Laila Atikah Sari - Analysis of Public Sentiment on Google Play Store Tije Application Users Using Naïve Bayes Classifier Method

by Layanan Perpustakaan UHAMKA

Submission date: 28-Dec-2023 10:54AM (UTC+0700)
Submission ID: 2265214584
File name: LAILA_ATIKAH_SARI_TI_TURNITIN_KE-1_-_Laila_Atikah_Sari.docx (780.44K)
Word count: 4004
Character count: 21878

Jurnal Teknik Informatika (JUTIF)
Vol. x, No. y, December 2023, pp. x-y

DOI: 10.52436/jutif...
p-ISSN: 2723-3863
e-ISSN: 2723-3871

Analysis of Public Sentiment on Google Play Store Tije Application Users Using Naïve Bayes Classifier Method

Laila Atikah Sari [1], Nindia Fitri Ramadhita[2], Firman Noor Hasan*[3]

[1,3]Informatics Engineering, Faculty of Industrial Technology and Informatics, Universitas Muhamadiyah Prof. Dr. Hamka, Indonesia
[2]International Relations, Faculty of Social Sciences, Sakarya Universitesi, Turkey
Email: [1]lailaatikah00@gmail.com, [2]nindiadhitaa@gmail.com, [3]firman.noorhasan@uhamka.ac.id

(Article received: date; Revision: date; published: date)

Abstract

Advances in information technology have an influence on companies and agencies to innovate. The Tije application is one of the innovations that has been made by PT Tranportasi Jakarta which is used by its users. Of course, each application has advantages and disadvantages that can trigger responses from users that can be submitted through the review column on the Google Play Store platform. This research was conducted to analyze the sentiment of community reviews of Tije application users on the Google Play Store platform using the Naïve Bayes Classifier method. Tije application review data collection is done by web scrapping techniques on the Google Play Store using Google Colab. Then, the collected data will be processed to remove inappropriate elements and obtain the sentiment content in each review, whether the review falls into the category of positive or negative sentiment towards the Tije application. The results of this study conclude that users are dissatisfied and disappointed with the services in the Tije application. This is evidenced by the number of negative sentiments that are more dominant and in the application of the Naive Bayes algorithm in this study, obtained quite good accuracy results of 85.88%.

Keywords: *Google Play Store, Naïve Bayes Classifier, Reviews, Sentiment Analysis, Tije App.*

1. INTRODUCTION

In the digital era, information technology continues to progress which changes the way people interact, thus bringing a number of new challenges for people to understand more about information technology [1]. With the progress that has occurred, it has influenced various companies and agencies to compete to create new innovations [2]. One of the companies that innovate is PT. Transportation Jakarta or commonly known as Transjakarta. Transjakarta is a public transportation system in Jakarta designed to provide quality services at affordable rates for its users [3]. One of the efforts to provide quality services by Transjakarta is by creating an application called Tije which can be used by customers as service support. The application itself is a ready-made program that has special functions based on its ability to solve user problems [4]. There are benefits gained in using applications, one of which is as a means of disseminating information easily for its users, ranging from local and world information [5]. The Tije application has various features that help Transjakarta service users, such as updated information on service routes that operate every day and ticket purchase features that can make it easier for Transjakarta service users. However, each application certainly has advantages and disadvantages, which can trigger responses from users to the application [6]. Usually, responses from application users can be given through the comments column on the Tije application on the Google Play Store platform. Based on the data taken on 5 November 2023, the Tije application has been downloaded by more than 500 thousand users with a rating of 1.7 and recorded more than 7 thousand user comment reviews [7].

This research will use Tije application user reviews as data for sentiment analysis. Sentiment analysis is a direct way to identify, extract, and process textual information to find emotional information [8]. Product sentiment analysis is carried out to classify user opinions of text-shaped products that contain positive or negative aspects which can later be used as a benchmark regarding the product [9]. Retrieval of application user review data information is carried out using data scrapping techniques which will be continued to the preprocessing stage until the evaluation stage. Sentiment analysis will be applied using the Naive Bayes Classifier method to identify whether a review has a positive meaning or a negative meaning. Naive Bayes is one part of a type of classification algorithm in data mining [10]. With the basic concept of using Bayes' theorem, Naive Bayes can be used to calculate the probability and find the accuracy of the data [11].

