World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Review Article)

Check for updates

Enhancement of random forest algorithm applied in fake news detection

Mary Justine R. Felicilda *, Aries James B. Geriane, Vivien A. Agustin, Mark Christopher R. Blanco, Jonathan C. Morano, Leisyl M. Mahusay and Jamillah S. Guialil

College of Information Systems and Technology Management, Pamantasan ng Lungsod ng Maynila, Sampaloc, Manila, Philippines.

World Journal of Advanced Research and Reviews, 2024, 22(02), 1075–1079

Publication history: Received on 02 April 2024; revised on 13 May 2024; accepted on 16 May 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.22.2.1439

Abstract

The widespread increase of fake news presents a serious obstacle in today's information-sharing situation, making it more difficult to distinguish fact from fabricated information. The study aims to improve machine learning algorithms' capacity to identify fake news to address this significant issue. To improve detection accuracy, Researchers applied a method that combines the Random Forest algorithm with Adaptive Boosting, or AdaBoost. By utilizing ensemble learning techniques, our method successfully leverages the collective intelligence of numerous decision trees, outperforming traditional approaches. By carefully experimenting with hyperparameters using the Random Search, The proposed method improved the algorithm's capacity to identify significant patterns suggestive of false information. The integration of AdaBoost to Random Forest produces a significant improvement, from 92.56% to 99.79% accuracy average difference, with an accuracy gain of 7% over the baseline Random Forest model. This improvement emphasizes how effective ensemble methods are at negotiating the complexities in terms of fake news detection. This study adds to the ongoing efforts to protect the integrity of information in the digital era by showcasing the efficacy of sophisticated machine-learning approaches in addressing the complex problem of fake news. The study underscores the critical importance of continuous innovation and adaptability in combating the increase of misinformation. Through interdisciplinary collaboration and technological advancements, the aim is to fortify defenses against the detrimental spread of false information, fostering an information ecosystem that prioritizes truth and resilience.

Keywords: Random Forest; AdaBoost; Hyperparameter Optimization; Fake News; Random Search

1. Introduction

In today's digital age, fake news has become widespread, and it is harder to assess what is true as information is readily available and can be shared quickly on social media platforms. Looking at the dark side of social media, it is observed that fake news is one of the serious issues in society [1]. Now it is becoming more common, causing harm to individuals, communities, and nations. The use of technology by people to spread and support lies, mislead, manipulate, and use propaganda, proved itself when online social networks such as Facebook and Twitter were taken advantage of for purposes for which is not originally intended [2].

The transition of social media into a platform replete with disinformation campaigns has a huge influence on the trustworthiness of the whole news ecosystem. Notably, one distinguishing element of news transmission on social media is the ability for anybody to register as a news publisher for free. Concurrently, firms are increasingly focusing on social media sites. Unsurprisingly, this transformation has raised worries about the spread of "fake" news articles by questionable news publishers who typically use deceptive strategies, such as the employment of false followers. The broad distribution of such fake news constitutes a significant concern, having the potential to have profound negative implications for individuals and society as a whole [3]. As a result, current research has emerged in this subject in order

* Corresponding author: Mary Justine R. Felicilda

to tackle key difficulties through the use of data. By categorizing how this spread and why it is effective, misleading netizens are crucial for building effective preventive algorithms.

In response to this growing threat, numerous studies have explored the application of machine learning techniques, with a notable focus on the Random Forest Algorithm. Random Forest, a versatile supervised learning technique, is adept at both Classification and Regression problems. It leverages multiple decision trees on distinct subsets of the input dataset, aggregating their results to enhance predictive accuracy. Despite its quick learning capability, the algorithm may take longer to make predictions due to the need for multiple trees. Examining feature importance becomes crucial for the potential removal of irrelevant features to mitigate overfitting [4].

To address the challenges associated with the Random Forest Algorithm, ensemble methods have been developed, representing a significant advancement in enhancing its efficiency and effectiveness. These ensemble methods work by combining the predictions of multiple individual models, such as decision trees, to achieve a more robust and accurate overall prediction. Despite these advantages, there are still limitations to it such as producing training sets by randomly picking cases totally at random from the original dataset, focusing instead on the underlying learning algorithm on previously misclassified training examples. In this research, we establish an improved random forest method that incorporates boosting ensemble learning to solve challenges and improve the algorithm's accuracy and performance.

