Jurnal Mantik, 6 (2) (2022) 1349-1357

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

SENTIMENT ANALYSIS ON SHOPEE E-COMMERCE USING THE NAÏVE BAYES CLASSIFIER ALGORITHM

Yulinar Rizkyani Saputri[1], Herny Februariyanti[2]

[1,2]Information Technology and Industry Faculty, Stikubank University, Jl. Tri Lomba Juang, Semarang, 50241, Indonesia.

E-mail: yrizkyanisaputri27@gmail.com[1], hernyfeb@edu.unisbank.ac.id[2]

ARTICLE INFO

Article history:
Received: Jun 15, 2022
Revised: Jul 23, 2022
Accepted: Aug 30, 2022

Keywords:
Naïve Bayes Classifier, RStudio, Text Mining, Text Processing, Shopee.

ABSTRACT

The growth of information and technology is causing individuals to live a more digital lifestyle, and one example of this is in the buying and selling activities that already use E-Commerce as a venue to facilitate them. Using the RStudio application and the Naive Bayes Classifier Algorithm, this study aims to find or analyze both positive and negative reviews of the e-commerce application through user reviews on Google Play. The technique employed is text mining and text processing, which involves stemming, tokenizing, case folding, normalization, and filtering steps. Review data for the Shopee app on Google Play is gathered through data scraping utilizing the web application appfollow, and review data may be saved in csv format. The acquired data will be processed using text processing, first by translating reviews in other languages to Indonesian, then by normalizing or cleaning the content to remove emoticons. The normalized data will then be uniformly converted to lower case letters as a whole in the case folding process, which comes next. Each word that has no influence or is independent, which is referred to as a token, will be isolated from the uniform data. Calculating the presence of words in the data and their frequency of occurrence is made easier by the tokenizing procedure. The Nave Bayes Classifier Method is used to compare training data and test data after text has undergone text processing to produce positive and negative sentiment classes based on the number of word frequencies.

1. Introduction

People's lives become more digital in the current era of digitalization, the rapid development of technology and information. The majority of people are very mobile in their regular routines throughout this pandemic. As a result, people frequently search online for anything useful to fulfill their everyday demands. The provision of a range of services that help the community satisfy its practical requirements comes next. E-commerce, or electronic commerce, is a marketing strategy that advertises things online directly. Companies provide benefits or advantages to customers so that they can stay in business. Shopee is one of the E-Commerce in Indonesia, and as the company's sales system becomes more appealing, simple, and profitable for customers, so do its earnings. As an e-commerce app that is frequently used, the Shopee app on the Google Play website was acknowledged to be in the top spot.

Based on comments made by users or consumers of the Tokopedia application, a number of prior research offered strategies for calculating the percentage level of comments and responses from Tokopedia application users. An accuracy value of 97.13 percent was obtained from the performance produced by Rapidminer's tests on 1,500 test data [1]. Additionally, the accuracy number on each test set of data from Facebook's Sentiment Analysis in Indonesian yields a different value. The accuracy from the 479 data evaluated is 87.1%, while the error is 12.9%. Meanwhile, an accuracy of 5% and a 95% error rate were acquired from 20 test data [2]. The Sentiment Analysis of Responses to the Lodging Information Service Application then revealed that consumers tended to use the Reddoorz application, which displays positive sentiment of 587 data, negative sentiment of 130 data, and has a data accuracy value of 92.67 percent, for the last six months in 2019. The other applications tested were Agoda, AiryRooms, Oyo, and Reddoorz [3]. Then, using automatic classifications techniques, conduct research on public complaints and reporting through Call Center Service 110. Researchers in this study used a re-selection step to test 33 documents and discovered mistakes in the officers' classification results as a result of their inaccurate use of report category writing. Four of the 33 documents under test—out of 33—were incorrectly classified manually. In this study, 10 papers will be examined in further detail in order to compare the outcomes of system and manual classification [4]. Next, based on 800 tweets, 300 training data, and 500 test data, the research by Nugroho et al. about sentiment analysis on Ojek Online services has the finding that the system can classify sentiment using Nave Bayes with an accuracy of 80% [5].

The collecting of review information or user reviews on the Shopee application before and after the Covid-19 epidemic is described in this study. Review data from February 2021 and April 2021 were collected, with a sample size of 1000 reviews per month. technique Data retrieval is done through the process of scraping. technique Scraping is the process of extracting data from a website, after which the data is often saved in a certain format. Using the Nave Bayes Classifier approach, the author attempts to categorize user review texts in order to determine which reviews are favorable and unfavorable.

