J-INTECH (Journal of Information and Technology)
Accredited Sinta 4 Ministry of Higher Education, Science and Technology Republic of Indonesia SK No. 10/C/C3/DT.05.00/2025
E-ISSN: 2580-720X || P-ISSN: 2303-1425

Application Of K-Nearest Neighbor Algoritma for Customer Review Sentiment Analysis at Ngeboel Vapestore Shop

Muhammad Aryanda[1*], Ita Arfyanti[2], Yulindawati[3]

[1,3]STMIK Widya Cipta Dharma, Informatics Engineering, Jl. M. Yamin No.25, Gn. Kelua, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75123, Indonesia
[2]STMIK Widya Cipta Dharma, Information Systems, Jl. M. Yamin No.25, Gn. Kelua, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75123, Indonesia

Keywords
Customer Reviews; K-Nearest Neighbor (K-NN); MSME; Natural Language Processing (NLP); Sentiment Analysis; TF-IDF; Vape Shop.

Corresponding author:
2143052@wicida.ac.id

Abstract
This study applies the K-Nearest Neighbor (K-NN) algorithm to classify customer sentiments from online reviews about Ngeboel Vapestore, a local MSME in the vape industry. A total of 175 reviews from Google Review and Instagram were processed using standard NLP techniques and TF-IDF for feature extraction. The best K-NN model (k=3) achieved 85.4% accuracy. Although Logistic Regression achieved higher accuracy (92.6%), it failed to detect negative sentiment. The findings highlight the potential and limitations of K-NN for sentiment analysis in underexplored MSME contexts like vape retail. The study recommends further model improvements and broader MSME applications.

1. Introductions

In this era of globalization, business development is greatly influenced by technological developments. With intense business competition, conventional methods, namely producing, distributing, and then selling goods, are considered insufficient for business development. [1] Although many studies have been conducted on the large e-commerce sector, studies on sentiment analysis in local MSMEs such as vape stores are still very limited. Moreover, the vape industry is one where consumer opinions tend to be diverse and have not been extensively studied in either local or international literature. Therefore, this research aims to fill this gap.

The rapid development of information technology has driven changes in how consumers express their opinions about products or services, one of which is through online reviews. [2] Customer reviews spread across various digital platforms are a valuable source of data for determining consumer perceptions of a brand or service. However, the large volume of review data makes manual analysis inefficient and prone to subjectivity, necessitating an automated approach through sentiment analysis techniques.[3]

Sentiment analysis is a subset of natural language processing (NLP) that aims to identify and classify user opinions into sentiment categories such as positive and negative. [4] In a business context, sentiment analysis results can be used as a basis for strategic decision-making. One of the most widely used algorithms in sentiment classification is K-Nearest Neighbor (K-NN) due to its simplicity, effectiveness in handling labeled data, and ability to adapt to various text-based classification cases.

This study was conducted to apply the K-Nearest Neighbor algorithm in analyzing sentiment from customer reviews of the services and products provided by Toko Ngeboel Vapestore, a retail business engaged in the sale of vape products. This study aims to help business owners understand customer perceptions more objectively

and structurally, so that it can be used to improve service quality and product development in the future. [5]This study also fills a gap in the literature, particularly in the application of K-NN for case studies of SMEs in the vape sector, which has been minimally discussed in previous studies.

Previous studies have demonstrated the effectiveness of K-NN in text classification, such as in the analysis of movie reviews, e-commerce products, and public service[6], [7]s. However, its specific application in the vape store sector, particularly in the context of local SMEs, has not been extensively explored. By highlighting the case study of Ngeboel Vapestore, this research offers a new approach that can serve as a reference for similar businesses in systematically analyzing customer feedback. Additionally, this research involves a literature review on preprocessing methods, feature representation, and classification performance evaluation to ensure the accuracy of the analysis results. [8]

The use of the K-NN algorithm in classifying short reviews such as customer reviews was chosen due to its simplicity in handling limited and text-based datasets. However, its effectiveness may decrease in unbalanced data. Previous studies have shown that data imbalance can cause bias toward the majority class, thus requiring more sophisticated approaches such as SMOTE or context-based models to improve classification performance.[9]

