Recursive Journal of Informatics
Vol. 2, No. 2, September 2024
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

Analysis of the Use of Nazief-Adriani Stemming and Porter Stemming in COVID-19 Twitter Sentiment Analysis with Term Frequency-Inverse Document Frequency Weighting Based on k-Nearest Neighbor Algorithm

Muhammad Fikri[1], Zaenal Abidin[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract.
This system was developed to determine the accuracy of sentiment analysis on Twitter regarding the COVID-19 issue using the Nazief-Adriani and Porter stemmers with TF-IDF weighting, along with a classification process using K-Nearest Neighbor (KNN) that resulted in a comparison of 48.24% for Nazief-Adriani and 48.24% for Porter.
Purpose: This research aims to determine the accuracy of the Nazief-Adriani and Porter stemmer algorithms in performing text preprocessing using a dataset from Indonesian-language Twitter. This research involves word weighting using TF-IDF and classification using the K-Nearest Neighbor (KNN) algorithm.
Methods/Study design/approach: The experimentation was conducted by applying the Nazief-Adriani and Porter stemmer algorithm methods, utilizing data sourced from Twitter related to COVID-19. Subsequently, the data underwent text preprocessing, stemming, TF-IDF weighting, accuracy testing of training and testing data using K-Nearest Neighbor (KNN) algorithm, and the accuracy of both stemmers was calculated employing a confusion matrix table.
Result/Findings: This study obtained reasonably accurate results in testing the Nazief-Adriani stemmer with an accuracy of 50.98%, applied to sentiment analysis of COVID-19-related Twitter data using the Indonesian language. As for the accuracy of the Porter stemmer, it achieved an accuracy rate of 48.24%.
Novelty/Originality/Value: Feature selection is crucial in stemmer accuracy testing. Therefore, in this study, feature selection is carried out using the Nazief-Adriani and Porter stemmers for testing purposes, and the accuracy data classification is conducted using the K-Nearest Neighbor (KNN) algorithm.

Keywords: Text mining, stemmer, Nazief-Adriani, Porter, KNN, Twitter
Received September 08, 2023 / **Revised** May 02, 2024 / **Accepted** September 17, 2024

INTRODUCTION

The COVID-19 pandemic has reshaped the global landscape, affecting various aspects of life and societies worldwide. In this context, social media, particularly Twitter, has become a platform where people express their views, feelings, and reactions to the evolving situation. As a result, sentiment analysis on this platform has become increasingly crucial in understanding collective sentiments and perspectives regarding this pandemic [1]. Therefore, this research aims to analyze COVID-19-related sentiments on Twitter using text processing techniques and advanced classification algorithms while identifying the impact of using the Nazief-Adriani and Porter stemmers in enhancing sentiment analysis accuracy. Thus, this study contributes to uncovering public perspectives through Twitter data related to this global pandemic [2].

This study involves three main stages in sentiment analysis. Firstly, COVID-19-related tweet data from Twitter is collected using web scraping techniques with the Snscrape library. Subsequently, the tweet texts undergo preprocessing stages, including case folding, link, symbol, and number removal. Following this, stemming is performed using the Nazief-Adriani and Porter stemming methods to reduce words to their base forms.
Next, TF-IDF weighting is applied to compute word weights within each tweet. These weights reflect the significance of words within the sentiment analysis context. The sentiment classification process is

[1]*Corresponding author.
Email addresses: muhfikri1998@students.unnes.ac.id (Fikri)
DOI: 10.15294/rji.v2i2.74267

conducted using the K-Nearest Neighbor (KNN) algorithm. The preprocessed and weighted tweet data is partitioned into training and testing datasets. The KNN model is then trained using the training data and subsequently tested using the testing data to classify the sentiment of each individual tweet [3].

Sentiment Analysis is a natural language processing methodology with the aim of extracting, recognizing, and categorizing sentiments or emotions present within the text. Its primary objective is to comprehend individuals' or groups' perspectives, emotions, and responses toward a particular subject, such as products, services, events, or specific issues.

Text Mining, also referred to as text data mining, is a computer science approach that concentrates on extracting valuable information from digital text or documents. This process involves recognizing patterns, unearthing knowledge, and analyzing data applied to textual content to acquire novel insights or useful information [4].

