Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 839−846
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1957

Comparison of Ensemble Methods for Detecting Hoax News

Delvanita Sri Wahyuni, Yuliant Sibaroni*

School of Computing, Telkom University, Bandung, Indonesia
Email: [1]delvanita@student.telkomuniversity.ac.id, [2,*]yuliantsibaroni@telkomuniversity.ac.id
Email Penulis Korespondensi: yuliantsibaroni@telkomuniversity.ac.id
Submitted:**25/07/2022**; Accepted:**18/08/2022**; Published: **30/09/2022**

Abstract−The spread of hoaxes in Indonesia has become a big concern for the public, especially now that the COVID-19 virus pandemic is hitting the whole world. Due to the large number of people who believe the hoax news regarding the COVID-19 vaccination that has spread on social media, many people refuse to carry out the COVID-19 vaccination as a form of government effort in dealing with this pandemic. Therefore, people need to be wiser when reading news on social networks. To help the public not to read hoaxes, it is necessary to classify the COVID-19 vaccine hoax. This study builds a system to classify hoax news on the COVID-19 vaccine. The model was built using the ensemble method by comparing the Random Forest and AdaBoost algorithms to choose a good classification for detecting hoaxes. In this research, there are use two test scenarios. The first scenario is an experiment using the Random Forest algorithm method and the second scenario is an experiment using the Adoboost algorithm method. The experimental results show that the first scenario produces a good accuracy value with the random forest algorithm method of 93.58%.

Keywords: Hoax; Covid-19 Vaccine; Ensemble; Random Forest; Adaboost

1. INTRODUCTION

On March 11, 2020, the WHO (World Health Organization) determined that a new pandemic had occurred in Wuhan, China. This virus is caused by infection with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which can affect or infect the respiratory system in humans [1]. The characteristics of people infected with the virus are increasingly diverse, ranging from high fever over 39°C, dry cough, difficulty breathing, shortness of breath, and loss of smell and taste [2]. However, many are unaware that they have been exposed to the coronavirus, so it is very easy to transmit it to each other. The coronavirus has spread in various parts of the world, including Indonesia. The Indonesian government has made multiple efforts to deal with the pandemic, such as limiting community activities, using masks when leaving the house, washing hands, maintaining distance, and self-quarantining if the COVID virus is detected [3]. The following solution by the government is to accelerate vaccination for all Indonesians. The information is disseminated to all news channels and social media, making it easier for the general public to get the news.

According to Silverman, hoax news is a series of pieces of information that are misled but are 'sold' as accurate news [4]. The amount of information from news sites and social media that spread hoax news about the side effects of the COVID-19 vaccine has made people nervous about vaccination. The spread of hoax news is very dangerous for the community, especially now that it is easy to find information on social media platforms such as Facebook, Instagram, Twitter, YouTube, TikTok, and WhatsApp. Hoax news is usually made to influence views in the political field or just be news [5]. This makes the government unable to contain the COVID virus in Indonesia because many people do not believe in spreading fake news. One of the efforts that can be made is to detect fake news to see whether the news is true or false.

Several studies related to the detection of hoax news have been carried out previously, such as research conducted by Awalina et al. [6]. Detected hoax news using Deep Learning Network Transformers with the BERT method and Transformer Network as the basis for reference and comparison between other methods using CNN, Hybrid CNN-BiLSTM, and BiLSTM. The results obtained by the BERT method show that the accuracy value reaches 90%, supported by the f1-score, which is 90% greater than the other methods. Santoso et al. [7], conducted research to classify fake and genuine Twitter accounts using the Naïve Bayes method has an accuracy value of 80%, using a 5% training set. Another study was conducted by Panjaitan et al. [8], using Random Forest, SVM, Gradient Boosting, Naïve Bayes, and Logistic Regression methods. The result with a good accuracy value is Random Forest, with an accuracy value of 96.05%. Das et al. [9] researched the hoax news detection system using the ensemble method, which yielded an f1-score accuracy value of 0.98%. Research by Gowthami et al. [10], has identified fake news by using a comparison of the SVM and Random Forest algorithm methods; the results obtained show that the Random Forest algorithm can predict by producing a value of 98% compared to the SVM algorithm, which produces an accuracy value of 70%. Therefore, the Random Forest algorithm has better and more efficient performance.

