Indonesian Language Hoax News Classification Based on Naïve Bayes

Ari Sudrajat*[1], **Ratna RIzky Wulandari**[2]
[1)]Technical Information, [2)]Building Construction, Politeknik TEDC Bandung
Pesantren KM. 2, Cibabat, Kec. Cimahi Utara, Cimahi City, Jawa Barat, 40513
E-mail : arisud@poltektedc.ac.id*[1], ratnarizky@poltektedc.ac.id[2]
**Corresponding author*

Elvathna Syafwan
[3)]Mechanical Engineering, Politeknik TEDC Bandung
Pesantren No.KM. 2, Cibabat, Kec. Cimahi Utara, Cimahi City, Jawa Barat, 40513
E-mail : elvathna@poltektedc.ac.id[3]

Abstract - Hoax news in Indonesia causes various problems, therefore it is necessary to classify whether a news is in the hoax category or is valid. Naive Bayes is an algorithm that can perform classification but has a weakness, namely the selection of attributes that can affect accuracy so that it needs to be optimized by giving weights to attributes using the TF-IDF method. Classification using Naive Bayes and using TF-IDF as attribute weighting on a dataset of 600 data resulted in 82% accuracy, 84% precision, and 89% recall. The suggestion put forward is that it is better to use a larger number of datasets in order to produce higher accuracy.

Keywords – Hoax, Classification, Naïve Bayes, TF-IDF

1. INTRODUCTION

Information technology is growing rapidly, it is easier to get information whether it is through social media, blogs, or websites, but often there are problems with hoax news. Hoax is information or news that contains things that are not certain or which really are not facts that happened. Thus, hoaxes can cause various problems and losses for people who get the hoax news [1], [2]. Hoax news can be spread either using social media or websites. According to Juditha, in 2018 [3] the hoax phenomenon in Indonesia was seen as causing various problems. They appear more and more during the presidential election or regional head elections. This can be seen during the 2017 DKI Jakarta Pilkada. One way to identify or classify news including hoax or not can be done by checking it on a site managed by the Indonesian Anti Hoax Society (MAFINDO) organization, namely turnbackhoax.id. However, according to Mustofa and Mahfudh [4], turnbackhoax did a manual check or search so that when there was more and more information it would be difficult to search because there is more and more information and if there is no check on turnbackhoax, it is necessary to do the search yourself and that will certainly bother. Therefore, it will be easier if you can use computer technology, namely machine learning to classify whether the news is a hoax or not. Hoax is news or information that contains something that is not certain to happen or is not a fact. With the internet and social media, the spread of hoaxes will certainly be easier and more numerous [5], [6]. Hoaxes can cause various problems because the news that is spread is not news that is in accordance with the facts and it will harm various parties. Hoaxes can manipulate people who can invite people to take action, be disturbing, or commit fraud.

