Play Store Data Scrapping and Preprocessing done as Sentiment Analysis Material

Rakyatul Hasanah [1,*], Sulistiani [2], Nurhikmayani [3], Zakiyah Hasanah [4], Setiawan Ardi Wijaya [5], Dahmani Abdennasser [6,7], Abdel-Nasser Sharkawy [8,9]

[1,2,3] Department of Computer Science, Universitas Qamarul Huda Badaruddin, Central Lombok 83562, Indonesia
[4] Bachelor of Economics, Albukhary International University, Kedah 05200, Malaysia
[5] Department of Information System, Universitas Muhammadiyah Riau, Pekanbaru 28294, Indonesia
[6] Department of Mechanical Engineering, Faculty of Sciences and Applied Sciences, University of Bouira, Bouira 10000, Algeria
[7] Laboratory of Biomaterials and Transport Phenomena (LBMPT), University of Medea, Urban Pole, 26000 Medea, Algeria
[8] Mechatronics Engineering, Mechanical Engineering Department, South Valley University, Qena, 83523, Egypt
[9] Mechanical Engineering Department, Fahad Bin Sultan University, Tabuk 47721, Saudi Arabia

ARTICLE INFO

Article history:

Received November 13, 2024
Revised December 28, 2024
Published January 18, 2025

Keywords:

e-commerce; preprocessing; sentiment analysis; scrapping; Shopee.

ABSTRACT

Sentiment analysis is a computational technique used to interpret user opinions about a product through textual reviews. This research aims to prepare useful data for further research, one of which is sentiment analysis. A total of 12,000 recent reviews from July 2024 - January 2025 were collected through web scrapping. The research process includes data preprocessing steps such as case folding and data cleaning to transform the raw data into a usable format. The raw data up to the given changes have been uploaded to the Mendeley data repository to be reprocessed into further research, one of which is the sentiment analysis approach.

Corresponding Author:

Rakyatol Hasanah, Department of Computer Science, Universitas Qamarul Huda Badaruddin, Central Lombok, Indonesia
Email: rokyatulhasanah77@gmail.com

1. INTRODUCTION

The rapid advancement of technology brings convenience to society in various aspects of life. The continuous growth of technology produces a very large amount of data, which can be a useful source of information if processed and utilized properly [1]. Electronic Commerce is a transaction process that involves sellers and buyers through the internet [2].

Sentiment analysis initially developed as a field in computer science and later expanded to social science and management studies. Because emotions, cognition, and behavior are intertwined, sentiment analysis can help researchers understand individual attitudes, predict human behavior, and provide guidance for preventive and corrective actions at the individual and societal levels. Sentiment refers to the feeling underlying an expressed positive or negative opinion, or the feeling implied in a neutral opinion. Therefore, this analysis is also known as opinion mining [3].

Sentiment analysis is a computational process that aims to automatically understand, extract, and process textual data in order to obtain information contained in a person's opinion or behavior. [4]. This process is used to analyze unstructured datasets, so as to provide relevan and useful information. Sentiment analysis has the benefit of understanding user responses to a product. By extracting text from reviews, it can reveal the user's emotions, whether the response is positive, negative or neutral. This helps in evaluating whether or not the user's response to the product is favorable [5].

Shopee is currently the most visited marketplace in Indonesia, with average monthly visits reaching 132.8 million in the first quarter of 2022, according to data from iPrice 2022. Shopee is a buying and selling and shopping application that can be downloaded through Google Playstore. On the Google Playstore page, there are various features, including ratings and reviews, that allow users to rate the app. Comments refer to text that provides a response to a particular idea or work. Reviews from previous users can provide effective information about the quality of products and services, so many internet users tend to trust the recommendations and opinions provided. However, there is no systematic and accurate method to classify reviews as positive or negative. Related to consumer affection can help developers in collecting emotional data from e-commerce application users [6].

Naïve Bayes algorithm is a simple classification algorithm that works by calculating probabilities based on summing and combining values from available datasets. In this research, the Naïve Bayes Algorithm method is applied to classify people's opinions on the Shopee application. Based on various references, the Naïve Bayes Algorithm is in high demand due to its simplicity and fast data processing capabilities. This algorithm can provide high accuracy while processing large amounts of data with high efficiency [7].

