CENTIVE
Conference on Electrical Engineering, Informatics
, Industrial Technology, and Creative Media
ISBN : 978-602-53004-0-0

Optimization of Random Forest Model via GridSearchCV for Hoax News Detection

Lutvi Riyandari[1*], Singgih Setia Andiko[2], Siti Delimasari[3], and Singgih Briandoko[4]
[1,2,3]STMIK Widya Utama
[1,2,3]Jl. Sunan Kalijaga, Purwokerto Selatan Banyumas, Jawa Tengah, 53146, Indonesia
[*]lutvi@swu.ac.id

Abstract — In this time of fast digital information growth, information sources can be helpful or harmful. The internet makes it easier for people to find information, but it also makes it easier for fake news and hoaxes to spread quickly and widely. This work seeks to combat the dissemination of false news in the digital age by employing text categorization through the Random Forest algorithm, coupled with hyperparameter optimization via Grid SearchCV.The dataset comprises both hoax and authentic news from Indonesia, subjected to various steps including text processing (case folding, tokenization, and stopword elimination) and feature weighting via TF-IDF.The study's results reveal that the Random Forest model does an impressive job of telling the difference between fake and real news when tested using a confusion matrix. The confusion matrix shows that the model works better after hyperparameter tweaking with GridSearchCV. This is shown by the fact that the number of accurate predictions (TN and TP) goes up and the number of wrong predictions (FP and FN) goes down. The evaluation measures (accuracy, recall, precision, and F1-Score) also demonstrate significant improvements, increasing from 96% to 97%.

Keywords – Hoax News, Text Classification, Random Forest, GridSearchCV

I. INTRODUCTION

The internet enables public access to knowledge; nevertheless, it also serves as a medium for the swift and extensive propagation of misinformation and hoaxes. A hoax is information intentionally created to obscure the actual truth. A hoax can be defined as an attempt to misrepresent facts using seemingly credible information that lacks verification. Additionally, a hoax can be characterized as the act of hiding facts by flooding the media with fake material to obscure the authentic message [1].

Fake news can significantly affect society by influencing election outcomes [2], inciting social tensions, and tarnishing the reputations of certain individuals or groups [3]. Indonesia, a nation with one of the highest populations of internet users in Southeast Asia, is also susceptible to the adverse effects of misinformation [4].

Fake news can upset social, cultural, political, and economic stability. Numerous companies have experienced stock losses that were attributed to the dissemination of falsehoods about them. Research conducted by the Indonesian Telematics Society indicates that 44.3% of the respondents acknowledged that they often saw false news or misinformation on a daily basis. The research by Newman et al. revealed that more than half (55%) of the online news articles in their global sample expressed concerns about the prevalence of false news. One method of minimizing the dissemination of false information is to create an automated system that is competent in its classification. A news report can be categorized as true or false throughout the categorization process.

Numerous prior research studies have sought to tackle the issue of misinformation. However, there is an opportunity to enhance our understanding of this issue and devise more effective strategies for the classification of fake news. This project aims to help people better understand the effects of fake news within the informatics domain and provide more effective remedies.

Text classification involves the allocation of categorical categories to documents or text according to their content. This approach entails the construction of a machine learning model capable of identifying patterns in text to facilitate category determinations. Classification denotes the procedure for categorizing

documents into one or more predetermined groups or classes based on shared characteristics [5].

The machine learning process, especially the classification, generally involves evaluating multiple models in a dataset and subsequently choosing the model with optimal performance. Choosing a model with precise predictions can be accomplished by several methods, one of which is hyperparameter adjustment, a deep learning technique that enhances the efficacy of predictive models [6]. Enhancing the efficacy of machine learning systems requires hyperparameter optimization. Altering a single hyperparameter and assessing its effect on model performance is inefficient and does not ensure optimal outcomes, as it overlooks the interactions among hyperparameters. Consequently, grid search methods are commonly employed as a solution [7].

