LONTAR KOMPUTER VOL. 15, NO. 1 APRIL 2024 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2024.v15.i01.p05 **e-ISSN 2541-5832**
Accredited Sinta 2 by RISTEKDIKTI Decree No. 158/E/KPT/2021

Quickly Assess the Acceptability Sentiment of White Paracetamol Intake Using KNN-SMOTE Based On Receptive Deciding

Rio Andika Malik[a1], Faizal Riza[b2], Sarjon Defit[b3]

[a]Digital of Business Department, University of Perintis Indonesia
Lubuk Buaya Padang, Indonesia
[1]rioandikamalik@upertis.ac.id (Corresponding author)

[b]Information Technology doctoral program, University of Putra Indonesia YPTK
Padang, Indonesia
[2]s3.faizal@gmail.com
[3]sarjon_defit@upiyptk.ac.id

Abstract

This research aims to develop a fast and adaptive sentiment evaluation approach related to the use of white paracetamol using a combination of the K-Nearest Neighbors (KNN) algorithm, Synthetic Minority Over-Sampling Technique (SMOTE), and the Receptive Deciding concept. Imbalances in the dataset, where positive sentiment may predominate, are addressed using SMOTE to synthesize minority class samples. The KNN algorithm is applied to build a sentiment classification model, while Receptive Deciding is used to provide adaptive intelligence to changes in sentiment. The SMOTE oversampling process is carried out to achieve class balance, while KNN is used to classify sentiment. Receptive Deciding is applied to increase the model's adaptability to changes in sentiment. The research results show that integrating the SMOTE, KNN, and Receptive Deciding methods effectively assesses sentiment accurately and adaptively. The developed model can recognize changes in sentiment over time and provide balanced evaluation results. These findings are expected to contribute to understanding public sentiment towards using white paracetamol and be the basis for developing more effective health communication strategies.

Keywords: *Machine Learning, K-NN, SMOTE, Acceptability Sentiment, Receptive Deciding.*

1. Introduction

As an ever-evolving field of study, computer science has made a major contribution to technological progress and significant transformation in various aspects of human life. One area that has received widespread attention is the development of machine learning techniques and sentiment analysis. Machine learning is becoming increasingly relevant in understanding complex patterns in data, while sentiment analysis enables the interpretation and measurement of human sentiment contained in text.

In the context of computer science, current research often explores the application of classification algorithms such as K-Nearest Neighbors (KNN) and oversampling techniques such as Synthetic Minority Over-Sampling Technique (SMOTE) to improve the performance of machine learning models. Utilizing these techniques is crucial in dealing with the problem of class imbalance in data, often encountered in sentiment analysis [1]. On the other hand, the concept of Receptive Deciding brings a deeper understanding of how decisions can be responsive to new information, providing the adaptive intelligence needed in complex and dynamic data processing. In this context, research strives to combine these concepts to increase the speed and effectiveness of sentiment analysis.

Assessing drug-related sentiment can provide valuable insight into user satisfaction levels, possible side effects, and general perception of a drug's efficacy. Paracetamol, or acetaminophen, is an analgesic and antipyretic drug commonly used to relieve pain and reduce fever [2][3]. As

one of the drugs frequently consumed by the public, assessing its acceptability and sentiment regarding its use is an essential aspect of developing and monitoring public health. In the case of paracetamol, sentiment assessment can influence consumption behavior and provide valuable information for health policy planning [3]–[5].

Previous research has shown that machine learning techniques, such as the K-Nearest Neighbors (KNN) classification algorithm and Synthetic Minority Over-Sampling Technique (SMOTE), can be used to analyze sentiment in text data [6]–[9]. However, in the context of paracetamol use, there has not been much research specifically addressing the application of these techniques to evaluate acceptance sentiment. Receptive Deciding, a concept involving decisions responsive to input or new information, is the main focus of this research. Integrating KNN-SMOTE with Receptive Deciding is expected to provide a fast and effective approach to evaluating acceptance sentiment regarding the use of white paracetamol. When applied to the context of drug use, especially white paracetamol, the need to understand the sentiment of public acceptance of its use becomes increasingly important. A quick evaluation of these sentiments can provide valuable information to support health policy and provide greater insight into user preferences.

This research will contribute to our understanding of public sentiment toward using white paracetamol and may provide a basis for developing more effective health communication strategies. Additionally, the proposed method's application can offer broader insight into using machine learning techniques in drug-related sentiment evaluation.

2. Research Methods

To provide robust and acute-sensing ML-based implementation, a meaningful portion of features must directly reflect the size and diversity of the collected observations. The research method used in this study was designed to carefully investigate public sentiment regarding the use of white paracetamol utilizing an approach that combines the K-Nearest Neighbors (KNN) algorithm, Synthetic Minority Over-Sampling Technique (SMOTE), and the concept of Receptive Deciding.

First, this research faces the challenge of imbalance in the dataset, where the number of samples between positive and negative sentiments may be unbalanced [8]. To overcome this, oversampling methods, especially SMOTE, are used to synthesize minority class samples so that the dataset becomes more balanced [6], [9]. The main goal of this step is to ensure that the developed model can recognize and assess sentiment in a balanced manner without being affected by class imbalance [9].

Next, the K-Nearest Neighbors (KNN) algorithm is applied to build a sentiment classification model. KNN was chosen because of its ability to handle classification problems on complex data and its ability to adapt to different data patterns [10], [11]. In this context, KNN is used to understand and classify sentiment based on the numerical representation of user reviews. Receptive Deciding is the main focus of this research. This concept is implemented to make the model more responsive to changes in sentiment that may occur over time. This provides adaptive intelligence to the model, allowing it to make more dynamic decisions based on new information received.

