Comparison of Naïve Bayes and Random Forest Algorithm in Webtoon Application Sentiment Analysis

Fadhila Tangguh Admojo[1], Slamet Risnanto[2], Ai Wulan Windiawati[3], Muhammad Innuddin[4], Desti Mualfah[5]

[1]*Malaysian Institute of Information Technology, Universiti Kuala Lumpur, 50250 Kuala Lumpur, Malaysia*

[2]*Department of Informatics, Sangga Buana University, 40124 Kota Bandung, Indonesia*

[3]*Institute of Advanced Informatics and Computing, 46115 Kota Tasikmalaya, Indonesia*

[4]*Department of Computer Science, Bumigora University, 83127 Kota Mataram, Indonesia*

[5]*Department of Informatics, Muhammadiyah Riau University, 28156 Kota Pekanbaru, Indonesia*

[1]*fadhila.tangguh@s.unikl.edu.my,* [2]*slamet.risnanto@usbypkp.ac.id,* [3]*aiwulan@iaico.org,* [4]*inn@universitasbumigora.ac.id,* [5]*destimualfah@umri.ac.id.*

ARTICLE INFORMATION

ABSTRACT

The Webtoon application has become one of the popular platforms for reading comics digitally. Webtoons, as a form of digital comics, present various types of comic content. The success of a Webtoon application depends greatly on understanding the preferences and views of its users. User evaluations of Webtoon applications can provide valuable insight into user satisfaction levels, as well as identify problems that need to be fixed by developers. In this research, Sentiment Analysis was applied to user reviews of the Webtoon Application on the Google Play Store. This research uses two different classification algorithms, namely Naïve Bayes and Random Forest, with the aim of comparing their performance in the context of sentiment analysis of user reviews of Webtoon applications. The results of this research are expected to provide an overview of the most suitable algorithm for conducting sentiment analysis classification in Webtoon applications. In collecting the dataset, we involved webtoon user reviews covering various sentiments, such as positive, negative, and neutral. However, in this analysis, the focus is given to two types of sentiment, namely positive and negative. We apply Naïve Bayes and Random Forest algorithms to perform sentiment classification on the reviews. Performance evaluation is carried out by considering metrics such as accuracy, precision, recall, and F1-score. The results of implementing these two algorithms are an accuracy of 74% Naïve Bayes, and 88% Random Forest. It can be concluded that the Random Forest algorithm is superior to the Naïve Bayes algorithm. With this, the Random Forest algorithm becomes a recommendation for classifying sentiment analysis for Webtoon applications with greater accuracy.

1. INTRODUCTION

In recent years, developments in digital technology have brought important changes, especially in the context of entertainment. One of the most important changes is the existence of digital-based reading comics, such as Webtoon applications or digital comics. Webtoon is an online platform that allows users to access thousands of comics works in a convenient and interactive way. Webtoons are divided by type, including genre and age [1].

The comics contained in the Webtoon application are not only made by professional comic artists [2], but also by beginners who can publish their work through this Webtoon application. From then until now, the development of digital comics or Webtoons has continued rapidly and has become increasingly popular globally [3].

The Webtoon application can be accessed or downloaded via the Google Play Store platform. Google Play Store is an Android user service that provides digital

content such as applications, games, films, music, and books [4]. Based on data from the Google Play site in 2023, this Webtoon application has been downloaded more than 100 million times and has a rating of 4.7. And on the Google Play site, it is recorded that this Webtoon application has 3 million review comments. Therefore, because the popularity of this Webtoon application continues to increase, understanding user reviews of this application is very important to see how developed this application is and see the shortcomings that exist in this application. Not only that, but currently user reviews are also a benchmark for the quality of the Webtoon application, so that positive and negative sentiments in these reviews can influence users.

To help overcome this, more in-depth research needs to be done on the Webtoon application [5]. The research process for this application can use sentiment analysis with the Naïve Bayes classification algorithm and the Random Forest classification algorithm. The results of this classification algorithm comparison will be a prediction of the user assessment level of this Webtoon application. Sentiment analysis is used in research on this application because sentiment analysis is a text data processing technique to extract information from a text [6]. And sentiment analysis is also a broad field of analysis that includes natural language processing and text evolution. Sentiment analysis is a good and appropriate tool for this research and allows gaining insights from unstructured data sets [7]. In the classification process, two algorithms are used to compare the accuracy of the two algorithms. The algorithms for classification in this research are the Naïve Bayes and Random Forest algorithms.