This challenge involves applying text classification to differentiate between news that conveys accurate information (non-hoax news) and news that disseminates inaccurate or misleading information (hoax news). The Random Forest technique, along with Grid Search hyperparameter optimization, will serve as the principal method to enhance the accuracy in text classification. This algorithm will facilitate the automatic and efficient classification of fraudulent news, thus aiding in the preservation of information authenticity in the contemporary digital age. This research also seeks to offer tangible advantages to the public in the fight against the proliferation of misinformation.

II. LITERATURE REVIEW

The Random Forest algorithm is an ensemble learning technique utilized for classification and regression tasks. Ensemble learning entails the aggregation of outputs from multiple machine learning models to enhance performance and forecast accuracy. Random Forest achieves these results by aggregating a substantial amount of randomly generated decision trees [8].

Classification is the process of creating a model designed to predict the category or class label of an object. This procedure relies on data from a collection of previously labeled or classed instances [9].

Machine learning is a domain of computer science that is dedicated to creating techniques and algorithms that allow computers to learn from data and experience, independent of explicit programming [10].

Natural Language Processing (NLP) is a subdivision of artificial intelligence that focuses on the interaction between computers and human language. In this domain, scientists and engineers create technologies and algorithms that allow computers to comprehend, process, and produce human language in a manner that machines can interpret and handle [11]. The study [12] examines the identification of fraudulent news in English using the Bernoulli Naive Bayes algorithm and feature extraction via TF-IDF.

This study aimed to assess the efficacy of the algorithm in detecting fraudulent news. The implementation findings indicated that the prediction model, constructed with 8,800 news data sets, achieved an accuracy rate of 98.5% of a total of 2,200 test data sets. The outcomes demonstrated an accuracy of 97.8% for the ‘false’ label and 99.1% for the ‘real’ label. Additionally, this model achieved a precision of 99.1%, a recall of 97.8%, and an F1-Score of 98.4%.

Subsequently, the study [13] examined the detection of news validation on Twitter using the Naive Bayes algorithm. The trials conducted revealed that the Naive Bayes algorithm is effective in classifying fake news using Twitter social media test data, obtaining a maximum accuracy of 92% on a test set of 309 elements.

Subsequent research [14] examines the detection of false news through the Random Forest algorithm. This study aimed to determine the efficacy of the algorithm in distinguishing between false news and authentic news. This study utilized a data set that encompassed many news genres, including both fake and authentic news, which had been examined and classified by truth specialists. The results showed that the Random Forest algorithm accurately classified news. The findings indicated that this system could differentiate between false news and authentic news with an average accuracy exceeding 87%.

Research [15] addressed the issue of misinformation regarding health. This work aimed to establish and refine a system capable of detecting hoaxes in Indonesian-language news, specifically concerning the health sector, utilizing the Naive Bayes classification method. This approach utilizes data obtained from the crawling results as training data. This system was developed using the CI 4 language and the Waterfall methodology, achieving an accuracy of 90.9%.

Paper [16] examines the identification of bogus news with the Multinomial Naive Bayes approach. This study utilized data obtained using the Twitter API data crawling method, which covered 500 preprocessed and labeled “News” searches, which were subsequently used for training and testing purposes. The model derived by this technique achieved an accuracy of 83%, while the accuracy of the training set was 94%.

The paper [17] examines the efficacy of the BERT text classification algorithm to identify false news. This research utilized data from three publicly available datasets obtained from websites including “onlineacademiccommunity.uvic.ca,” “HuggingFace.co,” and “Kaggle.com.” The dataset was partitioned into three segments: training (70%), validation (15%), and testing (15%). The BERT model was initialized, the AdamW

optimizer was employed with NLLLoss, and the model underwent training over many epochs. The model was assessed with the validation data to avoid overfitting. The model was employed to categorize the test data. The evaluation results indicated that the BERT classification model achieved an accuracy of 76% on the validation data set to classify hoax news, illustrating the efficacy of machine learning methods in this domain. Simultaneously, the BERT Multilingual categorization model had a diminished accuracy of 63%.