This study will concentrate on preprocessing the customer reviews of the Shopee e-commerce app, sourced from the Google Play Store. The preprocessing process will involve case folding and data cleaning. The goal of this research is to prepare the data for future analysis, ensuring that the reviews are properly cleaned and ready for sentiment analysis in the next steps.

2. METHODS

The general research method can be shown in Figure.

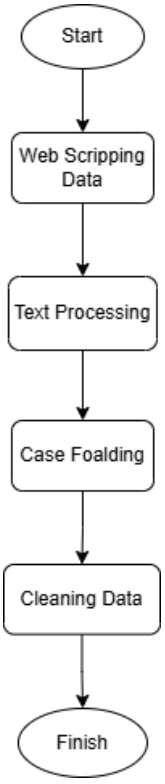

Fig. 1. Research Stages

2.1. Web Scrapping

Web scrapping data is a process carried out to collect review data from the google play store using the python programming language [8]. The results of this web scrapping obtained 12,000 records of review data on the shopee application on the Google Play Store.

2.2. Text Preprocessing

Preprocessing is the stage of converting raw data into data that is ready for processing [9]. the preprocessing stages carried out in this study include the stage:

- Case folding Lowercase is the process of converting capital letters (uppercase) in a sentence into lowercase letters [10]. Differences in capital letters can affect the analysis process because they can be considered as different words, so all words need to be converted to lowercase.
- Data cleaning is the process of cleaning data such as punctuation, and empty values in data [11]. Since any set of comments, whether on the web or on a social network, often contains non-letter characters and punctuation marks, they are deleted as their processing is also not essential for analysis.

3. RESULTS AND DISCUSSION

3.1. Web Scrapping

Web scraping is the process of retrieving semi-structured documents from the internet, usually web pages written in a markup language such as HTML or XHTML. The document is then analyzed to extract specific data that can be utilized in various contexts [12]. This process is also often referred to as screen scraping [13]. Web Scrapping conducted with Google Collaboratory tools and the Python programming language on the Shopee link on the Google Play Store obtained as much as 1000 of the latest review data in 2023. The procedure for performing web scraping on an application from the Google Play Store using a Jupyter Notebook [14]. This process is done by installing the google-play-scraper library, which is used to scrape review data on the Google Play Store. Simply by entering the app ID taken from the app link, the review data can be accessed. Examples of documents generated from web scrapping are as follows.

```
from google_play_scraper import Sort, reviews

# Scrape data ulasan aplikasi Shopee dari Google Play Store
result, continuation_token = reviews(
    'com.shopee.id',       # ID aplikasi Shopee di Google Play Store
    lang='id',             # Bahasa ulasan yang diambil adalah Bahasa Indonesia
    country='id',          # Negara diatur ke Indonesia
    sort=Sort.MOST_RELEVANT,  # Mengambil ulasan yang paling relevan
    count=2000,            # Jumlah ulasan yang diambil adalah 2000
    filter_score_with=None # Mengambil semua ulasan dari skor bintang 1 hingga 5
)

# Contoh manipulasi data hasil scrape:
# 1. Menampilkan ulasan pertama
print("Ulasan pertama:")
print(result[0])

# 2. Mengambil hanya ulasan dengan rating bintang 5
bintang_5 = [ulasan for ulasan in result if ulasan['score'] == 5]

print(f"\nJumlah ulasan bintang 5: {len(bintang_5)}")

# 3. Menyimpan data ulasan ke dalam file CSV
import pandas as pd

df = pd.DataFrame(result)
df.to_csv('ulasan_shopee.csv', index=False)
print("\nData ulasan berhasil disimpan dalam file 'ulasan_shopee.csv'")
```

Within 2 minutes, the program developed to perform web scrapping can collect around 12,000 review data on the shoppe application which includes the content of user reviews, the date the comment was made, and the score of the user for the application.

To perform web scrapping on the Google Play Store for the Shopee application, it takes 2 minutes and 20 seconds. The time is relative because it depends on the internet speed and computer specifications. However, in general, the time needed for data collection using the web scrapping method is shorter than manual data collection which takes more than one day. [15]. Moreover, web scrapping is carried out automatically using a program, making it more efficient in terms of human resources.

Besides being faster than manual data collection, the web scrapping method also has several advantages, including being able to minimize human error [16]. Repetitive manual work can cause boredom, leading to potential human errors such as typing errors or missing data to be inputted. Through the machine-run web scrapping method, these errors can be avoided [17]. In addition, the program for web scrapping can be further developed for other needs such as periodic reports [18].