This research focuses on comparing the Naïve Bayes and Random Forest classification algorithms, and seeing to what extent the accuracy is better in the sentiment analysis process. The results of the accuracy level of the classification algorithm will be the results of sentiment analysis of Webtoon application users. This research was conducted to see the extent of accuracy between the two classification algorithms Naïve Bayes and Random Forest, and to compare the extent of differences in accuracy results from this sentiment analysis. Not only that, but this sentiment analysis process is also to help evaluate the sentiment of Webtoon Application users, to find out the extent of user satisfaction with application services, identify advantages and disadvantages of the application, and provide insight to developers for further improvement and development.

2. RELATED WORK

Almost similar research was conducted by Hana et al. that aspect-based sentiment analysis of Tentrem Yogyakarta Hotel Reviews using the Random Forest Classifier algorithm [8]. This research discusses sentiment analysis using the Random Forest algorithm by conducting tests for determine how good the algorithm is at prediction. And the results of this analysis are 90 % of the accuracy value and f1 score. But this research suggests trying using other algorithmic methods. The difference with this research is that this research only uses Random Forest.

The results of research by Kurniawan et al. in 2023 about Comparison of Naïve Bayes and SVM Algorithms in

Marketplace Sentiment Analysis on Twitter [9]. The aim of this research is to analyze and find out which algorithms better at analyzing marketplace user sentiment. The results of this algorithm get the highest and better accuracy, namely the SVM algorithm.

Then research conducted by Aprilia et al. in 2021 about Sentiment Analysis of Twitter Users at Flash Sale Events using the K-NN, Random Forest, and Naïve Bayes Algorithms [9]. This research aims to find out Twitter users about public opinion which is related to the Flash sale event on e-commerce. The results of this research use the Naïve Bayes algorithm as a recommendation for sentiment analysis with greater accuracy.

Naïve Bayes Algorithm This algorithm is widely used and is considered very suitable for the process of analyzing sentiment contained in datasets [10]. Because the Naïve algorithm is considered to have very good performance in classification [11]. Therefore, because this method is considered good, a comparison is carried out with other classification algorithms, namely the Random Forest algorithm. The Random Forest algorithm is an algorithm that has been widely used in the process of analyzing user reviews [12], because it is a learning technique based on decision trees. The process of this algorithm works by making the most class decisions according to the results of the decision tree [13], [14].

3. METHODOLOGY

The research method used in this research uses several stages, as described below:
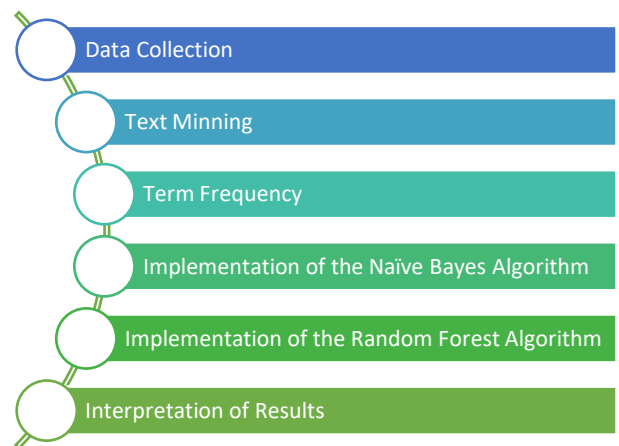

FIGURE 1. METHODOLOGY

3.1 Data Collection

The data collection process is the first step in this research. The data required is user reviews of the Webtoon Application on the Google Play Store. The data taken is in the form of review data, ratings, and other required information. This data collection process was carried out using web scrapping techniques from the Google Play application. Web scraping is very suitable for data retrieval because it can provide results that suit your needs. The data scraping process was carried out using Google collab and using the Python language to help the web scraping process run [15]. After successful data collection, this data must be cleaned and processed to ensure the quality of the data. The data taken in this research was 500 reviews or comments on the Webtoon application. The latest data was taken from October 2023 to November 2023.