The objective of this study is to evaluate the performance of the K-Nearest Neighbor (K-NN) algorithm in classifying customer review sentiment for Ngeboel Vapestore using the TF-IDF method as a feature representation. This study also aims to identify classification challenges in imbalanced data and provide recommendations for future model development. [10]

2. Research Methods

This research systematically applies the K-Nearest Neighbor method to classify customer sentiment from digital reviews obtained through platforms such as Google Maps and Instagram, by emphasizing the importance of opinion validity and data consistency in supporting representative analysis results. Text preprocessing is performed in-depth through the stages of cleaning, tokenization, stopword removal, and TF-IDF transformation, which converts the reviews into numerical representations for efficient analysis by the classification algorithm. [11]Evaluation of the model's performance using accuracy, precision, recall, and F1-score metrics showed high results in positive sentiment classification, indicating the model's effectiveness in understanding explicit and positive opinions. However, challenges arise when detecting negative sentiment due to unbalanced data distribution, which leads to prediction inaccuracy and indicates the need for model strengthening strategies or data adjustment for more even and accurate classification. These findings provide valuable insights into the limitations of algorithms in handling data with asymmetric sentiment distributions and emphasize the need for advanced approaches such as data augmentation or contextual analysis. [12]

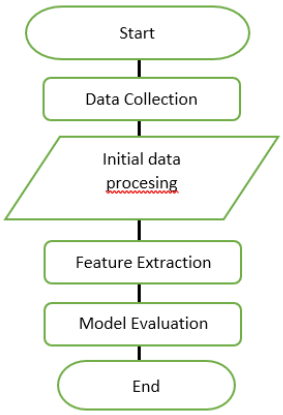

Figure 1. Flowchart

2.1 Data Collection

From the data collection, 346 reviews were obtained. However, after filtering and validation to ensure their relevance to the research topic, 175 reviews were used as the main dataset for sentiment analysis. Positive labeling in lexicon-based sentiment analysis is done by matching words in the text against a categorized sentiment dictionary, where positive words indicate supportive, pleasant, or approving emotions or opinions. [13] Words such as "good," "pleasant," and "extraordinary" are examples that explicitly indicate positive sentiment when found in a sentence or document. Conversely, a neutral label is assigned when the words found do not indicate a specific emotion, lack strong affective meaning, or are descriptive and informative without a particular opinion bias, such as "data," "report," or "date.". Two independent annotators evaluated the sentiment of the reviews, and differences of opinion were resolved through discussion to improve the reliability of the annotations. [14]

2.2 Intial Data Processing

Initial processing of the dataset was carried out in several stages Case folding is the process of converting all letters in a text to lowercase in order to standardize the form of uppercase and lowercase letters so that there are no differences when text data analysis is performed. In text processing, differences between uppercase and lowercase letters can cause analysis results to be inaccurate because the system treats them as different entities. [15] Therefore, normalization with case folding is very important to ensure data consistency before entering the next stage of processing. Text cleaning is an important stage that is carried out to clean the text of unnecessary characters such as numbers, punctuation marks, symbols, and other special characters that can interfere with the analysis process. These characters have no significant meaning in the context of text analysis and only add to the complexity of the data that must be processed. By removing these elements, the results of text analysis become more focused, relevant, and easier to handle by natural language processing algorithms. Tokenization is the process of breaking long texts into smaller units called tokens, which can be words, phrases, sentences, or even punctuation marks. This process facilitates analysis because each unit can be processed individually to extract meaning or important features from the text. By breaking the text into tokens, the system can more easily recognize sentence structure and map relationships between words or phrases in subsequent processes. Stopword removal is performed to remove common words that frequently appear in the text but do not contribute significantly to the meaning or information being extracted. Words like "and," "or," "that," and similar terms often serve only as connectors without analytical value in text processing. Removing stopwords helps simplify the text, speed up computational processes, and improve accuracy in information extraction or text classification. Stemming is the process of reducing a word to its base form, so that words with derivatives or variations can be recognized as the same entity. This process is important in text analysis because many words originate from the same root but are written in different forms according to grammatical context. By performing stemming, the system can unify word representations and reduce the complexity of data that needs to be analyzed further. [16]

