: THE CASE OF GOJEK IN INDONESIA," *Adv. Manag. Appl. Econ.*, vol. 10, no. 1, pp. 51–59, 2020.

[4] K. A. D. Putra, F. Hidayatullah, and N. Farida, "MEDIATISASI LAYANAN PESAN ANTAR MAKANAN DI INDONESIA MELALUI APLIKASI GO-FOOD," *Islam. Commun. J.*, vol. 5, no. 1, p. 114, Jun. 2020, doi: 10.21580/icj.2020.5.1.5416.

[5] N. E. Kartika, "FITUR APLIKASI GOJEK FAVORIT KONSUMEN PADA SAAT PANDEMI COVID-19 DI KOTA BANDUNG," *J. Communio J. Jur. Ilmu Komun.*, vol. 9, no. 2, pp. 1680–1695, Nov. 2020, doi: 10.35508/jikom.v9i2.2922.

[6] U. S, "ANALYSIS OF BIG DATA UTILIZATION IN TECHNOLOGY COMPANIES (GOJEK CASE STUDY: PT GOTO GOJEK TOKOPEDIA Tbk)," *J. Bus. Manag. Stud.*, vol. 4, no. 4, pp. 92–96, Sep. 2022, doi: 10.32996/jbms.2022.4.4.8.

[7] S. T. Muhammad Wali *et al.*, PENERAPAN & IMPLEMENTASI BIG DATA DI BERBAGAI SEKTOR (PEMBANGUNAN BERKELANJUTAN ERA INDUSTRI 4.0 DAN SOCIETY 5.0). PT. Sonpedia Publishing Indonesia, 2023.

[8] A. Rehman, S. Naz, and I. Razzak, "LEVERAGING BIG DATA ANALYTICS IN HEALTHCARE ENHANCEMENT:

TRENDS, CHALLENGES AND OPPORTUNITIES," *Multimed. Syst.*, vol. 28, no. 4, pp. 1339–1371, Aug. 2022, doi: 10.1007/s00530-020-00736-8.

[9] J. Basukie, Y. Wang, and S. Li, "BIG DATA GOVERNANCE AND ALGORITHMIC MANAGEMENT IN SHARING ECONOMY PLATFORMS: A CASE OF RIDESHARING IN EMERGING MARKETS," *Technol. Forecast. Soc. Change*, vol. 161, p. 120310, Dec. 2020, doi: 10.1016/j.techfore.2020.120310.

[10] Q. A. Nisar, N. Nasir, S. Jamshed, S. Naz, M. Ali, and S. Ali, "BIG DATA MANAGEMENT AND ENVIRONMENTAL PERFORMANCE: ROLE OF BIG DATA DECISION-MAKING CAPABILITIES AND DECISION-MAKING QUALITY," *J. Enterp. Inf. Manag.*, vol. 34, no. 4, pp. 1061–1096, Jul. 2021, doi: 10.1108/JEIM-04-2020-0137.

[11] S. V. Tsiu, L. Mathabela, and M. Ngobeni, "APPLICATIONS AND COMPETITIVE ADVANTAGES OF DATA MINING AND BUSINESS INTELLIGENCE IN SMES PERFORMANCE: A SYSTEMATIC REVIEW," *Available at SSRN 4958874*. 2024. doi: 10.2139/ssrn.4958874.

[12] A. Ligthart, C. Catal, and B. Tekinerdogan, "SYSTEMATIC REVIEWS IN SENTIMENT ANALYSIS: A TERTIARY STUDY," *Artif. Intell. Rev.*, vol. 54, no. 7, pp. 4997–5053, Oct. 2021, doi: 10.1007/s10462-021-09973-3.

[13] P. K. Jain, R. Pamula, and G. Srivastava, "A SYSTEMATIC LITERATURE REVIEW ON MACHINE LEARNING APPLICATIONS FOR CONSUMER SENTIMENT ANALYSIS USING ONLINE REVIEWS," *Comput. Sci. Rev.*, vol. 41, p. 100413, Aug. 2021, doi: 10.1016/j.cosrev.2021.100413.

