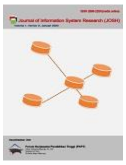


Journal of Information System Research (JOSH)
Volume 4, No. 3, April 2023, pp 914−919
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i3.3317

Hoax Detection on Indonesian Tweets using Naïve Bayes Classifier with TF-IDF

Ichwanul Muslim Karo Karo[1,*], Romia[2], Sri Dewi[1], Putri Maulidina Fadilah[3]

[1]Faculty of Mathematics and Natural Science, Computer Science, Medan State University, Medan
Jl. William Iskandar Ps. V, New Memories, Kec. Percut Sei Tuan, Deli Serdang Regency, North Sumatra, Indonesia
[2]Information System, STMIK Citra Mandiri Padangsidimpuan, Padangsidimpuan
Jl. West Sumatra Cross No. 341, Scales, Kec. North Padangsidimpuan, Padang Sidempuan City, North Sumatra, Indonesia
[3]Faculty of Mathematics and Natural Science, Statistics, Medan State University, Medan
Jl. William Iskandar Ps. V, New Memories, Kec. Percut Sei Tuan, Deli Serdang Regency, North Sumatra, Indonesia
Email: [1,*]ichwanul@unimed.ac.id, [2]Romialubis18@gmail.com, [3]sridewi@unimed.ac.id, [4]putrimaulidina@unimed.ac.id
Correspondence Author Email: ichwanul@unimed.ac.id

Submitted: **04/04/2023;** Accepted: **30/04/2023;** Published: **30/04/2023**

Abstract−Twitter is one of the most popular social media platforms in the world nowadays. Twitter users in Indonesia are the fifth largest in the world and are always active in expressing themselves and getting information through tweets. A hoax is a lie created as if it were true. Hoaxes are also often spread via tweets. The spread of hoaxes is extremely dangerous because it can cause social discord and even misunderstanding. Therefore, hoaxes must be resisted. This study aims to build a system to detect hoaxes on Indonesian tweets. The objective of this research is to identify hoax Indonesian tweets by using the Naïve Bayes classifier with Term Frequency Inverse Document Frequency (TF-IDF). This study collects and annotates tweets from hoax tweets post which sent by a user account. This study also applied several text preprocessing techniques to provide datasets. To provide the best hoax prediction model, this work splits datasets into training and testing datasets. There are four experimental scenarios that refer to splitting the dataset. The experimental results showed that the hoax prediction model using Naïve Bayes with TF-IDF had 64% accuracy and recall, 69% and 67% precision, and a F1-score respectively. This result is also superior to the hoax prediction model when using the Naïve Bayes classifier without the TF-IDF. It means that TF-IDF has made a positive contribution to improving model performance. Finally, this research contributes by detecting news with a proclivity for hoaxes and filtering what is classified as hoaxes or not.

Keywords: Hoax; Naïve Bayes; TF-IDF; Text Preprocessing; Performance

1. INTRODUCTION

Twitter is a social networking service that enables its users to post text, images and videos known as tweets[1]. A tweet is a short message of no more than 140 characters. The freedom of users to spread tweets is frequently used to spread fake news[2]. Hoaxes are pieces of information that are published without regard for their being factual or false[3], [4]. Hoaxes have general characteristics such as sources of information that are not credible or media (photos or videos) that are not original[3]. Hoaxes can arise from various topics such as social, economic, educational, and political, and of course this has a very negative impact on society [5]. Hoaxes have recently become a topic of discussion on Twitter because they are thought to be bothering the public with unreliable information[6].

The number of Twitter users in Indonesia is ranked 5th in the world, with a total of 24 million users [7]. In addition, there are 500 million tweets published every day and 350,000 tweets posted every minute[7]. Imagine if that number of users were exposed to hoax tweets. Negative impacts that may occur range from misunderstanding to hostility. So that the problem of hoax tweets is anticipated. In addition, the spread of hoaxes tweets is extremely dangerous because it can cause discord and even misunderstanding in social media user even society. Therefore, hoaxes must be resisted and important to anticipate its spread.

