NEWTON: Networking and Information Technology
Vol. 4 No. 3 February 2025, Page. 124 - 131
E-ISSN: 2797-0728

Sentiment Analysis of Online Game Clash of Clans Reviews Using the K-Nearest Neighbor Method

Achmad Agus Athok Miftachuddin[1], Andika Prastyo[2]
[1]Faculty of Information Technology, Universitas KH. A. Wahab Hasbullah
[2]Faculty of Information Technology, Universitas KH. A. Wahab Hasbullah
*Email: agusathok@unwaha.ac.id, andikaprstyo20@gmail.com

ABSTRACT

Clash of Clans is a popular strategy game with millions of players worldwide. User reviews for this game are available on various online platforms. Sentiment analysis of these reviews can provide valuable insights into players' experiences and opinions. In this study, the researchers used the K-Nearest Neighbor (KNN) algorithm to classify the sentiment of Clash of Clans player reviews collected from the Google Play Store. Experimental results show that with a 60:40 training and testing data split, the KNN model was able to classify review sentiment with an accuracy of 64.52%, a precision value of 68.4%, a recall value of 88%, and an F1-score of 76.97%. The application of TF-IDF word weighting produced high accuracy at k-2 with an accuracy of 95.55%, precision of 96.16%, recall of 95.55%, and F1-score of 95.59%. These results indicate that KNN can be an efficient tool for analyzing player sentiment towards the Clash of Clans game.

Keywords*: Sentiment Analysis; Clash Of Clans; K-Nearest Neighbor.*

INTRODUCTION

The world of gaming technology has become increasingly advanced globally, with one of the most well-known games being Clash of Clans, or COC for short. Clash of Clans is a mobile strategy game developed and published by the Finnish gaming company, Supercell. This game has become widely discussed and played by online game enthusiasts, from children to adults who enjoy playing it. The lightweight nature of the game, played on smartphones, has made it popular among online game lovers. Clash of Clans is one of the 3D games for smartphones, available on platforms like Android and iOS (AlHafiidh & Oktadini, 2023).

As the number of players grows, app distribution platforms like Google Play have become a place where users can share their experiences through reviews. According to data from Google Play on March 4, 2023, there were over 60.6 million user reviews for Clash of Clans. These reviews encompass various user experiences, complaints, and opinions about the game. Due to the large number of reviews, including both complaints and praise from users, it is essential for developers and researchers to understand the general sentiment of these reviews. Understanding this sentiment can provide valuable insights into the aspects of the game that users like or dislike, as well as areas that need improvement or further development (Putriani et al., 2022).

However, given the enormous number of reviews, the process of manually extracting and analyzing sentiment is inefficient and prone to errors. Therefore, a more efficient and automated sentiment analysis approach is required. One suitable method for this is the K-Nearest Neighbors (KNN) Algorithm. KNN is a machine learning algorithm that classifies data based on its similarity to other data whose categories are already known. This algorithm is chosen for its ability to handle large review datasets with diverse content and its flexibility in adjusting the k-value (number of nearest neighbors) to optimize classification accuracy (Putriani et al., 2022).

For instance, a study conducted by (Habibah et al., 2023) on Sentiment Analysis Regarding E-Wallet Usage on Google Play explained that the algorithms used to assist with this analysis were Lexicon-Based and K-Nearest Neighbor. The research results, based on the mentioned issues, showed public responses about the three applications and the accuracy rates from implementing Lexicon-Based and K-Nearest Neighbor for each E-Wallet. Dana achieved the highest accuracy of 78% with k = 6, Ovo achieved 75.33% accuracy with k = 9, and LinkAja achieved 73.5% accuracy with k = 8. The application with the most positive responses from users was LinkAja, with 6,037 positive reviews.

Thus, the researchers chose to use the K-Nearest Neighbor method to analyze the sentiment of Clash of Clans reviews. By utilizing the vast amount of review data available on the Google Play platform, the researchers aim to see how accurately this method can identify two types of sentiment, namely positive and negative. The reason for choosing this method is its ability to modify k-values to produce more stable and higher accuracy.