2. Methods

Collected data for this research's implementation involves data scraping. The data will be calculated for the presentation of accuracy and sensitivity using the Nave Bayes Classifier method, after which the data will be processed at the text step to obtain the word calculating in the review, which serves to determine whether the data is included in a positive or negative classification. Data collected, text processing, word weighting, application of the Naive Bayes classifier algorithm, and accuracy calculations are among the stages of the research methodology.

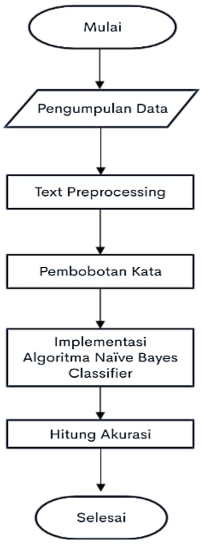
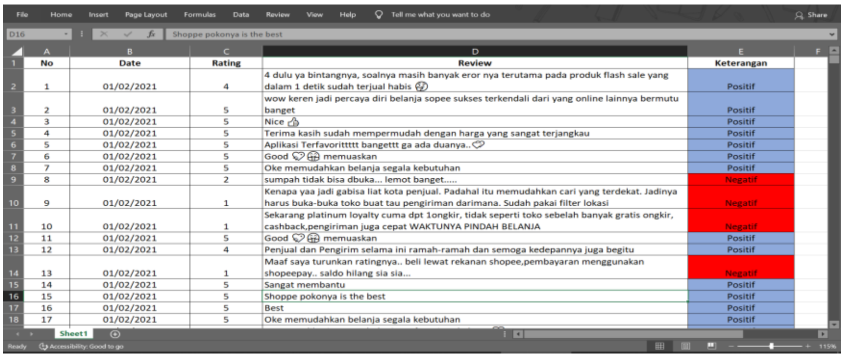


Figure 1. Flowchart for Research

2.1 Data Collect

The method used to get the data for this study was called web scraping, which is the process of scraping data from an online resource that makes data scraping options available. The information will then be saved in CSV format into a file.

The information gathered consists of comments made by Shopee application users in February 2021, before the covid epidemic, and in April 2021, during the covid pandemic. In this study, training data and test data are the two types of data required. Additionally, the sentiment class was manually assigned to the data in use. 1000 data were collected for the review in February, and 1000 more in April. Manually choosing, copying, and saving the review data into Microsoft Excel were all done by hand. Figure 2 displays the outcomes of the review data.


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
Tell me what you want to do
Share
D16
Shoppe pokonya is the best
Sheet1
Ready
Accessibility: Good to go
115%

	A	B	C	D	E
1	No	Date	Rating	Review	Keterangan
2	1	01/02/2021	4	4 dulu ya bintangnya, soalnya masih banyak eror nya terutama pada produk flash sale yang dalam 1 detik sudah terjual habis 😂	Positif
3	2	01/02/2021	5	wow keren jadi percaya diri belanja sopee sukses terkendali dari yang online lainnya bermutu banget	Positif
4	3	01/02/2021	5	Nice 👍	Positif
5	4	01/02/2021	5	Terima kasih sudah mempermudah dengan harga yang sangat terjangkau	Positif
6	5	01/02/2021	5	Aplikasi Terfavorittttt bangettt ga ada duanya..💙	Positif
7	6	01/02/2021	5	Good 💙🤗 memuaskan	Positif
8	7	01/02/2021	5	Oke memudahkan belanja segala kebutuhan	Positif
9	8	01/02/2021	2	sumpah tidak bisa dbuka... lemot banget.....	Negatif
10	9	01/02/2021	1	Kenapa yaa jadi gabisa liat kota penjual. Padahal itu memudahkan cari yang terdekat. Jadinya harus buka-buka toko buat tau pengiriman darimana. Sudah pakai filter lokasi	Negatif
11	10	01/02/2021	1	Sekarang platinum loyalty cuma dpt 1ongkir, tidak seperti toko sebelah banyak gratis ongkir, cashback,pengiriman juga cepat WAKTUNYA PINDAH BELANJA	Negatif
12	11	01/02/2021	5	Good 💙🤗 memuaskan	Positif
13	12	01/02/2021	4	Penjual dan Pengirim selama ini ramah-ramah dan semoga kedepannya juga begitu	Positif
14	13	01/02/2021	1	Maaf saya turunkan ratingnya.. beli lewat rekanan shopee,pembayaran menggunakan shopeepay.. saldo hilang sia sia...	Negatif
15	14	01/02/2021	5	Sangat membantu	Positif
16	15	01/02/2021	5	Shoppe pokonya is the best	Positif
17	16	01/02/2021	5	Best	Positif
18	17	01/02/2021	5	Oke memudahkan belanja segala kebutuhan	Positif

Figure 2. Scraping Review Results

2.2 *Text Processing*

Reviews in other languages have also been included in the collected review data, some of which use unusual sentence or text constructions. The application chooses the data to be processed for each document at the Text Preprocessing stage. Figure 3 below depicts the stages of text preparation.

