iBRICS
Journal of Multidisciplinary Study

Vol. 1, No. 2, 2025, Pg.117-132

Sentiment Analysis on The Shopee Platform Using The *Naive Bayes Algorithm*

Khilda Hania Rohmah [1] , Lolanda Hamim Annisa [2]
[1] Computer Science Study Program, Faculty of Science & Technology, Putra Bangsa University
[2] Data Science Study Program, Faculty of Science & Technology, Putra Bangsa University.
Email: khildahaniarohmah@gmail.com[1], lolanda@fst.universitasputrabangsa.ac.id[2]

ABSTRACT

Shopee is one of the *e-commerce applications* with the largest user base in Indonesia. The increase in the number of users is directly proportional to the number of reviews left on *the Google Play Store* , so these reviews can be used as an important source of information regarding user experience and satisfaction levels. This study was conducted to analyze the sentiment of these reviews using the *Naïve Bayes algorithm* . There are three main focuses of the study, namely: (1) explaining the *pre-processing stages* of review text data, (2) testing the performance of the *Naïve Bayes algorithm* in classifying sentiments into positive, negative, and neutral, and (3) interpreting the classification results to provide an overview of user perceptions and satisfaction with the Shopee application. The research data was collected using web scraping techniques on 4,500 reviews of the Shopee app on the Google Play *Store* . with ratio of 70% as training data and 30% as test data. Stages *pre-processing.* covering *cleaning, case folding, tokenizing, stopword removal,* normalization as well as *stemming.* Next, the data is labeled according to sentiment categories and then processed using the *Naïve Bayes algorithm* . Model performance is evaluated using accuracy, precision, *recall,* and *F1-score metrics.* Test results show that the *Naïve Bayes algorithm* is capable of providing sentiment classification with a good level of accuracy. The majority of user reviews are positive, reflecting a high level of satisfaction with the Shopee app. This research is expected to provide input for app developers and *e-commerce businesses* in designing service improvement strategies based on user perceptions.

Keywords: *Analysis Sentiment; Naive Bayes; Shopee; Google Play Store; Reviews User*

INTRODUCTION

The rapid development of information and communication technology (ICT) over the past two decades has driven a major transformation in the digital commerce sector. *E-commerce* allows consumers to conduct transactions easily and flexibly, without being limited by time and space. In Indonesia, *e-commerce has grown* rapidly due to high internet penetration and mobile device usage (Central Bureau of Statistics, 2023).

Shopee has become a leading *e-commerce platform,* providing not only online shopping services but also interactive features such as product reviews and ratings, which play a crucial role in influencing purchasing decisions.

The increase in user activity has also resulted in a growing number of reviews, particularly on *the Google Play Store.* These reviews vary widely in terms of language, text length, and context, creating challenges in managing them. Therefore, sentiment analysis is an effective method for understanding user opinions through the reviews they provide. Sentiment analysis is an automated process in natural language processing that aims to identify user emotions or opinions, such as positive, negative, or neutral, toward an entity (Kumar et al., 2023) . The results can then be used to improve services, marketing strategies, and product development.

One popular algorithm in sentiment analysis is *Naïve Bayes* , due to its simple, efficient, and accurate model for large-scale text processing. Despite assuming independence between features, which is rarely met in practice, *Naïve Bayes* is still capable of producing competitive classifications (Agustina et al., 2022) . Research by (Rahel Lina Simanjuntak et al., 2023) shows that this algorithm has advantages in efficiency and ease of implementation, and is relevant for large data such as reviews during National Online Shopping Day. Furthermore, (Wandani, 2021) shows that sentiment analysis can capture public perception in real-time through social media data during Shopee flash sale events.

Other research findings also reinforce the importance of sentiment analysis for business decision-making. (Siniwi et al., 2021) emphasize that the results of this analysis can provide constructive feedback for improving product and service quality. Meanwhile, (Sihombing et al., 2021) and (Darmawan et al., 2020) demonstrate its direct benefits in responding to customer expectations, and (Rani & Candra, 2023) state that customer sentiment can be an important indicator in evaluating online store performance. Based on these relevance and challenges, this study aims to implement the *Naïve Bayes algorithm* in sentiment analysis of Shopee user reviews, while evaluating its effectiveness through an enhanced preprocessing approach and analysis based on various review contexts.