METHOD

1. Sentiment Analysis

Sentiment analysis or opinion mining is one of the main tasks in Natural Language Processing (NLP), which involves studying a person's opinions about a specific entity. Sentiment analysis itself originates from the perspective of product users. In addition to being based on someone's opinions, sentiment analysis can be used to evaluate a person's behavior or emotions. The basic task when performing sentiment analysis is to classify the polarity of the obtained text into a document, sentence, or opinion with positive and negative aspects (Rahmattullah, 2022).

2. Labeling Data

Labeling aims to determine whether reviews are positive, negative, or neutral. In this process, researchers manually label the data using Microsoft Excel to determine the sentiment value based on the score category (rating from the reviews). The accuracy of labeling is crucial as it affects the model's performance in sentiment analysis. Inconsistent or inaccurate labels can lead to incorrect analysis. Data is categorized as positive if the score is > 4, negative if the score is < 2, and neutral if the score is = 3. However, in this analysis, only positive and negative reviews will be used (Rahayu et al., 2022).

3. Text Preprocessing

The process of transforming data into a structured format suitable for data mining, typically into numerical values, is known as Text Preprocessing (Siti Masturoh, 2019).Once the data is structured and converted into numerical values, it can be used as a source for further processing. The main processes involved are as follows:

a. Cleaning

Cleaning is the process of removing duplicates and irrelevant attributes from the comment columns. The primary goal is to ensure that the data is free from duplicates and unnecessary attributes.

b. Case Folding

Case Folding aims to convert all text into lowercase. This process changes all letters in the document to lowercase, accepting only the letters 'a' to 'z'. Non-letter characters are removed and treated as separators.

c. Tokenizing

Tokenizing involves breaking down sentences into individual words. This process parses sentences into individual words and removes separators such as periods (.), commas (,), quotation marks ("), parentheses (()), spaces, and numbers.

d. Stopwords Removal

Stopwords Removal aims to remove words classified as stopwords or non-essential words, such as conjunctions like "and," "with," etc. This stage involves filtering out non-essential words from the tokenization results, which can be done using a stoplist algorithm (removing less important words) or a wordlist (retaining important words). Stopwords are non-descriptive words that can be discarded, such as "which," "and," "in," etc.

e. Stemming

Stemming involves reducing words with affixes to their base form. This process removes prefixes or suffixes from words to achieve a more general root form.

4. K-Nearest Neighbor

The K-Nearest Neighbors (KNN) algorithm classifies based on basic examples that do not form an explicit declarative representation of categories but rely on the category labels of training documents similar to the test document. KNN is a classification method that determines the class of an object based on the nearest training data. The principle of KNN is straightforward: it works based on the nearest distance from the test sample to the training samples (Adhi Putra, 2021).The K-Nearest Neighbor algorithm involves the following steps:

a. Determine the parameter k (the number of nearest neighbors).

b. Assign weights to each term using Term Weighting TF-IDF.

c. Calculate the similarity between documents using cosine similarity:

$$cos(i,k) = \frac{\sum_k (d_1 d_k)}{\sqrt{\sum_k d_{ik}}^2 \sqrt{\sum_k d_{jk}}^2}$$

Where:

$\sum_k (d_1 d_k)$	= Dot product vector of i and k
$\sqrt{\sum_k d_{ik}}^2$	= Magnitude of vector i
$\sqrt{\sum_k d_{jk}}^2$	= Magnitude of vector k

d. Sort the results of the cosine similarity calculation from highest to lowest.

e. Select the top K most similar documents to the classified document and determine its class.

5. Evaluasi and Validasi

Cross Validation is a technique used to assess or validate the accuracy of a model built on a specific dataset. The purpose of model creation is often to perform prediction or classification on new data that may not have appeared in the dataset before. A confusion matrix is a commonly used method for calculating accuracy. In evaluating the accuracy of the results, metrics such as recall, precision, accuracy, and error rate are assessed (Adhi Putra, 2021).

RESULT AND DISCUSSION

The results and discussion for sentiment analysis on the Clash of Clans application using the K-Nearest Neighbor method, processed with Jupyter Notebook, will be presented in the following sections. This will include examples of scraped data, labeled data, preprocessing steps, and the accuracy values of each.

