International Journal Software Engineering and Computer Science (IJSECS) 3 (3), **2023, 194-204**
Published Online December 2023 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN : 2776-4869, E-ISSN : 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v3i3.1699.

The Shopee Application User Reviews Sentiment Analysis Employing Naïve Bayes Algorithm

Nur Adha Pasaribu *
Universitas Islam Negeri Sumatra Utara, Deli Serdang Regency, North Sumatra Province, Indonesia.
Email: nur03adha@gmail.com

Sriani
Universitas Islam Negeri Sumatra Utara, Deli Serdang Regency, North Sumatra Province, Indonesia.
Email: sriani@uinsu.ac.id

Received: 25 October 2023; Accepted: 14 November 2023; Published: 1 December 2023.

Abstract: With the significant growth of internet use in Indonesia, there has been a surge in online business activity. The convenience offered by online platforms is increasingly in demand because it allows consumers to shop without being bound by a certain time or location. Before making a purchase, consumers tend to look for information first through various sources such as reviews on blogs, Instagram, TikTok, or reviews on the YouTube platform which is integrated in the application. This research adopted a method that included planning, literature study, data collection, and data processing using a dataset from the Play Store application which was taken using the Python library with an initial amount of 5000 data. After a manual filtering process which involved removing slang words, eliminating duplications, and normalizing punctuation marks, the remaining data was 3946. The application of the Naïve Bayes algorithm in this research uses probability methods to classify and predict 3141 training data and 805 test data, with Python library help. The accuracy calculation results show satisfactory performance, with an accuracy of 86.00%, precision of 80.74%, recall of 78.13%, and f1-score of 79.00% on a dataset of 3946. Analysis from this research shows the dominance of positive sentiment in 2050 data, while sentiment negative amounted to 1199 data. The amount and quality of training data plays an important role in system predictions, where high data quality provides better accuracy in predicting sentiment classes.

Keywords: Sentiment Analysis; Shopee; Naive Bayes.

1. Introduction

In the current Modern Era in Indonesia, technological advancements, particularly in the field of information and telecommunications, have progressed rapidly. This has significantly facilitated consumers in various aspects, from shopping to ordering food and transportation, all within the grasp of a smartphone. With the increasing prevalence of the internet in Indonesia, a growing number of businesses are venturing into the online business realm. The convenience they offer is highly sought after by the public, allowing them to shop online from the comfort of their homes or offices, overcoming constraints of time and location. Before consumers make purchases on online platforms, they often seek information through various means, such as reviewing content accessible through blogs, Instagram, TikTok, and YouTube. These platforms serve as valuable sources for specifications, advantages, and disadvantages, as well as pricing details of products. Consumers need to gather comprehensive information about the products they intend to purchase to minimize potential negative impacts. To obtain this information, they rely on reviews, ratings, and star ratings commonly found on Shopee, a prominent online marketplace.

Online customer reviews constitute feedback provided by consumers regarding product evaluations covering various aspects. This information allows consumers to assess the quality of products based on the experiences shared by others who have purchased from the same online seller. Online customer reviews function as a form of word-of-mouth communication in the realm of online sales, where potential buyers gather insights about a product from those who have derived benefits from it. Therefore, information becomes a crucial determinant in decision-making, with consumer reviews and ratings playing a pivotal role. Online customer ratings are part of the product review system, expressed not in text but through star symbols. These stars symbolize customers' opinions about a product, reflecting their psychological and emotional experiences when interacting with virtual products in mediated environments. A higher star rating indicates a better ranking for the seller, as assigned by Shopee buyers who have made online purchases and published their reviews on the seller's website or storefront. Ratings are integral as a benchmark for product quality.

The Naïve Bayes method is a classification method in machine learning that excels in using training data samples to estimate parameters involved in the classification process rapidly, resulting in high accuracy. Various methods and algorithms have been proposed in related research on Naïve Bayes classification. For instance, Muhammad Muslimin achieved a Naïve Bayes accuracy rate of 94.38%, while Tri Herdiawan Apandi achieved 91.33%, and Didik Garbian Nugroho achieved 80.00% [1]. In a similar study, Muhammad Muslimin *et al*. (2023) conducted research titled "Sentiment Analysis of Rising Basic Commodity Prices Using Naïve Bayes Classifier," concluding that Sentiment analysis of rising basic commodity prices garnered negative responses from the public. The Naïve Bayes method used in the analysis achieved an accuracy of 94.38% [1]. Another study by Didik Garbian Nugroho *et al*. (2016) titled "Sentiment Analysis on Online Motorcycle Taxi Services Using Naïve Bayes" achieved an 80% accuracy rate using Sentiment data from Twitter, consisting of 3000 sentences categorized into 1000 neutral, 1000 positive, and 1000 negative sentences [2]. Naïve Bayes is one of the most effective and efficient machine learning and data mining algorithms. Despite the assumption that attributes in the data are independent, Naïve Bayes classification performance remains high. Although the assumption of attribute independence is rarely met in actual data, the algorithm can still produce good classification results. Naïve Bayes is the implementation of a probabilistic machine learning algorithm using Bayes' theorem calculation techniques [3].