Vol. 1, No. 2, 2025, Pg.117-132

LITERATURE REVIEW

Shopee *Platform*

According to Wikipedia, Shopee is an e-commerce website headquartered in Singapore owned by Sea Limited (formerly known as Garena), founded in 2009 by Forrest Li. It first launched in Singapore in 2015 and has since expanded its reach to Malaysia, Thailand, Taiwan, Indonesia, Vietnam, and the Philippines. Starting in 2019, Shopee also launched in Brazil, making it the first country in South America and outside Asia to be launched.

Sentiment Analysis

According to Mustopa et al. (2020), sentiment analysis is a technique for extracting information from a person's perspective on a problem. It is also a process of understanding and automatically processing textual data to generate information about the negative and positive categories contained within sentences. Sentiment analysis has tremendous impact and benefits, leading to rapid growth in sentiment analysis research.

Naive Bayes Algorithm

According to Syarli in (Mustopa et al., 2020) the Naive Bayes Algorithm is a statistical classifier that can predict the probability of data class membership, this classifier is classified into a certain class according to probability theory.

Text Pre-processing.

Text Pre-processing is an implementation of *text mining* that selects text data to make it more structured through a series of stages, namely cleaning, tokenizing, normalization, case folding, and stopword removal (Farasqa Nauval Akbar et al., 2023) .

Previous research used the Naive Bayes algorithm as its method, but there are also other algorithms, such as in the study conducted by (Agustina et al., 2022) entitled Implementation of Naive Bayes for sentiment analysis of Shopee reviews on *the Google Play Store* , an accuracy of 87.58% which means emphasizing the importance

of *pre-processing. . Research* conducted by (Rahel Lina Simanjuntak et al., 2023) entitled Comparison of Naive Bayes with other algorithms, and the results Naive Bayes is superior in efficiency and competitive results. Research conducted by (Beno et al., 2022) with title Sentiment *Analysis of Customer Reviews on E-commerce Platforms: A Machine Learning Approach* , research the compare algorithm other with Naive Bayes the result algorithm *Naive Bayes* accurate and capable handling big data No structured in analysis sentiment review *e-commerce.*

METHOD

This study uses a quantitative approach with secondary data in the form of Shopee user reviews from *the Google Play Store* . A total of 4,500 reviews were collected through web scraping, then *pre-processing was carried out,* including *cleaning, case folding, tokenizing, stopword removal,* normalization, and *stemming.* The data was labeled positive, negative, or neutral, then divided into 70% training data and 30% test data. The *Naïve Bayes algorithm* was used for classification, and performance was evaluated using accuracy, precision, *recall* , and *F1-score.*

RESULTS AND DISCUSSION

Data Scraping

Data scraping is an automated technique for extracting information from web pages and converting unstructured data into a structured format for further processing. In this study, the scraping process was performed using the Python programming language through the *Google Colab platform* because it supports cloud-based Python libraries without local installation. The Pandas library was used to organize the extraction results into tables before being saved in CSV format.

Stages *scraping* started with identification The target's *Uniform Resource Locator* (URL), i.e page review The Shopee app on *the Google Play Store* . Next, a Python script is run to access the page and auto-scroll to load the full review. Once all review elements are loaded, an extraction process is performed to retrieve necessary attributes, such as review text, number of stars, publication date, and user ID. The extracted data is then saved in CSV format for use in the data *preprocessing stage.*

SCRAPING DATA

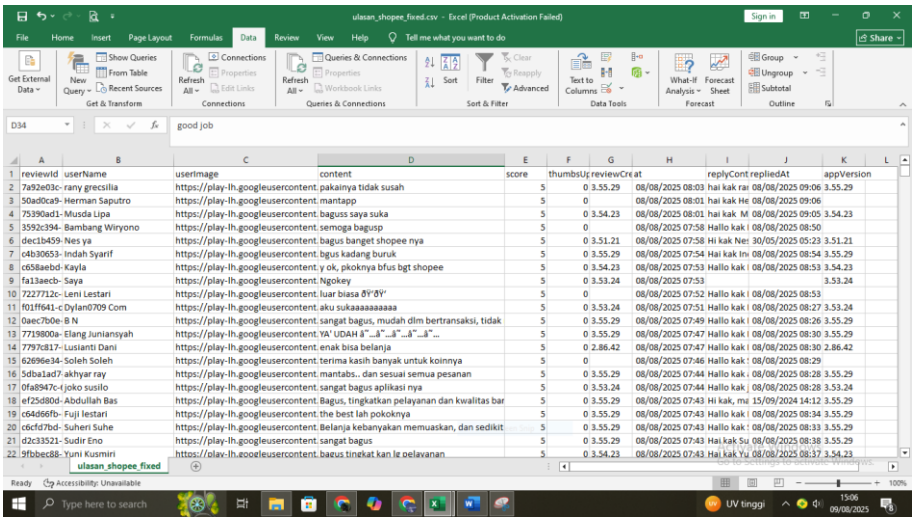
```python
from google_play_scraper import reviews, Sort
import pandas as pd
import time

app_id = 'com.shopee.id'
all_reviews = []
total_reviews = 10000
batch_size = 200

continuation_token = None

while len(all_reviews) < total_reviews:
    rvs, continuation_token = reviews(
        app_id,
        lang='id',
        country='id',
        sort=Sort.NEWEST,
        count=batch_size,
        filter_score_with=None,
        continuation_token=continuation_token
    )
    all_reviews.extend(rvs)
    if not continuation_token:
        break
    time.sleep(1)
# Simpan ke CSV
df = pd.DataFrame(all_reviews)
df.to_csv("ulasan_shopee_scrap.csv", index=False, encoding='utf-8')
```

