Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 4 (2022) 585 - 590 ISSN Media Electronic: 2580-0760

Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis

Najirah Umar[1], M. Adnan Nur[2]
[1,2]Teknik Informatika, STMIK Handayani Makassar
[1]najirah@handayani.ac.id, [2]adnan@handayani.ac.id

Abstract

Indonesian General Analysis Dataset is a dataset sourced from social media twitter by using keywords in the form of conjunctions to get a dataset that does not only focus on a particular topic. The use of Indonesian language datasets with general topics can be used to test the accuracy of the classification model so as to provide additional reference in choosing the right methods and parameters for sentiment analysis. One of the algorithms which in several studies produces the highest level of accuracy is naive Bayes which has several variations. This study aims to obtain the method with the best accuracy from the naive Bayes variation by setting the minimum and maximum document frequency parameters on the Indonesian General Analysis Dataset for sentiment analysis. The naive Bayes classifier variations used include Bernoulli naive Bayes, gaussian naive Bayes, complement naive Bayes and multinomial naive Bayes. The research stage begins with downloading the dataset. Preprocessing becomes the next stage which consists of tokenizing, stemming, converting abbreviations and eliminating conjunctions. In the preprocessed data, feature extraction is carried out by converting the dataset into vectors and applying the TF-IDF method before entering the sentiment analysis classification stage. Tests in this study were carried out by applying the minimum document frequency (min-df) and maximum document frequency (max-df) for each variation of naive Bayes to obtain the appropriate parameters. The test uses k-fold cross validation of the dataset to divide the training data and sentiment analysis test data. The next confusion matrix is made to evaluate the level of accuracy.

Keywords: *sentiment analysis, indonesian dataset, bernoulli nave bayes, gaussian nave bayes, multinomial nave bayes, complement nave bayes*

1. Introduction

Sentiment analysis has become one of the main technologies in obtaining information from social media. The field of sentiment analysis has grown and it is possible to explore various fields such as marketing, health, banking and politics[1]. Machine learning is a commonly used approach in the application of sentiment analysis[2]. There are two types of machine learning, namely supervised learning and unsupervised learning. Several studies have shown that supervised learning approaches such as support vector machine algorithms and nave Bayes algorithms are most often used and have the highest level of accuracy[3][4]. The nave Bayes algorithm and support vector machine are often compared for accuracy in the application of sentiment analysis which in several studies shows that nave Bayes has a higher level of accuracy.[5][6][7]. Naïve Bayes has several classic variants, namely multinomial, Bernoulli and Gaussian[8]. In addition, there is also the development of the classic variant of nave bayes and one of them is complement nave bayes which is the development and adaptation of multinomial nave bayes[9].

The application of nave Bayes in sentiment analysis requires a dataset that already has a sentiment label as training data to form patterns in classifying and predicting the sentiment of a text.[10]. These datasets are generally obtained from social media and the process of determining the label is done subjectively and has no concrete value. The data is based on human opinion which can differ from one another. This of course tends to be difficult in the machine learning training process[2]. In addition, most of the *data set* that are available or used in research and publications are only intended for certain topics whose labeling is subjective based on the opinion of the author on the topic so that *data set* affect the level of accuracy if it is used to classify or predict the sentiment of texts on other topics. Availability *data set* which topics are general in nature, not in English, are still minimal, especially

Accepted: 26-06-2022 | Received in revised: 15-08-2022 | Published: 22-08-2022

datasets in Indonesian. One of the studies that discusses the creation of Indonesian language datasets whose general topics are made by Ferdiana, Redi et al under the name *Indonesian General Sentiment Analysis Data set* [11]. *Data set* compiled containing text originated from *twitter* a total of 10,806 *tweets* and have been labeled with positive, negative and neutral values. The dataset is tested by comparing its accuracy value with a comparison dataset using an algorithm *support vector machine*,*k-nearest neighbors* and *stochastic gradient descent*. The results of the tests carried out show accuracy *Indonesian General Sentiment Analysis Data set* and comparable comparison datasets.

