Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2024
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 6 (2024) 730 - 736 e-ISSN: 2580-0760

The Impact of Feature Extraction in Random Forest Classifier for Fake News Detection

Dhani Ariatmanto[1]*, Anggi Muhammad Rifai[2]
[1]Magister of Informatics Engineering, Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia
[2]Department of Informatics Engineering, Faculty of Engineering, Universitas Pelita Bangsa, Bekasi, Indonesia
[1]dhaniari@amikom.ac.id, [2]anggimuhammad@pelitabangsa.ac.id

Abstract

The pervasive issue of fake news spreading rapidly on online platforms. causing a concerning dissemination of misinformation. The influence of fake news has become a pressing social problem, shaping public opinion in important events such as elections. This research focuses on detecting and classifying fake news using the Random Forest algorithm by investigating the impact of feature extraction techniques on classification accuracy, this study specifically employs the TF-IDF method. For this purpose, we used 44,898 English-language articles from the ISOT fake news dataset. The dataset is cleaned using tokenization and stemming then split into 75% training and 25% testing. The TF-IDF vectorizer technique was applied to convert text into numeric as feature extraction. This study has implemented a Random Forest classifier to predict real and fake news. The proposed model contributes to overall classification precision by comparing it to the existing models. This fake news detection highlights the efficacy of the TF-IDF vectorizer and Random Forest combination which achieved an impressive accuracy rate of 99.0%. This contribution highlights an effective strategy for combating misinformation through precise text classification.

Keywords: *feature extraction; fake news; machine learning; Random Forest; text classification*

How to Cite: D. Ariatmanto and A. M. Rifai, "The Impact of Feature Extraction in Random Forest Classifier for Fake News Detection", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 6, pp. 730 - 736, Dec. 2024.
DOI: https://doi.org/10.29207/resti.v8i6.6017

1. Introduction

In an era dominated by digitalization, the rise of fake news has become a pressing social problem, spreading misinformation rapidly through various online platforms [1]. Apart from the direct impact of loss of trust, the harmful influence of fake news also extends to shaping public opinion and influencing important events such as elections [2]. In response to this problem, machine learning can be a solution for classifying and filtering fake news. Machine learning has many algorithms and one of them is random forest. The random forest process is carried out by combining decision trees with training on the dataset. Previous research has discussed several algorithm models that utilize various feature extraction and machine learning for fake news classification [3], [4], [5], [6], [7]. However, due to the limited availability of data sets and the imbalance between negative and positive classes, the performance of the evaluation results of accuracy values may still be improved.

Previous research has discussed the classification of fake news. [3], proposed the use of n-gram models and various feature extraction techniques to classify fake news. From the research results, it was found that the model evaluation performance reached an accuracy value of 92% respectively. Researchers used unigram features with Linear Support Vector Machine (LSVM) to produce the best accuracy performance. However, the results of this research may still be improved. Research conducted by the researcher [4] discussed the application of four machine learning techniques to identify fake reviews, namely: naïve Bayes, support vector machine (SVM), random forest (RF), and adaptive boost. For the feature extraction, it is used TF-IDF method. Their experimental results show that the RF technique produces superior performance with an accuracy of 95%, which is 3% better than [3]. Researchers [4] suggested the importance of cleaning datasets that had a significant effect on the results. Furthermore, researcher [5] proposed a classification model based on the BERT method by assembled learning and text sentiment analysis to identify