3.2 Text Mining

This stage is the initial stage in the text mining process. Text mining is a stage in text preprocessing and is also a method commonly used to overcome classification problems [16], [17]. At this stage, the text will be changed into data which will be processed as follows:

a. The initial stage is to replace all existing letters in the data to lower case.

b. The Normalization Stage, in this process, is a cleaning process such as removing punctuation marks, lifting or special characters in the text.

c. The next stage is breaking down sentences into one word or called tokenizing.

d. The final stage here is the selection of words resulting from tokenizing. Such as selecting valuable words and deleting unnecessary words.

3.3 Term Frequency

Term Frequency or TF is a frequency that will appear from the number of times a word appears in existing data [18]. If the occurrence of a word is higher, the greater the value or weight of the word. Term Frequency (TF) is a fundamental metric used in natural language processing to quantify the importance of a term within a document or a corpus. It represents the frequency with which a particular word occurs in each piece of text relative to the total number of words in that document or corpus. Essentially, TF measures how often a term appears in the text.

3.4 Implementation of the Naïve Bayes Algorithm

At this classification stage, the Naïve Bayes algorithm is used because this algorithm will help to categorize sentiment into positive, negative, and neutral. However, this research only categorized them into two, namely positive and negative categories. This algorithm can classify and make probability predictions for appropriate class members based on predicted assumptions. The results of this classification will provide an overview of how users view the application being analyzed.

3.5 Implementation of the Random Forest Algorithm

In this process, Random Forest is implemented in sentiment analysis for the Webtoon application. Random forest is carried out to get the final value of the analysis process. The final value of the process is obtained from identifying most of each class that has been previously labeled. The process of this algorithm works by randomly splitting the data into a decision tree. Random Forest implementation is carried out after going through the Preprocessing and Tf-IDF stages.

3.6 Interpretation of Results

After successfully analyzing sentiment, the results of the sentiment analysis are interpreted to understand users' general views of the Webtoon application, and provide recommendations or suggestions based on the findings of the analysis that has been carried out. By leveraging the findings of sentiment analysis and translating them into actionable recommendations, developers and stakeholders can make informed decisions to drive continuous improvement and innovation, ultimately fostering greater user engagement and loyalty towards the Webtoon application.

4. RESULT AND DISCUSSION

In this research, the process of classifying Webtoon application review data was carried out. But before that, this process goes through several stages as explained in the research methodology section.

4.1 Data Preparation

In this process the research takes data or scraping from the Google Play Store application website. Web scraping is a technique for retrieving raw or unstructured data from websites without having to copy the data manually. The data taken is raw data in the form of user reviews of the application. In this scraping process, the Python programming language is used with the Jupyter Notebook and Google Colab libraries. Web scraping is a process. In this research, 500 webtoon application user reviews were taken

FIGURE 2. WEBTOON SCRAPING RESULTS

4.2 Text Preprocessing

Before the data is processed using the Naïve Bayes algorithm, the preprocessing stage is carried out first. This stage is included as part of text mining. This stage is a process of tokenizing, normalizing, stemming, and removing stop words. The results of this preprocessing stage can be seen in the image below.

FIGURE 3. PREPROCESSING STAGE

4.3 Naïve Bayes Classification

At this stage, the Naïve Bayes Classification stage is carried out using the Python programming language with the scikit-learn library. Before determining positive and negative sentiments, a labeling process is needed. The tagging process is the step of giving a label to each sentence. The marking step is carried out by calculating the frequency of words that appear to be labeled positive or negative. The results of calculations using the Naïve Bayes algorithm in this study produced an accuracy of 74%. In this sentiment analysis, it was stated that the results of sentiment analysis using 500 Webtoon application review data, produced a negative sentiment value of 26%, and a

positive sentiment of 85%. So, the user's view of this application is quite good because there are more positive sentiment accuracy values.

	precision	recall	f1-score	support
Negatif	1.00	0.15	0.26	27
Positif	0.73	1.00	0.85	63
accuracy			0.74	90
macro avg	0.87	0.57	0.55	90
weighted avg	0.81	0.74	0.67	90