[14] Q. Qiu, Z. Xie, L. Wu, and L. Tao, "AUTOMATIC SPATIOTEMPORAL AND SEMANTIC INFORMATION EXTRACTION FROM UNSTRUCTURED GEOSCIENCE REPORTS USING TEXT MINING TECHNIQUES," *Earth Sci. Informatics*, vol. 13, no. 4, pp. 1393–1410, Dec. 2020, doi: 10.1007/s12145-020-00527-9.

[15] S. Raschka, J. Patterson, and C. Nolet, "MACHINE LEARNING IN PYTHON: MAIN DEVELOPMENTS AND TECHNOLOGY TRENDS IN DATA SCIENCE, MACHINE LEARNING, AND ARTIFICIAL INTELLIGENCE," *Information*, vol. 11, no. 4, p. 193, Apr. 2020, doi: 10.3390/info11040193.

[16] A. K. Kushwaha, A. K. Kar, and Y. K. Dwivedi, "APPLICATIONS OF BIG DATA IN EMERGING MANAGEMENT DISCIPLINES: A LITERATURE REVIEW USING TEXT MINING," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100017, Nov. 2021, doi: 10.1016/j.jjimei.2021.100017.

[17] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "TEXT MINING IN BIG DATA ANALYTICS," *Big Data Cogn. Comput.*, vol. 4, no. 1, p. 1, 2020.

[18] D. Tao, P. Yang, and H. Feng, "UTILIZATION OF TEXT MINING AS A BIG DATA ANALYSIS TOOL FOR FOOD SCIENCE AND NUTRITION," *Compr. Rev. Food Sci. Food Saf.*, vol. 19, no. 2, pp. 875–894, Mar. 2020, doi: 10.1111/1541-4337.12540.

[19] S. Kumar, A. K. Kar, and P. V. Ilavarasan, "APPLICATIONS OF TEXT MINING IN SERVICES MANAGEMENT: A SYSTEMATIC LITERATURE REVIEW," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, p. 100008, 2021.

[20] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "TEXT PREPROCESSING FOR TEXT MINING IN ORGANIZATIONAL RESEARCH: REVIEW AND RECOMMENDATIONS," *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, 2022.

[21] P. Nandwani and R. Verma, "A REVIEW ON SENTIMENT ANALYSIS AND EMOTION DETECTION FROM TEXT," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 81, 2021.

[22] K. Thakur and V. Kumar, "APPLICATION OF TEXT MINING TECHNIQUES ON SCHOLARLY RESEARCH ARTICLES: METHODS AND TOOLS," *New Rev. Acad. Librariansh.*, vol. 28, no. 3, pp. 279–302, 2022.

[23] C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi, and M. Cannataro, "SENTIMENT ANALYSIS FOR MINING TEXTS AND SOCIAL NETWORKS DATA: METHODS AND TOOLS," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 1, p. e1333, 2020.

[24] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, "COMPREHENSIVE REVIEW OF TEXT-MINING APPLICATIONS IN FINANCE," *Financ. Innov.*, vol. 6, no. 1, p. 39, Dec. 2020, doi: 10.1186/s40854-020-00205-1.

[25] A. K. S. Yadav and M. Sora, "FRAUD DETECTION IN FINANCIAL STATEMENTS USING TEXT MINING METHODS: A REVIEW," in *IOP conference series: Materials science and engineering*, IOP Publishing, 2021, p. 12012.

[26] F. R. Lucini, L. M. Tonetto, F. S. Fogliatto, and M. J. Anzanello, "TEXT MINING APPROACH TO EXPLORE DIMENSIONS OF AIRLINE CUSTOMER SATISFACTION USING ONLINE CUSTOMER REVIEWS," *J. Air Transp. Manag.*, vol. 83, p. 101760, 2020.

[27] L. Serrano, A. Ariza-Montes, M. Nader, A. Sianes, and R. Law, "EXPLORING PREFERENCES AND SUSTAINABLE ATTITUDES OF AIRBNB GREEN USERS IN THE REVIEW COMMENTS AND RATINGS: A TEXT MINING APPROACH," in *Sustainable Consumer Behaviour and the Environment*, Routledge, 2021, pp. 114–132.

