Knowledge Engineering and Data Science (KEDS)
Vol 7, No 1, April 2024, pp. 13–26

pISSN 2597-4602
eISSN 2597-4637

Hybrid Method for User Review Sentiment Categorization in ChatGPT Application Using N-Gram and Word2Vec Features

Husna Luthfiatun Nisa [1,*], Atina Ahdika [2]

Department of Statistics, Universitas Islam Indonesia
Kaliurang St No.Km. 14,5, Krawitan, Umbulmartani, Ngemplak, Sleman, Yogyakarta 55584, Indonesia

*[1] husna.nisa@students.uii.ac.id *; [2] atina.a@uii.ac.id*
** corresponding author*

ARTICLE INFO

Article history:
Received 26 January 2024
Revised 25 March 2024
Accepted 18 April 2024
Published online 24 April 2024

Keywords:
ChatGPT Application
Sentiment Analysis
Naïve Bayes
K-Means

ABSTRACT

The rapid development of Artificial Intelligence (AI) has significantly influenced nearly all aspects of life. One AI product widely used by people worldwide is the Chat Generative Pre-Training Transformer (ChatGPT), which can respond to questions conversationally. Although data indicates that the use of ChatGPT in Indonesia is less widespread than in other countries, a Populix survey reveals that half of the respondents have utilized ChatGPT, using AI more than once a month. This indicates its crucial role among the Indonesian population. ChatGPT is not limited to browsers; it is also available as a downloadable application on the Google Play Store. The ChatGPT application has garnered various user reviews, particularly those from Indonesia. Therefore, this research employs the Naïve Bayes Classifier and K-Means Clustering to classify sentiments and group user reviews of the ChatGPT application originating from Indonesia. The study utilizes TF-IDF and Word2Vec as feature extraction methods, combining various N-Gram in data preprocessing to consider the context of sequentially arranged words that may carry meaning. The best classification results are obtained from the trigram classification model, as indicated by precision, recall, and accuracy values of 0.99 each, along with an F1-score of 1. Clustering also yields positive results, with some overlapping, yet words within clusters exhibit high similarity. Categorization results suggest that user reviews of the ChatGPT application from Indonesia tend to be positive, expressing satisfaction impressions, providing feedback for feature development, and expressing hope for the continued availability of the accessible version of ChatGPT due to its remarkable benefits.

I. Introduction

Artificial Intelligence (AI) was first introduced in 1956 by John McCarthy during a conference at Dartmouth College. Since then, the progress of AI has experienced rapid advancements, propelled by increasingly sophisticated technological developments. Dr Lukas, a lecturer and the chairman of the Indonesia Artificial Intelligence Society (IAIS), explained that the evolution of AI spans from theoretical foundations to the internet era [1]. The influence of AI permeates various aspects of human life.

One of the prominent AI products is the Chat Generative Pre-Training Transformer (ChatGPT), which OpenAI first released on November 30, 2022. Data from databoks indicates that from then until April 2023, Indonesia did not fall within the top five countries with the highest ChatGPT usage. However, more than half of the Indonesian population (52%), as surveyed by Populix (with 1014 respondents from various age groups), have used ChatGPT [2]. Interestingly, 40% have utilized AI more than once a month. Since then, ChatGPT users in Indonesia have continued to increase, so that in March 2024, Indonesia was ranked fourth with the most ChatGPT users [3].

The data indicates that ChatGPT is crucial among the Indonesian population, allowing users to pose various questions. The architecture of ChatGPT is a type of neural network capable of Natural

https://doi.org/10.17977/um018v7i12024p13-26

Language Processing (NLP) [4], resulting in responses that closely resemble human language. This aligns with ChatGPT's primary goal, as quoted from Britannica, which is to make ChatGPT applicable in various contexts such as chatbots, language translation, automated content creation, and even solving mathematical and programming problems.

In July 2023, OpenAI released the ChatGPT application on the Google Play Store, facilitating easier user access. This application garnered over 50 million downloads with a 4.7 out of 5 rating and 445 thousand user reviews worldwide, including Indonesia. User reviews provide positive or negative assessments and reveal preferences, expectations, and in-depth insights into how the public embraces this technology.

Like several preceding studies, this research employs the Naïve Bayes method to classify sentiment in reviews [5][6][7][8][9][10][11]. Using Naïve Bayes as a classifier is based on research by D. Oktavia and D.D.L.C. Pardede, who stated that the Naïve Bayes method is still superior when the test data used is categorical [12]. A study by Rafael E. Banchs also proves that the Naïve Bayes method has high speed and accuracy when applied to large volumes of data [7].

In addition, the K-Means method is used to group reviews and capture discussed topics [13][14][15][16][17][18]. The technique in K-Means clustering is adequate for handling large amounts of numerical data [19]. This is by the type of data used as grouping input (Word2Vec results), which is numerical. This data represents the words in the review, so the processed K-Means algorithm will form clusters that show the evaluation context.

In classification, the feature extraction process of text review data is carried out using TF-IDF [9][20] with N-Gram [5][6][7][8][21], as well as Word2Vec which produces word embedding vectors [16][17][22][23]. Dividing data into train and test data uses stratified random sampling [11]. Meanwhile, in the grouping process, the feature extraction used is Word2Vec and the distance matrix formed is calculated using cosine similarity [14][18].

Research results using Naïve Bayes reveal that the use of unigram [7][9], bigram [8], trigram [5], and the combination of unigram-bigram-trigram [6][21] yield the best outcomes based on precision, recall, F1-score, and accuracy values. The application of Word2Vec in classification also exhibits relatively high accuracy, reaching 77% [24]. That suggests that in the Naïve Bayes classification, TF-IDF performs better than Word2Vec [20]. Meanwhile, the text grouping results indicate that the application of K-Means effectively depicts the characteristics of data in each cluster [13][16][18] through visualization of the most frequently appearing words [17].

Therefore, this research aims to compare the evaluation metrics of the two methods used to determine their effectiveness in classifying and grouping reviews. This was done because of the significant use of the ChatGPT application in Indonesia, so this research is also aimed at understanding the sentiment and characteristics of reviews of ChatGPT application users from Indonesia, both generally and contextually. The research outcomes are anticipated to offer insights into the perceptions of the Indonesian community towards the ChatGPT application, serving as a foundation for ChatGPT application development. Furthermore, the findings can guide interested parties looking to adapt, integrate, and market similar technologies.