Text Preprocessing serves as the preliminary phase in text processing, designed to purify, modify, and organize raw text to prepare it for subsequent analysis. This phase encompasses actions like case folding (converting text to lowercase), eliminating hyperlinks, symbols, and numerical figures, as well as removing irrelevant words [5].

Stemming is a technique employed in text processing to reduce words within text to their fundamental or root form. This is achieved by removing prefixes or suffixes, thereby consolidating words with morphological variations into a unified basic form. This process intends to address word disparities and heighten resemblances among interconnected words in text analysis.

Classification involves the categorization of data or objects into pre-established groups or classes. Within the sphere of sentiment or text analysis, classification entails the categorization of specific texts into predetermined sentiment categories—such as positive, negative, or neutral—based on identifiable features or attributes found within the text.

A study was conducted by comparing the performance of Nazief-Adriani and Porter stemmers and calculating the precision and processing time of each stemmer by testing 30 sample documents. The stemmer evaluation was carried out on 30 Indonesian language text documents with varying document sizes. The findings indicated that stemming Indonesian language text documents using the Porter stemmer required less time than the Nazief-Adriani stemmer. However, stemming Indonesian language text documents using the Porter stemmer exhibited a lower accuracy percentage than the Nazief-Adriani stemmer.

Based on the explanation above, this research focuses on the application of the Nazief-Adriani and Porter stemmer algorithms in sentiment analysis of COVID-19-related tweets using TF-IDF weighting and evaluating their accuracy levels through K-Nearest Neighbor (KNN) classification. The aim is to determine which algorithm is more accurate in the stemming process for sentiment analysis of Twitter data.

METHODS

The testing is conducted by retrieving data from Twitter tweets, which are then subjected to preprocessing and stemming using the Nazief-Adriani and Porter stemmers for accuracy assessment. The following outlines the research stages to be undertaken. The following is the flowchart depicting the stages of the method as shown in Figure 1.

