Truth in the Age of Misinformation: Detecting Fake News

Jamila Sadikot, Mann Dangaria, Darshana Patel, Dishita Mashru, Aditiba Jadeja, Soniya Aghera
Department of Information Technology,
V.V.P. Engineering College, Rajkot.

Abstract :

Determination of fake news has become a crucial research subject, considering the voluminous and misleading social media content, hence, important to determine the authenticity of the news on the internet. This paper conducts a thorough analysis of current research endeavors in fake news detection, particularly emphasizing the effectiveness of traditional machine learning techniques. The primary aim is to develop a robust model using supervised machine learning algorithms to accurately classify news articles as authentic or deceptive. To achieve this objective, various tools and technologies have been used. Through this interdisciplinary approach, it has been tried to contribute to the ongoing efforts to combat the proliferation of fake news and maintain the integrity of information dissemination in the digital realm. In all the graphs developed for accuracy metrics like precision, recall, and F1-score, the Passive Aggressive Classifier consistently shows the best values. It proves to be the most effective classifier for distinguishing between real and fake news articles which proves to be beneficial to society.

Keywords : Machine Learning, Fake News, Passive Aggressive Classifier, TfidfVectorizer

1. Introduction :

In the realm of artificial intelligence, machine learning emerges as a potent tool to counter the spread of fake news. Through the application of intricate algorithms trained on extensive datasets, machine learning facilitates the automated identification of deceptive content, a task fraught with intricacies and challenges.[2,10]

This literature review meticulously examines the pivotal role of machine learning in detecting and mitigating the influence of fake news. It thoroughly explores the methodologies and algorithms employed in this pursuit, shedding light on the escalating significance of machine learning in tackling contemporary issues.

In today's digital landscape, the widespread dissemination of fake news presents formidable obstacles to individuals and institutions alike. [2] With social media platforms such as Facebook and Twitter becoming fertile grounds for misinformation, the propagation of false information has reached unprecedented levels.[13]

Subsequent sections of the review delineate the research methodology, address pivotal research inquiries, and present insightful findings and discussions. Ultimately, this analysis underscores the pressing

necessity for robust detection mechanisms to confront the escalating tide of misinformation pervading online platforms.

The applications of the proposed system include[14]:

- Fact-checking news articles before sharing on social media.
- Verifying sources and content integrity.
- Teaching students media literacy and critical thinking skills.
- Monitoring misinformation campaigns and disinformation tactics.

2. Methodology

A fake news detection system marks a critical stride in confronting the pervasive spread of misinformation across digital platforms. To predict the spreading of this fake news here are the few methods which can be used:

A. Decision Tree:

Decision trees are widely used in Fake News Detection for their interpretability and ability to capture non-linear relationships in data. They partition the dataset into subsets based on feature values, enabling straightforward interpretation of classification rules. However, decision trees may suffer from overfitting, especially with complex datasets containing noisy or irrelevant features.[1]

B. Random Forest:

In Fake News Detection projects, Random Forests offer improved performance over decision trees by aggregating predictions from multiple trees. By training on random subsets of the data and averaging predictions, Random Forests reduce overfitting and enhance generalization.[2] They are resilient to noisy data and handle high-dimensional feature spaces effectively.[15]

C. Passive Aggressive Classifier:

The Passive Aggressive Classifier is favored in Fake News Detection for its ability to handle streaming data and adapt to concept drift. It updates model parameters incrementally, making it suitable for real-time applications where data distribution may change over time. However, its performance may vary based on the nature of the dataset and the characteristics of the features.[3,9,12]

D. K-Nearest Neighbors (KNN):

KNN is a straightforward and intuitive algorithm used in Fake News Detection for its simplicity and ease of implementation. It classifies instances based on the majority class among their nearest neighbors in feature space. While KNN does not require model training, it may suffer from computational inefficiency with large datasets and struggles with high-dimensional feature spaces.[4,16]