Table 1. Example of Scrapped Documents

reviewId	username	userimage	content	score	thumbsUpCount	replyContent
130cb160-a061-4314-920e-bf7dd71f7d76	Chikka Risa	https://play-lh.googleusercontent.com/a/ACg8oc...	Aplikasi bagus, sayang di pengiriman lamaaaaaa...	1	153	hi kak, maaf ya buat kendala pesanan kakak, ha...
4e3ae6df-adb4-4b82-9954-11202eb3fad7	Dewi Nurlaela	https://play-lh.googleusercontent.com/a-/ALV-U...	Sangat muaskan belanja di shopee. selain lengk...	5	407	hi kak, makasih buat review bintang 5 nya, yuk...
e90791c8-915b-40f0-951a-08e88e51d6f4	Jernih Kurnia idawati Lahagu	https://play-lh.googleusercontent.com/a/ACg8oc...	Udh bertahun tahun Saya sering belanja di apli...	4	2	Hai kak, maaf yaa udh buat ga nyaman terkait k...

3.2. Text Preprocessing

Preprocessing is an important step that must be done before applying classification algorithms to documents [19]. The preprocessing stage includes the following processes:

a. Case folding

This process aims to convert all capital letters into lowercase letters so that the text has a variety of formats. In addition, this process also includes the removal of punctuation and irrelevant characters. An example of the results after case folding can be seen in Table 2.

Table 2. Process case folding

No	Content	Score	Label	Text_clean
1	Makin kesini proses pengirimannya mangkin lama…	1	Negatif	makin kesini proses pengirimannya makin lama…
2	Sakit ni aplikasi, lagi nonton youtube ngepaus…	1	Negatif	sakit ni aplikasi, lagi nonton youtube ngepaus…
3	Aplikasi ini mudah untuk belanja	5	Positif	aplikasi ini mudah untuk belanja

b. Data cleaning

This process aims to remove punctuation marks and empty values in the data. In addition, this process also includes the removal of punctuation and irrelevant characters. An example of the results of applying data cleaning can be seen in Table 3.

Table 3. Process data cleaning

No	Content	Score	Label	Text_Clean
1	Aplikasi bapuk. Dulu saya pernah bikin akun, d….	1	Negatif	aplikasi bapuk dulu saya pernah bikin akun dan…..
2	Sakit ni aplikasi, lagi nonton youtube ngepause…..	1	Negatif	sakit ni aplikasi laginonton youtobe ngepause……
3	Tetap kita yang harus pintar pintar pilih toko, s…..	5	Positif	tetap kitanya yang harus pintar pilih toko…..

In the research conducted by (Palamino, Marco, A. dkk), with the title "*Evaluating The Effectivenes Of Text Pre-Processing In Sentiment Analysis*" In his research, the researcher conducted quantitative pre-processing of sentiment analysis on twitter in order to analyze grammar and spelling in the twitter application because many words did not fit the rules. The method used by researchers is data collection taken directly from the twitter application and pre-processing such as case folding, removing URLs and twitter features, removing unnecessary spaces, removing punctuation marks, and removing stop words. The results show that the order of the pre-processing components is important and significantly improves the performance of the naive bayes classifier. Researchers also found lemmatization classifiers useful for improving index performance, but did not significantly improve the quality of sentiment analysis [20].