Figure - 1 *Data Scraping* Process



	reviewId	userName	userImage	content	score	thumbsUp	reviewCreat	at	replyCont	repliedAt	appVersion
2	7a92e03c-	rany grecsilia	https://play-lh.googleusercontent.	pakainya tidak susah	5	0	3.55.29	08/08/2025 08:03	hai kak rar	08/08/2025 09:06	3.55.29
3	50ad0ca9-	Herman Saputro	https://play-lh.googleusercontent.	mantapp	5	0		08/08/2025 08:01	hai kak He	08/08/2025 09:06	
4	75390ad1-	Musda Lipa	https://play-lh.googleusercontent.	baguss saya suka	5	0	3.54.23	08/08/2025 08:01	hai kak M	08/08/2025 09:05	3.54.23
5	3592c394-	Bambang Wiryono	https://play-lh.googleusercontent.	semoga bagusp	5	0		08/08/2025 07:58	Hallo kak (	08/08/2025 08:50	
6	dec1b459-	Nes ya	https://play-lh.googleusercontent.	bagus banget shopee nya	5	0	3.51.21	08/08/2025 07:58	Hi kak Nes	30/05/2025 05:23	3.51.21
7	c4b30653-	Indah Syarif	https://play-lh.googleusercontent.	bgus kadang buruk	5	0	3.55.29	08/08/2025 07:54	Hai kak Ind	08/08/2025 08:54	3.55.29
8	c658aebd-	Kayla	https://play-lh.googleusercontent.	y ok, pkoknya bfus bgt shopee	5	0	3.54.23	08/08/2025 07:53	Hallo kak (	08/08/2025 08:53	3.54.23
9	fa13aecb-	Saya	https://play-lh.googleusercontent.	Ngokey	5	0	3.53.24	08/08/2025 07:53			3.53.24
10	7227712c-	Leni Lestari	https://play-lh.googleusercontent.	luar biasa ðŸ‘ðŸ‘	5	0		08/08/2025 07:52	Hallo kak (	08/08/2025 08:53	
11	f01ff641-c	Dylan0709 Com	https://play-lh.googleusercontent.	aku sukaaaaaaaaaa	5	0	3.53.24	08/08/2025 07:51	Hallo kak (	08/08/2025 08:27	3.53.24
12	0aec7b0e-	B N	https://play-lh.googleusercontent.	sangat bagus, mudah dlm bertransaksi, tidak	5	0	3.55.29	08/08/2025 07:49	Hallo kak (	08/08/2025 08:26	3.55.29
13	7719800a-	Elang Juniansyah	https://play-lh.googleusercontent.	YA' UDAH â"...â"...â"...â"...	5	0	3.55.29	08/08/2025 07:47	Hallo kak (	08/08/2025 08:30	3.55.29
14	7797c817-	Lusianti Dani	https://play-lh.googleusercontent.	enak bisa belanja	5	0	2.86.42	08/08/2025 07:47	Hallo kak (	08/08/2025 08:30	2.86.42
15	62696e34-	Soleh Soleh	https://play-lh.googleusercontent.	terima kasih banyak untuk koinnya	5	0		08/08/2025 07:46	Hallo kak (	08/08/2025 08:29	
16	5dba1ad7-	akhyar ray	https://play-lh.googleusercontent.	mantabs.. dan sesuai semua pesanan	5	0	3.55.29	08/08/2025 07:44	Hallo kak (	08/08/2025 08:28	3.55.29
17	0fa8947c-	ijoko susilo	https://play-lh.googleusercontent.	sangat bagus aplikasi nya	5	0	3.53.24	08/08/2025 07:44	Hallo kak (	08/08/2025 08:28	3.53.24
18	ef25d80d-	Abdullah Bas	https://play-lh.googleusercontent.	Bagus, tingkatkan pelayanan dan kwalitas bar	5	0	3.55.29	08/08/2025 07:43	Hi kak, ma	15/09/2024 14:12	3.55.29
19	c64d66fb-	Fuji lestari	https://play-lh.googleusercontent.	the best lah pokoknya	5	0	3.55.29	08/08/2025 07:43	Hallo kak (	08/08/2025 08:34	3.55.29
20	c6cfd7bd-	Suheri Suhe	https://play-lh.googleusercontent.	Belanja kebanyakan memuaskan, dan sedikit	5	0	3.55.29	08/08/2025 07:43	Hallo kak (	08/08/2025 08:33	3.55.29
21	d2c33521-	Sudir Eno	https://play-lh.googleusercontent.	sangat bagus	5	0	3.55.29	08/08/2025 07:43	Hai kak Su	08/08/2025 08:38	3.55.29
22	9fbbec88-	Yuni Kusmiri	https://play-lh.googleusercontent.	bagus tingkat kan lg pelayanan	5	0	3.54.23	08/08/2025 07:43	Hai kak Yu	08/08/2025 08:37	3.54.23