From several series of studies that have been carried out, we have not found any topics regarding comparisons regarding the detection of hoax news related to the COVID-19 vaccine. This has prompted us to conduct research related to the classification of hoax news regarding the COVID-19 vaccine to find out whether the news is hoax news or fact.

In this study, we classify those related to the spread of hoax news about vaccination in the community. The dataset used comes from a government website specifically related to COVID-19 news, namely https://covid19.go.id/. The data is taken from the period 01 January 2020 to 28 February 2022[11]. In this case, we use two test scenarios, namely a data scenario and a classification scenario. In the data scenario, the data is used using three data types,

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 839−846
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1957

namely data that is different only in the "Title", the "News Content" section, and the combination of "Title + News Content". Then, in the classification scenario, we conducted an algorithm comparison experiment using the basic ensemble approach, namely Random Forest and AdaBoost. Forest Random Selection is known for its simplicity and effectiveness in generating good classifications [12], and the choice of the AdaBoost algorithm is because it is suitable for unbalanced data [13]. The purpose of this experiment is to find out which ensemble method effective in detecting is good to use in detecting hoaxes related to COVID-19 vaccine data.

2. RESEARCH METHODOLOGY

In this study, we built a system to detect fake news about the COVID-19 vaccine, which includes steps such as setting up a labeled news dataset. Then the dataset is processed at the preprocessing stage to get good data. Furthermore, the data is processed into a matrix form using Word2Vec which will later divide the data into training data and test data. In the classification, two models of the ensemble method are used, namely Random Forest and AdaBoost with labels from existing datasets. To build the system, there are several stages, namely retrieving data, conducting data preprocessing, performing feature extraction, modeling ensemble methods, and conducting an evaluation process. The process description of the system being built is illustrated in Figure 1.

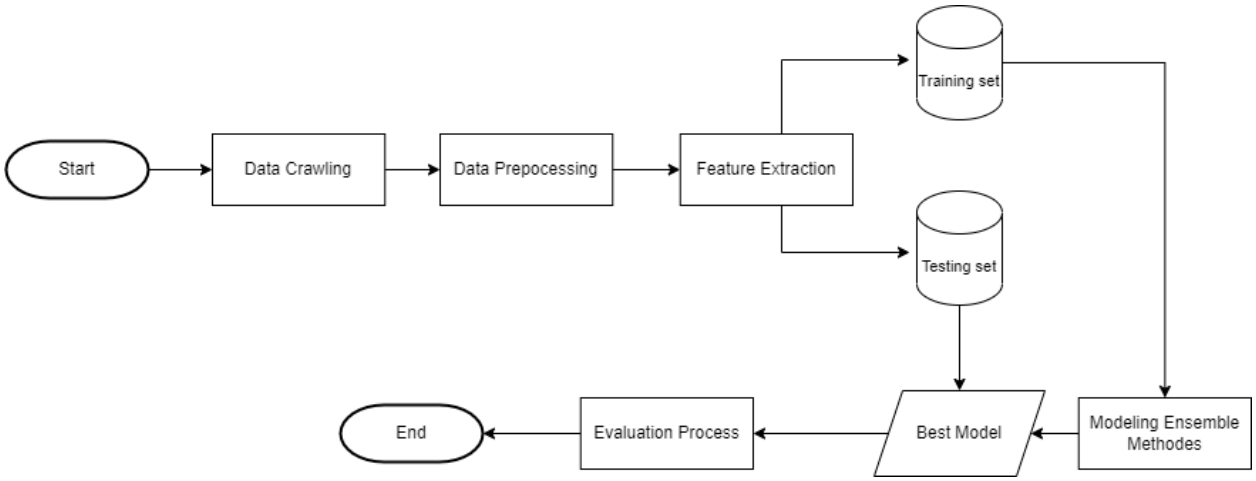

Figure 1. System Flowchart

2.1. Data Crawling

The dataset used only data that provided information regarding the COVID-19 vaccination through the government website, namely “covid19.go.id”. The data taken is news in the Indonesian language, which is combined into one in the form of an excel format “.csv” and has the attributes of title, news content, date, source, link, and “Hoax” and “Facts” category labels. The label given to “Hoax” news is “1”, and in fact, the news is labeled “0”. Hoax data is obtained from public reports that check the news being discussed on various social media. Meanwhile, real news is news provided by the government as a form of correct information related to the news of the COVID-19 vaccination.