The use of the Naïve Bayes Classifier method in research has proven its accuracy compared to several other classification methods which have the

highest accuracy result of 99.22% for the Naïve Bayes method [12]. This research was conducted to determine the satisfaction of Tije application users and to determine the effectiveness of using the Naïve Bayes Classifier algorithm in sentiment analysis research based on reviews obtained from comments on the Google Play Store. So that the results of this study are expected to be a source of information for companies, namely PT. Transportation Jakarta in making decisions regarding the right steps to develop the Tije application in the future.

2. RESEARCH METHODS

The researcher conducted several stages which are visualized through Figure 1.

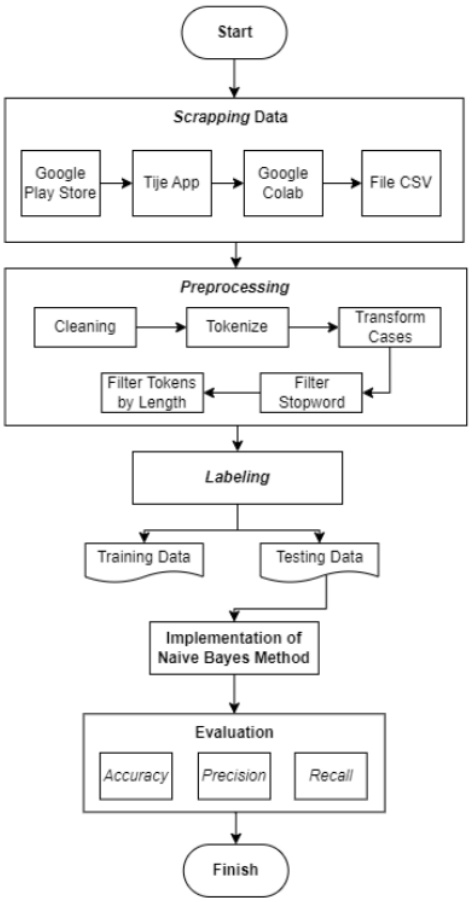

Figure 1. Stages of Research

In this research, it starts with scrapping the user review data of the Tije application through the Google Play Store using the help of Google Colab. The results of scrapping the data will be cleaned through the preprocessing process to produce a dataset. After the dataset is formed, a sample is made as training data by labeling the data which is a negative or positive opinion as much as 80% of the dataset results. Then, the training data will be used to train the test data, which is the rest of the dataset that has not been labeled. The test is done by implementing the Naïve Bayes Classifier method. After the implementation of the Naive Bayes Classifier is carried out, the final stage will continue, namely the evaluation process using the results of the confusion matrix to determine accuracy, precision, and recall.

2.1 Problem Identification

This research was conducted on the basis of the problems previously described, namely:

1. To find out the sentiment of Tije application users. Sentiment analysis will produce a percentage of the number of user responses in the form of positive and negative sentiments.
2. To identify the effectiveness of using the Naive Bayes Classifier method in assessing the sentiment of a topic of discussion, especially regarding user reviews of the Tije application.
3. Knowing what words often appear in each sentiment.

2.2 Scrapping Data

Figure 2. Flow of Scrapping Data

Figure 2 is a series of data collection processes, where data collection is taken based on relevant data to be analyzed in the next stage. this process starts from retrieving Tije application reviews with web scrapping techniques on the Google Play Store using Google Colab and the results will be saved into a CSV format file. Web scrapping is a stage used to collect application review data using the Python programming language with the aim of obtaining information [13]. This technique is only enough by entering the application link you want to retrieve data from and the desired amount of data [14].

