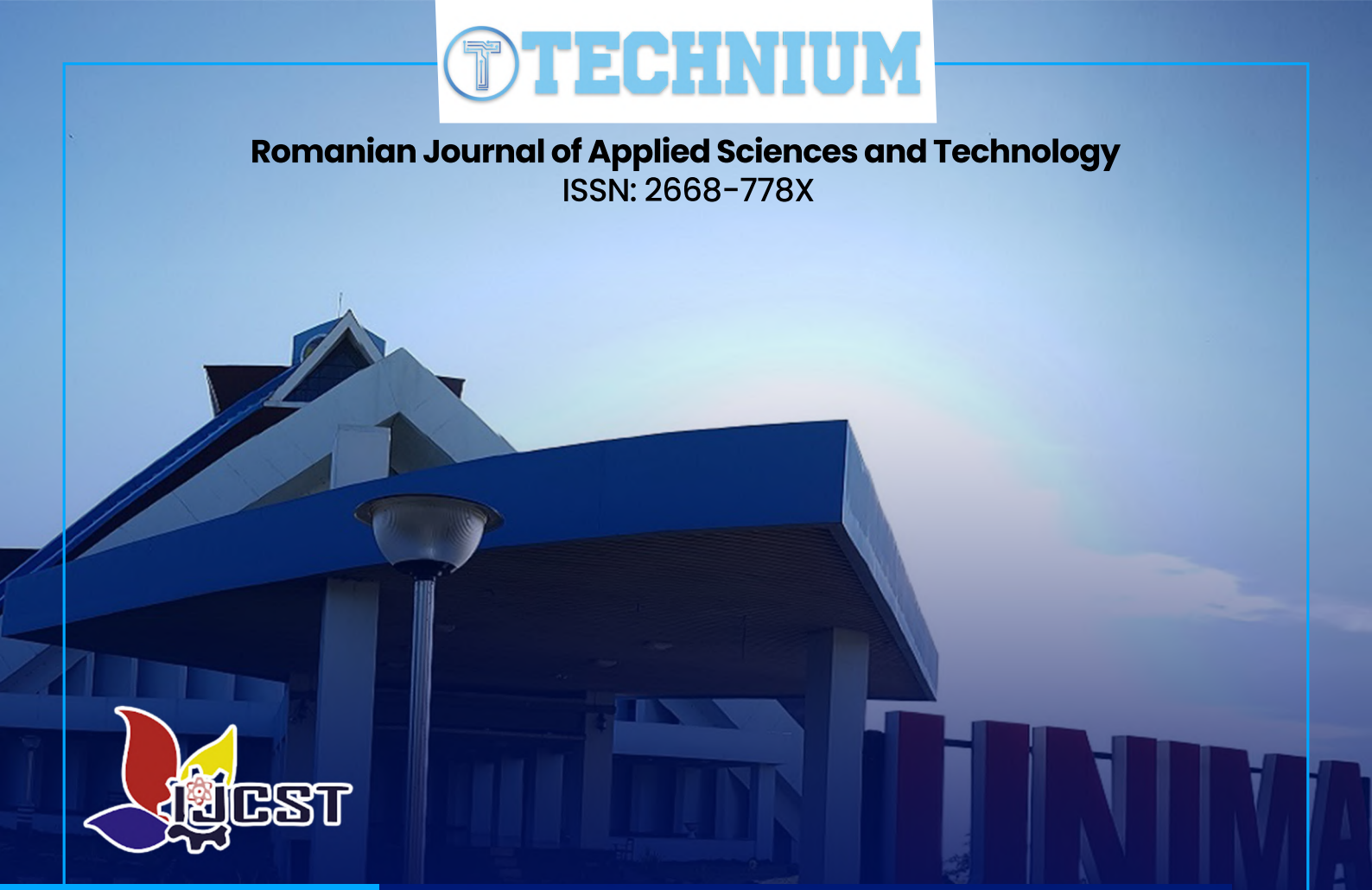
TECHNIUM

Romanian Journal of Applied Sciences and Technology
ISSN: 2668-778X



vol. 17 / 2023

The 7th International Conference on Science Technology

organized by
Faculty of Social Science and
Law Universitas Negeri Manado and
Consortium of International Conference
on Science and Technology

The Innovation Breakthrough in Digital and Disruptive Era

PLUS COMMUNICATION

International Communication & PR

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 17, 440-445, November, 2023
ISSN: 2668-778X
www.techniumscience.com

