Jurnal Informasi dan Teknologi

https://jidt.org/jidt

2024 Vol. 6 No. 2 Hal: 217-220 e-ISSN: 2714-9730

A Performance Comparison of Algorithms on the Indonesian Tweet Comment Labeled with ITE Law

Faisal Fahmi[1]✉, Ardila Nolla Romantike[2]

[1,2] Department of Information and Library Science, Airlangga University

faisalfahmi@fisip.unair.ac.id

Abstract

The presence of Twitter as an online forum causes everyone to be free to comment. This is one of the reasons the government issued the law on Electronic Information and transactions (ITE) to oversee all activities in cyberspace. However, the growing and growing amount of comment data is also increasingly difficult to analyze. Therefore, the application of data mining using the Rapid Miner application is proposed in this study to help in finding the most effective and efficient method, by looking at the accuracy, precision, and time lapse required when processing the comment data. In this study, a total of more than 12,000 data, containing comments and seven types of labels based on ITE were collected for analysis. After pre-processing the data, the researchers chose five of several classification methods, namely Naive Bayes, k-NN, Decision Tree, SVM, and Perceptron methods to be tested. From the tests that have been carried out through the Rapid Miner application, it was found that SVM became the best method used to classify comment data, with an accuracy of 55.32%. Meanwhile, the method with the lowest accuracy is occupied by Perceptron with a total accuracy of only 18.04%. Based on observations, the best accuracy results only reached 55.32% due to the large number of labels considered in the prediction.

Keywords: Data Mining, Rapidminer, Twitter, SVM.

1. Introduction

Freedom of access to information in the midst of advances in digital technology makes social media users can easily share their opinions with the public. [8] In a study it is said that from the results of interviews that have been conducted with several respondents, it is concluded that the increase in social media users in Indonesia causes everyone to be able to freely write their opinions, one of which is through comments. It is also mentioned that negative comments often cause a conflict that affects the emergence of hoaxes or disruption of the level of social stability. Seeing this condition, the government itself has struggled through the establishment of a law that contains laws related to all cyberspace activities, namely through the law on information and Electronic Transactions (UU ITE) [3]. In this case, comments on social media Twitter becomes more highlighted and often analyzed by researchers because of the many communities that can share comments in real time and widely [2].

In addition, the presence of Twitter as an online forum used to criticize a particular policy [3] causes the volume of data to increase. So, to classify the many types of comments, a technology in the form of data mining is needed. Data mining basically refers to the decision-making process obtained from unstructured data, then analyzed, both data in the form of text and images in order to produce the information needed [6], [1], [10]. In addition, data mining is also used in various fields to analyze, classify, and predict large and growing data through certain ways or techniques [10]. This process is then used to interpret the existing data into something that can be read. In this regard, Rapid Miner is one of the data mining tools proposed in this study.

Rapid Miner is considered the right tool because of its fast processing stage, and refers to all stages including pre-processing, modeling, methods used, model tests, to the visualization stage [10]. In addition, a literature [7] also mentions, that Rapid Miner becomes one of the right data mining tools for text mining by providing answers to whether the statement is negative or positive through labels. However, from the many modeling on Rapid Miner, it is necessary to choose an effective and efficient method to analyze a large data. This is motivated by the level of accuracy and precision of the method if applied to different data. For example, research [9] which uses the association rule method to look for patterns of relationships in each company's data, [7] chose the naive bayes method to analyze the sentiment of Covid-19 vaccines in Indonesia, [1] proposed the NLP method to classify student perceptions of e-learning at universities in Malaysia, the K-Means method to cluster Covid-19 cases in Lampung Province [5], and so on.

Receipt: 30-11-2023 | Revision: 29-06-2024 | Publish: 10-07-2024 | doi: 10.60083/jidt.v6i2.564

Thus, by referring to the explanation above, this study also uses Rapid Miner to compare five different classification methods to see how these methods work in classifying comments on Twitter social media that are labeled ITE Law in Indonesia. The five classification methods are intended to see the advantages and disadvantages, accuracy, precision, and time lapse required when processing data. Thus, the results of this study can provide an overview, suggestions, or alternative solutions to choose the right method in the process of data classification using Rapid Miner.

2. Research Methods

This research method is carried out through a Rapid Miner data mining tool with several stages that can be seen in Figure 1 [11] as follows.