A machine learning approach has also been developed to detect, analyse, and study information hoaxes on Twitter as early as possible. There have been many hoax detection studies using a machine learning approach to solve it. Research by [8] proposes Naive Bayes to classify fake news in Indonesian Language. The study used its own dataset, which included 600 valid and bogus articles. They assigned the dataset to three reviewers for manual grading. Taking the highest score from the three reviewers, the final tagging was determined. The results show that using the term frequency feature of the PHP-ML library component, Naive Bayes can classify Indonesian online news articles. Static testing accuracy was 82.6%, and dynamic testing accuracy was 68.33%. Other research by [9] explains how the Nave Bayes classifier is applied to detect fake news. BuzzFeed News' dataset consists of information from Facebook posts, which is represented by a news article. This study shows that classifiers using the Naïve Bayes method can show good results on important issues such as hoax news classification, with a total classification accuracy of 75.40%. Another study analyzed the Naïve Bayes and k-Nearest Neighbor algorithms for classifying health hoax news on social media Twitter. Treatment used without involving feature extraction on their research[10]. As a result, the accuracy of the two algorithms is below 68%, even the accuracy of the Naïve Bayes algorithm is only 66%. One effort to improve the performance of machine learning models to detect hoaxes is to use the TF-IDF feature extraction[6], [11]. In the research [11], the use of TF-IDF feature extraction in the Bernoulli Naive Bayes algorithm can increase accuracy by 16.08%, precision 15.7%, recall 16.22%, and f1-score

15.92% when compared to the results of previous studies. Based on the short review above and literature [12], Naïve bayes algorithm is one of the most popular machine learning algorithms for detecting hoaxes.

This study aims to detect hoaxes using the Term Frequency – Inverse Document Frequency (TF-IDF) and the Naïve Bayes classifier. TF-IDF weighting is used to extract text on tweet and to measure how important a term is in a tweet. TF is the frequency of occurrence of a word in a document, while IDF is a measure of a word's ability to differentiate categories. The Naïve Bayes method was chosen based on previous research which produced good accuracy in classifying hoax Indonesian tweets.

2. RESEARCH METHODOLOGY

Figure 1 shows a flowchart of the steps that must be taken to support this research. Broadly speaking, the stages begin with the process of crawling Indonesian tweets, followed by text pre-processing and feature extraction. After that the dataset split into two: training and testing daset. These phases are called data preparation. Next phases are the modelling process using Naïve Bayes classifier, and its model becomes a hoax prediction model. The last process is evaluation.

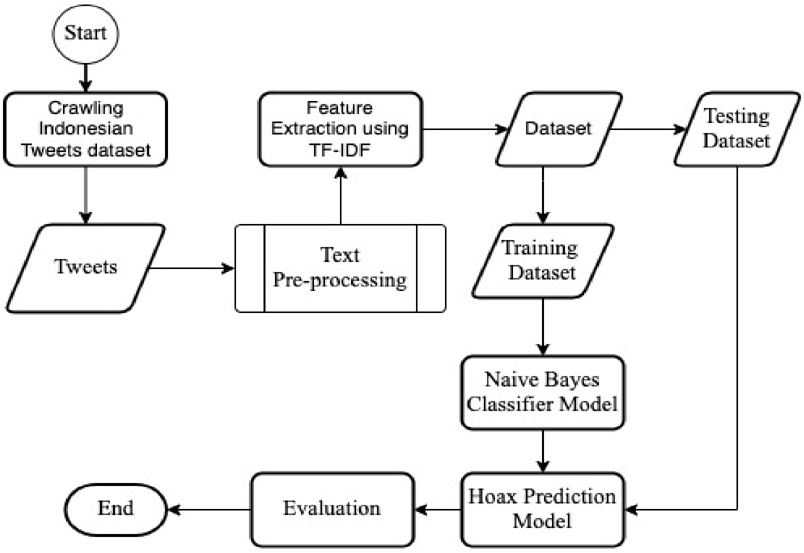


Figure 1. Flowchart Hoax Detection System

2.1 Crawling Dataset

The initial stage is crawling Tweets from Twitter. Crawling process uses the python programming language and application programming interface (API) from Twitter. There are several alternatives to get tweets in this study, by using keywords, hashtags, replies and mentions until thread tweets.