Sentiment Analysis for Indonesian Salt Policy uses Naïve Bayes and Information Gain Methods

Yeni Kustiyahningsih[1,*], *Ikromul* Islam [1], *Bain Khusnul* Khotimah[1] , and *Jaka* Purnama [2]

[1]Department of Informatics engineering, University of Trunojoyo Madura
[2]Department of Industrial Engineering, Universitas 17 Agustus 1945, Surabaya

Abstract. Salt production is one of the concerns of the Indonesian government. Several government policies such as salt imports have had a major impact on local salt farmers. Other factors are due to the increased demand for salt, decreased domestic salt production which is unfavorable due to weather factors, and the price of imported salt is lower than that of local salt. Many people express their opinions regarding the salt import policy, via Twitter social media. Sentiment analysis can be applied to analyze tweets or writings by the public regarding salt import policies and classify the data. This study uses the naïve Bayes classifier algorithm model as a sentiment classification algorithm on Twitter social media tweets. The classification process uses the Naïve Bayes algorithm. The feature extraction and weighting method is the TF-IDF method. Not all of the features resulting from the TF-IDF process are used, so feature selection is carried out using the information gain method. Model testing was carried out 5 times with 500 data, using feature selection and without feature selection. Without feature selection, the highest accuracy result is 84% at K=4, while without feature selection it produces an accuracy of 71% at K=3, so there is an increase of 13%.

1. Introduction

Indonesia is a country that has a long coastline with a total coastline length of 99,093 kilometers[1]. This is used by the Indonesian population as a source of livelihood. For example, the use of the sea or beach is to produce Indonesian local salt. Salt is one of the basic needs which is Indonesia's strategic commodity that is needed by all Indonesian people both individually and as an industrial raw material. Without salt, the production processes of the food/beverage and chemical industries will be hampered and will also hamper national economic growth[1]. Domestic salt production is one of the concerns of the Indonesian government. But on the other hand, policies issued by the government such as salt imports have had a major impact on salt farmers who are Indonesian local salt producers. According to a news source, liputan6.com shows that in 2021 total salt imports will reach 3.07 tons. This is influenced by several factors including the quality of local salt which is still inferior to imported salt. Other factors are caused by the increasing demand for salt, domestic salt production is not good due to weather factors, and the price of imported salt is lower than local salt. This certainly has an impact on decreasing the motivation and income of local salt farmers [1]. This salt import policy elicited many responses from the Indonesian people who were concerned about local salt farmers.

Many people express their opinions regarding the salt import policy, including through the social media Twitter. In terms of conveying an opinion, Twitter is one of the most popular text-based social media [2]. The ease of using Twitter in expressing opinions is a factor for social users to write responses regarding government policies. Public writings on Twitter social media can be used by the government to find out opinions or responses from the public towards policies made. The government can find out the positive and negative responses submitted by twitter users as a material for consideration in making policies including salt import policies. Tweet data from twitter can be classified into positive or negative opinion by using sentiment analysis. Sentiment analysis is the process of analyzing text data originating from the writings or opinions of the public in response to an issue or problem [3]. Recently, sentiment analysis has become a type of research that has received wide attention because it can produce important information such as knowing public opinion, politics, and decision-making processes [4]. The main purpose of sentiment analysis is to analyze a document, review, and comments and categorize it as a positive, negative, or neutral sentiment [2]. Sentiment analysis can be applied to analyze tweet data or public writings submitted via Twitter on salt import policies and classify the data.

Twitter data classification can use a data classification model. Several models of data classification algorithms that are often used include the Naïve Bayes algorithm, K-Nearest Neighbors (KNN),

Corresponding author: ykustiyahningsih@trunojoyo.ac.id

TECHNIUM
Technium: Romanian Journal of Applied Sciences and Technology
Vol. 17, 440-445, November, 2023
ISSN: 2668-778X
www.techniumscience.com