In this research, several distinctions from previous studies are identified. First, this research uses the latest data from the Play Store, while the last uses different data sources. This data allows research to focus directly on user reviews regarding the ChatGPT application. Second, the incorporation of N-Gram analysis (with N being 1 (unigram), 2 (bigram), 3 (trigram), and their combination), taking into consideration the potential meanings of various word combinations in the reviews. Third, feature extraction involves the application of both TF-IDF and Word2Vec, with the latter exclusively utilizing unigrams. Fourth, this study compares the evaluation of classification results by considering the data sharing method, namely using the stratified random sampling method and not using any other method, because the data randomization process can affect the classifier's performance. Fifth, the clustering results will be visualized based on frequently occurring words within the review documents of each cluster, and the evaluation will be conducted using the Dunn Index.

II. Methods

A. Data

The data employed in this research consists of user reviews of the ChatGPT application originating from Indonesia from the initial release until September 2023. The attributes utilized in the study include Text (user application reviews) and Labels (positive and negative categories for each review).

B. Methods

This research uses text mining to extract useful information from data sources by identifying and exploring interesting patterns [25]. The review sentences are classified using the Naïve Bayes Classifier, and the review grouping is achieved through K-Means Clustering. The research stages are illustrated in Figure 1.

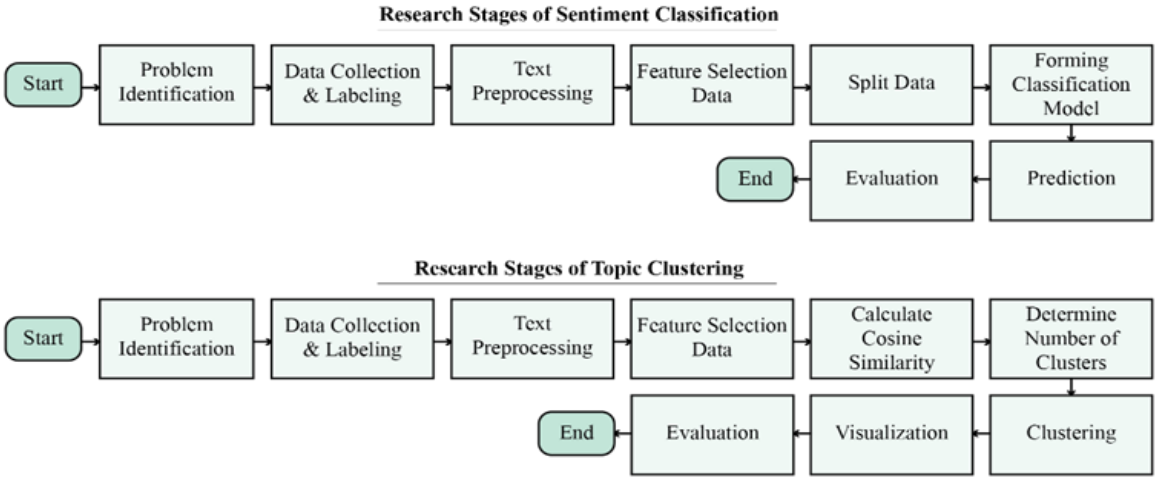

Fig. 1. Research stages

The data was collected by scraping the ChatGPT application page on the Google Play Store. A total of 1302 data points were labelled based on the ratings provided by users. Ratings 1 and 2 were labelled unfavourable, while ratings 4 and 5 were labelled positive. For rating 3, labels were assigned based on the substance of the review, whether it was more likely to be a positive or negative review.

Text preprocessing was conducted in four stages, namely N-gram tokenization (dividing the text into smaller units, such as single words or phrases), standardization and cleansing (standardizing and cleaning the text), stop-word removal (eliminating words that do not add value or information in the context of analysis), and stemming (reducing words to their base form) [26].

Unstructured data is very challenging for computers to process [26]. Therefore, the Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec methods in feature selection transform textual data into numerical form. TF-IDF can identify statistically significant words within a document collection[26]. Mathematically, the TF-IDF value is obtained using the following formula as in (1). Where tf denotes the number of times the word appears in that specific document, and idf across all documents.

$$tfidf = tf \times idf = tf \times \log\left(\frac{N}{df}\right) \quad (1)$$

Furthermore, Word2Vec is a method used to convert a word into a vector by training the words based on the words around them [27]. One commonly used method within Word2Vec is Continuous Bag Of Words (CBOW). The CBOW method aims to predict the central or target word from the context words surrounding it within a specific range [27].

After the data is prepared, to carry out sentiment classification, train data is needed to create a classification model and test data to test the model's performance. Data division into train and test data is carried out randomly without a specific method. Hence, the proportion of data for each label in the

train and test data tends to differ from the original data. Therefore, stratified sampling is also used because the principle of this method is to partition the population into strata so that the units in a stratum are similar [28]. Thus, the proportion of data for each label in the train and test data will be the same as the original data, and the units selected from each stratum will tend to represent the population as a whole.

Sentiment classification is carried out using the Naïve Bayes method, where the classifier model is based on Bayes's rules, which are used to estimate conditional probabilities [26]. For example, there is a dataset where each attribute is assumed to be a random variable. These attributes are denoted as $\{A_1, A_2, \dots, A_n\}$ and will be classified into class C. Classification is considered accurate when the conditional probability as in (2).

$$P\{C|A_1, A_2, \dots, A_n\} \tag{2}$$

Reaches its maximum value. The Bayes formula is applied as follows to maximize it.

1. Calculating the posterior probability $P\{C_j|A_1, A_2, \dots, A_n\}$ for all classes C_j using the formula as in (3).

$$P\{C_j|A_1, A_2, \dots, A_n\} = \frac{P\{A_1, A_2, \dots, A_n|C_j\}.\ P\{C_j\}}{P\{A_1, A_2, \dots, A_n\}} \tag{3}$$

2. Selecting the class C_k that maximizes $P\{C_j|A_1, A_2, \dots, A_n\}$ or $P\{A_1, A_2, \dots, A_n|C_j\}.P\{C_j\}$. In this case, the denominator $P\{A_1, A_2, \dots, A_n\}$ is the same for each class, so it can be ignored.