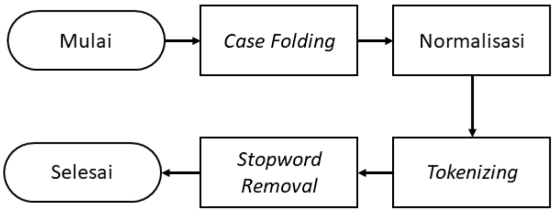


Figure 3. Text-processing phases

In the case folding procedure, the letters in the review sentence are uniformly reshaped into lower case letters. Punctuation marks and extra space characters can also be eliminated via case folding. It is advantageous in this procedure if uppercase and lowercase letters are not recognized as having distinct meanings. Additionally, normalizing or cleaning text is helpful for fixing abbreviated or misspelled words that have specific forms with the same intent. For instance, the word "no" has numerous ways to be written, including no, no, no, and so on. Similarly, misspellings of the word "available" include available, available, available, and so on. In this step, emoticons are also removed from the review data and reviews that contain foreign languages are converted into Bahasa.

Tokenizing, also known as tokenization, is the process of breaking up the text of a document into separate, unrelated word groups called tokens. This procedure simplifies the processes of determining a word's presence in a document and determining how frequently it appears in the corpus. As a filter, the Stopword Removal. the selection of key words from the token results, or the words that serve as the document's representation. Stopwords are used to get rid of words that don't add anything or take away from the information in the document yet are frequently present.

2.3 Word Weighting

Term Frequency (TF) refers to how frequently a word or term appears in the given material. A term's weight increases with the number of times it appears in the document (high TF). Word frequency analysis, or word weighing, is the process of calculating the value of each word versus the frequency of words in the document.

Data that has undergone preprocessing must take the form of numbers. Employing the TF-IDF weighting method to transform the data into a numeric format. The Term Frequency Inverse Document Frequency (TF-IDF) approach assigns a weight to each word in order to gauge how closely related the words are to the document. The TF-IDF method combines the inverse frequency of the document including the word with the frequency of occurrence of the term in the document. The TF value per word is determined first when determining the weights using the TF-IDF, with the weight of each word being 1. Although Equation generates the IDF value (1):

$$\mathrm{IDF}(word) = \log \frac{td}{df} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

2.4 Naive Bayes Classifier Algorithm Implementation

Two types of data classification training data and test data are required to implement the naive Bayes approach. A classifier model is created using the training data by making predictions for the new data class. In order to determine the outcomes of label data classification, measurements of the training data on the classifier model are made using the test data. When it comes to supervised learning or data sets with labels or classes, the Naive Bayes algorithm uses statistical science and probability theory to solve problems. The Nave Bayes Classifier procedure consists of two parts, including the creation of training data with an existing label and the creation of test data to discover new labels utilizing the outcomes of model calculations on training data via equation (2)

$$\mathrm{P}(V_j)=\frac{doc\ j}{training}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Description:

P(Vj) : document likelihood versus a word class or category.
doc j : the proportion of documents that are positive or negative in the training data for category or class "j".
training : the quantity of training data documents..

After that, use equation to calculate the probability value of each word in a document based on the probability value of the document compared to the class of documents (3)

$$\mathrm{P}(xi|Vj)=\frac{nk+1}{n+|kosa\ kata|}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

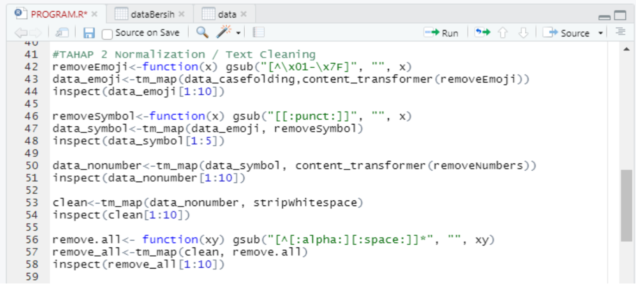
Description:

P(xi|Vj) : the probability that the word "xi" will appear in a document based on Vj.
nk : the quantity of times the word "xi" appears in the document in the group or class of documents.
Vj n : the document's overall word count in the Vj category or class of documents.
|kosakata| : how many words are in the training data.