and Decision Tree C4.5[5]. This study will use the naïve Bayes classifier algorithm model as a sentiment classification algorithm on social media tweets, Twitter. The Naive Bayes algorithm includes a simple data classification algorithm that is used to calculate the probability of each data against an existing data set to determine its classes [5]. Data that will be used for the classification process must be cleaned first to improve the data structure. The classification process for text data must go through a feature extraction process to determine the features or attributes to be used and the weight of each feature. The feature extraction and weighting method used in this study is the tf-idf method. The process of extraction and feature weighting with TF-IDF is carried out to determine the features produced from each sentence and the weight of each word to sentences or documents in a document (data) set [6]. Not all of the resulting features from the TF-IDF process will be used, so feature selection is carried out to determine the features that are considered the most important. In this study, the information gain method is used for the feature selection process. The feature selection process with Information Gain can determine the features that are considered most relevant for use in the classification process with the naïve Bayes algorithm. The Naïve Bayes algorithm has several advantages compared to other data classification algorithms, including the Naïve Bayes algorithm which is suiTable for classifying large amounts of data and produces a fairly high level of classification accuracy [3]. Therefore, this study uses the Naïve Bayes algorithm as a data classification method from Twitter, which is quite large in number. The results of the accuracy of the naïve Bayes classification model in previous studies reached 84% [7], and 75.42% [8]. This sentiment analysis research is expected to show the results of the accuracy of the classification model used and provide useful information to the public and the government regarding the opinions of Twitter users regarding salt import policies.

2. RESEARCH METHOD

2.1. Sentiment Analysis

Sentiment analysis is an analytical technique of emotions and opinions from social media, websites or documents studied since 2000s [9]. Sentiment analysis can also be defined as a process of processing data and classifying the data into several categories. In general, sentiment analysis is divided into 3 levels [10], namely:

1. Document level : Document level analysis is commonly known as document level classification. This level will classify documents in general showing a positive or negative sentiment.
2. Sentence level : This level analysis determines every sentence in a document including a positive, negative, or neutral sentiment. Sentiment analysis at the sentence level is similar to the document level, that is, it cannot analyze directly what people actually like or dislike
3. Aspect level : Aspect level analysis determines a sentiment more specifically, because it is focused on aspects that are assessed in sentences or documents.

2.2. Crawling Data

Crawling Data is the process of automatically retrieving data from the website according to the needs of the user [11,12,13,14,15]. Data can come from various websites according to the needs of data seekers, including from the Twitter website. The process of crawling data can use several assistive tools including programming languages in order to obtain data according to the needs of data seekers.

2.3. TF-IDF Extraction and Weighting

Feature extraction and weighting is a process to determine the features obtained from the words in each sentence and to give weight to each word. One method of feature extraction and weighting is TF-IDF (Term Frequency – Inverse Document Frequency). The concept of the TF-IDF method is to extract sentences to determine features and assign weights to each feature by finding the value of a relationship in a set of documents [11].

There are 3 calculation steps performed in the TF-IDF feature extraction and weighting process, namely :

1. Term Frequency (TF) is used to calculate the weight of a term/word in a sentence. The TF value of each word will be greater if the word appears more and more in a sentence [12]. The equation for calculating the TF value is as follows [13].

$$tf = \begin{cases} 1 + log10(f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases} \quad (1)$$

With:
tf = term frequency value
f_(t,d) = term frequency (t) in document (d)

2. Inverse document frequency (IDF) is used to reduce the weight of each word that appears frequently or occurs a lot in several sentences so that the weight will be lower[12]. The equation for calculating tf-idf is as follows [13] :

$$idf_t = log\left(\frac{D}{df_t}\right) + 1 \quad (2)$$

With :
D = total number of documents
idft = document containing the term t
idf = inverse document frequency

3. Information Gain feature selection
Information gain is a feature selection method to measure the importance of each feature or attribute in determining the class of a data [14]. The concept of the feature selection process is to select the features that are considered the most important and affect the determination of the

TECHNIUM
Technium: Romanian Journal of Applied Sciences and Technology
Vol. 17, 440-445, November, 2023
ISSN: 2668-778X
www.techniumscience.com

class of a data and remove features that are considered less important. The application of feature selection aims to speed up the system or model being built and improve accuracy [15]. The steps to calculate the information gain are as follows:

4. Calculating entropy values or class uncertainty measures

$$Entropy(S) = \sum_i^c -P_i log_2 P_i \quad 3$$

with:
c = the value in the classification class
Pi = number of samples for class i

5. Calculating information gain:

$$Gain(S, A) = Entropy(S) - \sum_{Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad 4$$

with:
A = Attribute
v= possible values for attribute A
Values(A)= set of possible values for A
|Sv|= number of samples for the value of v
|S|= the total number of data samples
Entropy(S_v) = entropy for samples that have a value of v

6. Naïve Bayes Classifier

The naïve Bayes classifier algorithm is a data classification algorithm proposed by Thomas Bayes. The concept of the Naïve Bayes classifier is to calculate the probability of a particular class of data in the future by referring to previous data so that the data can be classified [16]. Learning data that is used as a reference for classification is often referred to as training data. All training data have their respective classes which have been determined manually as a learning for the classification model created. Classifying data using the naïve Bayes algorithm is done by calculating the probability of each feature or attribute to certain classes so that the probability value can determine the class of each data [17].