Naïve Bayes Classifier assumes independence among events [29]. In this case, each attribute is assumed to be independent within a specific class C, as in (4).

$$P\{A_1, A_2, \dots, A_n|C\} = P\{A_1|C\}.P\{A_2|C\} \dots P\{A_n|C\} \tag{4}$$

Next, estimate the probabilities $P\{A_i|C_j\}$ for all attributes A_i and class C_j. Therefore, a new and unknown object will be classified into class C_k if the probability corresponding to that class is in (5).

$$P\{C_k\}.\prod P\{A_i|C_k\} \tag{5}$$

Evaluating the classification results is done by calculating several evaluation metrics such as precision, recall, F1-score, and accuracy, which are mathematically expressed as in (6) to (9).

$$Precision\ (P) = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{6}$$

$$Recall\ (R) = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{7}$$

$$F_{Measure}\ (F_i) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

$$Accuracy\ (A) = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \tag{9}$$

Apart from being classified, the initial dataset of Word2Vec results will be partitioned using K-Means to understand the topics discussed in the review. K-Means is a method in data analysis that aims to partition a set of input data into K groups or clusters, where K is a predetermined number of clusters [30]. K-means clustering in text analysis is also a method to discover natural groupings based on a document matrix, but no hierarchical structure or visualization type of dendrogram has been created. Instead, clustering is formed around cluster seeds or predefined starting points based on similarity or minimum distance [26].

The first process in grouping reviews is calculating cosine similarity, a metric for measuring the extent to which two n-dimensional vectors are similar. In document comparison, cosine similarity is

utilized to gauge the semantic similarity between clusters, reflecting the conceptual similarity between documents [31]. Mathematically, the calculation of cosine similarity is expressed as in (10).

$$Similarity(X,Y) = \frac{X \cdot Y}{\|X\| \times \|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \times \sum_{i=1}^{n} Y_i^2} \quad (10)$$

X and Y represent the documents to be compared, the range of cosine similarity values is between 0° and 180°. The smaller the angle, the more similar the two documents are in their semantic context [31].

Next, determine the optimal number of clusters using the Silhouette method. The silhouette coefficient, s_i, for the i-th data point is mathematically calculated as in (11) [26].

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (11)$$

Where a_i is the average distance from data point i to all other data points in the same cluster, b_i is determined by calculating the distance to all other data points in all other clusters and finding the minimum average distance from a data point to different clusters. The Silhouette score ranges from -1 to 1. Typically, an optimal cluster is based on a higher Silhouette score, as it indicates that objects within the cluster are more similar to each other and more dissimilar from objects in different clusters.

Once the clusters are formed, internal validity measurements (obtained from the data and the model created) are carried out to assess the suitability of the clustering solution. One measure is the Dunn index, the ratio of the minimum distance between inter-cluster data points to the maximum intra-cluster distance as in (12) [26].

$$Dunn\ Index = \frac{\min\{d_{inter}\}}{\max\{d_{intra}\}} \quad (12)$$

Where $\min\{d_{inter}\}$, minimum inter-cluster distance is the minimum distance between two points in different clusters, and $\max\{d_{intra}\}$, maximum intra-cluster distance is the maximum distance between two points in the same cluster. The distance calculation is carried out from the Cosine Similarity results but with slight modifications. The distance between two points is obtained from the difference between one and the cosine similarity value because the cosine similarity value reflects the level of similarity between two points, not the distance between them. A better clustering model will have a higher Dunn index value [26].

III. Result and Discussion

The review of the ChatGPT application in this research was written in Indonesian. Reviews constitute unstructured data that needs to be processed before further analysis can be conducted. The initial step involves assigning labels or categories of positive and negative classes for each review based on the rating. Reviews deemed neutral are not included in the analysis. Table 1 shows that most of the 1302 review documents in the corpus are categorized as positive reviews.

Table 1. Number of reviews

Label	Amount	Percentage
Positive	1198	92%
Negative	104	8%
Total	1302	100%

From Table 1, only 8% of the overall review documents fall into the harmful review category. This information adequately reflects many positive responses from ChatGPT application users. Nevertheless, it is also noteworthy that there may be reviews indicating application shortcomings that must be addressed.

The writing styles in the provided reviews vary significantly, encompassing capital letters, prefixes, slang words, abbreviations, conjunctions, punctuation, and emojis. Therefore, text

preprocessing is conducted to produce text in its base form and free from various punctuation and emojis. An example of the text preprocessing results is displayed in Table 2.

Table 2. Results of text preprocessing

Actual Texts	Result of Text Preprocessing
Aplikasi ini sangat membantu saya 👍	*aplikasi sangat bantu*
Jawaban ny ga sesuai dgn soal yg sy sampaikan.	*jawab tidak suai soal sampai*

Based on Table 2, the text resulting from preprocessing has been converted into lowercase form, which consistently impacts machine learning classifiers [32]. The preprocessing results also show the text without prefixes, abbreviated words, conjunctions, punctuation, and emojis. Some less valuable words in the analysis, such as "ini" and "saya", have also been removed. This preprocessed text greatly assists the machine in capturing unique words from all review documents. For example, the words "*Aplikasi*" and "*aplikasi*" will be detected as the same by the machine, as their preprocessed form is "*aplikasi*". Thus, the machine comprehends the content more effectively.

In the preprocessing stage, N-Gram tokenization is also performed, where the review text is segmented into single words and phrases of two or three words. Furthermore, combinations of word and phrase segments are done to assess their effectiveness in text review classification. An example of the tokenized text results for the sentence "bagus aplikasi sangat bantu gratis" with N-Gram is displayed in Table 3.