The research process began with collecting a dataset that included user reviews of white paracetamol and continued with the preprocessing stage to clean and prepare the data for analysis. After that, model building, using SMOTE, and applying Receptive Deciding are holistically integrated to achieve fast, effective, and adaptive sentiment evaluation. Using this method, it is hoped that this research can contribute to understanding public sentiment regarding the use of white paracetamol and provide a basis for developing more effective health communication strategies. The research stages carried out are presented in Figure 1 [12].

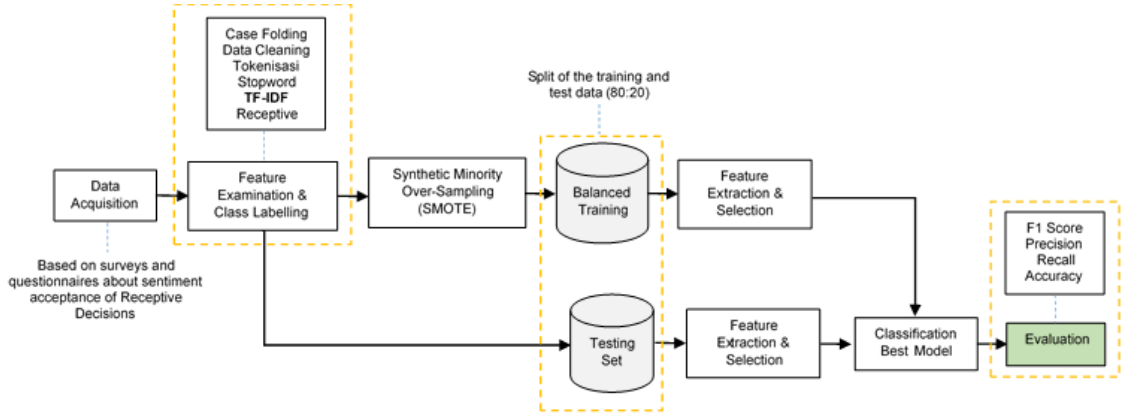


Figure 1. The Planned Methodology's Productivity [12]

The dataset used in this study includes information related to the use of white paracetamol and related sentiments. It was collected from user reviews through drug-related public health surveys. The main focus of this dataset is to understand how society responds to and accepts the use of white paracetamol as an analgesic and antipyretic drug.

This dataset contains several attributes, including review text containing users' views and experiences with white paracetamol. Other attributes include demographic information, such as age and gender, and additional information regarding medication use. This dataset contains various views and user experiences regarding white paracetamol, from safety and efficacy to possible side effects. This diversity allows models to train and identify complex sentiment patterns. This dataset has undergone a preprocessing process, including text cleaning, tokenization, and feature extraction, to ensure the data is ready for development models. Thus, this dataset provides a solid basis for sentiment evaluation research regarding white paracetamol using the proposed machine learning method on receptive deciding.

Receptive Deciding is a text processing technique that examines the sentiment or emotions expressed in the text. The Receptive Stage and Deciding Stage are the two primary phases of this approach. At this point, the text or sentences are processed using the TF-IDF technique to discover keywords or features that most help determine the sentiment or feeling present in the text. Receptive Deciding uses vector representations of words created at the reception step to classify sentiment using machine learning methods like K-Nearest Neighbours (KNN). This method compares recently processed text with previously analyzed material during the model training phase. By using the Receptive Deciding method, the model can understand and classify the sentiment or feelings contained in the text more accurately.

2.1. K-Nearest Neighbors (KNN)

For a variety of regression and classification scenarios, a well-liked machine learning method is the K-Nearest Neighbours (KNN) algorithm [13]–[15]. It is predicated on the notion that comparable data points typically have similar labels or values [16]. The KNN method employs the complete training dataset as a reference throughout the training phase [17], [18]. It uses a selected distance metric, such as Euclidean distance, to determine the distance between each training example and the input data point before making predictions [13], [19].

KNN has the drawbacks of being simple to develop and capable of handling complicated data patterns [20]. Over time, KNN becomes increasingly sensitive to emotional shifts when integrated with Receptive Deciding. This idea enables the model to adjust to societal preferences or viewpoints changes, leading to a more dynamic evaluation of sentiment [21]. The first step in using KNN is to choose the k value, which is the number of nearest neighbors used to determine the class or predicted value. The k value can be selected through cross-validation or other methods. After the ***k*** value is determined, KNN calculates the distance between the data to be classified or predicted and all the data in the dataset [3], [10]. This distance can be calculated using a metric using Euclidean distance via the following equation (1) [13], [15], [22], [23].

$$\boldsymbol{Euclidean\ Distance}\ (x_i, x_j) = \sqrt{\sum_{r=1}^{n}(x_{ir} - x_{ij})^2} \quad (1)$$

Where x_{ir} is the testing data and x_{ij} is the training data. Data that has the smallest distance to the data to be classified or predicted is selected as the nearest neighbor. The number of neighbors chosen is by the predetermined k value, and then the most common or majority class among the selected neighbors will be attributed as the class of the data to be classified

2.2. Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE is an oversampling technique used to overcome imbalances in a dataset, especially when there is a significant difference in the number of samples between the majority class and the minority class [8], [9], [24]. This imbalance can cause the model to produce biased or less accurate predictions towards minority classes [6], [7]. In this case, SMOTE works by synthesizing minority datasets by adding synthetic samples so that the number of datasets in the minority class becomes balanced with the majority class [25]. Class connectivity is found in the dataset, where positive sentiment may be more dominant than negative sentiment or vice versa. Therefore, the use of oversampling methods, especially the Synthetic Minority Over-Sampling Technique (SMOTE), is considered essential to achieve the necessary balance in sentiment analysis [10], [26].

