Delving into the Practices Involved in the Creation and Dissemination of Misinformation

Muhammad Ubaid Ur Rehman[1], Asma Javed[2], Hasna Arshad[1], Samia Ijaz[1]
[1]Department of Computer Science, HITEC University Taxila Cantt.
[2]Department of Computer Science, Capital University of Science and Technology, Islamabad Pakistan.
*** Correspondence:** mubaid63@gmail.com

Citation | Rehman. M. U. U, Javed. A, Arshad. H, Ijaz. S, "Delving into the Practices Involved in the Creation and Dissemination of Misinformation", IJIST, Vol. 7 Issue. 1 pp 462-478, Feb 2025
DOI | https://doi.org/10.33411/ijist/202571462478
Received| Jan 15, 2025 **Revised|** Feb 22, 2025 **Accepted|** Feb 25, 2025 **Published|** Feb 28, 2025.

This study investigates the authenticity of news with specific training features validating the same with specific machine-learning techniques. The contents of fake news are created to make credible information that would create mass opinions and provide a strong basis to convince the readers or confuse them utterly. The fake information is usually disseminated using numerous automated algorithms. Therefore, it is very quintessential to identify the sources and authenticity of such information. With recent advancements in information communication technology, there exists a cluster of deep knowledge from which a user intends to retrieve relevant information such as news articles. For data mining and classification tasks such as fake news classification, the approach of machine learning can be employed for effective experimentation. To address the raised issues in this study, a comprehensive and diversified dataset was required that must contain relevant knowledge with sentiment tags such as authentic and fake news. To fulfill the same, a corpus comprising over 44k authentic and fake news items is collected. The current study demonstrates that bagging with an extra tree classifier yielded better classification accuracy as compared with multiple existing studies and other classification algorithms. Moreover, this study emphasizes news classification as fake or authentic using data mining and analytics.

Keywords: Fake News; Misinformation; Feature Extraction; NLP; Ensembling.

Introduction:

The concept of fake news existed long before the emergence of the internet and modern computing technologies. The transition from traditional media to social media has significantly accelerated information dissemination, acting as a catalyst for the rapid spread of both authentic and fake news [1][2]. Fake news completely convinces the reader that authentic news is not credible and can lead to a negative impact on society at large. Fake news is deliberately crafted to appear credible, influencing public opinion and either persuading readers or causing complete confusion. Misinformation directly affects investment plans, the stock market, and reactions to natural calamities [3] During the 2016 U.S. Presidential elections, social media platforms played a significant role in spreading fake news and misinformation, potentially influencing the election outcome. Some examples include misleading claims about Hillary Clinton's health and the false report that Pope Francis had endorsed Donald Trump. [4]. Similarly, In the year 2020, During the COVID-19 Pandemic, The Misinformation regarding the origins of the virus, possible cures, and the safety of vaccines presented a lot of confusion, panic, and distress. Distorted facts regarding the discovery of various treatment techniques and the effectiveness of hydroxychloroquine contributed to medicine shortages and negatively impacted patients. Similarly, some automated algorithms also disseminate misinformation. Therefore, it is crucial to identify the content and sources of such information [5]. Multiple organizations such as MIT's CSAIL have developed algorithms to identify fake and authentic information [6].

With a recent advancement in the domain of information communication technology, a vast repository of knowledge has emerged, allowing users to retrieve relevant information efficiently [7]. To retrieve this knowledge, a user has to perform certain operations, such as data mining. Data mining has emerged as a powerful tool for solving analytical problems such as decision-making, which enables a particular organization to gain a competitive advantage in the corporate sector. Data mining algorithms can be implemented to extract desirous information from a large data repository, such as fake and authentic news [8]. With recent developments in the domain of deep learning [9], the capability of a learning algorithm to analyze a particular text has been improved significantly [10]. Using advanced learning techniques could be an effective tool for conducting in-depth research. Another study proposed a mechanism to investigate the credibility of tweets or news with a specific pool of features using certain learning algorithms [11].

Researchers in [12] proposed a novel algorithm that detects unauthentic information in multiple languages including German, Slavic, and Latin. The authors of this study evaluated their algorithm on different corpuses namely Fake-Br-Corpus, Twitter-BR, but-lifestyle, Fake-News-Data, and Fake-Or-Real-News respectively. Moreover, this study was conducted on an Italian language-based dataset containing 300k news and more than 50k posts extracted from multiple web pages and blogs providing fake or incorrect facts. The proposed technique achieved an accuracy of 91% testing accuracy with 77% training accuracy. However, prediction of fake news based on machine learning techniques can be enhanced by identifying the elements that negatively impact the information [13].

Similarly, another study conducted vigorous experiments on renowned social media platforms such as Weibo and Twitter. The proposed technique can detect fake news with 90% accuracy within 300 seconds of its dissemination. However, this technique was evaluated on a small news corpus containing 2282 news articles related to US elections which restricted the real-world potential of this work [14]. The evaluation was conducted using a random forest classifier which reported an 85% accuracy rate. To assess diversity, the proposed approach should be evaluated on a large corpus [15]. Similarly, another platform for the identification of fake news based on a deep learning model can be found in [16]. In this platform, authors have utilized publicly available LIAR datasets to classify different news items. As per study

results, the proposed approach has reported 86.12% accuracy with an average recall and precision of 86% [17][18]. The study evaluation also demonstrated that the proposed model achieved 85.86% accuracy on the BuzzFeed dataset and 88.64% on the PolitiFact dataset. The accuracy can be improved by adding more datasets. However, feature extraction is not demonstrated in this study [19]. Meanwhile, another work [20] proposes a technique to classify fake news using a deep neural network. The major aim of this approach is to classify fake news based on the length of the sentence.

Machine learning techniques generate higher accuracy [21] and therefore, this study examines if authentic news is more likely to be true than program-generated news. Therefore, for data mining and classification tasks; the machine learning approach has proven its significant effectiveness [22][23] such as feature extraction-based news classification [15][24]. Numerous studies have performed news classification using learning approaches; however, the availability of a proper dataset is a major concern [25] because of the evolving nature of news articles. Furthermore, the availability of a comprehensive and diversified dataset is a general requirement in almost every learning approach. Moreover, these techniques suffer from certain poor or insignificant accuracy. The reason for insignificant accuracy can be the usage of an imbalanced dataset or the selection of an unimportant pool of features. To address these concerns and evaluate the proposed approach, a comprehensive and diversified dataset was required that should contain labeled news items as authentic and fake. For that reason, a diversified corpus containing fully mapped 44,898 news items has been collected to evaluate the proposed approach. The reason for this selection is multi-fold: first, it contains a comprehensive and varied dataset and it is publicly available. In addition, the current study focuses on the classification of news items as fake or authentic using data mining. Moreover, the classification technique is implemented using enormous approaches to given testing and training datasets.

Problem Statement:

Fake news poses a significant threat to individuals, societies, and democratic processes by disseminating false or misleading information, eroding trust in credible news sources, and distorting public opinion. The challenge lies in designing and implementing sophisticated algorithms and techniques that can accurately and efficiently distinguish between genuine and fake news articles, considering the evolving nature of deceptive tactics employed by purveyors of misinformation. By addressing this problem, researchers aim to safeguard the integrity of information, restore trust in reliable sources, empower individuals to make informed decisions and preserve the foundations of a well-informed and democratic society.

Research Contribution:

- To propose an ensemble-based approach for the classification of fake news.
- To utilize enormous natural language approaches such as tokenization, stop-word removal, punctuation removal, and stemming for pre-processing of new datasets.
- To extract a large pool of feature sets containing both numerical and textual values from a large news dataset for learning algorithms.
- To employ an ensemble approach such as bagging to enhance the accuracy of fake news classification.
- To evaluate the performance of the ensemble approach on the fake news dataset.
- To compare the effectiveness of the ensemble approach with another state-of-the-art learning algorithm.
- To demonstrate the impact of ensemble-based classifiers over individual classifiers.