Table 3. Results of text preprocessing

N-Gram	Result of Tokenizing Text
Unigram	*"bagus", "aplikasi", "sangat", "bantu", "gratis"*
Bigram	*"bagus aplikasi", "aplikasi sangat", "sangat bantu", "bantu gratis"*
Trigram	*"bagus aplikasi sangat", "aplikasi sangat bantu", "sangat bantu gratis"*
Unigram – Bigram	*"bagus", "aplikasi", "sangat", "bantu", "gratis", "bagus aplikasi", "aplikasi sangat", "sangat bantu", "bantu gratis"*
Unigram – Trigram	*"bagus", "aplikasi", "sangat", "bantu", "gratis", "bagus aplikasi sangat", "aplikasi sangat bantu", "sangat bantu gratis"*
Bigram – Trigram	*"bagus aplikasi", "aplikasi sangat", "sangat bantu", "bantu gratis", "bagus aplikasi sangat", "aplikasi sangat bantu", "sangat bantu gratis"*
Unigram – Bigram – Trigram	*"bagus", "aplikasi", "sangat", "bantu", "gratis", "bagus aplikasi", "aplikasi sangat", "sangat bantu", "bantu gratis", "bagus aplikasi sangat", "aplikasi sangat bantu", "sangat bantu gratis"*

Subsequently, the tokenized text results from preprocessing are transformed into structured data, i.e., data in numerical form. Two feature selection methods are employed in converting textual data into numerical data: the TF-IDF method and Word2Vec. TF-IDF calculations are performed for all the data with N-Gram tokens. Thus, seven-word matrices are formed based on the combination of the N-gram used in this research. Each row represents a review document, and each column represents unique words from all review documents, except for the label column, which contains labels for each review document.

The number of rows for the entire word matrix is the same, namely 1302, corresponding to the number of review documents. Meanwhile, the number of columns or unique words for each matrix differs due to the influence of N-Gram tokens. The number of unique words for unigram tokens is 842, for bigram tokens is 3812, and for trigram tokens is 4363. Then, the number of unique words for a combination of unigram and bigram tokens is 4654; for a combination of unigram and trigram tokens, it is 5205. A combination of bigram and trigram tokens is 8175, and a combination of all three is 9017. Table 4 presents an example of the word matrix formed by combining unigram, bigram, and trigram tokens.

Table 4. Results of text preprocessing

Text	label	*aplikasi*	*benar*	…	*aplikasi_ benar*	*benar_ sangat*	…	*aplikasi_ benar_ sangat*	*benar_ sangat_ bagus*
1	positive	1.618697	4.036162	…	5.785362	6.478509	…	7.171656	7.171656
2	negative	0	0	…	0	0	…	0	0

Based on the information presented in Table 4, each numeric cell represents the TF-IDF value of the corresponding word. The higher the TF-IDF value, the more unique a word is within a document and the less frequently it appears in other documents in the corpus. Conversely, the lower the TF-IDF value, the more common and frequent a word is within a document and across the corpus. Meanwhile, a TF-IDF value of 0 indicates that a particular word is absent in the document. This is because the term frequency (TF), representing the number of times the word appears in that specific document, is 0, resulting in multiplication with the Inverse Document Frequency (IDF) of that word across all documents, yielding a value of 0.

Meanwhile, the results of the Word2Vec calculation are only presented for review data with unigram tokens. The output consists of vectors for each unique word in the corpus, ensuring that identical words in different documents have the same vector values. This differs from TF-IDF results, where the TF-IDF value for a word may vary in each document. Word2Vec aims to transform textual data into numerical data through a specific algorithm, in this case, using the Continuous Bag Of Words (CBOW) algorithm. Therefore, each word is input to predict a particular vector of words.

In this case study, a vector length of 200 will be formed, considering that reviews tend to be written with varying word counts. While some reviews may consist of only one to five words, others are written with more words. Determining the vector length considers reviews written with many words, ensuring that each word is expected to have its vector value. Since each review has a different word length, the number of vectors also varies. This inconsistency cannot proceed to the classification process. Therefore, a sentence vector will be formed for each review document, ensuring that the vector length for each review is the same. The algorithm for creating a sentence vector involves summing the vector values of each word in the review document and dividing by the number of words in that document. The result is a single vector of average values. Here is an example of a sentence vector for two review documents with different word counts in each review, for instance, text 1 being "*aplikasi sangat bagus*" and text 2 being "*bagus aplikasi sangat bantu gratis*".

$$text\ 1 : [-0.84 \quad -0.21 \quad 0.03 \quad \dots \quad -0.16 \quad -0.09 \quad 0.09 \quad 0.14]$$

$$text\ 2 : [-0.61 \quad -0.10 \quad 0.29 \quad \dots \quad 0.05 \quad -0.24 \quad -0.02 \quad 0.15]$$

The values in these vectors reflect the semantic relationships between words and how often these words appear together in the corpus. Generally, the closer the value is to 1, the more substantial the semantic relationship or the higher the similarity in meaning. Conversely, if the vector value approaches zero or is negative, these words rarely appear together or have a lower semantic relationship. In this case, these words may be more independent of each other in their usage within the corpus.

A. Naïve Bayes Classification Result

After going through the data preparation process, the next step is to shuffle the data so that each review document has an equal probability of being used as training or test data. The training data used to create the classification model comprises 80% of the corpus, while the remaining 20% is used as test data as present in Table 5.

Table 5. Number of training and test data based on label

Data	Without Stratification			With Stratification		
	Positive	Negative	Total	Positive	Negative	Total
Train	963	78	1041	958	83	1041
Test	235	26	261	240	21	261
Total	1198	104	1302	1198	104	1302

A critical calculation in creating a Naïve Bayes classification model is the prior probability, which is the probability of each class occurring in the training data. The likelihood of positive and negative sentiment appearing in the data train without stratification is 0.93 and 0.07, respectively, while in the data train with stratification, it is 0.92 and 0.08, respectively. This indicates that positive sentiment is more common than negative sentiment.

Furthermore, the classification model is formed using the Gaussian Naïve Bayes algorithm, where the probability of each feature in each class is calculated using the Probability Density Function

(PDF) of the normal distribution. Once the prior probability and the probability of each feature occurring in each class are calculated, the classification process can be executed. This stage involves using test data, where the likelihood of a review being classified into a particular class is calculated based on the known features. If the probability of a review being in the positive class is higher than the probability in the negative class, the review is classified as positive, and vice versa.

The test data classification results are presented in the confusion matrix in Table 6 and Table 7, where N denotes negative, and P denotes positive.