Use *data set* Indonesian language with general topics needs to be studied with several methods and parameters to get an effective model in its application. With the appropriate model, this dataset can be used as training data in applying sentiment analysis to various topics on social media.

In encouraging the use of this Indonesian language dataset and to contribute to the development of Indonesian sentiment analysis, this study aims to apply variations of the algorithm *naive bayes*on *Indonesian General Sentiment Analysis Dataset* by setting parameters *minimum document frequency* (min-df) and *maximum document frequency* (max-df) to find out and compare the resulting accuracy. Validation of accuracy level is done with 10 *fold cross validation* for training data and test data. The results of this study certainly provide additional references in choosing the appropriate method or approach if using the *Indonesian General Sentiment Analysis Data set* as training data in future sentiment analysis.

2. Research methods

In achieving the research objectives, the authors have designed and implemented the research stages which can be seen in Figure 1.

This research stage begins with preparing the downloaded dataset, then pre-processing which aims to reduce noise when classifying. Feature extraction from the preprocessing results is carried out with a count vectorizer, determination of min-df and max-df and weighting with TF-IDF. The results of term-frequency tweets are then divided into training data and test data using 10 fold cross validation to be tested on naive Bayes variations.

The following is a description of each stage of the research.

2.1. *Indonesian General Analysis Datasets*

The primary data used in this study is an Indonesian language dataset. The dataset comes from the results of research made by Ferdiana, Redi et al totaling 10,408 tweets with general topics downloaded via the linkhttp://ugm.id/idsadataset [11]. This downloaded dataset has been labeled with a value of 0 for neutral, 1 for positive and -1 for negative as shown in table 1 with a ratio of 2:1:1. The dataset provided has gone through several pre-processing stages such as cleaning symbols and disturbing characters (noise), stemming and deleting conjunctions or stop words.

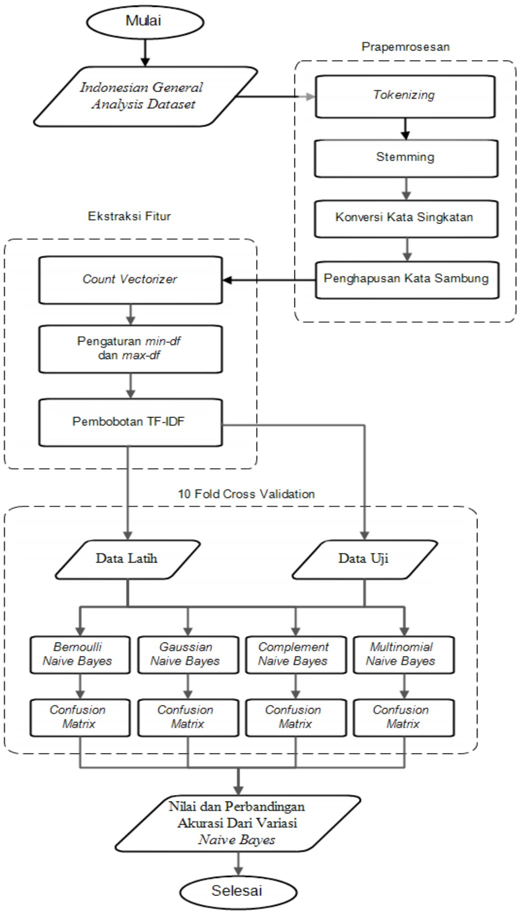

Figure 1. Research Stages

Table 1. Indonesian General Analysis Datasets

Label	*Tweet*
0	yg selama ini nungguin video dari aku nih retweet yg mau full
1	Mau bantu aku nabung
1	selamat hari jumat buat kamu yg selalu bikin rindu ini kumat
-1	aku tak paham betul lelaki tak suka perempuan pakai makeup
-1	iya hubungin aku aja ya butuh dosa berapa banyak lagi wkwk
0	aku ngatau nih bang definisi berhasil apa nga
1	kalau kamu mau dihormati harus mau menghormati