2.2 Dataset

The Indonesian tweets were crawled using Python programming with API. The tweets are manually labelled with a class. There are two classes: hoaxes and not hoaxes. This study successfully crawled and annotated 519 Indonesian tweets. hoax class as many as 306, non-hoax class 213 (shown in Figure 2). Fifty-nine percent of the dataset is hoax tweets, which is more than those who don't.

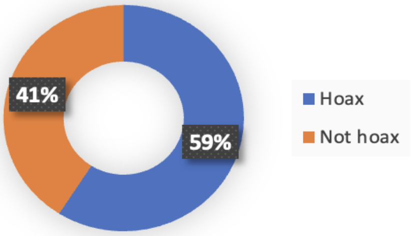


Figure 2. Percentage of Classes

2.3 Text Pre-processing

Text preprocessing plays a crucial role in hoax detection research [13]. Text preprocessing is the process of converting unstructured data into structured data using natural language processing (NLP)[14], [15]. This study applied five text pre-processing techniques: Case folding, tokenizing, Data cleansing, stemming and. stop word removal.

a. **Case folding** is converting all of the characters in a tweet into a single case, either all upper case or all lower case. The process also speeds up comparisons during the indexing process
b. **Tokenization** is separating process a piece of text into term. It's a fundamental step on NLP
c. **Data cleansing** is the process of cleaning data from noise. This study applied data cleansing to remove some symbols and emoji.
d. **Stemming** is a natural language processing technique that lowers inflection in words to their root forms, hence aiding in the preprocessing of text, words, and documents for text normalization. This study used sastrawi stemmer [16].
e. **Stopwods removal** is removing process the words that occur commonly across all the documents in the corpus. It could be conjunction or another term like '*the*'.

2.4 Feature Extraction using TF-IDF

Term Frequency - Invers Document Frequency (TF-IDF) is an empirical metric that indicates how important a term is to a document in a collection of documents. Term Frequency (TF) and Inverse Document Frequency (IDF) are combined in TF-IDF (IDF) [17]. Term Frequency (TF) measures the frequency with which words appear in a document. Because the length of each document varies, the TF value is usually divided by the length of the document (the total number of words in the document). The Inverse Document Frequency (IDF) is a value used to determine how important a term is. The IDF will evaluate terms based on how they appear throughout the document. The lower the IDF value, the less significant the word, and vice versa. Mathematically, the TF-IDF value for the term *t* in the document *d* from the document set *D* is calculated using equation (1).

$$tf\ idf(t,d,D) = tf(t,d).idf\ (t,D) \tag{1}$$

2.5 Naïve Bayes Classifier

Naive Bayes algorithm is an algorithm used to find the highest probability value to classify test data in the most appropriate category[18]. The Naïve Bayes classifier exhibits high accuracy and speed when applied to large databases. The Naïve Bayes algorithm is based on the Bayes theorem discovered by Thomas Bayes in the 18th century. The Bayes theorem occurs in equation (2).

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2}$$

P (X | H) or called likelihood, is the probability of hypothesis *X* based on the condition of *H*. Meanwhile *P (H)* and *P (X)* are the probability of texting data for which class is unknown and the probability of data *X* which is a specific class. In text classification cases, Naïve Bayes theorem adjust to several conditions. Let, *j* is text categories (*Cj*), each *X* contains word (w_i) and assume that each word in category is independent, so the Naïve Bayes calculation can be simplified further as equation (3)

$$P(X|C_j) = \prod_{i=1}^{n} P(w_i|C_j) \tag{3}$$

Meanwhile, to determine the class of a tweet, it is obtained from the maximum result value $P(w_i|C_j)$ (probaliblitas word *i* on class *j*) and probalibilitas class *j* ($P(C_j)$) in equation (4). Then Naïve bayes algorithm shown on Table 1.

