A Comparative Study of Hybrid Machine Learning Approaches for Fake News Detection that Combine Multi-Stage Ensemble Learning and NLP-based Framework

SUBMISSION DATE / POSTED DATE

10-01-2023 / 27-02-2023

CITATION

Singh, Gaurav; Selva, Kamal (2023): A Comparative Study of Hybrid Machine Learning Approaches for Fake News Detection that Combine Multi-Stage Ensemble Learning and NLP-based Framework. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.21856671.v3

DOI

10.36227/techrxiv.21856671.v3

A Comparative Study of Hybrid Machine Learning Approaches for Fake News Detection that Combine Multi-Stage Ensemble Learning and NLP-based Framework

Gaurav J. Singh
Department of Metallurgy and Materials Engineering, NIT-Tiruchirappalli

Dr. Selvakumar K
Department of Computer Applications, NIT-Tiruchirappalli

ABSTRACT: Fake News has been spreading widely throughout the world as the booming internet era has started worldwide. Now, more people have access to the Internet than ever, which has led to a significant rise in spreading fake news. So, to solve this issue, it would be highly impossible to manually remove every phony news article. To tackle the above problem of checking on information related to the source, content, or news publisher to categorize it as genuine or fake, we take the help of Machine Learning to classify the information on the web as True or False. Therefore this paper explores the different types of ML classifiers to detect fake news. Therefore, this study will use textual properties of the news dataset we took from Kaggle to distinguish a piece of news as fake or real. Furthermore, with these properties, we will train our model using different ML classification algorithms to evaluate the performance of the dataset collected.

KEYWORDS: *fake news detection, text classification, machine learning, NLP, ensemble classifiers*

1. INTRODUCTION

In the Internet age, most people spend the majority of their time on their mobile phones. While the younger generation consumes news through social media or online news blogs [1], the older generation spends their leisure time watching the news on TV or reading it in the newspaper [2], which is readily available with a single click. As a result, there is no longer a need to purchase a newspaper and read it. However, with such latitude, we have seen an all-time surge in the prevalence of fake news on the Internet and social media [3]. Anyone on the internet/social media may publish whatever they want, making traditional fact-checking nearly impossible. Along with modern-day journalism, it has resulted in an increase in fake news, which is easily accessible on the Internet for anybody to read.

Fake News, as defined by various other authors, comprises misleading content that may deceive readers and fabricated stories that appear to have legitimate sources [4][5]. Fake news is mainly disseminated via the internet, websites, or social media. These websites attempt to appear legitimate by naming themselves after legitimate websites. And some of these websites will vanish as soon as the intended results are obtained through fake news promotion.

Fake News has certain standard features or characteristics, making it easier to determine whether it is fake or real. Those include low facticity, such as misleading or deceptive material; journalistic styles, such as structural elements like headline, text, and body; and intention to deceive for personal benefits, such as financial, political, or to provoke someone are examples of these[6][7]. Other characteristics include grammatical and spelling errors or, if a reputable fact checker does not verify the articles, the content is not ascribed to any original news source [8].

Even though many websites, such as Politifact and Snopes, frequently check the news for legitimacy to keep the public informed about which news is fake or real. Many researchers have also been developing repositories to identify which web pages are fraudulent or genuine [9].

Any news that comes out of a publishing house contains information content that reaches out to consumers of daily news. The news is then put up on the internet by bloggers, online news agencies, and social media platforms. Sometimes the news is fabricated on social media or anonymous blogs on the internet, causing upheaval or riots in modern society [10]. As a result, it is vital to keep bogus news in check. However, the traditional fact-checking approach is labor-intensive, time-consuming, and inefficient. As a result, we require a more robust method, such as the usage of advanced Machine Learning algorithms, to assist us in classifying certain news as false or genuine based on the semantics or attributes of the news.

This paper has been divided into sections and subsections from hereon:

Section 1 - Introduction: Provides a quick overview of fake news and its characteristics as well as an introduction to the entire research literature that will be discussed in this paper.

Section 2 - Related Work: This section provides an overview of earlier studies and their contributions to the field of fake news detection.

Section 3 - Research Problem: This section discusses the difficulties with fake news detection that will be addressed in the paper.

Section 4 - Proposed Method: In this section, the project pipeline is discussed, along with a summary of the datasets used and the models that will be used to train the datasets.

Section 5 - Implementation and Results: This section discusses the best-performing model of all the models that have been proposed, as well as its performance indicators.