SMOTE starts by calculating the distance between data on minority data, determines the SMOTE percentage value, determines the number of k closest, and finally creates synthetic data, as seen in equation (2) [12].

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \quad (2)$$

With x_{syn} is the synthetic data that will be created x_i data to be replicated, x_{knn} data that has the shortest distance of the data to be replicated and δ random value between 0 and 1.

2.3. Evaluation

Classification evaluation is based on testing correct objects and incorrect objects to determine the best model type from the classification results. This validation is vital in measuring how much the model can recognize and classify data accurately. The Confusion Matrix is an evaluation tool commonly used in this research. It provides detailed information regarding the actual classification results that the classification system can predict.

The confusion matrix divides classification results into four main categories:

a. True Positive (TP): The model correctly predicted the number of objects that belong to the positive class.
b. True Negative (TN): The model correctly predicted the number of objects in the negative class.
c. False Positive (FP): The number of objects that belong to the negative class but were incorrectly predicted as positive by the model.
d. False Negative (FN): The number of objects that belong to the positive class but were incorrectly predicted as unfavorable by the model.

By using the information from the confusion matrix, several evaluation metrics can be calculated, such as [13] [27]:

a. Accuracy: Shows the extent to which the model can correctly classify all objects via equations

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

b. Precision: Shows the extent to which objects predicted as positive by the model belong to the positive class via the equation

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \quad (4)$$

c. Recall (Sensitivity or True Positive Rate): Shows the extent to which the model can identify all objects that belong to the positive class through the equation

$$Recall = \frac{TP}{TP+FP} \quad (5)$$

d. F1-Score: Combines precision and recall into one metric, useful when there is a trade-off between the two via an equation

$$F1 - Score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad (6)$$

Using the confusion matrix and evaluation metrics allows this research to understand the model's performance in classifying sentiment related to the use of white paracetamol. This helps researchers and practitioners evaluate the model's strengths and weaknesses, understand how well the model recognizes positive and negative sentiment, and provide a basis for future model improvements. Good evaluations also support more effective health policies and more accurate drug decision-making.

3. Result and Discussion

The initial step in this research was collecting a dataset consisting of user reviews regarding white paracetamol. The dataset used in this research contains data on the usage of white paracetamol and associated feelings. This dataset was gathered from user reviews via public health questionnaires about drugs (we conducted random case studies in several city pharmacies in West Sumatra Padang). This dataset's primary goal is to comprehend how society views and tolerates the usage of white paracetamol as an antipyretic and painkiller.

Based on user ratings, the class labeling process was implemented based on opinions about the acceptability of white paracetamol intake. It is possible to categorize the views expressed in these reviews as good, neutral, or negative. This procedure entails selecting a random dataset sample and evaluating the disclosed sentiments based on the researcher's subjective opinion. The correctness and comprehensive representation of the sentiment conveyed in the dataset were considered during the research procedure despite the researcher's subjectivity being required for the class labeling process. It is anticipated that the outcomes of this class labeling procedure would accurately depict user sentiment regarding the social approval of white people's paracetamol use. It is hoped that criticism or inquiries concerning the absence of explanation regarding these two features can be appropriately addressed by outlining the dataset's source and the class labeling processes.

A preprocessing process is then carried out to clean and prepare the data so that it is ready for analysis. This process involves steps such as removal of irrelevant data, normalization of text, and grouping of words. The text will be processed before the data is used to make the data more accurate. Data preprocessing includes deleting unused columns (attributes) and filling in null data (cleaning), changing capital letters in columns to lowercase (case folding), breaking sentences into words (tokenizing), and deleting abbreviated words that have unimportant meanings. (Stopword) as shown in Figure 2.

1 to 25 of 462 entries Filter

index	Mention	Sentimen
0	Menyenangkan karena warna obat paracetamol yang cerah.	Positif
1	Kurang suka dengan warna terlalu mencolok pada obat paracetamol.	Positif
2	Warna obat paracetamol memberikan kesan bersih dan aman.	Positif
3	Tidak terlalu mempedulikan warna obat asalkan efektif.	Netral
4	Warna obat paracetamol terlalu standar dan membosankan.	Negatif
5	Warna obat yang menenangkan membuat saya merasa lebih baik.	Positif
6	warna obat paracetamol tampak usang.	Negatif
7	Merasa positif karena obat paracetamol yang mudah dikenali.	Positif
8	Warna obat paracetamol memberikan kesan ramah anak.	Positif
9	Tidak suka dengan warna obat yang terlalu terang	Negatif
10	Merasa nyaman dengan warna putih pada obat paracetamol.	Positif
11	warna obat paracetamol memberikan kepercayaan diri karena terkesan bersih.	Positif
12	Warna obat paracetamol yang terang membuat suasana hati lebih baik.	Positif
13	Tidak suka dengan warna yang terlalu mencolok pada obat.	Negatif
14	Menyukai kesederhanaan warna pada obat paracetamol.	Positif
15	Warna obat paracetamol terlalu terang untuk konsumsi malam.	Negatif
16	Menganggap warna obat paracetamol tidak relevan dengan khasiatnya.	Netral
17	Desain warna obat memberikan kesan profesional.	Positif
18	Warna obat paracetamol yang lembut memberikan rasa kenyamanan.	Positif
19	warna obat terlihat kurang menarik.	Negatif
20	Warna obat paracetamol yang cerah membuat saya senang.	Negatif
21	Kurang suka dengan warna-warna yang terlalu banyak pada obat.	Negatif
22	Warna obat paracetamol yang terlalu terang membuat mata sakit.	Negatif
23	Terkesan dengan warna obat yang simpel dan elegan.	Positif
24	Warna obat paracetamol yang putih membuat saya waspada.	Negatif

Show 25 per page 1 2 10 19

Figure 2. Preprocessing Text Results

After the text preprocessing is complete, Figure 3 shows that the following bar-chart visualization of sentiment labels from the cleaned data will be calculated to determine how public the public reaction regarding the use of white paracetamol is.

