Naive Bayes Performance in Analysis of Public Opinion Sentiment Against COVID-19

Ayu Hendrati Rahayu*[1], **Ari Sudrajat**[2]
[1]*Rekam Medik dan Informasi Kesehatan, Politeknik TEDC Bandung*
[2]*Teknik Informatika, Politeknik TEDC Bandung*
E-mail : ayuhendrati@poltektedc.ac.id[1], arisud@poltektedc.ac.id[2]*
**Corresponding author*

Abstract - The huge impact caused by the COVID-19 pandemic has made many people express their opinions on Twitter social media. There are various responses given by the community that are negative and positive. The dataset comes from kaggle with more than 750 tweets of data. Classification designed by the Naive Bayes method. Implementation through preprocessing, case folding, tokenizing, stopword removal, TF-IDF, and cross validation has been able to produce quite high accuracy. After classification, validation will be carried out with Cross Fold Validation. The best value is on cv5 where accuracy = 0.847, precision = 0.855, recall = 0.83, and f1 score = 0.842.

Keywords - COVID-19, Twitter, Sentiment, Naive Bayes

1. INTRODUCTION

One of the efforts that are reported to be effective in dealing with COVID-19 is to provide vaccinations. However, there is no vaccine that can fight this virus effectively and safely because there is still little information about this virus [1]–[3]. The government has also taken an approach in dealing with the COVID-19 emergency, which varies widely, but there are two main types of policies, namely policies to strengthen the capacity of the hospital system, and policies aimed at reducing the possibility of the virus, such as lockdowns and social distancing [4]. The positive impact of this pandemic period is that people are becoming more aware of existing technologies such as the use of online meeting media and the use of social media.

Social media is an online platform for communicating and interacting remotely without the need to be limited by space and time [5]–[7]. Social media is also a platform where we can get various information. There are several social media that have been popular and used by most people across the country and it was recorded that in January 2021, the total users of social media reached 4.2 billion. Some social media that have many users include Facebook, Instagram, Whatsapp, Youtube, Twitter etc. One of the functions of social media is the microblog function, which is a new service in blogging, because microblog is an answer to the anxiety of users who want faster speeds through their smartphones. Twitter belongs to a type of social media that allows the general public to post their ideas, reviews of others where they may know each other [8]. A feature of Twitter that allows us to share our opinions is tweets. These Tweets are a source of data which, if handled properly, will be able to produce a variety of useful information such as conducting public sentiment analysis on the COVID-19 virus.

Sentiment Analysis is a technique of extracting text data that is used to receive information about a sentiment whether it has positive, neutral, and negative values [9].

Sentiment analysis can be used for various opinion data, such as customer satisfaction data and data on an assessment of a performance. The Naive Bayes algorithm method is the algorithm that will be used for this research. Naive Bayes algorithm [9] is a classification method that uses probability techniques. This algorithm will perform calculations on a set of probabilities by adding up the frequencies and combinations of a dataset.

2. RESEARCH METHOD

2.1. Previous Research

Several studies that have been done previously on sentiment analysis using the Naive Bayes Classifier algorithm. A study using Naïve Bayes to see community perception on smooking behavior [10]. In this study, it is explained that the selling price of cigarettes in Indonesia in 2017 has increased. Therefore, the authors of this study try to look at the polemic of the selling price of cigarettes that are reported to have increased using the sentiment analysis method. The author uses the twitter platform as a test material for his analysis to find out whether the public is more inclined to positive or negative sentiment, or will be more neutral, so that the results of the analysis, the government is expected to determine its policy on the polemic of rising cigarette selling prices. In this research, the author uses the Naive Bayes Classifier data mining algorithm by performing data preprocessing with tokenizing, cleansing, normalization, and case folding stages. neutral by 85% with the Naive Bayes Classifier classification. Another study has been done to classify of the corona virus using Naïve Bayes [11]. It was explained that the corona virus that occurred for the first time in Indonesia was on March 2, 2020, which originated from Wuhan, China in December 2019. This virus quickly spread to all corners of the world. The author realizes that the topic of this pandemic has become a trending topic on the Twitter platform. By utilizing the tweet feature of twitter. There are various kinds of public responses ranging from positive, negative, and neutral comments. Therefore, the author wants to conduct a sentiment analysis on COVID-19 using tweets as a dataset. In its design, Edigia Yuni Savitri uses the Naive Bayes Classifier data mining algorithm. In the design using the preprocessing method, starting with case folding, tokenizing, stemming. After going through the preprocessing process, the weighting process is carried out using the Term Frequenzy – Inverse Document Frequency (TF-IDF) technique and cross validating. In the design the author gets the highest accuracy of 90.825%.