Section 6 - Conclusion: Concluding the essay with anecdotal statements and possible future research in this area.

2. RELATED WORK

There has been much talk around the research community discussing how to detect fake news. The topic stated above has been addressed in numerous research articles that have been published. The paper “Fake News Detection Using Machine Learning Ensemble Methods” authored by Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad, used the ML ensemble methods to automate the classification of news articles. They used four real-world datasets from the Politifact dataset from Kaggle, 2 Kaggle datasets, and an IOST dataset. The ensemble method, such as XGBoost, gave a better result than the individual learners, such as SVM, KNN, Wang-CNN, and Wang-Bi-LSTM [11].

Another study by Barbara Probierza, Piotr Stefanski, and Jan Kozaka titled "Rapid identification of bogus news based on machine learning approaches." highlighted how one could use classical ensemble methods to attain a higher level of accuracy while using NLP to define the text and title of news articles [12]. The major objective was to identify fake news solely by looking at news article titles. SVM was the model that performed the best, with an accuracy of 94.19%. The same analysis was also performed using text as the only feature for comparison's sake, and Bagging was proven to be the best model, with a 99.64% accuracy rate. Along with the model's performance based on accuracy, the paper also focussed on the running time [13], on which random forest algorithm worked best. One of the research papers by Anjali Jain, Avinash Shakya, Harsh Khatter, Harsh Khatter, and Amit Kumar Gupta focussed on developing a hybrid Algorithm combining SVM and Naive Bayes classifier and using NLP to extract textual features to achieve an accuracy of 93.50% [14].

Z Khanam, B N Alwasel, H Sirafi, and M Rashid focused on detecting fake news by using Supervised ML and NLP for textual analysis [15]. They used a quantitative approach by adding more features known as POS textual analysis to the existing models. Only qualitative analysis was done previously using the title and word frequency [16][17]. The highest accuracy achieved was 73% through XGBOOST on the benchmark politics fake news "LIAR" dataset [18].

Ankit Kesarwani, Sudakar Singh Chauhan, and Anil Ramachandran Nair proposed solving the fake news detection problem using K-Nearest Neighbors on the Facebook news dataset from BuzzFeed. There were around 2000 datasets with five features: account_id, post_id, share_count, comment_count, and Rating. The model was trained by splitting the dataset into training and testing in the 80:20 ratio. The accuracy achieved was 79% with the KNN model on the test set and 75% with precision and recall is 79% [19].

Awf Abdulrahman; Muhammet Baykara, with an average accuracy of 91.23% across all the algorithms employed in the paper using NLP, ML, and Deep Learning Algorithms, concentrated on classifying fake news based on social networking websites [20]. By combining the usage of ML and NLP on a sizable and labeled corpus provided by Twitter, Chun-Ming Lai, Mei-Hua Chen, Endah Kristiani, Vinod Kumar Verma, and Chao-Tung Yang attempted to categorize the news data in another work. The conventional ML algorithms' accuracy was 85% and more than 90% with neural networks [21]. Abdulaziz Albahr and Marwan Albahr used the well-known "LIAR" dataset to create a model for detecting fake news. The paper examined four algorithms—Random Forest, Neural Network, Naive Bayes, and Decision Trees—were primarily examined. The Naive Bayes classifier was the model that performed the best [22]. Velivela Durga Lakshmi and Ch Sita Kumari used text, title, and author as parameters for their research in detecting fake news. They converted them to vectors that used Term Frequency - Inverse Document Frequency (TF-IDF) and Count Vectorizers. Furthermore, PCA was used to reduce the dimensions. The Random Forest method was subsequently constructed and was discovered to be the best model regarding accuracy, precision, and recall [23].

2.1. RESEARCH PROBLEM

The fundamental issue this paper addresses is detecting/filtering out misleading news from online websites, which are a user's primary source of news intake. Many research studies on the detection of fake news have yielded positive findings. The sole limitation is that the dataset utilized is limited to political data, as in IOST [24][25] or the benchmark "LIAR" dataset from the 2016 presidential elections [18]. The missing link was the diversity of the field of news taken in the dataset, which may make the entire model less accurate when a new input relating to news outside of the political domain is entered into the model. We will also investigate which ML algorithm best fits the dataset available online.