Received: 28-08-2024 | Accepted: 13-12-2024 | Published Online: 26-12-2024

dangerous news. These two methods aim to help readers recognize dangerous news and evaluate the neutrality of information. From the research results, the F1 score performance was 66.3%. Meanwhile, research [6] presents a new semi-supervised framework that combines Convolutional Neural Networks (CNN) with the concept of self-assembling. The research leverages linguistic and stylometric information from annotated news articles while exploring patterns in unlabeled data. The results of this research achieved an accuracy of 97.45% respectively. However, the dataset used only 50% of the labeled articles, the dataset came from Kaggle fake news dataset. In addition, the latest researcher [7] classified fake news based on the CNN method and feature extraction using the Easy Data Augmentation (EDA) technique. The dataset used in their research is 1000 Indonesian news articles. The EDA technique applied feature extraction such as random deletion, random insertion, random exchange, and synonymous replacement. Their experimental results achieved an accuracy of 82.61% respectively. However, their research may still be improved. Collectively, these existing studies highlighted the diversity of fake news classification approach by demonstrate the effectiveness of different algorithms and techniques. Although the result achieved high accuracy, the varying results underscore the influence of data set characteristics and algorithm in efforts to combat the spread of false information [8].

This research highlights its differences from previous research by significantly advances the field of fake news detection by underlining the importance of feature extraction. By investigating the impact of feature extraction can improve the classification accuracy [3]. Researchers not only underline the need to utilize Random Forest with appropriate feature extraction techniques, but also emphasize the potential social impact. Implementing updated fake news detection systems has the potential to effectively mitigate the spread of misinformation, safeguard information integrity and digital society [9]. Random forest provides good predictions on data and the ability to overcome overfitting [10]. This can be achieved using ensemble learning and randomization of variation in decision trees [11]. Random forest also provides an estimate of the importance of each feature in modeling that influences prediction results [12]. The combination of ensemble learning and decision trees proves that Random Forest is effective in a variety of modeling tasks. This invites a broader understanding of the evolving dynamics of information dissemination and ethical considerations surrounding the digital world [13]. By contributing diverse insights and innovative strategies, this research serves as a catalyst for ongoing efforts to fortify the digital landscape against the threat of fake news, thereby strengthening the foundations of trustworthy and resilient digital discourse.

Recognizing this impact, the need arises for efficient techniques to detect and counter fake news [4]. By investigating various kinds of fake news detection utilizing RF-based methods, the interaction of various feature extraction techniques influences the accuracy [14]. The main goal of the previous research is to find out the most effective technique in selecting relevant features, thereby improving the overall performance of the fake news detection system [8]. To achieve the objectives of our research, we used a carefully curated dataset, consisting of real and fake news articles, which served as a platform for training and testing RF classifiers. The classification performance assessment uses evaluation metrics such as accuracy, precision, recall, and F1-score. The comprehensive evaluation of the classification accurately differentiates between real news and fake news [6]. Some of experimental findings conclusively confirm the potential of Random Forest classification, especially when coupled with feature extraction methods, resulting in better accuracy rates [5]. The elaborately selected features capture the typical characteristics of fake news articles, thereby strengthening the overall classification performance [15].

Based on the problems described, this research is organized as follows. Second subsection present the research methods used. Third subsection discusses the result and analysis in classification method and discussion. The experimental result and suggestions for further research are shown in last subsection concludes the paper.

2. Research Methods

The research methodology employed in this study is quantitative, focusing on the analysis of numerical data to draw statistical inferences and conclusions. The dataset used in this research were hoax news and factual news from Information Security and Object Technology (ISOT) project at the University of Victoria [13]. Prior to analysis, the dataset undergoes preprocessing steps, including tokenization and stemming, to standardize and clean the data. The TF-IDF Vectorizer technique is then applied for feature extraction, enhancing the model's ability to identify significant terms within the dataset. TF-IDF can minimize noise and enhance the overall quality of data obtained from text data. the alternative approaches exhibit limitations in successfully capturing contextual information and introduce biases in the data as a result of their large model size and limited training data.This process contributes to the overall precision of the classification process by ensuring that relevant features are appropriately weighted.

Subsequently, the dataset is divided into training and test sets, with 75% of the data allocated for training and 25% for testing. The Random Forest classification algorithm is then employed to classify the news articles based on their content, enabling the proposed model to discern between Truthful news and hoax news. Following classification, an evaluation is conducted to

This is an open access article under the CC BY-4.0 license

measure the performance of the system using a confusion matrix.