The basic formula for the naïve Bayes algorithm used is the following [18] :

$$P(C|X) = \frac{p(x|c).p(C)}{p(x)} \quad 5$$

with :
x = Data whose class is unknown
c = class that is used as a hypothesis
P(C|X) = Probability of a sample to enter in class C (Posterior)
P(C) = Probability of a certain class (prior)
P(X|C) = Chance of appearance of features/characteristics in class C (likelihood)
P(X) = Probability of X

The steps in the naïve Bayes algorithm can be done as follows:

Calculating the likelihood or conditional probability

P(x|C)=P (x1,x2…xn|C) (7)

C = Class
x = vector of attribute values n
P(x|C) = probability document from class C that contains attribute value x

3. RESULTS AND DISCUSSION

The results of the stages carried out in this research are as follows:

1. Crawling data. The results of crawling data obtained 500 data taken in 2021 with the keywords used "Indonesian salt policy". The following data labelling can be seen in Table 2. Crawling Sentence Result

Table 1. Crawling Sentence Result

Example of sentences
In this case, @kkpgoid is among the most responsible. I suspect, is it possible that there is a salt mafia that benefits from this salt import activity? Just like the export of lobster seeds, which turned out to be bribery committed by Edhi Prabowo Cs.

2. Manual labelling. The result of manual labelling is crawling data whose class has been determined manually. The data that has been labelled used as learning data for classification model. The following data labelling can be seen in Table 2. Labelling Data.

Table 2. Labelling Data

Example of sentences	Labelling
In this case, @kkpgoid is among the most responsible. I suspect, is it possible that there is a salt mafia that benefits from this salt import activity? Just like the export of lobster seeds, which turned out to be bribery committed by Edhi Prabowo Cs.	Positif

3. Data preprocessing is a step taken to clean and improve the data structure. In the data preprocessing stage, there are 6 stages that are carried out, starting from cleansing, tokenizing, case folding, slang words conversion, stemming, stop words. The results of the preprocessing process are in the form of clean data and improved data structures. The preprocessing process will also affect the accuracy of the classification model to be built. The following is an example of data after carrying out several stages in data preprocessing.

Cleansing Data: The cleansing stage produces data that has removed URL, numbers, hashtags, and usernames. The following is an example of data after the cleaning process. The following data cleansing can be seen in Table 3.

Table 3. Data after cleansing

Example of sentences
In this case, kkpgoid is among the most responsible. I suspect that there should be no salt mafia that benefits from this salt import activity. Just like the export of lobster seeds, it turns out that there was an act of bribery committed by Edhy Prabowo Cs.

Tokenizing: The Tokenizing stage produces data that is cut into words by removing punctuation without paying attention to the relationship of each

word in the sentence. The following data after tokenizing can be seen in Table 4.

Table 4. Data After Tokenizing

Example of sentences
In this case kkpgoid is among the most responsible. I suspect that there should be no salt mafia that benefits from this salt import activity. Just like the export of lobster seeds, it turned out that there was an act of bribery committed by Edhy Prabowo Cs.

Case Folding: The Case folding stage produces data whose letters will change to all lowercase. The following an example of data after case folding process can be seen in Table 5.

Table 5. Data After Case Folding

Example of sentences
In this case kkpgoid is among the most responsible. I suspect that there should be no salt mafia that benefits from this salt import activity, just like the export of lobster seeds, which turned out to be an act of bribery committed by edhy prabowo cs

Convert Slang words

The slang words conversion stage will check each word. This stage will check whether the words are in the slang word dictionary or not. If a non-standard word is found according to the slang words dictionary, then it will be changed to a standard word according to the slang words dictionary. The following data after the slang word conversion process can be seen in Table 6

Table 6. Data after conversion of slang words

Example of sentences
In this case kkpgoid is among the most responsible. I suspect that there should be no salt mafia that benefits from this salt import activity, just like the export of lobster seeds, which turned out to be an act of bribery committed by edhy prabowo cs

Stemming

The Stemming stage will change all the words in the sentence into basic words by removing all the affixes. The stemming process uses assistance from the literary library with the Nazief and Adriani algorithms. The following Data after stemming process can be seen in Table 7.