A research [18] examines the optimization of Random Forest and Support Vector Machine using GridSearchCV hyperparameters for sentiment analysis of reviews related to the "PrimaKu" application. The dataset comprised 2,293 of the most pertinent reviews obtained from the Google Play Store. We identified the optimal models for Random Forest and Support Vector Machine by fine-tuning hyperparameters with GridSearchCV. The results of this study demonstrate that Random Forest achieves a higher classification success rate for PrimaKu user reviews, with an accuracy of 89%, a precision of 88%, a recall of 81% and an F1-score of 85%.

Based on the findings of many studies, this study expands on the use of the Random Forest model and optimizes its use of GridSearchCV to detect hoax news and assess the accuracy of the results.

III. RESEARCH METHOD

This research uses an experimental method. This study aims to investigate the application of Random Forest with hyperparameter configurations via the GridSearch method.

Fig. 1. Research stage model.

A. Data Collection

The dataset employed in this study pertains to Indonesian fake news. The dataset utilized is sourced from the Kaggle repository, specifically retrieved from https://www.kaggle.com/datasets/vijayandika/hoax-news-indonesia. This dataset has two files: the true data, named factual news detik.csv, which includes 2160 samples, and the false data, called fake news turnbackhoax.csv, which contains 2152 samples.

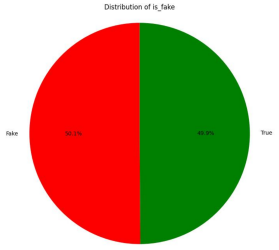

Fig. 2. Percentage of fake and true data.

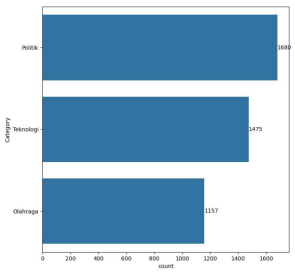

Fig. 3. Count of news articles by category.

Figure 3 illustrates three news categories: Politics, technology and Sports, with respective counts of 1680, 1475 and 1157.

B. Text Processing

To produce accurate and relevant data, the text processing procedure requires adherence to various methodological phases. The subsequent steps undertaken are as follows:

- Case folding is a technique employed to convert words from uppercase to lowercase [19]. This strategy also serves to eliminate punctuation [20]. Strategic Foresight for Presidential Aspirants. Presidential and Vice Presidential candidates must demonstrate strategic insight. Presidential and Vice Presidential candidates must demonstrate foresight.
- Tokenization is the process of dividing sentences into words and removing specific characters from words [19]. In this stage, punctuation, numerals, and symbols that are not letters of the alphabet are typically eliminated, as they are considered word separators (delimiters). Nevertheless, in certain instances, the numerals are not removed because they are still considered significant [20].
- Stopword is the process of eliminating words that are commonly used but lack significant meaning or information [20]. We employ a stopword list at this juncture, which includes words that lack significant meaning, including "yang," "di," "ke," "adalah," "akhir," "apabila," and "di." This phase enables us to acquire data that are more detailed and precise.

C. TF-IDF

TF-IDF, or Term frequency—inverse Document Frequency, is the most widely used weighting method and is known to provide satisfactory results [21]. This method relates the frequency of word occurrences (Term Frequency, TF) to the inverse document frequency (Inverse Document Frequency, IDF) [22]. As one of the feature extraction techniques, TF-IDF is used to determine feature vectors that are influential in creating class diversity [21]. In the context of TF, this term refers to the emphasis placed on words that appear frequently in the text, where the more often a word appears, the higher its value. On the other hand, the IDF reflects the inverse of the frequency of word

occurrences across documents; the more documents that contain the word, the lower its IDF value. The study by Wahyuni et al. [23] provides the formula to calculate TF-IDF.

$$\mathrm{TF}_i = \frac{N}{\text{word count in D}} \tag{1}$$

$$\mathrm{IDF}_i = \log(\text{number D/dfi}) \tag{2}$$

$$W_{ij} = \mathrm{TF_{ij}} \times \mathrm{IDF_i} \tag{3}$$

where D represents the document, N represents the frequency term i in D, dfi is the count of documents containing the term i, and W_{ij} is the weight of the term i in the document j.