Previous research by Ahmad *et al*. (2023) conducted Sentiment analysis on COVID-19 vaccine opinions on Twitter social media using the Naïve Bayes and Decision Tree algorithms. The research results show that in general, the public responds positively to the Indonesian government's vaccination policy [4]. Apandi and Sugianto (2019) applied the Naïve Bayes algorithm to predict satisfaction with e-KTP recording services. It was found that Naïve Bayes had an accuracy of 91.70%, while Decision Tree had an accuracy of 65.90% [5]. Aulia *et al*. (2023) conducted Sentiment analysis on e-Tilang via YouTube social media using the Naïve Bayes algorithm. This research resulted in an accuracy of 79.44% [6]. Budiman (2021) compared data mining classification algorithms for tracking the interests of prospective new students and found that the J48 Decision Tree had the highest accuracy, namely 90.3% [7]. Cahyaningtyas *et al*. (2021) conducted Sentiment analysis on the Shopee application rating using the SMOTE-based Decision Tree method, with an accuracy of 99.91% [8]. Depari *et al*. (2022) compared Decision Tree, Naïve Bayes, and Random Forest models for predicting heart disease classification, with Random Forest achieving the highest accuracy, namely 0.75% [9].

Research has revealed diverse applications and comparisons between different classification methods in various contexts. One study focused on improving sales performance with decision tree and naïve Bayes methods, finding that decision tree achieved 100% accuracy while naïve Bayes only 41% [10]. On the other hand, other research such as fish freshness classification using K-Nearest Neighbor and Hue Saturation Value managed to achieve 93% accuracy [11], while in Sentiment analysis of user reviews on Shopee, the Naïve Bayes algorithm gave high accuracy results of 99.5 % [12]. In addition, a comparison between the C4.5 and Naïve Bayes algorithms in predicting student graduation shows different performance, with C4.5 having an accuracy of 76.69% and Naïve Bayes of 72.95% [15]. Another study that compared the Naïve Bayes algorithm with Decision Tree in classifying thesis titles showed that the accuracy of Naïve Bayes was 80.33% and Decision Tree was 60.33% [16]. Not only that, but there is also a study on predicting company bankruptcy by comparing decision trees and naïve Bayes, where naïve Bayes achieved 100% accuracy [17]. A comparison between the C4.5 and Naïve Bayes algorithms for predicting the timeliness of student studies shows that Naïve Bayes has higher accuracy with 69.97% compared to C4.5 which only reaches 61.99% [18]. Furthermore, in the context of Sentiment classification for Starbucks customer satisfaction, decision trees show better performance with an accuracy of 83% compared to naïve Bayes which only has an accuracy of 74% [19]. Finally, research comparing the naïve Bayes and decision tree algorithms in classifying outstanding lecturers shows that the decision tree has an accuracy of 95.80% while naïve Bayes reaches 94.80% [20].

From the analysis of previous research, it appears that the choice of classification algorithm has a significant impact on the accuracy of results in various research contexts. Each method, from decision trees to naïve Bayes, has unique advantages and limitations in data processing and classification. With these various insights, the author will carry out a research study on Sentiment Analysis on User Reviews of the Shopee Application Using the Naïve Bayes Algorithm.

2. Research Method

The initial stage carried out in this research was research planning. This research was carried out by applying the Naïve Bayes method to user reviews of the Shopee application, and produced high and good method accuracy. The next stage is to carry out a literature study by reading and understanding several previous studies, whether in the form of journals, articles, theses or theses which are related to the research currently being carried out. After the literature study stage, the next stage is discussion and consultation with the lecturer. supervisor to get direction regarding the research to be carried out. At this stage, the process of collecting user review data regarding the Shopee application is carried out on the Google Play Store application using the python library as a data scraper. The amount of data used is 5000 data, which is then divided into 80% training data and 20% testing data. This means that 800 of the data is the training data group, and the rest is testing data.

The data processing process in Google Colab includes several steps, starting from the dataset, Preprocessing, dividing the data into two parts, namely testing and training data. The data processing process carried out will produce a result

which will be discussed and produce conclusions in the research process carried out. So, in this research method, the first step is to input data where the data that has been obtained through the Python Library as a data scraper in the Play Store application is 5000 data. After data importing, the next stage is data preprocessing, which is the step of obtaining relevant training data so that it is ready for use. The software used in preprocessing the data is Google Colab using the Python library as a storage area or library. After preprocessing the data, the results are obtained. After the results are obtained, the TF-IDF weighting stage is carried out again. This stage is carried out to calculate the word weights. TF-IDF is a calculation to measure how important words are in documents and corpus. After the TF-IDF weighting stage is complete, the next stage is the classification stage using the Naïve Bayes method. The data that is classified using the Naïve Bayes method is the preprocessing data that has been labeled according to the Sentiment of the opinion data. After the classification results of the Naïve Bayes method are obtained, it is complete.