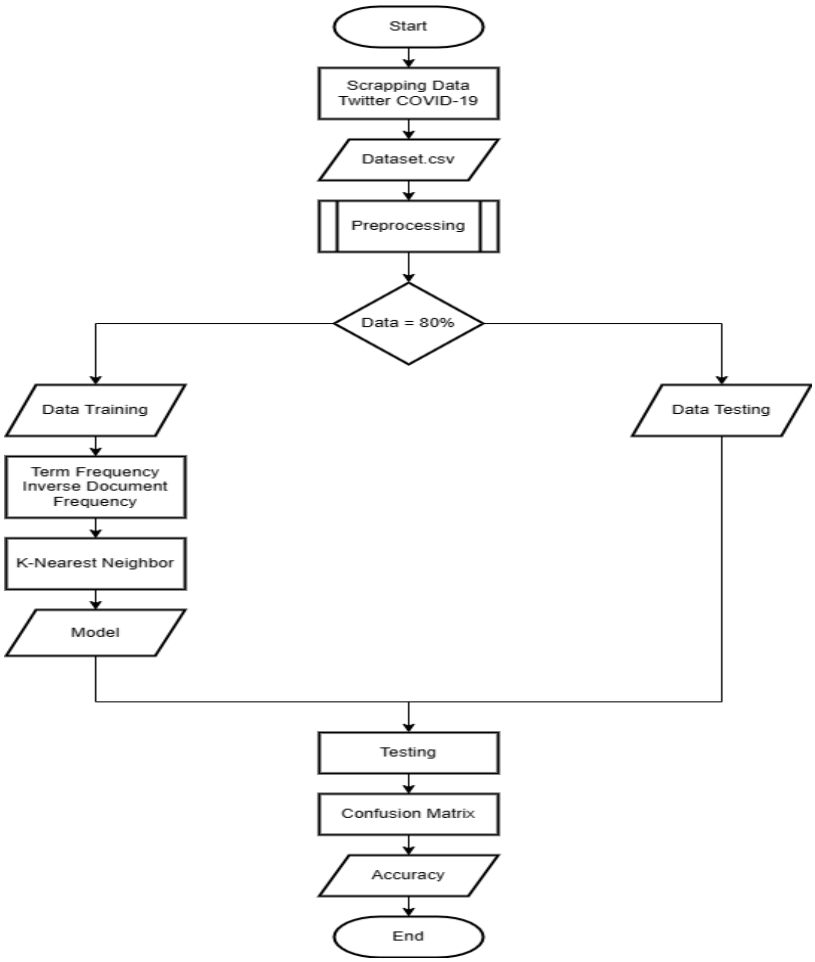

Figure 1. Research stages

Dataset Collection

The method employed in the initial stage involves extracting data from Twitter using data scraping techniques, utilizing the “snscrape” library in Python. The process of collecting Twitter data using web scraping involves extracting relevant information from Twitter's platform. This is typically achieved through libraries or tools designed for web scraping, such as "snscrape" The technique involves programmatically accessing Twitter's content and extracting tweets based on specific criteria, such as keywords, hashtags, or user profiles. The collected data is then stored in a structured format, such as CSV files, for further analysis. This method enables researchers to gather a substantial volume of tweets for various purposes, including sentiment analysis and trend identification.

Text Preprocessing

The initial stage of text mining is the text preprocessing process, which aims to transform a piece of text into structured data that can be further processed in subsequent stages. In the text preprocessing stage, it is expected to be effectively utilized for classification before proceeding to the COVID-19 Twitter data classification process. The data must go through several stages, including case folding, cleansing, stemming, stopword removal, and tokenization [6]. Common preprocessing steps include:

Case Folding

Converting all characters to lowercase to ensure uniformity.
Removing URLs: Eliminating URLs or hyperlinks from the text.
Removing Symbols and Numbers: Getting rid of special characters and numeric values that don't contribute to sentiment analysis.
Tokenization: Splitting the text into individual words or tokens [7].

Stemming
After preprocessing, stemming is applied to reduce words to their root forms. Stemming helps in dealing with different word variations and enhances the analysis by considering words with the same root as one. The Nazief-Adriani and Porter stemmers are algorithms specifically designed for this purpose. They remove prefixes and suffixes from words to obtain their base forms.

These preprocessing and stemming steps ensure that the text data is transformed into a standardized and manageable format, making it ready for further analysis, such as sentiment classification using machine learning algorithms like K-Nearest Neighbor (KNN).

Stemmer Nazief-Adriani
The Nazief-Adriani algorithm was first developed by Bobby Achirul Awal Nazief and Mirna Adriani. This algorithm adheres to the rules of Indonesian language morphology, which are compiled into a group and encapsulated into permissible and non-permissible affixes. It relies on a basic word dictionary and supports recoding, which involves reassembling words that have undergone excessive stemming [8]. The Nazief-Adriani Stemmer is an algorithm used for stemming in the Indonesian language. Stemming is the process of reducing a word to its root form to eliminate morphological variants of a word. The Nazief-Adriani algorithm is known as one of the effective stemmers for the Indonesian language, with a high level of accuracy and the ability to preserve the original meaning of words in sentences. Another advantage of this algorithm is its capability to perform real-time stemming and can be used in text-based applications such as text mining and NLP. However, the Nazief-Adriani algorithm also has a limitation, which is its inability to handle words that do not have affixes or suffixes [9].

Stemmer Porter
The Porter algorithm was first created and published by Martin Porter in 1980. The Porter stemmer is a context-sensitive algorithm for removing suffixes. Some definitions used in the Porter stemmer algorithm include consonants, which are letters other than vowels that are preceded by a consonant. There are five steps in this algorithm, and in each step, rules are applied until one of the rules satisfies the conditions. Once the conditions are met, the suffix is removed and processed in the next stage. The stem produced in the fifth step becomes the result. The advantage of using the Porter algorithm is its fast processing speed in stemming, but its drawback is the low level of accuracy [10]. This algorithm is very popular and widely used in the fields of text mining and informatics due to its good performance and effectiveness in processing text. In recent years, this algorithm has also been applied to other languages, although with some modifications to address unique issues in each language. The Porter stemmer is crucial because it can help reduce dimensions in text data and facilitate text analysis. In 1992, W.B. Frakes adapted this algorithm for stemming the Indonesian language. In the Porter algorithm, W.B. Frakes has made modifications that can be used according to the Indonesian language [11].

Term Frequency-Inverse Document Frequency
TF-IDF Weighting: In this stage, the Term Frequency-Inverse Document Frequency (TF-IDF) technique is applied to assign weights to words within each tweet, reflecting their significance in the context of sentiment analysis [12].