Table 6. Confusion matrix of classification result without data stratification

		Prediction															
		Unigram		**Bigram**		**Trigram**		**Unigram-Bigram**		**Unigram-Trigram**		**Bigram-Trigram**		**Unigram-Bigram-Trigram**		**Word2Vec**	
		N	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**
Actual	**N**	14	12	17	9	25	1	13	13	15	11	17	9	13	13	22	4
	P	2	233	1	234	1	234	1	234	2	233	1	234	1	234	33	202

Table 7. Confusion matrix of classification result with data stratification

		Prediction															
		Unigram		**Bigram**		**Trigram**		**Unigram-Bigram**		**Unigram-Trigram**		**Bigram-Trigram**		**Unigram-Bigram-Trigram**		**Word2Vec**	
		N	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**	**N**	**P**
Actual	**N**	7	14	12	9	20	1	7	14	9	12	12	9	8	13	14	7
	P	2	238	3	237	2	238	3	237	3	237	3	237	2	238	18	222

Based on the information in Table 6 and Table 7, it is evident that there are still misclassifications. This is likely due to a significant difference in the number of positive and negative sentiment reviews, making it challenging for the model to recognize negative sentiments. As a result, some negative sentiment reviews are misclassified as positive sentiment reviews. Therefore, the classification results are evaluated to assess the model's effectiveness in classifying reviews. The evaluation metrics used are precision (P), recall (R), F1-score, and accuracy (A). The evaluation of classification result can be seen in Table 8.

Table 8. Evaluation of classification result

Feature	**Without Stratification**				**With Stratification**			
	P	**R**	**F1-Score**	**A**	**P**	**R**	**F1-Score**	**A**
Unigram	0.95	0.99	0.97	0.95	0.94	0.99	0.97	0.94
Bigram	0.96	0.99	0.98	0.96	0.96	0.99	0.98	0.95
Trigram	**0.99**	**0.99**	**1**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
Unigram – Bigram	0.95	0.99	0.97	0.95	0.94	0.99	0.97	0.93
Unigram – Trigram	0.95	0.99	0.97	0.95	0.95	0.99	0.97	0.94
Bigram – Trigram	0.96	0.99	0.98	0.96	0.96	0.99	0.98	0.95
Unigram – Bigram – Trigram	0.95	0.99	0.97	0.95	0.95	0.99	0.97	0.94
Word2Vec	0.98	0.86	0.92	0.86	0.97	0.93	0.95	0.90

Based on the information in Table 8, the performance of classification results between data without stratification and stratified data does not show a significant difference. For N-gram features (except trigrams), the accuracy of classification results from stratified data is 1% smaller than that from data without stratification, as in study A. Somasundaram and S. Reddy show that the Naïve Bayes algorithm can still provide good results when the data for each class is not balanced [33][34]. In contrast to the Word2Vec feature, stratified data's classification performance is higher than those without stratification. This is due to the nature of the Word2Vec feature, which considers relationships between words. When the class distribution changes, the vector representation of words in the Word2Vec model also changes, thereby affecting the model in classifying text.

The trigram model demonstrates excellent performance based on the given evaluation metric values with stratified and unstratified data. With a precision of 0.99, the N-Gram TF-IDF classification model can predict positive outcomes with very high accuracy and produces few false positives. The equally high recall value (0.99) indicates that the model effectively identifies the most favourable

outcomes. The F1-Score, reaching the maximum value of 1, signifies an optimal balance between precision and recall. The model's accuracy, reaching 0.99, confirms that it provides overall predictions with high precision.

Nevertheless, the Word2Vec classification model performs better with stratified data. A precision of 0.97 indicates a high accuracy in predicting positive outcomes, and a recall of 0.93 demonstrates a high accuracy in predicting actual positive outcomes. The balanced nature of these two values is also optimal, as indicated by an F1-Score of 0.95. The model's precision in accurately predicting positive and negative outcomes is also high, at 0.90. With these outstanding evaluation metrics, the model with N-Gram Trigram is the preferred choice for classification in this context, offering highly optimal and reliable performance.

A. K-Means Clustering Result

After achieving satisfactory performance in classification, an equally crucial aspect is extracting context from user reviews of the application. Insights related to user reviews of the ChatGPT application originating from Indonesia, contextualized, are obtained through clustering. The silhouette score is the metric employed to determine the number of clusters formed in this study.

According to theory [26], a higher silhouette score indicates that objects within a cluster are more similar and more distinct between clusters. Therefore, the number of groups formed, as seen in Figure 2, is nine groups.

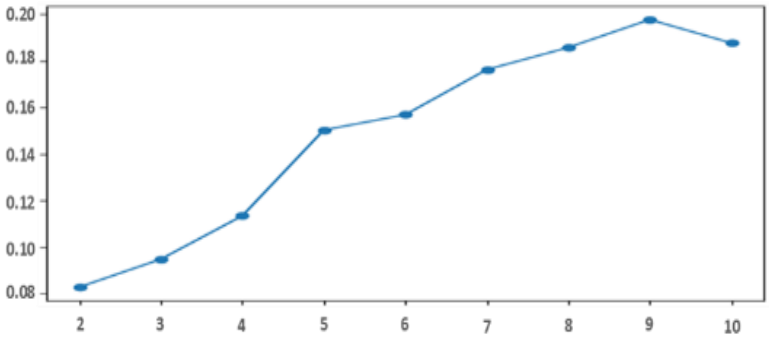

Fig. 2. Result of silhouette score

In the first cluster, there are 81 reviews from application users, where positive sentiments dominate more than negative sentiments. The context discussed in each sentiment is similar, focusing on the artificial intelligence of the ChatGPT application discussed from both negative and positive perspectives. On the negative side, users talk about the application being good, but its capabilities still need improvement and confusion. As a result, they expressed hope that the application will be updated soon. On the positive side, users perceive that the AI's capabilities are extraordinary regarding the information it possesses and ChatGPT's responses. Figure 3 show the visualization for cluster 1.