3. Results and Discusion

3.1 *Text Processing*

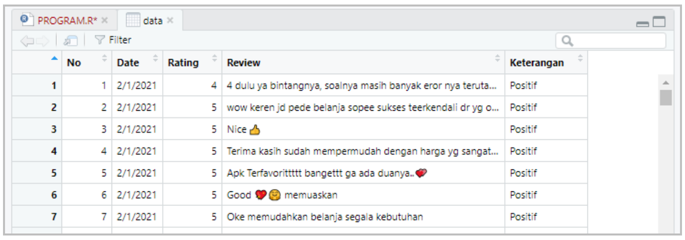
The RStudio application is used to complete each step of Text Processing while taking the author's research findings. The normalization stage of the Shopee application review data between February and April 2021 is where the text processing process begins. Figure 4 displays the program in the Rstudio application for the normalization stage.



```
#TAHAP 2 Normalization / Text Cleaning
removeEmoji<-function(x) gsub("[^\x01-\x7F]", "", x)
data_emoji<-tm_map(data_casefolding,content_transformer(removeEmoji))
inspect(data_emoji[1:10])

removeSymbol<-function(x) gsub("[[:punct:]]", "", x)
data_symbol<-tm_map(data_emoji, removeSymbol)
inspect(data_symbol[1:5])

data_nonumber<-tm_map(data_symbol, content_transformer(removeNumbers))
inspect(data_nonumber[1:10])

clean<-tm_map(data_nonumber, stripWhitespace)
inspect(clean[1:10])

remove.all<- function(xy) gsub("[^[:alpha:][:space:]]*", "", xy)
remove_all<-tm_map(clean, remove.all)
inspect(remove_all[1:10])
```

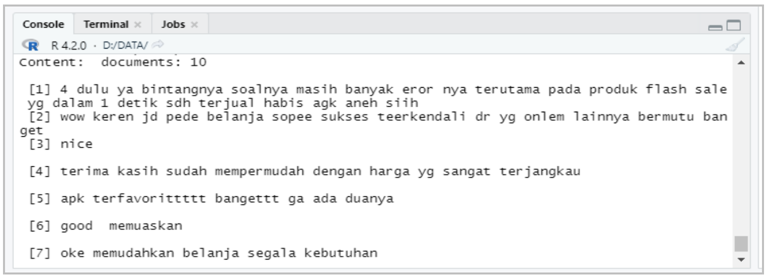
Figure 4. Normalization and Cleaning Text Program

This step's goal is to eliminate emoticons from the review data. Figure 5 shows the review data before normalization, with the emoticon still present. Figure 6 displays the normalization results after the computer has processed the data to eliminate emoticons during the review.



	No	Date	Rating	Review	Keterangan
1	1	2/1/2021	4	4 dulu ya bintangnya, soalnya masih banyak eror nya teruta...	Positif
2	2	2/1/2021	5	wow keren jd pede belanja sopee sukses teerkendali dr yg o...	Positif
3	3	2/1/2021	5	Nice 👍	Positif
4	4	2/1/2021	5	Terima kasih sudah mempermudah dengan harga yg sangat...	Positif
5	5	2/1/2021	5	Apk Terfavoritttttt bangettt ga ada duanya..❤	Positif
6	6	2/1/2021	5	Good ❤🤗 memuaskan	Positif
7	7	2/1/2021	5	Oke memudahkan belanja segala kebutuhan	Positif

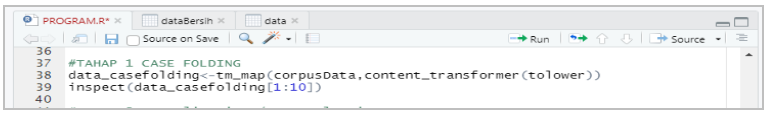
Figure 5. Before Normalization the Data



```
Content:  documents: 10

 [1] 4 dulu ya bintangnya soalnya masih banyak eror nya terutama pada produk flash sale
 yg dalam 1 detik sdh terjual habis agk aneh siih
 [2] wow keren jd pede belanja sopee sukses teerkendali dr yg onlem lainnya bermutu ban
get
 [3] nice

 [4] terima kasih sudah mempermudah dengan harga yg sangat terjangkau

 [5] apk terfavoritttttt bangettt ga ada duanya

 [6] good  memuaskan

 [7] oke memudahkan belanja segala kebutuhan
```

