Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 3 No. 2 2023, 76 - 83
DOI: *https://doi.org/10.22487/27765660.2023.v3.i2.16919*
eISSN: 2776-5660

SENTIMENT ANALYSIS OF REVIEW DATA OF THE RUANGGURU ONLINE LEARNING APPLICATION USING THE C5.0 ALGORITHM

Nurul Izzah[1], Nur'eni[2], Rizka Pitri[3*]

[1,2]Statistics Study Program, Tadulako University
[3]Islamic Library and Information Science Study Program, Raden Intan State Islamic University

***e-mail**: *rizka@radenintan.ac.id*

ABSTRACT

Sentiment analysis is process to determine the sentiment of a person that is manifested in the form of text. Internet users write their opinions and everything that concerns them in the google play store review column. Moreover, when the world of education could not carry out face-to-face learning due to the covid-19 pandemic, learning turned to e-learning applications. Through this innovation, many pros and cons flow from the community with the existence of Ruangguru online learning application in the world of education. This research was conducted with the aim of determining word cloud visualization and the accuracy of the results of sentiment analysis of review data on the Ruangguru application using the C5.0 algorithm. The word cloud visualization results are dominated by word such as "paham", "bagus", "mudah", "suka", "langganan", "seru", "nyaman", "senang", "menarik", "keren", "lancar", "sukses". This shows that Ruangguru Application is a good application because it is dominated by positive sentiment words which means that users find it helpful and easy to understand the learning material in Ruangguru. The results of the Confusion Matrix show that the model successfully classifies 0.8557 or 85.57% of the data correctly from all test data.

Keywords: *Ruangguru, Sentiment Analysis, Classification, C5.0 Algorithm*

Cite: Izzah, N., Nur'eni & Pitri, R. (2023). *Sentiment Analysis of Review Data of the Ruangguru Online Learning Application Using the C5.0 Algorithm*. *Parameter: Journal of Statistics, 3*(2), 76-83, https://doi.org/10.22487/27765660.2023.v3.i2.16919.

Copyright © 2023 Izzah et al. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[*]Received: Dec 2023; Reviewed: Dec 2023; Published: Dec 2023

INTRODUCTION

Education is one sector that is influenced by technological developments. Currently, rapid technological progress has become an important element in the world of education (Wahyono, 2019). The increasing development of information technology means that learning methods are required to follow technological developments. The learning method used should be able to utilize various media to improve the quality of learning outcomes. One of them, learning media can use e-learning or electronic-based learning using computers. Mobile learning or often referred to as m-learning is defined as e-learning via mobile computing devices or is the delivery of electronic learning materials on mobile computing devices with the aim of making it easier to use so that it can be accessed anywhere and at any time (Listyorini and Widodo, 2013). Moreover, at a time when the world of education could not carry out face-to-face learning due to the Covid-19 pandemic, this had a huge impact on the education sector so that learning shifted to e-learning applications. One of the existing non-formal education companies is Ruangguru. In Indonesia, Ruangguru is the most popular education startup as of the first quarter of 2022 according to a Daily Social survey (Databoks, 2022). Ruangguru has received the nickname as the best interactive tutoring application, making Ruangguru an online learning application that is in great demand. Ruangguru connects students with the right teachers to help students learn new knowledge and get learning solutions outside of school. The Ruangguru application can be downloaded on the *Google Play Store* and *iOS App Store*. Ruangguru is the largest technology company in Indonesia that focuses on education-based services and has more than 15 million users and manages 300,000 teachers who offer services in various subject areas. Ruangguru develops various technology-based learning services, including virtual class services, online exam platforms, subscription learning videos, private tutoring marketplace, and other educational content that can be accessed via the Ruangguru web and application (Ruangguru, 2020). With the birth of this innovation, there were many pros and cons flowing from society regarding the existence of Ruangguru in the world of education. Various comments have appeared in the Google Play Store review column. To find out the assessment of a mobile learning application, you can use sentiment analysis. Sentiment analysis is a process for analyzing opinions, sentiments, assessments and emotions from someone's statements regarding a domain (Dharmawan et al., 2020). This research applies the field of data mining, especially classification techniques using the decision tree method. Decision tree is a classification technique for a set of objects or data with tree representation, one of the decision tree algorithms is the C5.0 algorithm. The C5.0 algorithm is a decision tree algorithm that can analyze data into a set of rules which it is hoped can later be used as input in decision making. In research by Balamurugan and Kannan (2016) using the C5.0 algorithm, carrying out 2 experiments with different datasets and sampling techniques showed high accuracy, namely 78.79% and 93.82%. Then research by Albances et al (2018) carried out flu predictions by applying the C5.0 algorithm using Twitter data. In terms of precision and efficiency, the Naive Bayes algorithm is better than the C5.0 algorithm. However, the C5.0 is better in terms of accuracy coming in at 66%. So the solution proposed in this research, namely the C5.0 algorithm, has better accuracy compared to the Naive Bayes classifier. Based on this, researchers will conduct research on sentiment analysis of Google Play Store review data on the Ruangguru online learning application using the C5.0 algorithm.