Term Frequency (TF) represents the frequency of a word within a specific tweet. It is calculated by dividing the number of times a word appears in a tweet by the total number of words in that tweet. TF measures the relevance of a word within a single tweet.

Inverse Document Frequency (IDF) is a measure of how unique a word is across the entire dataset. It is calculated as the logarithm of the total number of documents divided by the number of documents containing the word. IDF helps to downweight common words and emphasize more meaningful and informative words.

The product of TF and IDF, known as TF-IDF, results in a weight assigned to each word in a tweet. This weight reflects the importance of the word within the entire dataset while considering its frequency within the specific tweet. Words that are frequent in a tweet but rare across the dataset receive higher TF-IDF weights, indicating their significance in representing the tweet's sentiment the formula can be seen in equation 1.

$$\text{TF} = \begin{cases} 1 + \log 10\ (f\ t,d), f\ t,d > 0 \\ 0, \quad f\ t,d = 0 \end{cases} \quad (1)$$

Here is an explanation of the formula.

f : Frequency.
t : Term.
d : Document

K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) Classification: In this step, the preprocessed and weighted data are divided into two subsets: a training set and a testing set. The training set is used to train a KNN classifier, which is a supervised machine learning algorithm. The KNN algorithm learns from the patterns in the training data and stores them as reference points in a multi-dimensional space, where each point represents a document (tweet) [13].

Once the KNN classifier is trained, it uses the labeled data in the training set to determine the class labels of new, unseen data points. For each data point in the testing set, the KNN algorithm calculates its distance to the reference points in the training set. The algorithm then identifies the k-nearest neighbors, where k is a user-defined parameter representing the number of neighbors to consider. The majority class among these k neighbors is assigned as the predicted class for the new data point the formula can be seen in equation 2.

$$\cos(\theta_{QD}) = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n} (Q_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (D_i)^2}} \quad (2)$$

$\text{Cos}\ (\theta_{QD})$: Similarity $(Q)$to document (D).
Q : Data test.
D : Data training.
n : Amount of data

The performance of the KNN classifier is evaluated using the testing set. The predicted class labels are compared with the true class labels in the testing set to calculate accuracy, which measures the classifier's ability to correctly predict sentiment labels. The accuracy score provides insights into how well the KNN algorithm performs in classifying sentiments for COVID-19-related tweets based on the processed and weighted features from the TF-IDF technique [14].

Confusion Matrix

The use of a confusion matrix is to visually represent the performance of a classification model. It provides a comprehensive understanding of how well the model's predictions align with the actual labels in a multiclass classification problem. The matrix consists of cells where each row represents the instances in an actual class, and each column represents the instances in a predicted class [15]. The diagonal cells of the matrix show the correct predictions, while off-diagonal cells indicate misclassifications. The values in the cells help in assessing various evaluation metrics such as accuracy, precision, recall, and F1-score. By analyzing the confusion matrix, one can gain insights into the strengths and weaknesses of the classification model's performance across different classes[16].

RESULT AND DISCUSSION

The experimental results showcase a notable level of accuracy in sentiment classification, highlighting the effectiveness of incorporating the Nazief-Adriani and Porter stemmers along with the TF-IDF weighting scheme and the K-Nearest Neighbor (KNN) algorithm. The obtained accuracy rates of 50.98% and 48.24% for the Nazief-Adriani and Porter stemmers provide a meaningful assessment of their performance on COVID-19-related tweets. These accuracy rates were determined by evaluating the model on both the data training and testing sets. The outcomes emphasize the intricacies inherent in sentiment analysis when dealing with tweets concerning the pandemic.

While these results indicate a certain level of accuracy, there remains potential for further enhancement. The research insights underscore the necessity for continued exploration and innovative methodologies to

address the nuanced challenges embedded within sentiment analysis during the prevailing pandemic context. This study sets the stage for advancing sentiment analysis techniques and contributing to the broader comprehension of public sentiment on social media platforms, particularly regarding global health crises such as the COVID-19 pandemic. The accuracy results for both data training and testing for each stemmer can be observed in the following Table 1 and 2.

Table 1. Accuracy Results of Training Data and Testing Data using Nazief-Adriani Stemmer with K-Nearest Neighbor (KNN).