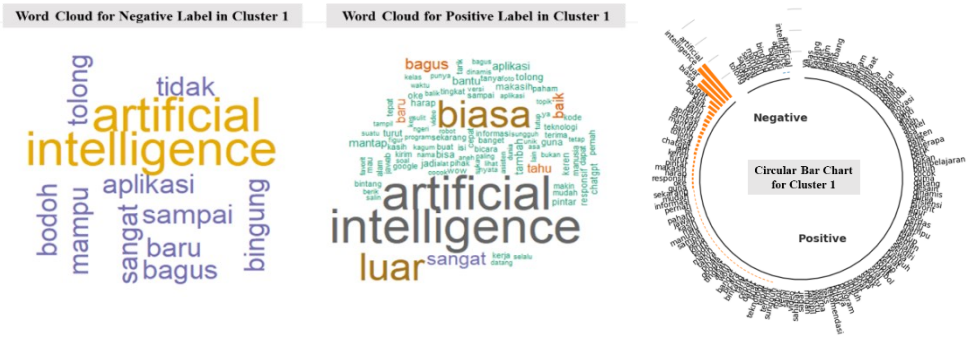

Fig. 3. Visualization for cluster 1

The second cluster has 514 reviews from application users, most of which are positive. The context discussed in positive sentiments revolves around the application being good because it provides fast and comprehensive responses, helpful information, and suggestions for developing application features. Some feature suggestions that users propose include message editing, message pinning, text scanning from images, and automatic text from voice notes. Additionally, users express the hope that the accessible version of ChatGPT will always be available as it benefits them. On the negative side, the context discussed in this cluster includes issues accessing the application, errors, difficulties in providing answers, and incomplete features. Figure 4 show the visualization for cluster 2.

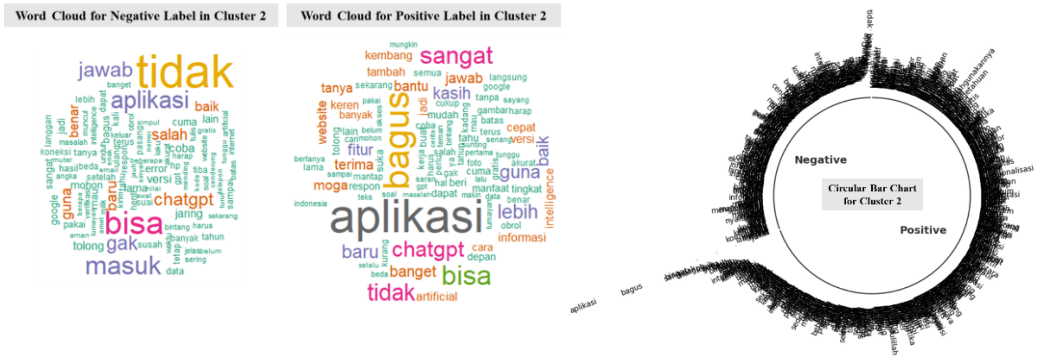

Fig. 4. Visualization for cluster 2

In the third cluster, there are 165 reviews from application users, with only one review expressing a negative sentiment. This cluster reflects overall satisfaction with the application. On the positive side, users provide reviews stating that the application is good overall and is recommended for use. On the contrary, no specific details were mentioned, which led one user to give a negative review of the application. Figure 5 show the visualization for cluster 3.

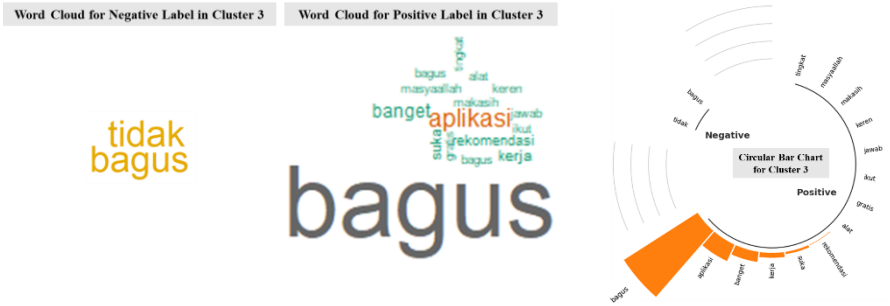

Fig. 5. Visualization for cluster 3

The fourth cluster has 54 reviews from application users, all expressing positive sentiments. Like the third cluster, the fourth cluster indicates the context of application users' reviews about their satisfaction with the application. However, the expressions of satisfaction are phrased differently. One aspect contributing to user satisfaction is the responses users obtain while using the application. Figure 6 show the visualization for cluster 4.

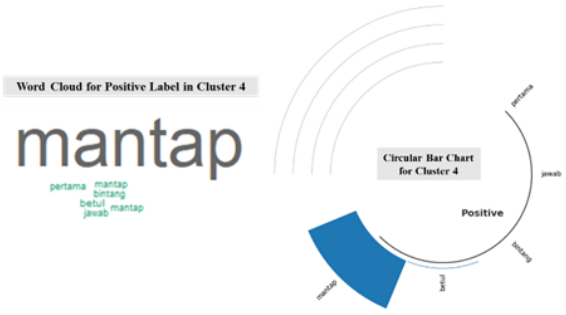

Fig. 6. Visualization for cluster 4

Like the fourth, in the fifth cluster, all the reviews are positive in the fifth clustering. Eighty-one reviews in this cluster provide information that application users are impressed with the application's performance. The reviews also refer to the application's advanced features, leading to user satisfaction. Figure 7 show the visualization for cluster 5.

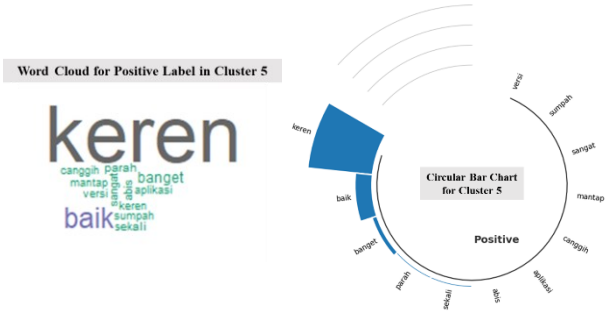

Fig. 7. Visualization for cluster 5

The sixth cluster has 372 reviews from application users, most of which are positive sentiments. In this cluster, users feel the benefits of the application in various aspects, especially in completing tasks and finding information, making them grateful for the existence of this application. Although the information generated is still limited, it remains relevant to date. This is due to the architecture of ChatGPT, which strives to learn from user responses. On the negative side, some users mention difficulties in the registration process when logging into the application. Figure 8 show the visualization for cluster 6.