$$C_{tweet\ j} = argmax\ P(C_j).\prod_{i=1}^{n} P(w_i|C_j) \tag{4}$$

Table 1. Naive Bayes Algorithm[17]

Naïve Bayes Algorithm
1. Prepare a dataset
2. Calculate the number of classes in the training
3. Count the same number of cases with the same class
4. Multiply all results according to the testing data that the class will look for
5. Compare results per class, the highest value is set as the latest class

2.6 Evaluation

Output of Naïve bayes classifier process on this study is hoax prediction model. The model was run on a test dataset and evaluated using a variety of metrics. Accuracy is the most commonly used evaluation metric. However, the accuracy metric is vulnerable to unbalanced classes[16], [17] . Consequence, the study provides evaluation metrics such as precision, recall, and F1 scores. The confusion matrix was used to calculate all metrics (Table 2).

Table 2. Confusion Matrix

		Predict	
		Hoax	**Not-hoax**
Actual	**Hoax**	True Positive (TP)	False Negative (FN)
	Not-hoax	False Positive (FP)	True Negative (TN)

Precision (P) is defined as the proportion of correctly predicted positive observations to all positively predicted observations [19]. The metric was calculated using formula in equation (5).

$$P = \frac{TP}{TP+FP} \,.\, 100\% \tag{5}$$

The proportion of correctly predicted positive observations to all available samples is referred to as recall[20]. The metric was calculated using formula in equation (6).

$$Recall(R) = \frac{TP}{TP+FN} \,.\, 100\% \tag{6}$$

The F1 score is a weighted average of Precision and Recall, as well as a technique for assessing the model's effectiveness. The metric can be calculated using formula in equation (7)

$$F1\ score = \frac{2PR}{P+R} \tag{7}$$

The proportion of correctly identified cases is referred to as accuracy. It calculates the proportion of correctly predicted observations to all observations. It is possible to calculate it using an equation (8).

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \,.\, 100\% \tag{8}$$

3. RESULT AND DISCUSSION

This section presents and discusses experiment results. To provide comprehensive analysis, this study ran a dataset using four scenario experiments. The scenario experiment refers to splitting the dataset (Table 3). This scenario is executed in two ways. The first scheme was run by a Naive Bayes classifier without TF-IDF. The second scheme was run with a Naive Bayes classifier with TF-IDF.

Table 3. Scenario Experiment

Scenario Experiment	Percentage Comparison of Training and Testing Datasets
I	90 :10
II	80 : 20
III	70 : 30
IV	60 : 40

3.1 Result of Naive Bayes without TF-IDF

This sub-section presents hoax prediction evaluations that use the Naïve Bayes classifier without the TF-IDF. The result is shown in Table 3. In general, the performance of the Naïve Bayes classifier model in detecting hoax Indonesia tweets is above 60%. This study believes that the model has not performed well, but the model is not really poor. The best model is obtained from experiment scenario IV. In other words, 60% of the training data is able to provide the best model for this scheme.

Table 4. Naive Bayes performance without TF-IDF (%)

Metric Performances	Scenario Experiment			
	I	II	III	IV
Accuracy	63	62	63	64
Precision	62.5	63	62	64
Recall	60.5	61.5	61	64
F1-score	61.25	62	61.5	64

3.2 Result of Naïve Bayes with TF-IDF

his sub-section presents hoax prediction results using Naïve Bayes classifier with TF-IDF. The result is shown in Table 4. There is a significant performance gap between scenario I and others, where the performance model produced by scenario experiment I is under 60%. Meanwhile, other experimental scenarios are able to provide models with performances above 60%. This study argues that using too largest training data is also not recommended for building hoax detection models. Furthermore, the best model in this scheme is obtained from the scenario IV experiment. In other words, 80% of the training data is able to provide the best model for this scheme.

