**Fake News Analysis Using Machine Learning**

Richa Singh [a], Sobhik Kumar Sawdagar [b] and Govind Kumar [c]

[a][b][c] MCA Scholar,

Guide Name: Brijendra Singh

Department of Information Technology,

School of Information Technology and Engineering,

Vellore Institute of Technology, Vellore – 632014, Tamil Nadu, India.

E-mail:

[a][richa.singh2022@vitstudent.ac.in](mailto:richa.singh2022@vitstudent.ac.in)

[b][sobhik.kumar2022@vitstudent.ac.in](mailto:sobhik.kumar2022@vitstudent.ac.in)

[c] [govind.kumar2022@vitstudent.ac.in](mailto:govind.kumar2022@vitstudent.ac.in)

**ABSTRACT**

In our day-to-day life, current news helps make us aware of what is happening worldwide. So, most people prefer watching the information on television. Some prefer reading the newspaper early in the morning with a cup of tea. In this digital age, where the vast majority have access to the internet, they get or read news from various online platforms. If the information is fake that will mislead people, sometimes fake news is used to spread rumours about things or it will affect some political leadership positions just because of fake news. With the recent growth of social media, the incidence of fake news has skyrocketed, and this false information is slowly making its way into the mainstream media. Many spammers use alluring news headlines to generate revenue using advertisements via click baits. In this paper, we aim to perform the detection of fake news spreading online using python and various machine-learning methods to provide users the ability to differentiate between fake and real news.

**KEYWORDS**

fake news, social media, python, machine-learning, text-classification.

**1. INTRODUCTION**

**1.1 TOPIC**

Social media and the internet today have a signiﬁcant inﬂuence on society.

Information is now accessible in a convenient manner. Therefore, some people might exploit this fact to spread false information among others who access information online. To make their website the one that gets all the attention, social media platforms nowadays have a tendency to decorate the original news report before releasing it. Fake news has become increasingly prevalent and harmful as a result of such instances.

To identify intentional fake news online, there must be technology. However, determining the validity of the material found online is more diﬃcult. People use different social media platforms, such as Facebook, Twitter, and others, to discuss events that occur anywhere in the globe online. They look up news and talk about it. People are exposed to a dosage of fake news, rumours, misleading news, and bogus conspiracy theories thanks to these unreliable sources of information. Unpredictable natural disasters like earthquakes and tsunamis are an example of fake news. Due to the nature of these unexpected events, fake news is also aired as they occur, which furthers uncertainty.

What is fake news exactly?

[[1]](https://en.wikipedia.org/wiki/Fake_news) False or misleading information presented as news is known as fake news. Most of the time, the goal of fake news is to hurt someone or something's reputation or make money from advertising. However, the term has been used to refer to any kind of false information, and there is no established definition. By competing with real news, fake news can lessen its impact. According to research, false political information tends to spread "three times" faster than other false news [2]. There are three types of fake news detection:

a) Manipulation: Fabricated news is intentionally withheld information that typically comes only from one source. It's likely that the source is aware of the story's errors. Fake news stories cannot succeed without clickbait. b) Hoax: In order to deceive the public, this type of reporting employs more sophisticated deception strategies. Fake news is disseminated by numerous outlets. There is a possibility that some people believe the story to be true. This kind of news can be found on a lot of different websites, like the false information about Donald Trump that was shared on blogs, Twitter, Facebook, and other social media during the election. This makes it more likely that the general public will believe it.

c) Humour: fake news that the source presents as humorous. if you want to share satire with people who don't know where the material came from. It could be erroneously assumed to be true by some people.

[[3]](https://thelogicalindian.com/trending/india-saw-214-rise-in-fake-news-rumour-cases-in-2020-report-30723) The most recent data from the National Crime Records Bureau (NCRB) indicate that misinformation and rumours-related cases in India increased by a staggering 214 percent in 2020, which is a threefold increase from 2019. When the coronavirus pandemic struck the nation, 1,527 instances of fake news were reported, compared to 486 cases in 2019. With 280 cases in 2018, the number was significantly lower. Telangana had the most misinformation incidents out of all the states, with 273 of them, followed by Tamil Nadu with 188 and Uttar Pradesh with 166.

[[4]](https://www.bbc.com/news/world-asia-india-53165436) Fake news dominated from January to early March (before the coronavirus outbreak took hold) thanks to the Citizenship Amendment Act, a new law that gives citizens of three neighbouring nations citizenship. Protests were held all over the country by those who claimed that the law would marginalize Muslims. This included fake messages, messages with fake attributions, doctored videos, fake images, and the use of previously recorded videos and images in a new setting.