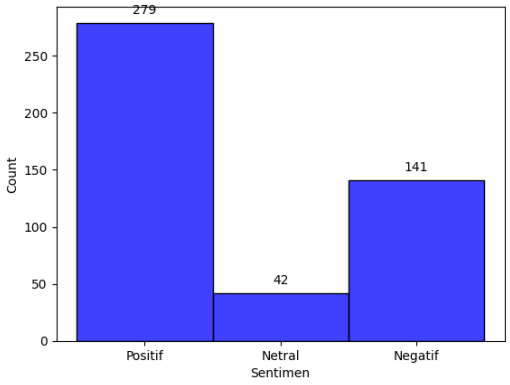


Figure 3. Proportion of Data

It can be seen in the picture above that there are 279 sentiments with positive labels, 42 sentiments with neutral labels, and 141 negative labels. In total, after preprocessing and labeling, there were 462 data. Figure 3 above also shows that most people react positively to white paracetamol. However, special attention must be paid to the data imbalance in these sentiment distributions [1]. Positive sentiment has a much higher frequency compared to neutral and negative sentiment. This condition creates an imbalance that can affect model performance, especially when dealing with cases where negative sentiment needs to be identified.

It is important to note that an imbalance in the dataset may cause the model to favor the majority of the class and, therefore, may be less sensitive to cases in the minority. Therefore, applying Synthetic Minority Over-Sampling Technique (SMOTE) is a critical step in this research. Before the SMOTE step is implemented, the dataset undergoes an additional processing stage by applying the Term Frequency-Inverse Document Frequency (TF-IDF) method [19], [28], [29]. This process is carried out to improve the representation of text in the dataset by giving weight to words based on their frequency [21], [30], [31]. With this step, the sentiment analysis process not only considers class imbalance but also ensures that the features extracted from the dataset have been appropriately processed to improve the accuracy and interpretability of the model, as shown in Figure 4.

	adalah	agar	air	aman	ampuh	anak	asal	asalkan	bahwa	baik	...	terlalu	terlihat	tidak	universal	untuk	usang	warna	warnanya	waspada	yang
0	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.0000	0.000000	0.0	0.0	0.0	0.116954	0.0	0.0	0.160715
1	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.301847	0.0000	0.000000	0.0	0.0	0.0	0.136947	0.0	0.0	0.000000
2	0.0	0.0	0.0	0.57443	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.0000	0.000000	0.0	0.0	0.0	0.128085	0.0	0.0	0.000000
3	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.592456	0.0	0.0	...	0.256456	0.0000	0.288999	0.0	0.0	0.0	0.116354	0.0	0.0	0.000000
4	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.247988	0.0000	0.000000	0.0	0.0	0.0	0.112512	0.0	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
457	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.0000	0.000000	0.0	0.0	0.0	0.136895	0.0	0.0	0.000000
458	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.0000	0.000000	0.0	0.0	0.0	0.145023	0.0	0.0	0.000000
459	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.3569	0.000000	0.0	0.0	0.0	0.118256	0.0	0.0	0.000000
460	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.0000	0.000000	0.0	0.0	0.0	0.130509	0.0	0.0	0.000000
461	0.0	0.0	0.0	0.00000	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.0000	0.000000	0.0	0.0	0.0	0.098270	0.0	0.0	0.000000

462 rows × 170 columns

Figure 4. TF-IDF Result before SMOTE

Following the TF-IDF procedure, the data is processed and ready for training and testing of the KNN model. Training and test data are the two categories into which the data is separated at this point. The model is trained using training data to pick up on patterns in the already existing data. Test data is used to evaluate the model's effectiveness and determine whether it can make

accurate predictions on data that has never been seen before. This segment guarantees the model can 'understand' generic data patterns rather than ' memorize' the training set. To this end, an 80:20 split of the training and test data has been experimented with. This way, the data is split into two sets: 80% for training the model and 20% for testing it. This section assists in making sure the model receives sufficient data to train itself and an adequate amount of data to evaluate its performance. Following splitting, the test data is predicted using the KNN model. When applying this model, test data is viewed, and the most prevalent label among the K data points closest to the test data points observed in the test is decided.

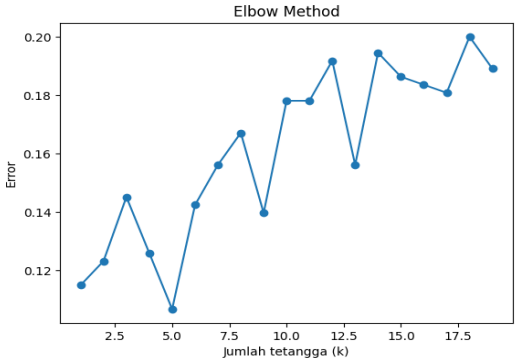
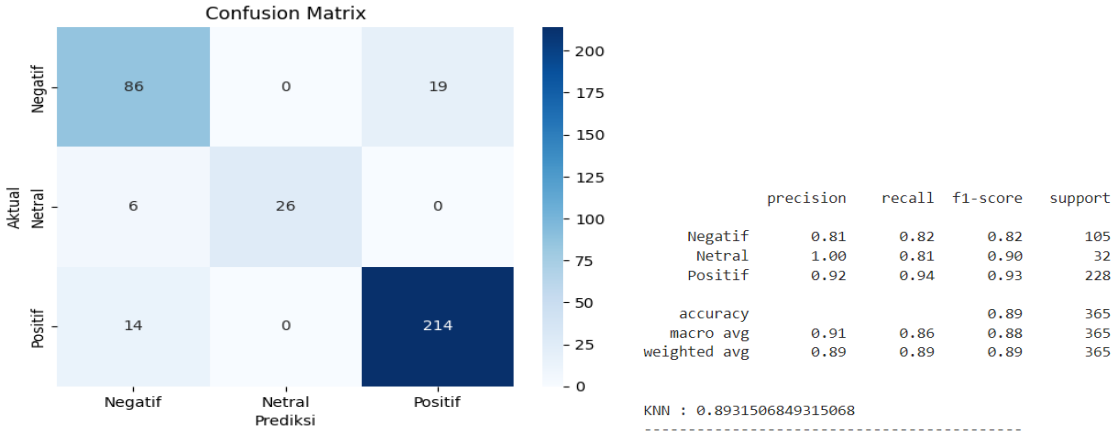