MATERIALS AND METHODS

C5.0 Algorithm

The C5.0 algorithm is an algorithm which is a refinement of the C4.5 algorithm which uses a tree-shaped representation where each node represents an attribute, then the branches present the value of the attribute and have what are called leaves where the function is class. Decision making is based on the largest gain value from the calculation results of all attributes. The following are the steps for using the C5.0 algorithm according to Dalbergio et al (2019) :

1. Calculate the entropy value for each variable category using the entropy formula.
2. Calculate the gain value of each variable using the gain formula.
3. Select the variable that has the highest gain value as the parent node.
4. Create a branch from each parent node category.
5. Check whether the entropy value of each node member has a value of zero. If you get a value of 0, then determine which leaves have formed. If the entropy value of each member node is completely zero, then the process stops.

The C5.0 algorithm model decision tree begins by calculating the entropy value as a determinant of the attribute impurity level and gain value. After that, the gain ratio value is calculated after the gain calculation above is carried out.

Data Source

This research uses secondary data regarding someone's review on the Google Play Store website regarding the Ruangguru application. Data was taken through scraping from the Google Play Store web using python software. The data scraping results are then labeled as positive sentiment class and negative sentiment class.

Analysis Method

Data analysis in this research used the C5.0 algorithm with the help of Python and R Studio software. The stages of analysis carried out are as follows:

1. Collection of review data from the Google Play Store website using the web scraping method.
2. Manually label the data with 2 categories, namely negative and positive.
3. Carry out text preprocessing, which includes case folding, removing punctuation, filtering, tokenizing and stemming.
4. Visualize data using Word Cloud.
5. Word weighting uses Term Frequenzy-Inverse Document (TF-IDF) to weight words that were originally text data into numerical data.
6. Divide training data and test data.
7. Carry out classification using the C5.0 algorithm
8. Evaluate the classification results using the confusion matrix.
9. Conclusion.

RESULTS AND DISCUSSION

Descriptive Analysis

Descriptive analysis in this research is used to see a general overview of information about the Ruangguru Application based on user review data from the Google Play Store. The following is a rating that describes the assessment given by users to the Ruangguru application on the Google Play Store, presented in Figure 1 below.

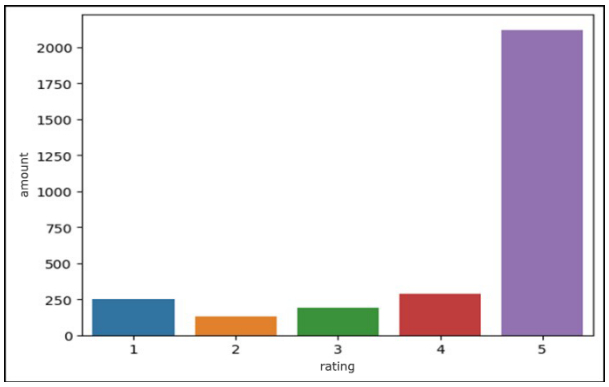

Figure 1. Number of Reviews Based on Ratings

Based on Figure 1. it can be seen that the majority of Ruangguru users have good ratings. This is proven based on the number of user ratings from 2,979 reviews, there are 2120 users giving the rating "Very Like" (rating 5), 288 users giving the rating "Like" (rating 4) and 192 users giving the rating "Quite Like" (rating 3) , while for the "Dislike" rating (rating 2) there were 129 and 250 reviews in the "Strongly Dislike" category (rating 1).