2.2. Preprocessing

*Indonesian General Analysis Data set s*used have passed the pre-processing stage, but after reviewing there are still some words that have affixes and abbreviations so that in this study the processing stage

DOI: https://doi.org/10.29207/resti.v6i4.4179

is still carried out. The preprocessing carried out is further divided into several stages as follows:

2.2.1 Tokenizing

Tokenizing is the stage to break the sentence into the words that compose it[12]. This stage is used to make it easier to carry out the next pre-processing stage which is word-oriented.

2.2.2 Stemming

Stemming is the process of changing affixed words into root words. This process is often used in research related to text mining. Stemming has an effect on increasing the accuracy of sentiment analysis[13]. One method that can be used for stemming Indonesian is the Nazief and Adriani algorithm. In several studies, this algorithm has the highest accuracy compared to other stemming algorithms[14][15][16]. In this study, to implement the Nazief and Adriani algorithm, the author uses the Sastrawi python Library

2.2.2. Abbreviation Conversion

Limiting the number of characters to 280 in messages that can be uploaded on Twitter makes users tend to use abbreviations. Commonly used abbreviations as in table 2 will be used to convert these abbreviations into standard words.

Table 2. Conversion of Abbreviations

Abbreviation	Raw Words
aja	saja
aq	aku
ato	atau
bhw	bahwa
blm	belum
brp	berapa
gak	tidak
tdk	tidak

The abbreviations provided in the study amounted to 116 words and the results were converted into standard words

2.2.3. Removal of conjunctions (stop words)

One of the methods in preprocessing text analysis is the removal of conjunctions or stop words removal. Words that are considered general and have little effect on text analysis will be removed. The application of stop words removal in preprocessing can improve the accuracy and performance of sentiment analysis classification[17]. In this study, the author uses the stop word remover function from the literary python library to implement the removal of conjunctions at the preprocessing stage. Table 3 shows an example of removing conjunctions in this study.

Table 3. Stop word Removal

Tweet	Stop word Removal Results
yang selama ini tunggu video dari aku ini retweet yang mau full	selama tunggu video aku retweet mau full
mau bantu aku nabung	Mau bantu aku nabung
selamat hari jumat buat kamu yang selalu bikin rindu ini kumat	selamat hari jumat buat kamu selalu bikin rindu kumat

2.3. Feature Extraction

In the implementation of the classification method in sentiment analysis, features are needed that become indicators in determining the class of a sentence. This feature is obtained by performing feature extraction that begins with tokenizing. Tokenizing aims to change sentences into simpler forms which in this study are formed into words or terms. In this study, the tokenizing results are placed in an array as shown in table 4.

Table 4. Tokenizing Results

Tweet	*Tokenizing*
selama tunggu video aku retweet mau full	[selama, tunggu, video, aku, retweet, mau, full]
mau bantu aku nabung	[mau, bantu, aku, nabung]
selamat hari jumat buat kamu selalu bikin rindu kumat	[selamat, hari, jumat, buat, kamu, selalu, bikin, rindu, kumat]

After the tokenizing results are obtained, the next step in feature extraction is the countvectorizer and term frequency - inverse document frequency (TF - IDF). Here's the description:

2.3.1. Countvectorizer

This stage is used to get the frequency of occurrence of words in a sentence and placed in a vector[18]. In the count vectorizer, words that rarely appear in tweets tend to be covered even though these words are important words in sentiment labeling in feature extraction, this is generally handled using TF-IDF[19]. An example of the form of the count vectorizer in this study can be seen in table 5.