2.2. Data Mining

Data mining is a knowledge mining procedure or the extraction of potential, implied, and unknown information from a set of data [3], [12]–[15]. Data Mining [12], [16] has several advantages, namely it can handle large amounts of data, and has high dimensions. There are several types of Data Mining techniques, namely:

1. Description: a technique that performs Data Mining by describing the definition and trend of patterns contained in the dataset owned.
2. Estimation: a technique that is almost the same as classification, but the target variable is more directed to numerical data than to categories.
3. Prediction: in prediction, data mining is tasked with predicting unknown values in the future.
4. Classification: is a technique that has a target variable category, such as long, medium, and short.
5. Clustering: this technique collects records, observes, and forms classes of objects that have similarities.

6. Association: is a technique that has the task of finding attributes that come out at one time.

2.3. Text Mining

Text is a means of exchanging information. Text Mining is an analysis process by conducting a continuous data mining and using certain methods in its application. Text Mining mines data in the form of text originating from documents [14], [16]–[19]. The initial text data is partially unstructured text data. Therefore, we need a processing technique called data pre-processing so that text data becomes structured. Data pre-processing serves to make text data that was originally unstructured into structured data. Some examples of data pre-processing stages that are often used are tokenizing, stopword removal, and stemming [10].

2.4. Sentimen Analysis

Sentiment Analysis, also known as option mining, is a computational study of the analysis of emotions, opinions, evaluations, attitudes, judgments, sentiments, and subjectivity of all forms of text. Sentiment analysis is also considered as an opinion mining and is a continuous field of research that lies between various fields such as Data Mining, Natural Language Processing (NLP), and Machine Learning whose task is to focus on extracting sentiments in a sentence based on its content [2], [20]–[22]. Sources of data from sentiment analysis can be found at:

1. Opinions or opinions on blogs, microblogs or forums.
2. Posting on social media.
3. Electronic mail.
4. Comment on articles, issues, trending topics or reviews.

Sentiment analysis of data is very useful for expressing opinions from the masses or groups, it is used as a helping tool for companies by observing customer responses to products being sold. As has been done in previous studies, this technique can be used as a tool to observe how the audience responds to the film that is currently showing or is used to observe the public's response to political issues or artists that are currently circulating. This technique is used to find people's sentiments by respecting the source of a particular content [12].

2.5. Term Frequency – Inverse Document Frequency

Term Frequency – Inverse Document Frequency (TF-IDF) [23], [24] is a method used to weight the relationship of a word (term) to a document. TF-IDF combines two term weighting concepts, namely Term Frequecy (TF) or the frequency of occurrence of a word. in a specific document and the Inverse Document Frequency (IDF) the frequency of the inverse document that has the word. Term Frequency is. This method will calculate the weight of each token t in document d [13]. The formula for the TF-IDF method is as follows:

$$idft = \log\left(\frac{D}{dft}\right) \qquad (1)$$

$$\mathrm{w}(t, d) = tf(t, d) * id\mathrm{f} \qquad (2)$$

Information:

tf = how many words are searched in a document

D = total of all documents

df = total documents that have term t

IDF = Inverse Document Frequency (log2(D/df))