3. PROPOSED METHOD

4.1. RESEARCH PIPELINE

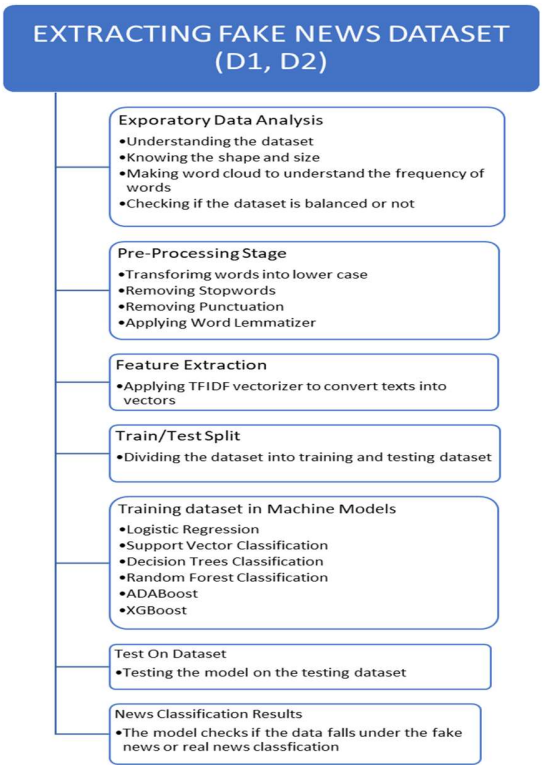

Fig 1: Research Pipeline

4.2. DATASET

Many real and fake news datasets are freely accessible on the internet. News websites such as the New York Times, Reuters, and the Washington Post, among others, are considered to be reliable sources of information. So the data we'll be using comes from websites like PolitiFact, Snopes, and Boom Live, all of which are fact-checking websites that host all kinds of news, from fake to real, from various sources on the internet, and also check to see if the articles are fake or real. We used two separate datasets for this article to be more inclusive of all forms of news. Previously, most papers centered around the PolitiFact website's Politics Dataset. As a result, we have included news from Entertainment, Politics, Sports, and other areas in this paper.

4.2.1 ISOT Dataset

ISOT Dataset (D1) contains fake and genuine news datasets. Real news, with 21417 rows, was gathered from websites such as Reuters.com, whereas Fraudulent news, with 23481 rows, was obtained from various sources, including websites highlighted by Politifact. The entire dataset includes 44898 news articles in total. The collection includes news from a variety of disciplines, although the majority of it is focused on Politics and World News [24][25]. The dataset has five columns: title, text, subject, date, and label.

Fig 2: Common Words in Fake News Titles

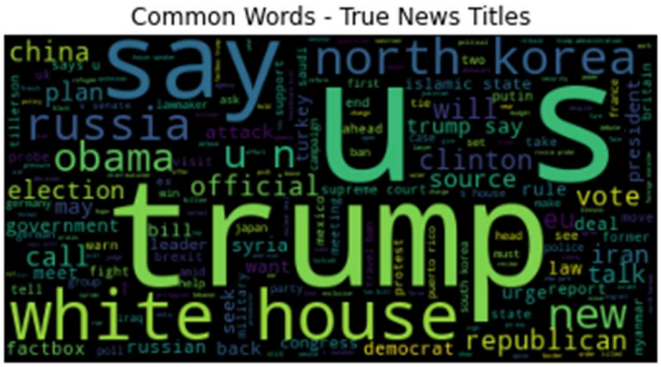

Fig 3: Common Words in True News Titles

4.2.2 Kaggle Dataset

The UTK Machine Learning Club maintains the Kaggle dataset (D2), which is collected from Kaggle [26]. This dataset is divided into two files: the training dataset (20386 articles) and the test dataset (5126 articles), which will be used to test the model's performance. This dataset includes news from numerous topics such as entertainment, crime, and sports and is not limited to Politics. The dataset has five columns: id, title, author, text, and label.

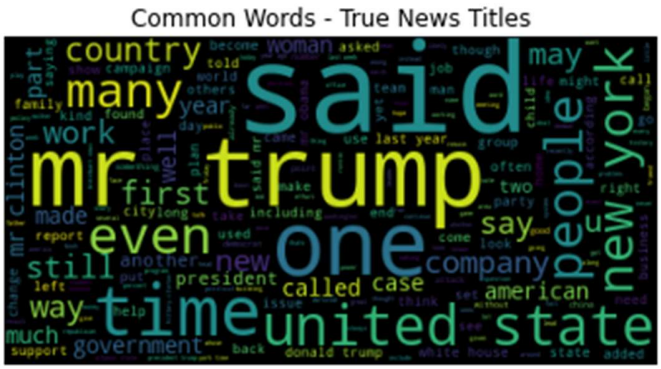

Fig 4: Common Words in Fake News Titles

Fig 5: Common Words in True News Titles

4.3. CLASSIFIER MODELS

4.3.1 Logistic Regression

Logistic Regression is a Supervised Machine Learning model that takes a series of independent variables as features and uses the Sigmoid Function to generate a probability value between 0 and 1 for a categorical dependent variable. It calculates the link between the dependent and independent variables. Linear Regression is a subset of Logistic Regression.

