Online Version Available at: https://jurnal.ubd.ac.id/

JOURNAL OF TECH-E

|2581-1916 (Online) | 2598-7585 (Printed) |

Article

Comparison of Seven Machine Learning Algorithms in the Classification of Public Opinion

Sri Redjeki[1], *Setyawan Widyarto*[2]

[1]*Informatics, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia*
[2]*FCVAC, University Selangor, Selangor, Malaysia*

SUBMISSION TRACK

Received: Mar 16, 2022
Final Revision: Mar 22, 2022
Available Online: Mar 25, 2022

KEYWORD

Classification, Machine Learning, Opinion, Sentiment Analysis, Twitter.

CORRESPONDENCE

E-mail: dzeky@utdi.ac.id

A B S T R A C T

Sentiment analysis is one way that is widely used to identify the beginning of public opinion in various fields of life which are associated with very massive and a lot of information through social media. This study aims to compare several algorithms in machine learning to see the best ability in sentiment classification. The research dataset uses a dataset of public opinion related to tourism in Indonesia. The number of datasets used is 10,228 twitter data that have been cleaned and labelled. The machine learning algorithm used is Logistic Regression, KNN, AdaBoost, Decision Tree, SVM, Random Forest and Gaussian. The seven algorithms for sentiment classification from the Twitter public opinion each produce a Gaussian accuracy of 0.52; SVM 0.78; KNN 0.98; Logistic Regression, Random Forest, Decision Tree, AdaBoost of 0.99. This study shows that the selection of the right machine learning algorithm will have a very good impact on the classification of public opinion through social media.

INTRODUCTION

Natural language processing, information retrieval, and text mining are all considered subfields of opinion mining. Opinion mining is the process of collecting human opinions and perspectives from unstructured texts, which has become a relevant, appealing, and demanding problem in light of the rise of online social media and the large volume of user comments [1]. The variety of public opinions that exists on social media is widely used to identify one of which is the sentiment towards various areas of community problems [2][3]. The increasing number of social media users creates a huge data abundance that can be used as an opportunity to conduct sentiment analysis. With rich crowd-sourced datasets that are easy to collect and evaluate, Twitter has been a boon to academic study. Its data has been used to predict crime, the stock market, political elections, public opinion polls, public health, and movie sales, among other things [4][5].Sentiment analysis has become a field of research studies in various fields of big data which is very helpful because sentiment analysis can be one of the initial

indicators in conducting analyzes involving the public [6].

The classification of public opinion sourced from social media has attracted the attention of researchers in the field of sentiment analysis [2][7][1][8]. The number of studies in the field of sentiment analysis provides researchers with opportunities for the development and optimization of algorithms used for sentiment analysis to provide the best results [2][9][6]. One of the most widely used algorithms in sentiment analysis is machine learning. Machine learning techniques have been widely adopted in several huge and sophisticated data-intensive sectors, such as medicine, astronomy, biology, and so on because these techniques offer potential solutions for extracting knowledge from data [10][11][12][13][14]. Several algorithms in machine learning have their respective advantages, this is a research opportunity to see the advantages of several algorithms in machine learning for classifying public opinion through social media [15][16][17]. The purpose of this study is to compare several algorithms in machine learning to determine the classification ability of public opinion data on Twitter. The machine learning algorithm used is Logistic Regression, KNN, AdaBoost, Decision Tree, SVM, Random Forest and Gaussian. The object of the data used is public opinion on Yogyakarta tourism. The writing is divided into several parts: introduction, research methodology, discussion, and conclusion.

I. METHODS

This section will explain the stages of research carried out to answer the research objectives, namely comparing 7 algorithms in machine learning to classify Twitter opinions about tourism in Yogyakarta through sentiment analysis. The Twitter dataset used is 10228 tweets that have been cleaned and labelled. The research stages for opinion classification using sentiment analysis are shown in Figure 1.

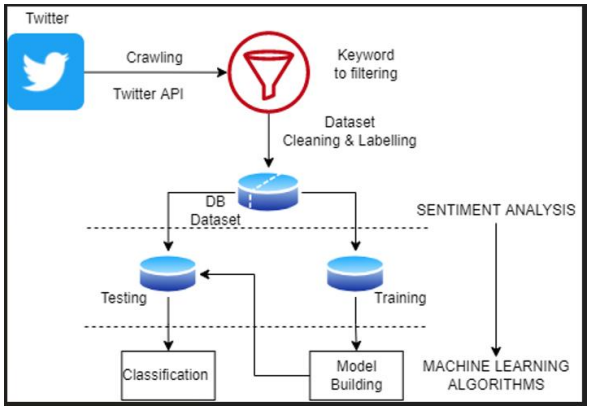

Fig 1: Classification System

The classification system shown in Figure 1 shows the overall stages of the research being carried out. From the Twitter data obtained using the Twitter API, crawling with keywords related to tourism in Yogyakarta was carried out. The data obtained were cleaned and given a positive, negative and neutral label. After the data is clean and labelled, the data is divided into 2 for training and testing. The distribution of datasets for training and testing is 80% and 20%.The training data is used to form a model using 7 algorithms in machine learning, namely Logistic Regression, KNN, AdaBoost, Decision Tree, SVM, Random Forest, and Gaussian (NBC). The training model will be used for testing. The performance of each algorithm will be seen from the accuracy of the model testing results. The stages of the data cleaning process carried out in this study are shown in Figure 2. The details of the stages are,