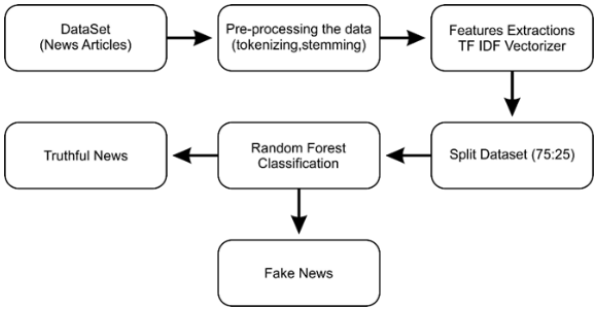
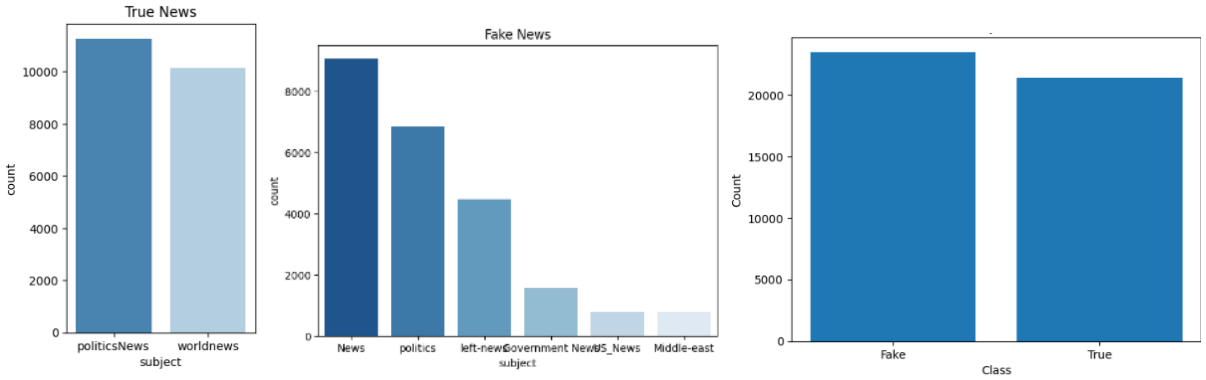


Figure 1. Research Flow

This evaluation provides insights into the accuracy and efficacy of the classification process, highlighting the proposed model's ability to correctly classify news articles as either genuine or hoax. The experimental result than compared with the existing model to the output of the research includes the classification results of the test data, indicating which news articles are classified as genuine and which are classified as hoax. The research flow system architecture is illustrated in Figure 1.

2.1 Dataset

The dataset consists of 44,898 English-language articles. There is a 21,417 dataset of valid news articles across two distinct topics: politics news and world news. There is a 23,481 dataset hoax news articles across five distinct topics: news, politics, left news, government news, and US news. The following graph compares the number of fake news and true news samples are shown in Figure 2. The dataset is specifically designed for fake news detection research and can be accessed through the following link [16].



(a) (b) (c)

Figure 2. Dataset Collection Graph (a) Fake: hoax news; (b) True: truthful news; (c) Total fake and true news

title	text	subject	date
Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017
Racist Alabama Cops Brutalize Black Boy While...	The number of cases of cops brutalizing and ki...	News	December 25, 2017
Fresh Off The Golf Course, Trump Lashes Out A...	Donald Trump spent a good portion of his day a...	News	December 23, 2017
Trump Said Some INSANELY Racist Stuff Inside ...	In the wake of yet another court decision that...	News	December 23, 2017
Former CIA Director Slams Trump Over UN Bully...	Many people have raised the alarm regarding th...	News	December 22, 2017
WATCH: Brand-New Pro-Trump Ad Features So Muc...	Just when you might have thought we d get a br...	News	December 21, 2017

Figure 3. Sample Dataset

The ISOT dataset provides a valuable resource for investigating the problem of fake news and evaluating the performance of different classification models [17]. It contains a diverse collection of news articles, both genuine and fake, which can be used for training and testing purposes. By utilizing this dataset, we aim to enhance our understanding of fake news detection and contribute to the development of effective detection techniques. The availability of such datasets is crucial for advancing research in this area and developing robust solutions to mitigate the spread of misinformation in online environments, a sample of the available datasets can be seen in Figure 3.