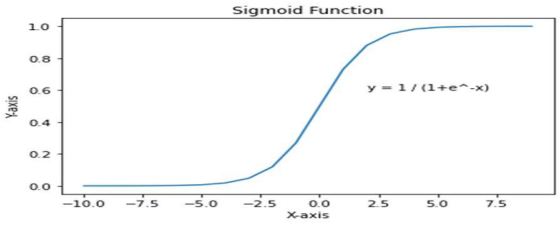

Fig 6: Sigmoid Function Graph

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots\ldots + b_nx_n$$

Equation 1: Linear Regression Equation

$$f(y) = \frac{1}{1 + e^{-y}}$$

Equation 2: Sigmoid Function

$$f(y)_1 = \frac{}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots\ldots + b_nx_n)}}$$

Equation 3: Hypothesis Equation/Mapping Linear regression in Sigmoid Function

$$Cost(h_\theta(x), y) = \begin{cases} \\ -log(1-h_\theta(x)) & if\ y = 0 \end{cases}$$

Equation 4: Cost Function for Logistic Regression

The sigmoid function helps us in mapping predicted values of different independent variables into probabilities called Hypothesis Equation. The main objective is to minimize the cost function for better accuracy. After mapping the values on the Sigmoid Function, a threshold value is set, so any value above the threshold value is labeled as one and less than the threshold value as 0. There are three types of Logistic Regression:

1. Binary Logistic Regression: During classification, the target variable can only attain two outcomes, i.e., True or False, 0 or 1, Spam or Not spam.
2. Multinomial Logistic Regression: During classification, when the target variable can attain three possible outcomes, i.e., classification of different types of news such as Business, Entertainment, Politics, etc.
3. Ordinal Logistic Regression: During classification, when the target variable can only attain two outcomes, i.e., feedback from people as Excellent, Good, Average, Bad, Very Bad, etc.

4.3.2 Support Vector Classification

Support Vector Machine (SVM) is a classification algorithm. The fundamental goal of SVM is to find the best-fit decision boundary line, also known as hyperplane, and use it to divide the n-dimensional space into separate classes. So, when a new dataset appears, we can quickly determine which category it belongs to.

$$J(\theta) = C\sum_{i=1}^{m} [y^{(i)}cost_1(\theta^T x^{(i)}) + (1 - y^{(i)})cost_0(\theta^T x^{(i)})] + \frac{1}{2}\sum_{i=1}^{n} \theta_i^2$$

Equation 5: Cost Function for SVM

In Fig 8, the equation of cost function is given. We find the best fit hyperplane by minimizing the cost function, i.e., J(θ) for SVM. Here $cost_1$ represents the cost when y=1 and $cost_0$ represents the cost when y = 0 [27], such that;

$$\theta^T x^{(i)} \geq +1\ if\ y^{(i)} = 1$$

$$\theta^T x^{(i)} \leq -1\ if\ y^{(i)} = 0$$

We have a range of possibilities for how hyperplanes can exist in the N-dimensional space, so to identify the best plane that can categorize the two classes, SVM works by choosing the data points (vectors) which are at extreme points also called as support vectors to make up the hyperplane between two classes, with the margin being the maximum distance between the hyperplane and the support vectors. An Optimal Hyperplane is a plane with the maximum margin. So, given a binary classification, such as distinguishing between fake and authentic news, the hyperplane will divide the plane into two distinct classes.

4.3.3 Decision Tree Classifier

Decision Tree is a Supervised ML algorithm. It is based on the structure of a tree-based classifier. The Decision tree has mainly two nodes:

1. Decision Nodes
2. Tree Nodes

Decision Nodes contain multiple branches representing the different features used to make a decision based on the if-else condition. The leaf node represents the outcome of the decisions taken in the decision node. The decision tree initiates at the root or parent node, compares the values to the real dataset, and then moves to the sub-node based on the results. This process is repeated until the last node, the leaf node, is reached.