2.2 Data Preprocessing

The data preprocessing stage consists of several steps, namely case folding, removing punctuation, stopwords, tokenizing, and stemming. Case Folding is used for uniformity of each word, and all capital letters will be changed to lowercase [14]. The Removing Punctuation step is used to remove special characters in the text of punctuation symbols such as commas (,), periods (.), question marks (?), exclamation marks (!), and others. Then use the eraser to remove numeric numbers (0–9) and other character symbols (for example, $, %, and &). [14]. Stopwords is a technique for eliminating frequently occurring but meaningless words, commonly known as the use of conjunctions. The procedure we use is in accordance with the Indonesian corpus (for example: “yang”, “itu”, “dan”, “karena”, “ini” and so on) [15]. Tokenizing is the process of separating text in sentences into tokens or word pieces [12], The stemming process uses a library from Python, namely Sastrawi to change the form of a sentence into its basic or original form by removing the added words (for example, “mendengarkan” to “dengar”, and “memakan” to “makan”, and so on) [15].

2.3 Feature Extraction

Word2vec is a word insertion method used to represent a word as a vector. Through the process of training vector representations with properties associated with related words are obtained [16]. This method takes text as input, then later the results of the input will produce a vector representation of each word in the text corpus as output. This word vector can also be used to measure the proximity between other word vectors [17]. Words that appear repeatedly in the context of N-grams have a tendency that is triggered by the same weight that will cause these words to be correlated

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 839−846
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1957

or connected. N-grams are pieces of a certain number of words or pieces of the number (n) of a sentence, and the features used in N-grams in this study are n = 1.

In word2Vec, there are two main learning algorithms, namely continuous bag-of-words (CBOW) and skip-gram. The CBOW model is a model that predicts when a certain word appears through the results of the analysis of neighboring words in the form of window sizes. Meanwhile, Skip-gram is a model that predicts surrounding words in the range of words before and after the current word in the same sentence [18]. In this study, we used the model from CBOW because it is faster and works well on repetitive words.

The weights of the input and hidden layers are represented by a matrix W of size V x N. V is the dimension of the input layer, while N is the dimension of the hidden layer. In addition, the W matrix is described as a matrix of size N x V[19]. Each line W is a representation of an N-dimensional vector v_w of the assembled word from the input layer. The input layer or context word is a one-hot encoded vector, which means that the given context word is only one of the vocabulary sizes that can be referred to as part V $\{x_1 \dots x_v\}$ will be 1 and the other part will be 0. v_{w_I} is a representation of the input word vector w_I This means that the function of the hidden layer part can directly pass the weighted input sum to the next layer. $v'_{w_j}T$ to calculate the total weighted score for each word in the vocabulary [18]. The following calculations can be seen in equation (1).

$$p(w_j|w_I) = \frac{exp\,(v'_{w_j}T\ v_{w_I})}{\sum_{j'=1}^{V} exp\,(v'_{w_{j'}}T\ v_{w_I})} \quad (1)$$

2.4 Modeling Ensemble Methods

The ensemble method model is a machine learning technique that combines a set of models to get a good or optimal predictive model [20]. The ensemble algorithm used, Random Forest, introduced in 2001 by L. Breiman, is a machine learning technique that combines the same tree for classification and regression problems. A random forest can be illustrated as a collection of structured classifications of trees. Random forests can produce good accuracy values because new data sets are created from the original data set with the replacement of training data to reduce model complexity. Then the tree will grow by choosing features randomly and usually the process of creating and training data will be carried out in several iterations. [21].

As a comparison algorithm, the algorithm used is AdaBoost or Adaptive Boosting, first introduced by Freund, Y., & Schapire, R. E in 1997. It is a supervised algorithm that is applied to create a classification model. AdaBoost is an algorithm that combines several weak tree models into a strong model to increase the complexity of the model. AdaBoost is also an effective algorithm for the problem of unbalanced data, and it is possible to increase the minority class so that the data is balanced and displays good model results [13]. Each later iteration will give a higher priority value to an observation that was predicted incorrectly by the previous model [22].