a. Conversion to Lower Case - The first pre-processing step we will take is to turn our tweets into a smaller case.
b. Removing Punctuations because when handling text data, it does not add any additional details.
c. Stop Word Removal should be omitted from the text data by frequently occurring words.
d. Correction of spelling - We have all seen tweets with a multitude of spelling errors. Our timelines are sometimes

loaded with tweets sent in a hurry that are often barely readable.

e. Tokenisation refers to breaking the text into a series of words or phrases.
f. Lemmatization, instead of only eliminating enough, transforms the word into its root word. To get the root word, it makes use of the vocabulary and conducts a morphological analysis.

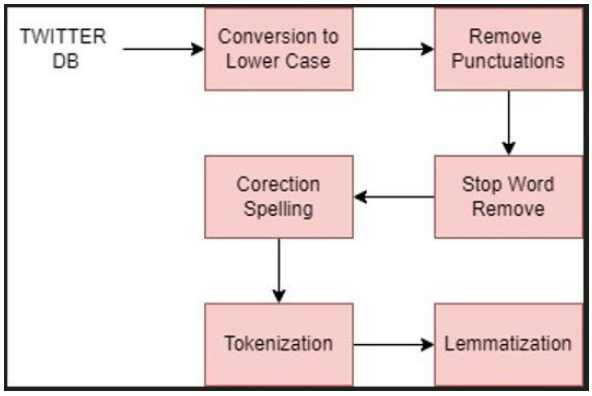

Fig 2: Cleaning Stages

All processes carried out in this research using the Python programming language along with all libraries related to sentiment analysis and machine learning classification algorithms.

II. RESULT

The results of this study answer the research objective, which is to compare the performance of seven machine learning algorithms in classifying opinions related to tourism in Yogyakarta. Algorithm performance is seen from the results of the confusion matrix for each existing algorithm. Figure 3 shows a snippet of the cleaned and labeled dataset displayed using python coding. The initial dataset obtained from Twitter has many attributes and which is used for sentiment classification based on opinions only tweets. While the label for each tweet is given using the sentiment analyzer in python.

Tweet	Cluster
kawasan malioboro nyari top	2
ada yg menganggap bahwa kuliah di kota yang ja...	2
jaman gw sd ngefansnya sama tatu dibelain beli...	2
pacaran udah pernah ngapain aja ngapain yaaaaa...	1
iya ni kehujanan di malioboro	2

Fig 3: Dataset Sample

Figure 4 shows the coding of twitter data crawling using the existing library in python programming.

Fig 4: Data Crawling

The Twitter data crawling process used in this study uses the python search tweets library.

The clean dataset will then be trained using the algorithms in machine learning. The training results of the seven machine learning algorithms Logistic Regression, KNN, AdaBoost, Decision Tree, SVM, Random Forest and Gaussian, as follows :

1. Logistic Regression Algorithm

This study uses multinomial logistic regression because it has an output of more than 2. Table 1 shows that the precision, recall, and f1-score values are almost the same, namely 97% -98%, which means that for the neutral sentiment (neutral = 0) the accuracy of the logistic regression model is very good. Because the value is above 90%. The negative sentiment (negative = 1) has a perfect precision value, namely 1, the recall value is 94% while the f1-score is 97%. A positive sentiment (positive = 2) has a precision value of 98%, a recall value of 1, and an f1-score value of 99%. The accuracy

value of the Logistic Regression model has a value of 99%, which means that the test results can determine the sentiment results well as many as 99 out of 100 test data.

Table 1.Logistic Regresion Performance

LogisticRegression

	precision	recall	f1-score	support
0	0.98	0.97	0.98	493
1	1.00	0.94	0.97	432
2	0.98	1.00	0.99	2144
accuracy			0.99	3069
macro avg	0.99	0.97	0.98	3069
weighted avg	0.99	0.99	0.99	3069

2. KNN Algorithm

Table 2 shows the performance for the neutral class (neutral = 0) with a precision value of 93%, a recall of 61%, and an F1-score of 74%. The positive sentiment class (positive = 2) had 90% precision, 98% recall and 94% f1-score, while negative sentiment (negative = 1) had 87% precision, 82% recall and 85% f1-score. The highest value of precision is shown for the classification of neutral sentiment, while the largest value of recall and f1-score is shown for the classification of positive sentiment. The accuracy produced by the KNeighbord Classifier model is 90%.