4.3.4 Random Forest Classifier

Random Forest is a supervised machine learning algorithm. It is employed in both regression and classification. The Random Forest approach is based on the notion of ensemble learning. It is a mix of multiple decision trees on subsets of the dataset that analyses the average of all the accuracies gained on each decision tree to enhance the model's overall accuracy. The bigger the number of decision trees, the greater the accuracy and the lower the possibility of overfitting.

4.3.5 Boosting Ensemble Classifiers

In the Machine Learning domain, we have many boosting ensemble classifiers. Although, for this paper, we will be using AdaBoost, which was first proposed by Freund and Schapire in 1997 [28], and the XGBoost algorithm, which Tianqi Chen and Carlos Guestrin first developed in 2016 [29]. It works by combining multiple classifiers to increase the overall accuracy of the classifier model. It combines various poor or weak classifiers to build a robust classifier with higher accuracy than a single weak classifier. The multiple weak learners present in AdaBoost can be Decision Trees, Neural Networks, SVM, etc. [30]. It works on the iterative steps. After each iteration, the model puts more focus or weightage on the training set where the prediction went wrong. It goes until the model achieves the highest accuracy. These are one of the most powerful algorithms which predict the target with very high accuracy.

5. IMPLEMENTATION AND RESULTS

5.1. METHODOLOGY

In this research paper, a new model is being developed to help us detect fake news beforehand. For the D1 dataset, we combined the two CSV files, namely fake.csv, which contains all the fake news labeled as 1, and true.csv, which contains all the authentic news labeled as 0, into one data frame. For the D2 dataset, we have one CSV which contains a labeled target variable and can be used as it is.

APPROACH / ALGORITHM / PSEUDO-CODE:

Input: *("TITLE", "TEXT")*

Modified Input: *"CONTENT" = "TITLE" + "TEXT"*

Output: *label*

```
for each row in "CONTENT":
        lower_case = lower(row)
        stopwords_free = remove_stopwords(lower_case
)
        punctuation_free =
remove_punctuation(stopwords_free)
        lemmatizer = lemmatize(punctuation_free)
content = vectorizer("CONTENT")
training_data, testing_data = test_train_split(content)
classifier_model = fit_model_training(training_data,
best_performing_algorithm)
testing_model = predict(testing_data)
return label
```

As mentioned in section 4.1, the methodology follows this structure:

1. Exploratory Data Analysis: to understand the data such as shape, names of columns, and value counts of the label column to understand if the dataset is balanced or not concerning the label column, making word clouds to know the maximum frequency of words that are part of fake or genuine news titles.
2. Data Pre-Processing:
    a. The dataset was cleaned before it could be used for making models.
    b. Removed all the rows with even one empty cell in the 'text' or 'title' column.
    c. We are mainly concerned with the 'text' and 'title' columns. Therefore we have combined these two columns into one

new column, ‘content’. More details about the news articles would help us train the model better.

d. We have transformed every character into the same lower case removed all the punctuation characters such as !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~” and numeric integers.
e. Tokenization is also performed to remove all the common stop words found in English, such as ‘i’, ‘me’, ‘we’, etc. The stopwords are not vital as they are common words used almost everywhere, so we discard these words for better model performance.
f. Furthermore, we use WordNetLemmatizer() to combine similar-meaning words as one term. In short, we have taken similar-meaning words only once in our modeling process.

3. The next step is to utilize TfIdfVectorizer() to convert texts into feature vectors that can be used as input. The TFIDF algorithm's primary objective is to ensure that the more times a word appears in a sentence, the greater the word's weightage when defining the sentence's meaning [31].
4. Train/Test Split: Now that the data is preprocessed and ready to be inserted into the model for training, we will split our dataset into train and test split with 80% as the training dataset and 20% as the training dataset.

Along with hyperparameter tuning within the different classifier models that are being used for the study, we achieved excellent results on both the selected datasets compared to other papers that have used the same datasets for the study of the detection of fake news [11][12]

5.2. PERFORMANCE METRICS

To compute or measure how good the model is, we need to have performance metrics that help us decide how good the model is performing. And for that, we have Accuracy, Precision, Recall, and F1 Score. And all of these metrics are based on the Confusion Matrix (represented below in Fig 1).