2. Literature Review

2.1. Fake News

Fake news has arisen as a serious and possibly detrimental phenomenon in contemporary culture, notably through social media outlets. One study proposes a thorough survey focusing on the identification of fake news in online spaces to address the rising problem of fake news [5]. The authors stress the wide range of ways in which false news may be harmful, including the potential manipulation of individuals' decisions in crucial elements of life such as financial markets, healthcare choices, online shopping, education, and even political elections. The poll emphasizes the critical relevance of the automatic detection of online false news, recognizing its complexities and the problems it provides to industry and academics.

2.2. Random Forest

The approach of choice for text classification problems has always been Random Forest [6]. However, several factors within the Random Forest classifier need to be carefully tuned. The performance of the classifier may be considerably enhanced with careful tweaking of these hyperparameters, producing better outcomes.

Decision trees are the building blocks of a random forest algorithm. The random forest algorithm creates a "forest" that is trained via bagging or bootstrap aggregating. The precision of machine learning algorithms is increased by bagging an ensemble meta-algorithm [7].

The trees in the forest prefer to use uninformative features for node splitting which uses randomization in both bagging samples and feature selection [8]. As a result, while working with high-dimensional data, RFs have low accuracy. Additionally, RFs favor multivalued features in the feature selection process, which is biased.

The majority of researchers concentrated on the categorization issues presented by imbalanced data sets, which are prevalent in both the industrial production and medical research sectors. The ensemble method based on over-sampling is one of the most competitive techniques in the current research for these severely unbalanced data sets. The wrong sampling technique, however, quickly had an impact on the model's performance, which increased the complexity of the training process and led to an overfitting issue [9].

2.3. Hyperparameter

In their 2018 work titled "Hyperparameters and Tuning Strategies for Random Forest," Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix delve into the essential aspects of the random forest (RF) algorithm, focusing on the various hyperparameters that users must set [10]. These hyperparameters include considerations such as the number of randomly drawn observations for each tree, the method of drawing observations with or without replacement, the number of randomly drawn variables for each split, the splitting rule, the minimum number of samples required in a node, and the total number of trees. The paper begins with a comprehensive literature review, highlighting the impact of these hyperparameters on prediction performance and variable importance measures.

2.4. Adaptive Boosting (AdaBoost)

In 2019, Rising O. Odegua undertook a research study to investigate the effectiveness of ensemble approaches, including Bagging, Boosting, and Stacking, in improving the prediction capacities of machine learning models. Ensembles combine predictions to categorize new examples. Ensembles are made up of separately trained base learners or models. Several previous studies have consistently shown that ensembles perform more accurately than single-base models on average. This paper explores the detailed assessment of Boosting, Stacking, and Bagging methods on nine different datasets.

The study's conclusions reveal several important discoveries. First of all, empirical data confirms that ensembles often outpredict single base model accuracy. Second, the research shows that Boosting ensembles perform better than Bagging ensembles on average. Remarkably, Stacking—which is described as a meta-learning strategy—becomes the most accurate ensemble technique out of the three, outperforming Bagging and Boosting in terms of average accuracy [11].

3. Methodology

The research conducted in this study is based on a thoroughly maintained dataset known as the "fake news" dataset. This dataset, obtained via Kaggle, stands out for its specialized composition, which has been meticulously developed to include a varied array of articles divided into two categories: true and fake. Each article in this dataset is meticulously labeled based on its authenticity, with genuine articles classified as "true" and spurious ones as "fake." The Dataset distinguishes its size into two CSV files named true and fake. Wherein true.csv contains 21,417 articles, while fake.csv has 23,481. Strategically chosen to facilitate a comprehensive exploration and evaluation framework for training and testing.

The simulation initiates with data-related tasks, starting with Data Collection to compile a comprehensive dataset of English-language messages pertinent to Fake News detection. Following this, the Data Pre-processing step is implemented to refine the dataset for machine learning. Subsequently, the simulation advances to the Random Forest Model stage, where the initial model for Fake News detection is trained. The phase includes the segmentation of the dataset into training and testing sets, model training, and performance evaluation. The evaluation stage involves assessing the enhanced model`s performance through tests and experimentation of various hyperparameters with the use of Random Search. The Result Analysis phase interprets outcomes, identifying the accuracy per fold and its overall average accuracy.

The researchers' deliberate selection of the "fake news" dataset shows their dedication to assuring relevance and specificity in their search to understand the complexities of fake news identification. Proponents want to capture the intricate qualities and patterns inherent in fake news by using a dataset specifically developed for this purpose, resulting in a more informative and focused study. The dataset's large size not only enables strong model training but also allows for thorough assessment, which improves the reliability and generalizability of the findings. In essence, the selection of this curated dataset acts as a fundamental element, anchoring the study in a rich and intentional environment favorable to furthering our understanding of false news detection approaches.