3. Result and Discussion

3.1 Results

In this research the author will analyze sentiment data on user reviews of the Shopee application on the Google Play Store. The first stage of research is collecting data. The process of collecting user review data regarding the Shopee application is carried out on the Google Play Store application using the python library as a data scraper. The data scrapping process carried out in this research does not depend on a specific time span. The scrapping data is the latest review of the Shopee application. The results of the scrapping process will be converted into a .csv extension and then processed using the Python programming language. The amount of data obtained from scrapping results is 5000 of the latest review data on the Shopee application.

3.1.1 Preprocessing Data

Scraping data must go through a preprocessing stage before it can be analyzed. The preprocessing stage aims to change a dataset that is initially unstructured and has a lot of noise into clean data that is ready to be processed. The scrapping results consist of the review_id, userName, userImage, thumbsUpCount, at, score and content columns. For the sentiment analysis process, only the content and score columns are needed. Where content is a column containing reviews from users while score is a rating given by users. For this reason, data structures and Sentiment structures will be cleaned on all data by deleting non-informative data. By using 5 training data samples and 1 test data. The following is a flow that explains the system stages during the data preprocessing process which includes case folding, filtering, tokenizing, slang word conversion, stopwords, stemming.

1) Case folding
   a. Description: This step involves converting uppercase letters to lowercase in all Sentiments. The process includes reading each line of data, detecting the presence of capital letters, and converting all detected capital letters to lowercase.
   b. Duplication Removal: This step aims to eliminate repetitive occurrences of words within Sentiments. It detects and handles word repetitions in all Sentiments. For example, the phrase might be transformed into to minimize redundancy.
2) Filtering
   The filtering stage is the stage of removing unnecessary characters. For example, numbers, symbols, emojis and punctuation. The following are the case folding steps. Read each line of data and detect whether there are letters, numbers, symbols, emojis and punctuation. Delete all numbers, symbols, emojis and text in the form of links and punctuation that are detected.
   a. Removal of links, hashtags and emojis on Sentiments
   b. Elimination of numbers on Sentiment
   c. Removal of punctuation on Sentiment
3) Tokenizing
   Tokenization is the process of breaking down text into smaller units, which are usually referred to as "tokens". Tokenization is important in Sentiment analysis and natural language processing (NLP) in general, as it helps in preparing raw text for further analysis.
4) Slangword conversion
   Slang words in the context of Sentiment analysis refer to words or phrases used in informal conversations and often have special meanings or connotations in certain circles. These slang words may not be standard or official, but are often used in everyday interactions, especially in communication on social media, online chats, or informal conversations. To interpret these slang words, a slang dictionary is used so that these slang words are converted into real words. For example, 'org' is changed to 'person'.
5) Stopword removal
   Stopwords are common words that appear frequently in text but tend to have no special meaning or significance in natural language analysis or Sentiment analysis. These words are generally ignored or removed from the text when

pre-processing the data before Sentiment analysis is performed. Stopwords usually consist of words such as "and", "or", "in", "to", "the", "this", and the like. These words are very commonly used in the language and appear in almost every text, but rarely have an important contribution to understanding Sentiment or text analysis. Removal of stopwords from text in Sentiment analysis has several benefits such as increased efficiency and emphasis on important words.

6) Stemming
Stemming is a morphological process in natural language processing that involves removing the beginning or ending of a word to produce the base form or root of a word. The main goal of stemming is to reduce the different variations of words in a text to a more general form, so that words with the same root can be identified as one entity. It helps in Sentiment analysis by reducing the dimensionality of words and identifying words that have the same meaning but in different forms. Examples of stemming "running", "running", "running", "running" will be changed to "running". "eat", "food", "eaten" will be changed to "eat". The aim of the stemming process is to reduce data dimensions, overcome keywords and increase accuracy.

After going through all the cleansing and preprocessing stages there will be the possibility of empty cells in the table due to the process selecting all the text in the cells. Therefore, checking and deleting rows is carried out if there are empty text cells.

3.1.2 TF-IDF weighting

The next stage after passing the labeling and Sentiment cleaning stages is the TF-IDF (Term Frequency-Inverse Document Frequency) weighting stage, where at this stage the weighting technique for each word (term) in the document data is used, calculated from each word and then each word. will be multiplied by idf. The following is a sample calculation of the TF-IDF value.