For feature extraction, we used Bag of words (CountVectorizer), Text Vectorization-Inverse Document Frequency (TF-IDF), and Logistic Regression Classifier, Naive Bayes Classifier, XGBoost Classifier, Decision Tree Classifier, Random Forest Classifier, and PassiveAggressive Classifier in this paper. For text classification, we used XGBoost Classifier, Decision Tree Classifier, Random Forest Classifier, and PassiveAggressive Classifier.

**1.2 RELATED WORK**

1. Atik Mahabub (2020), In this paper, Ensemble Voting Classiﬁer based, an intelligent detection system is proposed to deal with news classiﬁcation both real and fake tasks. The experimental outcomes conﬁrm that the proposed framework can accomplish about 94.5% of outcomes as far as accuracy. As it is actualized in a streaming way, this proposed structure can also be implemented in other classiﬁcation techniques to detect fake proﬁles, fake messages, etc
2. Nihel Fatima Baarir and Abdelhamid Djeffal (2021), Here this work, used term frequency inverse document frequency (TF-IDF) of bag of words and n-grams as feature extraction technique, and Support Vector Machine as a classifier. The work done is completed and continued in different aspects. It has to be relevant to extend this study with a larger dataset
3. Kartik Rajesh, Aditya Kumar, Dr. Rajesh Kadu (2019), This proposed paper used a bag of words technique to extract the features and passed those features into classifiers.  But It can be noted that working with its large feature sets made it difficult to understand the importance of each of the features, thereby hindering the analysis. and limiting the prediction accuracies.
4. Vian Sabeeh Mohammed Zohdy Rasha Al Bashaireh (2019) ,This paper proposed a CNIRI-FS (Contextual Negation Handling and Inherent Relation Identiﬁcation for Enhanced Feature Selection) model to detect fake information; A Genetic Algorithm (GA) was used to ﬁlter out unreliable features to understand the exact content of the users' text and to improve the classiﬁcation accuracy, the proposed model performs polarity identiﬁcation through negation handling. As a result, this model showed higher precision and 76% in terms of accuracy.
5. Anjali Jain 1, Avinash Shakya2, Harsh Khatter 3, Amit Kumar Gupta 4(2019)

With the help of Machine learning and natural language processing, the author tried to aggregate the news and later determine whether the news is real or fake using a Support Vector Machine. The results of the proposed model are compared with existing models. In the future, the ensuing algorithm may provide better results with hybrid approaches. It can be concluded that SVM can deal with high dimensional spaces and work well, deﬁning the correctness of results as 93.6% of accuracy.

1. Gowri, S., Jenila, J., Reddy, B. S., & Sheela, M. A. (2021), The proposed research work has been more focused on scrutinizing the fake news that compares different classiﬁcation methods for predicting accuracy among the proposed classiﬁers, based on the content provided by the dataset. We are initializing the data, then by using TF-IDF Vectorizer we split the training and testing dat TF-IDF vectorizer on the proposed dataset uses the classiﬁcation algorithm like SGD (Stochastic Gradient Descent) Classiﬁer. Finally, the project obtained a higher accuracy of about 93.29% by using the dataset for samples from the dataﬂair website.
2. Marco L. Della Vedova et al, in this proposed paper machine learning Fake News Detection method that comprises news content and social media features performs existing methods to enhance the accuracy up to 4.8%. They implement this method within Facebook messenger, Chabot and validate it in the real-time application which obtains the highest accuracy in fake news detection of 81.7%.
3. M. Granik and V. Mesyura (2017), This paper proposed a simple approach for the detection of fake news using the Naïve Bayes classifier

This model was tested against the dataset of Facebook new posts.

The accuracy observed is 74% on the test set.