Figure - 2 *Data Scraping* Results

Data Elimination

After collecting data through *web scraping* , the next step is filtering to ensure the quality of the dataset for analysis. One important step is removing reviews with 1- and 2-star ratings, as these tend to be extreme, emotional, short, or spammy. These types of reviews often lack relevant information and risk introducing *noise* into the training of sentiment classification models, potentially reducing their accuracy.

The filtering process was performed using the *Pandas library* in Python, leaving only reviews with 3, 4, and 5-star ratings. A rating of 3 represents a neutral opinion, while 4 and 5 stars indicate positive sentiment with varying degrees of intensity. After filtering, the data was reduced from 10,000 to 4,500 reviews, ready for use in the *pre-processing stage.*

ELIMINASI DATA

```
[20]
    df = pd.DataFrame(all_reviews)

    #  Baca file hasil scraping
    input_file = "/content/drive/MyDrive/SKRIPSIII/ulasan_shopee_scrap.csv"  #
    df = pd.read_csv(input_file)

    # Hapus ulasan dengan rating 1 dan 2
    df = df[df['score'] > 2]

    # Reset index
    df = df.reset_index(drop=True)

    print(f"Total ulasan setelah eliminasi rating 1 dan 2: {len(df)}")

    # Simpan hasil
    output_file = "ulasan_shopee_fixed.csv"
    df.to_csv(output_file, index=False, encoding='utf-8')

    df[['score','content',]].head(5)
```

Figure - 3 Data Elimination Process

	score	content
0	5	Aplikasi ini sangat berguna, jika tidak ada wa...
1	5	no hoak yah
2	5	Mantap
3	5	cumak shoppe yg kalo beli barang slalu gratis ...
4	5	aplikasinya sangat membantu

Figure - 4 Elimination Results

Cleaning

Cleaning stage is one of the from a number of stage *pre- processing.. Pre -processing* stage . is an important process in data- based analysis purposeful text For transform raw data into a more format clean , structured , and ready used in the classification process Sentiment . The *cleaning* stage aims to remove unnecessary characters, such as symbols, numbers, punctuation, emoticons, hashtags, and double spaces. *Figure* 5

shows the results of the review text after cleaning. The text is neater, cleaner, and ready for the next stage.

