Detecting Hate Speech Tweets And Abusive Tweets In Indonesian Language Using Random Forest And Support Vector Machine With Voting Classifier Technique

Dandi Indra Wijaya [1*], Riza Arifudin [1]

[1]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia
*Corresponding author: dandiindra29@gmail.com

ARTICLE INFO

Article history
Received: 4 Maret 2022
Revised: 15 Maret 2022
Accepted: 1 April 2022

Keywords
Sentiment Analysis
Support Vector Machine
Random Forest
Voting Classifier
TF-IDF
N-gram

ABSTRACT

The use of social media has become one of the main things in everyday life. This happens because the features provided make it easy for people to communicate and disseminate information. One of the social media used by many people is Twitter. the main feature of twitter is that its users can post posts that are termed tweets. There is a negative thing about the freedom to write a tweet, namely a tweet that does not contain things that harm other people or community. The problem that arises from this negative thing is to distinguish between hate speech tweets and abusive tweets. Hate speech and abusive speech are often the same thing. These differences need to be considered because they can have a negative impact on social life. Sentiment analysis is used to distinguish the two things. Sentiment analysis is an implementation of natural language processing which is part of machine learning. The algorithms used in this research are Support Vector Machine, Random Forest, and Voting Classifier with soft voting type. The estimator for the Voting Classifier is the Support Vector Machine and Random Forest. TF-IDF and N-gram were used as feature extraction. The data used is a tweet dataset that has been labeled neutral, hate speech, and rude speech. Measurement of model accuracy is done by using confusion matrix. The highest accuracy was produced by a combination of Voting Classifier technique with TF-IDF feature extraction and the amount of N-gram was 1 gram, which was 82.57% accuracy.

1 Introduction

The use of social media at this time is very well established. In 2015 the number of users reached around two billion (Kim & Jang, 2019). The increasingly sophisticated technology in the use of the internet and communication devices such as smartphones has encouraged the increasing number of sites that are useful for sending and receiving quickly. This makes it easier for people to communicate with each other regardless of space and time. Communication is a process to convey feelings or thoughts from one person to another using various meaningful symbols and using certain media. Twitter is a widely used social media. One of the problems with Twitter is tweets that contain hate (Fatahillah et al., 2018).

According to Widayati (2018), hate speech is different from speech in general. The difference can be seen from the intention of an utterance. The intention of hate speech is to cause a certain impact on a person or group that is addressed directly or indirectly. Hate Speech is one of the problems on social media platforms. Therefore, even if a sentence has a harsh tone, it is not necessarily an utterance of hatred or can be called just a rude speech. These differences need to be

fast, it is impossible to manually reduce hate speech content (Mishra et al., 2021). Thus, there is a need for a system that can detect hate speech tweets automatically. The system is also useful for distinguishing hate speech from abusive speech. Making the system can be done using sentiment analysis.

Social media sentiment analysis which aims to extract people's opinions, attitudes and emotions from social networks has become a center of research (Li et al., 2019). According to Taboada (2016), sentiment analysis is a developing field in computer science that tries to automatically determine the sentiments contained in the text. The purpose of sentiment analysis is to analyze the emotions, judgments, sentiments, evaluations, and opinions of people directed at an entity such as individuals, topics, problems, organizations, and services (Liu, 2012). Another term for sentiment analysis is opinion mining. According to Hussein (2018) sentiment analysis is the implementation of natural language processing which is part of machine learning. Natural Language Processing or abbreviated as NLP aims to process and analyze natural language data using computer programming techniques.

Several studies on the detection of hate speech tweets in Indonesian have been carried out. For example, Ibrohim & Budi's research (2019) discusses text detection with various labels for abusive speech and hate speech detection including detecting the target, category, and level of hate speech on Twitter Indonesia using a machine learning approach using Random Forest Algorithms, Naive Bayes, Support Vector Machines, And Binary Relevance. The study used several feature extraction methods before training. The random forest algorithm with the NLP data transformation method when using the word unigram feature gives the best performance with an accuracy of 76.16%.

