The accuracy comparison of vector support machine and decision tree methods in sentiment analysis

Nurfaizah¹, T Hariguna², Y I Romadon³
¹,²,³ Information System Amikom Purwokerto University
E-mail: nurfaizah@amikompurwokerto.ac.id

Abstract. Analysis of customer satisfaction with company service is one of the tools for the company to be able to find out the lack of services and be able to know the expectations of users of the company. This is very possible in real-time, especially companies that are engaged in buying and selling online because one of the features they provide is a comment column from the user. In addition to using websites or known as e-commerce sites that can be accessed by consumers through media websites, currently online buying and selling providers can also be accessed through Google Play. Consumers download e-commerce applications and install them on their smartphones. This study aims to process user comments using Support Vector Machine (SVM) and Decision Tree algorithms to see the accuracy of the algorithm used and see positive and negative reviews. In addition to producing sentiment analysis, SVM classification algorithms and Decision Tree methods will be compared to their accuracy values in calculating datasets.

1. Introduction
The development of Indonesian e-commerce is one of the effects of the development of internet technology today. The survey results from the Indonesian Internet Service Providers Association in 2017 showed that e-commerce sites were included in the top 15 services accessed by internet users in Indonesia [1]. One of the biggest e-commerce companies currently in Indonesia is Shopee. Based on a survey from the Google Play site per March 2019, the Shopee application has been downloaded 1,176,077 times. E-commerce companies continue to innovate in improving their services to be able to maintain their consumers. One measure of service success can be seen from the comments of consumers. Google Play provides features that can be used by consumers to review the products they buy. The results of reviews from consumers which we call sentiment analysis can provide concise results so that it makes easier for new users to make decisions [2] [3]. Another function of using sentiment analysis is that it can be used by companies to improve product and customer satisfaction [4]. Sentiment classification aims to optimize the task of classifying positive and negative comments [5].

The review process in this study was carried out by using the classification method namely supervised learning. The process was also carried out by using datasets that were used as learning references and training datasets whose values were processed based on the reference dataset [6] [7]. Text classification Support Vector Machine (SVM) is a classification method that has a high level of accuracy in carrying out the text classification process [8] [9] [10] [11].
In addition to using the SVM classification method, another method used was the Decision Tree. Decision Tree is a method used to map various alternatives. It estimates alternative choices that exist [12]. Decision trees can provide recommendations on the best, worst and other values for various scenarios [13]. Based on these data, this study will predict consumer comment reviews from Shopee on google play applications into negative or positive sentiments and produce predictive accuracy values using the SVM algorithm and Decision Tree algorithm.

2. Related Work
Research conducted by Zhouyang Li et al. [4] resulted in analytical sentiments that looked at user reviews into satisfied or dissatisfied opinions on services by developing algorithms. A similar study by Katerina Berezena et al. [14] which produced an analytical sentiment to measure customer satisfaction with hotel services using the text analysis module PAWS. Research on sentiment analysis by comparing algorithms was carried out by Rodrigo Moraes [5]. His study was about algorithms that were compared in making sentiment analysis using the SVM algorithm and Artificial Neural Network (ANN) algorithms. Similar research has also been carried out using SVM classification algorithms to process consumer comments, twitter comments and document text classification result in very high predictive accuracy [9] [11] [10] [7] [8]. In contrast to previous research, this study focused on seeing the results of SVM algorithm prediction accuracy and Decision Tree algorithm in sentiment analysis of consumers. Comment data was taken from Google Play so that they can see positive and negative comments and compare the two algorithms to see the performance of each algorithm in processing training data and testing data.

3. Proposed Method
3.1. Sentiment Analysis
Sentiment analysis is the process of understanding, extracting and processing textual data automatically to get information. In general, sentiment analysis is needed to find out the attitude of a speaker or writer to the topic or contextual polarity of the document [15]. There are 2 approaches used to analyze; the first, sentiment Machine Learning which requires datasets to be used as training data so it takes effort to collect and classify samples in dataset samples and the training process takes time. Second, Knowledge-Based is an approach to sentiment analysis on the word level, where the entity that is processed is the word. The methods entered in this stage are Lexicon-Based and Point Mutual Information [16].