There are several studies related to the classification or identification of hoax news, including research conducted by Granik and Mesyura in 2017 [7] conducting research on the detection of hoax news using Naive Bayes. This study uses a dataset containing information on Facebook posts, each of which represents an English news article. The results of this study are that the classification accuracy for true and hoax news articles is approximately the same, but the classification accuracy for hoax news is slightly worse. This may be due to the slope of the dataset, which is only 4.5% of which is hoax news and with a total accuracy of 75.40%. Then on the research conducted by Rahutomo, Faisal; Yanuar, Ingrid; Andrie Asmara, in 2017 [8] conducted research on hoax news detection using Naive Bayes using Indonesian. The study used its own dataset with 250 pages of valid hoax and news articles. The results of the study were based on three exercises and randomized data testing using the php-ml library component, the highest average was achieved in 70% of the training dataset and 30% of the test dataset with an accuracy of 78.6%, hoax precision was 67.1%, valid precision is 91.6%, hoax recall is 89.4% and valid recall is 71.4%. Then in the research of Poddar, Amali and Umadevi, in 2019 [9] a comparison was made of several methods, namely Support Vector Machine (SVM), Decision Tree Classifier, Naive Bayes, Logistic Regression, and Neural Networks using datasets from Kaggle to produce accuracy of weighting using Count Vectorizer with Naive Bayes method namely 86.3%, SVM is 89.1%, Logistic Regression is 91.6%, Decision Tree Classifier is 82.5%, and Neural Networks is 49.9%. Then the accuracy of the weighting using TF-IDF Vectorizer Naive Bayes is 85.4%, SVM is 92.8%, Logistic Regression is 91%, Decision Tree Classifier is 81.6%, and Neural Networks is 49.9%. Based on the journal published by Poddar, Amali and Umadevi in 2019, there are various classifiers such as Support Vector Machine (SVM), Decision Tree Classifier, Naive Bayes, and Logistic Regression that can be used to predict hoax news. Naive Bayes has the advantage of being simple but very efficient. The downside is that the features are usually not independent. Logistic regression has advantages such as probability modeling, features can depend, and can update the model with new data easily. However, the drawback is that logistic regression requires a large data set for higher accuracy. Decision trees have the advantages of dependent features, linear separation of classes is not required, efficient outlier handling, and easy interpretation of the decision tree. However, the drawback of a Decision tree is that it will be appropriate when there are a large number of sparse features, and therefore perform poorly on the test data. In the journal [10] it is also explained that the Support Vector Machine (SVM) has the advantage of being able to minimize errors by maximizing margins by separating hyper-lanes and a data set even with a small number of samples. However, it has drawbacks, namely it is difficult to choose the right features and optimal attribute weights.

Before doing the classification, a text mining stage is carried out first. Text mining is a technique that can be used to classify, where text mining is a variation of data mining that seeks to find interesting patterns from a large set of textual data. The news text data will be processed first through a preprocessing so that the data can be better and cleaner. At this preprocessing stage, several subprocesses are carried out so that documents can be used to carry out the grouping process. Subprocesses at the preprocessing stage consist of case folding, tokenizing, stopword filtering, and stemming processes. Case Folding is converting all letters in the text to lowercase. Tokenizing is a process to sort the contents of the text so that it becomes a unit of words. Filtering is the stage of taking important words from the token results. Can use a stoplist algorithm (remove words that are less important) or wordlist (save important words). Stemming is a process to reduce words to their basic form. Then at the word weighting stage this research will use a method, namely TF-IDF. In Devita's journal, Herwanto and Wibawa in 2018 [11], it was explained that naive bayes is a method that calculates a set of probabilities by adding up the frequency and combination of values from a given dataset. In classifying using the Naive Bayes method, the data that has been given a label or category is then processed and produces a probability value for each term or word. The Naive Bayes algorithm also has the advantage that it does not require a large number of datasets. Then other advantages are explained in research conducted by Pramudita, Putro and Makhmud in 2018 [12], that the naive Bayes algorithm has advantages, namely the concept is easy to understand, not sensitive to relevant features, and can handle real and discrete data. According to Muhammad, in his research in 2017 [13], Naive Bayes also has weaknesses. Naive Bayes Classifier has a weakness in attribute selection so that it can affect the accuracy value. Therefore, the Naive Bayes Classifier needs to be optimized by giving weights to the attributes so that the Naive Bayes Classifier can work more effectively. In another study by [14], it was also explained that Naive Bayes is an algorithm that is very simple, efficient and also has good performance in many domains. In addition, naive bayes is also a popular machine learning technique in classifying text where the statistical classification process works with bayes theorem which can be used to predict the probability of membership of a class.

Based on the things above, it is proposed to use the Naive Bayes algorithm. Because Naive Bayes has the advantage that it does not require a large number of datasets, is efficient and also has good performance and uses TF-IDF as word weighting so that it can work more optimally.

