Comparison of Sentiment Analysis Model for Shopee Comments on Google Play Store

Khuswatun Hasanah[1]*
Informatics Study Program, Faculty of Computer Science
Amikom University Yogyakarta
Yogyakarta, Indonesia
khuswatunhasanah@students.amikom.ac.id

***Abstract*— The current COVID-19 pandemic has greatly changed the order of consumption and the Indonesian economy. During the health crisis that hit Indonesia, the e-commerce sector experienced very rapid development because of changes in consumer behavior that are looking for safe and comfortable shopping alternatives. During the COVID-19 pandemic, Shopee became the number 1 online shopping site in Indonesia. However, this cannot be used as a standard for user satisfaction. User satisfaction can only be measured from comments by Shopee application users through the comments and rating features provided by the Google Play Store. Therefore, to be able to find out public opinion about Shopee, a sentiment analysis of the Shopee application will be carried out which can later be used by management to develop even better applications. In this study, the dataset taken is the rating and reviews of Shopee application users on the Google Play Store using the Multinomial Naïve Bayes method, Random Forest Classifier, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Extra Trees Classifier. This study uses 1000 comment and rating data which are processed using the Python language. The results of this study indicate that the method that has the highest level of accuracy is the Support Vector Machine algorithm with an accuracy of 88%, Extra Trees Classifier with an accuracy of 86%, Logistic Regression with an accuracy of 85%, Random Forest Classifier with an accuracy of 85%, K- Nearest Neighbors with an accuracy of 83%, and the last is Multinomial Naïve Bayes with an accuracy of 78%.**

***Keywords*— Shopee, Google Play Store, Sentiment Analysis**

I. INTRODUCTION

In 2020, Indonesian people live during a health crisis caused by the corona virus [1]. The Indonesian government, through the Ministry of Health, has implemented many methods to fight the spread of the corona virus [1]. Some of these methods are by implementing physical distancing, social distancing, isolation, lockdown, and large-scale social restrictions [1]. As a result of this regulation, all residents in the area are not allowed to leave the house, including carrying out buying and selling activities to meet their living needs [2]. Therefore, during the outbreak of the corona virus, many Indonesian people took advantage of e-commerce applications such as Shopee, Lazada, and Tokopedia which are available on the Google Play Store to support their buying and selling activities [3].

Shopee is the largest online shopping application in Indonesia which sells various kinds of products by offering good product quality, cheaper prices, lots of discounts, and of course free shipping throughout Indonesia [4]. This can be proven from the fourth quarter of 2019 to the fourth quarter of 2020. Shopee has always been ranked first in the 10th Indonesian e-commerce with the highest monthly visitors [5].

TABLE I. INDONESIAN E-COMMERCE MONTHLY VISITOR DATA (QUARTER II-2021)

No.	Name	Value/Visit
1.	Tokopedia	147.790.000
2.	Shopee	126.996.700
3.	Bukalapak	29.460.000
4.	Lazada	27.670.000
5.	Blibli	18.440.000
6.	Bhinneka	6.996.700
7.	Orami	6.260.000
8.	Ralali	5.123.300
9.	JD ID	3.763.300
10.	Zalora	3.366.700

In Table 1. Indonesian e-commerce monthly visitor data (quarter II-2021) taken from databox, in the second quarter - 2021 Shopee experienced a decline in visitors and was surpassed by Tokopedia with monthly visitors reaching 147,790,000[5]. Even though Tokopedia is in first place for the highest monthly visitors, on the Google Play Store Tokopedia is still inferior to Shopee, namely being in fourth place while Shopee is in first place in the Google Play Store [5]. However, Shopee's first ranking in the Google Play Store cannot be used as a standard for user satisfaction [6]. In this case, user satisfaction can only be seen from the comments and ratings given by Shopee application users on the Google Play Store [6]. Based on this problem, a sentiment analysis will be carried out to find out public opinion about the Shopee application on the Google Play Store [7]. Apart from that, this research will also compare the Multinomial Naïve Bayes algorithm, Random Forest Classifier, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Extra Trees Classifier to see which algorithm has the best level of accuracy [8]. By analyzing the sentiment of comments from users, Shopee will be able to understand user satisfaction more easily with the services and products provided [9]. This research will help Shopee identify what needs to be repaired and improve its performance to provide a better user experience [9].

To date, there has been a lot of research on sentiment

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

analysis, one of which is research conducted by Christian Cahyaningtyas, Yessica Nataliani, Indrastanti Ratna Widiasari [10]. This research was carried out using a SMOTE-based Decision Tree. They conducted sentiment analysis research by utilizing Shopee application rating data available on the Google Play Store. By using Decision Tree and SMOTE, the accuracy values obtained in this study reached 99.91%, AUC 0.999, recall 99.88%, and precision 99.98%. Meanwhile, for research using Decision Tree without using SMOTE, the accuracy value was 99.98%, AUC 0.950, recall 99.88% and precision 99.98% [10]. In another study conducted by Dany Pratmanto, Rousyati Rousyati, et al. Research related to sentiment analysis was carried out using the Naïve Bayes algorithm and the Knime application to measure data accuracy [11]. This research uses data from review comments on the Shopee application on the Google Play Store [11]. The data used in this research was 200 review data consisting of 100 negative review data and 100 positive review data [11]. By utilizing the Naïve Bayes algorithm and partition techniques, this research obtained an accuracy value of 96.667%, precision 100%, recall 93.33%, and AUC 1.00 [11]. Other research also discusses sentiment analysis using application comment data on the Google Play Store, for example research conducted by F. Bei and S. Saepudin [12]. This research utilized 1500 comment data from several online ticket applications and used Support Vector Machine for the classification process [12]. The results of this research show that the application that has the highest accuracy is Pegipegi with 78.21%, followed by the Agoda application with an accuracy of 77.00%, then there is Traveloka with an accuracy of 75.03%, then there is Mister Aladin with an accuracy 64.00%, and lastly there is tiket.com with the lowest accuracy, namely 58.68% [12]. In research conducted by U Kusnia and F. Kurniawan, this research utilized 5615 comment data from users of online news applications on the Google Play Store, as well as the SVM algorithm and Naïve Bayes algorithm for classification methods [13]. In the Naïve Bayes research, the accuracy value was 87%, then for the Support Vector Machine algorithm, the accuracy result was greater, namely 88% [13]. From this research it can be concluded that comments from online news media users tend to be positive, this can be seen because the total number of positive comments is 5160 data, while negative comments are only 455 data [13].