FIGURE 4. RESULTS OF NAÏVE BAYES IMPLEMENTATION

4.4 Random Forest Classification

At this stage, the sentiment analysis process in the Webtoon application involves classification using the Random Forest algorithm. This algorithm analyzes user feedback to categorize sentiments expressed in reviews. The results of this classification process indicate that the model achieves an accuracy rate of 88%. This high accuracy suggests that the Random Forest algorithm effectively predicts the sentiment of user reviews in the Webtoon application, providing valuable insights into user satisfaction and preferences.

	precision	recall	f1-score	support
0	0.78	0.70	0.74	20
1	0.90	0.94	0.92	69
2	0.00	0.00	0.00	1
accuracy			0.88	90
macro avg	0.56	0.55	0.55	90
weighted avg	0.86	0.88	0.87	90

FIGURE 5. RESULTS OF RANDOM FOREST IMPLEMENTATION

4.5 Visualization

This section provides a visual representation of user feedback on the Webtoon application. Utilizing a word cloud format, it showcases frequently occurring words within the reviews. The size of each word corresponds to its frequency in the reviews: the larger the word, the more often it appears. This visualization offers a quick and intuitive way to grasp common themes and sentiments expressed by users regarding their experiences with the Webtoon application.

FIGURE 6. POSITIVE SENTIMENT WORDCLOUD

In Figure 6 there are several words that often appear, namely "Webtoon", "and", "I", "Good", "comic", and many more. The words that appear are words that often appear in positive reviews of Webtoon applications. The word "Webtoon" itself serves as a central keyword, highlighting the primary focus of the

application. Its frequent appearance suggests that users often mention the application by name, indicating a strong association with positive sentiments.

FIGURE 7. NEGATIVE SENTIMENT WORDCLOUD

In figure 7 is a wordcloud image of negative reviews. Here are several words that often appear in negative reviews of Webtoon applications.

5. CONCLUSIONS

The results of this research, comparing the Naïve Bayes and Random Forest algorithms in Webtoon Application Sentiment Analysis, obtained accuracy results of 74% Naïve Bayes, 26% negative sentiment, and 85% positive sentiment in Webtoon application user reviews. And the Random Forest algorithm produces an accuracy of 88%, 74% negative sentiment and 92% positive sentiment for the Webtoon application. So, it can be concluded that the classification results using the Random Forest algorithm are superior, so that sentiment analysis in the Webtoon application is better using the Random Forest algorithm, the results of which have quite high accuracy. And the results of comparing the accuracy values between the two algorithms are also quite far apart. The results of this research will provide information regarding the views of Webtoon application users, so that if the results of the user's views are negative then it will be input for the application developer, but if the results are positive then it will help increase popularity and help to increase user satisfaction so that creating a better image on the Webtoon application. The suggestions in this research for further research are to carry out analysis using different algorithms such as the Support Vector Machine (SVM) algorithm or using KKN. So, you can compare the performance between these algorithms. And if more classifications are carried out, the results obtained can be more accurate and better.

REFERENCES

[1] A. F. Lestari and I. Irwansyah, "Line Webtoon Sebagai Industri Komik Digital," *SOURCE : Jurnal Ilmu Komunikasi*, vol. 6, no. 2, p. 134, Oct. 2020, doi: 10.35308/source.v6i2.1609.

[2] J.-H. Kim and J. Yu, "Platformizing Webtoons: The Impact on Creative and Digital Labor in South Korea," *Soc Media Soc*, vol. 5, no. 4, p. 205630511988017, Oct. 2019, doi: 10.1177/2056305119880174.

[3] M. T. Ghulam, H. L. Nurtaat, H. Lail, and I. M. Sujana, "The Effectiveness of Using Webtoon Applications in Teaching Reading Comprehension at The Eighth Grade of SMP Negeri 11 Mataram," *Jurnal Ilmiah Profesi*

Pendidikan, vol. 8, no. 2, pp. 1043–1049, May 2023, doi: 10.29303/jipp.v8i2.1411.

[4] 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, Apr. 2022, doi: 10.57152/malcom.v2i1.195.

[5] M. Ariel and D. Furwana, "Developing the Content of Webtoon Comic Application as Supporting Media in Learning English Grammar," 2020.

[6] R. Wahyudi and G. Kusumawardana, "Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine," *Jurnal Informatika*, vol. 8, no. 2, pp. 200–207, Sep. 2021, doi: 10.31294/ji.v8i2.9681.