index	content	cleaned_text
0	Aplikasi ini sangat berguna, jika tidak ada waktu luang untuk membeli barang, bisa membelinya di aplikasi ini kapan saja, dan aplikasi ini sangat bermanfaat.	Aplikasi ini sangat berguna jika tidak ada waktu luang untuk membeli barang bisa membelinya di aplikasi ini kapan saja dan aplikasi ini sangat bermanfaat
1	no hoak yah	no hoak yah
3	cumak shoppe yg kalo beli barang slalu gratis ongkir,, makasih pertahankan slalu	cumak shoppe yg kalo beli barang slalu gratis ongkir makasih pertahankan slalu
4	aplikasinya sangat membantu	aplikasinya sangat membantu
7	membantu memenuhi kebutuhan kita	membantu memenuhi kebutuhan kita
8	bagus sesuai pesenan	bagus sesuai pesenan
9	sudah pernah order dan barang sesuai foto... alhamdulillah amanah	sudah pernah order dan barang sesuai foto alhamdulillah amanah
10	tolong di cek Toko² yang ingin menipu pembeli, patroli ke tokonya dan liat feedback-nya.	tolong di cek Toko yang ingin menipu pembeli patroli ke tokonya dan liat feedback nya
15	sangat bagus & sangat membantu	sangat bagus sangat membantu
16	kenapa saya tidak bisa pesan barang yg saya minat coba kasi tangkapan soal masalah yg Saya dapat ini	kenapa saya tidak bisa pesan barang yg saya minat coba kasi tangkapan soal masalah yg Saya dapat ini
20	apk ini sangat dipercayai dan membantu ekonomi saya. kualitas terjamin dan jika ada pengembalian sllu diterima 🙏	apk ini sangat dipercayai dan membantu ekonomi saya kualitas terjamin dan jika ada pengembalian sllu diterima
21	semoga keamanan akun konsumen jadi prioritas	semoga keamanan akun konsumen jadi prioritas
22	e commerce terlanganan akuu pokoknyaaa... selalu beli disiniii ntah baju,celana,barang,perabotan lainnya selalu disiniii. bahkan pampers or kebutuhan bayi semua disiniii... enak bngttt klo gapunya waktu atau lagi gabisa buat keluar rumah aku selalu pesan lewat online... kecuali makanan atau ber food an disinj.. krna sistem di daerah aku belum bisa huhuuu😡😡😡🙏...	e commerce terlanganan akuu pokoknyaaa selalu beli disiniii ntah baju celana barang perabotan lainnya selalu disiniii bahkan pampers or kebutuhan bayi semua disiniii enak bngttt klo gapunya waktu atau lagi gabisa buat keluar rumah aku selalu pesan lewat online kecuali makanan atau ber food an disinj krna sistem di daerah aku belum bisa huhuuu
24	mantap sekali dan sangat membantu	mantap sekali dan sangat membantu
25	All running well all running more better	All running well all running more better

Figure - 5 *Cleaning* Results

Folding Case

This process change all over letter in text review become letter small or *lower case,* following is results *Folding Case* show text after the case folding process , where all letter Already uniform in small format letters . Here Ganbar – 6 *Cae Folding Results* show text After the case folding process, where all letter Already uniform in small format letters .

	casefolding_text
0	aplikasi ini sangat berguna jika tidak ada wak...
1	no hoak yah
2	cumak shoppe yg kalo beli barang slalu gratis ...
3	aplikasinya sangat membantu
4	membantu memenuhi kebutuhan kita

Figure - 6 *Folding Case* Results

Tokenizing

Tokenizing is a separation process text review into the smallest units in the form of tokens. Stages This play a role important in facilitate analysis at the word level, so that allows processing and understanding meaning text in a way more structured. The following Figure – 7 Tokenizing Results shows the list of words (tokens) of the results

separator text , so that can processed more continue to the next stage *stopword removal.*

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
```

	tokens
0	[aplikasi, ini, sangat, berguna, jika, tidak, ...
1	[no, hoak, yah]
2	[cumak, shoppe, yg, kalo, beli, barang, slalu,...
3	[aplikasinya, sangat, membantu]
4	[membantu, memenuhi, kebutuhan, kita]

Figure - 7 Tokenizing Results

Stopword Removal

This process removes common words that don't contribute significantly to sentiment analysis, such as "dan," "di," "yang," and "ke." The removal is performed using a list of Indonesian *stopwords available in the nltk* and *Sastrawi libraries.* Figure 8 shows the results of *stopword removal,* showing the text after removing unimportant words, leaving only meaningful words.

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

	tokens_no_stopword
0	[aplikasi, berguna, luang, membeli, barang, me...
1	[no, hoak, yah]
2	[cumak, shoppe, yg, kalo, beli, barang, slalu,...
3	[aplikasinya, membantu]
4	[membantu, memenuhi, kebutuhan]

Figure - 8*Stopword Removal* Results

Normalization

Stage furthermore is Normalization aim correct the words that are experiencing error writing or typo, duplication letters, or use Language No standard or slang to be form standard. For example, “bangeeet” is changed to “banget”, and “gk” becomes “tidak”.

The following Figure - 9 Normalization Results shows the results of the normalized review text.

	tokens_normalized
0	[aplikasi, berguna, luang, membeli, barang, me...
1	[no, hoak, yah]
2	[cumak, shoppe, yang, kalo, beli, barang, slal...
3	[aplikasinya, membantu]
4	[membantu, memenuhi, kebutuhan]

Figure - 9 Normalization Results

Stemming

Stemming converts affixed words to their base form by removing prefixes, suffixes, or infixes. For example, the word "berbelajar" is changed to "belanja." In this study, the stemming process was carried out using the Sastrawi library, which is specifically designed for Indonesian. Figure 10 shows the *stemming results* of the review text after the affixed words were converted to their base form.