Data Collection and Labeling

The first stage in carrying out the sentiment analysis process is data collection. In this research, data was taken from the Google Play Store website. The data collection technique used is a web scraping technique. The web scraping technique is a technique for getting information from a page automatically without having to copy it manually (Ayani et al, 2019). The overall dataset is 2979 review data. After the data has been successfully collected into a dataset, the next stage is data labeling. Labeling here is

intended to divide data into several sentiment classes that will be used in research. The number of sentiment classes used in this research are two classes, namely the negative class and the positive class. The purpose of this labeling process is to divide the dataset into two parts, namely training data and test data. The data labeling process is carried out manually and by sentiment rating.

Based on the results of web scraping 2979 review data on the Ruangguru Application, we obtained a positive sentiment class of 2408 reviews and a negative sentiment class of 571 reviews which is presented in Figure 2 below.

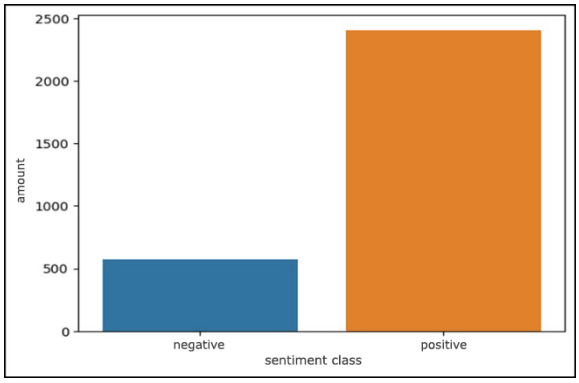

Figure 2. Number of Negative and Positive Sentiments

In the research carried out, examples of words belonging to the positive sentiment class and the negative sentiment class were obtained. Examples of words that fall into the positive sentiment class are “paham”, “bagus”, “mudah”, “suka”, “langganan”, “seru”, “nyaman”, “senang”, “menarik”, “keren”, “lancar”, “sukses”. Meanwhile, examples of words that fall into the negative sentiment class are “error”, “susah”, “ganggu”, “bosan”, “kecewa”.

Text Preprocessing

Data obtained through the web scraping process needs to be cleaned first before being processed by the machine. This process is called text preprocessing. The preprocessing steps carried out are case folding, removing punctuation, filtering, tokenizing and stemming.

Case Folding

Reviews before and after case folding are presented as examples Table 1

Table 1 Before and After Case Folding Review

Before	After
Aplikasi yg bermanfaat untuk mengajar dan teman2 yg belajar karena banyak manfaatnya	**a**plikasi yg bermanfaat untuk mengajar dan teman2 yg belajar karena banyak manfaatnya
Bagus banget Pas belajar pintar, cerdas banget	**b**agus banget **p**as belajar pintar, cerdas banget
Bagus, video pembelajarannya juga oke, aku suka, apalagi ada teman belajarnya, buat aku pengen belajar terus.	**b**agus, video pembelajarannya juga oke, aku suka, apalagi ada teman belajarnya, buat aku pengen belajar terus.

Remove Punctuation

Remove Punctuation is the stage for removing punctuation characters such as numbers, question marks, commas, colons and so on. Reviews before and after remove punctuation are presented as examples in Table 2.

Table 2 Before and After Remove Punctuation Review

Before	After
aplikasi yg bermanfaat untuk mengajar dan teman2 yg belajar karena banyak manfaatnya	aplikasi yg bermanfaat untuk mengajar dan teman yg belajar karena banyak manfaatnya
bagus banget pas belajar pintar, cerdas banget	bagus banget pas belajar pintar cerdas banget
bagus, video pembelajarannya juga oke, aku suka, apalagi ada teman belajarnya, buat aku pengen belajar terus.	bagus video pembelajarannya juga oke aku suka apalagi ada teman belajarnya buat aku pengen belajar terus

Filtering

The filtration stage is the stage of taking important words from the token results. The stoplist algorithm (removing less important words) or wordlist (keeping important words) can be used at this stage. Reviews before and after filtering are presented as examples in Table 3.

Table 3 Before and After Filtering Review

Before	After
aplikasi **yg** bermanfaat untuk mengajar **dan** teman **yg** belajar **karena banyak** manfaatnya	aplikasi bermanfaat mengajar teman belajar manfaatnya
bagus **banget pas** belajar pintar cerdas **banget**	bagus belajar pintar cerdas
bagus video pembelajarannya **juga oke aku** suka **apalagi ada** teman belajarnya **buat aku pengen** belajar **terus**	bagus video pembelajarannya suka teman belajarnya belajar

Tokenizing

In this tokenization process, the tokenizer carries out its task of dividing a sentence into several parts such as words, phrases or other meaningful elements. Reviews before and after tokenizing are presented as examples in Table 4.