Table 7. Data after stemming

Example of sentences
In this case, KKPGoid is the most responsible. I suspect that there should be no salt mafia that can profit from this salt import activity, the same as with the export of lobster seeds, which are clearly available for bribes that are being carried out by Edhy Prabowo cs

Stop words

The Stop words stage is last preprocessing stage. This stage will remove words that are considered unimportant or have no meaning in a sentence. The stop words process uses the help of previously created stop words data. The following data after the stop words process can be seen in Table 8.

Table 8. Data after stop words process

Example of sentences
['story', 'mafia', 'salt', 'for', 'activity', 'import', 'salt', 'export', 'seed', 'lobster', 'real', 'kickback', ' bribe', 'sell', 'prabowo']

4. Feature extraction. The feature extraction process using the TF-IDF method produces 500 features with each feature having its own weight. The following determining weight of each word using the TF-IDF method can be seen in Table 9.

Table 9. Determining value of IDF

Word	TF					IDF
	D1	D2	D3	Df	D/Df	
Impor	2	0	1	2	1,5	0,176
Salt	1	1	0	2	1,5	0,176
Mafia	0	1	1	2	1,5	0,176
Bride	1	1	0	2	1,5	0,176
No	0	1	0	1	3	0,477
Good	2	0	1	2	1,5	0,176

5. Feature selection. The feature selection process will produce features that are considered the most important and discard other features that are considered unimportant with the aim of increasing the accuracy of the classification model. This study uses the information gain method for the feature selection method and succeeds in reducing the number of features to 500 features. Then the classification model will be tested through the model testing phase using the k-fold cross validation method with a value of k = 5. The results of the model test will show the level of accuracy, precision, and recall of each test (Fold) and the average level. Following are the results of model testing conducted using 5-fold cross validation.
6. Classification process and model testing. The classification process was carried out by dividing the 500 data into training data and test data. The result of the calcification process is a new label for each data that is carried out by the classification model. The new label column shows labelling results provided by the naïve Bayes classification model.
7. Result of Testing
   Testing the accuracy of model is done using 500 data. The k-fold cross validation method is used to test accuracy of model with a fold value = 5. The 500 data will be divided into 2, namely training data and test data. The distribution of data can be seen in Table 1. K-Fold cross validation data, based on predetermined folds.]

Table. K-Fold data cross validation

Trial	Dataset				
K=1	**Testing**	Training	Training	Training	Training
K=2	Training	**Testing**	Training	Training	Training
K=3	Training	Training	**Testing**	Training	Training

K=4	Training	Training	Training	**Testing**	Training
K=5	Training	Training	Training	Training	**Testing**

TABLE 13. Results of model testing without feature selection

Nilai K	Akurasi	Precision	Recall
1	67.0%	62.0%	81.0%
2	61.0%	55.01%	87.0%
3	74.0%	66.0%	94.0%
4	71.0%	62.0%	98.0%
5	49.0%	47.0%	66.0%
Average	64.40%	58.40%	85.20%

TABLE 13. Results of model testing with feature selection

Nilai K	Akurasi	Precision	Recall
1	77.0%	72.5%	92.0%
2	75.0%	66.41%	85.0%
3	82.0%	81.0%	96.0%
4	84.0%	82.0%	99.0%
5	78.0%	76.0%	78.0%
Rata-rata	79.20%	75.58%	90.00%

Table 13, 14 shows the values for accuracy, precision, and recall from model testing conducted with 5-fold cross validation. The highest levels of accuracy, precision and recall are respectively 79.20%, 75.58% and 90.00% with feature selection. While the average level of accuracy produced is 64.4%, the average level of precision is 58.4%, and the average recall rate is 85.2%. Meanwhile, the resulting average accuracy rate is 64.4%, the average precision level is 58.4%, and the average recall rate is 85.2% for the model without feature selection.

4. CONCLUSION

The conclusion is that sentiment analysis using Naïve Bayes method and information gain can improve accuracy of data classification model. Crawling, data labeling, preprocessing, feature selection, classification and testing with k-fold cross validation are the steps in this research. Preprocessing consists of 6 stages, namely cleansing, tokenizing, case folding, conversion of slang words, stemming, and stop words. The model testing process was carried out 5 times with 500 data. The test results used naïve Bayes classification method without information gain feature selection having average accuracy 64.40%, while using naïve Bayes with information gain, the resulting accuracy is an average of 81.40%. There is an average increase of 17%.