4. Results and Discussion

Table 1 Cross Validation Scores of Enhanced Random Forest with AdaBoost

Existing method of random forest		**Proposed method of random forest with adaboost**	
Folds	**Accuracy**	**Folds**	**Accuracy**
1	94.77%	1	99.73%
2	89.87%	2	99.73%
3	90.55%	3	99.88%
4	92.54%	4	99.85%
5	91.80%	5	99.91%
6	94.24%	6	99.82%
7	95.51%	7	99.76%

8	94.03%	8	99.88%
9	91.83%	9	99.73%
10	90.43%	10	99.88%
Average	92.56%	Average	99.79%

Results have shown that the proposed method consistently outperforms the existing method in all 10 folds of cross-validation. The average accuracy of the proposed method is 99.79% which is much higher than the existing method's 92.56%. The significant increase in accuracy highlights the effectiveness of integrating AdaBoost into the Random Forest Algorithm. Furthermore, Researchers experimented with various combinations of hyperparameters between 5 to 50 for max_depth and 150 to 400 for n_estimator, using Random Search. The Random Search chose the best hyperparameter values, max_depth of 50 and n_estimator of 400. Examining the individual folding accuracy shows a consistent pattern of improvement with the proposed method. In almost all respects, the accuracy of the proposed method exceeds that of existing methods often by many times.

Overall, the results presented in Table 1 demonstrate the significant improvement achieved by integrating AdaBoost with Random Forest Algorithm. This improved method is expected to effectively detect fake news articles and contribute to ongoing efforts to combat misinformation in the digital age.

5. Conclusion

The integration of AdaBoost with Random Forest yielded promising results, manifesting in higher accuracy and improved classification performance compared to traditional approaches. Furthermore, meticulous hyperparameter tuning techniques optimized the Random Forest model's performance, enhancing its capabilities in detecting fake news articles. Using AdaBoost with the hyperparameters of 50 for max_depth and 400 for n_estimators, based on the selection from random search, the proposed method greatly impacts the outcome result. This proves that fine-tuning and finding the best hyperparameters results in the best outcome.

The research on the Enhancement of Random Forest has created opportunities beyond just identifying fake news. Researchers recommend exploring alternative applications to foresee the full capabilities of the enhanced algorithm. Additionally, it is advised to apply the proposed approach to multiple datasets to validate its effectiveness across diverse contexts. Additionally, investigating various feature selection and hyperparameter optimization techniques could further enhance the model's performance. These recommendations aim to guide future research in optimizing the potential of the enhanced Random Forest algorithm across a variety of domains and applications.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Patel, A., Tiwari, A. K., & Ahmad, S. S. (2021). Fake News Detection using Support Vector Machine. Proceedings of the 3rd International Conference on Advanced Computing and Software Engineering. https://doi.org/10.5220/0010562000003161

[2] Mateusz. (2022). Fake news as a threat to social resilience. European Research Studies, XXV(Issue 1), 765–782. https://doi.org/10.35808/ersj/2886

[3] Dr. S. Rama Krishna, Dr. S. V. Vasantha, K. Mani Deep, 2021, Survey on Fake News Detection using Machine learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICACT – 2021 (Volume 09 – Issue 08).

[4] Donges, N. (2024, March 8). Random Forest: A complete guide for machine learning. Built In. https://builtin.com/data-science/random-forest-algorithm

[5] Zhang, Xichen & Ghorbani, Ali. (2019). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management. 57. 10.1016/j.ipm.2019.03.004.

[6] George, S. C. G., & Sumathi, B. (2020). Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction. International Journal of Advanced Computer Science and Applications (IJACSA), 11(9). DOI: 10.14569/IJACSA.2020.0110920

[7] Mbaabu, O. (2020). Introduction to Random Forest in Machine Learning. Retrieved from https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

[8] Nguyen, T., Huang, J. Z., & Nguyễn, T. T. (2015). Unbiased feature selection in learning Random Forests for high-dimensional data. The Scientific World Journal, 2015, 1–18. https://doi.org/10.1155/2015/471371

[9] Gu, Q., Tian, J., Li, X., & Jiang, S. (2022). A novel Random Forest integrated model for imbalanced data classification problem. Knowledge-based Systems, 250, 109050. https://doi.org/10.1016/j.knosys.2022.109050

[10] Probst, P., Wright, M., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3). https://doi.org/10.1002/widm.1301

[11] Odegua, R. (2019). An empirical study of ensemble techniques (bagging, boosting, and stacking). In Proceedings of the Deep Learning IndabaX, Nairobi, Kenya, 25–31 August 2019.