2.3 Preprocessing

In the preprocessing stage, researchers clean the data that has been collected in the previous stage. This process is carried out to eliminate problems that exist in the data, because sometimes there is data that has problems that affect the final results of the data processing process [15]. In addition, this stage helps researchers to conclude whether comments are negative or positive at a later stage. This research, in preprocessing to evaluation, is assisted by

RapidMiner tools. RapidMiner is a software that provides tools for data processing, training machine learning models, text analysis, and making predictions [16]. The preprocessing stage is carried out through several series, namely:

a. Cleaning, the process of cleaning special characters or punctuation marks using the replace operator.
b. Subprocess, is a series of processes carried out at the preprocess stage starting from Tokenize, Tranforms Cases, Filter Stopwords, to Filter Tokens by Length.
c. Filter Example, used to eliminate datasets that do not have values or are empty (missing).
d. Remove Duplicate, is an operator used to eliminate data in the dataset that repeats.

2.4 Labeling

.Labeling is the process of forming sentiment. In this research, labeling was carried out manually using 2 sentiment categorizations, namely positive sentiment and negative sentiment. Usually this sentiment categorization is carried out by researchers to be used as training data.

Training data is a collection of data used to train a model or sentiment analysis method. Meanwhile, testing data is a collection of data used to test the performance of a model that has been trained using training data . Labeling is used to assign sentiment labels to each review entity in the dataset. This process is very crucial, because labels play an important role in determining accuracy results at the next stage.

2.5 Naïve Bayes Classifier

Naïve Bayes classification applies Bayes' Theorem. Bayes' theorem was first discovered by an English scientist named Thomas Bayes. In the context of sentiment analysis, many use the Naïve Bayes algorithm as a classification method. Naïve Bayes classification has the ability to predict the probability of class membership under the assumption of independence. Thus, it can identify future opportunities based on previous experience [2]. For example, when a new comment is entered, the step to classify it is to calculate its probability in the positive and negative classes, using the information from the previous training process [12].

2.6 Confusion Matrix

Confusion Matrix is a method involving the comparison of a prediction result matrix with the original class, which includes actual information and classification prediction values. After the system has successfully classified the tweets, a measure is needed to assess how accurate or precise the classification has been done by the system. Through evaluation using Confusion Matrix and these metrics, research can measure how well the classification model recognizes the desired classes and identify potential classification errors that may occur [17].

3. RESULTS AND DISCUSSION

3.1 Scrapping Data

```
from google_play_scraper import Sort, reviews

result, continuation_token = reviews(
    'com.transjakarta.tijeku',
    lang='id',
    country='id',
    sort=Sort.NEWEST,
    count=1000,
    filter_score_with=None
)

df_busu = pd.DataFrame(np.array(result),columns=['review'])
df_busu = df_busu.join(pd.DataFrame(df_busu.pop('review').tolist()))
df_busu.head()
len(df_busu.index)
df_busu[['userName', 'score','at', 'content']].head()
```

Figure 3. scrapping data with google colab

Figure 3 shows the data scrapping stage on Google Colab by entering the Tije application link on the Google Play Store, namely 'com.transjakarta.tijeku' by taking the amount of data as much as 1000 based on the latest category (NEWEST) with the attributes taken consisting of username 'username', rating 'score', date 'at', and review 'content'. Based on the given category, the data obtained from scrapping user reviews starting from April 11, 2022 to August 27, 2023. Then Figure 4 is the process of saving the scrapping results into the Comma Separated Value (CSV) File format with the name "scrapped_data_Tije.csv".

```
my_df.to_csv("scrapped_data_Tije.csv", index = False)
```

Figure 4. process of saving the scrapped data

3.2 Preprocessing

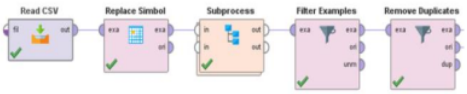

Figure 5. preprocessing stage

Figure 5 is the steps taken in preprocessing. At this stage, a series of processes are carried out, starting with processing the raw data that has been obtained through the data scrapping process, of course, it still needs to be cleaned, such as attributes and symbols that are not needed in this study. Therefore, data cleaning is done with the help of RapidMiner. Of the total 1000 review data that was successfully cleaned, 849 review data remained

which would be continued to the next stage. The results of the cleaned data are listed in Table 1.