2. RESEARCH METHOD

2.1. *Naïve Bayes*

Naive Bayes is one of the supervised learning algorithms to perform a classification in which Naive Bayes works based on the Bayes theorem to predict the probability of words in a category or class [15]. According to research [11], Naive Bayes is a simple probabilistic classification method. The advantage of using the Naive Bayes algorithm is that it only requires small training data. The stages in Naive Bayes are divided into 2, namely the training process and the classification process. In the Naive Bayes training process, it will calculate the probability value for each term or word from the training data that has been given a label or category [6]. The words generated fro m the process characterize a document into a certain category.

$$P(C|X) = \frac{P(X|C)XP(C)}{P(X)} \qquad (1)$$

Where :

P(C|X) : Probability of the hypothesis based on a condition

X : Data from unknown class

C : Hypothesis data from a certain class

P(C) : Hypothesis probability (prior probability)

P(X|C) : Probability based on the conditions in the hypothesis

P(X) : Probability of hypothesis X

Multinomial Naive Bayes is one of the specific methods of Naive Bayes. To calculate the probability of a document d in class C can be calculated by the following formula:

$$P(C|termdokumend) = P(X_1|C) \times P(X_2|C) \times \ldots \times P(X_n|C) \times P(C) \qquad (2)$$

Where :

P(C|term document d) : Probability of a document from class C

P(C) : Prior probability of class C

Xn : word n in document d
P(Xn | C) : Probability of the nth word of class C

Then to calculate the Prior Probability in class C, namely with the formula:

$$P(C) = \frac{N_c}{N} \qquad (3)$$

Where :
Nc : Number of Class C of all documents
N : Total of all documents

Furthermore, for the probability of a word or term n using the formula:

$$(X_n|C) = \frac{\sum tf(X_n, d \in C) + \alpha}{\sum N d \in C + V} \qquad (4)$$

Where :
$\sum \mathrm{tf}(\mathrm{X_n}, \mathrm{d} \in \mathrm{C})$:Total weighting (Weight) of the nth word of all training data documents in category C
$\sum \mathrm{N\,d} \in \mathrm{C}$: The total weight of all words in the training data from category C
α : Laplace Smoothing Value
V : The number of all words in the training data

2.2. Text Processing in Text Mining

Text mining is a text processing to extract and find information from a text or unstructured data that is useful through recognizing and exploring a pattern in a text. (Somantri, Wiyono and Dairoh, 2016) Thus text mining needs to be done in classifying a text in order to get the information needed for the data classification process [16]–[18]. Before classifying the data, we need a preprocessing which is part of text mining. Text Preprocessing is a method for changing the form of data that is still unstructured to be more structured in order. At this preprocessing stage, there are several sub-processes, including: case folding, tokenizing, filtering, and stemming.

2.3. Term Frequency Inverse Document

TF-IDF is a method that can be used to calculate the weight of each word or the frequency of occurrence of a word. According [1] Frequency (TF) is divided into several types of equations or formulas, namely:

1. Binary TF is to see whether a term or word appears or not in a document, if it appears then it is given a value of one (1), otherwise it will be given a value of zero (0).
2. RAW TF, which is how many times a word or term appears in a document. Suppose a word appears 7 times in a document, then the word will be worth 7 in the document.
3. Logarithmic TF, is used to avoid a dominant document that has few words in the query but has a high frequency. With the equation, namely:

   $$TF = 1 + log(TF) \qquad (5)$$

4. TF normalization, which is a comparison between the frequency in a term with the total number of terms in the document using :

$$TF = 0.5 + 0.5x\left(\frac{TF}{maxTF}\right) \qquad (6)$$

There is a formula to calculate TF-IDF, namely:

$$W_{d,t} = tf_{d,t} * IDF \qquad (7)$$

Where, W is the weight of the nth document, d is document, t is keyword, tf is the terms frequency (number of occurrences of the word), IDF is Inverse Document Frequency. To calculate the value of tf is by the formula:

$$tf_d = \frac{Jumlahmunculnyakatatdalamdokumen}{Totaljumlahseluruhkatadalamdokumen} \qquad (8)$$

To calculate the IDF value, that is with the formula:

$$IDF = log\left(\frac{D}{df}\right) \qquad (9)$$

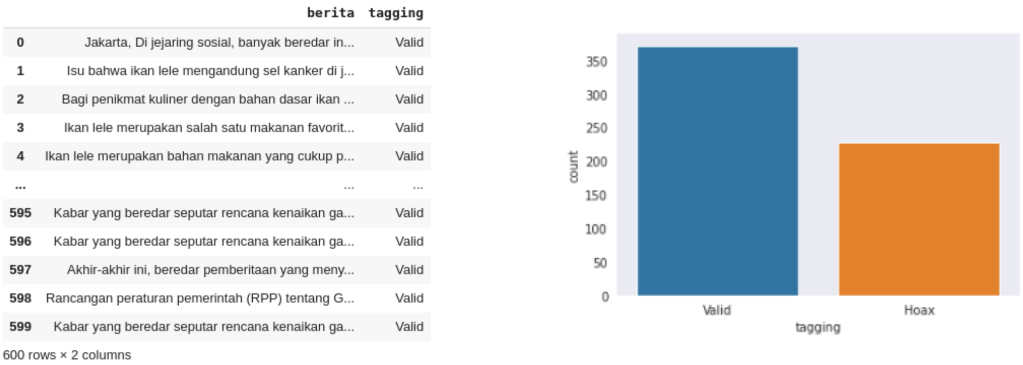
Where D is the total document and df is the number of documents containing the keyword.

2.4. Data Collection

The data collection process in this study used public data from https://data.mendeley.com/datasets/p3hfgr5j3m/1 which was part of the research by Rahutomo, Faisal; Yanuar, Inggrid; Andrie Asmara in 2017 that contains hoax news and valid news that have been labeled with a total of 600 data packaged in one .csv format file and consists of 2 columns, namely the news column and the tagging column. The data will be processed using pre-processing, the TF-IDF process, Naive Bayes for the classification of hoax news and a confusion matrix to determine the accuracy of the classification of hoax news.

3. RESULTS AND DISCUSSION

The dataset used in this study uses public data from https://data.mendeley.com/datasets/p3hfgr5j3m/1 containing 600 data in .csv file format which is divided into 2 columns, namely the news column and tagging. The label of this dataset is also available, namely in the tagging column as shown in Figure 1 point a. The dataset contains 2 categories, namely hoax news and valid news with 372 valid data and 228 hoax data as shown in Figure 1 point b.



	berita	tagging
0	Jakarta, Di jejaring sosial, banyak beredar in...	Valid
1	Isu bahwa ikan lele mengandung sel kanker di j...	Valid
2	Bagi penikmat kuliner dengan bahan dasar ikan ...	Valid
3	Ikan lele merupakan salah satu makanan favorit...	Valid
4	Ikan lele merupakan bahan makanan yang cukup p...	Valid
...	...	...
595	Kabar yang beredar seputar rencana kenaikan ga...	Valid
596	Kabar yang beredar seputar rencana kenaikan ga...	Valid
597	Akhir-akhir ini, beredar pemberitaan yang meny...	Valid
598	Rancangan peraturan pemerintah (RPP) tentang G...	Valid
599	Kabar yang beredar seputar rencana kenaikan ga...	Valid

600 rows × 2 columns

(a) (b)

Figure 1. (a) Sample Dataset, (b) Percentage of Valid and Hoax Data

In conducting the classification, pre-processing is carried out first, namely case folding, tokenizing, filtering, and stemming. In this study, 2 samples of data were used to perform sample calculations using pre-processing, TF-IDF, and the classification process using the Naive Bayes method.