Material and Methods:

To address the identified challenges in this study, the methodology outlined in Figure 1 is employed. It comprises multiple phases, including dataset collection, data pre-processing, textual and numerical feature extraction, and the implementation of data mining classifiers such as Random Forest, Decision Tree Classifier, and Bagging Decision Tree Classifier. In the last phase, mining algorithms were evaluated using multiple quantitative measures such as Accuracy, Precision, Recall, and F-Measure.

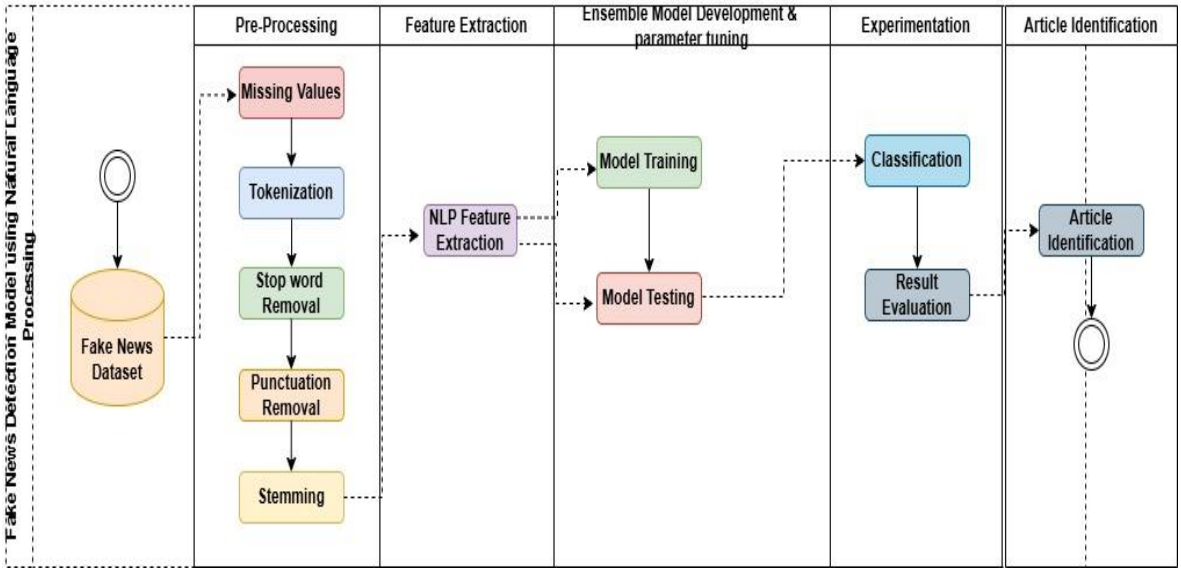

Figure 1: Proposed Methodology Diagram

To effectively evaluate a specific mining approach, a benchmark dataset is required [26][27], ensuring it encompasses all relevant types of information. Thus, a corpus comprising authentic and false news items was selected from Victoria University, which provides a fully mapped dataset. The rationale for this selection is multifaceted, as it offers a comprehensive corpus of over 44,000 false and authentic news items across multiple subjects, as illustrated in Figures 2 and 3. Additionally, it is an open-source dataset.

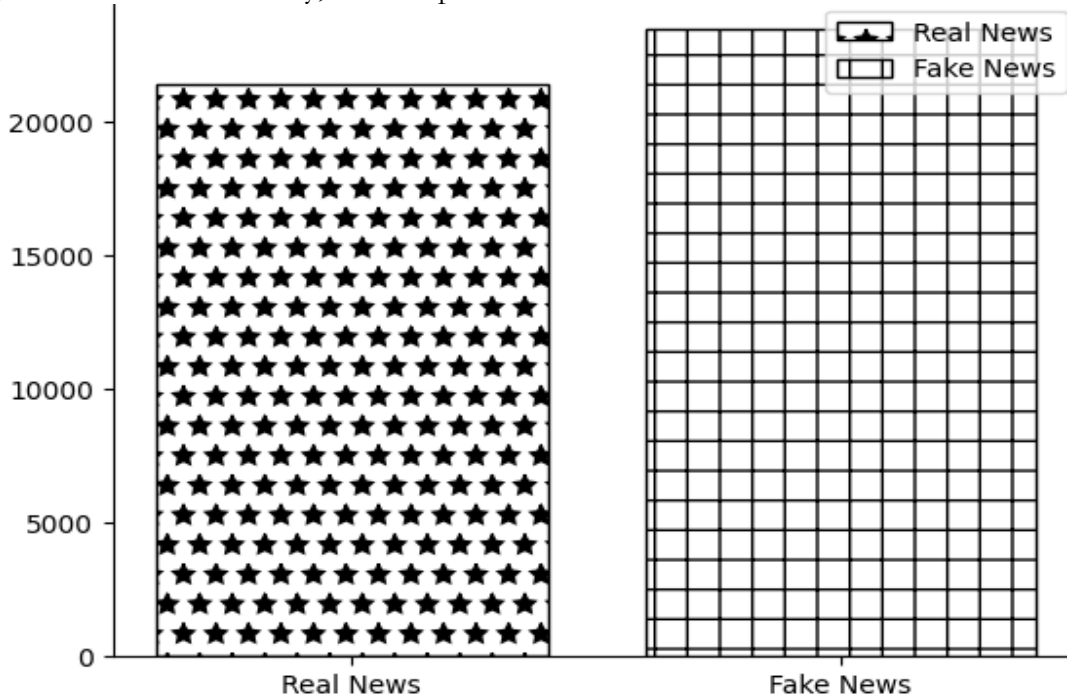

Figure 2. Visual Analytics of Fake News Dataset.

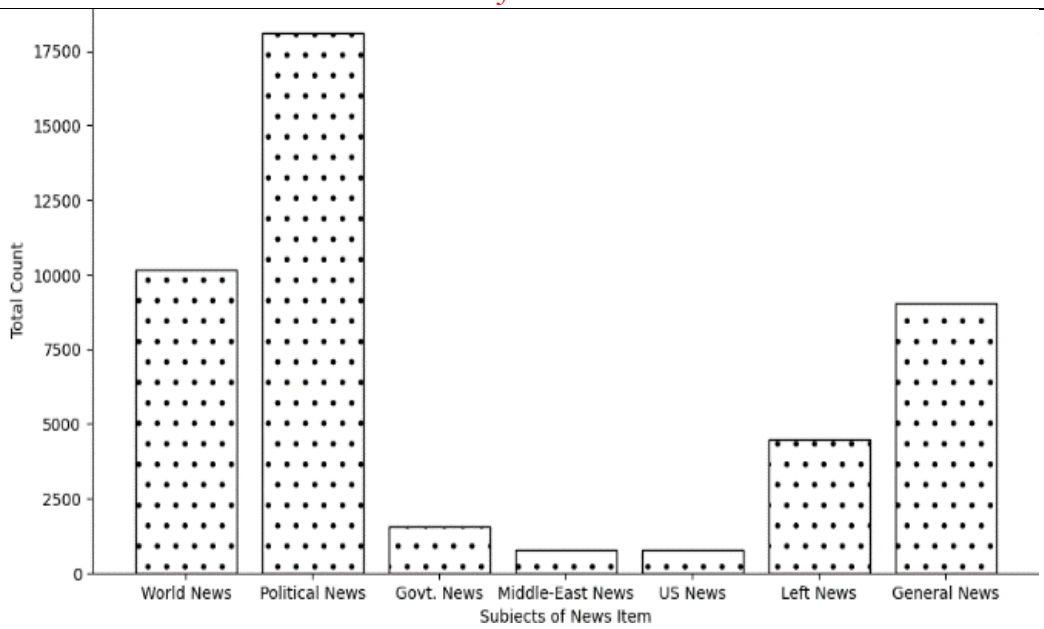

Figure 3. Subject-wise visual analytics of fake news dataset.