Fauzi & Yuniarti (2018) conducted research on hate speech in Indonesian with the same machine learning algorithm but using a different feature extraction method. His research used the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF makes it possible to give us a way to associate each word in a document with a number that represents how relevant each word in the document is. From the two studies, there are differences in the feature extraction stage. The feature extraction stage on the performance of the classification algorithm is very influential. In the research conducted by Fauzi & Yuniarti (2018), they did not use the N-gram method at the feature extraction stage which can improve performance as was done by Ibrohim & Budi (2019). Based on the research of Laoh et al. (2019) the use of the N-gram method in dataset feature extraction can increase accuracy in many literatures on sentiment analysis by an average of 94%.

Research conducted by Fauzi & Yuniarti (2018) also uses the ensemble method, namely voting. Based on the results of the study, the use of ensemble techniques using soft voting on the three best classifiers, namely Random Forest, Support Vector Machine, and Naive Bayes succeeded in improving classification performance. Classifier voting is one of the ensemble approaches where we can combine several models for better classification (Kumar et al., 2017). The way the voting classifier works is by choosing one of many alternatives, based on the class that is predicted with the most votes (Zhang et al., 2014).

Based on the description above, the learning algorithms that are commonly used and have good abilities in sentiment analysis are random forests and support vector machines. Random forest is an ensemble classifier that generates many decision tree algorithms, using a randomly selected subset of samples and training variables (Belgiu & Drăgu, 2016). Meanwhile, the support vector machine or commonly abbreviated as SVM is a method that makes statistical learning theory the basis for machine learning (Ding et al., 2017). The main idea of SVM is to separate different classes using a hyperplane (Tharwat et al., 2017). The research focuses on the preprocessing stage using TF-IDF and N-grams and the use of ensemble techniques, namely voting on the random forest algorithm and support vector machine.

2 Method

In this research, the classification of hate speech tweets and abusive tweets was carried out using the random forest algorithm, support vector machine, and a combination of the two using the voting classifier technique with TF-IDF and N-gram as feature extraction. The flowchart of the

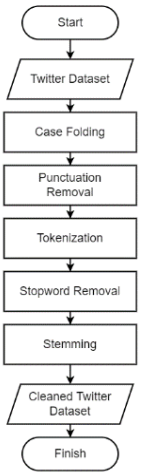

Figure 1. Preprocessing Flowchart

The flowchart of the method used in this study can be seen in Figure 2.

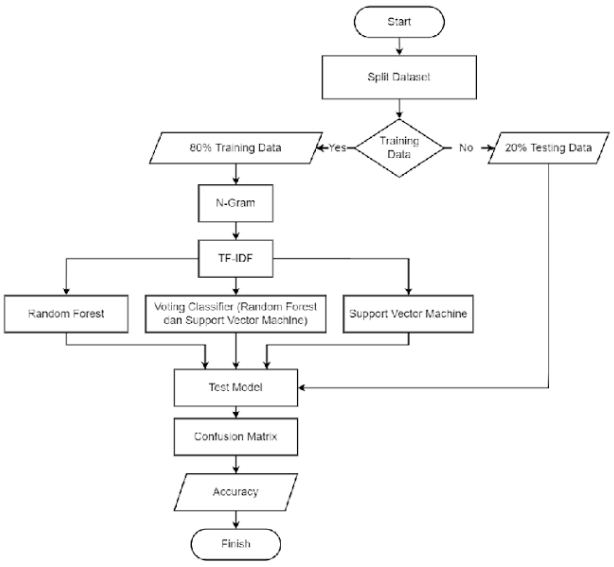

Figure 2. Flowchart of The Proposed Method

This research begins by preparing the dataset. The dataset in this study is the Indonesian language Twitter Hate Speech dataset labeled hate and abusive speech taken from Kaggle.com. The dataset obtained is then carried out in the preprocessing stage. Data preprocessing is divided into 5 stages, namely case folding, punctuation removal, tokenization, stopword removal, and stemming. At the case folding stage, all characters are converted to lowercase (lowercase). The punctuation removal stage removes all punctuation marks contained in the data. Tokenization stage each tweet data in the form of sentences or paragraphs is broken down into tokens. At the stopword removal stage, words or tokens that are considered not very important are removed. In the stemming stage, words that have affixes are changed to basic words.