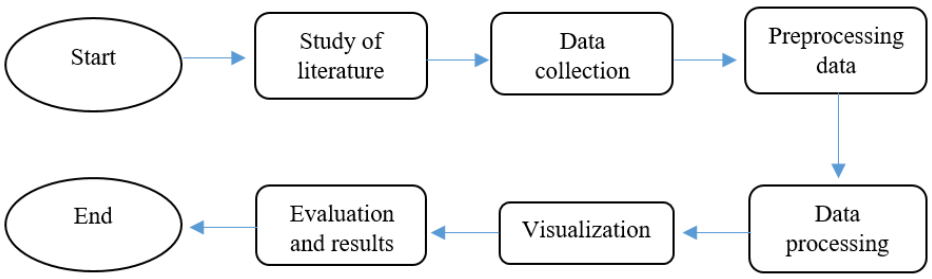

Figure 1. Research Methodology

First, the study of literature is conducted to find reference sources with the same topic. In this process, the researcher found at least 10 articles that discuss various methods in the Rapid Miner application where these are used to analyze the classification of data.

Furthermore, the process of searching data through a site kaggle.com and found a number of data sets that researchers managed to collect. Researchers chose data containing comments labeled ITE Law on social media Twitter because of several criteria, namely the suitability of the data with the topic you want to take, as well as covering more than 12,000 comments with 7 types of sentiment.

Pre-processing data sets, i.e. by using some tools such as tokenize to filter tokens, transform cases to change the characters in the data, filter stopwords to remove a certain list of tokens. In this case, the researchers used a type of filter stopwords (dictionary) because the data contains a collection of comments in Indonesian. Then, the researcher enters the Associated document a list of required words. Finally, filter tokens (by length) to filter tokens based on text length.

Then, data processing in which the researcher in this case uses several operators, such as split data to divide between training data and testing data and cross validation by setting the number of folds twice to test the data with a method. In this case, the researchers used five classification methods, namely Naive Bayes, k-NN, Decision Tree, SVM, and perceptron.

This study also presents data visualization that measures the accuracy, precision, and time lapse of the process through the five classification methods through the operator apply model and performance.

Finally, the results are evaluated by involving a Classification Report which serves to measure the quality of the prediction and display accuracy, precision, and recall.

3. Results and Discussion

The data mining process carried out through the Rapid Miner application in this study aims to explore the comments contained on Twitter social media, related to whether these comments can be labeled with the ITE Law or not. More than 12,000 Indonesian-language comments were collected for pre-processing before being visualized. In this case, data is obtained through kaggle.com and contains a collection of comments along with labels that can be seen in Table 1 as follows.

Table 1. Sentiment Labeled Categorized

Sentiment Labeled	Comment	Categorized
0	'barusan liat tulisan di belakang truk rela injek kopling demi kamu bisa shopping'	Neutral sentiment
1	'saya suka video aku tanya pak jokowi pakai bahasa jawa'	Positive sentiment
2	'Rakyat menolak KERAS!! Jika Indonesia dipimpin oleh komplotan MAFIA'	Negative sentiment

3	'Pejabat negeri yang kerjanya kayak babi (tidur makan)'	Insulting the government or public agency
4	'Warga kita kebanyakan kena penyakit DUNGU AKUT'	Insulting or defaming others
5	'Awas kalau kamu lapor ya saya bunuh kamu'	Threatening others
6	'Orang timur kurang pintar dari orang barat'	Alluding to the tribe, religion, race and intergroup

After the preprocessing stage in the Rapid Miner version 9.10.6 application, this study tried to compare the five classification methods tested against data related to comments labeled by the ITE Law on Twitter social media. From the analysis, obtained a comparison which can then be seen through Table 2 as follows.

Table 2. Results of the Application of Data Mining Comparison

Method	Accuracy	Precision	Recall	Time
Naive Bayes	42,65%	52.38%	0.57%	1:38
k-NN	43,90%	53.33%	1.88%	1:22:34
Decision Tree	44.60%	62.96%	2.03%	3:43
SVM	55.32%	52.38%	27.60%	7:46
Perceptron	18,04%	50.00%	1.19%	20:58

From the test results of the five methods above, namely Naive Baye, k-NN, Decision Tree, SVM, and Perceptron, it was found that the SVM value to be the highest with the lowest Perceptron. First, the Naive Bayes method has the advantage that it refers to the time interval that tends to be the fastest in processing data when compared to the other four methods. However, the value of accuracy, precision, and recall Naive Bayes is still lower than the other four methods.

Second, the k-NN method has a higher accuracy value than the Naive Bayes method, which is equal to 42.65% and 53.33% precision. Although the results are considered quite high when compared with the other two methods, k-NN is less suitable if it is said to be the most effective method used. This is because the time interval required to process data is quite long, which is up to more than 60 Minutes.