Figure 6. After Normalization the Data

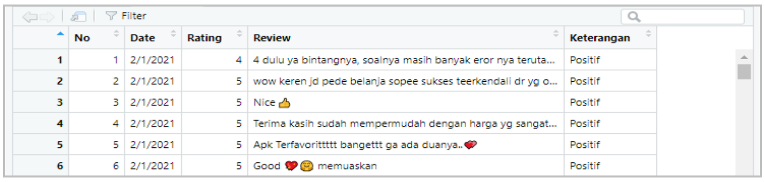
After that, the review data that was submitted to the RStudio application will have all of the letters fully reduced by this case folding. Figure 7 showa this program for letter reduction.



```

#TAHAP 1 CASE FOLDING
data_casefolding<-tm_map(corpusData,content_transformer(tolower))
inspect(data_casefolding[1:10])

```

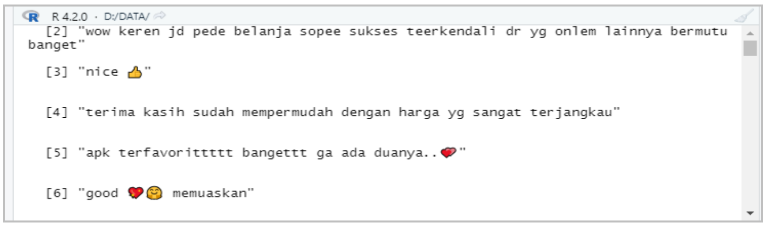
Figure 7. Case Folding Program

The lowercase letters in the review data can be modified, and the case folding method is effective in preventing the detection that uppercase and lowercase characters have different meanings. Figure 8 shows the data before case folding, and Figure 9 shows the outcome, with all the letters in the review being converted to lowercase.



	No	Date	Rating	Review	Keterangan
1	1	2/1/2021	4	4 dulu ya bintangnya, soalnya masih banyak eror nya teruta...	Positif
2	2	2/1/2021	5	wow keren jd pede belanja sopee sukses teerkendali dr yg o...	Positif
3	3	2/1/2021	5	Nice 👍	Positif
4	4	2/1/2021	5	Terima kasih sudah mempermudah dengan harga yg sangat...	Positif
5	5	2/1/2021	5	Apk Terfavoritttttt bangettt ga ada duanya..❤	Positif
6	6	2/1/2021	5	Good ❤🤗 memuaskan	Positif

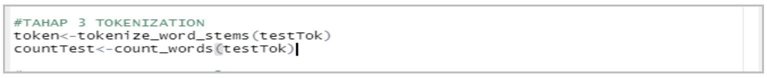
Figure 8. Before Folding Cases Data Reviews



```
  [2] "wow keren jd pede belanja sopee sukses teerkendali dr yg onlem lainnya bermutu
banget"

  [3] "nice 👍"

  [4] "terima kasih sudah mempermudah dengan harga yg sangat terjangkau"

  [5] "apk terfavoritttttt bangettt ga ada duanya..❤"

  [6] "good ❤🤗 memuaskan"
```

Figure 9. After Folding Cases Data Reviews

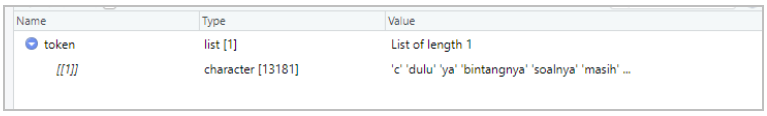
The next step is tokenizing, which is used to obtain word fragments or tokens that will later develop into entities useful for creating the document matrix in the following procedure. The 2000 review data has been tokenized in this study by counting the amount of words that have been gathered. Figure 10 depicts the tokenization of computer code.



```
#TAHAP 3 TOKENIZATION
token<-tokenize_word_stems(testTok)
countTest<-count_words(testTok)
```

Figure 10. Tokenizing Review Program

13,181 words from the submitted review data have been automatically calculated in the RStudio program; the results are shown in Figure 11.



Name	Type	Value
token	list [1]	List of length 1
[[1]]	character [13181]	'c' 'dulu' 'ya' 'bintangnya' 'soalnya' 'masih' ...

Figure 11. Result Tokenizing Data

The next step, called known as filtering, involves deleting words from the text or condensing word lengths in the corpus to create stopwords. In this investigation, we were able to get rid of words that didn't affect the review.