Table 2. KNN Classifier Performance

KNeighborsClassifier

	precision	recall	f1-score	support
0	0.93	0.61	0.74	493
1	0.87	0.82	0.85	432
2	0.90	0.98	0.94	2144
accuracy			0.90	3069
macro avg	0.90	0.80	0.84	3069
weighted avg	0.90	0.90	0.89	3069

3. AdaBoost Algorithm

Table 3 shows the highest precision value of 100% for the classification of negative and positive sentiments, the highest recall value of 100% for the classification of neutral and positive. The highest f1-score value for positive classification and the accuracy value of the AdaBoost classifier model is 99%.

Tabel 3. Ada Boost Classifier Performance

AdaBoostClassifier

	precision	recall	f1-score	support
0	0.96	1.00	0.98	493
1	1.00	0.96	0.98	432
2	1.00	1.00	1.00	2144
accuracy			0.99	3069
macro avg	0.99	0.99	0.99	3069
weighted avg	0.99	0.99	0.99	3069

4. Decision Tree Algorithm

Table 4 shows that the values for precision, recall, f1-score, and accuracy are almost the same, namely 100%, which means that the classification results are very good for all sentiment classes.

Table 4. Decision Tree Classifier Performance

DecisionTreeClassifier

	precision	recall	f1-score	support
0	0.99	1.00	1.00	493
1	1.00	0.99	1.00	432
2	1.00	1.00	1.00	2144
accuracy			1.00	3069
macro avg	1.00	1.00	1.00	3069
weighted avg	1.00	1.00	1.00	3069

5. SVM Algorithm

Table 5 shows the highest score of 100% precision for negative sentiment classification (negative = 1), the highest value for recall of 100%, and an f1-score of 86% for the classification of positive sentiment.

Table 5. SVM Classifier Performance

SVC

	precision	recall	f1-score	support
0	0.97	0.15	0.26	493
1	1.00	0.39	0.56	432
2	0.76	1.00	0.86	2144
accuracy			0.78	3069
macro avg	0.91	0.51	0.56	3069
weighted avg	0.83	0.78	0.72	3069

The accuracy produced by the support vector machine classifier model is 78%.

6. Random Forest Algorithm

Table 6 shows the highest precision value of 100% for negative and positive classifications, the highest recall value of 100% for the classification of neutral and positive sentiments. The f1-score value is 100% for positive classification and the accuracy of the random forest model is 99%.

Table 6. Random Forest Classifier Performance

RandomForestClassifier

	precision	recall	f1-score	support
0	0.97	1.00	0.98	493
1	1.00	0.96	0.98	432
2	1.00	1.00	1.00	2144
accuracy			0.99	3069
macro avg	0.99	0.99	0.99	3069
weighted avg	0.99	0.99	0.99	3069

7. Gaussian (NBC) Algorithm

Table 7 shows the highest precision value of 89% for positive sentiment classification, the highest recall value of 64% for negative sentiment classification. The highest f1-score value is 66% for positive classification and the accuracy value produced by the Naïve Bayes Classifier model is 52%. This accuracy value is the smallest value among the machine learning models used for sentiment classification.

Table 7. NBC Classifier Performance

GaussianNB

	precision	recall	f1-score	support
0	0.30	0.44	0.36	493
1	0.25	0.64	0.36	432
2	0.89	0.52	0.66	2144
accuracy			0.52	3069
macro avg	0.48	0.53	0.46	3069
weighted avg	0.70	0.52	0.57	3069

III. DISCUSSION

Figure 5 provide a comparison of the accuracy of the machine learning classification model employed in this study. The maximum accuracy above 99% is indicated by the Decision Tree method, Random Forest, AdaBoost, and logistic regression, according to the accuracy value acquired in the machine learning algorithm.

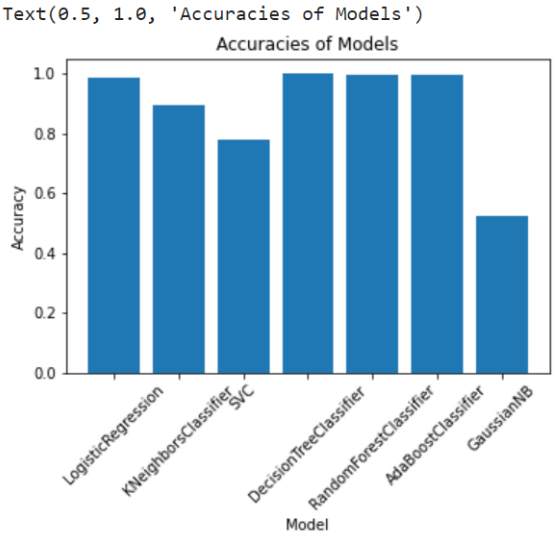

Fig 5: Histogram of Accuracy Model

The results of this study indicate that each algorithm used in the study has different performance according to the characteristics of the existing algorithm.