Initially, this dataset contains three attributes including news title, news body, and class. The dataset contains two types of articles; fake and real News. The dataset was collected from real-world sources; the truthful articles were obtained by crawling articles from Reuters. The fake news articles were collected from different sources including unreliable websites that were flagged by PolitiFact (a fact-checking organization in the USA) and Wikipedia. The dataset contains multiple articles on numerous topics; however, the majority of articles focus on political and World news topics. A class attribute presents the nature of each news article as authentic or false, which is very significant in classification as 'Fake' or 'Authentic'. The textual class can be encoded using the scheme proposed in Table 1.

Table 1. Encoding Scheme for Textual Class.

Sr. #	Class	Encoded Value
1.	Fake	0
2.	Authentic	1

Data Preprocessing:

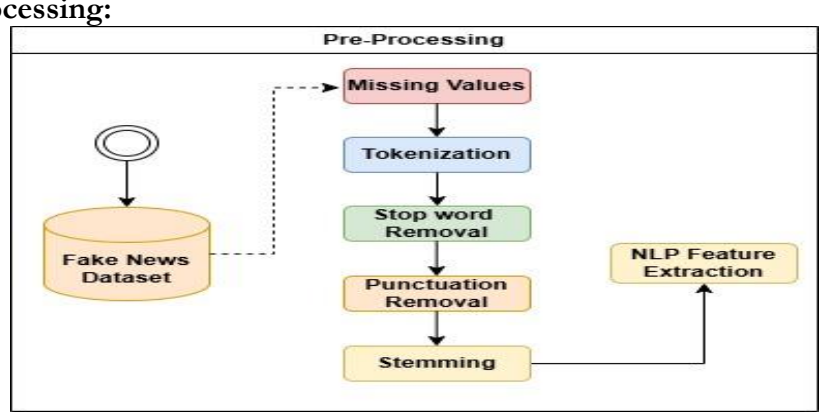

Figure 4. A flow chart explaining basic pre-processing steps.

- **Tokenization:** Tokenization is the process of splitting a text, document, or character sequence into multiple words [28]. These tokens can be in the form of sentences, characters, or individual words based on application requirements. For example, output from string 'He loves data science.' can be ['He', 'loves', 'data', 'science', '.'] where 'He', 'loves', 'data', 'science' and '.' are individual tokens. Tokenization plays a significant role in natural language and text-processing applications.

- **Stop-word Removal:** Stop-word removal is a technique used in text processing applications that involve the elimination of common words such as 'is', 'am', 'are', 'or', 'the', 'a', etc. from a text corpus [29]. Stop words are common words in language that do not convey any significant meanings. The stop-word removal technique is commonly used to eliminate noise and reduce high text dimensionality, making the text corpus more meaningful. For instance, applying this technique to the sentence "He loves data science." may result in "loves data science."
- **Punctuation Removal:** In a natural language, punctuation is a collection of symbols such as ", ', "!" etc. which are being utilized to make more meaningful sentences. In text mining or text processing applications, it is generally required to have a properly cleaned dataset in terms of removed punctuation marks because they create noise while processing any text. Thus, it is quintessential to clean the data corpus in terms of punctuation marks to have more effective results. For example, output from the string 'loves data science.' can be 'loves data science'. In this study, all punctuation marks are removed from the text corpus. Figure 5 presents a visual illustration of removed punctuation marks.

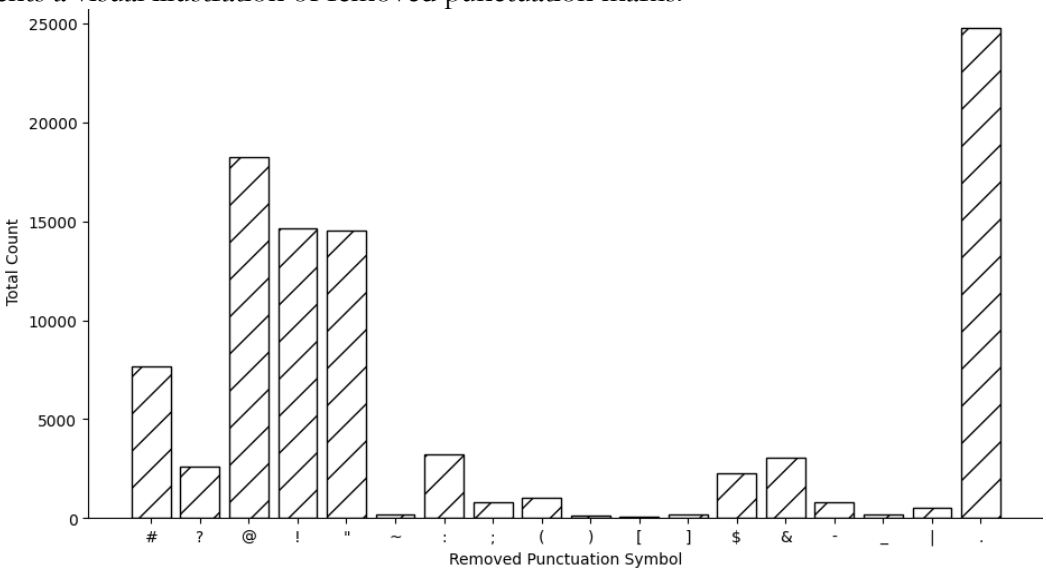

Figure 5. Visual Illustration of Removed Punctuation Marks.

- **Stemming:** Stemming is a text normalization technique that reduces words to their root form by retaining the core part of the word while removing suffixes, prefixes, and other variations [30]. The basic purpose of stemming is to convey its actual meaning. For example, the output from the string 'loves data science' can be 'loves data science'. The basic purpose of stemming is to consolidate words with the same meanings and reduce text dimensionalities. In this study, stemming is performed using snowball stemmer [31], which is ideal for fake news datasets as it reduces words to their root forms, standardizing informal language, slang, and mixed-language usage such as "gonna" to "go" or "wanna" to "want.". This capability improves text analysis tasks like sentiment analysis and topic modeling, enhancing content understanding and classification.
- **Feature Extraction:** The process of feature extraction refers to the transformation of raw data into numerical features because learning algorithms can only process numerical features. Therefore, it transforms raw data into numerical features while preserving information in its original dataset. This approach yields better accuracy than directly applying machine learning to raw data. Feature extraction can be further classified into two subcategories such as manual and automated feature extraction. Manual feature extraction is a process that requires identifying and describing features that are relevant to a given problem and implementing a way to extract those features. In numerous scenarios, a better

understanding of the problem domain or background can help in informed decisions to extract useful features. Over the decades, researchers have proposed feature extraction techniques for images, signals, and text.

Similarly, automated feature extraction is a process that requires specialized algorithms or deep neural networks for automatic feature extraction from text, signals, and images without the need for human intervention. This technique is very effective when you quickly want to input raw data to develop machine learning algorithms. The wavelet scattering is an example of automated feature extraction. Table 2 presents twenty-six numerical features that are automatically extracted using Spacey-API to conduct this study.

Table 2: List of Automated Extracted Numerical Features.

