Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Fake News Detection Using Machine Learning Algorithms

Priyanshu Garg[a], Priyanshu Sharma[b], Piyush Kumar[c], shivi Malik[d] , Mr. Md. Shahid

Department of CSE,Meerut Institute of Engineering & Technology, Meerut 250110, India

Abstract: In the era of information proliferation, the rise of fake news poses a significant threat to the integrity of digital content. This study uses machine learning algorithms[5] to address the important problem of detecting fake news. Leveraging a dataset comprising both genuine and fabricated news articles, we employ text processing techniques, including TF-IDF vectorization, to transform textual data into a machine-learning model-friendly format. Our study explores the effectiveness of Classifiers such as logistic regression, decision trees, gradient boosting, and random forest in discerning between real and fake news. A carefully selected dataset is used to train the models, and metrics like accuracy, precision, recall, and F1-score are used to thoroughly assess each model's performance. The models' levels of success vary, according to the results, and each has advantages and disadvantages. The study opens the door for further developments in this important field by providing a comparative analysis of machine learning models and insightful information about the state of fake news detection.

Keywords: Machine Learning, Logistic Regression, Decision Tree, Random Forest Classifier, Gradient Boosting Classifier, TF-IDF vectorization.

1. Introduction:

In recent times, one of the paramount research endeavors has centered on the identification of fake news, a phenomenon that has evolved significantly with the advent of digital technology. Previously, fake news thrived in the realm of yellow journalism, often disguised as sensational tales involving crimes, accidents, rumors, and humor. In the digital era, the propagation of fake news is facilitated by the unique attributes of social media, enabling users to easily disseminate misinformation to their networks. The digital landscape introduces complexities, such as variable comments on fake news, diminishing the reliability of authentic information. Fake news spreads swiftly, outpacing genuine news, and its influence can extend to entire populations and even governments.

Various techniques, including machine learning, language analysis, and knowledge-based methods, are employed to detect fake news. The pervasiveness of fake news is exacerbated by the lack of skills among individuals to critically assess and verify news accuracy, leading to a growing focus on fake news detection within the research community.

Recent studies have explored diverse methodologies for identifying fake news, with some concentrating on specific categories like e-commerce or politics. However, these studies often suffer from dataset bias and limited performance when applied to news from different domains. Consequently, there is a critical need to assess the suitability and efficiency of existing models across various news categories on diverse datasets propagated through social media platforms.

1.1. TF-IDF Vectorization:

TF-IDF Vectorization is a metric that quantifies the importance of a word to a document in a corpus or collection, taking into account the fact that some words are more commonly used than others[1]. Processed text data is converted into numerical vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This method weights terms according to their importance in the corpus, incorporating both term frequency and document frequency.

1.2. Machine Learning Algorithms:

- **Logistic Regression (LR):** The statistical technique known as logistic regression is mostly applied to binary classification. By using the sigmoid function to translate real values into probabilities between 0 and 1, it forecasts the likelihood that an instance will belong to a class. The classification result is determined by the decision boundary, which is usually set at 0.5.[4].

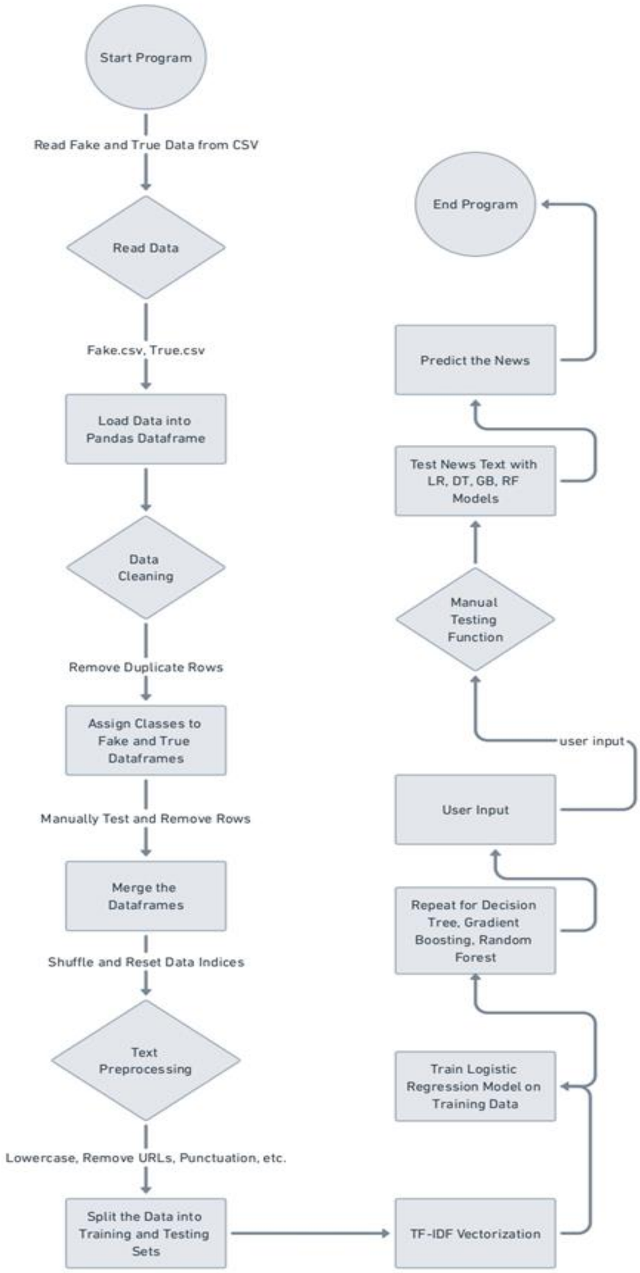
- **Decision Tree (DT):** A flexible algorithm applicable to both regression and classification tasks, the Decision Tree generates a tree-like structure with nodes representing decisions based on specific features. The tree is constructed by recursively splitting the dataset, and the final prediction is made at the leaves (terminal nodes)[3].