3.2. Support Vector Machine (SVM)
SVM was introduced by Vapnik in 1992 as an efficient classification technique for nonlinear problems. SVM has a concept that seeks to find a hyperplane by maximizing the distance between classes with the concept SVM which has a far more mature concept and is more mathematically clear [17]. The steps in calculating Support Vector Machine include:
- Determine vector, definition of each point \( X = (X_1, X_2), X \neq 0 \)
- The direction definition of vector \( u \) \((u_1, u_2)\) is vector \( W^T \left( \frac{u_1}{||u||}, \frac{u_2}{||u||} \right) \)
- Calculate dot products by formula \( x, y = x_1y_1 + x_2y_2 \sum_{t=1}^{2} (x_t y_t) \)
- Look for a hyperplane with a formula \( W^T x + b = 0 \)
3.3. Decision Tree
Decision Tree is one of the most popular classification methods because it is easy to interpret by humans. Decision Tree is used because it can accurately predict a label and is easy to use [18]. In addition, Decision Tree is a prediction model using a tree structure or hierarchical structure. The concept of a decision tree is to transform data into decision trees and decision rules. The main benefit of using a decision tree is its ability to reduce complex decision-making processes to be simpler so that decision-makers will better interpret the solution to the problem [19].

3.4. Methodology
The stages in this study are datasets through the pre-processing stage. The stages of text pre-processing are through the transformation phase, namely, read data and case folding. Data cleaning is also done, namely tokenization, spelling, filtering, and N-gram. The next stage is the process of sharing data into training data and testing data. Then the classification and comparison stages of the SVM method and Decision Tree. As in Figure 1 below.

Figure 1. Research Stages

4. Result and Discussion

4.1. Classification
Classification results used 500 training data and 500 data testing in order to give an accurate predictions. The number of fold values used in this study were 2 to 10 times. This was used to see the best performance of each algorithm.

4.1.1. SVM Classification
The accuracy of SVM gave results in word weights for each word. The result of weighting using SVM is shown in figure 2 below:
Figure 2. Results of word weighting using SVM

The word weighting result in Figure 2 shows that if the weight value showed a minus value (-) then it was categorized as a negative comment while the weighting produced a positive value indicating that the word included positive comments. Furthermore, the words were processed to predict the sentiment analysis on the results of testing data as shown in Figure 3 below.

In figure 3 shows that from 500 testing data used SVM produced 112 negative reviews and 388 positive reviews.

4.1.2. Classification of Decision Tree
In the use of the decision tree, it is necessary to evaluate all existing attributes to determine which attributes must first be checked and placed in the root. This can be done by using a statistical measure of the gain ratio in measuring the effectiveness of an attribute in classifying a collection of data samples. The minimum leaf size parameter from the tree formed was 2, while the minimum size of the split parameter determines that the broken node was a node that has a size of ≥ 4. Furthermore, the other parameters were standard values. The classification process carried out using the decision tree method produced a scheme of words that had the greatest influence in predicting sentiments in all review data. This can be seen in Figure 4.
The results of the best accuracy using the decision tree method produced a weighting scheme for each word that exists. From the word weighting, it was used in the classification process to predict the sentiment of the data to be tested. Furthermore, the scheme of these words was used to predict sentiment on testing data, the results obtained as shown in Figure 5.

In Figure show that from 500 reviews predicted using the decision tree algorithm produced 80 negative reviews and 420 positive reviews.

4.2. Comparison

Based on the tests carried out using the two classification algorithms and using cross-validation operators, they showed that the tests yielded the best accuracy of the two algorithms using fold = 2-10. A comparison of test results from each fold value can be seen in the table below.
Table 1. Comparison of Accuracy Values from SVM and Decision Tree

| Fold | Akurasi SVM | Akurasi Decision Tree |
|------|-------------|----------------------|
| 2    | 78,40%      | 69,40%               |
| 3    | 80,61%      | 73,21%               |
| 4    | 81,00%      | 70,00%               |
| 5    | 81,40%      | 72,40%               |
| 6    | 80,38%      | 72,78%               |
| 7    | 81,79%      | 72,60%               |
| 8    | 80,80%      | 70,39%               |
| 9    | 82,19%      | 72,01%               |
| 10   | 81,60%      | 73,00%               |

Based on Table 1 above from the experiments that have been carried out using the SVM method and the Decision Tree method, for the SVM method the value of fold = 9 will be used in the test model because fold = 9 shows the average of the accuracy results better than other fold values. As for the decision tree method, the value of fold = 3 will be used in the testing model, because fold = 3 shows the average of the results of accuracy that is better than other fold values. This study used confusion matrix in the process of calculating accuracy by knowing the amount of test data that were correctly classified and the number of test data that was wrong in its classification. Comparison of confusion matrix with testing using fold = 9 for SVM and fold = 3 for decision trees can be seen in Table 2 below:

Table 2. Confusion Matrix

| Prediction | SVM | Decision Tree |
|------------|-----|---------------|
|            | Positive | Negative | Positive | Negative |
| Positive   | 229   | 53         | 228      | 97       |
| Negative   | 36    | 182        | 37       | 138      |
| **Accuracy**| **82,19%** | **73,21%** |