D. Hyperparameter Tuning

The hyperparameter adjustment is essential to improve the performance of any machine learning system. Hyperparameter values cannot be directly derived from the data; rather, they are established during the model definition phase. Hyperparameters must be established prior to the initiation of the model's learning process. Hyperparameters are variables that affect a model's output. In the k-nearest neighbors (K-NN) model, the hyperparameter k, representing the number of nearest neighbors, can produce varying outputs despite identical inputs, such as k = 1 and k = 5. This study used the Grid Search method to identify the ideal hyperparameter values for the model. Grid Search Cross-Validation denotes a methodology that integrates Grid Search and Cross-Validation methodologies, wherein this approach identifies a model and hyperparameter combination by systematically evaluating and validating each individual combination. Figure 4 shows the hyperparameter optimization procedure with cross-validation of the grid search.

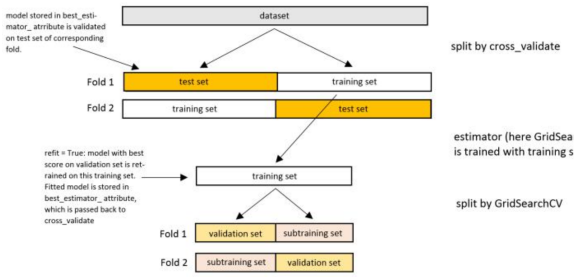

Fig. 4. Ilustrasi GridSearchCV.

E. Evaluation Metrics

The assessment of the classification technique in this study was performed using a confusion matrix. This matrix allows for the calculation of many metrics, including accuracy, precision, recall, and F1-Score. The formula used is delineated in equations (4)–(7) as follows [15].

$$\text{Accuracy} = \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN})} \times 100\% \tag{4}$$

$$\text{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \times 100\% \tag{5}$$

$$\text{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \times 100\% \tag{6}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision}+\text{Recall})} \times 100\% \tag{7}$$

IV. RESULT

This section discusses the research findings and analysis. The chaos matrix is an essential tool for assessing the effectiveness of a classification model. This matrix includes four main components. True Negatives (TN) indicate the number of cases that are accurately classified as negative, while True Positives (TP) represent the cases that are correctly identified as positive. False Negatives (FN) refer to cases that are genuinely positive but are misclassified as negative, and False Positives (FP) refer to cases that are actually negative but are misidentified as positive.

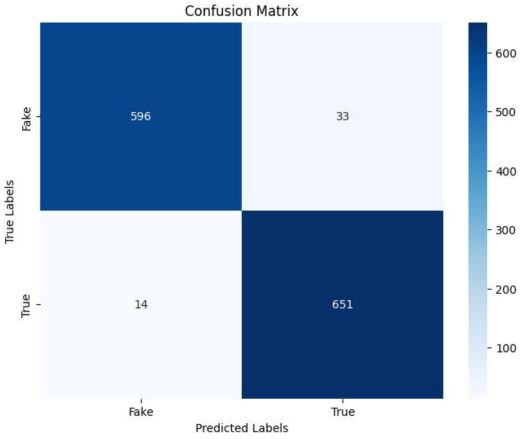

Fig. 5. Confusion matrix Random Forest.

In Figure 5, the model demonstrated 596 true negative predictions (TN), 651 true positive predictions (TP), 33 false positive predictions (FP) and 14 false negative predictions (FN).

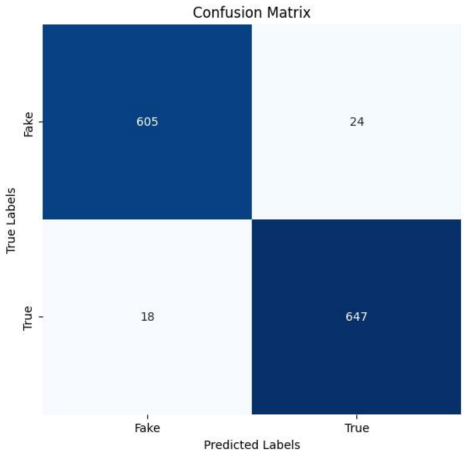

Fig. 6. Confusion matrix Random Forest with hyperparameter GridSearchCV.