Figure 5. Elbow Method KNN

To find the hyperparameter K value that best fits the data, the K-Nearest Neighbours (KNN) algorithm employs the Elbow approach. This method allows one to comprehend how system mistakes or losses respond to altered or controlled variables. In Figure 5, ***k*** =5 is the optimal ***k*** value selected for the first experiment utilizing the KNN model, according to the Elbow technique. This shows that the model error rate drastically drops at this value of K and that increasing K further might not considerably improve the model's quality. For your KNN model, ***k*** =5 is the appropriate value according to the Elbow technique.



	precision	recall	f1-score	support
Negatif	0.81	0.82	0.82	105
Netral	1.00	0.81	0.90	32
Positif	0.92	0.94	0.93	228
accuracy			0.89	365
macro avg	0.91	0.86	0.88	365
weighted avg	0.89	0.89	0.89	365

KNN : 0.8931506849315068

Figure 6. Confusion Matrix KNN

The model performance assessment metrics are displayed in Figure 6, which is the outcome of the confusion matrix. With a "weighted average" value of 0.89 for precision, recall, and F1-score measures, the model's overall accuracy is 0.89. The average precision, recall, and F1-score obtained using the "macro average" metric are 0.91, 0.86, and 0.88, respectively. These findings suggest that ***k***=5 is the ideal K number for KNN for categorizing sentiment. As a result, the model can accurately classify sentiment, particularly for the "Positive" class. The main goal could be to employ SMOTE to increase the precision and recall of the "Negative" and "Neutral" classes.

After the processing stage with KNN, the next step in this research is to apply the Synthetic Minority Over-Sampling Technique (SMOTE) method to overcome the class imbalance in the sentiment dataset related to white paracetamol. As seen in the sentiment distribution before implementing SMOTE, there is a significant imbalance between positive, neutral, and negative sentiment. This can cause the model to tend to be insensitive to minority sentiments, especially negative sentiments that have lower representation.

The application of SMOTE is carried out by duplicating and creating synthetic samples in the minority class. In this context, the minority class is a neutral and negative sentiment. This process helps create a more balanced distribution of positive, neutral, and negative sentiment, ensuring that the model can more effectively classify sentiment from all classes, as shown in Figure 7.

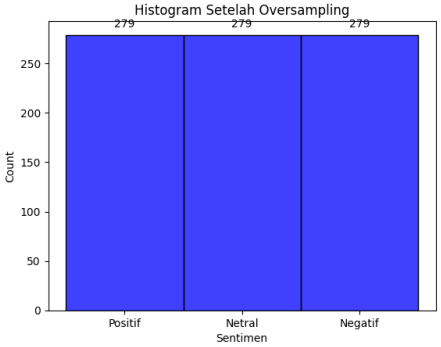


Figure 7. Histogram After Oversampling

After the SMOTE step is applied to overcome class imbalance, the sentiment analysis process involves an additional stage: the application of Term Frequency-Inverse Document Frequency (TF-IDF) on a dataset that has experienced oversampling. Next, the dataset involving SMOTE and TF-IDF is ready to be used in model building and sentiment evaluation to provide more accurate and comprehensive results, as shown in Figure 8.

	adalah	agar	air	aman	ampuh	anak	asal	asalkan	bahwa	baik	...	terlalu	terlihat	tidak	universal	untuk	usang	warna	warnanya	waspada	yang
0	0.0	0.0	0.0	0.00000	0.00000	0.0	0.000000	0.000000	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.116954	0.0	0.0	0.160715
1	0.0	0.0	0.0	0.00000	0.00000	0.0	0.000000	0.000000	0.0	0.0	...	0.301847	0.0	0.000000	0.0	0.0	0.0	0.136947	0.0	0.0	0.000000
2	0.0	0.0	0.0	0.57443	0.00000	0.0	0.000000	0.000000	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.128085	0.0	0.0	0.000000
3	0.0	0.0	0.0	0.00000	0.00000	0.0	0.000000	0.592456	0.0	0.0	...	0.256456	0.0	0.288999	0.0	0.0	0.0	0.116354	0.0	0.0	0.000000
4	0.0	0.0	0.0	0.00000	0.00000	0.0	0.000000	0.000000	0.0	0.0	...	0.247988	0.0	0.000000	0.0	0.0	0.0	0.112512	0.0	0.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
832	0.0	0.0	0.0	0.10264	0.36289	0.0	0.000000	0.000000	0.0	0.0	...	0.192244	0.0	0.216639	0.0	0.0	0.0	0.087221	0.0	0.0	0.119857
833	0.0	0.0	0.0	0.00000	0.00000	0.0	0.119878	0.000000	0.0	0.0	...	0.207341	0.0	0.233651	0.0	0.0	0.0	0.094070	0.0	0.0	0.000000
834	0.0	0.0	0.0	0.00000	0.00000	0.0	0.000000	0.000000	0.0	0.0	...	0.115980	0.0	0.241478	0.0	0.0	0.0	0.097221	0.0	0.0	0.000000
835	0.0	0.0	0.0	0.00000	0.00000	0.0	0.000000	0.592456	0.0	0.0	...	0.256456	0.0	0.288999	0.0	0.0	0.0	0.116354	0.0	0.0	0.000000
836	0.0	0.0	0.0	0.00000	0.00000	0.0	0.332007	0.000000	0.0	0.0	...	0.199006	0.0	0.224259	0.0	0.0	0.0	0.090288	0.0	0.0	0.000000