```
Requirement already satisfied: Sastrawi in /usr/local/lib/python3.11/dist-packages (1.0.1)
```

	tokens_stemmed
0	[aplikasi, guna, luang, beli, barang, bel, apl...
1	[no, hoak, yah]
2	[cumak, shoppe, yang, kamu, beli, barang, slal...
3	[aplikasi, bantu]
4	[bantu, penuh, butuh]

Image - 10*Stemming* Results

Labeling

Review data that has been processed previously Then labeled sentiment for needs modeling, with using two approaches, namely manual and semi- automatic. Manual labeling is done by researchers by reading the contents of the review directly and determining positive, negative, or neutral labels based on the meaning contained. Meanwhile, semi-automatic labeling uses *a sentiment dictionary* containing a list of positive and negative words; reviews will be labeled according to the dominance of

these words. If there are more positive words, the review is labeled positive; if there are more negative words, it is labeled negative; and if there is no dominance, the review is considered neutral.

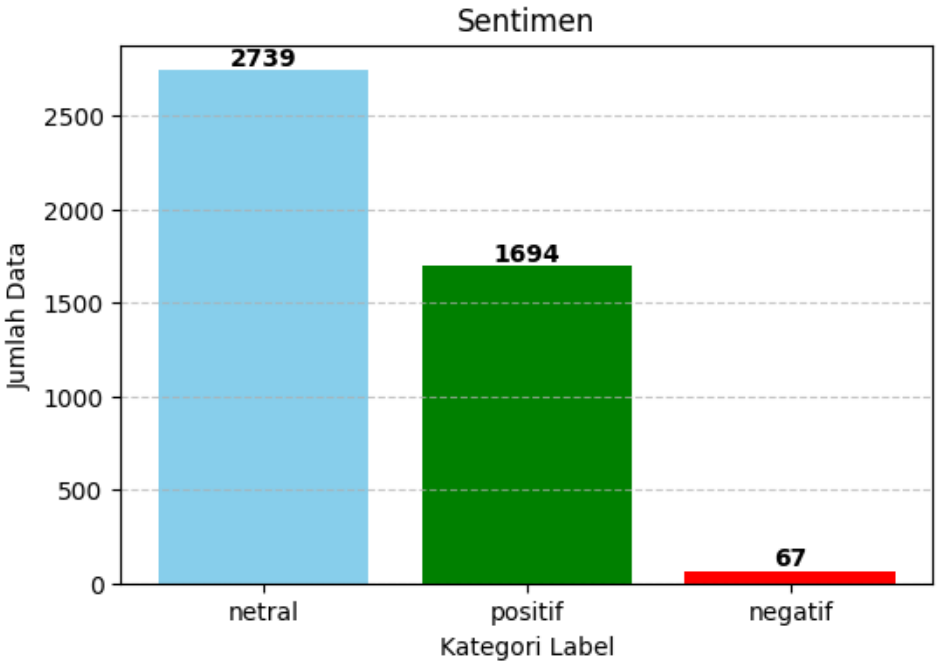


Figure - 11 *Data Labeling* Results

WordCloud

Wordcloud is visualization of the retrieved data set from frequent words appears on Shopee reviews.

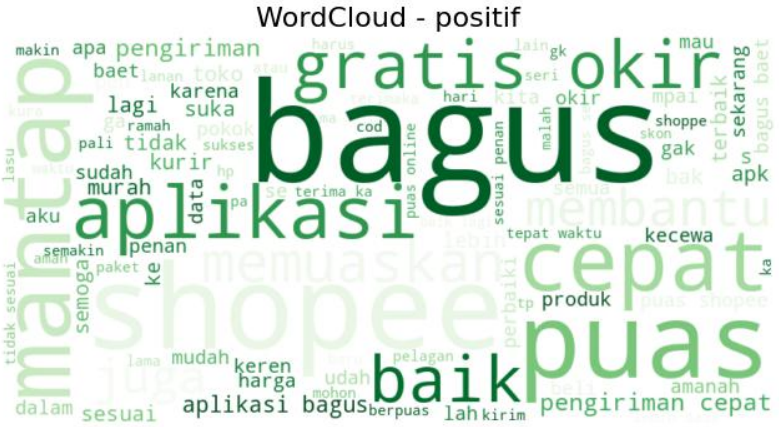


Figure - 12 *WordCloud* Sentiment Positive

Figure - 13 *Word Cloud* Sentiment Negative

Figure - 14 *Word Cloud* Sentiment Neutral

Division of Training Data and Test Data

Evaluation of the model in study This includes manual and semi- automatic data labeling processes as well as implementation method classification *Naïve Bayes*, dataset is divided into two parts, namely training data and test data, with ratio distribution 70 % for training data and 30% for test data. The division This aims to make *the Naïve Bayes* model get sufficient data portion big for the learning process, as well as provide adequate test data for evaluate model performance in general objective.