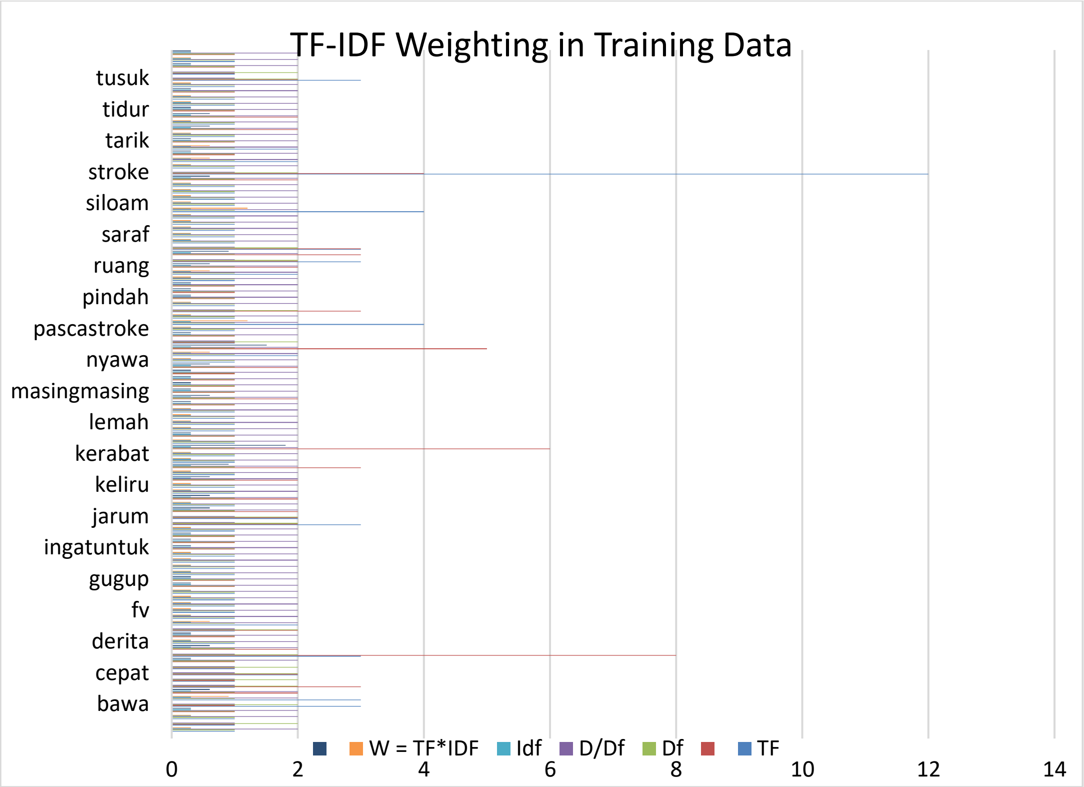


Figure 2. TF-IDF Weighting in Training Data

After getting the results in Figure 2, then do the same word search process with the model as shown in Figure 3.

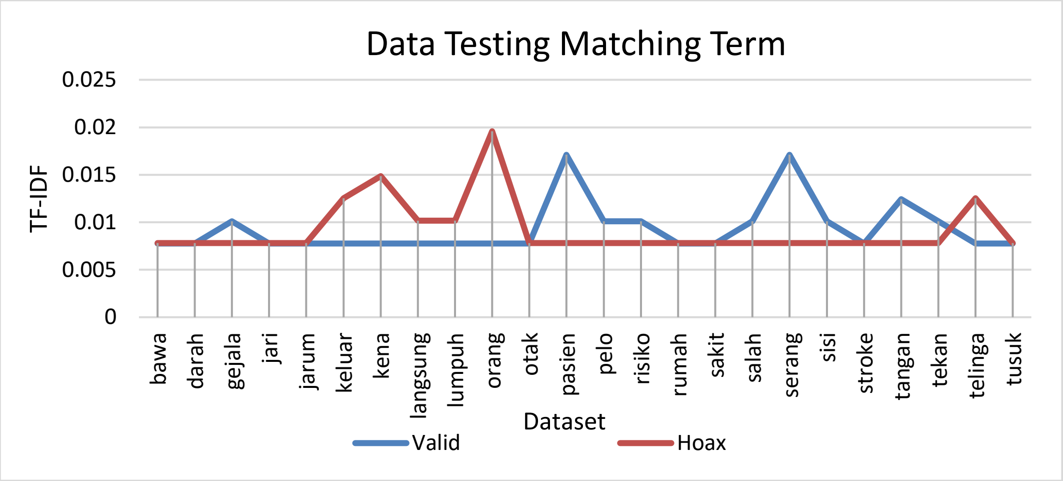


Figure 3. Calculate TF of Test Data

After getting the same words, then calculate the probability of each category by multiplying all the probabilities of words in the same test data category with the training data and also multiplying by the prior probability.

$$P(C|termdokumend) = P(X_1|C) \times P(X_2|C) \times \ldots \times P(X_n|C) \times P(C) \quad (10)$$

Valid Class
= Probability of the word Valid x Prior Probability Valid
= 8.59641574614452E-50 x 0.5
= 4.29820787307226E-50

Hoax Class
= Prior Probability Hoax x Prior Probability Hoax
= 5.51687100416534E-50 x 0.5
= 2.75843550208267E-50

Because 4.29820787307226E-50 is greater/max than 2.75843550208267E-50, the D1 test data is included in the Valid category or class.