d = d document

t = t-th word of keyword

W = weight of the d-document to word ke-t

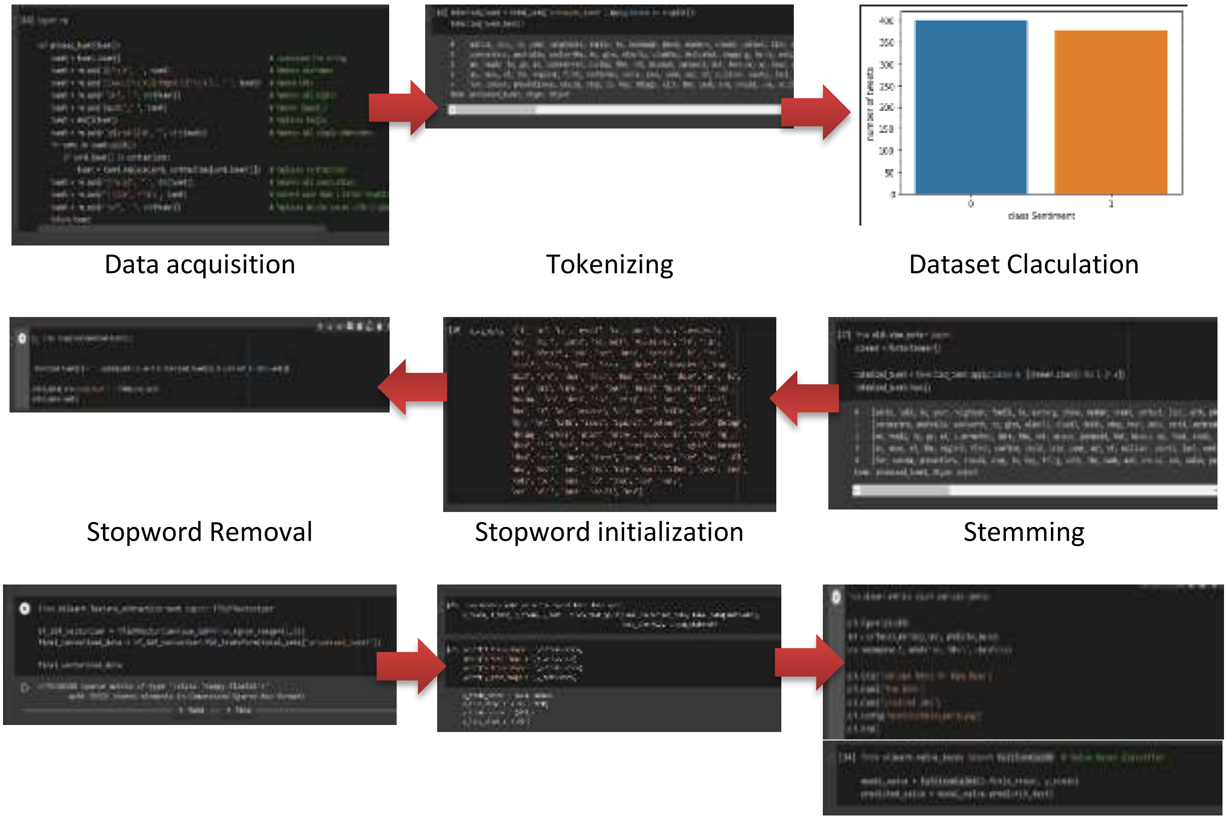
2.6. Naive Bayes Classification

The Naive Bayes Classifier (NBC) [2], [14], [25] algorithm is used for data classification. Naive Bayes Classifier is a classification algorithm based on Bayes theorem. This theorem was put forward by a British scientist named Thomas Bayes. Naive Bayes is a machine learning method that uses probability and statistical calculations. This algorithm shows excellent predictive performance and achieves comparable results with other classification methods such as decision trees and artificial neural networks. Naive Bayes bases its classification on a simplification estimate that the value possessed by an attribute is conditionally independent of each other if given a value that is output or expenditure. The advantage of using the Naive Bayes algorithm is that this method uses training data that is not large to be able to determine the approximate size of the dataset needed in the method for processing its classification. Naive Bayes in the process can work more accurately in the state of our world [14].