To test the scenario, the research was conducted by testing scenarios on previously processed data. The test was carried out during preprocessing and data classification using three different datasets, namely "title only", "only news content", and "Title and content of news". The first scenario is carried out using all preprocessing steps to determine the effect of preprocessing data on model performance. The following scenario is without using stopword removal and stemming steps. Classification in both scenarios is done using the ensemble method, with the algorithms chosen to be Random Forest and AdaBoost. Using a technique that combines hyperparameter variables with GridSearchCV to know the performance of a good combination of parameters is one way to improve model performance [23]. Table 1 shows the results of the parameter variables involved for each model.

Table 1. Hyperparameters Tuning Ranges

Method	Parameters	Ranges
Random Forest	N estimators	[50, 75, 100]
	Min_samples leaf	[1, 2, 3]
	Min_samples split	[2, 5, 10]
Adaboost	N estimators	[100, 150]
	Learning rate	[0.5, 0.01, 0.2, 0.3, 0.4]

2.5 Evaluation

The next step is the evaluation. The evaluation used in this research is the confusion matrix to see the evaluation considerations of a particular work from an algorithm. So, we need some metrics to get information on algorithm performance. The evaluations used are accuracy, precision, recall, and f1-score. Accuracy is the ratio of the correct predictions of the entire dataset, or it can be said that accuracy has a value close to the actual value. Precision is a ratio that predicts a true positive of all truly positive data. The recall is a prediction ratio that compares all actual positive data values. Finally, the f1-score is the average value of precision and recall as consideration for evaluating the algorithm [24]. To perform the calculations can be seen in equations (2)-(5).

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 839−846
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1957

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP + TN}{TP + FN} = \frac{TN}{P} \quad (4)$$

$$\text{F1} - \text{score} = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (5)$$

3. RESULT AND DISCUSSION

3.1 Dataset

The dataset used is different in each test scenario. The dataset used has a total of 1453 with the label number of hoax data being 446 and the correct data label is 1007 data. Because the datasets obtained are very different, or because it can be said that the data is not balanced, we use random oversampling on the minority class to get a balanced data value. The test is carried out with three data scenarios in the processed column, namely "Title", "News Content", and "Title + News Content". Then the data is split into train and test data with the ratio of 80:20. The dataset used can be seen in the example Table 2.

Table 2. Datasets

Title	News Content	Title + News Content
Presiden Memastikan Vaksin COVID-19 Gratis untuk Masyarakat	Presiden Joko Widodo telah memastikan vaksin COVID-19 untuk masyarakat Indonesia adalah gratis, tidak dikenakan biaya sama sekali. Tetaplah disiplin lakukan protokol kesehatan 3M (Memakai masker dengan benar, Menjaga jarak serta menjauhi kerumunan, Mencuci tangan pakai sabun) serta dukung upaya pemerintah 3T (Testing, Tracing, Treatment), dan siap untuk divaksin saat vaksin sudah siap.	Presiden Memastikan Vaksin COVID-19 Gratis untuk Masyarakat Presiden Joko Widodo telah memastikan vaksin COVID-19 untuk masyarakat Indonesia adalah gratis, tidak dikenakan biaya sama sekali. Tetaplah disiplin lakukan protokol kesehatan 3M (Memakai masker dengan benar, Menjaga jarak serta menjauhi kerumunan, Mencuci tangan pakai sabun) serta dukung upaya pemerintah 3T (Testing, Tracing, Treatment), dan siap untuk di vaksin saat vaksin sudah siap.

3.2 Preprocessing Data Result

The dataset used is different for each test scenario, and to get more structured data, each sentence or word is required to pass through the stages of data preprocessing: case folding, removing punctuation, tokenizing, stopword removal, and stemming. After passing the tokenizing stage, from 1453 data lines, 308628 tokens were generated, then 187529 tokens after the stopword removal and stemming process.

3.3 Feature Extraction

The word2vec model makes every word in the data it has turned into a vector. The vector in this model can be determined by the number of features. However, because word2vec has a default feature, we use the number of features of 100. The data is changed using the package from Gensim in the Python programming language to form the word2vec model as long as the number of features is 100 and the value of the word order or N-gram is n = 1, where the model only displays one word, which will generate 100 feature vectors. In Table 3, it can be seen that the model generated by word2vec is a word and the results are from vectors 1 to 100.