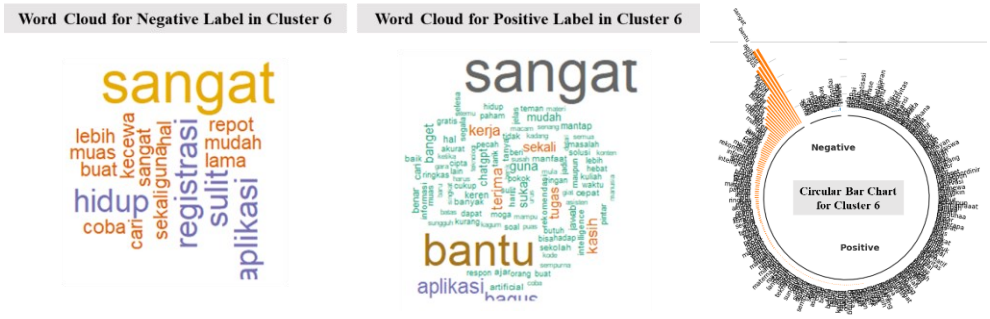

Fig. 8. Visualization for cluster 6

In the seventh cluster, all the reviews are positive. A total of nine reviews in this cluster provide information that application users are impressed with the technology adopted by the ChatGPT application. In cluster eight, out of 24 user reviews, only one negative sentiment exists. Detailed information about what issues caused the user to leave a negative review is unknown. Nevertheless, most application users in this cluster express their satisfaction with the application using words different from reviews in other clusters. Figure 9 show the visualization for cluster 7 and cluster 8.

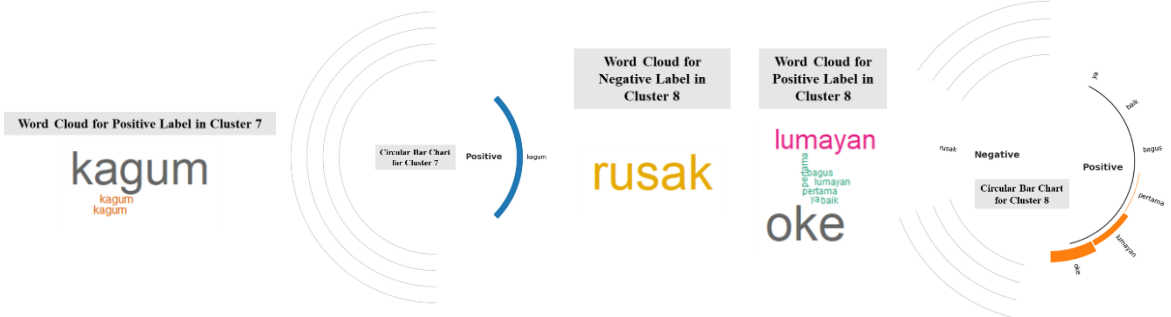

Fig. 9. Visualization for cluster 7 and cluster 8

In the last cluster, there are two positive reviews, where the context discussed by application users refers to the users's ability to manage the application, possibly in terms of registration, adjusting application permissions, or other security features. Figure 10 show the visualization for cluster 7 and cluster 9. Table 9 shows a summary of the core contexts of each cluster.

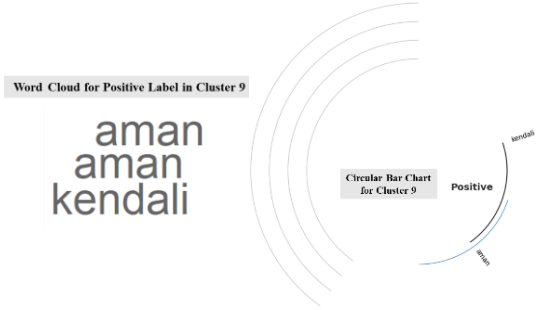

Fig. 10. Visualization for cluster 9

Table 9. Core contexts in each cluster

th Cluster	Core Contexts	th Cluster	Core Contexts
1	Artificial Intelligence application ChatGPT.	6	Benefits of the application and complaints during the registration process.
2	Appreciation, criticism, and suggestions for the application.	7	User expressions about the application.
3	User satisfaction with the application.	8	User satisfaction with the application.
4	User satisfaction with the application.	9	User ability to manage the application.
5	User impressions of the application's performance.		

Based on the clustering results from Table 9, it can be observed that some clusters share similar contexts, albeit differing in the choice of words. This is supported by the Dunn Index value of 0.05, indicating a low partition, thus overlapping. However, a low Dunn Index value only sometimes shows poor performance. Reviews have high subjectivity, and similarities among reviews can be more complex than other numerical data. Therefore, the low Dunn Index results from the difficulty in separating clusters.

IV. Conclusion

The performance of all formed classification models is excellent, as they produce precision, recall, F1-score, and accuracy values above 0.8. However, the best classification model in this study is the trigram model extracted using TF-IDF. The precision, recall, and accuracy values produced by this model are each 0.99, and the F1-score is 1. This indicates the appropriate use of methods in classifying text review data. Contrary to the low Dunn Index value generated in clustering (only 0.05), it indicates overlapping between clusters. Some clusters have similar contexts, but further analysis of review documents in each cluster shows that the clustering produces good results. Words within the clusters are identical, creating cohesive clusters despite the overlap. For future research, it may be beneficial to adjust the architecture of the Word2Vec model built to improve the clustering process.

Overall, user reviews of the ChatGPT application from Indonesia written in various language styles tend to have positive sentiments. Frequent reviews express satisfaction in using the application. This is supported by other reviews that appreciate the application's technology, benefits, and performance. Several criticisms and suggestions for application development were conveyed, as well as the hope that the accessible version of ChatGPT would still be available because of its extraordinary benefits. This shows the importance of maintaining and improving the app's quality so that negative reviews can be minimized in the future. Evaluation of incomplete application features and limited information can be used as a consideration for developing the ChatGPT application. Using sophisticated technology, conveying relevant information, fast conversational responses, and resembling human language can be used as a guide or standard for creating similar chatbot applications. Meanwhile,