2.7. Data Collection

At this stage of data collection, data collection uses a dataset obtained from the kaggle dataset provider site with the name "Coronavirus tweets NLP - Text Classification" with the site address https://www.kaggle.com/datatattle/covid-19-nlp-text-classification. This dataset is publicly accessible. This dataset was created by someone named Aman Miglani. Aman Miglani pulls data from Twitter and performs manual tagging. This dataset consists of 779 rows and has several columns such as Location or the origin of the location of a tweet, Tweet at column or tweet date. The Original Tweet column or the content of the tweet, and the Label column which contains several sentiments such as whether a tweet is positive or negative, and others.

2.8. Proposed Method



Data acquisition

Tokenizing

Dataset Claculation

Stopword Removal

Stopword initialization

Stemming

Weighting

Splitting

Confusion Matrix

Figure 1. Research Stages

The first research stage is data acquisition by collecting data and information, the second is tokenization which sends the data via an API call called a token then the data set is ready to be managed into new information and displayed in a bar chart. After that, parse the word into basic words in the stemming process, next is to find unimportant words in the data and so on the unimportant word is removed after that it is weighted and separated again and the last is the Confusion Matrix used to measure the performance of the classification model in machine learning.

3. RESULTS AND DISCUSSION

In this research, the writer uses Naive Bayes Multinomial variation. Multinomial Naive Bayes is a variation of Naive Bayes which assumes that all attributes will be related to each other given the class attributes, and ignores the dependencies between attributes [15]. Furthermore, after the data goes through the Naive Bayes classification method, the writer makes a Confusion Matrix, which is one of the calculations to measure the performance of the model used by the author, namely sentiment analysis using the Naive Bayes method. After getting the Confusion Matrix we can find out the number of True Negative = 75, False Negative = 11, True Positive = 60, and False Positive = 10. Then after finding the True Negative, False Negative, True Positive, and False Positive values we can determine the accuracy of Naive Bayes classification modeling used as in Figure 2.

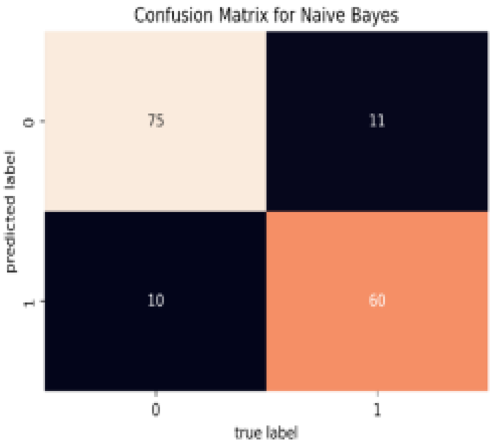


Figure 2. Confusion Matrix

Where :

TN (*True Negative*) = the actual calculation results and the algorithm classify negative

TP (*True Positive*) = the actual calculation result and the algorithm classify positive

FN (*False Negative*) = the actual result is positive but the algorithm classifies it as negative

FP (*False Positive*) = the actual result is negative but the algorithm classifies positive

Accuracy is an illustration of how strong the model that has been designed in the data classification process is correctly. From the calculations above, the writer can conclude that the accuracy obtained from the classification of sentiment analysis using the Naive Bayes method is 0.865 or 87%. Precision is a representation of the accuracy of the requested data with the prediction results that have been given by the algorithm that has been made. The precision of the calculation that has been done gets a value of 0.857 or if it is rounded it will be 86%. Sensitivity or what is called Recall from the success of the model in finding information data later. After calculating the recall from the Naive Bayes method used, the recall value was found to be 0.845 or 85% if rounded up and multiplied by 100%. F1 Score is a representation of the average comparison of recall and precision that has been weighted. The results obtained from the calculation of the f1 score above can be done after getting the precision and recall values first. Once known, it can be calculated from the f1 score which is worth 0.8509 or if it is rounded and multiplied by 100% it is 85%.