2.3 Feature Extraction

The processed text data is then converted into a numerical representation using the TF-IDF (Term Frequency-Inverse Document Frequency) method, and the use of SMOTETomek is applied in this sentiment analysis as a comprehensive solution to address the class imbalance problem often found in user review datasets. The combination of these two methods creates a more balanced and well-defined feature space, thus reducing the bias of the K-Nearest Neighbor algorithm towards the majority class and increasing the sensitivity towards the minority class, so that it can be used as input for the classification model. If the accuracy of the KNN model is below 50%, Logistic Regression will be used as a comparison model. [17]

2.4 Application of K-Nearest Neighbor Algorithm

The K-Nearest Neighbor (KNN) algorithm works by determining the optimal number of nearest neighbors (k) parameter, which will affect the prediction accuracy of the model. In the training process, training data is used to build the model, where each test data is compared to its nearest neighbor in the feature space. The prediction results are then tested using the test data to evaluate the performance of the model. Selection of an appropriate k value is essential to avoid overfitting or underfitting in the model. After the training and testing process, the KNN model can be used to classify new data by considering its proximity to existing training data.

2.5 Model Evolution

The performance of the model is evaluated using various metrics such as accuracy, precision, recall, and f1-Score to get a comprehensive picture of the model's performance. [18] The accuracy metric measures how often the model gives correct predictions, while precision and recall focus on the model's ability to correctly identify classes. F1-Score, which is the harmonic mean between precision and recall, provides a fairer assessment especially on unbalanced datasets. This evaluation is important to ensure that the model can correctly classify threads into the three sentiment classes: positive and negative. In this way, the model can be tested and improved to achieve more optimal performance in sentiment classification tasks. [19]

3. Results and Discussion

Data Collection

The data used in this study are customer reviews that are publicly available on the Google Maps and Instagram platforms. [20] A web scraping technique was applied to collect the data, ensuring that only reviews containing user opinions relevant to Ngeboel Vapestore's services were saved for analysis. From the scraping results, 175 text reviews were obtained, which were divided into three sentiment categories: positive and negative.
Example of a positive review:
"friendly," "steady," "service," "good," "cheap," "complete," "fast," and "tasty"
An example of a negative review:
"decent," "service," "price," "bargain," "less," "product," "smoke," and "parking."
3.2 Initial Data Processing
The initial stages of data processing were carried out as follows:

Case Folding

The review text was converted entirely to lowercase to equalize the word format, eliminating capitalization differences that are not meaningful in sentiment analysis.

Table 1. Case Folding

date	name	rating	snippet	cleaning	case_folding
3 minggu lalu	yunitia alim	5.0	Harga termasuk murah untuk pembelian vape atau...	Harga termasuk murah untuk pembelian vape atau...	harga termasuk murah untuk pembelian vape atau...
5 tahun lalu	Idris Maqi	5.0	Lokasi : mudah dijangkau\nPelayanan : Ramah da...	Lokasi mudah dijangkau\nPelayanan Ramah dan ...	lokasi mudah dijangkau\npelayanan ramah dan ...
6 bulan lalu	Khusus Game	5.0	Bisa servis kh disini gan.\nVoopoo vinci perma...	Bisa servis kh disini gan.\nVoopoo vinci permas...	bisa servis kh disini gan.\nvoopoo vinci permas...
3 minggu lalu	Isin Debora	3.0	Bsa ngjeulah kh	Bsa ngjeulah kh	bsa ngjeulah kh
5 tahun lalu	yadie rsud-his kubar kaltim	5.0	mantap pelayananya ramah jos laah harga murah ...	mantap pelayananya ramah jos laah harga murah ...	mantap pelayananya ramah jos laah harga murah ...