criticism and evaluation of the ChatGPT application can be used as a consideration for developing similar chatbot applications.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering and Informatics - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] Binus University, "Sejarah Singkat Tentang Kecerdasan Buatan (Artificial Intelligence)," Binus University Graduate Program. Binus University Graduate Program, 2022.
[2] C. M. Annur, "Survei: ChatGPT Jadi Aplikasi AI Paling Banyak Digunakan di Indonesia," databoks. databoks, 2023.
[3] F. Duarte, "Number of ChatGPT Users," 2024.
[4] E. Gregersen, "ChatGPT Software," Britannica. Britannica, 2023.
[5] M. Baygin, "Classification of Text Documents based on Naive Bayes using N-Gram Features," 2018 Int. Conf. Artif. Intell. Data Process. IDAP 2018, pp. 1–5, 2019.
[6] A. Solikhatun and E. Sugiharti, "Application of the Naïve Bayes Classifier Algorithm using N- Gram and Information Gain to Improve the Accuracy of Restaurant Review Sentiment Analysis," J. Adv. Inf. Syst. Technol., vol. 2, no. 2, pp. 1–12, 2020.
[7] I. E. Tiffani, "Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review," J. Soft Comput. Explor., vol. 1, no. 1, pp. 1–7, 2020.
[8] A. Z. Farmadiansyah, "Deteksi Surel Spam dan Non Spam Bahasa Indonesia Menggunakan Metode Naïve Bayes," vol. 2, no. 2, pp. 1–5, 2021.
[9] E. Hasibuan and E. A. Heriyanto, "Analisis Sentimen Pada Ulasan Aplikasi Amazon Shopping Di Google Play Store Menggunakan Naive Bayes Classifier," J. Tek. dan Sci., vol. 1, no. 3, pp. 13–24, 2022.
[10] A. Khan, D. Majumdar, and B. Mondal, "Machine Learning Approach to Sentiment Analysis from Movie Reviews Using Word2Vec," Proc. Res. Appl. Artif. Intell., p. 532, 2020.
[11] M. R. Nashrulloh, I. T. Julianto, and R. K. Muzaky, "Opinion Mining on Chat GPT based on Twitter Users," J. Appl. Intell. Syst., vol. 8, no. 2, pp. 183–192, Jul. 2023.
[12] A. P. Wibawa, M. G. A. Purnama, M. F. Akbar, and F. A. Dwiyanto, "Metode-metode Klasifikasi," Pros. Semin. Ilmu Komput. dan Teknol. Inf., vol. 3, no. 1, p. 134, 2018.
[13] V. Vargas-Calderón and J. E. Camargo, "Characterization of citizens using word2vec and latent topic analysis in a large set of tweets," Cities, vol. 92, no. March, pp. 187–196, 2019.
[14] F. Azmi, K. Utama, O. T. Gurning, and S. Ndraha, "Initial centroid optimization of k-means algorithm using cosine similarity," J. Informatics Telecommun. Eng., vol. 3, no. 2, pp. 224–231, 2020.
[15] J. Santoso, E. I. Setiawan, E. M. Yuniarno, M. Hariadi, and M. H. Purnomo, "Hybrid conditional random fields and k-means for named entity recognition on indonesian news documents," Int. J. Intell. Eng. Syst., vol. 13, no. 3, pp. 233–245, 2020.
[16] M. M. Haider, M. A. Hossin, H. R. Mahi, and H. Arif, "Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm," IEEE Xplore, no. June, pp. 283–286, 2020.
[17] M. M. J. Adnan, M. L. Hemmje, and M. A. Kaufmann, "Social Media Mining to Study Social User Group by Visualizing Tweet Clusters Using Word2Vec, PCA and K-Means," CEUR Workshop Proc., vol. 2863, pp. 40–51, 2021.
[18] A. Sandhu, A. Edara, F. Wajid, and A. Agrawala, "Temporal Analysis on Topics Using Word2Vec," pp. 1–11, 2023.
[19] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," Inf. Sci. (Ny)., vol. 622, pp. 178–210, Apr. 2023.
[20] N. E. Aoumeur, Z. Li, and E. M. Alshari, "Improving the Polarity of Text through word2vec Embedding for Primary Classical Arabic Sentiment Analysis," Neural Process. Lett., vol. 55, no. 3, pp. 2249–2264, 2023.

[21] A. Zakaria and M. Siallagan, "Predicting Customer Satisfaction through Sentiment Analysis on Online Review," Int. J. Curr. Sci. Res. Rev., vol. 06, no. 01, pp. 515–522, 2023.
[22] S. Ballı and O. Karasoy, "Development of content-based SMS classification application by using Word2Vec-based feature extraction," IET Softw., vol. 13, no. 4, pp. 295–304, Aug. 2019.
[23] F. Zhang, "A Hybrid Structured Deep Neural Network with Word2Vec for Construction Accident Causes Classification," Int. J. Constr. Manag., vol. 22, no. 6, pp. 1120–1140, 2022.
[24] A. C. Mazari and A. Djeffal, "Sentiment Analysis of Algerian Dialect Using Machine Learning and Deep Learning with Word2vec," Inform., vol. 46, no. 6, pp. 67–78, 2022.
[25] R. Feldman and J. Sanger, The Text Mining Handbook. New York: Cambridge University Press, 2006.
[26] M. Anandarajan, C. Hill, and T. Nolan, Practical Text Analytics, vol. 2. Cham: Springer International Publishing, 2019.
[27] S. Pattanayak, Pro Deep Learning with TensorFlow. Berkeley, CA: Apress, 2017.
[28] S. K. Thompson, Sampling, vol. 755. John Wiley & Sons, 2012.
[29] F. Gorunescu, "Data Mining Techniques and Models," 2011, pp. 185–317.
[30] M. Kubat, An Introduction to Machine Learning. Cham: Springer International Publishing, 2017.
[31] S. H. Haji, K. Jacksi, and R. M. Salah, "Systematic Review for Selecting Methods of Document Clustering on Semantic Similarity of Online Laboratories Repository," 2022, pp. 239–252.
[32] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," Organ. Res. Methods, vol. 25, no. 1, pp. 114–146, 2022.
[33] C. K. Aridas, S. Karlos, V. G. Kanas, N. Fazakis, and S. B. Kotsiantis, "Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers under Imbalanced Data Sets," IEEE Access, vol. 8, pp. 2122–2133, 2020.
[34] A. Somasundaram and S. Reddy, "Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance," Neural Comput. Appl., vol. 31, no. S1, pp. 3–14, Jan. 2019.