1. Scrapping Data

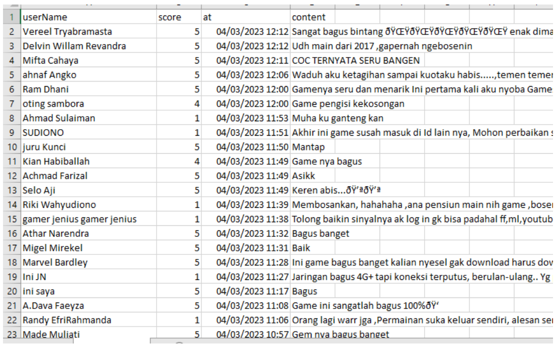
At this stage, data scraping is performed by including the link to the Clash of Clans application from the Play Store. The results of data scraping with the keyword "Clash of Clans," with a total of 2000 data points collected, are shown below using the script.

Scraping

```
from google_play_scraper import Sort, reviews

result, continuation_token =
    reviews(
    'com.supercell.clashofclans',
    lang='id',
    country='id
    ',
    sort=Sort.NEWEST,
    count=2000,
    filter_score_with=None
)
```

Following are the output results:



	userName	score	at	content
1	userName	score	at	content
2	Vereel Tryabramasta	5	04/03/2023 12:12	Sangat bagus bintang ðŸŒŸðŸŒŸðŸŒŸðŸŒŸðŸŒŸ enak dima
3	Delvin Willam Revandra	5	04/03/2023 12:12	Udh main dari 2017 ,gapernah ngebosenin
4	Mifta Cahaya	5	04/03/2023 12:11	COC TERNYATA SERU BANGEN
5	ahnaf Angko	5	04/03/2023 12:06	Waduh aku ketagihan sampai kuotaku habis.....,temen temen
6	Ram Dhani	5	04/03/2023 12:00	Gamenya seru dan menarik Ini pertama kali aku nyoba Game
7	oting sambora	4	04/03/2023 12:00	Game pengisi kekosongan
8	Ahmad Sulaiman	1	04/03/2023 11:53	Muha ku ganteng kan
9	SUDIONO	1	04/03/2023 11:51	Akhir ini game susah masuk di Id lain nya, Mohon perbaikan s
10	juru Kunci	5	04/03/2023 11:50	Mantap
11	Kian Habiballah	4	04/03/2023 11:49	Game nya bagus
12	Achmad Farizal	5	04/03/2023 11:49	Asikk
13	Selo Aji	5	04/03/2023 11:49	Keren abis...ðŸ'ªðŸ'ª
14	Riki Wahyudiono	1	04/03/2023 11:39	Membosankan, hahahaha ,ana pensiun main nih game ,bosen
15	gamer jenius gamer jenius	1	04/03/2023 11:38	Tolong baikin sinyalnya ak log in gk bisa padahal ff,ml,youtub
16	Athar Narendra	5	04/03/2023 11:32	Bagus banget
17	Migel Mirekel	5	04/03/2023 11:31	Baik
18	Marvel Bardley	5	04/03/2023 11:28	Ini game bagus banget kalian nyesel gak download harus dov
19	Ini JN	1	04/03/2023 11:27	Jaringan bagus 4G+ tapi koneksi terputus, berulan-ulang.. Yg
20	ini saya	5	04/03/2023 11:17	Bagus
21	A.Dava Faeyza	5	04/03/2023 11:08	Game ini sangatlah bagus 100%ðŸ'
22	Randy EfriRahmanda	1	04/03/2023 11:06	Orang lagi warr jga ,Permainan suka keluar sendiri, alesan se
23	Made Muliati	5	04/03/2023 10:57	Gem nya bagus banget

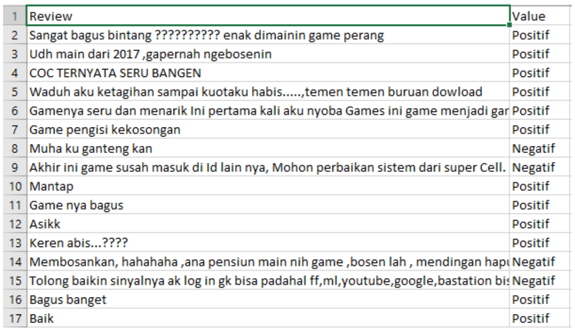
Figure 1 Data Scraping results

2. Labeling Data

After collecting the research data, the next step is labeling the Clash of Clans application reviews. The labeling results, using Excel formulas, are shown in Figure 1.