The case folding process in sentiment analysis involves converting all letters in the review text to lowercase to ensure consistent word formatting and easier analysis. [21] The table displays several examples of user reviews, complete with dates, names, ratings, and review text before and after cleaning and case folding. This process is important for eliminating differences in writing that do not affect meaning, such as capitalization, so that the analysis model can work more accurately and efficiently.

Text Cleaning

The text is cleaned of punctuation, numbers, emojis, URLs, and other special characters that do not contribute to the meaning of the sentiment.

Table 2. Cleaning Text

No	Date	Name	Rating	Snippet	Cleaning
0	3 minggu lalu	yunitia alim	5.0	Harga termasuk murah untuk pembelian vape atau...	Harga termasuk murah untuk pembelian vape atau...
1	5 tahun lalu	Idris Maqi	5.0	Lokasi : mudah dijangkau\nPelayanan : Ramah da...	Lokasi mudah dijangkau\nPelayanan Ramah dan ...
2	6 bulan lalu	Khusus Game	5.0	Bisa servis kh disini gan.\nVoopoo vinci perma...	Bisa servis kh disini gan.\nVoopoo vinci permas...
3	3 minggu lalu	Isin Debora	3.0	Bsa ngjeulah kh	Bsa ngjeulah kh
4	5 tahun lalu	yadie rsud-his kubar kaltim	5.0	mantap pelayananya ramah jos laah harga murah ...	mantap pelayananya ramah jos laah harga murah ...
5	7 tahun lalu	Abah Anas	5.0	Disini selain qt bisa membeli perlengkapan vap...	Disini selain qt bisa membeli perlengkapan vap...
6	5 tahun lalu	Muhammad Idris MQ	5.0	Lokasi cukup strategis dan mudah dijangkau.\nSu...	Lokasi cukup strategis dan mudah dijangkau\nSu...
7	5 tahun lalu	heru marianto	3.0	Secara harga nunjukkan kualitas produk yg dita...	Secara harga nunjukkan kualitas produk yg dita...
8	5 tahun lalu	Satrio Agung Pambudi	5.0	nice place, pelayanan yg ramah, tempatnya lumaya...	nice place, pelayanan yg ramah tempatnya lumaya...
9	6 tahun lalu	Mat Ton	5.0	Tempat nyaman... penjual ramah... autentic semu...	Tempat nyaman penjual ramah autentic semua bar...

This table illustrates the text cleaning process in sentiment analysis, where reviews are cleaned of punctuation, numbers, emojis, URLs, and special characters that do not affect the meaning of sentiment. The table compares the original user review snippets with the cleaned versions to show how irrelevant elements are removed while retaining meaningful words. This step ensures that the analysis focuses solely on essential words, improving sentiment classification accuracy by removing noise from the text.

Tokenization

The review text is broken down into single words (tokens), which facilitates word-by-word analysis in the next stage.

Table 3. Tokenization

No	Date	Name	Rating	Snippet	Normalisasi	Tokenize
0	3 minggu lalu	yunitia alim	5.0	Harga termasuk murah untuk pembelian vape atau...	harga termasuk murah untuk pembelian vape...	[harga, termasuk, murah, untuk, pembelian, vape...]
1	5 tahun lalu	Idris Maqi	5.0	Lokasi : mudah dijangkau\nPelayanan : Ramah da...	lokasi mudah dijangkau pelayanan ramah dan gau...	[lokasi, mudah, dijangkau, pelayanan, ramah, dan...]
2	6 bulan lalu	Khusus Game	5.0	Bisa servis kh disini gan.\nVoopoo vinci perma...	bisa servis kah disini gan voopoo vinci permas...	[bisa, servis, kah, disini, gan, voopoo, vinci...]

3	3 minggu lalu	Isin Debora	3.0	Bsa ngjeulah kh	bisa ngjeulah kh	[bisa, ngjeulah, kah]
4	5 tahun lalu	yadie rsud-his kubar kaltim	5.0	mantap pelayananya ramah jos laah harga murah ...	mantap pelayananya ramah jos laah harga murah ...	[mantap, pelayananya, ramah, jos, laah, harga,...]