$$accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (3)$$

$$precision = \frac{(TP)}{(TP + FP)} \quad (4)$$

$$recall = \frac{(TP)}{(TP + FN)} \quad (5)$$

$$f1\ score = 2\ x \frac{(Recall\ x\ Precision)}{(Recall + Precision)} \quad (6)$$

The Validation process has the aim of conducting a test of the results of the Naive Bayes classification that has been carried out to measure the performance of the system that has been created. The author uses the Cross Fold Validation validation method. This method is carried out by the author in order to get maximum accuracy results. This method will perform k times for one model used by the author, namely Naive Bayes. This process is carried out by the author because it has a higher level of accuracy, because the validation process using Cross Fold Validation is not like the usual splitting process which only divides two datasets into training data and testing data in one step, but Cross Fold Validation performs splitting repeatedly according to fold that has been determined, so the data used as training data and testing data does not only depend on one splitting. The author uses the values of cv = 3, cv = 5, and cv = 7 because the cv above has the highest average result. Then the calculation after cv = 7 has a tendency to have an average accuracy, precision, recall, f1 score which is not greater than cv = 3, cv = 5, and cv = 7.

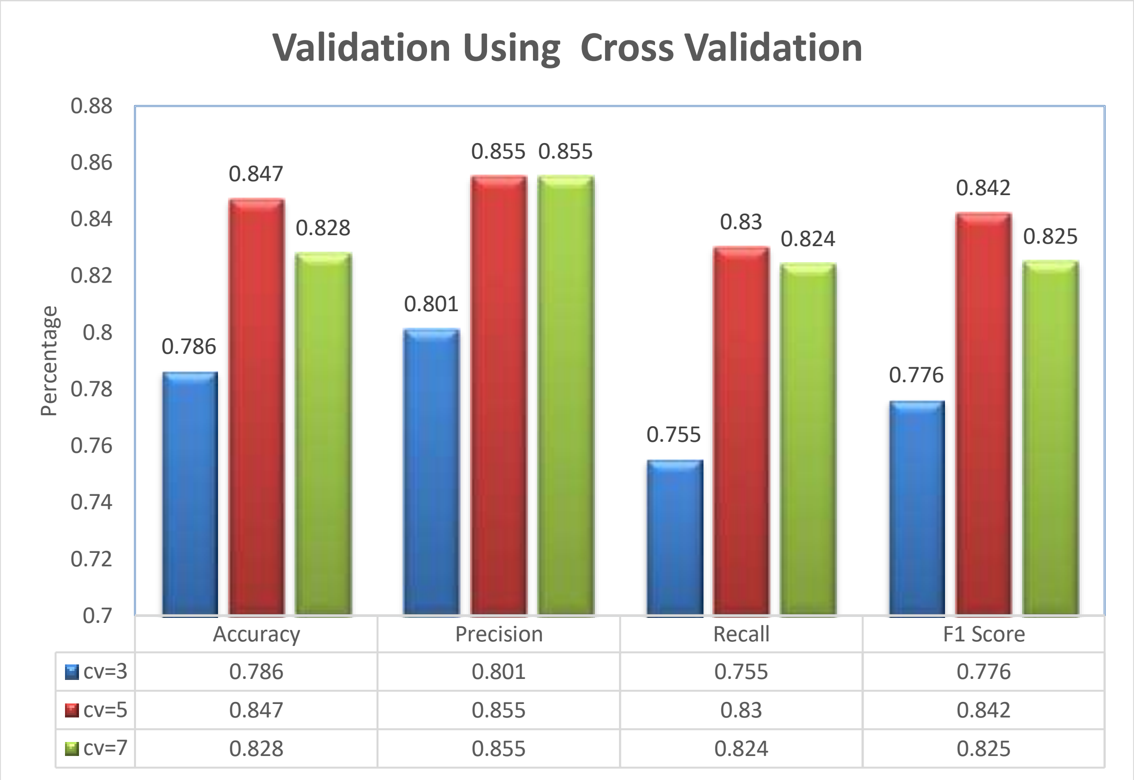


Figure 3. Comparison of Cross Validation

Based on the classification process that has been carried out using the Naive Bayes method and the results of the cross validation that has been passed with the values of cv = 3, cv = 5, and cv = 7, it is obtained as in Figure 3. In the classification process using the Naive Bayes method with cross validating, the results obtained from accuracy, precision, recall, f1 score. By using cv = 3, the accuracy value = 0.786, precision = 0.801, recall = 0.755, and f1 score = 0.776. In the classification process using the Naive Bayes method with cross validating, the results obtained from accuracy, precision, recall, f1 score. By using the value of cv = 5, the obtained value of accuracy = 0.847, precision = 0.855, recall = 0.83, and f1 score = 0.842. In the classification process using the Naive Bayes method with cross validating, the results obtained from accuracy, precision, recall, f1 score. By using the value of cv = 7, the obtained value of accuracy = 0.828, precision = 0.855, recall = 0.824, f1 score = 0.825.

4. CONCLUSION