Figure 6 indicates that the model produces 605 true negative predictions (TN), 647 true positive predictions

(TP), 24 false positive predictions (FP), and 18 false negative predictions (FN).

Fig. 7. Word cloud for fake news.

A word cloud for hoax news visually represents keywords that commonly occur in the context of misinformation. This tool underscores the proliferation of disinformation and stresses the necessity of verifying news prior to acceptance.

Fig. 8. Word cloud for real news.

The Word Cloud for Authentic News visually represents keywords that commonly occur in trustworthy news narratives. It underscores the need for the precision and dependability of information in news reporting.

Table 1. Comparison of Random Forests evaluation values before and after hyperparameter tuning GridSearchCV

Feature Selection	Random Forest			
	Accuracy	Recall	Precision	F1-Score
Before	96%	96%	96%	96%
After	97%	97%	97%	97%

Table 1 compares the assessment scores of the prediction model based on various metrics, including accuracy, recall, precision, and the F1-Score, both before and after the feature selection process. The results indicate a significant improvement in the model's assessment scores after feature selection. The model's accuracy increased from 96% to 97%. Additionally, recall, precision, and F1-Score also showed positive changes, each increasing from 96% to 97%. Therefore, it can be concluded that the hyperparameter tuning performed using GridSearchCV has proven effective in improving the performance of the predictive model.

V. Conclusions and Recommendation

A. Conclusion

The study's results indicated that the Random Forest model had better performance in distinguishing between hoax and legitimate news, as assessed using a confusion matrix, word cloud, and comparative evaluation metrics. The confusion matrix indicates enhanced model performance after hyperparameter tweaking with GridSearchCV, evidenced by an increase in correct predictions (TN and TP) and a reduction in wrong predictions (FP and FN). The word cloud visually represented commonly occurring keywords in both hoax and real news, highlighting the importance of information accuracy. All assessment measures (accuracy, recall, precision, and F1 score) exhibited notable enhancements, increasing from 96% to 97% after tuning, thus demonstrating the efficacy of feature selection and hyperparameter optimization in augmenting the prediction model's performance.

B. Recommendation

This study recommends employing alternative modeling techniques, such as BERT or deep learning, to improve the accuracy in identifying false news. Moreover, employing a more varied dataset that includes different languages can enhance the model's generalization capacity. Incorporating characteristics such as metadata or sentiment analysis can enhance the model's input. Evaluating models in real-world settings is essential to determine their efficacy in addressing intricate hoax news. In addition, initiatives aimed at improving digital literacy among the public must be intensified to cultivate awareness and the capacity to authenticate information, thus reducing the dissemination of false news.

References

[1] G. Gumilar, "LITERASI MEDIA: CERDAS MENGGUNAKAN MEDIA SOSIAL DALAM MENANGGULANGI BERITA PALSU (HOAX) OLEH SISWA SMA." https://www.semanticscholar.org/paper/LITERASI-MEDIA%3A-CERDAS-MENGGUNAKAN-MEDIA-SOSIAL-SMA-Gumilar/2a02d3c3ff46f7234c3870cad039a100ccf287a1, 2017. [Accessed 29-01-2026].

[2] A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on facebook," *Sci. Adv.*, vol. 5, p. eaau4586, Jan. 2019.

[3] G. Pennycook and D. G. Rand, "Assessing the effect of 'disputed' warnings and source salience on perceptions of fake news accuracy," *SSRN Electron. J.*, 2017.

[4] S. Kemp, "Digital 2021: the latest insights into the 'state of digital." https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/, 2024. [Accessed 29-01-2026].

[5] D. Sarkar, *Text analytics with python*. Berlin, Germany: APress, 1 ed., Dec. 2016.

[6] F. Hutter, L. Kotthoff, and J. Vanschoren, eds., *Automatic machine learning*. The Springer Series on Challenges in Machine Learning, Basel, Switzerland: Springer International Publishing, 1 ed., Mar. 2019.

[7] G. A. Lujan-Moreno, P. R. Howard, O. G. Rojas, and D. C. Montgomery, "Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study," *Expert Syst. Appl.*, vol. 109, pp. 195–205, Nov. 2018.