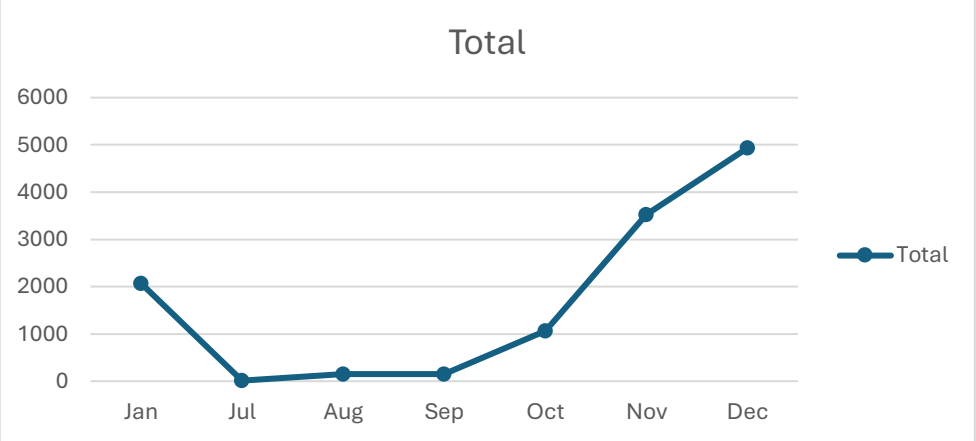

Fig. 2. Shopee review results

The graph above shows the total Shopee reviews for each month of the year. At the beginning of the year, the number of reviews reached a high number of around 2,000 reviews in January. However, there was a sharp drop to almost zero in July. In the period from August to September, the number of reviews remained at a very low level and stabilized. change in October, with a significant increase in the number of reviews. The upward trend continued until it peaked in December, with the total number of reviews exceeding 5,000. Here is the link to the Shopee review web scarping results dataset DOI:10.17632/9gkv3dpkgw.1 This pattern indicates that major promotions, such as Harbolnas in November and year-end discounts, are likely to be the driving factor for the increase in shopping activity and reviews on the platform.

4. CONCLUSION

The data obtained through the web scraping process of Shopee app reviews on the Google Play Store has great potential to be utilized in various studies, especially in the field of sentiment analysis. This pre-processed dataset can be used as a foundation to explore various sentiment analysis techniques, such as the use of other algorithms, such as Support Vector Machine (SVM), Random Forest, or deep learning-based approaches, to evaluate the performance and effectiveness in classifying user reviews. This dataset also has the potential to be used for other studies, such as consumer behavior analysis, user experience assessment, or the development of recommendation systems based on app reviews and ratings. Through this analysis, researchers can generate useful insights for app developers to improve their services.

To support research transparency and data accessibility, the results of this scraping are available through the Python google-play-scraper library. This library makes it easy for other researchers to replicate the research or extend the analysis with a larger dataset, simply by entering the app ID of the app you want to analyze to obtain the latest reviews from the Google Play Store. This dataset is not only relevant for research related to sentiment analysis on the Shopee app but can also make a broad contribution to other studies in the fields of e-commerce, user behavior analysis, and data-driven technology development.

Based on the results of the discussion along with the results of the analysis that has been presented, the conclusions include the process of collecting review data by performing web scrapping techniques on the Google Play Store obtained as much as 12,000 review data which then goes through the preprocessing process, but in this study researchers only preprocessed with two stages, namely, Case folding and data cleaning. Based on the results of the comparisons that have been made, it is found that out of 12,000 data there are negative labels higher than positive labels.

REFERENCES

[1] J. Homepage, N. C. Agustina, D. Herlina Citra, W. Purnama, C. Nisa, and A. Rozi Kurnia, "MALCOM: Indonesian Journal of Machine Learning and Computer Science The Implementation of Naïve Bayes Algorithm for Sentiment Analysis of Shopee Reviews on Google Play Store Implementasi Algoritma Naive Bayes untuk Analisis Sentimen Ulasan Shopee pada Google Play Store," vol. 2, pp. 47–54, 2022.

[2] Tania Puspa Rahayu Sanjaya, Ahmad Fauzi, and Anis Fitri Nur Masruriyah, "Analisis sentimen ulasan pada e-commerce shopee menggunakan algoritma naive bayes dan support vector machine," INFOTECH: Jurnal Informatika & Teknologi, vol. 4, no. 1, pp. 16–26, Jun. 2023, doi: 10.37373/infotech.v4i1.422.

[3] J. Y. M. Nip and B. Berthelier, "Social Media Sentiment Analysis," Encyclopedia, vol. 4, no. 4, pp. 1590–1598, Oct. 2024, doi: 10.3390/encyclopedia4040104.

[4] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiasari, "Analisis sentimen pada rating aplikasi Shopee menggunakan metode Decision Tree berbasis SMOTE," AITI: Jurnal Teknologi Informasi, vol. 18, no. Agustus, pp. 173–184, 2021.

[5] N. Agustina, D. H. Citra, W. Purnama, C. Nisa, and A. R. Kurnia, "Implementasi Algoritma Naive Bayes untuk Analisis Sentimen Ulasan Shopee pada Google Play Store," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 2, no. 1, pp. 47–54, 2022, doi: 10.57152/malcom.v2i1.195.