2.2 Pre processing

Tokenizing and stemming methods are widely used in text processing, natural language processing, and text analysis. In text processing, tokenization is used to separate words or other units in text, while stemming is used to change words to their basic form so that they can be processed efficiently [3]. Both of these techniques help in preparing the text for further analysis, such as text classification, topic modeling or information extraction show in Figure 4.

This is an open access article under the CC BY-4.0 license

	text_before
44873	the credibility of beverly young nelson the w...
44874	washington aboard air force one reuters pr...
44875	william shatner is such a smarty pants on twit...
44876	washington reuters u s president donald t...
44877	holy smokes that s billion with a b where di...

(a)

text_after
the credibility of beverly young nelson the w...
washington aboard air force one reuters pr...
william shatner is such a smarty pants on twit...
washington reuters u s president donald t...
holy smokes that s billion with a b where di...

(b)

Figure 3. Pre-Processing (a) Before; (b) After

2.3 Feature Extraction

This method converts text into a numeric representation suitable for machine learning models. The TF-IDF weighting scheme combines Term Frequency (TF), reflecting word frequency within a document, and Inverse Document Frequency (IDF), indicating how unique or common a word is across the entire collection of documents [3]. TF-IDF was selected over embeddings like Word2Vec or FastText because it captures the significance of each term within a specific context, making it well-suited for tasks where document-specific term importance is crucial [18]. Unlike Word2Vec or FastText, which focus more on semantic relationships between words, TF-IDF offers a straightforward approach to emphasize discriminative terms, which can improve classification performance in contexts where understanding document uniqueness is key. The TF-IDF Vectorizer calculates the value for each word, generating a numeric vector that represents the document [19], with each vector component indicating the term's importance within that document [15].

2.4 Random Forest Classification

The Random Forest method is a method in machine learning that is used to perform classification and regression [20]. This method is based on the concept of ensemble learning, in which several decision tree models are combined to produce the final prediction. The steps in using the Random Forest model include[10]:

Formation of a number of decision trees formed using bootstrap aggregating (bagging) techniques. Bagging selects a number of random samples with replacement from the original dataset to form a training set for each decision tree [21]. Each decision tree will have the variance obtained from this random sample.

Each decision tree is built by dividing the dataset into smaller subsets based on existing features [22]. This process is carried out recursively until it reaches a stopping condition, such as reaching the maximum number of nodes or reaching the maximum depth.

After all decision trees have been built, predictions from each decision tree are taken. In classification, majority voting is used to determine the final prediction, whereas in regression, the average prediction of each decision tree is taken as the final prediction [23].

Finally, the performance of the Random Forest model is evaluated using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score. This helps in measuring the extent to which the model can classify or regress correctly [24].

The steps for the Random Forest algorithm begin with selecting random data points from the training data. This involves randomly selecting K data points with replacement, ensuring that each subset may contain duplicate samples from the original dataset. Next, a decision tree is constructed using these K selected data points. This tree is built by recursively splitting the data based on a subset of features chosen at each node, optimizing for the best separation according to a criterion like Gini impurity for classification or mean squared error for regression. The Gini impurity G for a node is calculated by Equation 1.