[7] F. Aftab *et al.*, "A Comprehensive Survey on Sentiment Analysis Techniques," *International Journal of Technology*, vol. 14, no. 6, p. 1288, Oct. 2023, doi: 10.14716/ijtech.v14i6.6632.

[8] H. Chyntia Morama, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen berbasis Aspek terhadap Ulasan Hotel Tentrem Yogyakarta menggunakan Algoritma Random Forest Classifier," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (JPTIIK)*, vol. 6, no. 4, pp. 1702–1708, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[9] I. Kurniawan *et al.*, "Perbandingan Algoritma Naive Bayes Dan SVM Dalam Sentimen Analisis Marketplace Pada Twitter," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 10, no. 1, 2023, [Online]. Available: http://jurnal.mdp.ac.id

[10] Samsir *et al.*, "Naives Bayes Algorithm for Twitter Sentiment Analysis," *J Phys Conf Ser*, vol. 1933, no. 1, p. 012019, Jun. 2021, doi: 10.1088/1742-6596/1933/1/012019.

[11] Ernianti Hasibuan and Elmo Allistair Heriyanto, "ANALISIS SENTIMEN PADA ULASAN APLIKASI AMAZON SHOPPING DI GOOGLE PLAY STORE MENGGUNAKAN NAIVE BAYES CLASSIFIER," *Jurnal Teknik dan Science*, vol. 1, no. 3, pp. 13–24, Oct. 2022, doi: 10.56127/jts.v1i3.434.

[12] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst Appl*, vol. 237, p. 121549, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.

[13] F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, "Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (JPTIIK)*, vol. 6, no. 9, pp. 4305–4313, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[14] S. Lee, C. Lee, K. G. Mun, and D. Kim, "Decision Tree Algorithm Considering Distances Between Classes," *IEEE Access*, vol. 10, pp. 69750–69756, 2022, doi: 10.1109/ACCESS.2022.3187172.

[15] Fuad Amirullah, Syariful Alam, and M.Imam Sulistyo S, "Analisis Sentimen Terhadap Kinerja KPU Menjelang Pemilu 2024 Berdasarkan Opini Twitter Menggunakan Naïve Bayes," *STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer*, vol. 2, no. 3, pp. 69–76, Aug. 2023, doi: 10.55123/storage.v2i3.2293.

[16] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *KLIK: Kajian Ilmiah Informatika dan Komputer*, vol. 2, no. 5, pp. 178–185, Apr. 2022, doi: 10.30865/klik.v2i5.362.

[17] T. Aksoy, S. Gülseçen, and S. Çelik, "Data Pre-processing in Text Mining," in *Who Runs the World: Data*, Istanbul University Press, 2020, pp. 123–144. doi: 10.26650/B/ET06.2020.011.07.

[18] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *Int J Gen Syst*, vol. 46, no. 1, pp. 27–36, Jan. 2017, doi: 10.1080/03081079.2017.1291635.

AUTHORS

Fadhila Tangguh Admojo

He is a senior lecturer at the information systems department of Bina Darma University and pursuing a PhD at MIIT University of Kuala Lumpur. He researches field focuses on research Image Processing, NLP, and Information Retrieval.

Slamet Risnanto

He is a senior lecturer in the Computer Science department, Sangga Buana University YPKP Bandung Indonesia. He is the dean of the engineering faculty and is currently pursuing a PhD at UTEM, Malaysia. His research is in the fields of Software Engineering and Information Technology.

Ai Wulan Windiawati

Currently she is a researcher at the Institute of Advanced Informatics and Computing (IAICO), Tasikmalaya, Indonesia. Her research focuses on the fields of information systems, information technology and sentiment analysis.

Muhammad Innuddin

He is a lecturer in the Computer Science department, Faculty of Engineering, Bumigora University, Nusa Tenggara Barat, Indonesia. His research currently focuses on Data Mining, Classification Methods, and Data Imbalance.

Desti Mualfah

She is a dedicated lecturer in the Informatics Engineering study program at the Faculty of Computer Science, Muhammadiyah University of Riau, Indonesia. With a passion for academia, she actively engages in teaching and mentoring students while contributing to the advancement

of knowledge in her field. Her current research endeavors encompass various domains within information technology, including network security and sentiment analysis. Her expertise and make meaningful contributions to the field of computer science.