After preprocessing, the next step is to perform feature extraction with TF-IDF and N-grams (with n=1, n=2, and n=3). The data set with the selected features at the feature extraction stage

consists of 80% training data and 20% test data. The training data is used for training the random forest algorithm model, SVM, and voting classifier technique.

2.1 Data Collection

This research uses materials in the form of taking the 2019 Indonesian hate speech dataset from *kaggle.com/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text* with a total record of 13,169 and consists of 12 labels marked with labels 1 and 0. 1 means the tweet is included in the label, while if 0 means the tweet is not included in that label.

2.2 Support Vector Machine

Support vector machine (SVM) is a machine learning method based on statistical learning theory. SVM is a supervised learning algorithm that is trained to classify various categories of data from various disciplines. This algorithm can improve generalization performance by handling nonlinear classification by mapping inputs into high-dimensional areas and solving quadratic programming optimization problems (Polat et al., 2017). The main idea of a SVM is to separate different classes using a hyperplane.

2.3 Random Forest

The random forest machine learning method is used to develop predictive models in many research settings (Speiser et al., 2019). This algorithm is specially designed to analyze very high-dimensional data with many classes representing data, the well-known is the text corpus (Xu et al., 2012). Therefore, random forest is one of the algorithms that can be used to analyze sentiment. Random forest is an ensemble classifier that generates many decision tree algorithms, using a randomly selected subset of samples and training variables (Belgiu & Drăgu, 2016). Random processes are carried out so that each decision tree algorithm does not affect each other (Del Río et al., 2014).

2.4 Voting Classifier

Classifier voting is one of the ensemble approaches where we can combine several models for a better classification (Kumar et al., 2017) so as to balance the weaknesses of individual classifiers in certain datasets. Voting classifier is a meta classifier to bring together similar or unprecedented machine learning classifiers for classification and detection. A simple majority vote is a decision rule that selects one of many alternatives, based on the predicted class with the most votes (Zhang et al., 2014). The ensemble voting classifier performs both "hard" and "soft" voting (Ruta & Gabrys, 2001).

- Soft voting classifies the input data based on the probabilities of all predictions made by the different classifiers. The weights assigned to each classifier are applied appropriately. For example there are 3 binary classifiers clf_1, clf_2 and clf_3. For a given record, the classifier makes predictions with the class [0,1]. For example, the probability prediction results for each class in each classifier are clf_1 -> [0.3, 0.7], clf_2 -> [0.2, 0.8], clf_3 -> [0.8, 0.2]. The probability of each class with the same weight will be calculated as follows.
- An example of a hard voting with a majority of votes based on the same weight is if there are three classifiers clf_1, clf_2, clf_3 and the classification results of each classifier are [1, 1, 0] then the voting result is 1. Example of hard voting with a majority vote based on the same weight The difference is that if there are three classifiers clf_1, clf_2, clf_3 with each weight being [0.1, 0.3, 0.6] and the prediction result [0, 0, 1] then the prediction results are averaged to [0, 0, 0, 0, 1 , 1, 1, 1, 1, 1]. Thus the voting result is 1.

$$\text{Class } 0 = 0.33*0.3 + 0.33*0.2 + 0.33*0.8 = 0.429$$

With the above results, the voting result for the class [0, 1] with 3 classifiers is 1. The majority vote does not require setting any parameters after the individual classifier has been trained (Kuncheva & Rodríguez, 2014).

