Computing and Information Processing Letters
Vol. 1., No. 1, November 2021, pp. 25-30
http://jurnal.upnyk.ac.id/index.php/cip/index

Implementation of the c4.5 decision tree learning algorithm for sentiment analysis in e-commerce application reviews on google play store

Asrina Yuni Ikhsanti [a,1], Yuli Fauziah [a,2,*], Rifki Indra Perwira [a,3]

[a] Department of Informatics, Universitas Pembangunan Nasional Veteran Yogyakarta, 55281, Indonesia
[1] 12316008o@student.upnyk.ac.id; [2] yuli.fauziah@upnyk.ac.id*; [3] rifki@upnyk.ac.id
* Corresponding Author

ARTICLE INFO

Article history
Received: November 1, 2021
Revised: November 7, 2021
Accepted: November 9, 2021

Keywords
Sentiment Analysis
C4.5
E-commerce

ABSTRACT

Objective: Forknowing the level of accuracy of the C4.5 Decision Tree Learning Algorithm in sentiment analysis of reviews of e-commerce applications on the google play store.

Design/method/approach: Using C4.5 AlgorithmDecision Tree Learning.

Results: This study uses a confusion matrix test with a comparison of 80% for training data and 20% for test data, where 750 is used for training data and 190 is used for test data. This test obtained an average accuracy of 92.63%, precision 69.58%, and recall 69.99%.

Authenticity/state of the art: In this study using the C4.5 Algorithm Decision Tree Learningto conduct sentiment analysis of e-commerce reviews, which use the gain value to perform feature selection. There are four categories, namely display, service, access, and product. The data in this study were obtained from the google play store.

1. Introduction

The development of technology, information and communication today provides benefits in various aspects. One of them in the economic field is the buying and selling process. Research results from Bloomberg predict that by 2020 more than half of Indonesia's population will be involved in e-commerce activities[1]. One of the e-commerce application service providers is Google Play.

Google play as a provider of mobile application services, including e-commerce applications, already has at least 1,200,000 applications available[2]. With so many applications, the competition is getting tougher and consumers must be selective in choosing where to shop[3]. One way to choose the right e-commerce application is to read the reviews available on Google Play. Based on Bright Local in its report Local Consumer Review Survey, 84% of consumers trust online reviews more than recommendations from someone[4]. The purpose of the review is to evaluate and improve the quality of the product in the future[5]. The review feature provided on Google Play is a means for users to convey impressions, criticisms, and suggestions. There are two kinds of sections in Google Play reviews, namely ratings and textual comments reviews. Ratings can be in the form of overall

evaluation scores, while textual reviews can contain comments that can tell a deeper story. These reviews can be used as consideration for choosing an e-commerce mobile application.

There are several factors that influence a consumer to choose an e-commerce in purchasing an item or product they want. Among them are the quality of e-commerce which is characterized by easy communication with sellers or companies, easy to learn and operate applications, providing reliable information, and providing detailed information about products.[6].

To get information about an e-commerce application, it can be accessed through reviews on the Google Play Store. If reading the review as a whole can take time and vice versa if only a few reviews are read, the evaluation results will be biased[7]. Therefore, a research was made on sentiment analysis of google play reviews on e-commerce applications that can classify consumer reviews into positive, negative, and neutral opinions. The research on sentiment analysis of e-commerce applications that has been done previously is using the Naive Bayes Classifier method with an accuracy of 97.4%. However, the research only focuses on one e-commerce application, namely Shopee by classifying sentiment into two classes, namely positive and negative.[8].

Research that has been done to analyze sentiment using the C4.5 Algorithm to classify mass media reviews on the Yelp site. The results showed that the forward selection method for feature selection and the C4.5 algorithm produced better accuracy, compared to previous studies where the highest accuracy was 80.00%. In the analysis of the mass media review sentiment using C4.5 with the selection of the forward selection feature, the accuracy result is 84.00%. Thus, it can be concluded that classification research using the C4.5 algorithm on sentiment analysis of mass media reviews can be increased in accuracy by using a forward selection of 4.00%.[7].