Table 5. Example of Countvectorizer Results

tweets	I	help	for	day	you	want to	miss	videos
T1	1	0	0	0	0	1	0	1
T2	1	1	0	0	0	0	0	0
T3	0	0	1	1	1	0	1	0

To implement this method, the author uses the countervertorizer function from the sklearn python library

2.3.2. Determination of min-df and max-df

*Min-df (minimum document frequency)*It is used to remove words with a small frequency of occurrence. While max-df (maximum document frequency) is useful for eliminating words that often appear in tweets. This study tested several values of min-df and max-df to see the difference in the level of accuracy produced by the variation of nave Bayes. The min-df values tested are from 0 to 0.003 (0.3%) while the max-df is from 0.1 (10%) to 1 (100%).

2.3.3. *Term Frequency – Inverse Document Frequency*(TF-IDF)

TF-IDF is a method that aims to give weight to frequently used words. Term Frequency is the number of words in a word vector of a sentence divided by the total words in the word vector. As for the Inverse Document Frequency, it aims to reduce the weight of words if they exist in all documents[20]. In several studies, the use of TF-IDF in sentiment analysis shows an increase in accuracy[11][21]. Similar to the countvectorizer, the implementation of TF-IDF in this study uses the Tfidf Transformer function from the sklearn python library.

2.4. 10 Fold Cross Validation

*Dataset*used as primary data from this research amounted to 10,408 tweets that went through the pre-processing and feature extraction stages. To measure the level of accuracy of the classification model used, a test validation method is needed, namely K-fold cross validation because the number of datasets is quite large. The data will be tested on four variations of nave Bayes, namely Bernoulli, Gaussian, complement and multinomial nave Bayes. To share the training data and testing data from the dataset, K-Fold Cross Validation is used. K-Fold Cross Validation is a testing method by dividing the entire data into training data and testing data[22]. To see the performance of each nave Bayes variation, the author sets the value of K=10 on fold cross validation with 10 iterations of training and testing.

Testing results from each iteration of 10 fold cross validation on the Nave Bayes variation are then evaluated. The method used in the evaluation of this research model is the confusion matrix to get the accuracy value.

3. Results and Discussion

Data set which has gone through preprocessing and feature extraction with TF-IDF will be tested on each variation of nave Bayes. In this test, there are 30 datasets that have different combinations of min-df and max-df. The min-df values used consist of: 0; 0.00001; 0.000005; 0.0001; 0.0001; 0.0005; 0.001; 0.0015; 0.002; 0.0025 and 0.003. As for the max-df consists of: 0.1; 0.5 and 1. For the application of each variation of nave bayes in this study using the sklearn python library. The following are the results and discussion of the application of nave Bayes variations in the Indonesian General Analysis Dataset.

3.1. *Bernaoulli Naive Bayes*

The results of applying Bernoulli Nave Bayes to the dataset can be seen in the table 6.

Table 6. Bernoulli Naïve Bayes Evaluation Results

Min-df	*Max-df*		
	0.1	0.5	1
0	0.5967	0.5994	0.5994
0.00001	0.5967	0.5994	0.5994
0.000005	0.5967	0.5994	0.5994
0.0001	0.6189	0.6235	0.6235
0.0005	0.6226	0.6253	0.6253
0.001	**0.6337**	0.6263	0.6263
0.0015	0.6170	0.6198	0.6198
0.002	0.6068	0.6142	0.6142
0.0025	0.6105	0.6115	0.6115
0.003	0.6124	0.6170	0.6170

This nave Bayes variation produces the highest accuracy value of 0.6337 which is found at min-df 0.001 and max-df 0.1. The lowest accuracy was obtained at min-df 0 and 0.00001 and max-df 0.1 with a value of 0.5967.

3.2. *Gaussian Naive Bayes*

The results of applying gaussian nave bayes to the dataset can be seen in table 7.