2.5 TF-IDF (Term Frequency – Inverse Document Frequency)

TF-IDF examines the relevance of keywords to documents in the corpus (Qaiser & Ali, 2018). The TF-IDF algorithm can be used to solve the problem of calculating feature weights for automatic text classification (Fan & Qin, 2018). The concept of TF-IDF is to calculate the frequency of occurrence of a word in a document and the reverse frequency of the document containing the word. The more a word appears in a document, the more influence it has on the document. On the other hand, the less a word appears in a document, the less impact it has on the document. TF-IDF can be seen in Equation below.

$$W_{i,j} = tf_{i,j} \text{ x } log\frac{N}{df_i} \quad (1)$$

$tf_{i,j}$ = The number of i-words in the j-th document
N = Total documents
df_i = The number of documents containing the word i

2.6 N-gram

The N-gram method is a popular feature identification and analysis approach used in language modeling and the field of natural language processing or known as natural language processing (Ahmed et al., 2017). N-gram is a method for checking continuous 'n' words or sounds of a given sequence of text or speech (Tripathy et al., 2016). N-grams are useful for predicting the next item in sequence. To analyze the sentiment of a text or a document, the N-gram is used in sentiment analysis.

3 Result and Discussion

3.1 Results

The initial stage in this research is to preprocess the dataset. The preprocessing stages in this study are case folding, punctuation removal, tokenization, stopword removal, and stemming. One of preprocessing methods result with the tweet

"USER USER Kitab suciku yg diturunkan oleh Allah dianggsp fiksi. Pake sok2an harus dilihat dr segi filaafat. Eh tapi kalo misalnya Quranku fiksi, maka tidak bisa dong Quran menjadi dasar hukum di dalam Hukum Islam. Jadi harus pk apa? Ada"

can be seen in Table 1.

Table 1. Preprocessing Method Result

Tweet	Method
user user kitab suciku yg diturunkan oleh allah dianggsp fiksi. pake sok2an harus dilihat dr segi filaafat. eh tapi kalo misalnya quranku fiksi, maka tidak bisa dong quran menjadi dasar hukum di dalam hukum islam. jadi harus pk apa? ada	Case Folding
user user kitab suciku yg diturunkan oleh allah dianggsp fiksi pake sok2an harus dilihat dr segi filaafat eh tapi kalo misalnya quranku fiksi maka tidak bisa dong quran menjadi	Punctuation Removal

Tweet	Method
dasar hukum di dalam hukum islam jadi harus pk apa ada	
['user', 'user', 'kitab', 'suciku', 'yg', 'diturunkan' oleh', 'allah', 'dianggsp', 'fiksi', 'pake', 'sok2an' harus', 'dilihat', 'dr', 'segi', 'filaafat', 'eh', 'tapi' kalo', 'misalnya', 'quranku', 'fiksi', 'maka', 'tidak' bisa', 'dong', 'quran', 'menjadi', 'dasar', 'hukum' di', 'dalam', 'hukum', 'islam', 'jadi', 'harus', 'pk' apa', 'ada']	Tokenization
['user', 'user', 'kitab', 'suciku', 'yg', 'diturunkan' allah', 'dianggsp', 'fiksi', 'pake', 'sok2an', 'dr' segi', 'filaafat', 'eh', 'kalo', 'quranku', 'fiksi' quran', 'dasar', 'hukum', 'hukum', 'islam', 'pk']	Stopword Removal
user user kitab suci yg turun allah dianggsp fiks pake sok2an dr segi filaafat eh kalo quran fiks quran dasar hukum hukum islam pk	Stemming

Stemming method in this research is using Sastrawi. Sastrawi is a simple python library that functions to reduce inflectional words in Indonesian to their basic form. The next step is labeling. In the dataset, each tweet is labeled with the numbers 0 and 1. The number 0 indicates the tweet is not included in the label, while the number 1 indicates the tweet is part of the label. Tweets with labels can be seen in Table 2.