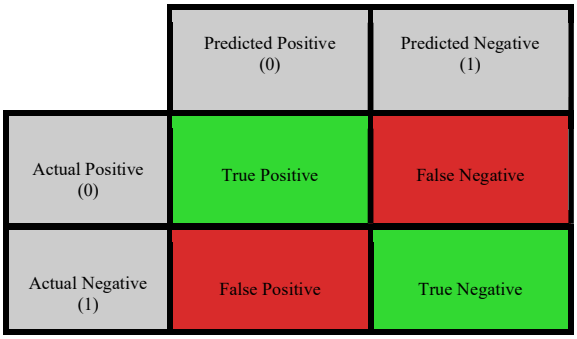

Fig 7: Confusion Matrix

Machine Learning, which uses Classification models to segregate data in a binary classification (Yes/No, 0/1, True/False), is defined in a table as shown in Fig 1. The Confusion Matrix consists of four different parameters:

1. True Positive (TP): This block represents all the cases where the target variables are predicted as Positive and are also actually Positive.
2. True Negative (TN): This block represents all the cases where the target variables are predicted as Negative and are also actually Negative.
3. False Positive (FP): This block represents all the cases where the target variables are predicted as Positive but are actually Negative. These are also termed Type I Error.
4. False Negative (FN): This block represents all the cases where the target variables are predicted as Negative but are actually Positive. These are also termed Type II Error.

5.2.1 Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 6: Accuracy Calculation

Accuracy is one of the most used performance metrics out of the four. It tells us about the percentage of correct predictions out of the total that were made using a particular Machine Learning model. The equation to calculate the Accuracy is given in Fig 2.

A high accuracy value indicates that the model is good. However, this is not always the case; for example, if the dataset is imbalanced, a better accuracy may not lead to a successful model with a newer dataset. And, in classification tasks, a news article that is genuinely accurate but is predicted as False may have a negative impact. As a result, more performance indicators, including Precision, Recall, and F1-score, are required to determine how well a model performs. As a result, in this work, we will address the aforementioned performance indicators and accuracy to compute a model's overall performance.

5.2.2 Precision

$$Precision = \frac{TP}{TP + FP}$$

Equation 7: Precision Calculation

Precision is defined as the ratio of correct positive predictions to total cases in which the target variable was predicted to be positive. Precision less than 0.5 indicates that the number of False Positives is extremely high.

5.2.3 Recall

$$Recall = \frac{TP}{TP + FN}$$

Equation 8: Recall Calculation

Recall is defined as the ratio of valid positive predictions to the total number of instances where the target variable was actually positive. Recall values less than 0.5 indicate a significant number of False Negatives.

5.2.4 F1 Score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Equation 9: F1 Score Calculation

Since Precision and Recall alone are not very good performance metrics for evaluating a model's performance, we have F1 score that incorporates both Precision and Recall values by taking the Harmonic Mean of the data. In other words, it considers both the False Positive (FP) and False Negative (FN) into account.

5.3. COMPARISON OF MODELS

SR. NO	CLASSIFICATION MODELS	ACCURACY	PRECISION	RECALL	F1 - SCORE
1	Logistic Regression	0.9892	0.9899	0.9874	0.9886
2	SVC	0.9941	0.9950	0.9925	0.9938
3	Decision Trees	0.9952	0.9930	0.9970	0.9950
4	**Random Forest**	**0.9987**	**0.9989**	**0.9984**	**0.9987**
5	AdaBoost	0.9979	0.9978	0.9978	0.9978
6	XGBoost	0.9983	0.9980	0.9984	0.9982

Fig 7: Performance Metrics of D1 dataset

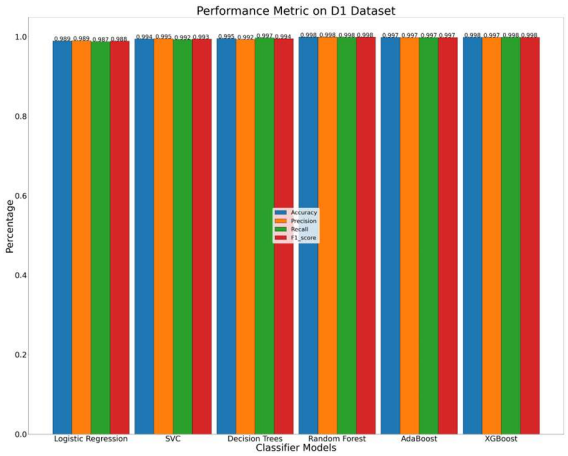

Fig 8: Performance Metrics of various models visualized graphically on the D1 dataset.

Fig 10 represents the accuracies achieved with different models on the ISOT dataset (D1). We can see from the figure that the highest accuracy achieved was 99.87% via Random Forest (Bagging classifier), closely followed by Ensemble Boosting classifier XGBoost with an accuracy of 99.83%. Also, the highest precision, recall, and F1-score was achieved by Random Forest, marked in bold color.