Table 1. Training Data Sample

Train Sentiment	Class
['bantu', 'mudah', 'belanja']	Positive
['enak', 'belanja']	Positive
['cepat', 'mudah', 'bayar', 'pay', 'senang']	Positive
['banyak', 'iklan', 'aplikasi', 'jual', 'iklan', 'iklan']	Negative
['aplikasi', 'busuk', 'banyak', 'iklan', 'ganggu']	Negative

Table 2. DF Values from Training Data

Term	TF					DF
	D1	D2	D3	D4	D5	
Bantu	1	0	0	0	0	1
Mudah	1	0	1	0	0	2
Belanja	1	1	0	0	0	2
Enak	0	1	0	0	0	1
Cepat	0	0	1	0	0	1
Bayar	0	0	1	0	0	1
Pay	0	0	1	0	0	1
Senang	0	0	1	0	0	1
Banyak	0	0	0	1	1	2
Iklan	0	0	0	3	1	2
Aplikasi	0	0	0	1	1	2
Jual	0	0	0	1	0	1
Busuk	0	0	0	0	1	1
Ganggu	0	0	0	0	1	1

After the TF (term frequency) value is obtained, the next step is to find the value of the IDF. Below is a formula to determine the IDF value of each word.

$IDF = \log\left(\frac{D+1}{df+1}\right) + 1$ ..... (1)

Table 3. IDF Values from Training Data

Term	DF	IDF
Bantu	1	2.098
Mudah	2	1.693
Belanja	2	1.693
Enak	1	2.098
Cepat	1	2.098
Bayar	1	2.098
Pay	1	2.098
Senang	1	2.098
Banyak	2	1.693
Iklan	2	1.693
Aplikasi	2	1.693
Jual	1	2.098
Busuk	1	2.098
Ganggu	1	2.098

After the TF and IDF values are obtained, the TF-IDF value can be calculated. To find the TF-IDF value, use the equation below.

$W = TF\ x\ IDF$ ..... (2)

Table 4. TF-IDF values from training data

No.	*Term*	TF-IDF				
		D1	D2	D3	D4	D5
1	Bantu	2.098	0	0	0	0
2	Mudah	1.693	0	1.693	0	0
3	Belanja	1.693	1.693	0	0	0
4	Enak	0	2.098	0	0	0
5	Cepat	0	0	2.098	0	0
6	Bayar	0	0	2.098	0	0
7	Pay	0	0	2.098	0	0
8	Senang	0	0	2.098	0	0
9	Banyak	0	0	0	1.693	1.693
10	Iklan	0	0	0	5.079	1.693

11	Aplikasi	0	0	0	1.693	1.693
12	Jual	0	0	0	2.098	0
13	Busuk	0	0	0	0	2.098
14	Ganggu	0	0	0	0	2.098

Next, the TF-IDF values are normalized to equalize the interval of each data. The equation used to normalize the data is as follows.

$$TF_{norm}(t,d) = \frac{TF(t,d)}{\sqrt{\sum_i (TF(t,d))^2}} \quad (3)$$

The following are the results of the data normalization calculations carried out.

Table 5. Data Normalization

No.	D1	D2	D3	D4	D5
1	0.2197	0	0	0	0
2	0.1773	0	0.1773	0	0
3	0. 1773	0.1773	0	0	0
4	0	0.2197	0	0	0
5	0	0	0.2197	0	0
6	0	0	0.2197	0	0
7	0	0	0.2197	0	0
8	0	0	0.2197	0	0
9	0	0	0	0.1773	0.1773
10	0	0	0	0.53182	0.1773
11	0	0	0	0.1773	0.1773
12	0	0	0	0.2197	0
13	0	0	0	0	0.2197
14	0	0	0	0	0.2197

3.1.3 Naïve Bayes Classification

After the data has been cleaned and structured, the next step is to carry out classification using the Naïve Bayes algorithm. The first stage of the classification process is dividing the data into training data and test data. In this study, a ratio of 8:2 was used to compare training data and test data. Training data is used to study the characteristics and differences between the two positive and negative classes, while test data is used to see the percentage of success when classifying correctly. The total initial dataset for Sentiment analysis research on the Shopee application on the Play Store using the Naïve Bayes algorithm was 5000 pieces of data. However, after going through the preprocessing stage, the remaining data amounted to 3946 pieces of data. The ratio of training data to test data in this research is 8: 2. So a total of 3141 pieces of training data will be obtained, while for test data there will be 805 pieces of data. The training process will produce a weight value for each word in each class using the TF-IDF weighting method.