Table 2. Tweets with Labels

Tweet	Hate Speech	Abusive
sinden banci kocak princes aprilia amp mimin onggo inggi url user	0	1
user user kitab suci yg turun allah dianggsp fiksi pake sok2an dr segi filaafat eh kalo quran fiksi quran dasar hukum hukum islam pk	0	0
mobile langend babi anjeng xnk kasi rank kimakkk team noob babi bocahh setan	1	1
user user jahahaha cebong kapir alay goblok munafik jijik banget dah iyuuhhh	1	1

The table above were changed to three categories for research purposes, namely neutral tweets with a number label 0, abusive tweets labeled with number 1, and hateful speech tweets with number 2 labels. and label the number 0 in the hate speech category. Meanwhile, hate speech tweets are tweets with a label 1 in the hate speech category and label 0 in the abusive speech category or tweets with a label 1 in the hate speech category and label 1 in the abusive speech category. The tweet with the updated labels can be seen in Table 3.

Table 3. Tweets with Updated Labels

Stemming	Label
sinden banci kocak princes aprilia amp mimin onggo inggi url user	1
user user kitab suci yg turun allah dianggsp fiksi pake sok2an dr segi filaafat eh kalo quran fiksi quran dasar hukum hukum islam pk	0
mobile langend babi anjeng xnk kasi rank kimakkk team noob babi bocahh setan	2

user user jahahaha cebong kapir alay goblok munafik jijik banget dah iyuuhhh	2

The dataset is used in the training model process using random forest, support vector machine, voting classifier with random forest estimator and support vector machine with TF-IDF and N-Gram as feature extraction. Confusion Matrix is used to measure the accuracy of the model. Measurement with the Confusion Matrix is important to evaluate the accuracy of a model. The results of the accuracy of the existing models can be seen in Table 4.

Table 4. Models Accuracy

Feature Extraction	Gram	Classifier	Accuracy
TF-IDF	1	Random Forest	80.75%
TF-IDF	2	Random Forest	80.41%
TF-IDF	3	Random Forest	79.91%
TF-IDF	1	Support Vector Machine	81.39%
TF-IDF	2	Support Vector Machine	80.10%
TF-IDF	3	Support Vector Machine	78.54%
TF-IDF	1	Voting Classifier	82.57%
TF-IDF	2	Voting Classifier	80.86%
TF-IDF	3	Voting Classifier	79.46%

Table 4. shows the highest accuracy, which is 82.57%, obtained in the soft voting classifier algorithm with support vector machine and random forest estimators and the use of a combination of TF-IDF and N-gram feature extraction as much as 1 gram. The second highest level of accuracy is the support vector machine algorithm, which is 81.39% with 1 gram of N-gram. Random forest has a lower level of accuracy compared to other algorithms, namely 80.75% with 1 gram of N-gram. The use of N-grams combined with TF-IDF in this study shows that the higher the gram used, the lower the accuracy obtained for each algorithm.

3.2 Discussion

Based on the accuracy results in this study, the application of TF-IDF and N-gram feature extraction on the support vector machine and random forest algorithms was able to test sentiment analysis on the Indonesian hate speech tweet dataset well. The method used in this study is better than the existing methods in related research. Comparison with previous studies using the same dataset and algorithm can be seen in Table 5.

Table 5. Comparison of Results with Previous Research

Writer	Method	Accuracy
Ibrohim & Budi (2019)	Label power set + Unigram + Random forest	76.16%
Proposed Method	TF-IDF + Unigram + Voting classifier dengan estimator Support vector machine dan Random forest	82.57%
Proposed Method	TF-IDF + Unigram + Random forest	80.76%

In Table 5 the combination of the use of the voting classifier soft voting algorithm with the support vector machine estimator and the random forest feature extraction of TF-IDF and N-grams amounting to 1 gram resulted in the highest accuracy of 82.57%. The research conducted by Ibrohim & Budi (2019) did not use TF-IDF feature extraction and only used N-grams in several algorithms that produced the highest accuracy of 76.16%, namely the use of the random forest algorithm with 1 gram of N-grams. When compared with this study, the use of the random forest algorithm produces higher accuracy because it uses TF-IDF feature extraction. Based on the results of this study, when compared with previous studies, the use of TF-IDF and N-gram feature extraction has a good effect on increasing accuracy in the algorithm used. In addition, the use of the voting classifier algorithm can produce higher accuracy