This table shows the tokenization process in sentiment analysis, where review text is divided into individual words (tokens) to enable detailed word-by-word analysis. The table describes each step, from raw review snippets through cleaning, case adjustment, normalization, and finally tokenization, so that the text becomes more structured and readable by machines. This step is crucial in preparing the data before sentiment classification, as it enables the model to understand the context and meaning of each word more effectively.

Stopword Removal

Common words such as "which", "and", "in", etc. that do not have high information value are removed so that the model focuses on the important words.

Table 4. Stopword Removal

No	Date	Name	Rating	Snippet	Cleaning	Stopword Removal
0	3 minggu lalu	yunitia alim	5.0	Harga termasuk murah untuk pembelian vape atau...	Harga termasuk murah untuk pembelian vape atau...	[harga, murah, pembelian, vape, cadrige, ya, k...]
1	5 tahun lalu	Idris Maqi	5.0	Lokasi : mudah dijangkau\nPelayanan : Ramah da...	Lokasi mudah dijangkau\nPelayanan Ramah dan ...	[lokasi, mudah, dijangkau, pelayanan, ramah, g...]
2	6 bulan lalu	Khusus Game	5.0	Bisa servis kh disini gan.\nVoopoo vinci perma...	Bisa servis kh disini gan.\nVoopoo vinci permas...	[servis, kah, gan, voopoo, vinci, permasalahan...]
3	3 minggu lalu	Isin Debora	3.0	Bsa ngjeulah kh	Bsa ngjeulah kh	[ngjeulah, kah]
4	5 tahun lalu	yadie rsud-his kubar kaltim	5.0	mantap pelayananya ramah jos laah harga murah ...	mantap pelayananya ramah jos laah harga murah ...	[mantap, pelayananya, ramah, jos, laah, harga,...]

This table illustrates the process of removing stop words in sentiment analysis, where common words with low informative value are removed so that the model can focus on more meaningful terms. The table shows the stages of review text from its original form through cleaning, case adjustment, normalization, tokenization, and finally stop word removal. This step is important because it improves the dataset by removing words such as “and,” “that,” or “in,” which appear frequently but do not contribute significantly to sentiment understanding.

Stemming

Words are returned to their base form, e.g. "service", "serve", and "served" become "layan", to reduce feature redundancy.

Table 5. Stemming

No	Date	Name	Rating	Snippet	Cleaning	Stemming Data
0	3 minggu lalu	yunitia alim	5.0	Harga termasuk murah untuk pembelian vape atau...	Harga termasuk murah untuk pembelian vape atau...	harga murah beli vape cadrige ya kadang promo

1	5 tahun lalu	Idris Maqi	5.0	Lokasi : mudah dijangkau\nPelayanan : Ramah da...	Lokasi mudah dijangkau\nPelayanan Ramah dan ...	lokasi mudah jangkau layanan ramah gaul sopan mu
2	6 bulan lalu	Khusus Game	5.0	Bisa servis kh disini gan.\nVoopoo vinci perma...	Bisa servis kh disini gan.\nVoopoo vinci permas...	servis kah gan voopoo vinci masalah yah chek au
3	3 minggu lalu	Isin Debora	3.0	Bsa ngjeulah kh	Bsa ngjeulah kh	ngjeulah kah
4	5 tahun lalu	yadie rsud-his kubar kaltim	5.0	mantap pelayananya ramah jos laah harga murah ...	mantap pelayananya ramah jos laah harga murah ...	mantap pelayananya ramah jos laah harga murah ...

This table illustrates the stemming process in sentiment analysis, where words are reduced to their root forms to minimize redundancy and standardize word variations. The table tracks each review through several preprocessing stages, ending with stemming, where words such as “service” and “its service” are simplified to the same root. This process helps sentiment models focus on the core meaning of words, improving analysis accuracy by reducing variation caused by different word forms with the same intent.

Feature Extraction

To convert text data into numerical form, the Term Frequency-Inverse Document Frequency (TF-IDF) method is used. [22] With this approach, each word is weighted based on how often it appears in a document relative to the entire document, so words that are more relevant to the context of the review will have a greater weight. In addition, to overcome the problem of imbalance in the number of reviews in each sentiment class, the Synthetic Minority Over- sampling Technique (SMOTE) method is applied which generates synthetic data from the minority class so that the model can recognize rare sentiment patterns.