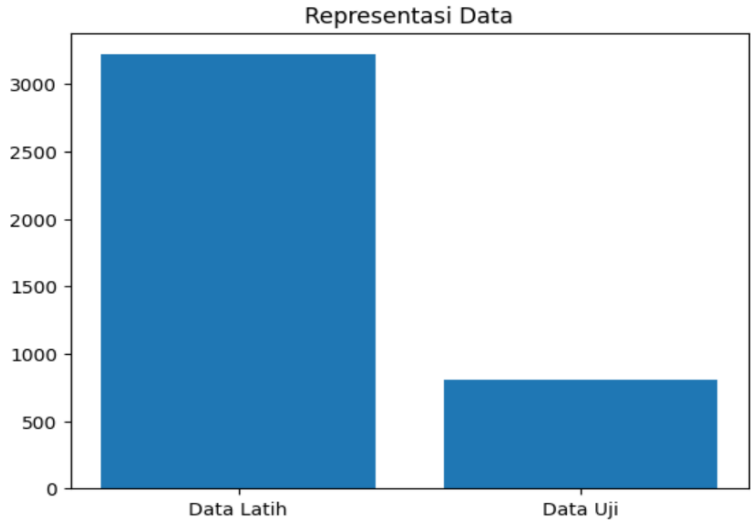

Figure 1. Distribution of training data and test data

By using 5 previous sample data as training data, 1 test data is determined as follows.

Table 6. Sample Test Data

Test Sentiment	Class
['belanja', 'aman', 'senang']	?

Sentiment determination is carried out by calculating the probability of the testing data document by referring to the probability of the training data words. This Sentiment classification is carried out automatically by implementing the Naïve Bayes algorithm. This process is implemented in the MultinomialNB function by comparing each weight in the testing data with words in the training data. The result of each training document is the sum of the positive probability and negative probability word weights. Next, the weights of the test documents are compared. If the positive probability document weight is greater than the Sentiment classification result is positive, whereas if the probability weight is negative then the Sentiment classification result is negative. The stages in class classification begin by calculating prior probability, conditional probability, and posterior probability. The stages of the classification process using the Naïve Bayes algorithm on test data are as follows.

1) Calculation of prior probability values

$$P\ (Kelas\ |Sentiment) = \frac{Number\ Classes\ X}{Number\ of\ Sentiments} \quad (4)$$

Using the equation above, we will obtain the probability of each class in Sentiment.

a. P (Positive | Sentiment) $= \frac{3}{5} = 0.6$

b. P (Negative | Sentiment) $= \frac{2}{5} = 0.4$

2) Calculation of conditional probability values

$$P\ (Term\ |Class) = \frac{Total\ TF{-}IDF\ Term\ Weight\ in\ Class{+}1}{TF{-}IDF\ Weight\ Class{+}\ Total\ TF{-}IDF\ Weight} \quad (5)$$

Using the equation above, we will obtain the probability of the terms in each class in Sentiment.

a. Positive

P (belanja | Positive) $= \frac{0.3546+1}{1.8501+3.927} = \frac{1.354}{5.777} = 0.234$

P (aman | Positive) $= \frac{0+1}{1.8501+3.927} = \frac{1}{5.777} = 0.173$

P (senang | Positive) $= \frac{0.2197+1}{1.8501+3.927} = \frac{1.2197}{5.777} = 0.211$

a. Negative

P (belanja | Negative) $= \frac{0+1}{2.077+3.927} = \frac{1}{6.004} = 0.167$

P (aman | Negative) $= \frac{0+1}{2.077+3.927} = \frac{1}{6.004} = 0.167$

P (senang | Negative) $= \frac{0+1}{2.077+3.927} = \frac{1}{6.004} = 0.167$

3) Calculation of posterior probability values

$$P\ (Sentiment|\ Class) =\ P_{term\ 1}\ x\ ..\ x\ P_{term\ n}\ x\ P\ (\ Class|\ Sentiment) \quad (6)$$

By using the equation above, the posterior probability value for each class will be calculated.
P (Sentiment | Positive) = 0.234 * 0.173 * 0.211 * 0.6
= 0.0051250
P (Sentiment | Negative) = 0.167 * 0.167 * 0.167 * 0.4
= 0.0018629

Based on the example calculations carried out, the highest calculated value from all classes and the highest value from the test data on Sentiment Positive with a value of **0.0051250** were selected. So, the classification result for the Sentiment Belanja aman senang' is **Positive**.

3.1.4 Application
This research was carried out with the aim of developing a system that can conduct sentiment analysis and categorizing reviews that users post on the Play Store regarding the Shopee application. The method applied in the data classification process is the Naïve Bayes Classifier (NBC), with the focus of this research being to test the level of accuracy of the system that has been built. The results of this research show that NBC has succeeded in providing an extraordinary level of accuracy, which will provide a positive contribution for students and researchers in analyzing public views regarding this topic.