- **Gradient Boosting (GB):** Gradient Boosting is a technique for ensemble learning that combines predictions from multiple weak learners, often decision trees. It builds trees sequentially, with each tree correcting errors made by the previous ones. The process aims to minimize residuals (differences between predictions and actual values) using gradient descent. Hyperparameters such as learning rate control the contribution of each tree to the ensemble.

- **Random Forest (RF):** With the use of multiple decision trees built during training, Random Forest is an ensemble learning technique that produces the mean prediction in regression or the class mode in classification. Each tree is trained on a bootstrapped sample with feature subsetting, introducing randomness. The final prediction is determined by majority voting, while regression is determined by averaging. The method leverages bagging (Bootstrap Aggregating) to reduce variance by combining predictions from independently trained models.

2. Method Used:

The flowchart in Figure 1 provides an overview of the steps involved in constructing the model and evaluating its performance. It starts with Data Acquisition and Processing, where information is extracted from "Fake.csv" and "True.csv" datasets representing distinct news article classes. Subsequent Data Exploration unravels dataset structures. The process includes Data Labelling and Shuffling, introducing binary labels, and optional subset removal for controlled testing. Text Processing involves a systematic pipeline, utilizing wordopt for lowercasing, URL removal, and punctuation removal. Transitioning to Model Training and Testing, the dataset is split, and models undergo evaluation using comprehensive metrics. Model Comparison identifies the most effective algorithm, considering strengths and weaknesses. Manual Testing integrates real-time user input for practical scenarios, completing a comprehensive framework for effective fake news detection.



2.1 Data Acquisition and Processing:

In the initial phase of our methodology, we delved into the pivotal step of data acquisition, laying the groundwork for discerning the authenticity of news articles. This involved the meticulous extraction of information from two distinct datasets: "Fake.csv" and "True.csv." These datasets were carefully curated to represent specific classes of news articles, setting the stage for a comprehensive understanding of the broader corpus.

Subsequently, we conducted an insightful exploration of the datasets, examining the initial entries to unravel their structures, features, and contents. This exploration served as a crucial precursor to further steps in our methodology.

Data Labelling and Shuffling:

To enhance the datasets, we introduced a binary label, 'class,' categorizing news articles as either genuine or fake (0). An optional yet valuable step involved the removal of specific subsets of data for controlled manual testing, allowing for an isolated examination of model robustness. The datasets were then merged to create a consolidated corpus, and random shuffling of entries ensued. This shuffling process was implemented to eliminate biases and ensure a well-distributed dataset, setting the stage for subsequent model training and testing.

Text Processing:

Our systematic text processing pipeline played a pivotal role in preparing the textual data for analysis. Leveraging the wordopt function, this pipeline encompassed crucial steps such as lowercasing, ensuring uniformity in the text by converting it all to lowercase. Furthermore, the pipeline included the removal of URLs to reduce noise in the data, ensuring a cleaner and more focused dataset for subsequent analysis.

As we navigate through these key stages of data acquisition and processing, we lay a solid foundation for the subsequent phases of our methodology, including model training, testing, and result analysis. Each step is designed to contribute to the overall goal of developing a robust system for the detection of fake news, grounded in thorough data understanding and effective preprocessing.

2.2 Traning and testing model:

In the critical phase of Model Training and Testing, our methodology prioritized effective evaluation strategies and model selection. First, the dataset underwent meticulous Data Splitting, with a division into training (75%) and testing (25%) sets. This separation facilitated the evaluation of model performance on unseen data.

Each machine learning model underwent multiple Training and Testing Iterations, where it was trained on the designated training set and rigorously evaluated on the reserved test set. This iterative process ensured a thorough understanding of each model's behavior and predictive capabilities.

In the subsequent Evaluation Metrics stage, a suite of comprehensive metrics was employed to assess model performance rigorously. Accuracy (overall prediction correctness), Precision (proportion of true positive predictions among all positive predictions), Recall (proportion of true positive predictions among all actual positive instances), and F1-Score (the harmonic mean of precision and recall) were among the metrics used.

. This thorough evaluation laid the foundation for informed decision-making in subsequent steps.