837 rows × 170 columns

Figure 8. TF-IDF Result after SMOTE

Table 1 shows the class proportions for the final data, divided into two scenarios: using the K-Nearest Neighbors (KNN) algorithm alone and using KNN with the SMOTE (Synthetic Minority Over-Sampling Technique) method. In the KNN scenario alone, 279 data are labeled "Positive," 141 data are labeled "Negative," and 42 data are labeled "Neutral." The total data processed in this scenario is 279 + 141 + 42 = 462.

In the KNN scenario with the SMOTE method, the amount of data remains the same for the "Positive" label, namely 279 data. However, the "Negative" and "Neutral" labels were resampled (regenerated) using the SMOTE method so that the number was the same as the "Positive" label, namely 279 data for each label. The total data processed in this scenario is 279+279+279 = 837.

Table 1. Proportion of Final Data

Label	KNN	KNN+SMOTE
Positive	279	279
Negative	141	279
Neutral	42	279

Using the SMOTE method, the proportion of classes in the final data becomes balanced and increases the amount of data the KNN model will process. This can improve the quality of the model, especially in classifying data from minority classes such as "Negative" and "Neutral." In the KNN scenario alone, there is synchronization in the amount of data for each label, where the "Positive" label has a much more significant amount of data compared to the "Negative" and "Neutral" labels.

After going through the TF-IDF stages and implementing SMOTE, the next step involves selecting the optimal k value for the KNN model. This process can be done through cross-validation or other evaluation methods to minimize overfitting and ensure good model generalization. After the k value is determined, the KNN model is trained using a dataset that involves SMOTE and TF-IDF. The model building uses features generated through TF-IDF extraction, including information from synthetic and original user reviews. One method commonly used to determine the optimal k value is the Elbow Method. The Elbow Method helps identify where increasing the k value does not significantly improve model performance. Plotting the k value against the cost or inertia value, we look for where the curve forms an "elbow." Increasing the k value does not provide significant benefits at this point, and the model is considered good enough, as shown in Figure 9.

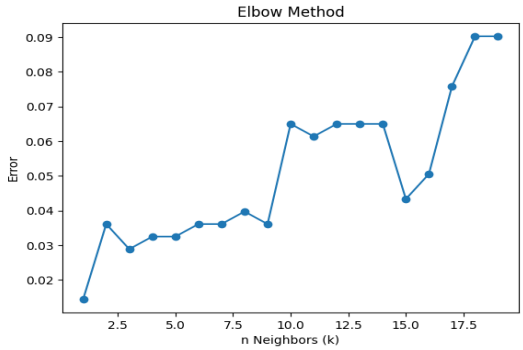


Figure 9. Elbow Method KNN+SMOTE

It was found that the "elbow" point on the curve occurs at the value ***k***=1. This means that increasing the value of k after ***k***=1 does not significantly improve model performance. Furthermore, with the value k=1, which has been determined as the optimal value, the KNN model is ready for sentiment analysis on SMOTE and TF-IDF datasets. Further evaluation involved using a Confusion Matrix to gain more detailed insight into the model's ability to classify positive, neutral, and negative sentiment, as shown in Figure 10.

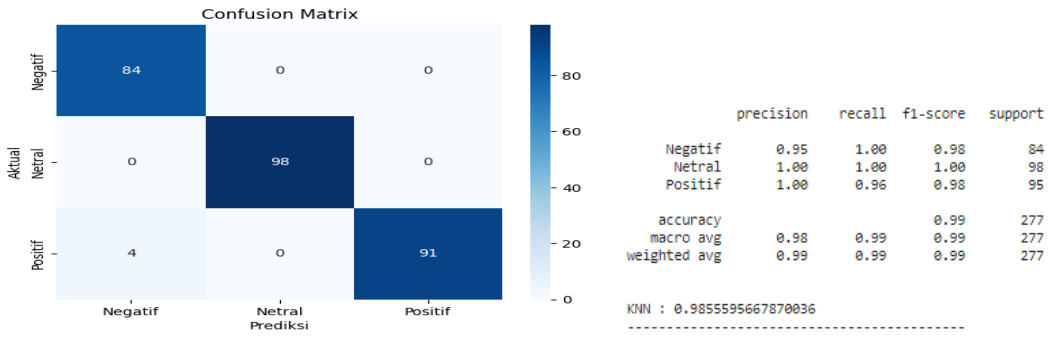


Figure 10. Confusion Matrix KNN+SMOTE

Evaluation of the KNN+SMOTE model results for sentiment analysis regarding white paracetamol provides a very satisfactory performance picture. We can describe the model's accuracy in detail using precision, recall, and F1-score metrics, along with information from the Confusion Matrix. The model can classify Negative sentiment with an accuracy of 0.95 and a recall of 1.00, indicating the ability to correctly and comprehensively identify reviews that reflect negative sentiment. The model reaches perfection for Neutral sentiment with precision, recall, and an F1-score of 1.00. Meanwhile, precision and recall reached 1.00 and 0.96 for Positive sentiment, indicating the model's ability to classify positive reviews very well. With an accuracy of 98%, the KNN+SMOTE model succeeded in classifying overall sentiment. Even though there is an imbalance in the number of samples between classes, the high F1-score values, especially for Negative and Neutral sentiment, indicate that the model can achieve a good balance between precision and recall.