3.2 Discussion
The presented research leverages the Naïve Bayes algorithm for Sentiment analysis on user reviews of the Shopee application from the Google Play Store. The study encompasses several crucial stages, starting with data collection through web scraping, followed by meticulous data preprocessing, TF-IDF weighting, and eventual classification using Naïve Bayes. The discussion section interprets and contextualizes the results obtained in the previous sections.

3.2.1 Data Collection and Preprocessing
The initial phase involved scraping user reviews from the Google Play Store, resulting in a dataset comprising 5000 of the latest Shopee application reviews. The subsequent preprocessing steps were vital in refining this raw dataset. Case folding ensured uniformity in text by converting all uppercase letters to lowercase. Reduplication removal and filtering removed unnecessary repetitions and characters, enhancing the dataset's informativeness. Further steps, including tokenization, conversion of slang words, removal of stopwords, and stemming, contributed to preparing the dataset for meaningful analysis. The thorough cleansing process ensured that the subsequent analysis was conducted on a refined and representative dataset.

3.2.2 TF-IDF Weighting
The TF-IDF weighting process assigned weights to each term, emphasizing the significance of words in distinguishing Sentiments. The sample training data underwent calculations to determine TF, DF (Document Frequency), and IDF values. The resulting TF-IDF values were then normalized, ensuring a consistent interval for each data point. This normalized TF-IDF dataset served as a foundation for subsequent Sentiment classification.

3.2.3 Naïve Bayes Classification
The core of the study lies in the application of the Naïve Bayes algorithm for Sentiment classification. The training data, derived from the preprocessed and weighted dataset, facilitated the learning of class characteristics. The subsequent testing phase, employing a 80:20 training-testing split, aimed to evaluate the model's accuracy. The classification process involved calculating prior probabilities, conditional probabilities, and posterior probabilities. A sample test, 'shop safe happy,' demonstrated the algorithm's ability to classify Sentiments accurately, yielding a positive Sentiment classification.

3.2.4 Application and Implications
The overarching objective of this research is to develop a system capable of Sentiment analysis, particularly focusing on user reviews of the Shopee application. The chosen Naïve Bayes Classifier exhibited remarkable accuracy, as evidenced by the positive Sentiment classification in the sample test. This achievement underscores the potential of the developed system to effectively categorize and analyze user Sentiments. The implications of this research extend beyond the immediate application to Shopee reviews. The methodology employed, from data collection to classification, establishes a robust framework applicable to Sentiment analysis in various domains. Researchers, practitioners, and students can

benefit from the insights gained in this study, particularly in understanding public Sentiment towards applications and services.

3.2.5 Limitations and Future Directions

While the Naïve Bayes algorithm demonstrated high accuracy, it's crucial to acknowledge certain limitations. The algorithm relies on the assumption of independence between features, which might not always hold true in real-world scenarios. Additionally, the study's success is contingent on the representativeness of the training data. Future research could explore advanced machine learning techniques, considering neural network-based approaches for Sentiment analysis. Expanding the dataset size and incorporating temporal aspects could enhance the system's robustness. Furthermore, feedback from end-users and continuous model refinement would contribute to an evolving and adaptive Sentiment analysis system.