Table 1. results of the preprocessing stage

Sebelum	Sesudah
Indri Yanto,1,2023-08-17 10:36:28,Aplikasi tolol.....di buka ga bisa sama sekali.....mending di hapus aja	Aplikasi tololdi buka ga bisa sama sekalimending di hapus aja
Nadia Khoirunnisa,1,2023-08-02 14:24:56,Metode bayarnya tolong di banyakin jangan hanya as***pay	Metode bayarnya tolong di banyakin jangan hanya aspay
Rizky Aulia Rosman,1,2023-05-05 00:39:59,Tije (tidak jelas) udah paling bener trafi dulu.	Tije tidak jelas udah paling bener trafi dulu
Heru Putra,5,2023-08-22 09:38:21,Ok bngeeet....sangatx meembantu	Ok bngeeetsangatx meembantu
ukhti melati,1,2023-06-26 00:14:26,"Pas dibuka ""terkendala masalah koneksi"" terus. Pdahal internet kenceng"	Pas dibuka terkendala masalah koneksi terus Pdahal internet kenceng

After cleaning characters or punctuation marks, the process will enter the Subproces operator series, which contains several series, as in Figure 6.

Figure 6. Proses Operator Subproses

There is a tokenize process, which is the process of breaking sentences into pieces of words that suit the needs to facilitate text processing [18]. Table 2 shows an example of tokenize results.

Table 2. tokenize results

Sebelum	Sesudah
Aplikasi tololdi buka ga bisa sama sekalimending di hapus aja	Aplikasi, tololdi, buka, ga, bisa, sama, sekalimending, di, hapus, aja
Metode bayarnya tolong di banyakin jangan hanya aspay	Metode, bayarnya, tolong, di, banyakin, jangan, hanya, aspay
Tije tidak jelas udah paling bener trafi dulu	Tije, tidak, jelas, udah, paling, bener, trafi, dulu
Ok bngeeetsangatx meembantu	Ok, bngeeetsangatx, meembantu
Pas dibuka terkendala masalah koneksi terus Pdahal internet kenceng	Pas, dibuka, terkendala, masalah koneksi, terus, Pdahal, internet, kenceng

Then, the transform cases process is performed, which is the process of converting uppercase letters in the text into lowercase letters used to eliminate letter differences in the dataset [19]. Table 3 shows an example of the transform cases result.

Table 3. transform cases result

Sebelum	Sesudah
Aplikasi, tololdi, buka, ga, bisa, sama, sekalimending, di, hapus, aja	aplikasi, tololdi, buka, ga, bisa, sama, sekalimending, di, hapus, aja
Metode, bayarnya, tolong, di, banyakin, jangan, hanya, aspay	metode, bayarnya, tolong, di, banyakin, jangan, hanya, aspay
Tije, tidak, jelas, udah, paling, bener, trafi, dulu	tije, tidak, jelas, udah, paling, bener, trafi, dulu
Ok, bngeeetsangatx, meembantu	ok, bngeeetsangatx, meembantu
Pas, dibuka, terkendala, masalah koneksi, terus, Pdahal, internet, kenceng	pas, dibuka, terkendala, masalah koneksi, terus, pdahal, internet, kenceng

After all words have become lowercase letters, a stopword filter process is carried out. This process is done to eliminate words that have no meaning and have no effect on sentences such as "continue" "used to" "only" [19]. Researchers used a stopword dictionary that had been downloaded via the website www.kaggle.com. Table 4 shows an example of the stopword filter results.

Table 4. stopword result

Sebelum	Sesudah
aplikasi, tololdi, buka, ga, bisa, sama, sekalimending, di, hapus, aja	aplikasi, tololdi, buka, ga, sekalimending, hapus, aja
metode, bayarnya, tolong, di, banyakin, jangan, hanya, aspay	metode, bayarnya, tolong, banyakin, aspay
tije, tidak, jelas, udah, paling, bener, trafi, dulu	tije, udah, bener, trafi
ok, bngeeetsangatx, meembantu	ok, bngeeetsangatx, meembantu
pas, dibuka, terkendala, masalah koneksi, terus, pdahal, internet, kenceng	pas, dibuka, terkendala, koneksi, pdahal, internet, kenceng

The process of filtering tokens by length, removes words that have a minimum and maximum character length that has been determined [19]. In this research, a minimum length parameter of 4 characters and a maximum of 25 characters is determined as in Figure 7. Table 5 shows examples of token results based on the length filter.