$$G = 1 - \sum_{i=1}^{C} P_i^2 \quad (1)$$

P_i is the probability of a data point belonging to class *i* and *C* is the total number of classes (1). Before repeating the selection and tree construction steps, the number of trees (*NTree*) to be created must be determined. This number, a hyperparameter, influences the ensemble's overall performance and robustness. Once the number of trees is set, steps one and two are repeated *NTree* times to build a forest of decision trees. For making predictions, each of the *NTree* decision trees provides a prediction for a new data point. In classification tasks, the final prediction $\hat{y}$ for a data point *x* is determined by majority voting among the trees. If $T_i(x)$ is the prediction of the *i*-th tree for data point *x*, the final prediction is formulated by Equation 2.

$$\hat{y} = mode\{T_1(x), T_2(x), \dots, T_{NTree}(x)\} \quad (2)$$

In regression tasks, the final prediction $\hat{y}$ is the average of all the predicted values from the trees which is defined by Equation 3.

$$\hat{y} = \frac{1}{NTree} \sum_{i=1}^{NTree} T_i(x) \quad (3)$$

This ensemble approach helps improve the accuracy and stability of the model's predictions.

3. Result and Discussion

After the dataset is determined, the researcher carries out random Shuffling Frame Data, which serves to randomize the order of data frames, ensuring that the

This is an open access article under the CC BY-4.0 license

model does not learn any unintended patterns from the sequence of the data. This step is crucial to prevent any potential bias and improve the generalization of the model. The process can be seen in Figure 5.

	text	class
12026		0
15307	Hillary in prison orange kinda like an early C...	0
17128	Don t you just love the absolute boldness of t...	0
10659	Judge Jeanine commented this morning on the FB...	0
15737	The radical Baltimore mayor who ordered the Ba...	0

(a)

	text	class
0		0
1	Hillary in prison orange kinda like an early C...	0
2	Don t you just love the absolute boldness of t...	0
3	Judge Jeanine commented this morning on the FB...	0
4	The radical Baltimore mayor who ordered the Ba...	0

(b)

Figure 5. (a) Before Random Shuffling Frame Data; (b) After Random Shuffing Frame Data

The next step is tokenizing process, it's break down the text into smaller units called tokens. The tokenizing process begins with a function such as `re.sub("¥¥W", " ", text)`, which replaces non-alphanumeric characters with spaces. Tokenization is essential for breaking down the text into individual words or tokens that the model can process. For stemming, the process involved `text.lower()`, which converts all the text to lowercase. While not a formal stemming algorithm, this is a simple way to reduce word variations by treating uppercase and lowercase letters equally. However, no specific stemming algorithms, such as the Porter Stemmer or Snowball Stemmer, were used in this step. Instead, the `wordopt` function simplifies the text without further morphological reduction. Before and after tokenizing and stemming show in Figure 6.

	text_before
0	washington reuters while attention in asia...
1	washington reuters the top republican tax ...
2	washington reuters the u s supreme court ...
3	tegucigalpa reuters eight years after a co...
4	native new yorker and republican front runner ...

(a)

text_after
washington reuters while attention in asia...
washington reuters the top republican tax ...
washington reuters the u s supreme court ...
tegucigalpa reuters eight years after a co...
native new yorker and republican front runner ...

(b)

Figure 6. (a) Before tokenizing and stemming; (b) After tokenizing and stemming

We proceed to the Feature Extractions stage with the TF-IDF Vectorizer which aims to convert text into a numerical representation that can be used in machine learning models. Furthermore, we divided the datasets by 75% for training and 25% for random testing. The random forest algorithm classified fake news to find out whether the model can recognize and distinguish the fake news appropriately and accurately.

3.1 Evaluation Performance Matrix

The evaluation performance is essential to determine the model's results accurately. We evaluated the Random Forest algorithm using a confusion matrix, as it provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, offering a more comprehensive understanding of the model's performance on individual classification metrics. Unlike cross-validation, which provides an average accuracy across multiple folds, the confusion matrix allows us to pinpoint the model's strengths and weaknesses in distinguishing between fake news and truthful news. The ability to identify and classify news effectively can be observed in Figure 7.