[28] E. Y. Boateng, J. Otoo, and D. A. Abaye, "BASIC TENETS OF CLASSIFICATION ALGORITHMS K-NEAREST-NEIGHBOR, SUPPORT VECTOR MACHINE, RANDOM FOREST AND NEURAL NETWORK: A REVIEW," *J. Data Anal. Inf. Process.*, vol. 8, no. 4, pp. 341–357, 2020.

[29] A. Salma and W. Silfianti, "SENTIMENT ANALYSIS OF USER REVIEWS ON COVID-19 INFORMATION APPLICATIONS USING NAIVE BAYES CLASSIFIER, SUPPORT VECTOR MACHINE, AND K-NEAREST NEIGHBOR," *Int. Res. J. Adv. Eng. Sci.*, vol. 6, no. 4, pp. 158–162, 2021.

[30] M. Bansal, A. Goyal, and A. Choudhary, "A COMPARATIVE ANALYSIS OF K-NEAREST NEIGHBOR, GENETIC, SUPPORT VECTOR MACHINE, DECISION TREE, AND LONG SHORT TERM MEMORY ALGORITHMS IN MACHINE LEARNING," *Decis. Anal. J.*, vol. 3, p. 100071, 2022.

[31] D. M. Abdullah and A. M. Abdulazeez, "MACHINE LEARNING APPLICATIONS BASED ON SVM CLASSIFICATION A REVIEW," *Qubahan Acad. J.*, vol. 1, no. 2, pp. 81–90, 2021.

[32] M. Arhami, M. Kom, and S. T. Muhammad Nasir, *DATA MINING-ALGORITMA DAN IMPLEMENTASI*. Penerbit Andi, 2020.

[33] H. Wang, P. Xu, and J. Zhao, "IMPROVED KNN ALGORITHMS OF SPHERICAL REGIONS BASED ON CLUSTERING AND REGION DIVISION," *Alexandria Eng. J.*, vol. 61, no. 5, pp. 3571–3585, 2022.

[34] S. U. Hassan, J. Ahamed, and K. Ahmad, "ANALYTICS OF MACHINE LEARNING-BASED ALGORITHMS FOR TEXT CLASSIFICATION," *Sustain. Oper. Comput.*, vol. 3, no. July 2021, pp. 238–248, 2022, doi: 10.1016/j.susoc.2022.03.001.

[35] S. H. Hasanah, M. Permatasari, and C. Author, "BACKPROPAGATION ARTIFICIAL NEURAL NETWORK

CLASSIFICATION METHOD IN STATISTICS STUDENTS OF OPEN UNIVERSITY," *BAREKENG J. Ilmu Mat. dan Terap.* , vol. 14, no. 2, pp. 243–252, 2020, [Online]. Available: https://ojs3.unpatti.ac.id/index.php/barekeng/

[36] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "IMPLEMENTATION OF K-NEAREST NEIGHBOR (K-NN) ALGORITHM FOR PUBLIC SENTIMENT ANALYSIS OF ONLINE LEARNING," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, p. 121, 2021, doi: 10.22146/ijccs.65176.

[37] S. H. Hasanah, "CLASSIFICATION SUPPORT VECTOR MACHINE IN BREAST CANCER PATIENTS," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 1, pp. 129–136, Mar. 2022, doi: 10.30598/barekengvol16iss1pp129-136.

[38] N. Kalcheva, M. Karova, and I. Penev, "COMPARISON OF THE ACCURACY OF SVM KEMEL FUNCTIONS IN TEXT CLASSIFICATION," in *2020 International Conference on Biomedical Innovations and Applications (BIA)*, IEEE, 2020, pp. 141–145.

[39] X. Luo, "EFFICIENT ENGLISH TEXT CLASSIFICATION USING SELECTED MACHINE LEARNING TECHNIQUES," *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021, doi: 10.1016/j.aej.2021.02.009.

[40] A. A. Akinyelu, "ADVANCES IN SPAM DETECTION FOR EMAIL SPAM, WEB SPAM, SOCIAL NETWORK SPAM, AND REVIEW SPAM: ML-BASED AND NATURE-INSPIRED-BASED TECHNIQUES," *J. Comput. Secur.*, vol. 29, no. 5, pp. 473–529, 2021.