Third, the Decision Tree method becomes the next tested technique. In this method, the results obtained a fairly good accuracy, which is 44.60% with the greatest precision among the four other methods of 62.96%. In addition, the time required to process data is also quite short, only about 4 minutes. However, although efficient, the recall value of the Decision Tree method is still quite low, so it can not be said to be the most effective method.

Furthermore, the Support Vector Machine (SVM) method was tested against the data in this study. Although the precision value of 52.38% and not as big as the Decision Tree method, produced 55.32% accuracy value and recall value of 27.60%, which is the highest value of the other four methods were also tested. In addition, the time required for this method to process data also tends to be long. Thus, of the five methods that have been tested and analyzed, it can be said that the Support Vector Machine (SVM) method is the most effective and efficient method to be applied to the classification of text-shaped data, especially data that analyzes comments.

Finally, the Perceptron method became the fifth method that this study tested using the Rapid Miner application. Where, obtained accuracy value is only 18.04%, precision value is 50.00%, and recall is only 1.19%. In addition to the acquisition of the value of the three indicators, the Perceptron method also requires a long time in processing the data with almost half an hour. So, by looking at these results, Support Vector Machine (SVM) becomes the best method to use, as well as a suggestion or alternative solution to perform classification tests on data in text form.

4. Conclusion

From the testing of five methods in the Rapid Miner application, it was found that the Support Vector Machine (SVM) method became the most effective and efficient method if used to classify text-shaped data. This is evidenced in the results of SVM accuracy tends to be higher when compared with the other four methods, namely Naive Bayes. k-NN. Decision Tree, and Perceptron. In addition, the SVM method also does not take too long to process the data. However, the result of SVM accuracy which touched 55.32% was considered not a high value. Based on observations, These results can not be separated from the reason that the labels used in this study data predicted quite a lot. So, the value cannot be a reference to be said to be the best.

References

[1] Baragash, R. S., Aldowah, H., and Umar, I. N., "Students' Perceptions of E-Learning in Malaysian Universities:

Sentiment Analysis Based Machine Learning Approach," *J. Inf. Technol. Educ. Res.*, *21*, 439-463, 2022.
[2] Efendi, A., and Shasrini, T., "Communication Ethics of Criticism in the Public Space of Twitter Social Media," *Experimental Student Experiences*, *1*(5), 440-445, 2023.
[3] Hakim, L., Kusumasari, T. F., and Lubis, M.. "Text mining of UU-ITE implementation in Indonesia," *Journal of physics: conference series* (Vol. 1007, No. 1, p. 012038). IOP Publishing, 2018.
[4] Hartono, P. C., and Widiantoro, A. D, "Analisis Prediksi Harga Saham Unilever Menggunakan Regresi Linier dengan RapidMiner," *Journal of Computer and Information Systems Ampera*, *5*(3), 174-190. 10.51519/journalcisa.v5i3.481, 2024.
[5] Nabila, Z., Isnain, A. R., Permata, P., and Abidin, Z., "Analisis data mining untuk clustering kasus covid-19 di Provinsi Lampung dengan algoritma k-means," *Jurnal Teknologi Dan Sistem Informasi*, *2*(2), 100-108, 2021.
[6] Nurhachita, N., and Negara, E. S., "A comparison between deep learning, naïve bayes and random forest for the application of data mining on the admission of new students," *IAES International Journal of Artificial Intelligence*, *10*(2), 324, 2021.
[7] Pristiyono, Ritonga, M., Ihsan, M. A. A., Anjar, A., and Rambe, F. H., "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," In *IOP Conference Series: Materials Science and Engineering* (Vol. 1088, No. 1, p. 012045). IOP Publishing, 2021.
[8] Rafiq, A., "Dampak media sosial terhadap perubahan sosial suatu masyarakat," *Global Komunika: Jurnal Ilmu Sosial Dan Ilmu Politik*, *3*(1), 18-29, 2020.
[9] Santoso, M. H., "Application of association rule method using apriori algorithm to find sales patterns case study of indomaret tanjung anom," *Brilliance: Research of Artificial Intelligence*, *1*(2), 54-66., 2021.
[10] Taranto-Vera, G., Galindo-Villardón, P., Merchán-Sánchez-Jara, J., Salazar-Pozo, J., Moreno-Salazar, A., and Salazar-Villalva, V., "Algorithms and software for data mining and machine learning: a critical comparative view from a systematic review of the literature," *The Journal of Supercomputing*, *77*, 11481-11513, 2021.
[11] Utomo, W., "The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia," *ILKOM Jurnal Ilmiah*, *13*(1), 31-35, 2021.