No	Sample Data	Training Data	Testing Data
1	20	0.56299	0.50980
2	30	0.57817	0.51308

Table 2. Accuracy Results of Training Data and Testing Data using Porter Stemmer with K-Nearest Neighbor (KNN).

No	Sample Data	Training Data	Testing Data
1	20	0.56003	0.48235
2	30	0.57592	0.51047

The accuracy calculations follow the same pattern in both above data tests. The accuracy of the model trained on the test data is calculated by comparing the number of instances predicted correctly to the total number of instances in the test data. The accuracy percentage indicates the model's performance in accurately classifying sentiments based on data that has undergone Nazief-Adriani and Porter stemming processes with KNN classification.

In both data tests, accuracy is calculated in decimal form (from 0.00 to 1.00), demonstrating the model's success rate in classifying sentiments in the testing and training data using the Nazief-Adriani and Porter stemmers with the KNN algorithm. The diagram representation can be viewed in the Figure 2 and Table 3 for stemming accuracy results

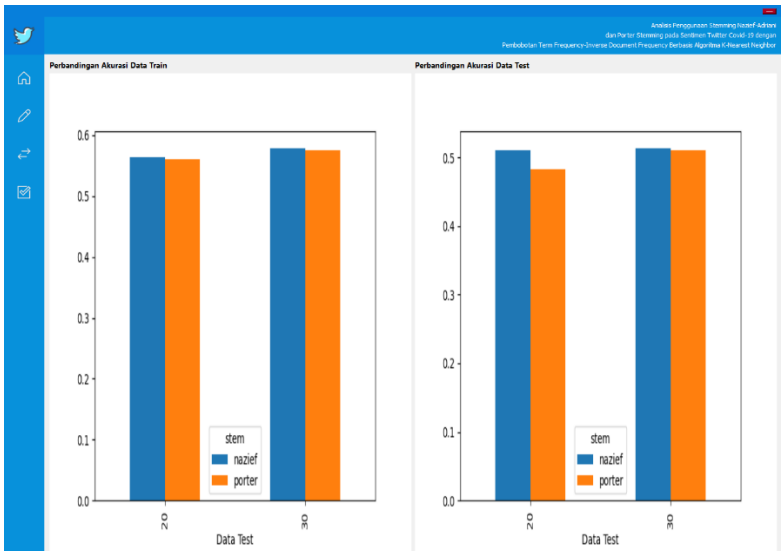

Figure 2. The accuracy diagram of training and testing data using the Nazief-Adriani and Porter stemmers.

The use of Nazief-Adriani stemming, and Porter Stemming has an impact on accuracy results. Nazief-Adriani stemming achieves an accuracy of 50.98%, while Porter Stemming achieves an accuracy of 48.24%. The difference in accuracy can be attributed to the characteristics of each stemmer and the match of words in the text. Although the accuracy is not optimal, these results provide insights into the challenges of analyzing sentiment on social media platforms, which often contain informal words and words that are not present in dictionaries. The accuracy results can be seen in the Table 3.

Table 3. The accuracy stemmer Nazief-Adriani and stemmer Porter.

	Accuracy Results	
	Nazief-Adriani + TF-IDF + KNN	Porter + TF-IDF + KNN
Accuracy	50,98%	48,24%

CONCLUSION

This study sheds light on the intricate dynamics of sentiment analysis when applied to the context of COVID-19-related tweets using the Nazief-Adriani and Porter stemming techniques. The exploration of these two stemming methods, combined with TF-IDF weighting and the K-Nearest Neighbor (KNN) algorithm, has yielded distinctive accuracy outcomes. The Nazief-Adriani stemming method achieved an accuracy rate of 50.98%, showcasing its adeptness in capturing nuanced sentiments expressed in Indonesian text. On the other hand, the Porter stemming method achieved an accuracy rate of 48.24%, indicating its potential to contribute to sentiment analysis tasks in similar contexts.

These accuracy rates underscore the role of stemmers in enhancing the analysis of sentiments on social media platforms like Twitter during the COVID-19 pandemic. The Nazief-Adriani stemmer, with its higher accuracy rate, proves to be particularly effective in deciphering and categorizing a diverse range of sentiments, thereby providing a more comprehensive understanding of public reactions and emotions.