[41] H. Najadat, M. A. Alzubaidi, and I. Qarqaz, "DETECTING ARABIC SPAM REVIEWS IN SOCIAL NETWORKS BASED ON CLASSIFICATION ALGORITHMS," *Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 1, pp. 1–13, 2021.

[42] S. Kaddoura, G. Chandrasekaran, D. E. Popescu, and J. H. Duraisamy, "A SYSTEMATIC LITERATURE REVIEW ON SPAM CONTENT DETECTION AND CLASSIFICATION," *PeerJ Comput. Sci.*, vol. 8, p. e830, 2022.

[43] Z. Chen, L. J. Zhou, X. Da Li, J. N. Zhang, and W. J. Huo, "THE LAO TEXT CLASSIFICATION METHOD BASED ON KNN," *Procedia Comput. Sci.*, vol. 166, pp. 523–528, 2020, doi: 10.1016/j.procs.2020.02.053.

[44] M. R. Alam, A. Akter, M. A. Shafin, M. M. Hasan, and A. Mahmud, "SOCIAL MEDIA CONTENT CATEGORIZATION USING SUPERVISED BASED MACHINE LEARNING METHODS AND NATURAL LANGUAGE PROCESSING IN BANGLA LANGUAGE," in *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, IEEE, 2020, pp. 270–273.

[45] T. Kanan, A. Mughaid, R. Al-Shalabi, M. Al-Ayyoub, M. Elbes, and O. Sadaqa, "BUSINESS INTELLIGENCE USING DEEP LEARNING TECHNIQUES FOR SOCIAL MEDIA CONTENTS," *Cluster Comput.*, vol. 26, no. 2, pp. 1285–1296, 2023.

[46] N. Yan, X. Xu, T. Tong, and L. Huang, "EXAMINING CONSUMER COMPLAINTS FROM AN ON-DEMAND SERVICE PLATFORM," *Int. J. Prod. Econ.*, vol. 237, p. 108153, Jul. 2021, doi: 10.1016/j.ijpe.2021.108153.

[47] R. Pereira and C. Tam, "IMPACT OF ENJOYMENT ON THE USAGE CONTINUANCE INTENTION OF VIDEO-ON-DEMAND SERVICES," *Inf. Manag.*, vol. 58, no. 7, p. 103501, Nov. 2021, doi: 10.1016/j.im.2021.103501.

[48] E. O. Omuya, G. O. Okeyo, and M. W. Kimwele, "FEATURE SELECTION FOR CLASSIFICATION USING PRINCIPAL COMPONENT ANALYSIS AND INFORMATION GAIN," *Expert Syst. Appl.*, vol. 174, p. 114765, 2021.

[49] D. A. Pisner and D. M. Schnyer, "SUPPORT VECTOR MACHINE," in *Machine learning*, Elsevier, 2020, pp. 101–121.

[50] W. Xie, Y. She, and Q. Guo, "RESEARCH ON MULTIPLE CLASSIFICATION BASED ON IMPROVED SVM ALGORITHM FOR BALANCED BINARY DECISION TREE," *Sci. Program.*, vol. 2021, no. 1, p. 5560465, 2021.

[51] S. Dong, "MULTI CLASS SVM ALGORITHM WITH ACTIVE LEARNING FOR NETWORK TRAFFIC CLASSIFICATION," *Expert Syst. Appl.*, vol. 176, p. 114885, Aug. 2021, doi: 10.1016/j.eswa.2021.114885.

[52] K.-X. Han, W. Chien, C.-C. Chiu, and Y.-T. Cheng, "APPLICATION OF SUPPORT VECTOR MACHINE (SVM) IN THE SENTIMENT ANALYSIS OF TWITTER DATASET," *Appl. Sci.*, vol. 10, no. 3, p. 1125, 2020.

[53] V. D. P. Jasti *et al.*, "RELEVANT-BASED FEATURE RANKING (RBFR) METHOD FOR TEXT CLASSIFICATION BASED ON MACHINE LEARNING ALGORITHM," *J. Nanomater.*, vol. 2022, no. 1, p. 9238968, 2022.

[54] F. Firmansyah *et al.*, "COMPARING SENTIMENT ANALYSIS OF INDONESIAN PRESIDENTIAL ELECTION 2019 WITH SUPPORT VECTOR MACHINE AND K-NEAREST NEIGHBOR ALGORITHM," in *2020 6th International Conference on Computing Engineering and Design (ICCED)*, IEEE, 2020, pp. 1–6.