Even though there has been a lot of research related to sentiment analysis that compares several algorithms, there are still very few studies that compare six different algorithms in sentiment analysis [7]. Therefore, the aim of this research is to fulfill this research gap by using six different algorithms, namely Multinomial Naïve Bayes, Random Forest Classifier, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Extra Trees Classifier for sentiment analysis of Shopee comments. on Google Play Store [7]. By using various algorithms, this research aims to provide a more comprehensive comparison regarding the performance and accuracy of each algorithm for sentiment analysis of Shopee comments on the Google Play Store [7].

II. RESEARCH METHODS

This research was carried out using Google Collaboratory to assist in the data processing and classification process. The machine learning methods used in this research are Multinomial Naïve Bayes, Random Forest Classifier, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Extra Trees Classifier.

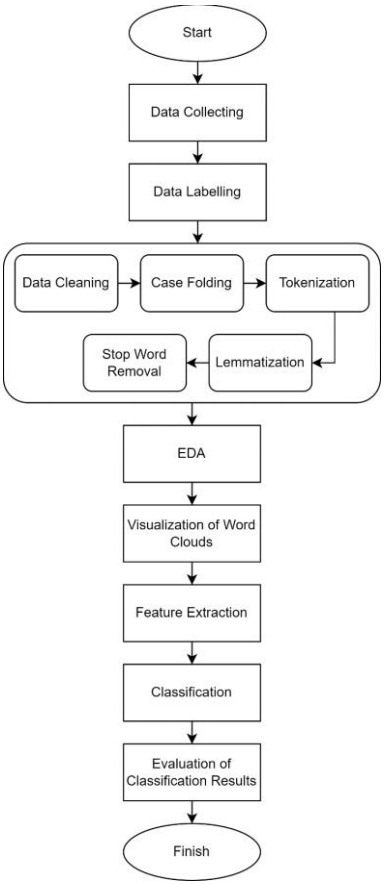

Figure 1. Research Flow

In Figure 1. Research Flow, this research method is divided into several processes, namely, Data Collection, Data Labeling, Data Preprocessing, EDA, Word Cloud Visualization, Feature Extraction, Classification using the Multinomial Naïve Bayes algorithm, Random Forest Classifier, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Extra Trees Classifier, and finally evaluation of classification results. Then the data preprocessing process is divided into several more parts, namely data cleaning, adding additional features such as text length and percentage of punctuation in the text, tokenization, lemmatization, and removing stop words.

A. Data Collection

The data used in this research are Shopee user comments and ratings on the Google Play Store website which were taken by Web Scraping [14]. The amount of data taken was 1000 data which were the most relevant comments and ratings with a time limit of 25 July 2023 at 00:18:38 to 02 August 2023 at 05:12:03[14]. The data scraping process is carried out by utilizing the google-play-scraper library in Python and using Google Collaboratory for the data retrieval process [14].

B. Data Labeling

Labeling is a process of calculating the polarity of

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

comments taken from Web Scrapping. The labeling method used in this research is that if the rating values given by Shopee application users are numbers 1, 2, and 3 then they will be labeled the number 0 which means negative, whereas if the ratings given by Shopee application users are numbers 4 and 5 then will be labeled with the number 1 which means positive [15].

C. Preprocessing

Preprocessing is the most important process in this research because the results of this preprocessing will greatly influence the performance results of the algorithm used [14]. In this research, preprocessing is used to process raw data, improve raw data, select raw data, and ensure that the raw data is ready to be used for the analysis process [14]. The preprocessing stages in this research are divided into 5 processes, namely:

1) *Data Cleaning*
   At this stage, the comment data resulting from the scraping process will be cleaned by replacing and deleting all characters other than letters (a-z and A-Z) such as emojis, numbers, punctuation marks and other special characters [16]. Apart from that, in the data cleaning process all uppercase letters will be changed to all lowercase letters [16].
2) *Added Features*
   This stage is used to see how often punctuation occurs in the review text and calculate the percentage [17].
3) *Tokenization*
   Tokenization is a stage for separating long sentences into words or better known as tokens [17].
4) *Lemmatization*
   Lemmatization is a stage for changing affixed or inflectional words to their basic form [18].
5) *Removing Stop words*
   Stop word removal is a stage of deleting words that are not useful, and are considered meaningless, non-standard words and have no effect on the analysis [17].

D. Exploratory Data Analysis

EDA (Exploratory Data Analysis) is a stage of analyzing data whose aim is to analyze and understand the characteristics of data visually and descriptively. By using EDA, this research data will be formed into a data frame and divide the data according to its group, for example rating 1 is entered in class 1, rating 2 is included in class, rating 3 is included in class 3, rating 4 is included in class 4, and rating 5 is included in class 5[19]. After this process, the next step is to see whether there are missing values in the data used, then visualize them using a bar chart [19].

E. Visualisasi Word Clouds

Word Clouds are an image consisting of a collection of words where the size of the word represents the appearance of the word, the more a word appears, the more often the word is mentioned in a text document [15]. In this study, the word clouds were divided into 2, namely, positive comment word clouds and negative comment word clouds [15].

F. Feature Extraction

Feature Extraction is a dimension reduction process that converts original data into a dataset with a smaller number of variables. In this research, the extraction process was carried out using a Vectorizer: TF-IDF. Term Frequency (TF) is a method used to see how often a word appears in a text document. The more often the word appears, the greater the TF value [17]. The TF formula itself can be seen in equation (1).

$$W(t,d) = TF(t,d) \quad (1)$$

Information :
TF(t,d) : Frequency of appearance of the word "t" in document "d".