Table 3. Word2Vec Model

Word	v1	v..	v100
Vaksin	-1.11629866	-3.64977300	4.92604971
Masyarakat	-6.57811016	-1.76345706	1.82826314
Covid	-9.90428552	-2.80601472	3.33659127
Indonesia	-6.57811016	-1.76345706	1.82826314
Infeksi	-6.27990486	-2.00168118	-5.46948053

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 839−846
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1957

3.4 Result

Hyperparameter tuning is used to find the best parameter values for all models. We carried out two scenarios, the first data scenario that uses the overall processing step and the second scenario, namely the data that does not use stopword and stemming steps. This scenario test was run five times using random sample values to find the best parameter values. The parameter values generated in the five trials are different, so the predicted values that appear are also different in each iteration.

Table 4. Summary Best Values of Hyperparameter Tuning Result (Scenario 1: Full Pre-processing)

Parameter	Random Forest (Iteration 1-5)			AdaBoost (Iteration 1-5)		
	Title	News Content	Titles + News Content	Title	News Content	Titles + News Content
N estimator	[75, 75, 75, 100, 50]	[100, 100, 100, 75, 100]	[50, 100, 50, 75, 100]	[150, 150, 150, 150, 150]	[150, 100, 150, 150, 150]	[150, 150, 150, 150, 150]
Min samples leaf	[1, 1, 1, 1, 1]	[1, 1, 1, 1, 2]	[1, 1, 1, 1, 1]	-	-	-
Min samples split	[5, 2, 5, 2, 2]	[2, 2, 2, 5, 10]	[2, 2, 2, 5, 5]	-	-	-
Learning rate	-	-	-	[0.5, 0.5, 0.5, 0.5,0.5, 0.5]	[0.3, 0.2, 0.5, 0.4, 0.2]	[0.5, 0.5, 0.5, 0.5, 0.5]

The best values of the tuning hyperparameters for five iterations are shown in Table 4. Table 5 is the result of the average value of scenario 1, which shows the accuracy performance value of the three different process data. As can be seen in Table 5, among the three data scenarios processed, the Random Forest algorithm method has the best average accuracy value, with the "News Content" data scenario achieving an accuracy value of 93.58%, which is higher than the "Title" and "Title + News Content" data scenarios. The f1-score value of 89.65 percent supports the high accuracy value. In the AdaBoost algorithm, the best accuracy value is also obtained in the "News Content" data scenario, which has an accuracy value of 91.65% with an f1-score of 86.70%. The lowest accuracy value from the algorithm test is in the "Title" data scenario with the results obtained, namely 68.58% for the Random Forest algorithm and 71.34% for the AdaBoost algorithm.

Table 5. Average Results of Scenario 1 (Scenario 1: Full Preprocessing)

Method	Scenario Data	Average	
		Accuracy	F1-score
Random Forest	Title	68,58 %	45,50 %
	News Content	**93,58 %**	**89,65 %**
	Title + News Content	86,87 %	79,52 %
AdaBoost	Title	71,34 %	56,68 %
	News Content	**91,65 %**	**86,70 %**
	Title + News Content	88,10 %	81,13 %

Table 6. Summary Best Values of Hyperparameter Tuning Result (Scenario 2: without Stopword Removal and Stemming)

Parameter	Random Forest (Iteration 1-5)			AdaBoost (Iteration 1-5)		
	Title	News Content	Titles + News Content	Title	News Content	Titles + News Content
N estimator	[100, 50, 100, 100, 50]	[50, 50, 50, 75, 75]	[50, 75, 100, 100, 50]	[150, 150, 150, 150, 150]	[150, 150, 150, 100, 150]	[150, 150, 150, 150, 150]
Min samples leaf	[1, 1, 1, 1, 1]	[1, 1, 1, 1, 1]	[1, 1, 1, 1, 1]	-	-	-
Min samples split	[2, 2, 2, 2, 2]	[2, 2, 2, 2, 5]	[5, 5, 2, 2, 5]	-	-	-
Learning rate	-	-	-	[0.5, 0.5, 0.5, 0.5,0.5, 0.5]	[0.2, 0.4, 0.4, 0.5, 0.4]	[0.5, 0.5, 0.5, 0.5, 0.5]