[55] S. Al Sulaimani and A. Starkey, "SHORT TEXT CLASSIFICATION USING CONTEXTUAL ANALYSIS," *IEEE Access*, vol. 9, pp. 149619–149629, 2021, doi: 10.1109/ACCESS.2021.3125768.

[56] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A COMPARATIVE ANALYSIS OF LOGISTIC REGRESSION, RANDOM FOREST AND KNN MODELS FOR THE TEXT CLASSIFICATION," *Augment. Hum. Res.*, vol. 5, no. 1, p. 12, 2020.

[57] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, "A NEW K-NEAREST NEIGHBORS CLASSIFIER FOR BIG DATA BASED ON EFFICIENT DATA PRUNING," *Mathematics*, vol. 8, no. 2, p. 286, 2020.

[58] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "MULTI-LABEL CLASSIFIER PERFORMANCE EVALUATION WITH CONFUSION MATRIX," in *Computer Science & Information Technology*, AIRCC Publishing Corporation, Jun. 2020, pp. 01–14. doi: 10.5121/csit.2020.100801.

[59] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: MULTI-LABEL CONFUSION MATRIX," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[60] D. D. Nada, S. Soehardjoepri, and R. M. Atok, "PERBANDINGAN ANALISIS SENTIMEN MENGENAI BPJS PADA MEDIA SOSIAL TWITTER MENGGUNAKAN NAÏVE BAYES CLASSIFIER (NBC) DAN SUPPORT VECTOR MACHINE (SVM)," *J. Sains dan Seni ITS*, vol. 11, no. 6, pp. D480–D485, May 2023, doi: 10.12962/j23373520.v11i6.96330.

[61] F. S. Nahm, "RECEIVER OPERATING CHARACTERISTIC CURVE: OVERVIEW AND PRACTICAL USE FOR CLINICIANS," *Korean J. Anesthesiol.*, vol. 75, no. 1, pp. 25–36, Feb. 2022, doi: 10.4097/kja.21209.

[62] D. Chicco and G. Jurman, "THE MATTHEWS CORRELATION COEFFICIENT (MCC) SHOULD REPLACE THE ROC AUC AS THE STANDARD METRIC FOR ASSESSING BINARY CLASSIFICATION," *BioData Min.*, vol. 16, no. 1, p. 4, Feb. 2023, doi: 10.1186/s13040-023-00322-4.

[63] D. Antons, E. Grünwald, P. Cichy, and T. O. Salge, "THE APPLICATION OF TEXT MINING METHODS IN INNOVATION RESEARCH: CURRENT STATE, EVOLUTION PATTERNS, AND DEVELOPMENT PRIORITIES," *R&D Manag.*, vol. 50, no. 3, pp. 329–351, Jun. 2020, doi: 10.1111/radm.12408.

[64] P. Föll and F. Thiesse, "EXPLORING INFORMATION SYSTEMS CURRICULA," *Bus. Inf. Syst. Eng.*, vol. 63, no. 6, pp. 711–732, Dec. 2021, doi: 10.1007/s12599-021-00702-2.

[65] S. M. Mohammad, "SENTIMENT ANALYSIS," in *Emotion Measurement*, Elsevier, 2021, pp. 323–379. doi: 10.1016/B978-0-12-821124-3.00011-9.

[66] M. Birjali, M. Kasri, and A. Beni-Hssane, "A COMPREHENSIVE SURVEY ON SENTIMENT ANALYSIS: APPROACHES, CHALLENGES AND TRENDS," *Knowledge-Based Syst.*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[67] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A SURVEY ON SENTIMENT ANALYSIS METHODS, APPLICATIONS, AND CHALLENGES," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[68] M. Alzate, M. Arce-Urriza, and J. Cebollada, "MINING THE TEXT OF ONLINE CONSUMER REVIEWS TO ANALYZE BRAND IMAGE AND BRAND POSITIONING," *J. Retail. Consum. Serv.*, vol. 67, p. 102989, Jul. 2022, doi: 10.1016/j.jretconser.2022.102989.