Table 7. Gaussian Naïve Bayes Evaluation Results

Min-df	*Max-df*		
	0.1	0.5	1
0	0.4903	0.4903	0.4903
0.00001	0.4903	0.4903	0.4903
0.000005	0.4903	0.4903	0.4903
0.0001	0.4940	0.4940	0.4940
0.0005	0.4172	0.4154	0.4154
0.001	0.4128	0.4274	0.4274
0.0015	0.4736	0.4709	0.4709
0.002	0.4783	0.4783	0.4792
0.0025	0.4792	**0.5051**	0.5014
0.003	0.5014	0.4968	**0.5051**

The application of gaussian nave bayes on the dataset produces the highest value for accuracy at min-df 0.0025 and max-df 0.5 with a value of 0.5051. Meanwhile, the lowest accuracy value is 0.4154 at min-df 0.0005 and max-df 0.5.

3.3. *Complement Naive Bayes*

The results of the nave Bayes complement evaluation of the dataset can be seen in the following table.

Table 8. Evaluation Results of Complement Naïve Bayes

Min-df	*Max-df*		
	0.1	0.5	1
0	0.5819	0.5819	0.5819
0.00001	0.5819	0.5819	0.5819
0.000005	0.5819	0.5819	0.5819
0.0001	0.6068	0.6105	0.6105
0.0005	0.6216	**0.6235**	**0.6235**
0.001	0.6189	0.6115	0.6115
0.0015	0.6004	0.5930	0.5930
0.002	0.5920	0.5902	0.5902
0.0025	0.5874	0.5772	0.5772
0.003	0.5874	0.5809	0.5809

Accuracy the highest in table 8 is 0.6235 with min-df 0.0005 and max-df 0.5. for the lowest accuracy value is in the combination of min-df 0.003 and max-df 0.5 with an accuracy of 0.5809.

DOI: https://doi.org/10.29207/resti.v6i4.4179

3.4. *Nave Bayes Multinomial*

The results of the evaluation of the application of multinomial nave Bayes can be seen in table 9.

Table 9. Evaluation Results of Nave Bayes Multinomial

Min-df		*Max-df*	
	0.1	0.5	1
0	0.6133	0.6105	0.6105
0.00001	0.6133	0.6105	0.6105
0.000005	0.6133	0.6105	0.6105
0.0001	0.6216	0.6179	0.6179
0.0005	0.6346	**0.6374**	**0.6374**
0.001	0.6346	0.6337	0.6337
0.0015	0.6300	0.6300	0.6300
0.002	0.6290	0.6281	0.6281
0.0025	0.6189	0.6180	0.6180
0.003	0.6180	0.6190	0.6190

The highest accuracy value from the application of multinomial nave Bayes was obtained at min-df of 0.0005 and max-df 0.5 with an accuracy value of 0.6374. The lowest accuracy value is 0.6105 at min-df 0.0005 and max-df 0.5.

3.5. Comparison of Evaluation Results

The results of the application of nave Bayes variation obtained the highest accuracy value on multinomial nave Bayes of 0.6374 at min-df 0.0005 and max-df 0.5. The highest accuracy value was then obtained with Bernoulli nave Bayes with an accuracy of 0.6337, then complement of nave Bayes of 0.6235 and the lowest was Gaussian nave Bayes with an accuracy of 0.5051. The following graph shows the comparison of the highest accuracy resulting from the determination of min-df and max-df .

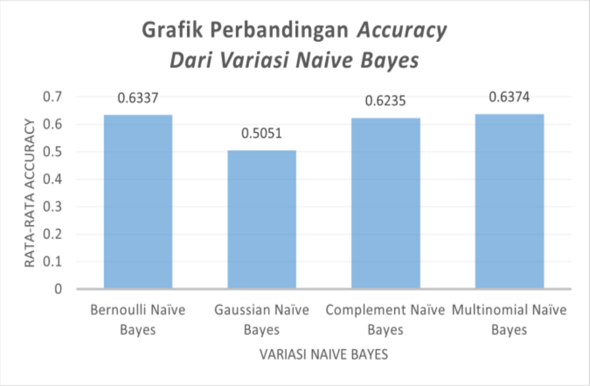

Figure 2. Accuracy Comparison Chart

4. Conclusion

Based on the results of the application of nave Bayes variations in the Indonesian General Analysis Dataset with the determination of min-df and max-df, the highest accuracy value is obtained in multinomial nave Bayes. The average value of min-df which has high accuray is 0.0005 and max-df 0.5. For further research, optimization on the pre-processing side can be done by normalizing spelling errors and slang words in Indonesian General Analysis Datasets.