IDF (Inverse Document Frequency) is a method that is inversely proportional to TF, if in TF words that appear frequently have a high value, then in IDF words that frequently appear in a text document will get a low IDF value while words that appear less a text document will get a large IDF value [17]. The IDF formula itself can be seen in equation (2).

$$IDF(t) = \log(N/df(t)) \quad (2)$$

Information :
N : Is the total number of documents.
df(t) : Is the document frequency of the word "t".

After completing calculating the TF and IDF values one by one, the next step is to multiply the calculated values to get the TF-IDF value for each word in the text document. In TF-IDF (Term Frequency - Inverse Document Frequency) words that often appear in certain text documents but rarely appear in other text documents will have a large TF-IDF value [17]. The TF-IDF formula itself can be seen in equation (3).

$$TF - IDF = TF(t,d) \; x \; IDF(t) \quad (3)$$

Information :
TF(t,d) : Frequency of appearance of the word "t" in document "d.
IDF(t) : IDF for a word "t".
TF – IDF : TF – IDF weight for a term "t" in a document "d".

G. Classification

After performing feature extraction, the next stage is carrying out the classification process. This research uses the Multinomial Naïve Bayes algorithm, Random Forest Classifier, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Extra Trees Classifier which were carried out separately.

1) *Multinomial Naïve Bayes*
   Multinomial Naïve Bayes is a type of Bayes Theorem classification method which assumes that all features or attributes used in classification are independent of each other. The Multinomial Naïve Bayes algorithm is very suitable for classification modeling because it has a very fast computing time and only requires a little memory. The formula for the Multinomial Naïve Bayes algorithm itself can be seen in equation (4) [20].

$$P(X|c) = log\frac{N_C}{N} + \sum_i^n =_1 log\frac{t_i+\alpha}{\sum_i^n =_1 t_i+\alpha} \quad (4)$$

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

Information :
$P(X|c)$: probability of feature X from class c
N_c : the total number of features from class c.
N : the total number of features.
t_i : bobot kata t.
$\sum_{t}^{n} =_1 t$: the total weight of words in class c.
α : smoothing parameter value.

2) Random Forest Classifier

Random Forest Classifier is a machine learning algorithm that is part of the ensemble algorithm group [21]. The Random Forest Classifier algorithm works by creating many decision trees and collecting prediction results from these trees to achieve accurate results [21]. For the classification process, the Random Forest Classifier algorithm can handle imbalance data, reduce overfitting and is able to provide good accuracy values.

3) Logistic Regression

Logistic Regression is a machine learning algorithm that can be used for the sentiment analysis process. In sentiment analysis, the Logistic Regression algorithm works by using statistical methods to predict the probability of positive and negative sentiment from text based on input variables [22]. As a machine learning algorithm, Logistic Regression can be used to classify data into 2 groups, namely negative and positive sentiment groups.

4) Support Vector Machine (SVM)

The SVM algorithm is a machine learning algorithm that is usually used for classification and regression. In the classification process, the SVM algorithm can produce good performance even with small datasets. Apart from that, SVM also could overcome overfitting problems by using regularization techniques. The SVM algorithm used in this research is a non-linear SVM with a kernel. Non-linear SVMs use kernels to map data into higher dimensions, allowing better separation between classes. The kernel used in classification modeling using SVM is a linear kernel. The prediction function in a non-linear SVM with a kernel can be written in equation (5) as follows:

$$y(x) = \Sigma(\alpha\_i*y\_i*K(x\_i,x)) + b \quad (5)$$

Information :
y(x) : prediction function
α_i : lagrange coefficient
y_i : class label
K : kernel function
x_i,x : input feature vector
b : biased
The linear kernel formula can be seen in equation (6).

$$K(x\_i,x) = x\_i\text{^}T*x \quad (6)$$

5) K-Nearest Neighbor

K-Nearest Neighbor is a machine learning algorithm that is non-parametric and lazy learning. Non-parametric means the K-Nearest Neighbor algorithm will not make any assumptions about the underlying distribution. Meanwhile, lazy learning means that K-Nearest Neighbor does not use training data points to create a model. In classification, the K-Nearest Neighbor algorithm works by determining the value of K, namely the closest neighbors that will be used to carry out classification. Then the K-Nearest Neighbor algorithm will calculate the distance between new and old data in the training set using a certain distance metric. After that K-Nearest Neighbor will select the K closest neighbors with the shortest distance from the new data. Next, the K-Nearest Neighbor algorithm will carry out majority voting based on the class labels of the K closest neighbors that have been selected. And finally, K-Nearest Neighbor will classify new data based on the majority labels obtained [22].

6) Extra Trees Classifier

Extra Trees Classifier is a machine learning algorithm that is part of the ensemble algorithm group of Random Forest algorithms [23]. The way the Extra Trees Classifier algorithm works is by creating many decision trees with random samples and features, which then carry out a majority vote to get the final prediction [23].

H. Evaluation of Classification Results

The final process of this research is to look at the performance of the algorithm used in the classification process. The purpose of this evaluation is to see how well the algorithm used in predicting the dataset performs [17]. The technique used to evaluate the algorithm is Confusion Matrix. Confusion Matrix is a table that is often used to measure the performance of classification in machine learning which can display and compare actual data with model predicted values [17].

TABLE II. CONFUSION MATRIX

Prediction	Ground Truth	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Tabel II. It is a confusion matrix table with the following information:

TP (True Positive) : This is the number whose prediction is positive, and the ground truth is also positive.
TN (True Negative) : This is the number where the prediction is negative, and the ground truth is also negative.
FP (False Positive) : This is the number where the prediction is positive, but the ground truth is negative.
FN (False Negative) : This is the number whose prediction is negative, but the ground truth is positive.