Scenario 2 testing is a test that does not use all of the preprocessing stages. In the second scenario, only Case Folding, Removing Punctuations and Tokenizing without Stopword Removal and Stemming. Table 6 shows the best values of the tuning hyperparameters obtained. For the second scenario, five iterations are also carried out so that the parameter values obtained will produce different values for each iteration.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 839−846
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1957

Table 7. Average Results of Scenario 2 (Scenario 2: without Stopword Removal and Stemming)

Method	Scenario Data	Average (%)	
		Accuracy	F1-score
Random Forest	Title	70,44 %	44,13 %
	News Content	**92,41 %**	**87,98 %**
	Title + News Content	86,94 %	79,59 %
AdaBoost	Title	68,58 %	56,03 %
	News Content	**92,68 %**	**88,42 %**
	Title + News Content	88,38 %	81,17 %

The results in Table 7 show that in second scenario testing, the average value of accuracy performance gets a high accuracy value in the Random Forest algorithm, namely the "News Content" data scenario with an accuracy value of 92.41%, which is supported by an f1-score value of 87.98%. Meanwhile, in the AdaBoost algorithm, the "News Content" data scenario yields results with a high accuracy value of 92.68%, supported by an f1-score value of 88.42%.

3.5 Analysis

Analysis of the results of the two tested scenarios can be seen in Table 6, which shows the test first scenario, where the value of the data scenario that gets the best performance value lies in the "News Content" data scenario in the two algorithms Forest and AdaBoost. In Figure 2, the bar graph height is slightly different between the two algorithms, Random Forest and AdaBoost, at about 1.93%. It shows that the Random Forest algorithm has good predictive accuracy compared to the AdaBoost algorithm in the first scenario testing, which performs a full preprocessing step.

Figure 3 shows the second scenario's graph with the test without using stopword removal and stemming steps. As seen in Table 8, the data scenario that has a high accuracy average value is also in the "News Content" data scenario. The comparison of the two algorithms in the data scenario can be very thin, at around 0.27%. The second scenario shows that the AdaBoost algorithm is superior or has an excellent predictive accuracy value compared to the Random Forest algorithm with incomplete preprocessing stages.

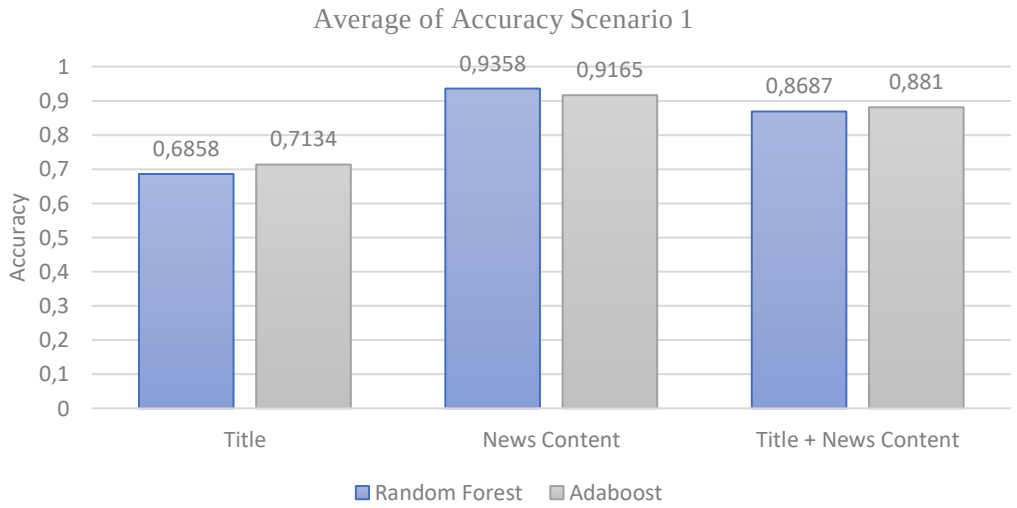

Figure 2. Comparison of Testing Results for Scenario 1 Full Pre-processing.

It can be seen from Figure 2 and Figure 3 that from the three scenarios, the data that appears to have a good accuracy value is only in the "News Content" data scenario. On the other hand, the comparison between testing the first and second scenarios can be seen in the accuracy value of the "Title" data scenario, where the "Title" data can be very influential in getting a good accuracy value.