[6] B. Bayu Baskoro et al., "Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs TRIPADVISOR)," Journal Of Informatics, information system, software engineering application, vol. 3, no. 2, pp. 21–029, May 2021, doi: 10.20895/INISTA.V3I2.

[7] S. Khomsah and Agus Sasmito Aribowo, "Text-Preprocessing Model Youtube Comments in Indonesian," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, no. 4, pp. 648–654, Aug. 2020, doi: 10.29207/resti.v4i4.2035.

[8] Tania Puspa Rahayu Sanjaya, Ahmad Fauzi, and Anis Fitri Nur Masruriyah, "Analisis sentimen ulasan pada e-commerce shopee menggunakan algoritma naive bayes dan support vector machine," INFOTECH: Jurnal Informatika & Teknologi, vol. 4, no. 1, pp. 16–26, 2023, doi: 10.37373/infotech.v4i1.422.

[9] H. Benhar, A. Idri, and J. L. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review," Comput Methods Programs Biomed, vol. 195, p. 105635, Oct. 2020, doi: 10.1016/j.cmpb.2020.105635.

[10] H. Utami, "Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network," Indonesian Journal of Applied Statistics, vol. 5, no. 1, p. 31, May 2022, doi: 10.13057/ijas.v5i1.56825.

[11] Irma Surya Kumala Idris, Yasin Aril Mustofa, and Irvan Abraham Salihi, "Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM)," Jambura Journal of Electrical and Electronics Engineering, vol. 6, no. 6, pp. 823–848, Jan. 2023, doi: 10.1177/0165551510388123.

[12] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," Nov. 01, 2020, Elsevier B.V. doi: 10.1016/j.trac.2020.116045.

[13] U. Mufidah, M. Siahaan, and S. Informasi, "PERANCANGAN APLIKASI PERBANNDINGAN HARGA PRODUK (HISTORICAL DATA) MENGGUNAKAN TEKNIK SCRAPING WEB," 2021. Accessed: Jan. 05, 2025. [Online]. Available: http://pusdansi.org/index.php/pusdansi/article/view/12/12

[14] Y. A. Hafiz and E. Sudarmilah, "IMPLEMENTASI WEB SCRAPING PADA PORTAL BERITA ONLINE," Inisiasi, pp. 55–60, Nov. 2023, doi: 10.59344/inisiasi.v12i1.120.

[15] L. Hidayati, L. P. Kusuma, D. Agustini, and V. Y. P. Ardhana, "IMPLEMENTASI WEB SCRAPING UNTUK PENGUMPULAN DATA MEDIA SOSIAL LINGKUP PEMERINTAH PROVINSI NTB," Jurnal Sistem Informasi dan Informatika (Simika), vol. 7, no. 1, pp. 63–72, Mar. 2024, doi: 10.47080/simika.v7i1.3200.

[16] A. S. Yondra, D. Triyanto, and S. Bahri, "IMPLEMENTASI WEB SCRAPING UNTUK MENGUMPULKAN INFORMASI PRODUK DARI SITUS E-COMMERCE DAN MARKETPLACE DENGAN TEKNIK PEMROSESAN PARALEL," Coding Jurnal Komputer dan Aplikasi, vol. 10, no. 01, p. 93, May 2022, doi: 10.26418/coding.v10i01.52722.

[17] S. Wang et al., "Advances in Data Preprocessing for Biomedical Data Fusion: An Overview of the Methods, Challenges, and Prospects," Information Fusion, vol. 76, pp. 376–421, Dec. 2021, doi: 10.1016/j.inffus.2021.07.001.

[18] A. Z. Rizquina and C. I. Ratnasari, "Implementasi Web Scraping untuk Pengambilan Data Pada Website E-Commerce," Jurnal Teknologi Dan Sistem Informasi Bisnis, vol. 5, no. 4, pp. 377–383, Oct. 2023, doi: 10.47233/jteksis.v5i4.913.

[19] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," Pamukkale University Journal of Engineering Sciences, vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/pajes.2021.62687.

[20] M. A. Palomino and F. Aider, "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," Applied Sciences (Switzerland), vol. 12, no. 17, Sep. 2022, doi: 10.3390/app12178765.