With the TP, TN, FP and FN values, researchers can calculate many model performance evaluation matrices such as accuracy, precision, recall and F1-Score. The accuracy formula can be seen in equation (7), precision in equation (8), recall in equation (9), and F1-Score in equation (10).

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

$$Precission = \frac{TP}{TP + FN} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1 - score = 2\,x\,\frac{recall.precission}{recall + precission} \quad (10)$$

III. RESULT AND DISCUSSION

This chapter will describe one by one the results of the research that has been carried out starting from Data Collection, Data Preprocessing, Data Labeling, EDA, Visualization of Word Clouds, Feature Extraction, Classification, and Evaluation of Classification Results.

A. Data Collection

By using Web Scraping techniques, this research succeeded in obtaining 1000 rating data and comments from Shopee application users on the Google Play Store Website.

TABLE III. WEB SCRAPING RESULTS DATA

Comments	Ratings
Untuk opsi Live Bagian bawah Berubah menjadi agak ribet, lebih baik versi sebelumnya.	5
Tolong dong spaylater sy diaktifkan kembali,,pembayaran cicilan sudah lunas,,apalagi spinjam ndak pernah bisa diaktifkan dari awal buka akun shopee,,kalau diaktifkan ke2nya sy full kan bintangnya.	2
Spaylater saya tiba tiba di batasi, padahal pemakaian wajar dan tidak bertindak kecurangan, aplikasinya juga lemot bgt dan nambah berat.. Padahal spek hp saya juga memadai.. Pathetic	1
Banyak fitur yang sebenernya gak perlu-perlu banget buat ada di shopee yang terkadang bikin aplikasinya jadi lambat dan gampang ngelag.	3
Kemarin kemarin games shoppie capit yg agak susah di buka...nah sekarang setelah di update...shoppie bubble nya pas di buka cuma layar kuning muda aja yg nongol...😔	2
Sudah mantap semuanya ... Usul poin penilaiannya ditingkatkan ke ratusan ya .... Kalo cuma puluhan kelamaan dan susah ..... Lebih ditingkatkan lagi kemampuan kecepatan aksesnya ya SHOPEE, biar sangat lancar mengaksesnya.	5

Table III. is the result of the data collection process carried out using web scraping. From this data collection, comment and rating data were obtained from the Shopee application.

TABLE IV. DATA FILTERING RESULTS

Comments	Ratings
Untuk opsi Live Bagian bawah Berubah menjadi agak ribet, lebih baik versi sebelumnya.	5
Tolong dong spaylater sy diaktifkan kembali,,pembayaran cicilan sudah lunas,,apalagi spinjam ndak pernah bisa diaktifkan dari awal buka akun shopee,,kalau diaktifkan ke2nya sy full kan bintangnya.	2
Spaylater saya tiba tiba di batasi, padahal pemakaian wajar dan tidak bertindak kecurangan, aplikasinya juga lemot bgt dan nambah berat.. Padahal spek hp saya juga memadai.. Pathetic	1
Setelah update lapak yang ada tampilan video nya tidak bisa di tampilkan. Hanya yang foto dan live yg bisa tampil. Sudah clear cache, log out dan login. Bahkan install ulang apk shopee ...masih aja lapak yg ada videonya gak bisa di klik	3
Kemarin kemarin games shoppie capit yg agak susah di buka...nah sekarang setelah di update...shoppie bubble nya pas di buka cuma layar kuning muda aja yg nongol...😔	2
Tadi nya saya kasing rating tinggi, karena gara gara sekarang nih, klo pembayaran harus verifikasi wajah terus, ujung ujung nya malah gagal ga bisa membayar, padahal saldo ada, tapi pembayaran malah ga bisa, jelek sekali shopee sekarang,..!!!!!!!?	1
Mantap.. racun nya banyak.. bawaannya pengen jajan mulu.. diskon dan promo nya melimpah.. Spaylater nya bikin nagih belanja.. SPinjam nya juga memudahkan kalo lagi kepepet butuh uang.. semakin mantap shopee	3

As seen in Table IV. This research only uses 2 data, namely comment and rating data from Shopee users on the Google Play Store website. Because after carrying out the Web Scraping process, some of the data that has been obtained will be filtered and deleted so that only comment and rating data will be left.

B. Data Labeling

After carrying out the scraping process, the data that has been collected goes to the data labeling stage to determine whether the comments fall into the negative or positive sentiment class.

TABLE V. LABELING RESULTS

Comments	Ratings	Labels
Untuk opsi Live Bagian bawah Berubah menjadi agak ribet, lebih baik versi sebelumnya.	5	1
Tolong dong spaylater sy diaktifkan kembali,,pembayaran cicilan sudah lunas,,apalagi spinjam ndak pernah bisa diaktifkan dari awal buka akun shopee,,kalau diaktifkan ke2nya sy full kan bintangnya.	2	0
Spaylater saya tiba tiba di batasi, padahal pemakaian wajar dan tidak bertindak kecurangan, aplikasinya juga lemot bgt dan nambah berat.. Padahal spek hp saya juga memadai.. Pathetic	1	0
Banyak fitur yang sebenernya gak perlu-perlu banget buat ada di shopee yang terkadang bikin aplikasinya jadi lambat dan gampang ngelag.	3	0
Kemarin kemarin games shoppie capit yg agak susah di buka...nah sekarang setelah di update...shoppie bubble nya pas di buka cuma layar kuning muda aja yg nongol...😔	2	0
Sudah mantap semuanya ... Usul poin penilaiannya ditingkatkan ke ratusan ya .... Kalo cuma puluhan kelamaan dan susah ..... Lebih ditingkatkan lagi kemampuan kecepatan aksesnya ya SHOPEE, biar sangat lancar mengaksesnya.	5	1

In Table V in this study ratings 1, 2, and 3 received a value of 0, which means a negative label, and ratings 4 and 5 received a value of 1, which means a positive label.