3.2 Labeling Sentiment by Class

The number of positive reviews had a higher frequency in this study's sentiment class classification results than the number of negative reviews. Positive reviews make up 7,216 of the 8,712 overall reviews, while negative reviews make up 1,496 of the reviews. Figure 12 shows a comparison of the quantity of sentiment data for Shopee application reviews based on negative and positive categories in February and April 2021.

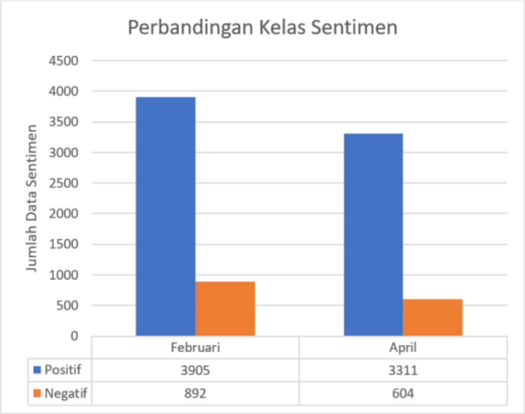


Figure 11. Comparison of the Number of Classes in Graphical Data

3.3 The Nave Bayes Method for Classification

In the classification process, machine learning is used to carry out the classification process using training data and random test data with iterations of experiments. To achieve the best possible prediction accuracy value, cross-validate four times on the dataset. Calculating training and test data using a comparison of 60:40, 70:30, 80:20, and 90:10 on the February and April review data is how iterations are carried out. Figures 4.12 and 4.13 show accurate and sentimental data.

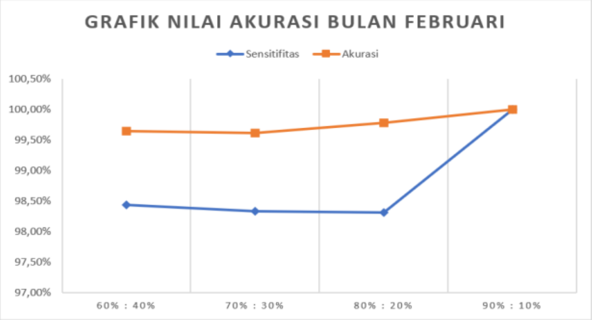


Figure 12. February Accuracy Of The Graph

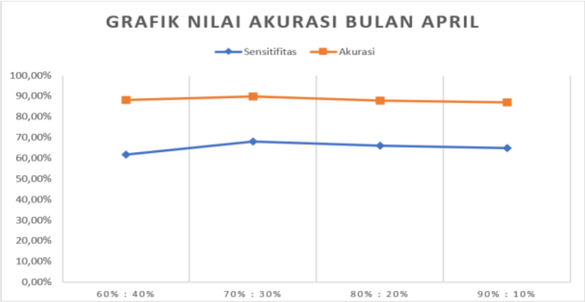


Figure 13. February Accuracy Of The Graph

In February, the difference between accuracy and sensitivity levels is less than 2%, but in April, the difference is more than 20%. Table 1 shows the data accuracy and sensitivity values for the months of February and April.

Table 1. Accuracy Results Using the Naïve Bayes Method

		February			
Perbandingan	60% : 40%	70% : 30%	80% : 20%	90% : 10%	Rata-rata
Akurasi	99,64%	99,61%	99,78%	100%	99,76%
Sensitifitas	98,44%	98,33%	98,31%	100%	98,77%
		April			
Perbandingan	60% : 40%	70% : 30%	80% : 20%	90% : 10%	Rata-rata
Akurasi	87,98%	89,71%	87,94%	87,05%	88,17%
Sensitifitas	61,79%	68%	65,98%	64,84%	65,15%

The accuracy results using the Nave Bayes Classifier (NBC) method have significant variations in accuracy, with an average accuracy value of 99.76 % for February data, as shown in Table 1. The Nave Bayes Classifier (NBC) method is suitable for classifying reviews on Shopee that use Bahasa with multiclass subjects, with an average accuracy value of 99.76 %.

3.4 Association and Visualization

The data from the labeling performed using the lexicon dictionary in the RStudio application is the review data on the positive classification use. The following is a depiction of the information extracted from visitor reviews with positive classifications for February and April 2021 in the Shopee application. The reviews are recognized based on the frequency of words in the reviews. Figure 14 shows words that frequently appear in the data.