4. Related Work

The Naïve Bayes method has proven to be a formidable force in machine learning classification, boasting rapid parameter estimation from training data samples, leading to remarkable accuracy rates. In the realm of Naïve Bayes classification, diverse studies have contributed unique methodologies and algorithms, each achieving noteworthy results. Muhammad Muslimin, Tri Herdiawan Apandi, and Didik Garbian Nugroho presented studies focusing on Naïve Bayes classification, attaining accuracy rates of 94.38%, 91.33%, and 80.00%, respectively [2]. A study by Muhammad Muslimin *et al*. in 2023, titled "sentiment Analysis of Rising Basic Commodity Prices Using Naïve Bayes Classifier," concluded that sentiment analysis on increasing basic commodity prices yielded negative responses, achieving an accuracy of 94.38% [1]. Didik Garbian Nugroho's 2016 study, "sentiment Analysis on Online Motorcycle Taxi Services Using Naïve Bayes," achieved an 80% accuracy rate utilizing sentiment data from Twitter [2]. While Naïve Bayes operates under the assumption of attribute independence, it remains a highly effective and efficient machine learning and data mining algorithm. Despite this assumption seldom holding in actual data, the algorithm consistently produces robust classification results. The method has been successfully applied in various contexts, such as sentiment analysis on COVID-19 vaccine opinions on Twitter [4], predicting satisfaction in e-KTP service recording [5], and analyzing sentiment on e-Tilang through YouTube [6]. Studies by Ahmad *et al*. in 2023, Apandi and Sugianto in 2019, Aulia *et al*. in 2023, Budiman in 2021, Cahyaningtyas *et al*. in 2021, and Depari *et al*. in 2022 have explored sentiment analysis, decision tree models, and disease classification, showcasing the versatility of these techniques across different domains[4][5][6][7][8][9][14][15]. Furthermore, diverse comparisons between classification methods have been conducted. Noteworthy studies include the improvement of sales performance with decision tree achieving 100% accuracy compared to Naïve Bayes with 41% [10], classification of fish freshness using K-Nearest Neighbor and Hue Saturation Value achieving 93% accuracy [11], and Sentiment analysis on Shopee user reviews, where Naïve Bayes displayed high accuracy at 99.5% [12]. Comparisons between C4.5 and Naïve Bayes in predicting student graduation showed distinct performances, with C4.5 having 76.69% accuracy and Naïve Bayes with 72.95% [13]. Other studies comparing Naïve Bayes with Decision Tree in classifying thesis titles demonstrated Naïve Bayes with 80.33% accuracy and Decision Tree with 60.33% [16]. The prediction of company bankruptcy using decision tree and Naïve Bayes showed Naïve Bayes achieving 100% accuracy [17]. When predicting student study completion on time, Naïve Bayes outperformed C4.5, boasting 69.97% accuracy compared to C4.5's 61.99% [18]. In the realm of Sentiment classification on Starbucks customer satisfaction, decision tree excelled with 83% accuracy, surpassing Naïve Bayes at 74% [19]. Finally, a study comparing Naïve Bayes and decision tree in classifying outstanding faculty members showed decision tree with 95.80% accuracy and Naïve Bayes with 94.80% [20]. These studies collectively contribute to the rich tapestry of methodologies in the field of Sentiment analysis and machine learning classification.

5. Conclusion

Based on the results of the Sentiment analysis on user reviews of the Shopee application using the Naïve Bayes algorithm that has been carried out, the following conclusions can be drawn: This research uses a dataset obtained from the Play Store Application using the Python Library as a data scrapper of 5000 data. After manual filtering by removing slang words, removing reduplication, punctuation, 3946 data were obtained. The application of the Naïve Bayes algorithm in this research is to use the probability method to classify and make predictions on 3141 training data and 805 test data using the Python library. The results of calculating the level of accuracy using the Naïve Bayes method for Sentiment classification can be said to be good, this can be seen from the accuracy results on a dataset of 3946 with a training data and test data calculation of 8:2, an accuracy value of 86.00%, a precision of 80.74%, recall of 78.13% and f1-score of 79.00%. The results of this research analysis resulted in superior Positive Sentiment amounting to 2050 and the number of Negative Sentiment 1199. The amount of training data in the system has an influence on the prediction system. Apart

from the amount of data, the quality of the training data also plays a role because the higher the quality of the data, the greater the understanding the system will have so that it will be more precise in predicting Sentiment classes.

References

[1] Muslimin, M. and Lusiana, V., 2023. Analisis Sentiment Terhadap Kenaikan Harga Bahan Pokok Menggunakan Metode Naive Bayes Classifier. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *7*(3), pp.1200-1209. DOI: http://dx.doi.org/10.30865/mib.v7i3.6418

[2] Nugroho, D.G., Chrisnanto, Y.H. and Wahana, A., 2016, September. Analisis Sentiment Pada Jasa Ojek Online Menggunakan Metode Naive Bayes. In *Prosiding Seminar Sains Nasional dan Teknologi*. 1(1), pp. 156-161. DOI: http://dx.doi.org/10.36499/psnst.v1i1.1526.

[3] Firmansyach, W.A., Hayati, U. and Wijaya, Y.A., 2023. Analisa Terjadinya Overfitting Dan Underfitting Pada Algoritma Naive Bayes Dan Decision Tree Dengan Teknik Cross Validation. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), pp.262-269. DOI: https://doi.org/10.36040/jati.v7i1.6329.

[4] Ahmad, A.Z., Asril, E., Sadar, M. and Turnandes, Y., 2023. Analisis Sentiment Opini Terhadap Vaksin Covid-19 Pada Media Sosial Twitter Menggunakan Naïve Bayes Dan Decision Tree. *ZONAsi: Jurnal Sistem Informasi*, *5*(1), pp.100-110. DOI: https://doi.org/10.31849/zn.v5i1.5553

[5] Apandi, T.H. and Sugianto, C.A., 2019. Algoritma Naive Bayes untuk Prediksi Kepuasan Pelayanan Perekaman e-KTP. *Juita*, *7*(2), pp.125-128. DOI: https://doi.org/10.30595/juita.v7i2.3608

[6] Aulia, Z.N., Jati, G.K. and Santoso, I., 2023. ANALISIS SENTIMENT TANGGAPANPUBLIC MENGENAI E-TILANG MELALUI MEDIA SOSIAL YOUTUBE MENGGUNAKAN ALGORITMA NAIVE BAYES. *IKRA-ITH Informatika: Jurnal Komputer dan Informatika*, *7*(2), pp.150-156.