C. Preprocessing Data

As explained in the research method, the preprocessing process of this research consists of several stages, namely data cleaning, adding features, tokenization, lemmatization, and removing stop words.

TABLE VI. DATA CLEANING RESULTS

Comments Before Data Cleaning	Comments After Data Cleaning
Untuk opsi Live Bagian bawah	untuk opsi live bagian bawah

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

Berubah menjadi agak ribet, lebih baik versi sebelumnya.	berubah menjadi agak ribet lebih baik versi sebelumnya
Mantap.. racun nya banyak.. bawaannya pengen jajan mulu.. diskon dan promo nya melimpah.. Spaylater nya bikin nagih belanja.. SPinjam nya juga memudahkan kalo lagi kepepet butuh uang.. semakin mantap shopee	mantap racun nya banyak bawaannya pengen jajan mulu diskon dan promo nya melimpah spaylater nya bikin nagih belanja spinjam nya juga memudahkan kalo lagi kepepet butuh uang semakin mantap shopee
Kemarin kemarin games shoppie capit yg agak susah di buka...nah sekarang setelah di update...shoppie bubble nya pas di buka cuma layar kuning muda aja yg nongol...😔	kemarin kemarin games shoppie capit yg agak susah di buka nah sekarang setelah di update shoppie bubble nya pas di buka cuma layar kuning muda aja yg nongol
Sudah mantap semuanya ... Usul poin penilaiannya ditingkatkan ke ratusan ya .... Kalo cuma puluhan kelamaan dan susah ..... Lebih ditingkatkan lagi kemampuan kecepatan aksesnya ya SHOPEE, biar sangat lancar mengaksesnya.	sudah mantap semuanya usul poin penilaiannya ditingkatkan ke ratusan ya kalo cuma puluhan kelamaan dan susah lebih ditingkatkan lagi kemampuan kecepatan aksesnya ya shopee biar sangat lancar mengaksesnya

Table VI is the result of the data cleaning process. You can see in the table that the data has gone through the cleaning text stage, all capital letters have been changed to lower case and all characters other than letters (a – z and A – Z) have been deleted and replaced with spaces.

TABLE VII. RESULTS OF ADDITIONAL FEATURES

Comments	Content_len	Punct
Untuk opsi Live Bagian bawah Berubah menjadi agak ribet, lebih baik versi sebelumnya.	73	2.7
Spaylater saya tiba tiba di batasi, padahal pemakaian wajar dan tidak bertindak kecurangan, aplikasinya juga lemot bgt dan nambah berat.. Padahal spek hp saya juga memadai.. Pathetic	156	3.8
Senang belanja di shopee....barang-barang yang di tawarkan sesuai dengan gambar real pict...pelayanan kiriman juga sering dapat bebas biaya ongkir jdi bantu banget buat keuangan yang mepet... Ha..ha..ha..	178	9.6
Kemarin kemarin games shoppie capit yg agak susah di buka...nah sekarang setelah di update...shoppie bubble nya pas di buka cuma layar kuning muda aja yg nongol...😔	139	6.5
shopee simpan pinjam tidak pernah telat dan paylater hanya telat sekali sudah tidak di gunakan lagi.... jadi tidak ada artinya selama bertahun2 saya tepat waktu... kalah sama satu kali tunggakan....	169	6.5

Table VII is the result of adding features. There are 2 features that have been successfully added, the first is the content_len feature whose function is to count the number of letters and characters in text or sentences, and the last is the Punct feature whose function is to calculate the percentage of punctuation in each line of comment text.

TABLE VIII. TOKENIZATION RESULTS

Comments	Comments After Tokenization
Untuk opsi Live Bagian bawah Berubah menjadi agak ribet, lebih baik versi sebelumnya.	untuk,opsi,live,bagian,bawah, berubah,menjadi,agak,ribet, lebih,baik, versi,sebelumnya
Tolong dong spaylater sy diaktifkan kembali,,pembayaran cicilan sudah lunas,,apalagi spinjam ndak pernah bisa	tolong,dong,spaylater,sy,diaktifkan, kembali,pembayaran,cicilan,sudah, lunas,apalagi,spinjam,ndak,pernah, bisa,diaktifkan,dari,awal,buka,akun,
diaktifkan dari awal buka akun shopee,,kalau diaktifkan ke2nya sy full kan bintangnya.	shopee,kalau,diaktifkan,ke,nya,sy, full,kan,bintangnya
Spaylater saya tiba tiba di batasi, padahal pemakaian wajar dan tidak bertindak kecurangan, aplikasinya juga lemot bgt dan nambah berat.. Padahal spek hp saya juga memadai.. Pathetic	spaylater,saya,tiba,tiba,di,batasi, padahal,pemakaian,wajar,dan,tidak, bertindak,kecurangan,aplikasinya, juga,lemot,bgt,dan,nambah,berat, padahal,spek,hp,saya,juga, memadai,pathetic
Mantap.. racun nya banyak.. bawaannya pengen jajan mulu.. diskon dan promo nya melimpah.. Spaylater nya bikin nagih belanja.. SPinjam nya juga memudahkan kalo lagi kepepet butuh uang.. semakin mantap shopee	mantap,racun,nya,banyak, bawaannya,pengen,jajan,mulu, diskon,dan,promo,nya,melimpah, spaylater,nya,bikin,nagih,belanja, spinjam,nya,juga,memudahkan, kalo,lagi,kepepet,butuh,uang, semakin,mantap shopee

Table VIII is the result of the tokenization process. At the tokenization stage, this research succeeded in breaking long sentences into words or what are often called tokens. The token is in the form of a word, and a comma (,).