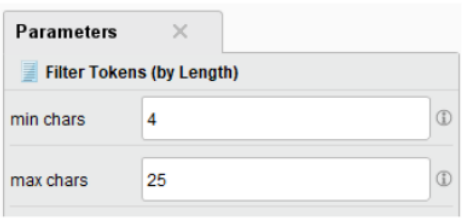

Figure 7. token filter parameters used

Table 5. filter tokens by length result

Sebelum	Sesudah
aplikasi, tololdi, buka, ga, sekalimending, hapus, aja	aplikasi, tololdi, buka, sekalimending, hapus
metode, bayarnya, tolong, banyakin, aspay	metode, bayarnya, tolong, banyakin, aspay
tije, udah, bener, trafi	tije, udah, bener, trafi
ok, bngeeetsangatx, meembantu	bngeeetsangatx, meembantu
pas, dibuka, terkendala, koneksi, pdahal, internet, kenceng	dibuka, terkendala, koneksi, pdahal, internet, kenceng

3.3 Labeling

After all data preprocessing stages have been completed, the data labeling process continues. Labeling is done by dividing the data into two parts, namely training data and test data, where this division uses a ratio of 80:20 with 80% of the total 849 data as training data and the remaining 20% as test data. In 679 training data, the process of manually labeling reviews from Tije application users using two labeling categories, namely labeling with positive categories and negative categories. Positive labels are given to user reviews that contain praise or user satisfaction. While negative labels are given to user reviews that contain criticism or user dissatisfaction. Then the rest of the dataset of 170 user review data will be used as test data in the Naïve Bayes implementation process.

Table 6. labeling process

Sebelum	Sesudah
Aplikasi tololdi buka ga bisa sama sekalimending di hapus aja	Negatif
Metode bayarnya tolong di banyakin jangan hanya aspay	Positif
Tije tidak jelas udah paling bener trafi dulu	Negatif
Ok bngeeetsangatx meembantu	Positif
Pas dibuka terkendala masalah koneksi terus Pdahal internet kenceng	Negatif

3.4 Implementation of Naïve Bayes Classifier Method

The series of method implementation processes begins with creating training data, such as the process in Figure 8, where the dataset file contains data that has been labeled and which has not been entered in the read CSV operator which has been connected to the set role operator to make the column read as a label. Then the filter examples operator is used to separate data that is labeled (is not missing). The process document or TF-IDF weighting operator is a subprocess that contains tokenize, transform cases, filter stopwords, and filter tokens by length operators. Then the Naïve Bayes Classifier operator is used to analyze the data that will be stored in the store named store model and training.

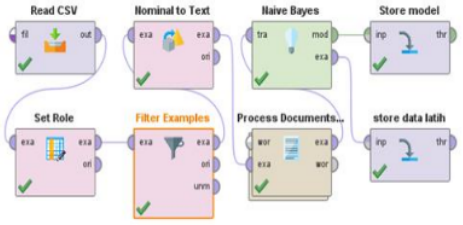

Figure 8. Training Data Creation Process

The results of the training data creation process will be used in the Naïve Bayes Classifier implementation. Figure 9 represents the stages in the Naïve Bayes Classifier implementation process, starting with using the read CSV operator which is connected with the filter examples operator to filter out data that has not been labeled (is missing). Furthermore, the analyzed data in Figure 8 named Tije training is unified with the results of TF-IDF weighting and connected to the filter examples 2 operator. While the analysis result data named store model is directly connected to the apply model operator and combined with the previous analysis result data.

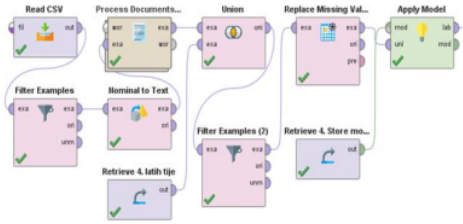

Figure 9. Naive Bayes Classifier Implementation Process

The results of the implementation have been visualized in Figure 10 which is a pie chart of the results of the review test of the total data of 849, with a total of 728 reviews as negative sentiment and 121 reviews as positive sentiment. From the results of sentiment prediction that has been done, it is found that user sentiment towards the Tije application tends to be negative. This is because many experience problems in using the Tije application, such as problems entering the application, purchasing tickets, and scanning barcodes and regretting that the payment method only uses one method, namely astrapay. So that users feel dissatisfied.