E. Gradient Boosting Classifier:

Gradient Boosting is a powerful technique in Fake News Detection, known for its high predictive accuracy and ability to handle imbalanced datasets. By iteratively training weak learners and correcting errors, Gradient Boosting creates strong ensemble models. However, it can be computationally intensive and requires careful hyperparameter tuning to prevent overfitting.[6,14]

2.1 Related Work :

This condensed comparative analysis presents a synthesis of noteworthy findings from a range of research papers. Across diverse studies, common themes emerge regarding emerging technologies, societal trends, and scientific advancements. Through this analysis, patterns and insights provide valuable perspectives for understanding contemporary research trajectories and their implications[1-5]

Reference Paper	Methods	Pros	Cons
End-to-End Learning of Decision Trees and Forests[1]	Decision Tree Classifier	Simple to understand and interpret Can handle both numerical and categorical data Requires little data preprocessing	lack of end-to-end trainability
FAKE NEWS DETECTION USING MACHINE LEARNING EMPLOYING RANDOM FOREST ALGORITHM[2]	Random Forest Classifier	provide stable and reliable predictions compared to single decision tree	However, its interpretability is reduced owing to the amalgamation of decisions from multiple trees.
Automated Identification of False Information in Romanian Online News[3]	Passive Aggressive Classifier	Fast Training Easy adaptability to real world data	Vulnerability to Noisy Data
Fake News Detection System Using Modified Random Forest[4]	K Nearest Neighbour Classifier	is valued for its simplicity and effectiveness in utilizing the k-nearest neighbor algorithm for classification.	its high time complexity limits its performance in big data scenarios
Identification of hyperpartisan news articles employing natural language processing methodologies.[7]	Gradient Boosting Classifier	Gradient boosting includes its ability to handle complex data and its high predictive accuracy.	it may be prone to overfitting and requires careful tuning of parameters.