```python
from sklearn.model_selection import train_test_split

# Pisahkan fitur dan label
X = df['content_clean']
y = df['label']

# Split train, test
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)

print("Data Train:", len(X_train))
print("Data Test:", len(X_test))

df.head(len(X_test))
```

Figure - 15 *Split Data*

Implementation Algorithm Naive Bayes

After stage division of training data and test data stages furthermore is implementation algorithm *Naive Bayes* to the stage model done For evaluate performance of the model that has been was built. In this study, *the Multinomial Naïve Bayes method* was implemented by utilizing the *scikit-learn library* . The review that has been through the *pre-processing* process. Then converted become form numeric use Term Frequency – Inverse Document Frequency (TF-IDF) technique via TfidfVectorizer.

```
=== Naive Bayes ===
Accuracy: 0.8755555555555555
              precision    recall  f1-score   support

     negatif       0.00      0.00      0.00        20
      netral       0.85      0.97      0.91       822
     positif       0.94      0.76      0.84       508

    accuracy                           0.88      1350
   macro avg       0.60      0.58      0.58      1350
weighted avg       0.87      0.88      0.87      1350
```

Figure - 16TF-IDF Results

The implementation results show that the model has an accuracy of 87.55%. The neutral category achieved a *precision* of 0.85 and a recall of 0.97, indicating that the model was able to identify almost all neutral reviews well. In the positive category,

the precision reached 0.94 but *the recall* was only 0.76, indicating that some positive reviews were classified into other categories. Meanwhile, the negative category achieved a *precision* and *recall* of 0.00 due to the very limited amount of negative data, which was only 20 reviews, so the model did not have enough information to learn relevant patterns.

The macro average precision and *recall* values were 0.60 and 0.58, respectively, which were considered low due to the unbalanced data distribution between classes, while the weighted average precision and *recall values* were 0.87 and 0.88, respectively, which remained high due to the dominance of the number of data in the neutral and positive categories.

Analysis of Data Processing Results

Pre-processing process. in study This aim cleaning and preparing review data Shopee users on *Google Play Store* before implemented algorithm *Naïve Bayes*. The stages covering *cleaning* (removing) character No relevant), *case folding* (change letter become small), *tokenizing* (breaking down sentence become a word), *stopword removal* (removing a word that does not important) and *stemming* (returning words to their original form) form This step results in cleaner and more consistent review text, supporting optimal model learning of word patterns. Evaluation using Multinomial *Naive Bayes and TF-IDF weighting on 1,350 reviews showed an accuracy of 87.55%, with the best performance for neutral and positive sentiments,* although still weak in the negative category due to the very small amount of data.

In general, the classification results indicate that most users have a positive to neutral perception of the Shopee app, as evidenced by the dominance of these two categories in the data. Neutral reviews tend to be informative, while positive reviews reflect user satisfaction levels. However, even in small numbers, the presence of negative reviews is an important signal for developers to address service deficiencies. The model's weakness in classifying negative sentiment can be addressed by adding negative data or implementing data balancing techniques such as SMOTE, which enable the model to more accurately recognize the characteristics of negative reviews.

Vol. 1, No. 2, 2025, Pg.117-132

CONCLUSION