In the first scenario, it is clear from the scenario data "Title + News Content" that there is a decrease in value of about 6.71% in the Random Forest algorithm and a decrease in the accuracy of the AdaBoost algorithm by about 3.55%. In the second scenario, the same data scenario shows a decrease in the accuracy value of about 5.47% for the Random Forest algorithm and about 4.3% for the AdaBoost algorithm.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 839−846
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1957

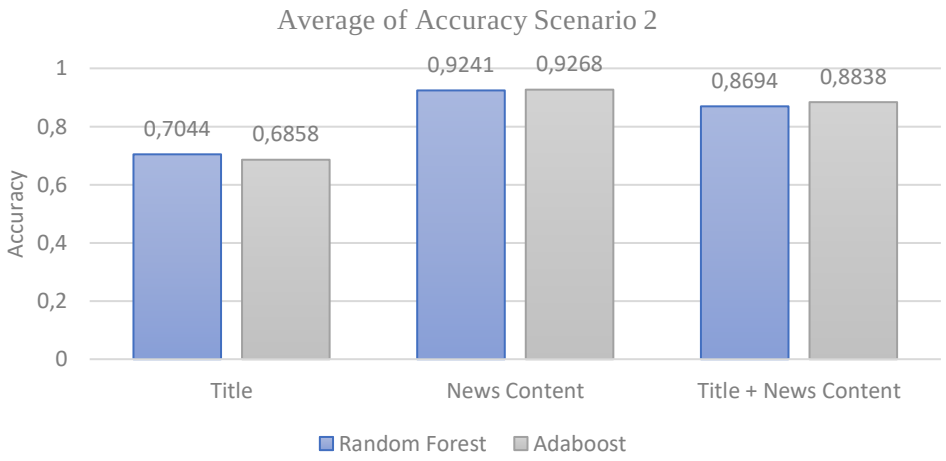

Figure 3. Comparison of Testing Results for Scenario 2 without Stop word Removal and Stemming

4. CONCLUSION

In this study, a classification system was built to compare two ensemble algorithms, Random Forest and AdaBoost, to detect hoax news related to the COVID-19 vaccine. Two scenarios were carried out, the first using full preprocessing and the second without using stopword removal and stemming. Each scenario uses three datasets, namely "Title", "News Content" and "Title + News Content". The experimental results show that all models produce the best accuracy using the "News Content" dataset. The first test scenario is obtained from the Random Forest algorithm with an accuracy value of 93.58% and the f1-score value of 89.65%. Compared to the AdaBoost algorithm, the prediction results are lower than those of Random Forest, where the accuracy value is 91.65%, and the f1-score is 86.70%. In addition, for testing the second scenario, the AdaBoost algorithm is slightly more accurate than the Random Forest algorithm. AdaBoost accuracy value is 92.68%, and the f1-score value is 88.42%. Besides, Random Forest has an accuracy of 92.41% and an f1 score of 87.98%. Based on the test analysis results, the ensemble method can provide good model results for detecting hoax news related to the COVID-19 vaccine case. For further research, it is hoped that they can try to classify hoax news using deep learning methods such as RNN to improve performance.