2.2 Workflow:

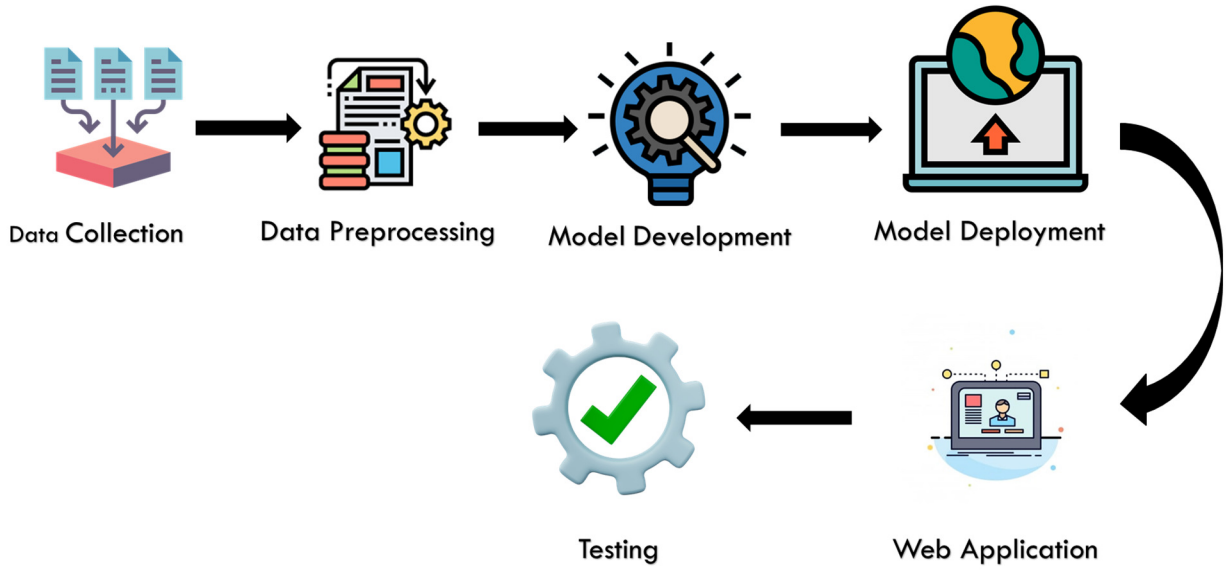

Figure 1: Workflow

The fake news detection system is constructed through a series of systematic stages. Initially, labeled news articles are gathered to create a dataset for training the machine learning model, which aims to differentiate between real and fake news. Following this, the text data undergoes preprocessing steps including stop word removal, tokenization, and vectorization using the TfidfVectorizer.[17] Within the Jupyter Notebook environment, training of the Passive Aggressive Classifier occurs, enhancing its capability to distinguish genuine news from fabricated content. Upon completion, the trained model is integrated into a Flask web application, offering prediction functionalities and a user interface. Thorough testing procedures are conducted to ensure the application's functionality and accuracy, validating its effectiveness in combating misinformation online.

3. Implementation:

This phase in the Fake News Detection process involves selecting and initializing a machine learning model, such as the Passive Aggressive Classifier (PAC) as depicted in the diagram. The chosen model is then trained on the training data to learn the patterns and features of fake and real news articles.

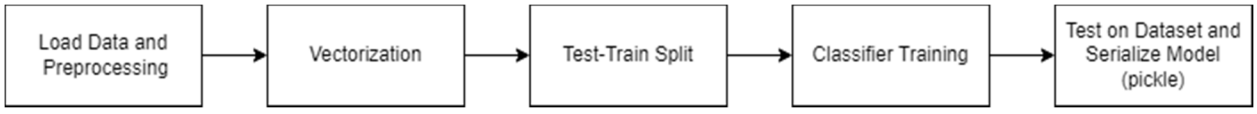

Figure 2 Model Development[18]

Step-by-Step Description about Model Development Workflow:

1. Loading Data and Preprocessing:

In the first step of the Fake News Detection process, the data is sourced from the 'news.csv' dataset, which can be obtained from Kaggle. This dataset contains news articles along with their labels (fake or real). Using Pandas, the 'news.csv' file is loaded into a Pandas DataFrame for further processing. The preprocessing step involves checking for null values in the DataFrame and handling any missing data appropriately to ensure data cleanliness and accuracy.

2. Vectorization:

In the Vectorization step, the text data is transformed into numerical features using the TfidfVectorizer from the sklearn library. This process converts the raw text data into a format that machine learning models can understand and process. TfidfVectorizer stands for Term Frequency-Inverse Document Frequency Vectorizer, which calculates the importance of each word in a document relative to a collection of documents.[8] It creates a matrix where rows represent documents, columns represent unique words, and each cell contains a numerical value representing the importance of that word in the document. This conversion is crucial for training machine learning models on text data, as it allows the algorithms to operate on the numerical data rather than raw text.[19]

3. Test-Train Split:

The Test-Train Split step involves dividing the dataset into two subsets: a training set and a testing set. This is done using the train_test_split function from the sklearn library. The training dataset is utilized to train the machine learning model, whereas the testing dataset is employed to assess its performance. The purpose of this split is to assess how well the model generalizes to new, unseen data. The train_test_split function randomly shuffles the dataset and splits it into specified proportions, typically with a larger portion allocated to the training set (e.g., 80% for training and 20% for testing). By separating the data this way, we can train the model on one portion and validate its performance on another, ensuring that it can effectively classify new, unseen news articles.

4. Classifier Training:

In the Classifier Training step, a machine learning model is selected and initialized, such as the Passive Aggressive Classifier (PAC). The choice of model depends on the problem at hand, and the PAC is often used for its ability to handle online learning and classify data into two classes (fake or real news in this case). Once the model is chosen, it is trained using the training data previously split. During training, the model learns the patterns and features in the training data that distinguish between fake and real news articles[11]. The goal is to create a model that can accurately classify new, unseen news articles based on the patterns it has learned. The training process involves adjusting the model's parameters to minimize errors and improve its predictive performance.[20]

5. Test on Dataset and Serialize Model:

After preprocessing and splitting the data, the machine learning model like the Passive Aggressive Classifier is trained on the training set to learn patterns in fake and real news. It then predicts on the testing set, generating accuracy metrics and a confusion matrix. The model and TfidfVectorizer used for text vectorization are serialized with pickle. For deployment, a Python file ('app.py') is created with Flask for the web app. Running 'app.py' starts the Flask app, accessible at http://localhost:5000, where users can input news text to instantly receive a classification result indicating whether it's likely fake or real. This streamlined process efficiently handles data, model training, serialization, and deployment for a user-friendly fake news detection web application.