pre-processing process for Shopee user reviews on *the Google Play Store* was systematically conducted to ensure the data was ready for processing using the *Naive Bayes algorithm* . This process successfully produced a clean, consistent, and representative dataset, thus supporting optimal model performance. Model testing with a 70% training data and 30% test data split yielded an accuracy of 87.55%, indicating that the majority of model predictions matched the actual sentiment labels. The distribution of classification results shows that most user reviews fall into the neutral and positive categories, which indicates that the general perception of the Shopee application tends to be good to satisfactory. Future research could incorporate methods such as *the Synthetic Minority Oversampling Technique* (SMOTE) or oversampling to improve data distribution between classes. Further research could use datasets from different *e-commerce platforms* to test the model's generalization ability across broader domains. Further research could increase the amount of data in each sentiment category to achieve more accurate and balanced classification results, particularly to address class imbalance issues in categories with limited data.

REFERENCES

Agustina, N., Citra, DH, Purnama, W., Nisa, C., & Kurnia, AR (2022). Implementation of the Naive Bayes Algorithm for Sentiment Analysis of Shopee Reviews on *the Google Play Store* . *MALCOM: Indonesian Journal of Machine Learning and Computer Science* , *2* (1), 47–54. https://doi.org/10.57152/malcom.v2i1.195

Central Bureau of Statistics. (2025, January 30). *E-commerce Statistics 2023.* Central Bureau of Statistics. https://www.bps.go.id/id/publication/2025/01/30/d52af11843aee401403ecfa6/e-commerce-statistics-2023.html

Beno, J., Silen, A. ., & Yanti, M. (2022). Sentiment Analysis of Customer Reviews on E-commerce Platforms: A Machine Learning Approach. *Braz Dent J.* , *33* (1), 1–12.

Darmawan, I., Pratiwi, ON, Industri, FR, & Telkom, U. (2020). Sentiment Analysis of Rubylicious Online Store Product Reviews For. *E-Proceeding of Engineering* , *7* (2), 7026–7034.

Farasqa Nauval Akbar, B., Arifianto, D., Maryam Zakiyyah, A., & Milu Susetyo, A. (2023). Sentiment Analysis of Anies Baswedan Using the Support Vector Machine Method Case Study of Twitter Social Media. *Jurnal Smart Teknologi* , *4* (6), 2774–1702. http://jurnal.unmuhjember.ac.id/index.php/JST

Kumar, S., Roy, P.P., Dogra, D.P., & Kim, B.-G. (2023). *A Comprehensive Review on Sentiment Analysis: Tasks, Approaches and Applications* . 1–29. http://arxiv.org/abs/2311.11250

Mustopa, A., Hermanto, Anna, Pratama, EB, Hendini, A., & Risdiansyah, D. (2020). Analysis of user reviews for the carelindungi application on google play using the support vector machine and naive bayes algorithm based on particle swarm optimization. *2020 5th International Conference on Informatics and Computing, ICIC 2020* , *2* . https://doi.org/10.1109/ICIC50835.2020.9288655

Rahel Lina Simanjuntak, Theresia Romauli Siagian, Vina Anggriani, & Arnita Arnita. (2023). Sentiment Analysis of Reviews on the Shopee E-Commerce Application Using the *Naive Bayes Algorithm* . *Journal of Mechanical, Electrical and Computer Engineering* , *3* (3), 23–39. https://doi.org/10.55606/teknik.v3i3.2411

Rani, MM, & Candra, F. (2023). Sentiment Analysis of Halona Beauty Care Online Store Reviews for Service Improvement Using the *Naive Bayes Algorithm* . *INOVTEK Polbeng - Informatics Series* , *8* (2), 215. https://doi.org/10.35314/isi.v8i2.3348

Sihombing, LO, Hannie, H., & Dermawan, BA (2021). Sentiment Analysis of

Shopee Indonesia Product Customer Reviews Using the *Naive Bayes* Classifier Algorithm. *Edumatic: Journal of Informatics Education* , *5* (2), 233–242. https://doi.org/10.29408/edumatic.v5i2.4089

Siniwi, LM, Prahutama, A., & Hakim, AR (2021). Query Expansion Ranking in Sentiment Analysis Using Multinomial *Naive Bayes Classification* (Case Study: Shopee Application Review on National Online Shopping Day 2020). *Gaussian Journal* , *10* (3), 377–387. https://doi.org/10.14710/j.gauss.v10i3.32795

Wandani, A. (2021). Sentiment Analysis of Twitter Users on Flash Sale Events Using K-NN, Random Forest, and Naive Bayes Algorithms. *Journal of Computer Science & Informatics (J-SAKTI)* , *5* (2), 651–665.