Table 1. Prediction Results

Data	Prediction
D1	Valid
D2	Valid
D3	Hoax
D4	Valid

To get the value of accuracy, precision, and recall using the confusion matrix method.

Table 2. Prediction and Actual Results of Test Data News

Data	Prediction	Actual
D1	Valid	Valid
D2	Valid	Hoax
D3	Hoax	Hoax
D4	Valid	Valid

Table 2 is the result of the classification (prediction) of news 4 test data using the Naive Bayes algorithm and the results of the actual category data (Actual).

Table 3. Confusion Matrix of Prediction and Actual Results of Test Data News

	Actual Valid	Actual Hoax
Prediction Valid	2 (TP)	1 (FP)
Prediction Hoax	0 (FN)	1 (TN)

Accuracy : $\frac{\text{tp+tn}}{\text{tp+fp+fn+tn}}=\frac{2+1}{2+1+1+0} = 0.75$X 100% = 75%

Presision : $\frac{\text{tp}}{\text{tp+fp}}=\frac{2}{2+1} = 0.67$X 100% = 67%

Recall : $\frac{\text{tp}}{\text{tp+fn}}=\frac{2}{2+0} = 1$X 100% = 100%

Based on the results of the above calculations, obtained the results of an accuracy of 76%. then Precision is 67%, and Recall is 100%. The distribution of data is done randomly using a module from the sklearn library, namely train_test_split. Based on the test results obtained as shown in Table 4.

Table 4. Confusion Matrix of Prediction and Actual Results of Test Data News

Metode	Test Size 20%			Test Size 30%			Test Size 40%			Test Size 50%			Test Size 60%		
	A	P	A	P	A	P	R	R	A	P	R	R	A	P	R
Naive bayes	82	83	89	82	84	89	82	85	88	76	81	82	68	78	71
SVM Linear	71	73	83	74	77	83	70	76	79	72	79	78	67	74	75
SVM Rbf	68	67	93	73	73	93	74	74	93	71	72	89	70	71	87
SVM Polynomial	66	65	97	71	69	98	70	69	98	67	67	95	66	66	95
SVM Sigmoid	67	66	94	72	72	94	74	73	94	67	70	87	66	68	89

Based on table 4, The results of this test are also compared with the support vector machine method. The support vector machine method is compared with the proposed method, namely Naive Bayes because in related studies that have compared several methods to produce the SVM method with the TF-IDF vectorizer, it gets the best results with the highest accuracy, but when applied to the dataset of this study, the results of SVM are still inferior compared to naive bayes method. Naive Bayes method gets the best results with an accuracy of 82% while SVM produces the highest accuracy, which is 74%. The svm method has a linear kernel and a non-linear kernel, for the linear kernel is used when the classified data can be separated easily through a line or hyperplane, so the classified data is separated linearly. Hyperplane functions as a separator of two classes by calculating the margin which is the distance between the hyperplane and the support vector. And Support vector is a pattern closest to each class. For non-linear kernels such as the rbf kernel, polynomial, and sigmoid are used when the data used cannot be separated by a line, then they are separated using curved lines or on a plane in a high-dimensional space. The curved line or plane in this high-dimensional space is a modification of SVM by inserting kernel functions into non-linear SVM. These kernel functions can be defined as input kernel tricks. Kernel trick is part of learning in svm to find out kernel functions without having to know the existence of non-linear functions. Then the naive bayes method works by predicting the probability of each word in a class. Based on this, the SVM method has the opportunity to get the highest accuracy because it has several kernels that can be used to adjust the existing data, while Naive Bayes only relies on word probabilities. But in fact the results in this study show that Naive Bayes has the highest accuracy. This may be because the data used is difficult to separate using a hyperplane or curved line from SVM, either using a linear kernel, rbf, polynomial, or sigmoid. Meanwhile, Naive Bayes which uses probability may be able to work more optimally on the data used. Then based on the composition of the testing data and training data, the percentage of training data is higher and the testing data is less, resulting in better accuracy. The Naive Bayes method has an accuracy of 82% in the distribution of training data compared to testing data, which are 80:20, 70:30, and 60:40. While svm produces the highest accuracy, namely 74% in the distribution of 70% training data and 30% testing data with linear kernels and 74% accuracy also in 60% training data distribution and 40% testing data with rbf and sigmoid kernels.