Table 4 Before and After Tokenizing Review

Before	After
aplikasi bermanfaat mengajar teman belajar manfaatnya	['aplikasi','bermanfaat','mengajar', 'teman', 'belajar', 'manfaatnya']
bagus belajar pintar cerdas	['bagus', 'belajar', 'pintar', cerdas']
bagus video pembelajarannya suka teman belajarnya belajar	['bagus','video','pembelajarannya', 'suka','teman','belajarnya','belajar']

Stemming

Stemming is done by removing affixes that start and end words so that the basic form of the word is obtained. Reviews before and after case folding are presented as examples in Table 5.

Table 5 Before and After Stemming Review

Before	After
['aplikasi','bermanfaat','mengajar', 'teman', 'belajar', 'manfaatnya']	aplikasi manfaat ajar teman ajar manfaat
['bagus', 'belajar', 'pintar', cerdas']	bagus ajar pintar cerdas
['bagus','video', 'pembelajarannya', 'suka','teman','belajarnya','belajar']	bagus video ajar suka teman ajar ajar

Word Cloud Formation

After going through the text preprocessing stage, the Ruangguru application review data can be visualized in Word Cloud form. The following are the word cloud results of the Ruangguru Application review which can be seen in Figure 3.

Figure 3. Word Cloud Ruangguru Application Review

Based on the word cloud results in Figure 3, there are several words with a larger size than other words. This shows that a larger word size indicates that the frequency of occurrence of the word is quite high.

TF-IDF weighting

The text data will first be converted into vector form using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. The following is a TF-IDF calculation using three documents, namely:

Document 1 (d_1) = mudah mengerti gampang bosan
Document 2 (d_2) = mudah ajar seru mengerti simak
Document 3 (d_3) = bantu selesai soal kerja

Next, the weight value can be calculated for each term in the query in each document as an example as in Table 6.

Table 6 Example of TF-IDF Weight Calculation Results

Q	tf_{ij}			dj	D/d_j	$\log D/d_j$	W_{ij}		
	d_1	d_2	d_3				d_1	d_2	d_3
mudah	1	1	0	2	1.5	0.1760913	0.176091	0.176091	0
bosan	1	0	0	1	3	0.4771213	0.477121	0	0
kerja	0	0	1	1	3	0.4771213	0	0	0.477121
Nilai bobot masing-masing dokumen							0.653213	0.176091	0.477121

Apart from that, it can also be seen that the weight with the first rank is on d_1, namely 0.65321, followed by the weight with the second rank, which is on d_3, namely 0.47712. Meanwhile, the weight with the third rank is d_2, namely 0.17609. The weight value of a document is directly proportional to the level of similarity of the document to the query being searched for. Therefore, among the three documents that have the highest level of similarity to the queries "mudah", "bosan", "kerja" is d_1. In other words, the highest level of relevance is owned by d_1 and the lowest level of relevance is owned by d_2.

Distribution of Training Data and Test Data

The data sharing ratio is 80% for training data, the remaining 20% will be test data. The following details of the distribution of training data and test data are presented in Table 7.

Table 7 Data Distribution

Data Distribution	Amount	Percentage
Data Latih	2383	80%
Data Uji	596	20%
Jumlah	2979	100%

Decision Tree Classification Model Using the C5.0 Algorithm

Decision Tree begins by calculating the entropy value as a determinant of the gain value. After the gain value calculation is carried out, the highest gain value or gain ratio will be calculated. Below, a more complete decision tree can be seen in the resulting tree as in Figure 4 below.

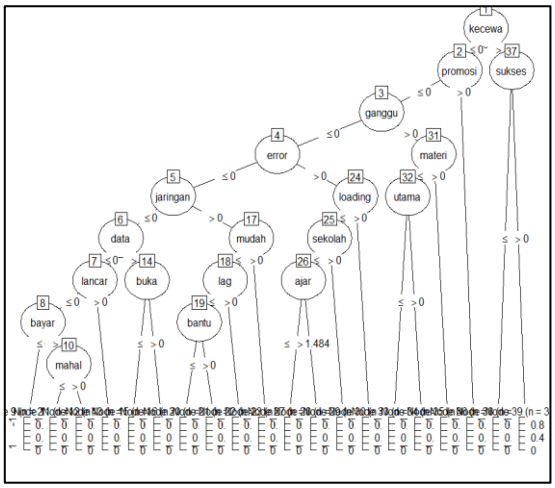

Figure 4. Decision Tree for Ruangguru Application Review

Based on Figure 4, it can be seen that the variable that plays the biggest role in the model is "disappointed" where this variable is used as the parent node.