Sr. #	Feature Name	Feature Type	Category
1.	Word Count	**General Features**	Numerical
2.	Sentence Count		Numerical
3.	Character Count		Numerical
4.	Sentence Length		Numerical
5.	Average Word Length		Numerical
6.	Average Sentence Length		Numerical
7.	Count of Countries	**Name Entry Recognition Features**	Numerical
8.	Persons		Numerical
9.	Products		Numerical
10.	Work of Arts		Numerical
11.	Languages		Numerical
12.	Time		Numerical
13.	Money		Numerical
14.	Cardinal		Numerical
15.	NORP		Numerical
16.	Organizations		Numerical
17.	Locations		Numerical
18.	Events		Numerical
19.	Date		Numerical
20.	Law		Numerical
21.	Quantity		Numerical
22.	Ordinal		Numerical
23.	Polarity		Numerical
24.	FAC		Numerical
25.	GPE		Numerical
26.	Class		Numerical

Learning Algorithm: Machine learning is an effective approach that can be employed for prediction and classification tasks [22] such as feature extraction-based news classification. Therefore, in this phase, we implemented a machine learning approach utilizing various

methodologies, including Decision Tree, Random Forest, and Regression, for experimentation.

Decision Tree: A decision Tree is a renowned classification and prediction algorithm that follows a tree-like data structure for the decision-making process [32]. These algorithms are developed by partitioning data through recursion, where each node indicates a decision based on relevant features and the leaf node leads to a particular outcome. Decision trees can effectively handle both numeric and categorical values. It also provides robustness to noise and missing values.

- **Random Forest:** Random Forest is an effective and hybrid machine learning algorithm that utilizes both decision trees and an ensemble learning model [33]. A random forest is usually a cluster of decision trees where each tree is constructed independently using randomly selected features at each node. During prediction, outcomes from all decision trees are combined through averaging. Random forests are also known for their capability to handle complex and diverse datasets and overfitting problems.
- **Extra Tree:** The extra Tree learning algorithm is a variant of random forest that enhances the tree-building process through randomization [34]. Unlike random forests that consider a subset of features at each node, Extra Trees randomly select feature thresholds without evaluating various splitting points. This additional randomization in the split selection makes Extra Trees even more robust against overfitting and noise in the data.
- **Assembling Approach:** An ensemble approach is a hybrid approach that combines individual learning models to make enhanced prediction or classification [35]. An ensemble leverages the diversity and collective intelligence of multiple models to enhance overall performance and robustness. In machine learning, an ensemble approach can be implemented using multiple sub-approaches such as Bagging, Boosting, Voting, and Stacking. In this study, we implemented an ensemble approach using bagging with Decision Tree and Bagging Extra Tree to enhance classification performance and robustness.

Quantitative Evaluation:

In this study, machine learning approaches were implemented using multiple algorithms that include Random Forest, Bagging Decision Tree, Bagging Extra Tree, and Decision Tree. For testing and training purposes, the dataset was converted into two parts a training dataset and a testing dataset. All implemented learning algorithms were evaluated using four mathematical-based evaluation measures such as accuracy, f-measure, precision, and recall scores using equations 1 – 3.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F\text{-}Measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (3)$$

Experimental Results:

Figure 6 represents the performance of the Decision Tree model using four quantitative evaluation metrics such as accuracy, precision, recall and F-measure. The y-axis indicates a score that is currently ranging from 0.0 to 0.9 where each bar illustrates the highest possible score for each measure. Similarly, each evaluation metric is distinguished using a distinguishable hatching pattern that visually illustrates a performance variance. Notably, a Recall of 0.85, while Accuracy = 0.80, Precision = 0.88, and F-Measure = 0.82 were recorded from this model.

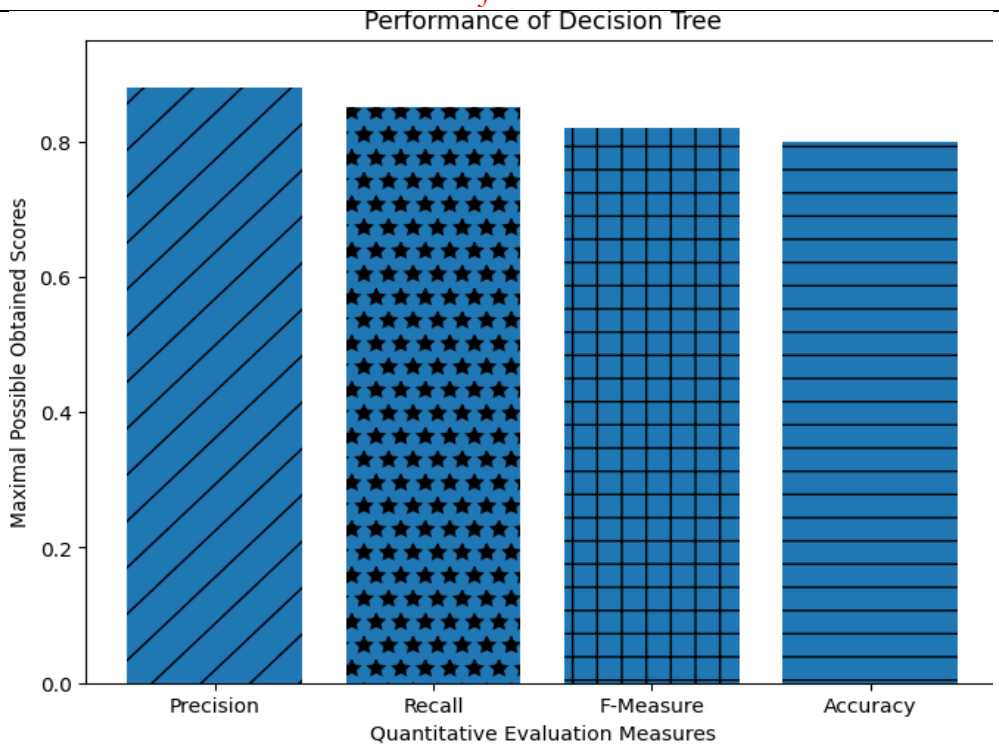

Figure 6: Performance graph of Decision Tree.

Figure 7 illustrates the performance of the Random Forest model which shows that this model has reported an accuracy score of 0.88. Meanwhile, precision = 0.86, recall = 0.86, and the f-measure score of 0.85.

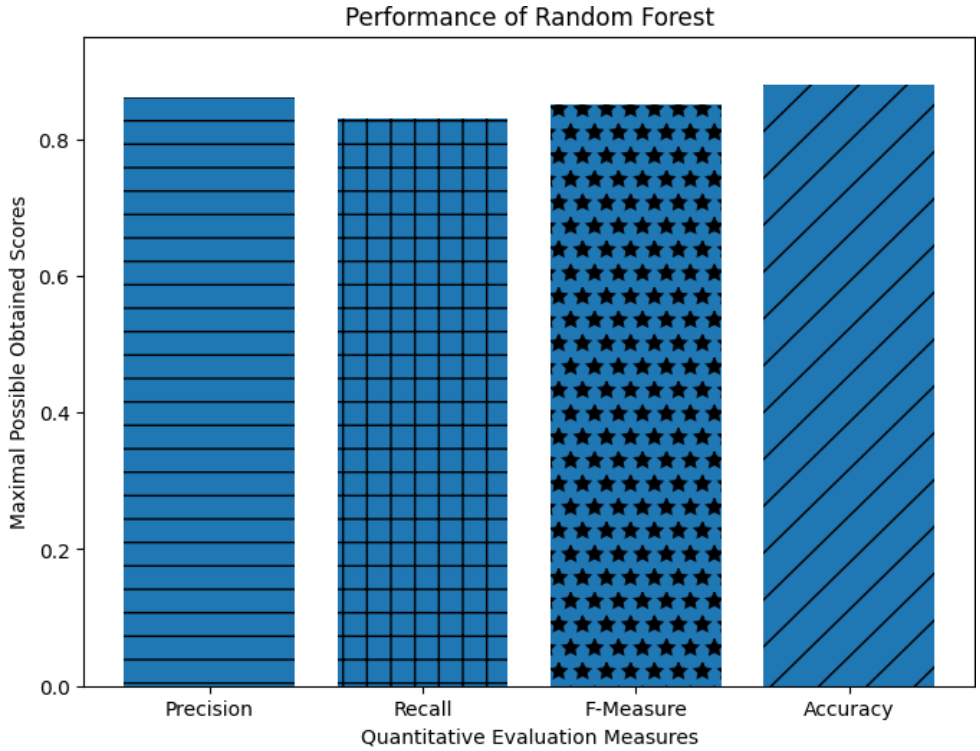

Figure 7: Performance graph of Random Forest

Figure 8 highlights the maximal performance of the Bagging Extra Tree model using a visual illustration below. From the obtained chart it can be observed that this model has reported the highest accuracy score of 0.95. Meanwhile, 0.90 as a precision score, a higher possible recall and f-measure scores of 0.95 and 0.93 significantly.