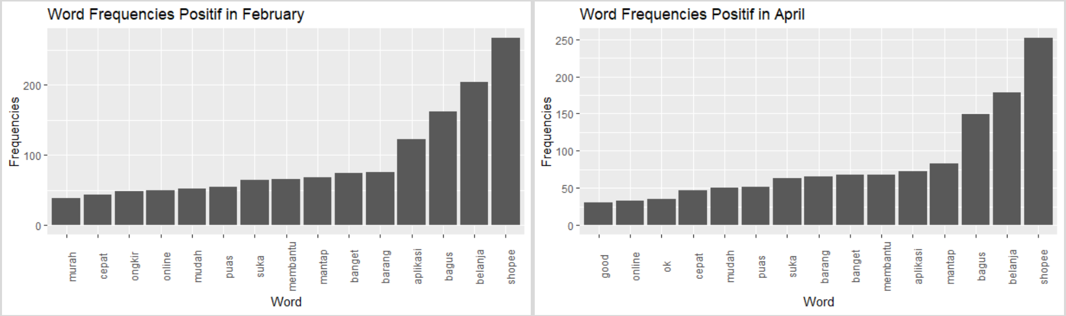


Figure 14. Graph Of Frequently Used Positive Classification Words

To get better information, these words are then applied as a platform for identifying relationships with additional terms. Shown in Figure 15, the group of words that recur frequently can also be visualized as a wordcloud.

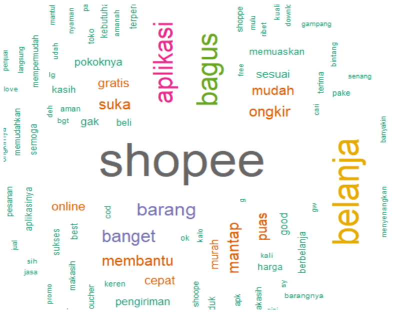


Figure 15. Positive Review Wordcloud

The information has been collected from visitor reviews with negative classifications for February and April 2021 in Shopee application reviews, and the findings are visualized below. Figure 16 shows the words that frequently appear in the data.

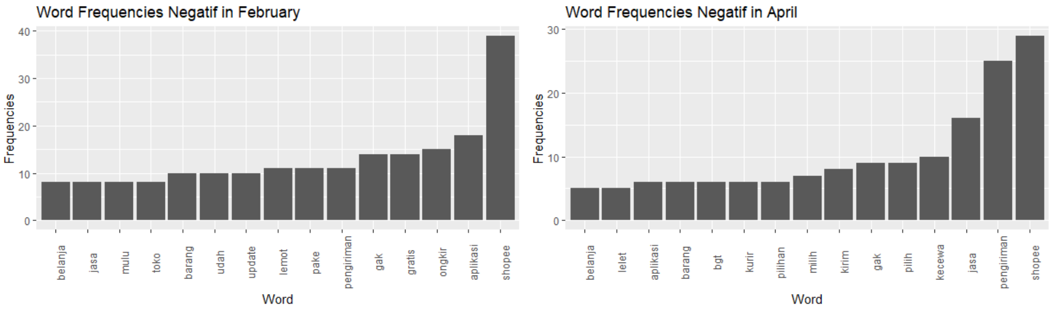


Figure 16. Graph Of Frequently Used Negative Classification Words

To get better information, these words are then used as a platform for identifying relationships with additional terms. As shown in the Figure 17, the list of terms that recur frequently can also be presented as a word cloud.

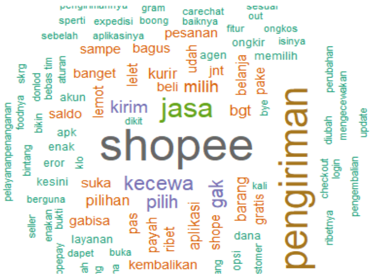


Figure 17. Negative Review Wordcloud

4. Concluison

Based on the results of this study, it can be said that Sentiment Analysis Using the Naive Bayes Classifier Algorithm has been successfully implemented and that text processing, word weighting, and the process of collecting review data on the Shopee application have gone well. The text processing steps, which included case folding, normalization, tokenization, and stopword removal on review data for February and April 2021, went successfully using the RStudio app. The data sentiment labeling process went smoothly, obtaining positive classification data for 7,216 words and negative classification data for 1,496 words.

According to the labeling results, there are 7,216 more positive reviews than negative reviews on the Shopee application, through a total of 8,712 reviews. This shows that a large number of people have positive views of the Shopee app. The accuracy of the classification using the Naive Bayes Classifier approach ranges from 88.17 % in the review data from April to 99.76 % in the data from February. Since the accuracy results are greater than 88 percent after the pandemic and greater than 99 percent before the pandemic, the application of the Naive Bayes Classifier method to Shopee application review data before and after the pandemic is a wise choice.