[7] Budiman, B., 2021. Perbandingan Algoritma Klasifikasi Data Mining untuk Penelusuran Minat Calon Mahasiswa Baru. *NUANSA INFORMATIKA*, *15*(2), pp.37-52. DOI: https://doi.org/10.25134/nuansa.v15i2.4162

[8] Cahyaningtyas, C., Nataliani, Y. and Widiasari, I.R., 2021. Analisis Sentiment pada rating aplikasi Shopee menggunakan metode Decision Tree berbasis SMOTE. *AITI*, *18*(2), pp.173-184. DOI: https://doi.org/10.24246/aiti.v18i2.173-184

[9] Depari, D.H., Widiastiwi, Y. and Santoni, M.M., 2022. Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung. *Informatik: Jurnal Ilmu Komputer*, *18*(3), pp.239-248. DOI: https://doi.org/10.52958/iftk.v18i3.4694

[10] Shabrilianti, S.S., Triayudi, A. and Lantana, D.A., 2023. Analisis Klasifikasi Perfomance KPI Salesman Menggunakan Metode Decision Tree Dan Naïve Bayes. *JURIKOM (Jurnal Riset Komputer)*, *10*(1), pp.182-191. DOI: http://dx.doi.org/10.30865/jurikom.v10i1.5628.

[11] Situmorang, R.N., 2021. *Klasifikasi Kesegaran Ikan Berdasarkan Ekstraksi Fitur Menggunakan Metode K-Nearest Neighbor Dan Hue Saturation Value* (Thesis, Universitas Islam Negeri Sumatera Utara Medan)..

[12] Zukhoiriyah, D., 2022. *Analisis Sentiment Pada Review Pengguna ECommerce Menggunakan Algoritma Naïve Bayes (Studi Kasus: Shopee)* (Thesis, Universitas Islam Negeri Sumatera Utara).

[13] Firdaus, A.L.I., WIDODO, S., Sutrisman, A.D.I., GADING, S. and MARDIANA, R., 2019. Rancang bangun sistem informasi perpustakaan menggunakan web service pada jurusan teknik komputer polsri. *INFORMANIKA*, *5*(2).

[14] Ginting, V.S., Kusrini, K. and Taufiq, E., 2020. Implementasi Algoritma C4. 5 untuk Memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan Sekolah Menggunakan Python. *Inspiration: Jurnal Teknologi Informasi dan Komunikasi*, *10*(1), pp.36-44. DOI: https://doi.org/10.35585/inspir.v10i1.2535

[15] Kamil, M. and Cholil, W., 2020. Analisis Perbandingan Algoritma C4. 5 dan Naive Bayes pada Lulusan Tepat Waktu Mahasiswa di Universitas Islam Negeri Raden Fatah Palembang. *Jurnal Informatika*, *7*(2), pp.97-106. DOI: https://doi.org/10.31294/ji.v7i2.7723.

[16] Nuraeni, R., Sudiarjo, A. and Rizal, R., 2021. Perbandingan Algoritma Naïve Bayes Classifier dan Algoritma Decision Tree untuk Analisa Sistem Klasifikasi Judul Skripsi. *Innovation in Research of Informatics (INNOVATICS)*, *3*(1), pp. 26-31. DOI: https://doi.org/10.37058/innovatics.v3i1.2976

[17] Oktafia, D. and Pardede, D.L., 2010. Perbandingan Kinerja Algoritma Decision Tree dan Naïve Bayes dalam Prediksi Kebangkrutan. Universitas Gunadarma.

[18] Permana, J.N., Goejantoro, R. and Prangga, S., 2023. Perbandingan Algoritma C4. 5 Dan Naïve Bayes Untuk Prediksi Ketepatan Waktu Studi Mahasiswa. *EKSPONENSIAL*, *13*(2), pp.161-170. DOI: https://doi.org/10.30872/eksponensial.v13i2.947.

[19] Putri, T.A.Q., Triayudi, A. and Aldisa, R.T., 2023. Implementasi Algoritma Decision Tree dan Naïve Bayes Untuk Klasifikasi Sentiment Terhadap Kepuasan Pelanggan Starbucks. *Journal of Information System Research (JOSH)*, *4*(2), pp.641-649. DOI: https://doi.org/10.47065/josh.v4i2.2949

[20] Supriyadi, A., 2023. Perbandingan Algoritma Naive Bayes dan Decision Tree (C4. 5) dalam Klasifikasi Dosen Berprestasi. *Generation Journal*, *7*(1), pp.39-49. DOI: https://doi.org/10.29407/gj.v7i1.19797.