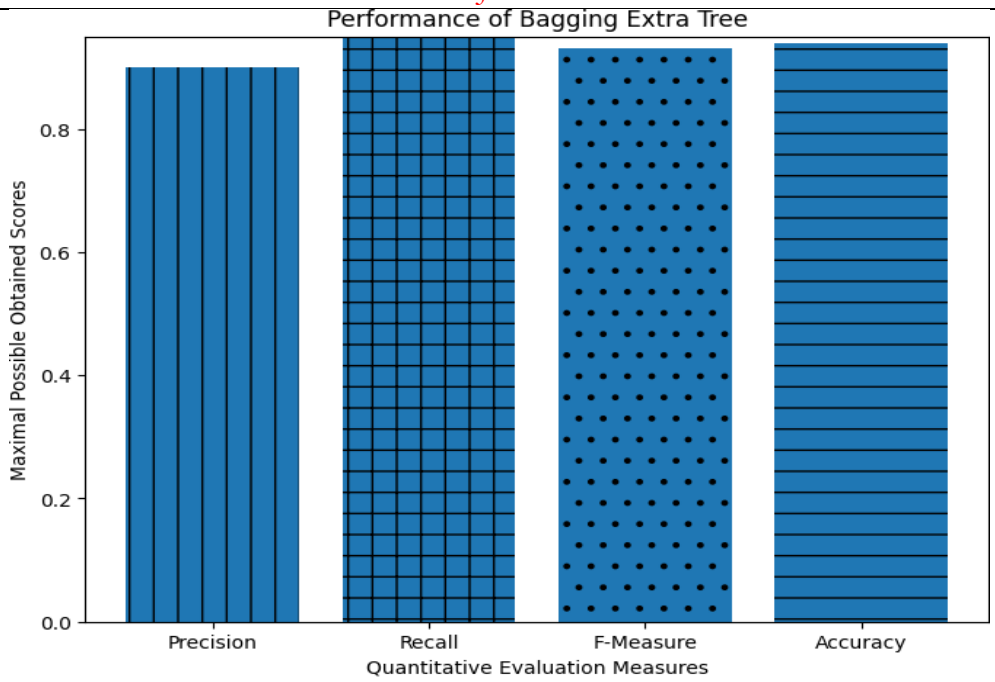

Figure 8: Performance graph of Bagging Extra Tee.

Figure 9 presents a comparison of obtained quantitative measures from the Bagging Decision Tree model. From this graph it can be observed that it has reported a precision of 0.86, meanwhile, 0.89 as the recall score, an f-measure score of 0.88, and an accuracy of 0.85 simultaneously.

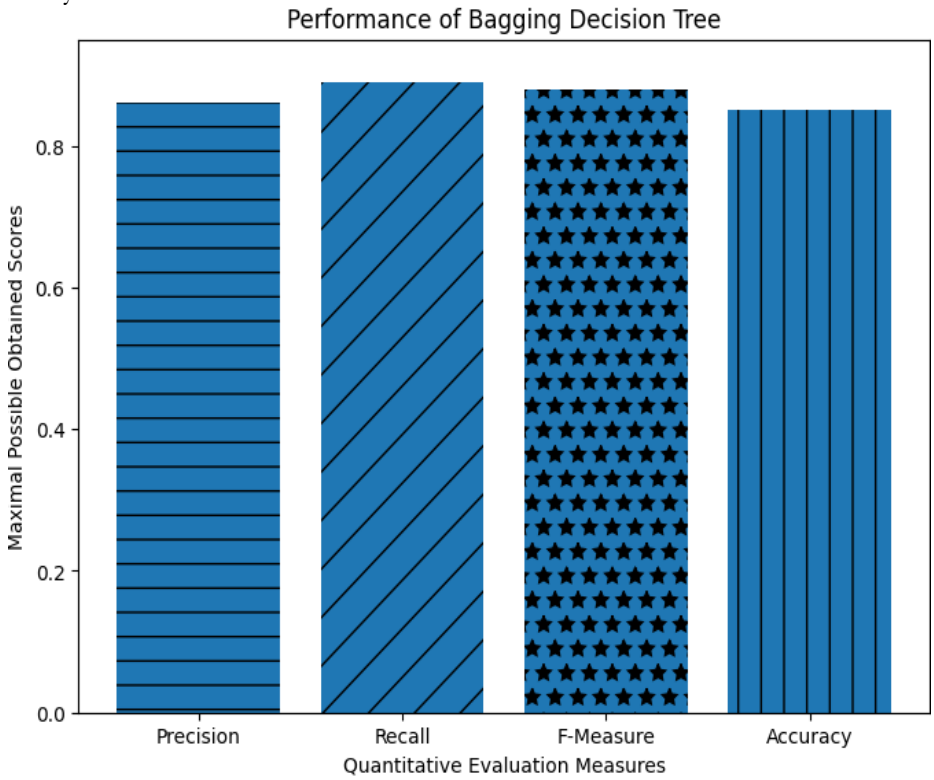

Figure 9: Performance graph of Bagging Decision Tee.

Figure 10 presents a precision-based comparison of all implemented approaches. For instance, Random Forest achieved a precision of 0.86, while Decision Tree scored a precision of 0.88. Similarly, the Bagging Decision Tree achieved a precision of 0.86, while the highest precision score of 0.90 was recorded for the Bagging Extra Tree Classifier.