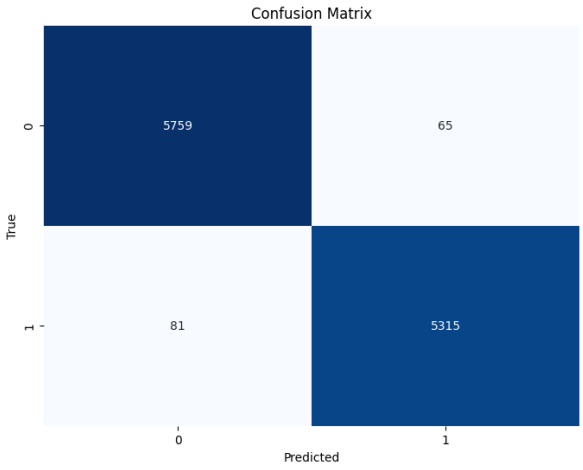


Figure 7. Confusion Matrik, 0 = fake news; 1= real news

From Figure 7, It shows that 5759 fake news test datasets were correctly identified as fake news and 81 data was misidentified as fake news. For the real news test dataset, 5315 real news were correctly identified as real news and 65 data was misidentified as real news.

To investigate the classify performance, we used four metrics evaluation based on the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) in the predictions of the binary classifiers:

Accuracy, which is the percentage of True (i.e. correct) predictions.

Recall, which captures the ability of the classifier to find all the positive samples.

Precision, which is the ability of the classifier not to label a negative sample positive.

This is an open access article under the CC BY-4.0 license

The score, which is the harmonic mean of precision and recall, computes values in the range [0,1].

Equations 4-7 compute the metrics:

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

$$precision = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$recall = \frac{TP}{TP+FN} \quad (6)$$

$$F_1 score = \frac{2^*(precision \times recall)}{precission+recall} \quad (7)$$

The proposed model result performance metrics of the Random Forest model is listed in Table 1.

Table 1. Results of Random Forest model

Metrics	Result
Accuracy	0.99
Precision	0.99
Recall	0.99
F1 Score	0.99

Referring to Table 1, the proposed Random Forest model with TF-IDF vectorizer can achieved accuracy of 0.99 respectively. This indicated that the proposed model show high ability and effectiveness in term of fake news classification.

3.2 Comparison Result

In order to get a better understanding of the result, our proposed Random Forest model based on TF-IDF vectorizer is compared with the other existing models. The existing models used implemented different extraction features than the proposed model as seen in Table 2.

Table 2. Results of Random Forest model

Author	Feature	Model	Accuracy
[3]	TF-IDF	LSV	0.9200
[25]	TF-IDF	SVM	0.9505
[12]	TF-IDF	BI-LSTM	0.9151
Ours	TF-IDF Vectorizer	RF	0.9900

Referring to Table 2, the proposed model produces higher accuracy value compared with models by Ahmed et.al [3], Shaik and Patil [25] and Sharma et.al [12]. The visual comparison accuracy are shown in Figure 8.

The Proposed model out performed existing models by Shaik and Patil [25] in term of accuracy by 0.04 respectively. The result demonstrated that the proposed RF model able to classified good test accuracy of fake news under 25% datasets.

The proposed model applied TF-IDF vectorizer to convert text into a numerical representation. Before training the model, the dataset was pre-processed by feature class target, cleansing data, random shuffling frame data and splitting data. The result obtained from the training model is accuracy and confusion matrix.

Classified fake news using the impact of feature extraction TF-IDF vectorizer has better performance compared to the existing models. This show that the proposed model is a suitable for fake news classification. Furthermore, using different feature extraction techniques, like word embeddings, could improve the model's resilience and avoid overfitting.