Here is the UI Interface for fake news detection web application.

Figure 3: UI for the Fake News Detection System (Prediction: Fake News)

Figure 4: UI for the Fake News Detection System (Prediction: Real News)

4. Detailed Analysis:

Here is the Accuracy measurement for the data set using various classification algorithms and found the graph as mentioned below:

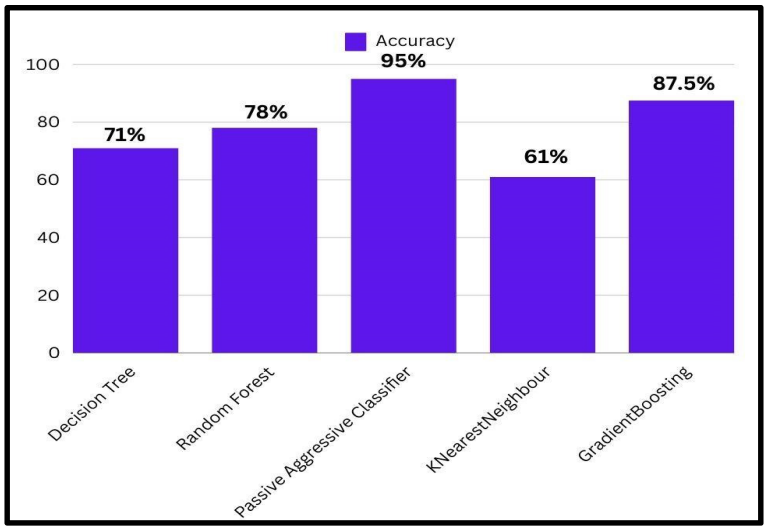

Figure 5. Accuracy measure for various classification algorithms

Figure 5 shows the observed values of accuracy over the various classification algorithms used, viz. Decision tree, Random Forest, Passive Aggressive Classifier, K Nearest Neighbour and Gradient Boosting Algorithm, where Passive Aggressive classifier shows the highest accuracy for the system.[21]

Also, here is the classification report graphs which comprises various accuracy measures such as Precision, Recall, and F1-score. This helps to understand the accuracy of the model and provide the detailed analysis for the various accuracy measures.