Labeling Data
=IF(B2>=4;"Positif";IF(B2=3;"Netral";"Negatif"))

Following are the output results:



	Review	Value
1	Review	Value
2	Sangat bagus bintang ?????????? enak dimainin game perang	Positif
3	Udh main dari 2017 ,gapernah ngebosenin	Positif
4	COC TERNYATA SERU BANGEN	Positif
5	Waduh aku ketagihan sampai kuotaku habis.....,temen temen buruan dowload	Positif
6	Gamenya seru dan menarik Ini pertama kali aku nyoba Games ini game menjadi gar	Positif
7	Game pengisi kekosongan	Positif
8	Muha ku ganteng kan	Negatif
9	Akhir ini game susah masuk di Id lain nya, Mohon perbaikan sistem dari super Cell.	Negatif
10	Mantap	Positif
11	Game nya bagus	Positif
12	Asikk	Positif
13	Keren abis...????	Positif
14	Membosankan, hahahaha ,ana pensiun main nih game ,bosen lah , mendingan hapu	Negatif
15	Tolong baikin sinyalnya ak log in gk bisa padahal ff,ml,youtube,google,bastation bis	Negatif
16	Bagus banget	Positif
17	Baik	Positif

Figure 2 Labeling Data Results

Figure 2 shows the labeled data. The previous data only had the columns user, score, at, and content. In this research, only the score and content columns are used. The score column will be renamed to value, and the content column will be renamed to review.

3. Text Preprocessing

In the text preprocessing stage, the goal is to transform unstructured data into a structured format suitable for analysis by cleaning and preparing the data. The text preprocessing process consists of cleaning, case folding, tokenizing, stopwords removal, and stemming.

Cleaning, this step cleans the comment column data by removing duplicates and irrelevant attributes. The goal is to ensure the data is free from duplicates and irrelevant attributes. Table 1 shows the results of the cleaning process.

Table 1 The result of the cleaning process

Before Cleansing	**After Cleansing**
PENGEMBALIAN SISTEM FITUR CHAT PUBLIK GLOBAL CLASH OF CLANS TOLONG DI PERTIMBANGKAN	PENGEMBALIAN SISTEM FITUR CHAT PUBLIK GLOBAL CLASH OF CLANS TOLONG DI PERTIMBANGKAN

The second text preprocessing process is case folding. In this step, the text is converted to lowercase. This is done by changing all the letters in the document to lowercase. Only the letters 'a' to 'z' are accepted. Characters other than letters are removed and considered delimiters. Table 2 shows the results of the case folding process.

Table 2 The result of the case folding process

Before Case Folding	**After Case Folding**
PENGEMBALIAN SISTEM FITUR CHAT PUBLIK GLOBAL CLASH OF CLANS TOLONG DI PERTIMBANGKAN	pengembalian sistem fitur chat publik global clash of clans tolong di pertimbangkan

The third text preprocessing process is tokenizing. The goal of this process is to break down sentences into individual words. This involves parsing descriptions, initially in the form of sentences, into words and removing delimiters such as periods (.), commas (,), quotation marks (“), parentheses (()), spaces, and numeric characters. Table 3 shows the results of the tokenizing process.

Table 3 The result of the Tokenizing process

Before Tokenizing	**After Tokenizing**
pengembalian sistem fitur chat publik global clash of clans tolong di pertimbangkan	[pengembalian, sistem, fitur, chat, publik, global, clash, of, clans, tolong, di, pertimbangkan]

The fourth text preprocessing process is stopwords removal. In this step, words categorized as stopwords, or words that are considered non-essential, are removed. Examples of such words include conjunctions like "and," "with," and others. Table 4 shows the results of the stopwords removal process.

Table 4 The result of the stopwords process

Before Stopwords Removal	**After Stopwords Removal**
[pengembalian, sistem, fitur, chat, publik, global, clash, of, clans, tolong, di, pertimbangkan]	[pengembalian, sistem, fitur, chat, publik, global, clash, clans, tolong, pertimbangkan]

The final text preprocessing process is stemming. In this step, words with affixes are reduced to their base form. This is done to produce the root form of the words by removing prefixes or suffixes, resulting in a more general representation. Table 5 shows the results of the stemming process.

Table 5 The result of the stemming process

Before Stemming	**After Stemming**
pengembalian sistem fitur chat publik global clash of clans tolong di pertimbangkan	kembali sistem fitur chat publik global clash of clan tolong timbang

4. Accuracy Results

In the classification stage, a machine learning model is developed using training and testing data randomly selected from the entire dataset to perform cross-validation and generate accuracy prediction values.