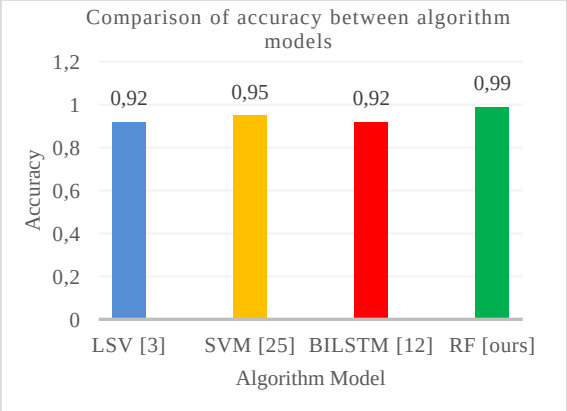


Figure 8. Comparison of accuracy value between models by Ahmed et.al [3], Shaik and Patil [25], Sherma et.al [12] and Proposed model.

4. Conclusion

This research presented Random Forest algorithm for fake news classification. A dataset was obtained from ISOT dataset. The proposed model based on the impact of TF-IDF vectorizer. In this model, tokenizing and stemming are used to separate and change words to basic form. Pre-processed is carried out by feature class target, cleansing data, random shuffling frame data and splitting data. The result demonstrate that the proposed model produces high accuracy value of classification fake news than the existing schemes. The result show that our model maintains the accuracy with a quality value of 0.99.

Acknowledgements

This work was supported by research Grant from Department of Research and Community Service, Universitas Amikom Yogyakarta, for supporting and funding this research. The author declares no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

References

[1] S. Mishra, P. Shukla, and R. Agarwal, "Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets," *Wirel. Commun. Mob. Comput.*, vol. 2022, Mar. 2022, doi: 10.1155/2022/1575365.

[2] A. Merryton and M. G. Augasta, "An Attribute-wise Attention model with BiLSTM for an efficient Fake News Detection," *Multimed. Tools Appl.*, vol. 83, pp. 1–18, Oct. 2023, doi: 10.1007/s11042-023-16824-6.

[3] H. Ahmed, I. Traore, and S. Saad, *Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques*. 2017. doi: 10.1007/978-3-319-69155-8_9.

[4] S. Alsubari *et al.*, "Data Analytics for the Identification of Fake

This is an open access article under the CC BY-4.0 license

Reviews Using Supervised Learning," *Comput. Mater. Contin.*, vol. 70, Sep. 2021, doi: 10.32604/cmc.2022.019625.

[5] S.-Y. Lin, Y.-C. Kung, and F.-Y. Leu, "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis," *Inf. Process. Manag.*, vol. 59, p. 102872, Mar. 2022, doi: 10.1016/j.ipm.2022.102872.

[6] P. Meel and D. Vishwakarma, "A Temporal Ensembling based Semi-supervised ConvNet for the Detection of Fake News Articles," *Expert Syst. Appl.*, vol. 177, p. 115002, Apr. 2021, doi: 10.1016/j.eswa.2021.115002.

[7] A. Noor, R. Gernowo, and O. Nurhayati, "Data Augmentation for Hoax Detection through the Method of Convolutional Neural Network in Indonesian News," *J. Penelit. Pendidik. IPA*, vol. 9, pp. 5078–5084, Jul. 2023, doi: 10.29303/jppipa.v9i7.4214.

[8] S. S. Dhanyal and S. Nandyal, "An Effective Machine Learning-based Segmentation and Feature Extraction Technique for Muscular-Disorder," in *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, 2023, pp. 475–482. doi: 10.1109/ICPCSN58827.2023.00083.

[9] M. J. Awan *et al.*, "Fake News Data Exploration and Analytics," *Electronics*, vol. 10, no. 19. 2021. doi: 10.3390/electronics10192326.

[10] M. Aufar, R. Andreswari, and D. Pramesti, "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," in *2020 International Conference on Data Science and Its Applications (ICoDSA)*, 2020, pp. 1–7. doi: 10.1109/ICoDSA50139.2020.9213078.