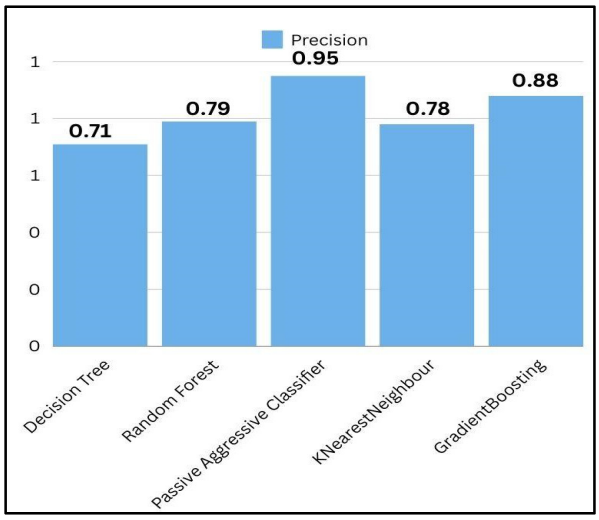

Figure 6 Precision for various classification algorithms

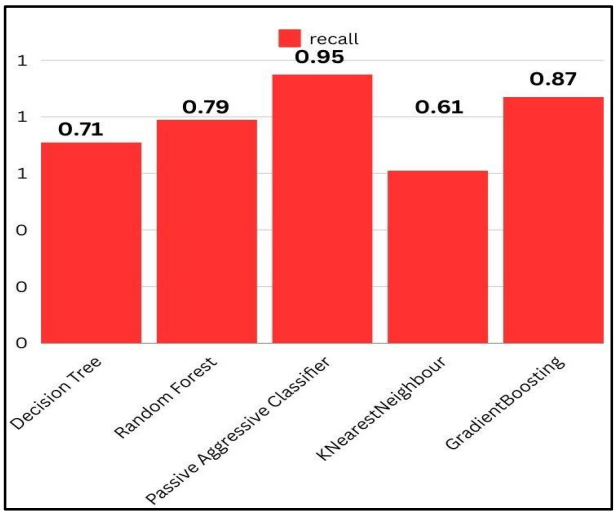

Figure 7 Recall for various classification algorithms

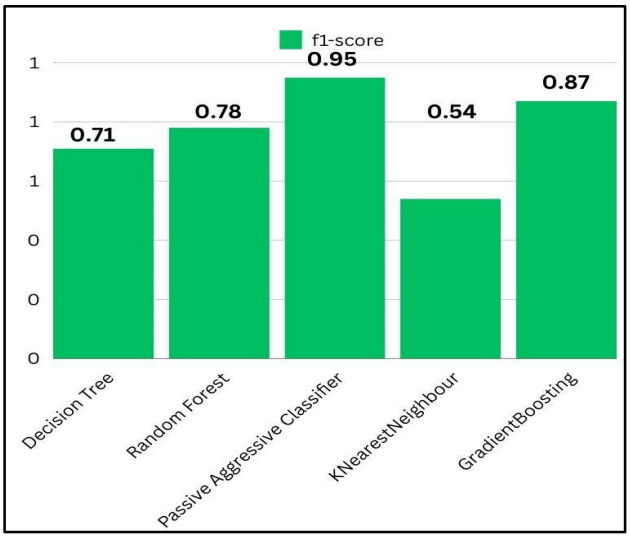

Figure 8 F1-score for various classification algorithms

In all the graphs developed for accuracy metrics like precision, recall, and F1-score, the Passive Aggressive Classifier consistently shows the best values. It proves to be the most effective classifier for distinguishing between real and fake news articles. These findings confirm that the Passive Aggressive Classifier is a reliable choice for detecting fake news.

5. Conclusion:

This study highlights the pressing need to tackle the pervasive issue of fake news in today's digital landscape. Through an extensive analysis of current research trends and methodologies, it aims to contribute to the ongoing efforts to develop effective solutions for detecting fake news. This methodology relies on established machine learning techniques and interdisciplinary collaboration, showcasing the potential of leveraging traditional approaches to address contemporary challenges. By utilizing supervised learning algorithms like the Passive Aggressive Classifier and advanced text processing tools such as the TfidfVectorizer, it aims to equip the model with the ability to accurately distinguish between authentic and deceptive news articles.

Moreover, the incorporation of Flask for web application development and HTML/CSS for user interface design underscores the commitment to creating accessible and user-friendly solutions, ensuring widespread adoption across diverse user groups. As navigating the intricate landscape of misinformation, it is crucial to continually refine the methodologies and embrace emerging technologies and collaborative efforts. By fostering an environment of innovation and cooperation, it can collectively work towards safeguarding the integrity of information dissemination and enhancing societal resilience against the dangers of fake news.

In essence, this research represents a significant step forward in addressing the multifaceted challenges posed by fake news, emphasizing the importance of interdisciplinary collaboration and technological advancement in fostering a more informed and resilient society.