Based on the problems described above, the application of the C4.5 Decision Tree Learning Algorithm is the solution offered in this study. From the process of doing sentiment analysis on e-commerce, it will provide recommendations and assist consumers in choosing the appropriate e-commerce application. There are ten e-commerce applications that will be analyzed later, including Tokopedia, Shopee, Bukalapak, Lazada, Blibli, JD.ID, Blanja, Elevenia, OLX, and Zalora. All of these e-commerce will be compared using predetermined methods to find out which e-commerce is good in the eyes of consumers through google play reviews. In this study, we will use data sourced from Google Play reviews on e-commerce applications. Other than that, It is hoped that this research will produce an intelligent system that can classify comment reviews into positive, neutral, and negative sentiments with high accuracy. So that the results of sentiment analysis can be used as consideration to determine the appropriate e-commerce application.

2. Method

The research method used is a quantitative method. The stages of the research that will be carried out are web scraping, labeling datasets, text preprocessing, modeling with the C4.5 Algorithm, and testing. The stages of the research carried out can be seen in Fig. 1.

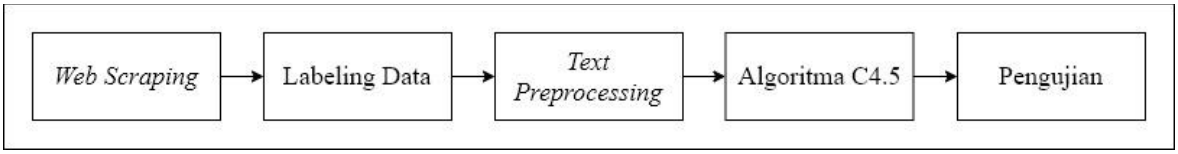

Fig. 1. Research Stages

2.1. Data Collection

The data in this study uses secondary data obtained from the web scraping process from e-commerce reviews on the google play store. The collected data is then manually purchased labels in the form of positive, negative, or neutral labels. The dataset used in this study amounted to 940 data, with the number of positive labels 415, negative 471, and neutral 54. Furthermore, the dataset

will be divided into 80% training data and 20% test data. The following results of web scraping reviews that have been carried out by the labeling process can be seen in Table 1.

Table 1. Output Class Label

No.	Review	Sentiment	Category
1.	First, because yesterday using the application was quite complicated later, if it's good, I'll add a star, okay?	Negative	Appearance
2.	Always satisfied shopping at shopee... Delivery is safe, fast and I really like the point of shopping... The best, really recommended.👍👍👍👍	Positive	Service
3.	disappointed with the shopee network.. this week I can't open again... slow and can't choose shopping items... since the last update shopee can't use it for shopping... forced me shopping at lazada...	Negative	Access
4.	Very helpful, the cheapest price compared to other applications	Positive	Product

2.2. Text Preprocessing

Text preprocessing are the steps taken so that a text / document is ready to be processed for the next process. The purpose of preprocessing is to prepare text documents into data that is ready to be processed at the next stage and to reduce noise[9]. In addition, the purpose of preprocessing is to clean and homogenize words so that they are ready to be extracted to the next stage[10]. The stages of text preprocessing in this study are case folding (converting uppercase letters to lowercase letters as a whole), cleansing (removing characters or punctuation that are not needed in the analysis process), tokenizing (breaking sentences into tokens), spelling normalization (replacing words that are not appropriate). spelling or abbreviations into the original word), stopword removal (removing words that are considered to have no effect in a sentence), and stemming (changing words into basic words by removing affixes in words). The following results from the text preprocessing process can be seen in Table 2.

Table 2. Output Class Label

Preprocessing	Text
Input	Very satisfied shopping at shopee, the goods are good and cheaper than other apks. Highly recommend shopping here!!! 👍
Case Folding	very satisfied shopping at shopee, the goods are good and cheaper than other apks. highly recommend shopping here!!! 👍
Cleansing	very satisfied shopping at shopee the goods are good and cheaper than other apks. I highly recommend shopping here
Tokenizing	[satisfied] [very] [shopping] [dishopee] [goods] [good] [and] [more] [cheap] [than] [apk] [other] [recommend] [very good] [lah] [shopping] [here]
Spelling Normalization	[satisfied] [very] [shopping] [dishopee] [goods] [good] [and] [more] [cheap] [from] [app] [other] [recommend] [very good] [lah] [shopping] [here]
Stopword Removal	[satisfied] [shopping] [dishopee] [goods] [good] [cheap] [app] [recommendation] [shopping]
Stemming	[satisfied] [shopping] [shopee] [goods] [good] [cheap] [app] [recommendation] [shopping]

2.3. C4.5 Decision Tree Algorithm

The C4.5 algorithm is a development of the ID3 algorithm, where the ID3 algorithm only handles a few and discrete attribute values. As for the C4.5 algorithm, it is able to handle continuous attribute values [11]. The advantages of the C4.5 algorithm are that it can handle continuous and discrete attributes, then it can handle training data with missing values, can process large and complex

datasets, and use gain ratios to improve information gain [11]. The disadvantages of the C4.5 Algorithm are:bias towards small distribution[12].