[8] N. G. Ramadhan, F. D. Adhinata, A. J. T. Segara, and D. P. Rakhmadani, "Deteksi berita palsu menggunakan metode random forest dan logistic regression," *Jur. Ris. Kom.*, vol. 9, p. 251, Apr. 2022.

[9] N. Widjiyati, "Implementasi algoritme random forest pada klasifikasi dataset credit approval," *J. Janitra Inform. Sis. Inf.*, vol. 1, pp. 1–7, Apr. 2021.

[10] I. I. Sholikhah, A. T. J. Harjanta, and K. Latifah, "Machine learning untuk deteksi berita hoax menggunakan bert," in *Prosiding Seminar Nasional Informatika*, vol. 1, pp. 524–531, 2023.

[11] N. Agustina, A. Adrian, and M. Hermawati, "Implementasi algoritma naïve bayes classifier untuk mendeteksi berita palsu pada sosial media," *Fakt. Exacta*, vol. 14, p. 206, Jan. 2022.

[12] A. Yodi Prayoga, A. Id Hadiana, and F. Rakhmat Umbara, "Deteksi hoax pada berita online bahasa inggris menggunakan bernoulli naïve bayes dengan ekstraksi fitur Tf-Idf," *Jurnal Syntax Admiration*, vol. 2, pp. 1808–1823, Oct. 2021.

[13] E. I. Setiawan, S. Johanes, A. T. Hermawan, and Y. Yamasari, "Deteksi validitas berita pada media sosial twitter dengan algoritma naive bayes," *Journal of Intelligent System and Computation*, vol. 3, pp. 55–60, Oct. 2021.

[14] S. Nurohanisah, R. Astuti, and F. Muhammad Basysyar, "DETEKSI BERITA PALSU MENGGUNAKAN ALGORITMA RANDOM FOREST," *jati*, vol. 8, pp. 422–428, Feb. 2024.

[15] T. A. Roshinta, E. Kumala, and I. F. Dinata, "Sistem deteksi berita hoax berbahasa indonesia bidang kesehatan," *remik*, vol. 7, pp. 1167–1173, Apr. 2023.

[16] A. Zahra and M. N. Fauzan, "Sistem identifikasi "fake news" menggunakan metode multinomial naïve bayes," *J. Sist. Dan Teknol. Inf. (JustIN)*, vol. 10, p. 489, Dec. 2022.

[17] A. R. Hanum, I. A. Zetha, S. C. Putri, R. A. Wulandari, S. P. Andina, J. N. Fajrina, and N. Yudistira, "Analisis kinerja algoritma klasifikasi teks bert dalam mendeteksi berita hoaks," *Jurnal teknologi informasi dan ilmu komputer*, vol. 11, pp. 537–546, July 2024.

[18] T. Misriati and R. Aryanti, "Optimalisasi random forest dan support vector machine dengan hyperparameter gridsearchcv untuk analisis sentimen ulasan primaku," *Journal of Information System Research (JOSH)*, vol. 5, pp. 1333–1341, Jul. 2024.

[19] P. D. Utami and R. Sari, "Filtering hoax menggunakan naive bayes classifier," *MULTINETICS*, vol. 4, p. 57, May 2018.

[20] F. Rahutomo, I. Y. R. Pratiwi, and D. M. Ramadhani, "Eksperimen naïve bayes pada deteksi berita hoax berbahasa indonesia," *Jurnal PKOP*, vol. 23, July 2019.

[21] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on indonesian news sites based on text classification using svm and sgd," in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pp. 45–49, 2017.

[22] P. Anandhi and D. E. Nathiya, "Application of linear regression with their advantages, disadvantages, assumption and limitations," *Int. J. Stat. Appl. Math.*, vol. 8, pp. 133–137, Nov. 2023.

[23] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan algoritma cosine similarity dan pembobotan tf-idf pada sistem klasifikasi dokumen skripsi," *Jurnal Teknik Elektro*, vol. 9, no. 1, pp. 18–23, 2017.