References

[1] CA Iglesias and A. Moreno, “Sentiment Analysis for Social Media,”*Applied Sciences 2019, Vol. 9, Page 5037*, vol. 9, no. 23, p. 5037, Nov. 2019, doi:10.3390/APP9235037.

[2] A. Mittal and S. Patidar, “Sentiment Analysis on Twitter Data: A Survey,”*Proceedings of the 2019 7th International Conference on Computer and Communications Management*, 2019, doi:10.1145/3348445.

[3] S. Pandya and P. Mehta, “A ReviewOn Sentiment Analysis Methodologies, Practices And Applications Dog Acoustic Analysis View project Pollution Monitoring System View project A Review On Sentiment Analysis Methodologies, Practices And Applications,” International Journal Of Scientific & Technology Research, vol. 9, p. 2, 2020.

[4] I. Odun-Ayo, R. Goddy-Worlu, L. Ajayi, al -, R. Baragash, and H. Aldowah, “Sentiment analysis in higher education: a systematic mapping review,”*Journal of Physics: Conference Series*, vol. 1860, no. 1, p. 012002, Mar. 2021, doi: 10.1088/1742-6596/1860/1/012002.

[5] R. Ardianto, T. Rivanie, Y. Alkhalifi, FS Nugraha, and W. Gata, “Sentiment Analysis On e-SportsFor Education Curriculum Using Naive Bayes And Support Vector Machine,” Journal of Computer and Information Science, vol. 13, no. 2, pp. 109–122, Jul. 2020, doi:10.21609/JIKI.V13I2.885.

[6] S. Dyah Anggita and Ikmah, “Algorithm Comparation of Naive Bayes and Support Vector Machine based on Particle Swarm Optimization in Sentiment Analysis of Freight Forwarding Services,”*RESTI Journal (System Engineering and Information Technology)*, vol. 4, no. 2, pp. 362–369, Apr. 2020, doi:10.29207/RESTI.V4I2.1840.

[7] S. Tamrakar, BK Bal, and RB Thapa, “Aspect Based Sentiment Analysis of Nepali Text Using Support Vector Machine and Naive Bayes,”*Technical Journal*, vol. 2, no. 1, pp. 22–29, Nov. 2020, doi:10.3126/TJ.V2I1.32824.

[8] S. Xu, “Bayesian Naïve Bayes classifiers to textclassification:," https://doi.org/10.1177/0165551516677946, vol. 44, no. 1, pp. 48–59, Nov. 2016, doi:10.1177/0165551516677946.

[9] CH Yutika, A. Adiwijaya, and S. al Faraby, "Aspect-Based Sentiment Analysis on Female Daily Review Using TF-IDF and Naïve Bayes,"*Journal of Media Informatics BUDIDARMA*, vol. 5, no. 2, pp. 422–430, Apr. 2021, doi:10.30865/MIB.V5I2.2845.

[10] M. Adnan Nur, "Comparison of Levenshtein Distance and Jaro-Winkler Distance for Word Correction in Preprocessing Twitter User Sentiment Analysis,"*Journal of Electrode Focus: Electrical Energy, Telecommunications, Computers, Electronics and Controls)*, vol. 6, no. 2, pp. 88–93, Jun. 2021, doi:10.33772/JFE.V6I2.17751.