SR. NO	CLASSIFICATION MODELS	ACCURACY	PRECISION	RECALL	F1 - SCORE
1	Logistic Regression	0.9514	0.9456	0.9591	0.9523
2	SVC	0.9613	0.9634	0.9612	0.9623
3	Decision Trees	0.9385	0.9408	0.9390	0.9399
4	Random Forest	0.9536	0.9550	0.9545	0.9547
5	AdaBoost	0.9683	0.9686	0.9695	0.9690
6	**XGBoost**	**0.9752**	**0.9705**	**0.9810**	**0.9757**

Fig 9: Performance Metrics of D2 dataset

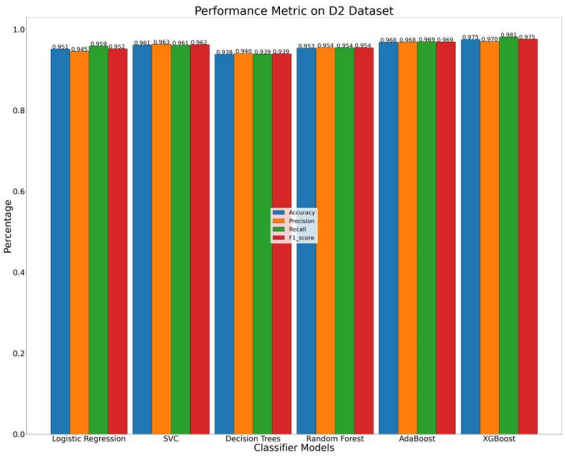

Fig 18: Performance Metrics of various models visualized graphically on the D2 dataset

While, for the D2 dataset, a more generalized dataset, contains the news from domains other than politics too, for which FIg 11 represents the accuracy, precision, recall, and F1 score. The highest accuracy, precision, recall, and F1 score achieved was 97.52%, 97.05%, 98.10%, and 97.57%, respectively, with the Ensemble Learner XGBoost algorithm. At the same time, the other ensemble learner AdaBoost algorithm was the second best performing model with an accuracy of 96.83%.

We can see from the above figures, i.e., Fig 10 and 11, that Ensemble classifier learners, i.e., XGBoost and AdaBoost, performed relatively better than individual weak learners. Also, on more generalized data, such as the D2 dataset, the Random Forest, with an accuracy of 95.36%, performed poorly compared to the D1 dataset. The overall difference between the Random Forest classifier and the XGBoost on the D2 dataset was 2.16%; this means XGBoost performed relatively better than Random Forest on a more generalized dataset.

XGBoost performs better with the D2 dataset because with every iteration in the XGBoost algorithm, it tries to minimize the error as it combines multiple weak learners, as defined in Section 4.3.6. It assigns more weightage to the data points which are wrongly classified.

Previous research papers only focussed on using either text or title + author from the news articles to build a classifier model. We used both features to build the model, which gave us higher accuracy than other papers [11][12]. As indicated in the Research Problem section, where prior studies exclusively focused on the political area, we employed a more generic dataset as D2 in this paper. One of the research showed that the model could be skewed if it is exclusively trained on a specific domain, in this example, political news [32], where the model predicted newer datasets pertaining to politics rather well but was virtually always erroneous if the new data was from a different subject. As a result, we attempted to be as fair as possible by including a more generic D2 dataset alongside the D1 dataset.

6. CONCLUSION

In this paper, we attempted to implement multiple Machine Learning algorithms, including single learning classifiers and ensemble learners, as well as Natural Language Processing, on available datasets pertaining to news articles to identify the best-performing algorithms capable of classifying the news articles as fake or authentic with the highest accuracy.

The model might have been applied to other datasets, making it even more versatile than we created, but this research was constrained to only two datasets. Furthermore, this report did not address the dissemination of fake news via other channels such as YouTube videos, Facebook, or Whatsapp content.

CITATIONS

[1] Boulianne, S., Shehata, A., 2021. Age Differences in Online News Consumption and Online Political Expression in the United States, United Kingdom, and France. The International Journal of Press/Politics 27, 763-783.

[2] MITCHELL, A., 2022. Younger adults more likely than their elders to prefer reading news [WWW Document]. Pew Research Center. URL https://www.pewresearch.org/fact-tank/2016/10/06/younger-adults-more-likely-than-their-elders-to-prefer-reading-news/.