The C4.5 algorithm works by separating samples based on the attributes that produce the highest information gain value. Model C4.5 can separate samples based on the highest information gain value, a subset of samples obtained from the previous separation will be separated afterwards. The process will continue until the sample subset is inseparable and usually matches other attributes. Finally, check for the lowest level split, the subset of samples that do not contribute to the model will be rejected[13]. To calculate the information gain, the first step is to calculate the class entropy value with the following equation:

$$\text{Entropy}(\text{S}) = -\log 2 - \log 2 \frac{\text{Pos}(\text{S})}{\text{S}} \frac{\text{Pos}(\text{S})}{\text{S}} \frac{\text{Neg}(\text{S})}{\text{S}} \frac{\text{Neg}(\text{S})}{\text{S}} \quad (1)$$

After calculating the class entropy, the next step is to calculate the entropy of each attribute using equation (1). Then calculate the value of informaton gain with the following equation:

$$Gain(\text{S}, \text{A}) = \text{Entropy}(\text{S}) - (\frac{\text{S}_0}{\text{S}}) * Entropy(\text{S}_0) + \frac{\text{S}_1}{\text{S}} * Entropy(\text{S}_1) \quad (2)$$

After calculating the information gain value of each attribute, then selecting the largest gain value to be used as a node in the resulting tree. Then recalculate the entropy and information gain values of each attribute based on the previously selected node.

2.4. Testing

Testing is done to measure the performance of a system. In this study, it is used to measure the accuracy of the text document classification method. Tests carried out using a confusion matrix. The confusion matrix is a useful tool for analyzing how well the classifier recognizes tuples of different classes[14]. The confusion matrix is in the form of a table that states the level of truth of the classification process. The confusion matrix table can be seen in Fig. 2.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Fig. 2. Confusion Matrix Design

3. Results and Discussion

This section will explain the tests carried out on the results of the research that have been made. The testing phase aims to determine the accuracy of the method that has been made. The implementation of testing in this study was carried out by testing the confusion matrix. Through the confusion matrix, it can be seen the value of the evaluation metrics used in the tests in this study, namely accuracy, precision, and recall. The test was carried out with 190 data, with a dataset ratio of 80% of test data and 20% of training data.

3.1. Confusion Matrix Test

The confusion matrix test is carried out to measure the performance of sentiment analysis with the Decision Tree C4.5 Algorithm. From the confusion matrix test, the accuracy, precision, and recall values will be obtained. The confusion matrix is done by obtaining True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) values, where these values are obtained from the prediction results of the Decision Tree C4.5 model and compared with labels from the data obtained. actually. The test results can be seen in Table 3.

Table 3. Confusion Matrix Test Results

		Prediction		
		Positive	**Negative**	**Neutral**
current	Positive	75	4	2
	Negative	2	100	2
	Neutral	0	4	1

The level of accuracy is a value that represents the level of closeness between the predicted value and the actual value. Based on Table 3, the accuracy value can be calculated using equation (3).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} x100\% \quad (3)$$

From (3), the accuracy calculation result is 92.63%. While the precision value describes the number of positive data categories that are classified correctly divided by the number of data that are classified as positive.The following is Table 4 of the results of testing the precision calculation of the confusion matrix.

Table 4. Precision Calculation Results

	Positive	**Negative**	**Neutral**
TP	75	100	1
FP	6	4	4
Precision	0.925926	0.961538	0.2
Average	0.695821		

Meanwhile, the recall value is the ratio of true positive predictions compared to the overall data that are true positive. The following is Table 5 the results of testing the recall confusion matrix calculation.

Table 5. Recall Calculation Results

	Positive	**Negative**	**Neutral**
TP	75	100	1
FN	2	8	4
Recall	0.974026	0.925926	0.2
Average	0.699984		

From Table 3, Table 4 and Table 5 it is obtained The accuracy of the Decision Tree C4.5 Algorithm in this study was 92.63%, with a precision of 69.58% and a recall of 69.99%. The values of accuracy, precision, and recall have quite a large difference in values becausethe number of datasets with neutral sentiment is very small so that the classification results for neutral sentiment are still less accurate.