TABLE IX. RESULTS OF LEMMATIZATION AND REMOVING STOP WORDS

Comments	After Lemmatized & Stop Word
Sip bagus mudah. Cuma kadang kacau. Suka berkedip. Dan saat telusuri agak jauh scrol kebawah langsung kembali ke awal dan hilangkan filter nya Sip bagus mudah. Cuma kadang kacau. Suka berkedip. Dan saat telusuri agak jauh scrol kebawah langsung kembali ke awal dan hilangkan filter nya	sip bagus mudah kadang kacau suka berkedip telusuri scrol kebawah langsung hilangkan filter nya
Kalau buka gambar produk lebih besar setelah 3× gak bisa dibuka, loading teruus. Harus keluar aplikasi & masuk lagi hasilnya sama. Padahal internetnya kencang banget.	buka gambar produk gak dibuka loading teruus aplikasi masuk hasilnya internetnya kencang banget
Kemarin kemarin games shoppie capit yg agak susah di buka...nah sekarang setelah di update...shoppie bubble nya pas di buka cuma layar kuning muda aja yg nongol...😔	kemarin kemarin game shoppie capit yg susah buka update shoppie bubble nya pa buka layar kuning muda aja yg nongol
Senang belanja di shopee....barang-barang yang di tawarkan sesuai dengan gambar real pict...pelayanan kiriman juga sering dapat bebas biaya ongkir jdi bantu banget buat keuangan yang mepet... Ha..ha..ha..	senang belanja shopee barang barang tawarkan sesuai gambar real pict pelayanan kiriman bebas biaya ongkir jdi bantu banget keuangan mepet ha ha ha
Banyak fitur yang sebenernya gak perlu-perlu banget buat ada di shopee yang terkadang bikin aplikasinya jadi lambat dan gampang ngelag.	fitur sebenernya gak banget shopee terkadang bikin aplikasinya lambat gampang ngelag
Shopee skarang sudah seperti aplikasi sampah yang sering error, ntah di saat proses pembayaran, pencarian barang, dan masih banyak lagi...mendingan di tingkatkan lagi daripada merugikan yang instal apk ini	shopee skarang aplikasi sampah error ntah proses pembayaran pencarian barang mendingan tingkatkan merugikan instal apk

Table IX is the result of the lemmatization and stop word removal process. At this stage, this research succeeded in removing words that often appear, but these words have no

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

meaning in providing understanding of a text. Some examples of words that were deleted were (and, I, but, already, like, which, often, when, rather than, many, really, make, make, so, this, etc.).

D. Exploratory Data Analysis (EDA)

At this stage, this research succeeded in creating research data in the form of a data frame and dividing the data according to its groups, namely rating 1 was entered in class 1, rating 2 was entered in class 2, rating 3 was entered in class 3, rating 4 was entered in class 4, and a rating of 5 is entered in class 5.

```
Input data has 1000 rows and 8 columns
rating 1.0 = 466 rows
rating 2.0 = 132 rows
rating 3.0 = 118 rows
rating 4.0 = 79 rows
rating 5.0 = 205 rows
```

Figure 2. Class Details Data

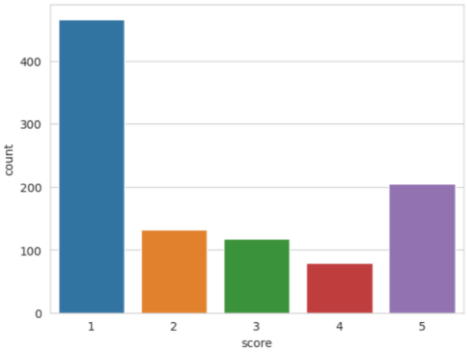

Figure 3. Class Details Data Visualization

Figure 2 and Figure 3 provide information that this research has input data of 1000 rows, 8 columns and the rating with the highest value is rating 1 with a value of 466 rows, followed by rating 5 with a value of 205 rows, rating 2 with a value of 132 lines, next there is rating 3 with a value of 118 lines, and finally there is rating 4 with the smallest value, namely 79 lines.

E. Visualisasi Word Clouds

Figure 4. Word Clouds Positive Comments

Figure 5. Word Clouds Negative Comments

As explained in Research Methods, this research has succeeded in creating 2 visualizations of word clouds, namely word clouds of positive comments and word clouds of negative comments. From Figure 4 and Figure 5 the words that often appear in the word clouds of positive and negative comments are the word Shopee.

F. Classification and Evaluation of Classification Results

Before entering the classification process, the research data will be divided into 2 parts, namely training data and testing data with a division proportion of 80:20, which means 80% of the data will be used for model training, and 20% will be used for testing data. By dividing the data, the following classification results are obtained:

	precision	recall	f1-score	support
0	0.78	1.00	0.87	153
1	1.00	0.06	0.12	47
accuracy			0.78	200
macro avg	0.89	0.53	0.50	200
weighted avg	0.83	0.78	0.70	200

Figure 6. Classification of Multinomial Naïve Bayes Algorithms

From Figure 6, it can be seen that the classification using the Multinomial Naïve Bayes algorithm obtained an accuracy of 78%, with a precision value for negative comments of 78%, recall of negative comments of 100%, and f1-score of negative comments of 87%. Meanwhile, positive comments get a precision value of 100%, recall 6%, f1-score 12%. With the following evaluation results:

TABLE X. EVALUATION OF MULTINOMIAL NAÏVE BAYES

Prediction	Ground Truth	
	Positive	Negative
Positive	153	0
Negative	44	3

From Table X, the evaluation of the Multinomial Naïve Bayes algorithm using a confusion matrix obtained a value of True Positive 153, False Negative 0, False Positive 44 and True Negative 3.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

	precision	recall	f1-score	support
0	0.85	0.98	0.91	153
1	0.87	0.43	0.57	47
accuracy			0.85	200
macro avg	0.86	0.70	0.74	200
weighted avg	0.85	0.85	0.83	200

Figure 7. Random Forest Classifier Algorithm Classification

From Figure 7, the classification using the Random Forest Classifier algorithm obtained an accuracy of 85%, with a precision value for negative comments of 85%, recall of negative comments of 98%, and f1-score of negative comments of 91%. Meanwhile, for positive comments, the precision value was 87%, recall 43%, f1-score 57%. With the following evaluation results:

TABLE XI. EVALUATION OF RANDOM FOREST CLASSIFIER

Prediction	Ground Truth	
	Positive	Negative
Positive	150	3
Negative	27	20

From table XI, It can be seen that the evaluation of the Random Forest Classifier algorithm using a confusion matrix obtained a value of True Positive 150, False Negative 3, False Positive 27 and True Negative 20.

	precision	recall	f1-score	support
0	0.84	1.00	0.91	153
1	1.00	0.36	0.53	47
accuracy			0.85	200
macro avg	0.92	0.68	0.72	200
weighted avg	0.87	0.85	0.82	200

Figure 8. Classification of Logistic Regression Algorithms

From Figure 8, the classification using the Logistic Regression algorithm obtained an accuracy of 85%, with a precision value for negative comments of 84%, recall of negative comments of 100%, and f1-score of negative comments of 91%. Meanwhile, positive comments obtained a precision value of 100%, recall 36%, f1-score 53%. With the following evaluation results:

TABLE XII. EVALUATION OF LOGISTIC REGRESSION

Prediction	Ground Truth	
	Positive	Negative
Positive	153	0
Negative	30	17

From table XII, the evaluation of the Logistic Regression algorithm using the confusion matrix obtained a value of True Positive 153, False Negative 0, False Positive 30 and True Negative 17.

	precision	recall	f1-score	support
0	0.86	1.00	0.93	153
1	1.00	0.49	0.66	47
accuracy			0.88	200
macro avg	0.93	0.74	0.79	200
weighted avg	0.90	0.88	0.86	200

Figure 9. Support Vector Machine Algorithm Classification

From Figure 9, the classification using the Support Vector Machine algorithm obtained an accuracy result of 88%, with a precision value for negative comments of 86%, recall of negative comments of 100%, and f1-score of negative comments of 93%. Meanwhile, positive comments obtained a precision value of 100%, recall 49%, f1-score 66%. With the following evaluation results:

TABLE XIII. EVALUATION OF SUPPORT VECTOR MACHINE

Prediction	Ground Truth	
	Positive	Negative
Positive	153	0
Negative	24	23

From table XIII, the evaluation of the Support Vector Machine algorithm using a confusion matrix obtained a value of True Positive 153, False Negative 0, False Positive 24, and True Negative 23.

	precision	recall	f1-score	support
0	0.86	0.93	0.89	153
1	0.70	0.49	0.57	47
accuracy			0.83	200
macro avg	0.78	0.71	0.73	200
weighted avg	0.82	0.83	0.82	200

Figure 10. K-Nearest Neighbors Algorithm Classification

From Figure 10, classification using the K-Nearest Neighbors (K-NN) algorithm gets an accuracy result of 83%, with a precision value for negative comments of 86%, recall of negative comments of 93%, and f1-score of negative comments of 89%. Meanwhile, positive comments obtained a precision value of 70%, recall 49%, f1-score 57%. With the following evaluation results:

TABLE XIV. EVALUATION OF *K*-NEAREST NEIGHBORS

Prediction	Ground Truth	
	Positive	Negative
Positive	143	10
Negative	24	23

From table XIV, the evaluation of the K-Nearest Neighbors algorithm using the confusion matrix obtained a value of True Positive 143, False Negative 10, False Positive 24, and True Negative 23.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

	precision	recall	f1-score	support
0	0.85	0.99	0.92	153
1	0.91	0.45	0.60	47
accuracy			0.86	200
macro avg	0.88	0.72	0.76	200
weighted avg	0.87	0.86	0.84	200

Figure 11. Extra Trees Classifier Algorithm Classification

From Figure 11, the classification using the Extra Trees Classifier algorithm obtained an accuracy result of 86%, with a precision value for negative comments of 85%, recall of negative comments of 99%, and f1-score of negative comments of 92%. Meanwhile, for positive comments, the precision value was 91%, recall 45%, f1-score 60%. With the following evaluation results:

TABLE XV. EVALUATION OF EXTRA TREES CLASSIFIER

Prediction	Ground Truth	
	Positive	Negative
Positive	151	2
Negative	26	21

From table XV, the evaluation of the K-Nearest Neighbors algorithm using the confusion matrix obtained a value of True Positive 151, False Negative 2, False Positive 26, and True Negative 21.

IV. CONCLUSION

This research has identified a gap in the literature regarding sentiment analysis, where there is still very little research that compares using six different algorithms in the context of sentiment analysis of Shopee comments on the Google Play Store. Previous research tends to only make comparisons using two or three algorithms. An example is the research conducted by Muhammad Fadli Asshiddiqi and Kemas Muslim Lhaksmana which only carried out a comparison with two algorithms, namely the Support Vector Machine and Decision Tree algorithms [17]. Then in another research conducted by Ulfa Kusnia and Fachrul Kurniawan, they also only used the Support Vector Machine and Naïve Bayes algorithms as comparison material in their research [13]. Therefore, to fill this gap, this research was carried out a comparison using six different algorithms, namely Multinomial Naïve Bayes, Random Forest Classifier, Logistic Regression, Support Vector Machine K-Nearest Neighbors, and Extra Trees Classifier. From the research that has been carried out, it shows that the Support Vector Machine algorithm has the highest accuracy, namely 88%, in second place is the Extra Trees Classifier algorithm with an accuracy of 86%, then there is Logistic Regression with an accuracy of 85%, Random Forest Classifier with an accuracy of 85%, K-Nearest Neighbors with an accuracy of 83%, and finally Multinomial Naïve Bayes with an accuracy of 78%.

REFERENCES

[1] N. R. Yunus and A. Rezki, "Kebijakan pemberlakuan lock down sebagai antisipasi penyebaran corona virus Covid-19," *Salam: Jurnal Sosial dan Budaya Syar-i*, vol. 7, no. 3, pp. 227–238, 2020.