REFERENCES

1. 1. Safrida, I. Afriani, and Fajri, "The Impact of Salt Imports on Domestic Salt Production and Prices in Indonesia," *J. Bisnis Tani*, vol. 7, no. 1, pp. 25–36, 2021.
2. A. S. Rusydiana, I. Firmansyah, and L. Marlina, "Sentiment Analysis Of Microtakaful Industry: Comparison Of Indonesia And Malaysia," vol. 06, no. 01, pp. 20–34, 2018.
3. A. P. Natasuwarna, "Sentiment Analysis Of The Decision Of Moving The Capital City Using Naive Bayes Classification," pp. 47–53, 2019.
4. J. Zhou, Y. U. E. Lu, H. Dai, S. Member, H. A. O. Wang, and H. Xiao, "Sentiment Analysis of Chinese Microblog Based on Stacked Bidirectional LSTM," *IEEE Access*, vol. 7, pp. 38856–38866, 2019, doi: 10.1109/ACCESS.2019.2905048.
5. M. F. A. Saputra, T. Widiyaningtyas, and A. P. Wibawa, "Illiteracy Classification Using K Means-Naïve Bayes Algorithm," *JOIV Int. J. Informatics Vis.*, vol. 2, no. 3, p. 153, 2018, doi: 10.30630/joiv.2.3.129.
6. K. A. Nugraha and D. Sebastian, "Formation of Indonesian Word Topic Datasets on Twitter Using TF-IDF & Cosine Similarity," vol. 4, pp. 376–386, 2018.
7. T. A. Lorosae, B. D. Prakoso, Saifudin, and Kusrini, "Sentiment Analysis Based on Public Opinion on Twitter Using Naive Bayes," pp. 25–30, 2018.
8. E. M. Sipayung, H. Maharani, and I. Zefanya, "Designing a Customer Comment Sentiment Analysis System Using the Naive Bayes Classifier Method," vol. 8, no. 1, pp. 958–965, 2016.
9. F. Ali *et al.*, "Transportation sentiment analysis using word embedding and ontology-based topic modeling," *Knowledge-Based Syst.*, vol. 174, no. xxxx, pp. 27–42, 2019, doi: 10.1016/j.knosys.2019.02.033.
10. R. V. P. Siwabessy, A. Herdiani, and A. Romadhony, "Analysis of Public Sentiment of Incumbents' Work Results in Relation to the 2019 Presidential Election on Social Media Twitter Using Support Vector Machine (SVM)," vol. 6, no. 2, pp. 8625–8636, 2019.
11. A. Deolika, Kusrini, and E. T. Luthfi, "Word Weighting Analysis on Text Mining Classification," *J. Teknol. Inf.*, vol. 3, no. 2, p. 179, 2019, doi: 10.36294/jurti.v3i2.1077.
12. B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sentiment Analysis System for Product Reviews Using the Naive Bayes Method," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.
13. Informatikalogi., 2016. Word Weighting or Term Weighting TF-IDF, URL: https://informatikalogi.com/term-weighting-tf-idf/.
14. A. B. P. Negara, H. Muhardi, and I. M. Putri, "Airline Sentiment Analysis Using the Naive Bayes Method and Feature Selection Information Gain," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, p. 599, 2020, doi: 10.25126/jtiik.2020711947.
15. S. H. A. Aini, Y. A. Sari, and A. Arwan, "SSelection of Information Gain Features for Classification of Heart Disease Using a Combination of K-Nearest Neighbor and Naïve Bayes Methods," *J. Pengemb. Teknol. Inf. dan*

Ilmu Komput., vol. 2, no. 9, pp. 2546–2554, 2018.

16. N. Wijaya, "Application of the Naive Bayes Classification Algorithm to Data on Occupied Status of Housing Assistance for Post-Eruption of Mount Merapi Rehabilitation and Reconstruction Fund 2010," pp. 1–10, 2019.
17. L. A. Andika, P. A. N. Azizah, and Respatiwulan, "Analysis of Public Sentiment on Quick Count Results of the 2019 Indonesian Presidential Election on Social Media Twitter Using the Naive Bayes Classifier Method," vol. 2, no. 1, pp. 34–41, 2019.
18. A. P. Wijaya and D. Wardhani, "Sentiment Analysis and Classification of Positive Comments on Twitter with Naïve Bayes Classification Sentiment Analysis and Classification of Positive Comments on Twitter with Naive Bayes Classification," vol. 1, no. 2, 2020.