6. References:

1. Thomas M. Hehn1, Julian F. P. Kooij1 , Fred A. Hamprecht2 | End-to-End Learning of Decision Trees and Forests | International Journal of Computer Vision (2020) 128:997–1011

2. Dr.M.Rajeswari, K.Dhana Priya | FAKE NEWS DETECTION USING MACHINE LEARNING EMPLOYING RANDOM FOREST ALGORITHM | © 2022 JETIR June 2022, Volume 9, Issue 6

3. Marius Cristian Buzea, Stefan Trausan-Matu Traian Rebedea | Automatic Fake News Detection for Romanian Online News | Novel Methods and Applications in Natural Language Processing

4. Ikbal Gazalba, Mustakim Mustakim and Nurul Gayatri Indah Reza | Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification | 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) | 10.1109/ICITISEE.2017.8285514

5. S. Selva Birunda; R. Kanniga Devi | A Novel Score-Based Multi-Source Fake News Detection using Gradient Boosting Algorithm | IEEE

6. Tavishee Chauhan, Hemant Palivela |Optimization and improvement of fake news detection using deep learning| International Journal of Information Management Data Insights 1 (2021) 100051

7. Suwarna Gothane, Vydehi Kavali, S. Dinesh Kumar, Ramya Jampani | Fake News Detection and Classification using Natural Language Processing | International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 9 Issue V May 2021

8. Jayashree M Kudari,Varsha V, Monica BG, Archana R | Fake News Detection using Passive Aggressive and TF-IDF Vectorizer | International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 09 | Sep 2020

9. Ms. Aarti Chugh, Dr. Yojna Arora, Mr. Jaivardhan Singh, Mr. Shobhit, Mr. Ronak | Media Manipulation Detection System Using Passive Aggressive | International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-9, Issue-3, May 2021

10. Mr. M.S.V.V.Ramesh, P.L.Surekha, A.Sindhuja, K.Divyaharshini, N.Sekhar | FAKE NEWS DETECTION USING MACHINE LEARNING | International Research Journal of Modernization in Engineering Technology and Science Volume:04/Issue:01/January-2022

11. Sakshi Puranik, Suzan Khan Pathan, Ronak Patel, Prof. Yogita Shelar | Fake News Detection using Passive Aggressive and Naive Bayes | International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 04 | Apr 2022

12. Vaishnavi Kesharwani , Vaishnavi Ladole , Shraddha Tak , Vaishnavi Gaigol , Prof. V. B. Bhagat , Dr. V. R. Thakare | Real and Fake News Detection Smart System Using Passive Aggressive Algorithm (Supervised Machine Learning) | International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) Volume 2, Issue 2, May 2022

13. Aishika Pal, Pranav, Moumita Pradhan | Survey of fake news detection using machine intelligence approach | Data & Knowledge Engineering Volume 144, March 2023, 102118
14. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu | Fake News Detection on Social Media: A Data Mining Perspective | ACM SIGKDD Explorations NewsletterVolume 19 Issue 1 01 September 2017

15. Ken MISHIMA, Nonmember and Hayato YAMANA , Member | A Survey on Explainable Fake News Detection | IEICE TRANS. INF. & SYST., VOL.E105–D, NO.7 JULY 2022

16. Aswini Thota, Priyanka Tilak, Simeratjeet Ahluwalia, Nibhrat Lohia | Fake News Detection: A Deep Learning Approach | SMU Data Science Review Volume 1 Number 3 Article 10

17. Abdulaziz Albahr, Marwan Albahar | An Empirical Comparison of Fake News Detection Using different Machine Learning Algorithms | International Journal of Advanced Computer Science and Applications,Vol. 11, No. 9, 2020

18. Zhang, X., and Ghorbani, A. A. (2019) | An overview of online fake news: Characterization, detection, and discussion | Information Processing & Management, 57, 102025.

19. Conroy, N., Rubin, V., & Chen, Y. (2015) | Automatic deception detection: Methods for finding fake news. | Proceedings of the Association for Information Science and Technology, 52(1)

20. Nida Aslam , Irfan Ullah Khan, Farah Salem Alotaibi , Lama Abdulaziz Aldaej and Asma Khaled Aldubaikil | Fake Detect: A Deep Learning Ensemble Model for Fake News Detection | Hindawi Complexity Volume 2021, Article ID 5557784, 8 pages

21. Reham Jehad and Suhad A. Yousif2 | Fake News Classification Using Random Forest and Decision Tree (J48) | Al-Nahrain Journal of Science ANJS, Vol.23 (4), December, 2020, pp. 4955