[2] A. Putri, A. Pebriani, M. J. Rumi, and J. H. Siregar, "Pemaanfaatan Aplikasi Toko Online Terhadap Kebutuhan Konsumen Selama Pandemi Covid-19," in *Prosiding Seminar Nasional Pengabdian Masyarakat LPPM UMJ*, 2021.

[3] A. D. Cahya, F. A. Aqdella, A. Z. Jannah, and H. Setyawati, "Memanfaatkan marketplace sebagai media promosi untuk meningkatkan penjualan di tengah pandemi Covid-19," *Scientific Journal Of Reflection: Economic, Accounting, Management and Business*, vol. 4, no. 3, pp. 503–510, 2021.

[4] D. Chong and H. Ali, "Literature Review: Competitive Strategy, Competitive Advantages, and Marketing Performance on E-Commerce Shopee Indonesia," *Dinasti International Journal of Digital Business Management*, vol. 3, no. 2, pp. 299–309, 2022.

[5] R. U. Erza, A. M. Ramdan, and N. Norisanti, "Analisis Online Customer Review Dan Seller Reputation Terhadap Keputusan Belanja Online Dimasa Pandemi Covid-19," *Management Studies and Entrepreneurship Journal (MSEJ)*, vol. 3, no. 3, pp. 1629–1634, 2022.

[6] U. W. Saputra, "The role of user experience towards customer loyalty with mediating role of customer satisfaction at Shopee," *REVIEW OF MANAGEMENT, ACCOUNTING, AND BUSINESS STUDIES*, vol. 2, no. 2, pp. 104–113, 2021.

[7] S. Saepudin, S. Widiastuti, and C. Irawan, "Sentiment Analysis of Social Media Platform Reviews Using the Naïve Bayes Classifier Algorithm," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 2, pp. 236–243, 2023.

[8] R. Ardianto, T. Rivanie, Y. Alkhalifi, F. S. Nugraha, and W. Gata, "Sentiment analysis on E-sports for education curriculum using naive Bayes and support vector machine," *Jurnal Ilmu Komputer dan Informasi*, vol. 13, no. 2, pp. 109–122, 2020.

[9] L. O. Sihombing, H. Hannie, and B. A. Dermawan, "Sentimen Analisis Customer Review Produk Shopee Indonesia Menggunakan Algortima Naïve Bayes Classifier," *Edumatic: Jurnal Pendidikan Informatika*, vol. 5, no. 2, pp. 233–242, 2021.

[10] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiasari, "Analisis sentimen pada rating aplikasi Shopee menggunakan metode Decision Tree berbasis SMOTE," *AITI*, vol. 18, no. 2, pp. 173–184, 2021.

[11] D. Pratmanto, R. Rousyati, F. F. Wati, A. E. Widodo, S. Suleman, and R. Wijianto, "App Review Sentiment Analysis Shopee Application in Google Play Store Using Naive Bayes Algorithm," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1742-6596/1641/1/012043.

[12] F. Bei and S. Saepudin, "Analisis Sentimen Aplikasi Tiket Online Di Play Store Menggunakan Metode

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024

Support Vector Machine (Svm)," in *Seminar Nasional Sistem Informasi dan Manajemen Informatika Universitas Nusa Putra*, 2021, pp. 91–97.

[13] U. Kusnia and F. Kurniawan, "Analisis Sentimen Review Aplikasi Media Berita Online Pada Google Play menggunakan Metode Algoritma Support Vector Machines (SVM) Dan Naive Bayes," *Explore IT!: Jurnal Keilmuan dan Aplikasi Teknik Informatika*, vol. 14, no. 1, pp. 24–28, 2022.

[14] F. F. Irfani, M. Triyanto, and A. D. Hartanto, "Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine," *JBMI (Jurnal Bisnis, Manajemen, dan Inform., vol. 16, no. 3, p. 258, 2020, doi: 10.26487/jbmi. v16i3. 8607*, 2020.

[15] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 9, no. 2, pp. 785–795, 2022.

[16] P. Aditiya, U. Enri, and I. Maulana, "Analisis Sentimen Ulasan Pengguna Aplikasi Myim3 Pada Situs Google Play Menggunakan Support Vector Machine," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 4, pp. 1020–1028, 2022.

[17] M. F. Asshiddiqi and K. M. Lhaksmana, "Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI," *eProceedings of Engineering*, vol. 7, no. 3, 2020.

[18] R. Y. L. Lesmana and R. Andarsyah, "Model Klasifikasi Multinomial Naive Bayes Untuk Analisis Sentiment Terkait Non-Fungible Token," *Jurnal Teknik Informatika*, vol. 14, no. 3, pp. 135–139, 2022.

[19] D. A. Agustina and F. Rahmah, "Analisis Sentimen pada Sosial Media Twitter terhadap MRT Jakarta Menggunakan Machine Learning," *Insearch: Information System Research Journal*, vol. 2, no. 01, pp. 1–6, 2022.

[20] N. L. P. M. Putu and A. Z. Amrullah, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 123–131, 2021.

[21] S. Budiman, A. Sunyoto, and A. Nasiri, "Analisa Performa Penggunaan Feature Selection untuk Mendeteksi Intrusion Detection Systems dengan Algoritma Random Forest Classifier," *SISTEMASI: Jurnal Sistem Informasi*, vol. 10, no. 3, pp. 754–760, 2021.

[22] F. Fazrin, O. N. Pratiwi, and R. Andreswari, "Perbandingan Algoritma K-Nearest Neighbor dan Logistic Regression pada Analisis Sentimen terhadap Vaksinasi Covid-19 pada Media Sosial Twitter dengan Pelabelan Vader dan Textblob," *eProceedings of Engineering*, vol. 10, no. 2, 2023.

[23] S. Khomsah and A. S. Aribowo, "Text-preprocessing model youtube comments in indonesian," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 4, no. 4, pp. 648–654, 2020.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i1.1916, Copyright ©2024
Submitted : August 11, 2023, Revised : January 16, 2024 , Accepted : January 18, 2024 , Published : February 15, 2024