Evaluation of Classification Results with Confusion Matrix

The classification process is carried out on the training data, then the classification accuracy is calculated on the test data. Evaluation of the classification model results is presented in Table 8.

Tabel 8 Evaluation of Classification Model Results

Actual	**Predict**	
	Positive	**Negative**
Positif	469	70
Negatif	16	41

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% = \frac{469 + 41}{469 + 70 + 16 + 41} \times 100\% = 85.57\ \%$$

$$\text{Specitivity} = \frac{TP}{TP + FP} \times 100\% = \frac{469}{469 + 16} \times 100\% = 96.70\ \%$$

$$\text{Sensitivity} = \frac{TN}{TN + FN} \times 100\% = \frac{41}{41 + 70} \times 100\% = 36.93\ \%$$

Based on Table 8, the total number of test data is 596 review data. The model correctly classifies 469 predicted positive sentiments as positive class and 41 predicted negative sentiments as negative class. However, there are 16 negative sentiment classes which are predicted as positive sentiment classes and there are 70 positive sentiment classes which are predicted as negative classes. The accuracy value shows that the model succeeded in predicting 85.57% of the data correctly from all test data. The resulting specificity value shows that the model succeeded in classifying data with a truly negative class of 96.70%. Furthermore, the resulting sensitivity value shows that the model succeeded in classifying data with a truly positive class of 36.93%.

CONCLUSION

Based on the results and discussions that have been carried out previously, it can be concluded that the word cloud visualization results are dominated by the words "paham", "bagus", "mudah", "suka", "langganan", "seru", "nyaman", "senang", “menarik”, “keren”, “lancar”, “sukses”. This shows that the Ruangguru application is a good application because it is dominated by positive sentiment words, which means users feel helped and easily understand the learning material on the Ruangguru application. The confusion matrix results show that the model succeeded in classifying 0.8557 or 85.57% of the data correctly from all test data.

REFERENCES

Albances, L. Z., Bungar, B. A., Patio, J. P., Sevilla, R. J. M., and Acula, D. (2018). Application of C5. 0 Algorithm to Flu Prediction Using Twitter Data. *2018 International Conference on Platform Technology and Service (PlatCon)*, 1-6.

Ayani, D. D., Pratiwi, H. S., dan Muhardi, H. (2019). Implementation of Web Scraping for Data Retrieval on Marketplace Sites. *Journal of Information Systems and Technology (JUSTIN)*, 7(4), 257.

Balamurugan, M., dan Kannan, S. (2016). Performance Analysis of Cart and C5.0 using Sampling Techniques. *International Conference on Advances in Computer Applications (ICACA)*, 72-75.

Dalbergio, D., Hayati, M. N., dan Nasution, Y. N. (2019). Classification of Student Length of Study Using the C5.0 Method in Case Studies of Student Graduation Data from the Faculty of Mathematics and Natural Sciences, Mulawarman University in 2017. *Pros. Semin. Nas. Mat. Stat. dan Apl*, 1(1), 36–42.

Dharmawan, L.R., Arwani, I., dan Ratnawati, D.E. (2020). "Sentiment Analysis on Twitter Social Media on Brawijaya University Student Academic Information System Services using the KNearest Neighbor Method". *Journal of Information Technology and Computer Science Development*, 4(3), 959–965.

Databoks. (2022). *Survey: Ruangguru is the most popular educational startup in Indonesia*. Retrivied from https://databoks.katadata.co.id/.

Listyorini, T., dan Widodo, A. (2013). Designing Mobile Learning for Android-Based Operating System Courses. *Symmetric J. Tech. Mechanical, Electrical and Computer Science*, 3(1), 25.

Ruangguru. (2020). *About Ruangguru*. Retrivied from https://www.ruangguru.com/about/.

Wahyono, H. (2019). Utilization of Information Technology in Assessing Learning Outcomes of the Millennial Generation in the Era of Industrial Revolution 4.0. *Proceeding Biol. Educ.*, 3(1), 192–201.