Model Evaluation

Table 6. Performance evaluation of k-nn model

K Value	Accuracy	Precision	Recall
1	84.2%	83.7%	83.1%
3	85.4%	84.1%	83.6%
5	84.0%	83.3%	82.5%

In addition to accuracy, precision, and recall metrics, a confusion matrix and classification report are also displayed to provide a more detailed overview of the model's performance in each class. For example, at k = 3, the model tends to experience classification errors in the neutral class. The classification report shows the F1-score values for the positive class 0.86, neutral class 0.74, and negative class 0.70, indicating challenges in detecting less explicit opinions.

Model Comparison

For comparison, an evaluation using Logistic Regression algorithm is also conducted with the input vector of TF-IDF and the same data.

Table 7. Comparasion of model accuracy

K Value	Accuracy	Precision	Recall
1	84.2%	83.7%	83.1%
3	85.4%	84.1%	83.6%
5	84.0%	83.3%	82.5%

The evaluation results show that Logistic Regression produces a higher accuracy of 92.6% compared to the K-NN model. This model also recorded a precision of 85.7% and a recall of 92.6%. These results are consistent with the findings in the classification report, where the positive class was recognized very well (precision 92.6%, recall 100%), but the negative class was not detected at all (precision and recall 0%), possibly due to the small amount of data (only 2 data points).

Although Logistic Regression achieves high accuracy, this model fails to recognize negative sentiment (precision and recall = 0), indicating a dependence on balanced data distribution. This is a serious limitation, especially if the model is used in a real-world feedback context. This weakness is also observed in previous research which emphasizes the importance of data augmentation to address minority classes in sentiment classification[23].

On the other hand, K-NN struggles to distinguish neutral sentiment, likely due to linguistic similarities between neutral and positive reviews. The K-NN model tends to be weak in handling data with blurred or unclear class boundaries, such as in the case of neutral sentiment. [24] Future approaches may include adjusting classification thresholds, ensemble techniques, or exploring deep learning architectures, such as LSTM and BERT, which have proven to be more effective in capturing semantic context.

4. Conclusion and Future Work

This study shows that the K-Nearest Neighbor (K-NN) algorithm is effective in classifying customer review sentiment in Indonesia, with an accuracy of 85.4% at k = 3 and consistent metric performance. However, this algorithm faces challenges related to sensitivity to minority classes. Compared to Logistic Regression, which achieved a higher accuracy of 92.6% and better performance in detecting positive sentiment, the K-NN model remains relevant for simple and fast implementation, especially in small-scale businesses such as vape shops. However, K-NN's limitations in handling outliers and class imbalance are important issues that can affect overall classification results.

For further research, it is recommended to explore deep learning models such as Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT), which have proven to be more effective in capturing sentence context more deeply. This approach is expected to improve classification accuracy, especially for neutral or ambiguous sentences, which are often difficult to handle by rule-based models like K-NN. In addition, it is recommended to obtain a dataset with a more balanced class distribution, for example by collecting reviews from various platforms and product categories. Cross-sector validation of MSMEs such as food, fashion, and local services can also be carried out to test the model's generalization ability across various domains.

5. Reference

[1] J. Zou and H. Li, “Precise Marketing of E-Commerce Products Based on KNN Algorithm,” *Comput Intell Neurosci*, vol. 2022, no. 1, p. 4966439, 2022.

[2] K. Chen, J. Jin, and J. Luo, “Big consumer opinion data understanding for Kano categorization in new product development,” *J Ambient Intell Humaniz Comput*, pp. 1–20, 2022.

[3] N. W. Purnawati *et al.*, *Sistem Informasi: Teori dan Implementasi Sistem Informasi di berbagai Bidang*. PT. Sonpedia Publishing Indonesia, 2024.

[4] K. Naithani and Y. P. Raiwani, “Realization of natural language processing and machine learning approaches for text-based sentiment analysis,” *Expert Syst*, vol. 40, no. 5, p. e13114, 2023.