Table 6 Test results divide training data and testing data

Data	*Recall*	*Precision*	*Accuracy*	F1-*Score*
50% - 50%	83,27%	64,81%	59,03%	72,89%

60% - 40%	88%	68,4%	64,52%	76,97%
70% - 30%	80,66%	64,8%	58,39%	71,92%
80% - 20%	87,39%	64,39%	60,27%	74,15%
90% - 10%	87,93%	62,19%	58,46%	72,85%

Below is an illustration of the classification results using the K-Nearest Neighbor algorithm script.

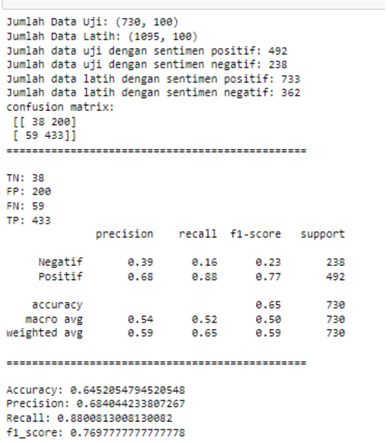
K-Nearest Neighbor

```
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.neighbors import KNeighborsClassifier

clf = KNeighborsClassifier(n_neighbors=7).fit(X_train, y_train)
predicted = clf.predict(X_test)
print(f'confusion matrix:\n {confusion_matrix(y_test, predicted)}')
print('===============================================\n')
tn, fp, fn, tp = confusion_matrix(y_test, predicted).ravel()
print("TN:", tn)
print("FP:", fp)
print("FN:", fn)
print("TP:", tp)
print(classification_report(y_test, predicted, zero_division=0))
print('===============================================\n')
print("Hasil Klasifikasi Sentimen Analisis Clash Of Clans:")
accuracy = accuracy_score(y_test, predicted)
precision = precision_score(y_test, predicted, average="binary", pos_label="Positif")
recall = recall_score(y_test, predicted, average="binary", pos_label="Positif")
f1 = f1_score(y_test, predicted, average="binary", pos_label="Positif")

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("f1_score:", f1)
```

The output of the script above can be seen in Figure 3



```
Jumlah Data Uji: (730, 100)
Jumlah Data Latih: (1095, 100)
Jumlah data uji dengan sentimen positif: 492
Jumlah data uji dengan sentimen negatif: 238
Jumlah data latih dengan sentimen positif: 733
Jumlah data latih dengan sentimen negatif: 362
confusion matrix:
 [[ 38 200]
 [ 59 433]]
===============================================

TN: 38
FP: 200
FN: 59
TP: 433
              precision    recall  f1-score   support

     Negatif       0.39      0.16      0.23       238
     Positif       0.68      0.88      0.77       492

    accuracy                           0.65       730
   macro avg       0.54      0.52      0.50       730
weighted avg       0.59      0.65      0.59       730

===============================================

Accuracy: 0.6452054794520548
Precision: 0.684044233807267
Recall: 0.8800813008130082
f1_score: 0.7697777777777778
```

Figure 3 Accuracy Result

Table 6 presents the validation results of applying the K-Nearest Neighbor algorithm to the Clash of Clans application. It shows that the highest accuracy is achieved in the 60:40 scenario, with an accuracy of 59,03%, precision of 64,81%, recall of 83,27%, and an F1-score of 72,89%.

Result

The second test uses the Fold value as a parameter. In this process, 10 scenarios will be conducted based on the determination of the k-value in the KNN algorithm, with values ranging from 2 to 11. This test uses a 60:40 dataset combination, which in the previous test showed the highest accuracy. Table 4.15 below presents the results of Cross Validation.

Table 7 Test results of k-Fold Cross Validation

Nilai K-*Fold*	*Recall*	*Precision*	*Accuracy*	F1-*Score*
2	95,55%	96,16%	95,55%	95,59%
3	92,22%	93,15%	92,22%	92,24%
4	93,28%	95,07%	93,28%	93,41%
5	91,11%	92,76%	91,11%	91,2%
6	93,33%	95,37%	93,33%	93,47%
7	92,3%	93,95%	92,3%	92,43%
8	92,23%	94,12%	92,23%	92,41%
9	92,22%	94,97%	92,22%	92,66%
10	92,22%	94,29%	92,22%	92,47%
11	92,29%	94,52%	92,29%	92,61%

From the table, it is shown that the highest accuracy is at K-Fold value 2, with an accuracy of 95.55%. At K-Fold, the precision is also higher at K-Fold value 2, with a precision of 96.61%.