The Model Comparison and Selection phase involved a meticulous analysis of results from each model. Strengths and weaknesses were carefully scrutinized, providing valuable insights into their applicability and performance across different scenarios. This comparative analysis was pivotal in identifying the most effective algorithm for fake news detection.

To ensure real-world applicability, our methodology incorporated Manual Testing. This feature allowed real-time user input of news articles, enabling practical testing scenarios. Users could input news articles for immediate evaluation using all trained models. This real-time user interaction not only enhanced the user experience but also offered crucial insights into the models' performance on diverse inputs.

In summary, our approach to Model Training and Testing emphasizes a comprehensive evaluation, model comparison, and real-world applicability through manual testing. This robust methodology ensures the selection of a highly effective algorithm for fake news detection, grounded in thorough analysis and user engagement.

3. Experimental Results:

The results are derived from training a model on a dataset comprising both True and False News. Models are effective at learning features for True News, especially when trained on larger datasets, leading to improved results. The specific dataset used for this purpose consists of more than Forty thousand five hundred true news and false news. We have trained the model news dataset.

It is clearly depicted by the figure that the accuracy score has reached more than 99% for all four models.

Fig. 2 Evaluation table for logistic regression

Accuracy: On the test set, the logistic regression model had an accuracy score of [0.986096256684492].

Classification Report:

Insights: Logistic regression, being a simple and interpretable model, demonstrated:

Strengths: Simple and interpretable, well-suited for binary classification tasks.

Considerations: May struggle to capture complex non- linear relationships in the data.

```
accuracy score for logistic regression : 0.986096256684492
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      5958
           1       0.98      0.99      0.99      5262

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

Fig. 3 Evaluation table for decision tree

Accuracy: On the test set, the decision tree model had an accuracy score of [0.9959001782531194].

Insights: Decision trees, with their ability to capture non-linear relationships.

Strengths: Can capture non-linear relationships and is interpretable.

Considerations: Prone to overfitting, might not generalize well to unseen data without proper tuning.

```
accuracy score for DecisionTreeClassifier : 0.9959001782531194
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      5958
           1       0.99      1.00      1.00      5262

    accuracy                           1.00     11220
   macro avg       1.00      1.00      1.00     11220
weighted avg       1.00      1.00      1.00     11220
```

Fig. 4 Evaluation table for gradient boosting classifier

Accuracy: On the test set, the gradient boosting classifier had an accuracy score of [0.9959001782531194].

Insights: Gradient boosting, through its sequential learning process, showcased:

Strengths: Sequential learning minimizes errors, and handles complex relationships well.Considerations:Computationally expensive,may require tuning of hyperparameters.

```
accuracy score for GradientBoostingClassifier : 0.9959001782531194
              precision    recall  f1-score   support

           0       1.00      0.99      1.00      5958
           1       0.99      1.00      1.00      5262

    accuracy                           1.00     11220
   macro avg       1.00      1.00      1.00     11220
weighted avg       1.00      1.00      1.00     11220
```

Fig. 5 Evaluation table for random forest classifier

Accuracy: On the test set, the random forest model had an accuracy score of [0.9864527629233512].

Insights: Random forest, with its ensemble of decision trees, demonstrated:

Strengths: Robust, handles overfitting, and provides feature importance.

Considerations: The ensemble approach might be computationally intensive.

```
accuracy score for RandomForestClassifier : 0.9864527629233512
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5958
           1       0.99      0.99      0.99      5262

    accuracy                           0.99     11220
   macro avg       0.99      0.99      0.99     11220
weighted avg       0.99      0.99      0.99     11220
```

4. Conclusion:

The news detection models leverage diverse algorithms, including Logistic Regression, Decision Tree, Gradient Boosting, and Random Forest, to effectively identify and classify news articles as genuine or fake. These models integrate the expertise of machine learning specialists with the ability to extract significant features indicative of fake news. The evaluation results showcase remarkable accuracy for each algorithm, with Decision Tree and Gradient Boosting models achieving particularly high accuracy levels. The trend in employing different machine learning algorithms for news detection is evident, emphasizing continuous efforts to enhance accuracy and effectiveness compared to conventional methods.

References:

[1] Wikipedia contributors. (2024). TF-IDF vectorization. Wikipedia, The Free Encyclopedia.https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[2] C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," 2017 IEEE International Conference on Smart Cloud (SmartCloud), 2017, pp. 208- 215, DOI: 10.1109/SmartCloud.2017.40.

[3] Gupta and R. Kaushal, "Improving spam detection in Online Social Networks," 2015 International Conference on Cognitive Computing and Information Processing(CCIP), 2015, pp. 1-6, DOI: 10.1109/CCIP.2015.7100738.

[4] Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake News

[5] Detection Using Machine Learning Ensemble Methods,” Complexity, 17-

[6] Oct-2020.

[7] De Beer, Dylan, and Machdel Matthee. “Approaches to Identify Fake News: A Systematic Literature Review.” Integrated Science in Digital Age 2020 vol. 136 13–22. 5 May. 2020, doi:10.1007/978-3-030-49264-9_2