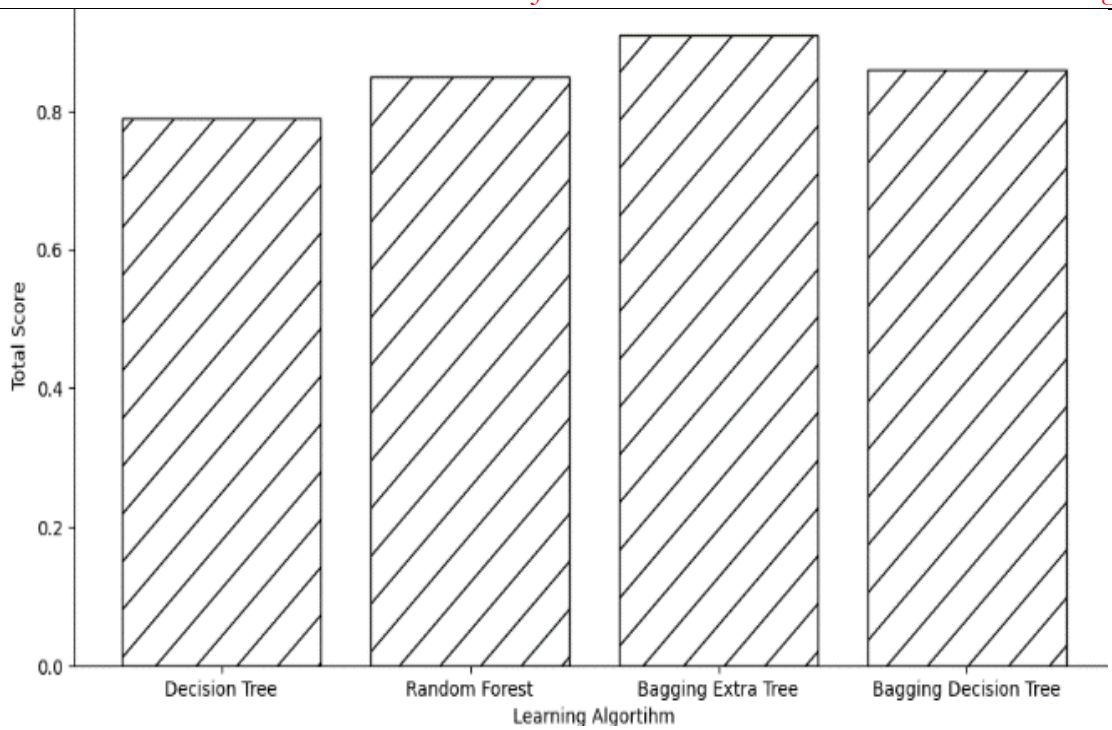

Figure 10. Comparison of Learning Approaches in Terms of Precision.

Figure 11 illustrates a recall score-based comparison of learning algorithms wherein the lowest recall score of 0.85 was recorded for the Decision Tree. Similarly, a recall score of 0.86 from the random forest and 0.89 from the bagging decision tree. However, the highest possible score of 0.95 was observed from bagging the extra tree classifier.

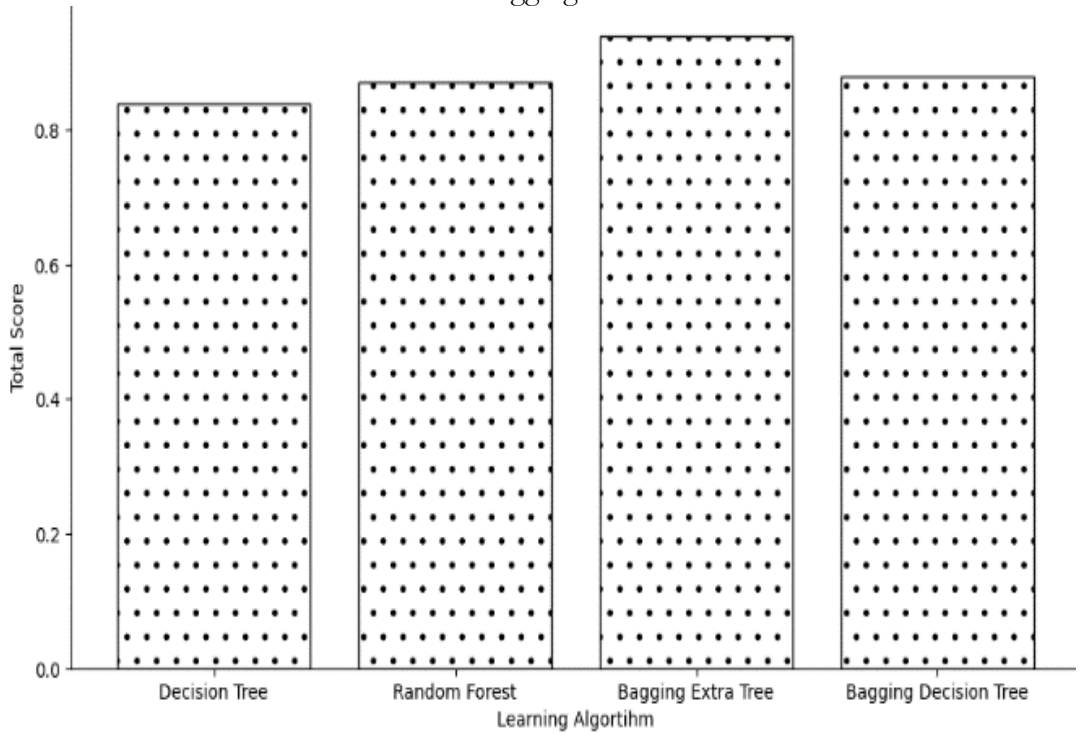

Figure 11. Comparison of Learning Approaches in Terms of Recall.

The f-measure score-based comparison is shown in Figure 12 which explains that the highest possible f-measure score of 0.93 was recorded from bagging extra tree classifier. Moreover, the decision tree reported a lowest f-measure score of 0.82. The f-measure scores of 0.85 and 0.88 were noted from random forest and bagging decision tree classifiers.

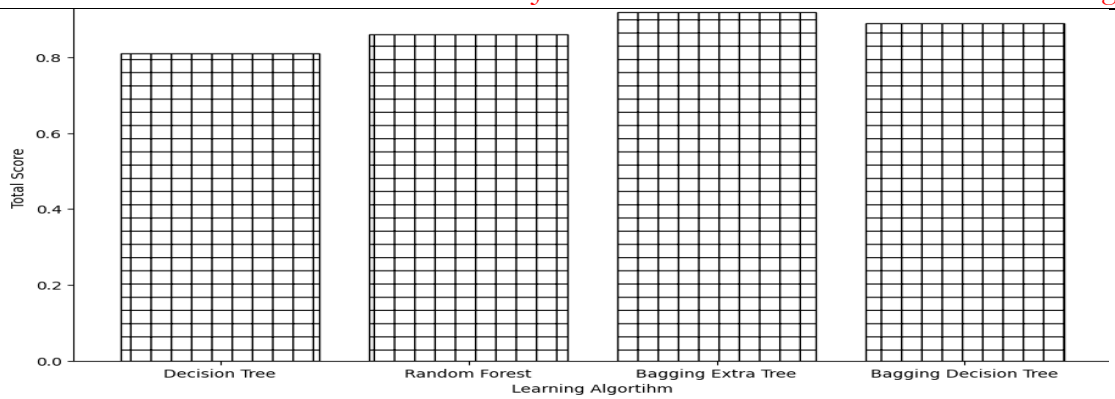

Figure 12. F-measure score-based comparison of all implemented approaches.

All obtained accuracies from the implemented techniques are visualized in Figure 13 which shows that a decision tree reported a lowest accuracy score of 0.80. Meanwhile, bagging decision trees and random forests reported accuracy scores of 0.85 and 0.88. However, the bagging extra tree classifier reported the highest score of 0.94.

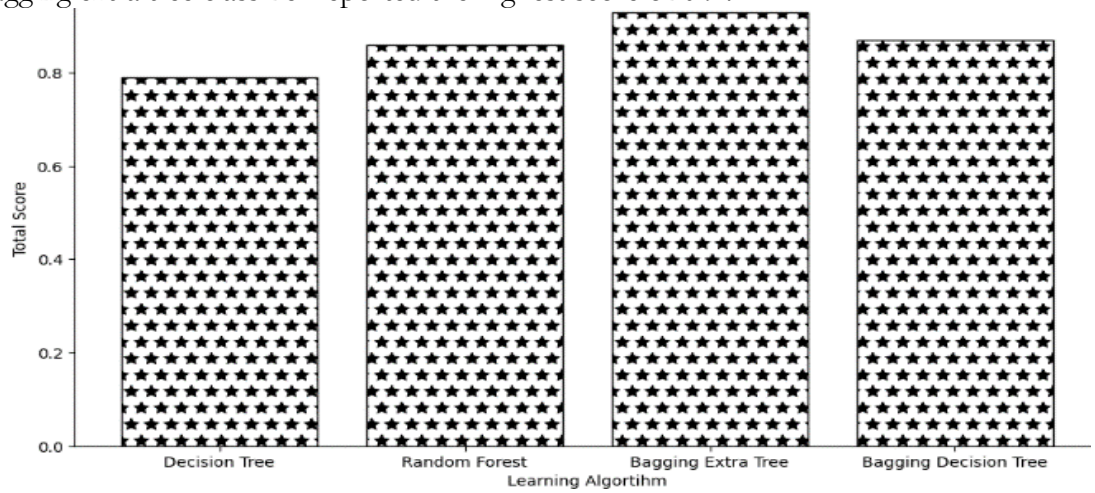

Figure 13. Comparison of Learning Approaches in Terms of Accuracies.