4. Conclusion

In this study, sentiment analysis of e-commerce applications has been carried out with the Decision Tree C4.5 Algorithm. The results of this study indicate that the Decision Tree C4.5 Algorithm can perform the sentiment analysis process. In testing sentiment analysis using the Decision Tree C4.5 Algorithm which uses a dataset of 940 data, with 750 training data and 190 test data. In testing the dataset using the confusion matrix, the accuracy results are 92.63%, the average precision is 69.58% and the recall average is 69.99%. The results of precision and recall tests show that they are lower than the accuracy values because the number of datasets with neutral sentiments is very small, so the classification results for neutral sentiments are still less accurate.

The suggestions that can be used for the development of further research are that it can be developed by adding datasets, especially with neutral sentiments so that the classification results

can be more accurate and dictionary words can be reproduced in the database, so that the decision tree formed will be more accurate in processing test data.

References

[1] Y. Jahja, "Pengaruh service performance value, emotional value, monetary value, social value terhadap customer loyalty melalui perceived value dan customer satisfaction pada konsumen zalora di Surabaya," pp. 1–13, 2018.

[2] Ilmawan, "Aplikasi Mobile untuk Analisis Sentimen pada Google Play," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 9, no. 1, pp. 53–64, 2015.

[3] A. Kumar and Y. K. Kim, "The store-as-a-brand strategy: The effect of store environment on customer responses," J. Retail. Consum. Serv., vol. 21, no. 5, pp. 685–695, 2014.

[4] Murphy, Rosie. 2020. "Local Consumer Review Survey: How Customer Reviews Affect Behavior", https://www.brightlocal.com/Research/Local-Consumer-Review-Survey/ <accessed 25 December 2020 >.

[5] F. Gunawan, M. A. Fauzi, and P. P. Adikara, "Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan Naive Bayes dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi BCA Mobile)," Syst. Inf. Syst. Informatics J., vol. 3, no. 2, pp. 1–6, 2017.

[6] H. Zulfa, "ANALISIS PENGARUH PERSEPSI RISIKO, KUALITAS SITUS WEB, DAN KEPERCAYAAN KONSUMEN TERHADAP KEPUTUSAN PEMBELIAN KONSUMEN E-COMMERCE SHOPEE DI KOTA SEMARANG," JSHP J. Sos. Hum. dan Pendidik., vol. 4, no. 2, pp. 1–11, 2020.

[7] A. Rakhman and M. R. Tsani, "Analisis Sentimen Review Media Massa," Smart Comp, vol. 8, no. 2, pp. 78–82, 2019.

[8] Z. A. Gumilang, "IMPLEMENTASI NAÏVE BAYES CLASSIFIER DAN ASOSIASI UNTUK ANALISIS SENTIMEN DATA ULASAN APLIKASI E-COMMERCE SHOPEE PADA SITUS GOOGLE PLAY TUGAS," 經濟研究, 2018.

[9] M. Lailiyah, "Sentiment Analysis Menggunakan Rule Based Method Pada Data Pengaduan Publik Berbasis Lexical Resources," 2017.

[10] And M. A. F. P. Antinasari, R. S. Perdana, "'Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku,'" מים והשקייה, vol. 549, no. 12, pp. 40–42, 2017.

[11] Y. Kustiyahningsih and E. Rahmanita, "Aplikasi Sistem Pendukung Keputusan Menggunakan Algoritma C4.5. untuk Penjurusan SMA," J. Semantec, vol. 5, no. 2, pp. 101–108, 2016.

[12] Mulholland, "Application of the C4 . 5 classifier to building an expert system for ionchromatography, 27, 95–104," עלון הנוטע, vol. 66, no. September, pp. 37–39, 1995.

[13] And P. V. C. P. N. Patil, P. R. Lathi, "'Customer Card Classification Based on C5 . 0 & CART Algorithms,'" עלון הנוטע, vol. 66, no. 3, pp. 37–39, 2012.

[14] J. Han, J., Kamber, M., & Pei, "Data Mining Concepts and Techniques Third Edition. Waltham: Elsevier Inc.," p. 2011, 2011.