IV. CONCLUSION

Sentiment analysis is one of the big data research topics that is widely used to identify or identify public opinion. One of the methods used to classify opinions in the form of sentiment is machine learning. This study uses seven algorithms in machine learning and the best performance for the dataset on public opinion related to tourism in Yogyakarta is 99% for the algorithm: Decision Tree, Random Forest, AdaBoost, and logistic regression. Algorithm performance results are influenced by several factors, one of which is the amount of data used for training and testing.

REFERENCES

[1] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1495–1545, 2019, doi: 10.1007/s10462-017-9599-6.

[2] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *AI Soc.*, vol. 30, no. 1, pp. 89–116, 2014, doi: 10.1007/s00146-014-0549-4.

[3] A. Emam and M. Alzahrani, "Opinion Mining Techniques and Tools: A Case Study on an Arab Newspaper," *Proc. - 2017 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2017*, pp. 292–296, 2018, doi: 10.1109/CSCI.2017.49.

[4] R. P. Schumaker, A. T. Jarmoszko, and C. S. Labedz, "Predicting wins and spread in the Premier League using a sentiment analysis of twitter," *Decis. Support Syst.*, vol. 88, pp. 76–84, 2016, doi: 10.1016/j.dss.2016.05.010.

[5] R. Singh and A. Bhatia, "Sentiment analysis using machine learning techniques to predict outbreaks and epidemics," *Int. J. Adv. Sci. Res.*, no. May, pp. 19–24, 2018.

[6] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27. Elsevier Ireland Ltd, pp. 16–32, 2018, doi: 10.1016/j.cosrev.2017.10.002.

[7] S. Ahmed and A. Danti, "Effective sentimental analysis and opinion mining of web reviews using rule based classifiers," *Adv. Intell. Syst. Comput.*, vol. 410, pp. 171–179, 2016, doi: 10.1007/978-81-322-2734-2_18.

[8] H. woo An and N. Moon, "Design of recommendation system for tourist spot using sentiment analysis based on CNN-LSTM," *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 2019, doi: 10.1007/s12652-019-01521-w.

[9] H. Thakkar and D. Patel, "Approaches for Sentiment Analysis on Twitter: A State-of-Art study," 2015.

[10] I. Hemalatha, D. G. P. S. Varma, and D. a. Govardhan, "Sentiment Analysis Tool using Machine Learning Algorithms," *Int. J. Emerg. Trends Technol. Comput. Sci.*, 2013.

[11] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process.*, 2016, doi: 10.1186/s13634-016-0355-x.

[12] S. Rana and A. Singh, "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques," *Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016*, no. October, pp. 106–111, 2017, doi: 10.1109/NGCT.2016.7877399.

[13] F. Thabtah, N. Abdelhamid, and D. Peebles, "A machine learning autism classification based on logistic regression analysis," *Heal. Inf. Sci. Syst.*, vol. 7, no. 1, pp. 1–11, 2019, doi: 10.1007/s13755-019-0073-5.

[14] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment Analysis in Tourism: Capitalizing on Big Data," *J. Travel Res.*, vol. 58, no. 2, pp. 175–191, 2019, doi: 10.1177/0047287517747753.

[15] C. Troussas, M. Virvou, and S. Mesaretzidis, "Comparative analysis of algorithms for student characteristics classification using a methodological framework," *IISA 2015 - 6th Int. Conf. Information, Intell. Syst. Appl.*, 2016, doi: 10.1109/IISA.2015.7388038.

[16] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," *2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017*, 2017, doi: 10.1109/CITSM.2017.8089231.

[17] M. Somvanshi, P. Chavan, S. Tambade, and S. V. Shinde, "A review of machine learning techniques using decision tree and support vector machine," *Proc. - 2nd Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2016*, 2017, doi: 10.1109/ICCUBEA.2016.7860040.

BIOGRAPHY

Sri Redjeki. One of the Ph.D. students at Universiti Selangor and also a lecturer at the Indonesian Digital Technology University. The research fields are data mining, big data analytics, applied statistics, and artificial neural networks. Journals are written mostly about sentiment analysis to predict or identify public opinion, especially in the social field. Because of this research, the author received an invitation from Berkeley University in California in 2019. Researchers received a research grant from DIKTI for research related to poverty prediction using a big data approach through sentiment analysis.

Setyawan Widyarto. Research fields that are mostly done are simulation and computer science. Currently, he is a lecturer at Universiti Selangor and teaches at several universities in Indonesia in the fields of simulation and computer science.