Table 3 provides a comparison between the proposed approach and the previously established methods in this domain. The proposed model outperformed previous approaches with a higher accuracy of 94% subsequently. These studies have multiple issues like lower identification accuracy, incorrect identification of news articles, low number of tuples in employed datasets, and a lack of proper feature extraction techniques.

Table 3: Comparison of Proposed Work with Existing Works.

Ref.	Paper Title	Dataset	Accuracy
[12]	Fake news detection in multiple platforms and languages	TwitterBR, FakeBrCorpus, Fake News Data1, Fake Or Real News, and btv lifestyle dataset	79%
[13]	Polarization and Fake News: Early Warning of Potential Misinformation Targets	Italian Facebook dataset with 300K official media news, 50K incorrect information	91%
[14]	FNED: A Deep Network for Fake News Early Detection on social media	Twitter and Weibo datasets	90%

[15]	Supervised Learning for Fake News Detection	2282 news articles related to the US election	85%
[16]	Exploring deep neural networks for rumor detection	5800 Twitter tweets	86.12%
[19]	DeepFakE: improving fake news detection using tensor decomposition-based deep neural network	BuzzFeed dataset, the PolitiFact dataset	85.86%
[36]	Analyzing and distinguishing fake and real news to mitigate the problem of disinformation	Extracted using FakeNewsNet tool	75% LSTM 45% GRU 62%-RNN
	Proposed	Fake News Dataset	94%

Table 4 below illustrates three possible outcomes of the current study related to misinformation and disinformation, each with a corresponding result. The first outcome explains the development of a robust framework for analyzing and classifying misinformation, leading to a clear definition and typology, along with specific criteria and classification systems. The second outcome encompasses an evidence-based strategy for mitigating the spread and impact of misinformation, with the study evaluating existing methodologies and recommending policy enhancements. The third outcome exploits the psychological and social factors delving into the creation and dissemination of misinformation, using research data and case examples to understand these elements and propose solutions to foster skepticism and critical thinking.

Table 4. Possible study outcomes with detailed obtained results.

Sr #.	Possible Outcome	Result
1.	A robust framework for identifying and classifying misinformation.	The results illustrate a clear definition and a typology for misinformation and disinformation. It proposes a specific criterion and classification systems to systematically identify and categorize various forms of misinformation.
2.	Evidence-based strategies for mitigating the spread and impact of misinformation.	The current study also evaluates multiple existing evidence-based methodologies to counter the spread of misinformation effectively. It also recommends ground for policymakers to enhance the resilience of the information ecosystem.
3.	A deeper understanding of the psychological and social factors driving the creation and dissemination of misinformation.	This study looks at the social dynamics and cognitive biases that lead to the production and dissemination of false information. It uses research data and case examples to demonstrate these elements and offers solutions to encourage skepticism and critical thinking.

Discussions:

The current study demonstrates that bagging with an extra tree classifier yielded better classification accuracy as compared with multiple existing studies and other classification algorithms such as Decision Tree, Random Forest, and Bagging with Decision Tree, because

of increased diversity and randomness introduced in model development. The proposed approach also introduced randomness in data selection for model training and provides robustness against noisy data and overfitting problems. In this approach, multiple trees were created where each was trained on a different data subset which inherits higher data diversity and captures multiple aspects of data. Furthermore, ensemble averaging in bagging significantly mitigates individual errors and enhances the overall classification accuracy of the proposed approach. The fusion of ensemble averaging, randomness, and reduction of overfitting problems makes it a powerful approach to achieve higher accuracy as compared with implemented algorithms such as Decision Tree, Random Forest, and Bagging with Decision Tree.

One of the key advantages of this method is its robustness against noisy data and overfitting. The introduction of randomness in both feature selection and data sampling helps the model generalize better to unseen data, reducing the likelihood of memorizing the training set. This characteristic is particularly valuable in real-world applications where data quality can vary significantly. Additionally, the ensemble averaging technique employed in bagging mitigates individual model errors, ensuring that the overall classification performance remains high even if some trees make incorrect predictions.

The proposed approach not only demonstrates superior performance but also offers practical benefits across various domains, including healthcare, finance, and marketing. Its versatility allows it to adapt to different datasets and classification tasks, making it a valuable tool for practitioners. Furthermore, the potential for future optimization and the ability to provide insights into the decision-making process through individual trees enhance its applicability. Overall, the combination of improved accuracy, robustness, and adaptability positions this method as a powerful solution for achieving reliable classification outcomes.

Conclusion:

The approach of data mining or machine learning is considered an effective technique that has proven its effectiveness in the analysis and visualization of enormous data corpora. The mining technique plays a pivotal role by encompassing multiple mathematical and statistical-based models in the identification and classification of hidden data patterns in comprehensive and diversified datasets. Because of its significant effectiveness, this approach can be employed for prediction and classification tasks such as real and fake news classification problems.

In this study, the classification is performed using multiple learning techniques such as Decision Tree, Random Forest, Bagging Extra Tree, and Bagging Decision Tree classifiers. The learning algorithms require a comprehensive and diversified dataset for effective analysis. To solve this problem, the authors collected a comprehensive and diversified data corpus of more than 44k real and fake news instances in the English language. Moreover, learning algorithms require all features in numerical form. Therefore, 26 automated numerical features are extracted using natural language processing from a preprocessed dataset.

All learning approaches are implemented using Python language which significantly affects class predictability in terms of accuracy. According to study findings, bagging extra tree classifiers can effectively discriminate between real and fake news. In the future, the authors would like to implement and evaluate this approach in a real-time environment with higher data variations.

Author's Contribution: Mr. Muhammad Ubaid Ur Rehman is a graduate student and performed this research during one of his graduate courses. He proposed the notion of this idea and collected the fake news dataset. He also authored the initial draft that was polished later by the other two authors i.e. Dr. Hasna Arshad and Dr. Samia Ijaz. Ms. Asma Javed collaborated with Mr. Muhammad Ubaid Ur Rehman to implement the aforementioned strategies in the Python programming language. She also assisted with her expertise related to

the visualization of different aspects of this research. Dr. Hasna Arshad contributed by identifying the various categories of features that are important for the detection of misinformation. She also presented her expertise related to the presentation of various tables (conceptual and assessment) and figures in the draft. Dr. Hasna and Dr. Samia assisted Mr. Ubaid in building sound arguments for the analysis of figures in the result section and also assisted in polishing the structure of the paper.

Conflict of interest: The author(s) declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Project details: N/A

Reference:

[1] S. A.-H. Fadia Shah, Aamir Anwar, Ijaz ul haq, Hussain AlSalman, Saddam Hussain, "Artificial Intelligence as a Service for Immoral Content Detection and Eradication," *Sci. Program.*, vol. 1, no. 1, 2022, doi: https://doi.org/10.1155/2022/6825228.

[2] C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," *Proc. - 2nd IEEE Int. Conf. Smart Cloud, SmartCloud 2017*, pp. 208–215, Nov. 2017, doi: 10.1109/SMARTCLOUD.2017.40.

[3] M. Alazab *et al.*, "A Hybrid Wrapper-Filter Approach for Malware Detection," *J. Networks*, vol. 9, no. 11, Dec. 1969, doi: 10.4304/JNW.9.11.2878-2891.