REFERENCES

[1] "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020." https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed Jul. 07, 2022).
[2] D. Cucinotta and M. Vanelli, "WHO Declares COVID-19 a Pandemic," Acta Biomed, vol. 91, pp. 157–160, 2020, doi: 10.23750/abm.v91i1.9397.
[3] "View of Behavior of Following Health Protocols (Keeping Distance, Washing Hands and Wearing Masks) as a Form of State Defense in the Era of COVID-19." https://cerdika.publikasiindonesia.id/index.php/cerdika/article/view/9/64 (accessed Jul. 07, 2022).
[4] K. Lutfiyah, "HOAX AND FAKE NEWS DURING COVID-19: IS THE LAW EFFECTIVE IN OVERCOMING IT?," 2020, doi: 10.15294/ijicle.v2i3.38422.
[5] M. D. Molina, S. S. Sundar, T. Le, and D. Lee, "'Fake News' Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content," American Behavioral Scientist, vol. 65, no. 2, pp. 180–212, Feb. 2021, doi: 10.1177/0002764219878224.
[6] A. Awalina, J. Fawaid, R. Yunus Krisnabayu, and N. Yudistira, "Indonesia's Fake News Detection using Transformer Network." [Online]. Available: https://github.com/JibranFawaid/turnbackhoax-dataset.
[7] H. A. Santoso, E. H. Rachmawanto, and U. Hidayati, "Fake Twitter Account Classification of Fake News Spreading Using Naïve Bayes," Scientific Journal of Informatics, vol. 7, no. 2, pp. 2407–7658, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji
[8] A. T. B. Panjaitan and D. I. Santoso, "Deteksi Hoaks Pada Berita Berbahasa Indonesia Seputar COVID-19," FORMAT: Jurnal Ilmiah Teknik Informatika, vol. 10, no. 1, pp. 76–85, 2021, [Online]. Available: https://turnbackhoax.id.
[9] S. D. Das, A. Basak, and S. Dutta, "A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection," Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.03545
[10] S. Ramani, M. S. Kumar, and A. Professor, "Identification of Fake News through SVM and Random Forest," 2020. [Online]. Available: http://ijesc.org/
[11] "Beranda | Covid19.go.id." https://covid19.go.id/ (accessed Jul. 14, 2022).
[12] N. Mahdi Abdulkareem and A. Mohsin Abdulazeez, "Machine Learning Classification Based on Radom Forest Algorithm: A Review," International Journal of Science and Business, vol. 5, no. 2, pp. 128–142, 2021, doi: 10.5281/zenodo.4471118.
[13] K. Li, G. Zhou, J. Zhai, F. Li, and M. Shao, "Improved PSO_AdaBoost ensemble algorithm for imbalanced data," Sensors (Switzerland), vol. 19, no. 6, Mar. 2019, doi: 10.3390/s19061476.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 839−846
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1957

[14] W. Hidayat, E. Utami, and A. Sunyoto, "Selection of the Best K-Gram Value on Modified Rabin-Karp Algorithm," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 16, no. 1, p. 11, Jan. 2022, doi: 10.22146/ijccs.63686.
[15] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information (Switzerland), vol. 10, no. 4. MDPI AG, 2019. doi: 10.3390/info10040150.
[16] M. Rusli et al., "EKSTRAKSI FITUR MENGGUNAKAN MODEL WORD2VEC PADA SENTIMENT ANALYSIS KOLOM KOMENTAR KUISIONER EVALUASI DOSEN OLEH MAHASISWA," vol. 07, no. 1, 2020.
[17] N. Nanda Widyastuti, A. Bijaksana, and I. Lukmana Sardi, "Analisis Word2vec untuk Perhitungan Kesamaan Semantik antar Kata," e-Proceeding of Engineering, vol. 5, no. 3, pp. 7603–7612, 2018.
[18] P. Compagnon and K. Ollivier, "Graph Embeddings for Social Network Analysis: State of the Art PhD Thesis-Toward unsupervised activity monitoring with sequence metric learning View project," 2017. [Online]. Available: https://www.researchgate.net/publication/331714802
[19] E. M. Dharma, F. Lumban Gaol, H. Leslie, H. S. Warnars, and B. Soewito, "THE ACCURACY COMPARISON AMONG WORD2VEC, GLOVE, AND FASTTEXT TOWARDS CONVOLUTION NEURAL NETWORK (CNN) TEXT CLASSIFICATION," J Theor Appl Inf Technol, vol. 31, no. 2, 2022, [Online]. Available: www.jatit.org
[20] R. Singh, "Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 3, pp. 3970–3976, Jun. 2020, doi: 10.30534/ijatcse/2020/221932020.
[21] B. Bahrawi, "SENTIMENT ANALYSIS USING RANDOM FOREST ALGORITHM ONLINE SOCIAL MEDIA BASED SENTIMENT ANALYSIS USING RANDOM FOREST ALGORITHM-ONLINE SOCIAL MEDIA BASED," 2019. [Online]. Available: https://www.researchgate.net/publication/338548518
[22] C. Albon, "Machine Learning with Python Cookbook Practical Solutions from Preprocessing to Deep Learning," 2019.
[23] L. Malem Ginting, M. M. Sigiro, G. Yola Lumbantoruan, and dan Januard Lumbangaol, "Prediksi Indikator Perawatan Menggunakan Random Forest Classification dan Classification and Regression Trees," 2022. [Online]. Available: http://journal-jati.del.ac.id/
[24] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification," 2020.