Based on 779 tweet data that has been obtained regarding the public's response to the current pandemic health crisis, namely Coronavirus Disease 2019 (2019-nCov), as many as 401 public tweets gave negative responses, and 378 public tweets gave positive responses. From the data above, it can be concluded that public opinion or response with a dataset of 779 tweets has a negative response to the health crisis of the COVID-19 pandemic. The results of the Twitter sentiment analysis that have been obtained through the Kaggle website with the owner of the Aman Miglani dataset about the COVID-19 pandemic health crisis have been designed using the Naive Bayes data mining algorithm with the Term Frequency – Inverse Document Frequency (TF-IDF) weighting method using cross validating with the value of cv = 3, the value of accuracy = 0.786, precision = 0.801, recall = 0.755, and f1 score = 0.776, using the value of cv = 5, the value of accuracy = 0.847, precision = 0.855, recall = 0.83, and f1 score = 0.842, and accuracy = 0.828, precision = 0.855, recall = 0.824, f1 score = 0.825 using cv = 7. It is hoped that in further system development this system can be applied to all platforms, and can be integrated with various other algorithms such as K-Nearest Neighbor (KNN), and Super Vector Machine (SVM). It is also hoped that further research can use the weighting method by considering the data linkage seen from the frequency of occurrence of terms in various related categories using the Term Frequency – Relevance Frequency (TF-RF) weighting method.

REFERENCES

[1] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, dan Y. Samuel, "COVID-19 public sentiment insights and machine learning for tweets classification," *Inf.*, vol. 11, no. 6, hal. 1–22, 2020.

[2] R. Vijay, B. Vangara, K. Thirupathur, dan S. P. Vangara, "Opinion Mining Classification using Naive Bayes Algorithm," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 5, hal. 495–498, 2020.

[3] V. M., J. Vala, dan P. Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining," *Int. J. Comput. Appl.*, vol. 133, no. 9, hal. 7–11, 2016.

[4] N. S. Sattar dan S. Arifuzzaman, "Covid-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the usa," *Appl. Sci.*, vol. 11, no. 13, 2021.

[5] A. M. Almars, E. S. Atlam, T. H. Noor, G. ELmarhomy, R. Alagamy, dan I. Gad, "Users opinion and emotion understanding in social media regarding COVID-19 vaccine," *Computing*, vol. 104, no. 6, hal. 1481–1496, 2022.

[6] A. Umair dan E. Masciari, "Sentimental and spatial analysis of COVID-19 vaccines tweets," *J. Intell. Inf. Syst.*, 2022.

[7] S. A. Jafar Zaidi, I. Chatterjee, dan S. Brahim Belhaouari, "COVID-19 Tweets Classification during Lockdown Period Using Machine Learning Classifiers," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, hal. 1–8, Jul 2022.