[11] P. T. Putra, A. Anggrawan, and H. Hairani, "Comparison of Machine Learning Methods for Classifying User Satisfaction Opinions of the PeduliLindungi Application," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 3 SE-Articles, Jun. 2023, doi: https://doi.org/10.30812/matrik.v22i3.2860.

[12] M. Ula, V. Ilhadi, and Z. Sidek, "Comparing Long Short-Term Memory and Random Forest Accuracy for Bitcoin Price Forecasting," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, pp. 259–272, Jan. 2024, doi: 10.30812/matrik.v23i2.3267.

[13] I. Ennejjai, A. Ariss, N. Kharmoum, W. Rhalem, S. Ziti, and M. Ezziyyani, "Artificial Intelligence for Fake News BT - International Conference on Advanced Intelligent Systems for Sustainable Development," J. Kacprzyk, M. Ezziyyani, and V. E. Balas, Eds., Cham: Springer Nature Switzerland, 2023, pp. 77–91.

[14] M. Alkaff, M. Miqdad, M. Fachrurrazi, M. Abdi, A. Abidin, and R. Amalia, "Hate Speech Detection for Banjarese Languages on Instagram Using Machine Learning Methods," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 3 SE-Articles, Jul. 2023, doi: https://doi.org/10.30812/matrik.v22i3.2939.

[15] D. K. Sharma, S. Garg, and P. Shrivastava, "Evaluation of Tools and Extension for Fake News Detection," in *2021 International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 2021, pp. 227–232. doi: 10.1109/ICIPTM52218.2021.9388356.

[16] University of Victoria, "The ISOT Fake News dataset is a compilation of several thousands fake news and truthful articles," Canada, 2022. [Online]. Available: https://onlineacademiccommunity.uvic.ca/isot/?utm_medium=redirect&utm_source=%2Fdatasets%2Ffake-news%2Findex.php&utm_campaign=redirect-usage

[17] K. Muthuvelu, S. Rajkumar, and R. Bhuvanya, "Fake News Detection Using Machine Learning Algorithms," 2022, pp. 181–207. doi: 10.1002/9781119763499.ch10.

[18] S. Khomsah, "Sentiment Analysis On YouTube Comments Using Word2Vec and Random Forest," *Telematika*, vol. 18, p. 61, Mar. 2021, doi: 10.31315/telematika.v18i1.4493.

[19] P. Karthika, R. Murugeswari, and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," in *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2019, pp. 1–5. doi: 10.1109/INCOS45849.2019.8951367.

[20] M. Azizah, A. Yanuar, and F. Firdayani, "Dimensional Reduction of QSAR Features Using a Machine Learning Approach on the SARS-Cov-2 Inhibitor Database," *J. Penelit. Pendidik. IPA*, vol. 8, no. 6 SE-Research Articles, pp. 3095–3101, Dec. 2022, doi: 10.29303/jppipa.v8i6.2432.

[21] S. Khomsah, R. D. Ramadhani, and S. Wijaya, "The Accuracy Comparison Between Word2Vec and FastText On Sentiment Analysis of Hotel Reviews," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 3 SE-Information Systems Engineering Articles, Jun. 2022, doi: 10.29207/resti.v6i3.3711.

[22] A. Setiawan, E. Utami, and D. Ariatmanto, "Cattle Weight Estimation Using Linear Regression and Random Forest Regressor," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 8, no. 1 SE-Information Systems Engineering Articles, Feb. 2024, doi: 10.29207/resti.v8i1.5494.

[23] E. Fitri, "Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine," *J. Transform.*, vol. 18, p. 71, Jul. 2020, doi: 10.26623/transformatika.v18i1.2317.

[24] B. Bahrawi, "SENTIMENT ANALYSIS USING RANDOM FOREST ALGORITHM ONLINE SOCIAL MEDIA BASED," vol. 2, p. h.29-33, Dec. 2019.

This is an open access article under the CC BY-4.0 license