References

[1] R. Apriani *et al*., "ANALISIS SENTIMEN DENGAN NAÏVE BAYES TERHADAP KOMENTAR APLIKASI TOKOPEDIA," 2019.

[2] P. S. M. Suryani, L. Linawati, and K. O. Saputra, "Penggunaan Metode Naïve Bayes Classifier pada Analisis Sentimen Facebook Berbahasa Indonesia," *Majalah Ilmiah Teknologi Elektro*, vol. 18, no. 1, p. 145, May 2019, doi: 10.24843/mite.2019.v18i01.p22.

[3] H. Februariyanti, M. Firmansyah, J. S. Wibowo, and M. S. Utomo, "Terakreditasi 'Peringkat 4 (Sinta 4)' oleh Kemenristekdikti," vol. 6, no. 2, pp. 1–5, doi: 10.5281/zenodo.4399381.

[4] F. Handayani, D. Feddy, and S. Pribadi, "Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110."

[5] D. Garbian Nugroho, Y. Herry Chrisnanto, A. Wahana Jurusan Informatika, and F. Matematika dan Ilmu Pengetahuan Alam Universitas Jenderal Achmad Yani Jalan Terusan Jenderal Sudirman, *ANALISIS SENTIMEN PADA JASA OJEK ONLINE MENGGUNAKAN METODE NAÏVE BAYES*. 2016.

[6] RStudio, "Download RStudio IDE," *rstudio.com*, 2022. https://www.rstudio.com/products/rstudio/download/ (accessed May 25, 2022).

[7] Stackoverflow, "Remove all special characters from a string in R," *stackoverflow.com*, 2021. https://stackoverflow.com/questions/10294284/remove-all-special-characters-from-a-string-in-r (accessed May 27, 2022).

[8] Nur Andi Setiabudi, "Stemming Bahasa Indonesia dengan R," *nurandi.id*, 2015. https://www.nurandi.id/blog/katadasar-stemming-bahasa-indonesia-dengan-r/ (accessed Jun. 15, 2022).

[9] Yann Ryan, "Calculating tf-idf Scores with Tidytext," *bookdown.org*, 2021. https://bookdown.org/yann_ryan/r-for-newspaper-data/calculating-tf-idf-scores-with-tidytext.html (accessed May 27, 2022).

[10] R Project, "Strip Whitespace from a Text Document," *r-project.org*, 2021. https://search.r-project.org/CRAN/refmans/tm/html/stripWhitespace.html (accessed May 27, 2022).

[11] John McIntosh, "Creating Word Clouds Vignette," *rpubs.com*, Apr. 03, 2017. https://rpubs.com/Johnmac1967/265334 (accessed May 27, 2022).

[12] Michael W. Kearney, "Wordcloud," *https://rpubs.com/*, Aug. 23, 2015. https://rpubs.com/mkearney/104366 (accessed May 27, 2022).

[13] RColorBrewer, "R/ColorBrewer.R," *https://rdrr.io/*, Apr. 04, 2022. https://rdrr.io/cran/RColorBrewer/src/R/ColorBrewer.R (accessed May 28, 2022).

[14] sudhanshublaze, "Filter data by multiple conditions in R using Dplyr," *https://www.geeksforgeeks.org*, Jan. 25, 2022. https://www.geeksforgeeks.org/filter-data-by-multiple-conditions-in-r-using-dplyr/ (accessed May 28, 2022).

[15] Suhartono and U'un Setiawati, "Metode Klasifikasi Naive Bayes, Random Forest dan Decicion Tree untuk Memprediksi Kanker Payudara Menggunakan Rstudio," *https://rpubs.com*, Jan. 12, 2021. https://rpubs.com/uuns/klasifikasi (accessed May 30, 2022).

[16] RB Fajriya Hakim, "Contoh Sederhana Aplikasi Naive Bayes dengan R," *https://medium.com*, Sep. 23, 2019. https://medium.com/@986110101/naive-bayes-classifier-65422fa14362 (accessed May 30, 2022).

[17] genediazjr, "remaining stopwrods from second batch," *https://github.com*, Oct. 10, 2016. https://github.com/stopwords-iso/stopwords-id (accessed May 31, 2022).

[18] Nur Andi Setiabudi, "Stemming Bahasa Indonesia dengan R," *https://www.nurandi.id*, Dec. 16, 2015. https://www.nurandi.id/blog/katadasar-stemming-bahasa-indonesia-dengan-r/ (accessed May 31, 2022).