Table 5. Naive Bayes performance with TF-IDF (%)

Metric Performances	Scenario Experiment			
	I	II	III	IV
Accuracy	59	64	64	63
Precision	58	69	66	65
Recall	55	64	60	60.5
F1-score	56.5	66.5	63	62.5

Based on the two test scenarios that have been carried out, this study compares the performances of the two. The comparison results are presented in Figure 4. The performance of the hoax prediction model using Naïve Bayes classifier is 64%. Based on this figure, the model obtained from the Naïve Bayes classifier with TF-IDF is superior based on metric precision and F1-score, while other metrics are not. This study argues that the TF-IDF has a positive effect on producing a hoax prediction model.

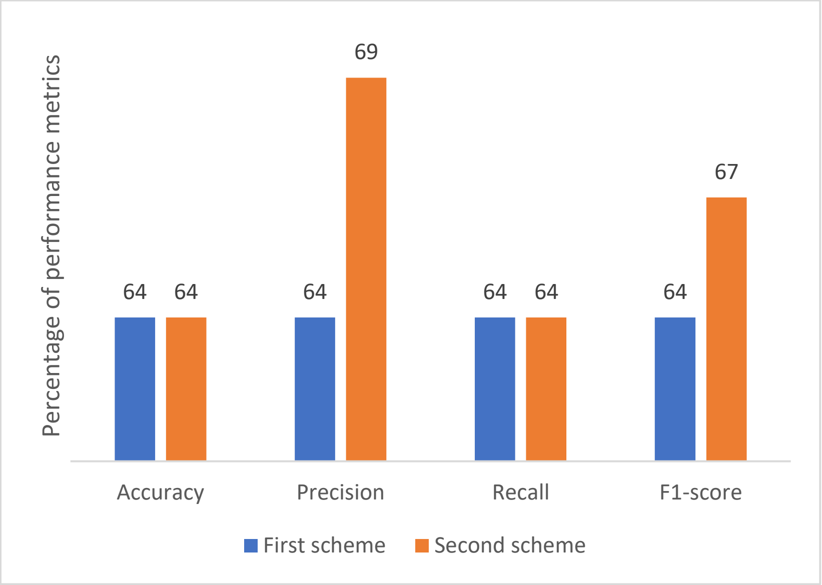


Figure 3. Performance Comparison of Both Scheme

4. CONCLUSION

Based on the results of the experiment, it can be concluded that identifying hoax Indonesian tweets can be accomplished using machine learning. Nave Bayes is a popular machine learning. To provide the best hoax prediction model, this work splits datasets into training and testing datasets. There are four experimental scenarios that refer to splitting the dataset. The experimental results showed that the hoax prediction model using Naïve Bayes with TF-IDF had 64% accuracy and recall, 69% and 67% precision, and a F1-score respectively. This result is also superior to the hoax prediction model when using the Naïve Bayes classifier without the TF-IDF. It means that TF-IDF has made a positive contribution to improving model performance. Lastly, the contribution of this research is that it can detect news that has a tendency towards hoaxes and can filter what is classified as hoaxes or not.

REFERENCES

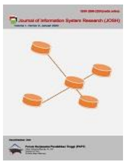
[1] U. R. Hodeghatta and S. Sahney, “Understanding Twitter as an e-WOM,” *Journal of Systems and Information Technology*, vol. 18, no. 1, 2016, doi: 10.1108/JSIT-12-2014-0074.

[2] T. Widaretna, J. Tirtawangsa, and A. Romadhony, “Hoax Identification on Tweets in Indonesia Using Doc2Vec,” in *2021 9th International Conference on Information and Communication Technology, ICoICT 2021*, 2021. doi: 10.1109/ICoICT52021.2021.9527515.

[3] Y. Priatna, “Hoax: An Information Society Challenge,” *Record and Library Journal*, vol. 4, no. 2, 2018.

[4] G. E. Dowd, *Groundless: Rumors, legends, and hoaxes on the early American frontier*. 2015. doi: 10.1093/jahist/jaw367.

[5] M. A. Hasbullah, “Hoax in legal perspective and literacy education in digital era,” *International Seminar and Call for Paper 2017 Darul Ulum Islamic University of Lamongan*, 2017.