4 Conclusion

The study was conducted to test the random forest algorithm, support vector machine, and voting classifier with TF-IDF and N-gram feature extraction in conducting sentiment analysis on the twitter dataset of hate speech and rude speech in Indonesian. The way the voting classifier algorithm works in this research uses support vector machine and random forest as an estimator. Voting classifier performs both "hard" and "soft" voting. The author uses soft voting during the training process because it takes into account more information than hard voting. Soft voting uses the uncertainty of each classifier in the final decision. The results of the highest accuracy in this research is found in the use of the voting classifier algorithm with a random forest estimator and the SVM for TF-IDF and N-gram as feature extraction, amounting to 1 gram, which is 82.57%. This result is higher than the method used in related research (random forest and unigram method) which is 76.16%.

References

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. *ISDDC*, 10618, 127–138.

Azar, A. T., Elshazly, H. I., Hassanien, A. E., & Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer methods and programs in biomedicine*, *113*(2), 465-473.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24-31.

Ding, S., Zhu, Z., & Zhang, X. (2017). An overview on semi-supervised support vector machine. *Neural Computing and Applications*, *28*(5), 969-978.

el Río, S., López, V., Benítez, J. M., & Herrera, F. (2014). On the use of MapReduce for imbalanced big data using Random forest. *Information Sciences*, 285(1), 112–137.

Fan, H., & Qin, Y. (2018). Research on text classification based on improved tf-idf algorithm. *International Conference on Network, Communication, Computer Engineering (NCCE),147,* 501-506.

Fatahillah, N. R., Suryati, P., & Haryawan, C. (2017). Implementation of naive bayes classifier algorithm on social media (Twitter) to the teaching of indonesian hate speech. *International Conference on Sustainable Information Engineering and Technology (SIET)*, 128-131.

Fauzi, M. A., & Yuniarti, A. (2018). Ensemble method for indonesian twitter hate speech detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *11(1*), 294-299.

Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University Engineering Sciences, 30(4)*, 330-338.

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian twitter. *Proceedings of the Third Workshop on Abusive Language Online,* 46-57.

Kim, D., & Jang, S. S. (2019). The psychological and motivational aspects of restaurant experience sharing behavior on social networking sites. *Service Business, 13(1),* 25-49.

Kumar, U. K., Nikhil, M. S., & Sumangali, K. (2017). Prediction of breast cancer using voting classifier technique. *IEEE international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM)*, 108-114.

Kuncheva, L. I., & Rodríguez, J. J. (2014). A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems, 38(2),* 259–275.

Laoh, E., Surjandari, I., & Prabaningtyas, N. I. (2019). Enhancing hospitality sentiment reviews analysis performance using SVM N-Grams method. *16th International Conference on Service Systems and Service Management (ICSSSM),* 1-5.

Li, Z, Fan, Y, Jiang, B, Lei, T & Liu. (2019). A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications, 78,* 6939–6967.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 5*, 1–167.

Mishra, S, Prasad, S & Mishra, S. (2021). Exploring multi-task multi-lingual learning of transformer models for hate speech and ofensive speech identifcation in social media. *SN Computer Science, 2(72), 1-19.*

Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *Journal of medical systems, 41(4)*, 55.

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181(1)*, 25-29.

Ruta, D., & Gabrys, B. (2001). Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems. Proceedings of the 4th International Symposium on Soft Computing.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications, (134)*, 93-101.

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, (*2)*. 325-347.

Tharwat, A., Hassanien, A. E., & Elnaghi, B. E. (2017). A BA-based algorithm for parameter optimization of support vector machine. *Pattern Recognition Letters (93)*, 13-22.

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117-126.

Widayati, L. S. (2018). Ujaran kebencian: Batasan pengertian dan larangannya. *Info Singkat: Kajian singkat terhadap isu aktual dan strategis, 5(6)*, 1-6.

Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. *Journal of Computers*, *7*(12), 2913-2920.

Zhang, Y., Zhang, H., Cai, J., & Yang, B. (2014). A weighted voting classifier based on differential evolution. *Abstract and Applied Analysis, 2014(2)*. 1-6.