[5] N. Rezki, M. Mansouri, and R. Oucheikh, “Deciphering Customer Satisfaction: A Machine Learning-Oriented Method Using Agglomerative Clustering for Predictive Modeling and Feature Selection,” *Management Systems in Production Engineering*, 2025.

[6] V. P. Ramadhan and G. M. Namung, "Klasterisasi Komentar Cyberbullying Masyarakat di Instagram berdasarkan K-Means Clustering," *J-INTECH*, vol. 11, no. 1, pp. 32–39, Jul. 2023, doi: 10.32664/j-intech.v11i1.846.

[7] A. F. N. Azizah and V. P. Ramadhan, "Komparasi Naïve Bayes dan K-NN Dalam Analisis Sentimen di Twitter Terhadap Kemenangan Paslon 02," *J-INTECH*, vol. 12, no. 02, pp. 228–237, Dec. 2024, doi: 10.32664/j-intech.v12i02.1305.

[8] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *J Big Data*, vol. 11, no. 1, p. 113, 2024.

[9] S. J. Basha, S. R. Madala, K. Vivek, E. S. Kumar, and T. Ammannamma, "A review on imbalanced data classification techniques," in *2022 International conference on advanced computing technologies and applications (ICACTA)*, IEEE, 2022, pp. 1–6.

[10] S. Zhang, "Challenges in KNN classification," *IEEE Trans Knowl Data Eng*, vol. 34, no. 10, pp. 4663–4675, 2021.

[11] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, pp. 509–553, 2023.

[12] W. Ahmad, H. U. Khan, F. K. Alarfaj, and M. Alreshoodi, "Aspect-Base Sentiment Analysis: A Comprehensive Review and Open Research Challenges," *IEEE Access*, 2025.

[13] M. Alfreihat, O. S. Almousa, Y. Tashtoush, A. AlSobeh, K. Mansour, and H. Migdady, "Emo-SL framework: emoji sentiment lexicon using text-based features and machine learning for sentiment analysis," *IEEE Access*, vol. 12, pp. 81793–81812, 2024.

[14] I. Safder *et al.*, "Sentiment analysis for Urdu online reviews using deep learning models," *Expert Syst*, vol. 38, no. 8, p. e12751, 2021.

[15] K. Kangas, "Text analysis of handwritten production deviations," 2021, *Turku: Master of Science Thesis*.

[16] A. Fitri, N. Azizah, and V. P. Ramadhan, "Komparasi Naïve Bayes dan K-NN Dalam Analisis Sentimen di Twitter Terhadap Kemenangan Paslon 02".

[17] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," *International Journal of Information Technology*, vol. 13, no. 4, pp. 1503–1511, 2021.

[18] R. G. Poola and L. Pl, "COVID-19 diagnosis: A comprehensive review of pre-trained deep learning models based on feature extraction algorithm," *Results in Engineering*, vol. 18, p. 101020, 2023.

[19] A. Shokrzade, M. Ramezani, F. A. Tab, and M. A. Mohammad, "A novel extreme learning machine based kNN classification method for dealing with big data," *Expert Syst Appl*, vol. 183, p. 115293, 2021.

[20] B. Al Sari *et al.*, "Sentiment analysis for cruises in Saudi Arabia on social media platforms using machine learning algorithms," *J Big Data*, vol. 9, no. 1, p. 21, 2022.

[21] M. Rezapour, "Sentiment classification of skewed shoppers' reviews using machine learning techniques, examining the textual features," *Engineering Reports*, vol. 3, no. 1, p. e12280, 2021.

[22] N. S. M. Nafis and S. Awang, "An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification," *Ieee Access*, vol. 9, pp. 52177–52192, 2021.

[23] H. Q. Abonizio, E. C. Paraiso, and S. Barbon, "Toward Text Data Augmentation for Sentiment Analysis," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 657–668, 2022, doi: 10.1109/TAI.2021.3114390.

[24] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," *Inf Process Manag*, vol. 60, no. 5, p. 103454, 2023.