[3] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake News Detection on Social Media. ACM SIGKDD Explorations Newsletter 19, 22-36.

[4] Allcott, H., Gentzkow, M., 2017. Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives 31, 211-236.

[5] Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., Rand, D., 2021. Shifting attention to accuracy can reduce misinformation online. Nature 592, 590-595.

[6] Wardle, C., Derakhshan, H., 2022. Information disorder: Toward an interdisciplinary framework for research and policy making [WWW Document].Council of Europe Publishing. URL https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html.

[7] Molina, M., Sundar, S., Le, T., Lee, D., 2019. “Fake News” Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. American Behavioral Scientist 65, 180-212.

[8] Molina, M., Sundar, S., Le, T., Lee, D., 2019. “Fake News” Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content. American Behavioral Scientist 65, 180-212.

[9] Taboada, M., Nielsen, D., 2022. GitHub - sfu-discourse-lab/MisInfoText: Datasets for fake news and misinformation detection [WWW Document]. GitHub. URL https://github.com/sfu-discourse-lab/MisInfoText.

[10] B. Narwal, "Fake News in Digital Media," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 977-981, doi: 10.1109/ICACCCN.2018.8748586.

[11] Ahmad, I., Yousaf, M., Yousaf, S., Ahmad, M., 2020. Fake News Detection Using Machine Learning Ensemble Methods. Complexity 2020, 1-11.

[12] Probierz, B., Stefański, P., Kozak, J., 2021. Rapid detection of fake news based on machine learning methods. Procedia Computer Science 192, 2893-2902.

[13] Amaris, M., de Camargo, R., Dyab, M., Goldman, A., Trystram, D., 2016. A comparison of GPU execution time prediction using machine learning and analytical modeling. 2016 IEEE 15th International Symposium on Network Computing and Applications (NCA).

[14] Jain, A., Shakya, A., Khatter, H., Gupta, A., 2019. A Smart System for Fake News Detection Using Machine Learning. International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT).

[15] Khanam, Z., Alwasel, B., Sirafi, H., Rashid, M., 2021. Fake News Detection Using Machine Learning Approaches. IOP Conference Series: Materials Science and Engineering 1099, 012040.

[16] Kirner-Ludwig, M., 2020. Creation, dissemination and uptake of fake-quotes in lay political discourse on Facebook and Twitter. Journal of Pragmatics 157, 101-118.

[17] Cardoso Durier da Silva, F., Vieira, R., Garcia, A., 2022. Can Machines Learn to Detect Fake News? A Survey Focused on Social Media [WWW Document]. Hdl.handle.net. URL http://hdl.handle.net/10125/59713.

[18] Wang, W., 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

[19] Kesarwani, A., Chauhan, S., Nair, A., 2020. Fake News Detection on Social Media using K-Nearest Neighbor Classifier. 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE).

[20] Abdulrahman, A., Baykara, M., 2020. Fake News Detection Using Machine Learning and Deep Learning Algorithms. 2020 International Conference on Advanced Science and Engineering (ICOASE).

[21] Lai, C., Chen, M., Kristiani, E., Verma, V., Yang, C., 2022. Fake News Classification Based on Content Level Features. Applied Sciences 12, 1116.

[22] Albahr, A., Albahar, M., 2020. An Empirical Comparison of Fake News Detection using different Machine Learning Algorithms. International Journal of Advanced Computer Science and Applications 11.

[23] Lakshmi, V., Kumari, C., 2022. Detection of Fake News using Machine Learning Models. International Journal of Computer Applications 183, 22-27.

[24] Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

[25] Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).

[26] Fake News | Kaggle [WWW Document], 2022. [WWW Document]. Kaggle.com. URL https://www.kaggle.com/competitions/fake-news/data.

[27] Evgeniou, T., Pontil, M., 2001. Support Vector Machines: Theory and Applications. Machine Learning and Its Applications 249-257.

[28] Freund, Y., Schapire, R., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 55, 119-139.

[29] Chen, T., Guestrin, C., 2016. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[30] Gao, C., Sang, N., Tang, Q., 2010. On selection and combination of weak learners in AdaBoost. Pattern Recognition Letters 31, 991-1001.

[31] Guo, A., Yang, T., 2016. Research and improvement of feature words weight based on TFIDF algorithm. 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference.

[32] Agarwal, V., Sultana, H., Malhotra, S., Sarkar, A., 2019. Analysis of Classifiers for Fake News Detection. Procedia Computer Science 165, 377-383.