The test findings indicate that the KNN+SMOTE model exhibits a significant improvement over the KNN model that is typically utilized, as shown in Table 2.

Table 2. Comparison of the Results

Model	Receptive Deciding	Result			
		F1 Score	Recall	Precision	Accuracy
KNN	Negatif	0.81	0.82	0.81	
	Neutral	1	0.81	1	**89,315%**
	Positif	0.92	0.94	0.92	
KNN+SMOTE	Negatif	0.98	1	0.95	
	Neutral	1	1	1	**98,556%**
	Positif	0.98	0.96	1	

Table 2 demonstrates that, compared to the KNN model, the KNN+SMOTE model performs significantly better across several model performance criteria, particularly in the "Negative" class. With KNN+SMOTE, the overall accuracy rises from 89.315% to 98.556%. This demonstrates that KNN+SMOTE is more accurate in predicting the attitude toward white people's use of paracetamol, particularly when bridging the class divide. The fundamental cause of this substantial improvement in accuracy is the SMOTE method's increased production of "Negative" class data.

Based on these findings, it can be said that the KNN model performs significantly better when the SMOTE approach is applied, particularly when it comes to the model's capacity to classify data from the minority class ("Negative"). The overall rise in F1-score, precision, recall, and accuracy indicates that the KNN+SMOTE model is more dependable in predicting the acceptance sentiment for consuming white paracetamol despite the modest loss in recall for the "Positive" class. Thus, the results of this evaluation provide confidence that the KNN model, which involves the SMOTE and TF-IDF processes, can provide a deep understanding of public sentiment toward the use of white paracetamol.

4. Conclusion

The performance evaluation results of the K-Nearest Neighbors (KNN) model in sentiment analysis regarding the use of white paracetamol show extraordinary achievements. By using precision, recall, and F1-score metrics, as well as information from the Confusion Matrix, the model successfully classifies sentiment with a very high level of accuracy. Negative sentiment can be identified with a precision of 0.95 and recall of 1.00, while Neutral sentiment is achieved with perfection in precision, recall, and F1-score of 1.00. Although Positive sentiment has a recall of slightly below 1.00 (0.96), the model still shows excellent ability in classifying positive reviews. An overall accuracy of 98,556% indicates the model's success in classifying sentiment comprehensively. The success of this model can be attributed to the selection of the optimal ***k*** value (***k***=1), the use of the SMOTE oversampling technique to overcome class imbalance, and the application of TF-IDF to improve feature representation in the dataset. The evaluation results

show that the model can capture nuances of sentiment well, even in cases of significant class imbalance.

This research contributes significantly to understanding public sentiment regarding using white paracetamol using approaches based on KNN, SMOTE, and Receptive Deciding. The research results show that the use of SMOTE successfully overcomes class continuity in the dataset, improving the performance of the KNN model in classifying positive and negative sentiments. Integration with Receptive Deciding provides additional adaptability to changing sentiment over time.

The developed model can provide sentiment assessments with high accuracy, and the results of confusion analysis show the model's ability to differentiate well between positive and negative sentiments regarding the use of white paracetamol. Validation of results on different datasets strengthens model generalization. The importance of rapid evaluation of public acceptance sentiment towards white paracetamol is becoming increasingly clear, and these findings may provide a basis for designing more effective health communication strategies. The research results also provide insight into the factors most influential in assessing sentiment, such as drug safety, efficacy, and side effects.

Although this research has made significant progress, it is essential to acknowledge its limitations, including those in the data collection and methods used. In future research, improved analysis techniques or the inclusion of more complex features could increase the accuracy and robustness of the model. Thus, this study contributes to understanding sentiment related to white paracetamol and opens the door for further exploration in developing machine learning-based adaptive sentiment models for public health contexts.

References

[1] M. Bach, “New Undersampling Method Based on the kNN Approach,” *Procedia Computer Science*, vol. 207, pp. 3397–3406, 2022, doi: 10.1016/j.procs.2022.09.399.
[2] M. J. Groot *et al.*, “4-acetaminophen (Paracetamol) levels in treated and untreated veal calves, an update,” *Food Control*, vol. 147, no. August 2022, p. 109577, 2023, doi: 10.1016/j.foodcont.2022.109577.
[3] A. Eliasen, S. Otnes, M. Matz, L. Aunsholt, and R. Mathiasen, “Safety of rapid intravenous paracetamol infusion in pediatric patients,” *Current Research Pharmacology Drug Discovery*, vol. 3, no. July 2021, pp. 2–5, 2022, doi: 10.1016/j.crphar.2021.100077.
[4] G. P. Milani, A. Mercante, D. Cattaneo, I. Alberti, C. Agostoni, and F. Benini, “Safety and efficacy of non-standard posology of paracetamol to manage pain in pediatric patients,” *Pharmacological Research*, vol. 197, no. September, pp. 1–3, 2023, doi: 10.1016/j.phrs.2023.106981.
[5] J. Augustino, F. Moshi, A. Joho, J. Faustine, and K. Mageda, “Dataset comparing the effectiveness of perineal cold pack application over oral paracetamol 1000mg on postpartum perineal pain among women after spontaneous vaginal delivery in Dodoma region,” *Data in Brief*, vol. 51, p. 109766, 2023, doi: 10.1016/j.dib.2023.109766.
[6] K. Kilic, H. Ikeda, T. Adachi, and Y. Kawamura, “Soft ground tunnel lithology classification using clustering-guided light gradient boosting machine,” *Journal of Rock Mechanics Geotechnical Engineering*, vol. 15, no. 11, pp. 2857–2867, 2023, doi: 10.1016/j.jrmge.2023.02.013.
[7] J. C. Macuácua, J. A. S. Centeno, and C. Amisse, “Data mining approach for dry bean seeds classification,” *Smart Agricultural Technology*, vol. 5, no. April, 2023, doi: 10.1016/j.atech.2023.100240.
[8] M. Umer *et al.*, “Scientific papers citation analysis using textual features and SMOTE resampling techniques,” *Pattern Recognition Letters*, vol. 150, pp. 250–257, 2021, doi: 10.1016/j.patrec.2021.07.009.
[9] J. Fonseca and F. Bacao, “Geometric SMOTE for imbalanced datasets with nominal and continuous features,” *Expert System with Application*, vol. 234, no. July, p. 121053, 2023, doi: 10.1016/j.eswa.2023.121053.
[10] A. N. Kasanah, M. Muladi, and U. Pujianto, “Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN,”

Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 3, no. 2, pp. 196–201, 2019, doi: 10.29207/resti.v3i2.945.

[11] S. G. Barus, "Klasifikasi Sentimen Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor Pada Ulasan Pengguna Aplikasi Pedulilindungi," *Senamika (Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasi)*, pp. 162–173, 2022.

[12] D. Gonzalez-Cuautle *et al.*, "Synthetic minority oversampling technique for optimizing classification tasks in botnet and intrusion-detection-system datasets," *Applied Science*, vol. 10, no. 3, 2020, doi: 10.3390/app10030794.

[13] Z. Chen, L. J. Zhou, X. Da Li, J. N. Zhang, and W. J. Huo, "The Lao text classification method based on KNN," *Procedia Computer Science*, vol. 166, pp. 523–528, 2020, doi: 10.1016/j.procs.2020.02.053.

[14] M. Suvarna and N. Venkategowda, "Performance Measure and Efficiency of Chemical Skin Burn Classification Using KNN Method," *Procedia Computer Science*, vol. 70, pp. 48–54, 2015, doi: 10.1016/j.procs.2015.10.028.

[15] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Applied Computing and Informatics*, vol. 12, no. 1, pp. 90–108, 2016, doi: 10.1016/j.aci.2014.10.001.

[16] I. R. Heruwidagdo, Suharjito, N. Hanafiah, and Y. Setiawan, "Performance of Information Technology Infrastructure Prediction using Machine Learning," *Procedia Computer Science*, vol. 179, no. 2020, pp. 515–523, 2021, doi: 10.1016/j.procs.2021.01.035.

[17] H. Zhu *et al.*, "Visualizing large-scale high-dimensional data via hierarchical embedding of KNN graphs," *Visual Informatics*, vol. 5, no. 2, pp. 51–59, 2021, doi: 10.1016/j.visinf.2021.06.002.

[18] D. S. Jodas, L. A. Passos, A. Adeel, and J. P. Papa, "PL-kNN: A Python-based implementation of a parameterless k-Nearest Neighbors classifier [Formula presented]," *Software Impacts*, vol. 15, no. December 2022, p. 100459, 2023, doi: 10.1016/j.simpa.2022.100459.

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

[20] M. Yuvalı, B. Yaman, and Ö. Tosun, "Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets," *Mathematics*, vol. 10, no. 3, 2022, doi: 10.3390/math10030311.

[21] S. Kumar and T. D. Singh, "Fake news detection on Hindi news dataset," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 289–297, 2022, doi: 10.1016/j.gltp.2022.03.014.

[22] Z. E. Fitri, L. N. Sahenda, P. S. D. Puspitasari, P. Destarianto, D. L. Rukmi, and A. M. N. Imron, "The The Classification of Acute Respiratory Infection (ARI) Bacteria Based on K-Nearest Neighbor," *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 12, no. 2, p. 91, 2021, doi: 10.24843/lkjiti.2021.v12.i02.p03.

[23] S. Wazarkar, B. N. Keshavamurthy, and A. Hussain, "Region-based Segmentation of Social Images Using Soft KNN Algorithm," *Procedia Computer Science*, vol. 125, pp. 93–98, 2018, doi: 10.1016/j.procs.2017.12.014.

[24] A. Imakura, M. Kihira, Y. Okada, and T. Sakurai, "Another use of SMOTE for interpretable data collaboration analysis," *Expert System with Application*, vol. 228, no. August 2022, p. 120385, 2023, doi: 10.1016/j.eswa.2023.120385.

[25] A. Kummer, T. Ruppert, T. Medvegy, and J. Abonyi, "Machine learning-based software sensors for machine state monitoring - The role of SMOTE-based data augmentation," *Results in Engineering*, vol. 16, no. October, 2022, doi: 10.1016/j.rineng.2022.100778.

[26] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *Journal of King Saud University - Computer and Information Science*, vol. 34, no. 6, pp. 3413–3423, 2022, doi: 10.1016/j.jksuci.2021.01.014.

[27] T. M. Mohamed, "Pulsar selection using fuzzy knn classifier," *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 1–6, 2018, doi: 10.1016/j.fcij.2017.11.001.

[28] M. Liang and T. Niu, "Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs," *Procedia Computer Science*, vol. 208, pp. 460–470, 2022, doi: 10.1016/j.procs.2022.10.064.

[29] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," *Procedia Computer Science*, vol. 116, pp. 137–143,

2017, doi: 10.1016/j.procs.2017.10.039.
[30] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, “Sentiment analysis and classification of Indian farmers’ protest using twitter data,” *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100019, 2021, doi: 10.1016/j.jjimei.2021.100019.
[31] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, “The impact of features extraction on the sentiment analysis,” *Procedia Computer Science*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.