[11] R. Ferdiana, F. Jatmiko, DD Purwanti, A. Sekar, T. Ayu, and WF Dicka, “Indonesian Datasets for Sentiment Analysis,”*National Journal of Electrical Engineering and Information Technology (JNTETI)*, vol. 8, no. 4, pp. 334–339, Nov. 2019, doi:10.22146/JNTETI.V8I4.533.

[12] J. Resti and F. Selva Jumeilah, “Implementation of Support Vector Machine (SVM)for Research Categorization,” RESTI Journal (System Engineering and Information Technology), vol. 1, no. 1, pp. 19–25, Jul. 2017, doi:10.29207/RESTI.V1I1.11.

[13] N. Saputra, TB Adji, and AE Permanasari, "Analysis of President Jokowi's Data Sentiment with Normalization and Stemming Preprocessing Using Naive Bayes and SVM methods,"*Journal of Informatics Dynamics*, vol. 5, no. 1, 2015.

[14] D. Wahyudi, T. Susyanto, D. Nugroho, P. Informatics Engineering Studies, S. Sinar Nusantara Surakarta, and P. Information Systems Studies, "Implementation and Analysis of Nazief & Adriani Dan Porter Stemming Algorithms in Indonesian Language Documents,"*SINUS Scientific Journal*, vol. 15, no. 2, Jul. 2017, doi:10.30646/SINUS.V15I2.305.

[15] A. Rahmatullah*et al.*, “Comparison between the Stemmer Porter Effect and Nazief-Adriani on the Performance of Winnowing Algorithms for Measuring Plagiarism ,” Article in International Journal on Advanced Science Engineering and Information Technology, 2019, doi: 10.18517/ijaseit.9.4.8844.

DOI: https://doi.org/10.29207/resti.v6i4.4179

[16] J. Jumadi, DS Maylawati, LD Pratiwi, and MA Ramdhani, “Comparison of Nazief-Adriani and Paice-Husk algorithm for Indonesian text stemming process,”*IOP Conference Series: Materials Science and Engineering*, vol. 1098, no. 3, p. 032044, Mar. 2021, doi:10.1088/1757-899X/1098/3/032044.

[17] AF Hidayatullah, “The Effect of Stopwords on Tweet Classification Performancein Indonesian,” JISKA (Journal of Informatics Sunan Kalijaga), vol. 1, no. 1, pp. 1–4, May 2016, doi: 10.14421/jiska.2016.11-01.

[18] M. Priandi and Painem, "Analysis of Community Sentiment Against Online Learning in the Era of the Covid-19 Pandemic on Twitter Social Media Using Countvectorizer Feature Extraction and the K-Nearest Neighbor Algorithm,"*National Seminar on Computer Science Students and Its Applications (SENAMIKA) Jakarta-Indonesia*, pp. 311–319, 2021.

[19] A. Turmudi and K. Syarief Yasah, "Indonesian Tweet Sentiment Analysis Using Extraction Features and Cross Validation Techniques Against Naive Bayes Models,"*SIGMA Information Technology Journal*, vol. 10, no. 4, pp. 2407–3903, 2020.

[20] SA Pratomo, S. al Faraby, and MD Purbolaksono, "Sentiment Analysis of the Effect of Combination of TF-IDF and Lexicon Feature Extraction on Film Reviews Using the KNN Method," in*eProceedings of Engineering*, 2021, pp. 10116–10126.

[21] S. Fransiska, R. Rianto, and AI Gufroni, “Sentiment Analysis ProviderBy.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method,” Scientific Journal of Informatics, vol. 7, no. 2, pp. 203–212, Nov. 2020, doi: 10.15294/SJI.V7I2.25596.

[22] K. Gde Sukarsa and I. Gusti Ayu Made Srinadi, “Discriminant Analysis on Classification of Villages in KabupatenTabanan Using the K-Fold Cross Validation Method,” E-Jurnal Mathematics, vol. 6, no. 2, pp. 106–115, 2017, doi:10.24843/MTK.2017.v06.i02.p154.

DOI: https://doi.org/10.29207/resti.v6i4.4179