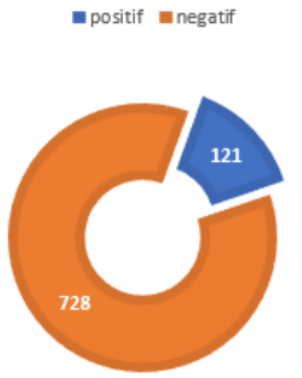

Figure 10. pie chart results

3.5 Evaluation

The final stage in this research is the evaluation stage which is used to determine the results of research performance using confusion matrix. Confusion matrix is used to represent the results of data prediction with the actual conditions of using the algorithm [20].

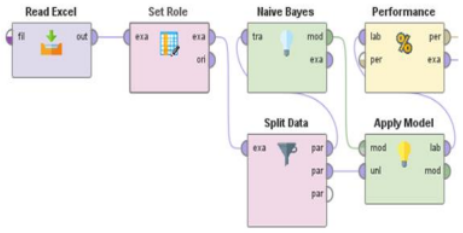

Figure 11. Confusion Matrix Process

Figure 11 shows the process carried out in the confusion matrix, starting with entering the prediction result file into the excel operator and connected with the set role operator to make the prediction column read as a label. The split data operator is used to split the data with a ratio of 80% to 20% which is then connected to the apply model operator. Then the apply model operator is connected to the performance operator to see the accuracy results.

accuracy: 85.88%

	true negatif	true positif	class precision
pred. negatif	146	24	85.88%
pred. positif	0	0	0.00%
class recall	100.00%	0.00%	

Figure 12. Accuracy Results from Performance Vector

Figure 12 shows the results of testing the Tije application user review dataset using the Naive Bayes Classifier algorithm method in the RapidMiner application which is then visualized as Figure 13. The results achieved an accuracy of 85.88%, with a precision value of 85.88%, and a recall value of 100%.

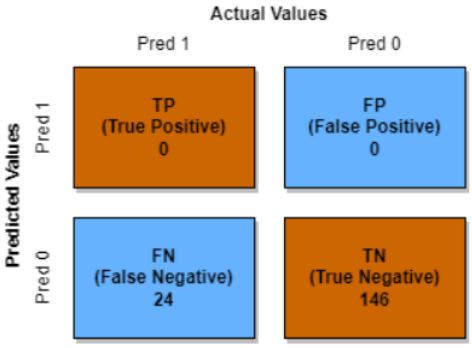

Figure 13. Confusion Matrix Visualization

3.6 Creating the Wordcloud

Figure 14. Wordcloud Creation Process

Wordcloud creation is done by weighting the Tije application user review dataset that has gone through the preprocessing stage. Figure 14 shows the stages of the wordcloud creation process. The process is carried out using the TF-IDF vector in the process documents from data, which includes several processes, namely tokenize, transform cases, filter stopwords, and filter tokens by length to take into account the frequency of words in the dataset [21]. Figure 15 shows the parameters used in the process documents from data operator. Then, the next process is connected to the WordList to Data operator to calculate the weight and frequency of words in the review.

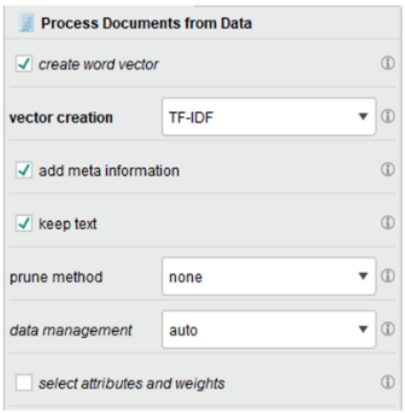

Figure 15. Parameters Process Documents from Data

Figure 16 is a visualization of the results created with wordcloud to display the 20 words with the highest frequency of occurrence. In wordcloud, it can be seen that the word 'aplikasi' (aplication) is the word that appears the most with the number of occurrences of 237 times. Then in second place is the word 'tiket' (ticket) with the number of occurrences 78 times. And in third place is the word 'pembayaran' (payment) with 71 occurrences.