[8] F. M. J. M. Shamrat *et al.*, "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 1, hal. 463–470, 2021.

[9] N. G. Ramadhan dan F. D. Adhinata, "Sentiment analysis on vaccine COVID-19 using word count and Gaussian Naïve Bayes," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 3, hal. 1765, 2022.

[10] M. Myslín, S.-H. Zhu, W. Chapman, dan M. Conway, "Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products," *J. Med. Internet Res.*, vol. 15, no. 8, hal. e174, Agu 2013.

[11] K. B. Priya Iyer dan S. Kumaresh, "Twitter sentiment analysis on coronavirus outbreak using machine learning algorithms," *Eur. J. Mol. Clin. Med.*, vol. 7, no. 3, hal. 2663–2676, 2020.

[12] N. M. Abdulkareem, A. Mohsin Abdulazeez, D. Qader Zeebaree, dan D. A. Hasan, "COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms," *Qubahan Acad. J.*, vol. 1, no. 2, hal. 100–105, 2021.

[13] Z. Ali, S. K. Shahzad, dan W. Shahzad, "Performance Analysis of Statistical Pattern Recognition Methods in KEEL," *Procedia Comput. Sci.*, vol. 112, no. 2017, hal. 2022–2030, 2017.

[14] P. Sharma dan T.-S. Moh, "Prediction of Indian election using sentiment analysis on Hindi

Twitter," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, hal. 1966–1971.

[15] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, dan S. Mukherjee, "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques," *CSI Trans. ICT*, vol. 4, no. 2–4, hal. 313–319, 2016.

[16] T. Mustaqim, K. Umam, dan M. A. Muslim, "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," *J. Phys. Conf. Ser.*, vol. 1567, no. 3, hal. 8–15, 2020.

[17] W. B. Zulfikar, M. Irfan, C. N. Alam, dan M. Indra, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," in *2017 5th International Conference on Cyber and IT Service Management, CITSM 2017*, 2017.

[18] P. Liu, H. han Zhao, J. yu Teng, Y. yan Yang, Y. feng Liu, dan Z. wei Zhu, "Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark," *J. Cent. South Univ.*, vol. 26, no. 1, hal. 1–12, 2019.

[19] K. M. A. Hasan, M. S. Sabuj, dan Z. Afrin, "Opinion mining using Naïve Bayes," in *2015 IEEE International WIE Conference on Electrical and Computer Engineering, WIECON-ECE 2015*, 2016, hal. 511–514.

[20] V. Dhanalakshmi dan D. Bino, "Opinion mining from student feedback data using supervised learning algorithms," in *2016 3rd MEC International Conference on Big Data and Smart City, ICBDSC 2016*, 2016, hal. 1–5.

[21] S. Dey Sarkar, S. Goswami, A. Agarwal, dan J. Aktar, "A Novel Feature Selection Technique for Text Classification Using Naïve Bayes," *Int. Sch. Res. Not.*, vol. 2014, hal. 1–10, 2014.

[22] S. Kaur, G. Sikka, dan L. K. Awasthi, "Sentiment Analysis Approach Based on N-gram and KNN Classifier," *ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun.*, hal. 13–16, 2018.

[23] B. Bhutani, N. Rastogi, P. Sehgal, dan A. Purwar, "Fake News Detection Using Sentiment Analysis," *2019 12th Int. Conf. Contemp. Comput. IC3 2019*, hal. 1–5, 2019.

[24] K. Poddar, G. B. D. Amali, dan K. S. Umadevi, "Comparison of Various Machine Learning Models for Accurate Detection of Fake News," *2019 Innov. Power Adv. Comput. Technol. i-PACT 2019*, hal. 1–5, 2019.

[25] F. C. Permana, Y. Rosmansyah, dan A. S. Abdullah, "Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media," *J. Phys. Conf. Ser.*, vol. 893, no. 1, 2017.