Table 2 illustrates how large the percentage of data predicted by the SVM method and the Decision Tree Method. True negative was the number of observation data with negative categories that can be predicted negatively (correctly predicted) by the method. True positive was the number of observation data with a positive category that can be positively predicted (correctly predicted) by the method. False-negative was the amount of observation data that was categorized as negative but there was a prediction error. While false positive was the number of observation data that was categorized as positive but there were prediction errors. In the SVM method, the prediction results showed that in the positive class there were 265 positive reviews tested, there were 229 reviews that had been correctly classified and there were prediction errors of 36 reviews that had entered negative reviews. Then in the negative review with a total of 235 tested, there were 182 reviews that had been classified correctly as negative reviews and there were prediction errors of 53 reviews that had entered positive reviews. Furthermore, from the confusion matrix value, it was obtained an accuracy of 82.19%. Whereas in the decision tree method, the prediction results showed that in the positive class there were 265 positive reviews tested, there were 228 reviews that had been correctly classified and there were prediction errors of 37 reviews which had entered negative reviews. Then in the negative review with a total of 235 tested, there were 138 reviews that had been correctly classified as negative reviews and there were prediction errors of 97 reviews that were included in the positive reviews. Furthermore, from the confusion matrix value, it was obtained an accuracy of 73.21%.
5. Conclusion
The calculation results using the classification algorithm method using 500 training data and 500 testing data from shopee consumer comment data taken from Google Play indicated that using the SVM method means that from 500 review data tested, there were 411 correct reviews, while the Decision Tree method produced 366 correct reviews in the classification by the decision tree model. Based on the results of the accuracy in this study if the decision tree method was compared with the Support Vector Machine (SVM) method, the Support Vector Machine (SVM) method had a better and higher level of accuracy in predicting analysis sentiment.

References
[1] APJII 2018 Buletin APJII Edisi 22 - Maret 2018
[2] M Thawakar 2015 An Intelligent Opinion Mining for Customer Reviews IJCSA vol. 8, no. 1, pp. 20–25
[3] W Songpan 2017 The Analysis and Prediction of Customer Review Rating Using Opinion Mining IEEE 15th Int. Conf. Softw. Eng. Res. Manag. Appl pp. 71–77
[4] Z Li and L Liu 2015 Analysis of Customer Satisfaction from Chinese Reviews using Opinion Mining IEEE pp. 95-99
[5] R Moraes, J F Valiati, and W P G Neto 2013 Expert Systems with Applications Document-level sentiment classification: An empirical comparison between SVM and ANN Expert Syst. Appl vol.40 no.2 pp. 621–633
[6] S B Kotsiantis 2007 Supervised Machine Learning: A Review of Classification Techniques Informatica vol. 31 pp. 249–268
[7] A Sarkar, S Chatterjee, W Das and D Datta 2015 Text Classification using Support Vector Machine IJESI vol. 4 no. 11 pp. 33–37
[8] A R Naradhipa and A Purwarianti 2011 Sentiment Classification for Indonesian Message in Social Media ICEEI pp. 2–5
[9] J B Chrysal and S Joseph 2015 Text M Ining And C Lassification Of P Roduct R Eviews U Sing S Tructured CS&IT pp. 21–31
[10] S Fatima, B Srinivasu, and S F M. 2017 Text Document categorization using support vector machine IRJET pp. 141–147
[11] M Ahmad and I Ali 2017 Sentiment Analysis of Tweets using SVM Sentiment Analysis of Tweets using SVM LJCA Vol. 177 No.5 pp. 25-29
[12] J George and S George 2014 Classification problem in text mining IJRAE vol.1, no. 8, pp. 333–341
[13] J Jotheeswaran and Kumaraswamy 2013 Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure J. Theor. Appl. Inf. Technol., vol. 58, pp. 72–80
[14] K Berezina, A Bilghian, C Cobanoglu, and F Okumus 2015 Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Journal of Hospitality and Management vol. 8623, pp. 1-24
[15] B Pang and L Lee 2014 Opinion mining and sentiment analysis vol. 2, no.1
[16] J Cao, K Zeng, H Wang, and J Cheng 2004 Web-Based Traffic Sentiment Analysis: Methods and Applications IEEE Transaction on Intelligent Transportation System vol. 15, no. 2, pp. 844–853
[17] B Santosa and A Umam 2018 Data Mining dan Big Data Analytics, Edisi 2. Yogyakarta: Penebar Media Pustaka
[18] T D Nguyen, T B Ho, and H Shimodaira 2001 A Scalable Algorithm for Rule Post-pruning pp. 467–476
[19] J Han, M Kamber, and J Pei 2012 Data Mining Concept and Technique Third USA: Elsevier