4. CONCLUSION

Based on this research, it can be concluded that the Naive Bayes algorithm can classify news text data with a total of 600 data consisting of hoax and valid categories and TF-IDF can perform attribute weighting and can produce 82% accuracy, 84% precision, and recall 89%. With these results, the text data model will be more suitable to use the Naive Bayes method because it has a fairly far distance of accuracy with SVM, which is 8%. However, it is not impossible that SVM has higher accuracy than Naive Bayes when applied to different text data. To improve results, in future research it is better to use a larger number of datasets in order to produce higher accuracy, can also improve the Naive Bayes algorithm or use other algorithms to increase accuracy.

REFERENCES

[1] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, “Fake News Detection Using Sentiment Analysis,” *2019 12th Int. Conf. Contemp. Comput. IC3 2019*, pp. 1–5, 2019.

[2] D. Katsaros, G. Stavropoulos, and D. Papakostas, “Which machine learning paradigm for fake news detection?,” *Proc. - 2019 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2019*, pp. 383–387, 2019.

[3] C. Juditha, “Interaksi Komunikasi Hoax di Media Sosial Serta Antisipasinya,” *J. Pekommas*, vol. 3, no. 1, pp. 31–34, 2018.

[4] H. Mustofa and A. A. Mahfudh, “Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes,” *Walisongo J. Inf. Technol.*, vol. 1, no. 1, p. 1, 2019.

[5] M. G. Hussain, M. Rashidul Hasan, M. Rahman, J. Protim, and S. Al Hasan, “Detection of Bangla Fake News using MNB and SVM Classifier,” *Proc. - 2020 Int. Conf. Comput. Electron. Commun. Eng. iCCECE 2020*, pp. 81–85, 2020.

[6] M. Singh, M. Wasim Bhatt, H. S. Bedi, and U. Mishra, “Performance of bernoulli’s naive bayes classifier in the detection of fake news,” *Mater. Today Proc.*, no. xxxx, 2020.

[7] M. Granik and V. Mesyura, “Fake news detection using naive Bayes classifier,” *2017 IEEE 1st Ukr. Conf. Electr. Comput. Eng. UKRCON 2017 - Proc.*, pp. 900–903, 2017.

[8] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, “Study of hoax news detection using naïve bayes classifier in Indonesian language,” in *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, 2017, pp. 73–78.

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

[10] E. Pudjiarti, “PREDIKSI SPAM EMAIL MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DAN PARTICLE SWARM OPTIMIZATION,” *J. Pilar Nusa Mandiri*, vol. XII, no. 2, pp. 171–181, 2016.

[11] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, “Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, p. 427, 2018.

[12] Y. D. Pramudita, S. S. Putro, and N. Makhmud, “Klasifikasi Berita Olahraga

Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 3, p. 269, 2018.

[13] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, “Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 3, p. 180, 2017.

[14] D. A. Muthia, “Analisis Sentimen Pada Review Restoran Dengan Teks Bahasa Indonesia Mengunakan Algoritma Naive Bayes,” *J. Ilmu Pengetah. Dan Teknol. Komput.*, vol. 2, no. 2, pp. 39–45, 2017.

[15] T. Herdiawan Apandi and C. Agus Sugianto, “Analisis Komparasi Machine Learning Pada Data Spam Sms,” *TEDC*, vol. 12, no. 1, p. 58, 2018.

[16] C. A. Sugianto and T. H. Apandi, “Pengaruh Tokenisasi Kata N-Grams Spam SMS Menggunakan Support Vector Machine,” in *CITISEE 2017*, 2017, pp. 5–9.

[17] T. H. Apandi and C. A. Sugianto, “Penyaringan Spam Short Message Service Menggunakan Support Vector Machine,” *Semin. Nas. Teknol. Inf. dan Komun. Terap.*, pp. 111–116, 2015.

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