Discussion

The researchers chose to use the K-Nearest Neighbor method to analyze the sentiment of Clash of Clans reviews. By utilizing the vast amount of review data available on the Google Play platform, the researchers aim to assess how accurately this method can identify two types of sentiment, namely positive and negative. The reason for choosing this method is its ability to modify k-values, which leads to more stable and higher accuracy.

CONCLUSIONS

1. This study conducted a classification process through several stages, including data labeling, text preprocessing, word weighting, and classification using the K-Nearest Neighbor (KNN) algorithm. A total of 2,000 reviews were used in this research. In the data labeling stage, the researchers categorized the sentiment into positive, neutral, and negative. However, neutral sentiment was excluded from the process to ensure more accurate results. After labeling, 1,825 reviews were ready for analysis. The preprocessing stage aimed to clean and prepare the data for analysis. Next, word weighting was performed using the TF-IDF method. The weighted data was then classified using the KNN algorithm. The classification results showed 1,225 reviews with positive sentiment and 600 reviews with negative sentiment. The classification outcomes were evaluated using a confusion matrix to determine accuracy. Additionally, the performance of the K-Nearest Neighbor algorithm was tested using the k-fold cross-validation method to evaluate its effectiveness.

2. The ratio between training and test data significantly influenced the improvement in accuracy. The best accuracy was achieved with a 60:40 ratio, with an accuracy of 59,03%, precision of 64,81%, recall of 83,27%, and an F1-score of 72,89%. The application of TF-IDF word weighting resulted in high accuracy at k-2, with an accuracy of 95.55%, precision of 96.16%, recall of 95.55%, and F1-score of 95.59%. The test results demonstrated that the K-Nearest Neighbor method is efficient for analyzing the sentiment of user reviews for the online game Clash of Clans. These results provide valuable insights for game developers in understanding user perceptions and feedback.

REFERENCES

Adhi Putra, A. D. (2021). Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(2), 636–646. https://doi.org/10.35957/jatisi.v8i2.962

AlHafiidh, A., & Oktadini, N. R. (2023). Analisis User Experience (UX) Pada Aplikasi Game Clash Of Clans Menggunakan Metode User Experience Questionnaire (UEQ). *INFORMATION SYSTEM FOR EDUCATORS AND PROFESSIONALS : Journal of Information System*, *8*(2), 219. https://doi.org/10.51211/isbi.v8i2.2694

Habibah, N., Budianita, E., Fikry, M., & Iskandar, I. (2023). Analisis Sentimen Mengenai Penggunaan E-Wallet Pada Google Play Menggunakan Lexicon Based dan K-Nearest Neighbor. *Jurnal Riset Komputer)*, *10*(1), 2407–389. https://doi.org/10.30865/jurikom.v10i1.5429

Putriani, N., Umbara, F. R., & Sabrina, P. N. (2022). Analisis Sentimen pada Aplikasi PeduliLindungi dengan Menggunakan Metode Improved K-Nearest Neighbor dan Lexicon Based. *Jurnal Teknologi Informatika Dan Komputer*, *8*(1), 350–364. https://doi.org/10.37012/jtik.v8i1.1107

Rahayu, S., MZ, Y., Bororing, J. E., & Hadiyat, R. (2022). Implementasi Metode K-Nearest Neighbor (K-NN) untuk Analisis Sentimen Kepuasan Pengguna Aplikasi Teknologi Finansial FLIP. *Edumatic: Jurnal Pendidikan Informatika*, *6*(1), 98–106. https://doi.org/10.29408/edumatic.v6i1.5433

Rahmattullah, R. (2022). Analisis sentimen persepsi pengguna pedulilindungi menggunakan algoritma. *Skripsi Repository Universitas Islam Indonesia Yogyakarta*.

Siti Masturoh. (2019). *Analisis Sentimen E-Wallet Ovo Dan Dana Pada Ulasan Google Play Menggunakan Algoritma k-Nearest Neighbor*.