[4] "Stanford study examines fake news and the 2016 presidential election | Stanford Report." Accessed: Feb. 22, 2025. [Online]. Available: https://news.stanford.edu/stories/2017/01/stanford-study-examines-fake-news-2016-presidential-election

[5] X. Zhou and R. Zafarani, "A Survey of Fake News," *ACM Comput. Surv.*, vol. 53, no. 5, Sep. 2020, doi: 10.1145/3395046.

[6] Ashish Gupta, Han Li, Wenting Jiang, "Understanding patterns of COVID infodemic: A systematic and pragmatic approach to curb fake news," *J. Bus. Res.*, vol. 140, pp. 670–683, 2022, doi: https://doi.org/10.1016/j.jbusres.2021.11.032.

[7] Y. Cheng, K. Chen, H. Sun, Y. Zhang, and F. Tao, "Data and knowledge mining with big data towards smart production," *J. Ind. Inf. Integr.*, vol. 9, pp. 1–13, Mar. 2018, doi: 10.1016/J.JII.2017.08.001.

[8] R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica, and A. N. Krishna, "Identification of Fake News Using Machine Learning," *Proc. CONECCT 2020 - 6th IEEE Int. Conf. Electron. Comput. Commun. Technol.*, Jul. 2020, doi: 10.1109/CONECCT50063.2020.9198610.

[9] S. S. U. Muhammad Mazhar Bukhari, Bader Fahad Alkhamees, Saddam Hussain, Abdu Gumaei, Adel Assiri, "An Improved Artificial Neural Network Model for Effective Diabetes Prediction," *Complexity*, 2021, doi: https://doi.org/10.1155/2021/5525271.

[10] S. S. U. Faiza Shah, Yumin Liu, Aamir Anwar, Yasir Shah, Roobaea Alroobaea, Saddam Hussain, "Machine Learning: The Backbone of Intelligent Trade Credit-Based Systems," *Secur. Commun. Networks*, 2022, doi: https://doi.org/10.1155/2022/7149902.

[11] N. X. Nyow and H. N. Chua, "Detecting Fake News with Tweets' Properties," *2019 IEEE Conf. Appl. Inf. Netw. Secur. AINS 2019*, pp. 24–29, Nov. 2019, doi: 10.1109/AINS47559.2019.8968706.

[12] P. H. A. Faustini and T. F. Covões, "Fake news detection in multiple platforms and languages," *Expert Syst. Appl.*, vol. 158, p. 113503, Nov. 2020, doi: 10.1016/J.ESWA.2020.113503.

[13] M. Del Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: early warning of potential misinformation targets," *ACM Trans Web*, vol. 13,

no. 2, pp. 1–22, Apr. 2019, doi: 10.1145/3316809.

[14] Y. Liu and Y. F. B. Wu, "FNED," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, May 2020, doi: 10.1145/3386253.

[15] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, and E. Cambria, "Supervised Learning for Fake News Detection," *IEEE Intell. Syst.*, vol. 34, no. 2, pp. 76–81, Mar. 2019, doi: 10.1109/MIS.2019.2899143.

[16] M. Z. Asghar, A. Habib, A. Habib, A. Khan, R. Ali, and A. Khattak, "Exploring deep neural networks for rumor detection," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 4, pp. 4315–4333, Apr. 2021, doi: 10.1007/S12652-019-01527-4/METRICS.

[17] G. Raja, Y. Manaswini, G. D. Vivekanandan, H. Sampath, K. Dev, and A. K. Bashir, "AI-Powered blockchain - A decentralized secure multiparty computation protocol for IoV," *IEEE INFOCOM 2020 - IEEE Conf. Comput. Commun. Work. INFOCOM WKSHPS 2020*, pp. 865–870, Jul. 2020, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162866.

[18] S. S. Zehra, R. Qureshi, K. Dev, S. Shahid, and N. A. Bhatti, "Comparative Analysis of Bio-Inspired Algorithms for Underwater Wireless Sensor Networks," *Wirel. Pers. Commun.*, vol. 116, no. 2, pp. 1311–1323, Jan. 2021, doi: 10.1007/S11277-020-07418-8/METRICS.

[19] R. K. Kaliyar, A. Goswami, and P. Narang, "DeepFakE: improving fake news detection using tensor decomposition-based deep neural network," *J. Supercomput.*, vol. 77, no. 2, pp. 1015–1037, Feb. 2021, doi: 10.1007/S11227-020-03294-Y/METRICS.

[20] M. H. Goldani, S. Momtazi, and R. Safabakhsh, "Detecting fake news with capsule neural networks," *Appl. Soft Comput.*, vol. 101, p. 106991, Mar. 2021, doi: 10.1016/J.ASOC.2020.106991.

[21] B. Tejaswini, V., "Depression Detection from Social Media Text Analysis using Natural Language Processing Techniques and Hybrid Deep Learning Model.," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, pp. 1–20, 2024.

[22] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006, doi: 10.1146/ANNUREV.BIOENG.8.061505.095802.

[23] S. S. U. Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain, Hussain AlSalman, Mogeeb A. A. Mosleh, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," *Math. Probl. Eng.*, 2021, doi: https://doi.org/10.1155/2021/1162553.

[24] A. Jain and A. Kasbe, "Fake News Detection," *2018 IEEE Int. Students' Conf. Electr. Electron. Comput. Sci. SCEECS 2018*, Nov. 2018, doi: 10.1109/SCEECS.2018.8546944.

[25] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020, doi: 10.1007/S10462-019-09794-5/METRICS.

[26] H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," *Secur. Priv.*, vol. 1, no. 1, p. e9, Jan. 2018, doi: 10.1002/SPY2.9.

[27] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10618 LNCS, pp. 127–138, 2017, doi: 10.1007/978-3-319-69155-8_9.

[28] A. Oussous, A. A. Lahcen, and S. Belfkih, "Impact of text pre-processing and ensemble learning on Arabic sentiment analysis," *ACM Int. Conf. Proceeding Ser.*, vol. Part F148154, 2019, doi: 10.1145/3320326.3320399.

[29] "Stopwords in Technical Language Processing." Accessed: Feb. 22, 2025. [Online].

Available: https://www.researchgate.net/publication/341926808_Stopwords_in_Technical_Language_Processing

[30] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman, "Stemming as a feature reduction technique for Arabic text categorization," *Proc. 10th Int. Symp. Program. Syst. ISPS' 2011*, pp. 128–133, 2011, doi: 10.1109/ISPS.2011.5898874.

[31] M. Bounabi, K. El Moutaouakil, and K. Satori, "A comparison of text classification methods using different stemming techniques," *Int. J. Comput. Appl. Technol.*, vol. 60, no. 4, pp. 298–306, 2019, doi: 10.1504/IJCAT.2019.101171.

[32] A. M. A. Bahzad Taha Jijo, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 1, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[33] A. B. Shaik and S. Srinivasan, "A brief survey on random forest ensembles in classification model," *Lect. Notes Networks Syst.*, vol. 56, pp. 253–260, 2019, doi: 10.1007/978-981-13-2354-6_27.

[34] N. Wahid, A. Zaidi, G. Dhiman, M. Manwal, D. Soni, and R. R. Maaliw, "Identification of Coronary Artery Disease using Extra Tree Classification," *6th Int. Conf. Inven. Comput. Technol. ICICT 2023 - Proc.*, pp. 787–792, 2023, doi: 10.1109/ICICT57646.2023.10134338.

[35] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, Apr. 2020, doi: 10.1007/S11704-019-8208-Z/METRICS.

[36] A. Vereshchaka, S. Cosimini, and W. Dong, "Analyzing and distinguishing fake and real news to mitigate the problem of disinformation," *Comput. Math. Organ. Theory*, vol. 26, no. 3, pp. 350–364, Sep. 2020, doi: 10.1007/S10588-020-09307-8/METRICS.