However, it is important to note that while the attained accuracy rates offer valuable insights, they also indicate areas for further refinement. The complexities of informal language, slang, and unique expressions within the Twitter sphere can still pose challenges for sentiment analysis. These findings emphasize the ongoing need for continuous development and enhancement of text processing techniques in sentiment analysis tasks, especially in the realm of evolving public discourse, such as the pandemic.

Future research directions may explore the fusion of advanced natural language processing techniques with the identified stemmers to potentially achieve higher accuracy levels. Moreover, incorporating domain adaptation and considering temporal dynamics could contribute to an even more accurate reflection of sentiment changes and public attitudes over time.

REFERENCES

[1] Y. Affandi and E. Sugiharti, "Sentiment Analysis of student on Online Lectured During Covid-19 Pandemic Using K-Means and Naïve Bayes Classifier," *Journal of Advances in Information Systems and Technology*, vol. 5, no. 1, pp. 38–49, 2023.

[2] F. F. Rachman and S. Pramana, "Analisis sentimen pro dan kontra masyarakat Indonesia tentang vaksin COVID-19 pada media sosial Twitter," *Indonesian of Health Information Management Journal (INOHIM)*, vol. 8, no. 2, pp. 100–109, 2020.

[3] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *INSYST: Journal of Intelligent System and Computation*, vol. 1, no. 1, pp. 43–49, 2019.

[4] N. Anggraini, E. S. N. Harahap, and T. B. Kurniawan, "Text Mining-Analisis Teks Terkait Isu Vaksinasi COVID-19 (Text Mining-Text Analysis Related to COVID-19 Vaccination Issues)," *JURNAL IPTEKKOM (Jurnal Ilmu Pengetahuan & Teknologi Informasi)*, vol. 23, no. 2, pp. 141–153, 2021.

[5] A. T. J. Harjanta, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *Jurnal Informatika Upgris*, vol. 1, no. 1 Juni, 2015.

[6] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[7] H. M. Keerthi Kumar and B. S. Harish, "Classification of short text using various preprocessing techniques: An empirical evaluation," in *Recent Findings in Intelligent Computing Techniques: Proceedings of the 5th ICACNI 2017, Volume 3*, Springer, 2018, pp. 19–30.

[8] A. C. Herlingga, I. G. L. P. E. Prismana, D. R. Prehanto, and D. A. Dermawan, "Algoritma Stemming Nazief & Adriani Dengan Metode Cosine Similarity Untuk Chatbot Telegram Terintegrasi Dengan E-layanan," *Journal of Informatics and Computer Science (JINACS)*, vol. 2, no. 1, 2020.

[9] R. Rosnelly, "The Similarity of Essay Examination Results using Preprocessing Text Mining with Cosine Similarity and Nazief-Adriani Algorithms," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 3, pp. 1415–1422, 2021.

[10] B. V. Indriyono, E. Utami, and A. Sunyoto, "Pemanfaatan Algoritma Porter Stemmer Untuk Bahasa Indonesia Dalam Proses Klasifikasi Jenis Buku," *Jurnal Buana Informatika*, vol. 6, no. 4, 2015.

[11] Y. Darnita, "Pengaruh Algoritma Stemming Porter Terhadap Kinerja Algoritma Rabin Karp Untuk Mendeteksi Plagiarisme Teks Bahasa Indonesia," *JTIS, Volume 3 Nomor 2, Juli 2020* , vol. 3, 2020.

[12] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries."

[13] Y. M. Elgammal, M. A. Zahran, and M. M. Abdelsalam, "A new strategy for the early detection of alzheimer disease stages using multifractal geometry analysis based on K-Nearest Neighbor algorithm," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-26958-6.

[14] F. Arsyadani and A. Purwinarko, "Implementation of Synthetic Minority Oversampling Technique and Two-phase Mutation Grey Wolf Optimization on Early Diagnosis of Diabetes using K-Nearest Neighbors," *Recursive Journal of Informatics*, vol. 1, no. 1, pp. 9–17, 2023.

[15] M. A. Rohman and D. Arifianto, "Penerapan Metode Euclidean Probality dan Confusion Matrix dalam Diagnosa Penyakit Koi," *Jurnal Smart Teknologi*, vol. 2, no. 2, pp. 122–130, 2021.

[16] H. G. Lewis and M. Brown, "A generalized confusion matrix for assessing area estimates from remotely sensed data," 2001. [Online]. Available: http://www.tandf.co.uk/journals