Figure 16. wordcloud

4. DISCUSSION

Previous research has produced many sources that can be used as a basis for this research, because supporting data is needed from research that has been done before. Usually supporting data is called a reference as material to further review the topic of Sentiment Analysis with the Naive Bayes Method using application user reviews on the Google Play Store. One of the studies that conducted sentiment analysis of application user reviews is research on the application of the Naive Bayes method in analyzing the sentiment of video conferencing application users with the results showing that the highest level of accuracy is in the Webex Meeting application with 87% accuracy, Zoom Cloud Meetings with 80% accuracy, and Google Meet with 79% accuracy [22].

In another study that conducted research by applying the Naive Bayes and K-Nearest Neighbors methods to analyze sentiment related to COVID-19 by collecting public sentiment data through the Twitter platform resulted in an accuracy rate of 63.21% for Naïve Bayes and K-Nearest Neighbors of 58.10% with the tendency of public opinion in the positive category [23]. In addition, there is also research that uses Naive Bayes with three scenarios of training and testing data ratios on e-commerce reviews. The results showed that the first scenario, with a ratio of 80:20 between training data and testing data had the highest accuracy of 92% [24]. Furthermore, there is also research that conducts sentiment analysis based on Fore Coffe application reviews related to customer satisfaction. This research was conducted with web scrapping techniques in taking data from July 2019 to December 2022, analyzing reviews by applying the Naive Bayes method and showing that the application of the Naive Bayes method with testing using 10-fold cross-validation achieved a good level of accuracy, amounting to 74.28% (±2.68%) sari). From several previous studies, it shows that the value of community sentiment analysis using the Naive Bayes Classifier method has varying accuracy results. These varying accuracy values can be caused by several factors, including the quality of the data taken, the preprocessing steps used, and the way of labeling the training data.

5. CONCLUSION

Based on the analysis of research that has been conducted on 849 data processing results of datasets obtained from April 11, 2022 to August 27, 2023 through web scrapping techniques on the Tije Application on the Google Play Store platform, it is found that user reviews of the Tije application tend to be dominant in sentiments with negative categories, as many as 729, while there are only 121 data that are positive sentiment categories. Meanwhile, the results of the TF-IDF weighting process show 3 words that have the highest frequency of occurrence, namely 'application', 'ticket', and 'payment'. This shows that many Tije application users are dissatisfied because they experience many problems in using the Tije application, such as problems entering the application, purchasing tickets, and scanning barcodes and regretting that the payment method only uses one method, namely astrapay. In applying the Naive Bayes Classifier algorithm, the accuracy result is 85.88%, with a precision value of 85.88%, and a recall value of 100%. In addition, there are True Positive (TP) results of 0, True Negative (TN) of 146, False Positive (FP) of 0, and False Negative (FN) of 24. This proves that the use of the Naïve Bayes Classifier method has high accuracy when

implemented on large amounts of data. However, in order for future research to produce better confusion matrix results, it is hoped that future researchers can pay more attention to understanding feelings or opinions on user reviews at the training data labeling stage and can use a better preprocessing sequence.

Laila Atikah Sari - Analysis of Public Sentiment on Google Play Store Tije Application Users Using Naïve Bayes Classifier Method

ORIGINALITY REPORT

2%	1%	1%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dimas Ananda, Teguh Ammar Taqiyyuddin, IIyen Nugraha Faqih, Raihan Badrahadipura, Anindya Apriliyanti Pravitasari. "Application of Bidirectional Gated Recurrent Unit (BiGRU) in Sentiment Analysis of Tokopedia Application Users", 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021 Publication	1%
2	community.rapidminer.com Internet Source	<1%
3	Submitted to Deakin University Student Paper	<1%
4	Submitted to Waubonsie Valley High School Student Paper	<1%
5	www.ncbi.nlm.nih.gov Internet Source	<1%

6	Jiang Tao, Wang Chen, Boliang Wang, Xie Jiezhen, Jiao Nianzhi, Tingqwei Luo. "Real-Time Red Tide Algae Classification Using Naive Bayes Classifier and SVM", 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, 2008 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		