[6] A. Fauzi, E. B. Setiawan, and Z. K. A. Baizal, “Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method,” in *Journal of Physics: Conference Series*, 2019. doi: 10.1088/1742-6596/1192/1/012025.

[7] D. J. Bayu, “Jumlah Pengguna Media Sosial di Dunia Capai 4,2 Miliar | Databoks,” *Databoks*, 2021.

[8] F. Rahutomo, I. Yanuar Risca Pratiwi, D. Mayangsari Ramadhani, and P. Negeri Malang Jalan Soekarno Hatta No, “Naïve bayes’s experiment on hoax news detection in Indonesian language,” *JURNAL PENELITIAN KOMUNIKASI DAN OPINI PUBLIK*, vol. 23, no. 1, 2019.

[9] M. Granik and V. Mesyura, “Fake news detection using naive Bayes classifier,” in *2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 - Proceedings*, 2017. doi: 10.1109/UKRCON.2017.8100379.

[10] D. A. N. Krisna and U. Salamah, “Perbandingan Algoritma Naïve Bayes Dan K-Nearest Neighbor Untuk Klasifikasi Berita Hoax Kesehatan Di Media Sosial Twitter,” *Jurnal Teknik Informatika Kaputama (JTIK)*, vol. 6, no. 2, 2022.

[11] A. Yodi Prayoga, A. Id Hadiana, and F. Rakhmat Umbara, “Deteksi Hoax pada Berita Online Bahasa Inggris Menggunakan Bernoulli Naïve Bayes dengan Ekstraksi Fitur Tf-Idf,” *Jurnal Health Sains*, vol. 2, no. 10, 2021, doi: 10.46799/jsa.v2i10.327.

[12] G. Bonaccorso, *Machine Learning Algorithms: Reference guide for popular algorithms for data science and machine learning*. 2017.

[13] A. Rusli, J. C. Young, and N. M. S. Iswari, “Identifying fake news in indonesian via supervised binary text classification,” in *Proceedings - 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2020*, 2020. doi: 10.1109/IAICT50021.2020.9172020.

[14] I. M. Karo Karo, M. F. M. Fudzee, S. Kasim, and A. A. Ramli, “Sentiment Analysis in Karonese Tweet using Machine Learning,” *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 219–231, Mar. 2022, doi: 10.52549/ijeei.v10i1.3565.

[15] J. Perkins, *Python 3 Text Processing With NLTK 3 Cookbook*. 2014.

[16] S. Fahmi, L. Purnamawati, G. F. Shidik, M. Muljono, and A. Z. Fanani, “Sentiment analysis of student review in learning management system based on sastrawi stemmer and SVM-PSO,” in *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, 2020. doi: 10.1109/iSemantic50169.2020.9234291.

[17] I. M. Karo Karo, M. Farhan, M. Fudzee, S. Kasim, and A. A. Ramli, “Karonese Sentiment Analysis: A New Dataset and Preliminary Result,” *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 2–2, pp. 523–530, 2022, [Online]. Available: www.joiv.org/index.php/joiv

[18] I. M. K. Karo, M. Y. Fajari, N. U. Fadhilah, and W. Y. Wardani, “Benchmarking Naïve Bayes and ID3 Algorithm for Prediction Student Scholarship,” *IOP Conf Ser Mater Sci Eng*, vol. 1232, no. 1, p. 012002, Mar. 2022, doi: 10.1088/1757-899X/1232/1/012002.

[19] I. M. K. Karo, A. Khosuri, and R. Setiawan, “Effects of Distance Measurement Methods in K-Nearest Neighbor Algorithm to Select Indonesia Smart Card Recipient,” in *2021 International Conference on Data Science and Its Applications, ICoDSA 2021*, 2021. doi: 10.1109/ICoDSA53588.2021.9617476.

[20] N. Z. Salih and W. Khalaf, “Prediction of student’s performance through educational data mining techniques,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, 2021, doi: 10.11591/ijeecs.v22.i3.pp1708-1715.