1. Jasmine Shaikh, Rupali Patil (2020), in this paper work the problem for fake news detection using different algorithm like Support Vector Machine(SVM), Naïve Bayes, Passive Aggressive Classifier.The dataset here took is a news dataset and the result shows the Support Vector Machine (SVM) as classifier, has accuracy of 95.05%.
2. Rahul R Mandical, N Mamatha and N Shiva Kumar (2020) this paper attempts to detect the fake news using Naive Bayes, Passive Aggressive Classifier and Deep Neural Networks have been used in eight different dataset acquired from different sources -Kaggle, LIAR, superset, politifact, jru etc. Different algorithms tested on these sources shows various accuracies. future work obtained is that Pre-trained word embeddings such as word2Vec and GloVe could be used.
3. Vaishali Vaibhav Hirelkar, Arun Kumar (2020) this paper discuss the general approach for fake news detection using SVM, Linear Logistic Regression, Decision Tree, Random Forest,XG-Boost Gradient Boosting .the dataset is collected from Kaggle .In this survey paper has been carried out from 2017 onwards. It has been observed that various standard machine learning algorithms perform well and give better results compared to other techniques. This study reviewed, compared, contrasted, and evaluated existing research on counterfeit news, including quantitative and qualitative analysis of counterfeit news.
4. Okuhle Ngada, Bertram Haskins (2020) this paper explores the fake news prediction by selecting 6 different algorithm KNN, Random Forest -RF, Support Vector Machine - SVM XGBoost as XGB applied on the datasets collected from the Kaggle site. As the result Support Vector Machine shows the highest accuracy of 99.4%. Limitation here observed is that Data normalization dimensionality reduction techniques were not considered. Future work related to the study include the use of deep learning approaches for fake news detection.
5. Anjali Jain, Avinash Shakya, Harsh Khatter, Amit Kumar Gupta (2019) This paper demonstrates a model and the methodology for fake news detection. With the help of Machine learning and natural language processing, the author tried to aggregate the news and later determine whether the news is real or fake using Naïve Bayes classifier, Support Vector Machines, NLP. The proposed model is working well and defining the correctness of results 93.6% of accuracy. Limitations observed was use of the hybrid approaches, enhancing the user interface of the proposed model and combining attribution feature extraction.
6. Ranojoy Barua, Rajdeep Maity, Dipankar Minj, Tarang Barua, Ashish Kumar Lavek (2019) The work proposed here is to tackle this issue and it aims to identify news articles whether it is true or false. This is achieved using ensemble techniques - LSTM (Long short-term memory) and GRU (Gated recurrent units). The source of data is Google News API (real), Web scraping (fake). The highest accuracy of this result is 80.29%. Some limitations shown by this paper is that it expects decent size articles, the article is not large enough to form 10 lines of summary, can’t function to classify tweets posted on microblogging sites like twitter, and similar grammatical constructs as real news fail to classify.
7. Karishnu Poddar, Geraldine Bessie Amali D, K S Umadevi (2019) The authors here used different algorithms namely are Tech-Decision Trees, Naive Bayes, Logistic Regression, Support Vector Machines and Neural Networks. the dataset in this paper is picked from Kaggle site and the whole paper shows the highest accuracy of 91%. Limitations to be keep in mind in future work is the use of multiple attributes and the drawback is that logistic regression and SVM model gives the same accuracy.
8. Hardik Sodhani and Sanket Jayesh Muchhala (2021), this paper discusses the use of natural language techniques to identify false news. using data obtained from signal media and a list of sources from OpenSources.co, the database was tested on multi-phase algorithms Vector Support Machines, Stochastic Gradient Decreases, Gradient Enlargement, Determined Decision Trees, and Random Forests. The paper concluded that the Stochastic Gradient Descent model identifies unreliable sources with 77.2% accuracy.
9. Uma Sharma, Sidarth Saran, Shankar M. Patil (2021), this paper aims to performs binary classification for prediction of machine learning. the data used for this project is LIAR and REAL\_OR\_FAKE.CSV which was tested using using four different algorithms -Naive Bayes Classifier, Random Forest, Logistic Regression and Passive Aggressive algorithms. The best model, with highest accuracy of 75% is shown by Logistic regression.
10. 1.Nerissa Pereira, 2. Simran Dabreo, 3. Linnet Rodrigues, 4.Prof. Merly Thomas (2020) in this paper authors have come up with the applications of NLP and machine learning algorithm for detecting fake news. the dataset used to test the efficiency is produced by kaggle which containing 25000 news articles, was noisy and required cleaning which include -stop words removal, stemming, tokenization and POS tagging. Various classifiers used here to train the model are - Naive Bayes, Support vector machine,LSTM . On calculating the accuracies LSTM shows the high result as it uses the concept of memory. Future scope is to increase the size of the sample at least 1000 tweets. Also we need to verify the images and audio included in the content.
11. Afrin Jaman Bonny, Puja Bhowmik, Md Shihab Mahmud, and Abdus Sattar (2022) used ML algorithms Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (GB), Gradient Boosting (GB), Multinomial Naive Bayes (MNB), and K Nearest Neighbours (KNN) to identify fake news in this study. In this case, we trained a machine learning model using a total of 44898 distinct news articles from a dataset of real and fake news. The model with the highest accuracy, LR, performed the best.
12. Waqas Haider Bangyal (2021), In this paper the author applied eight machine-learning algorithms such as Naive Bayesian, Adaboost, -nearest neighbors, random forest, logistic regression, decision tree, neural networks, and support vector machine. The fake news datasets contain fake news on covid -19. highest accuracy is shown by CNN and BiLSTM. Future aim is to large and complex dataset, with increasing the number of labels too. Also use of twitter streaming API and also include other languages and use of special character and numeric values as well.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ref. No. & Year | Technology Used | Dataset Sources | Limitations of Work | Accuracy |
| [[1]](https://sci-hub.hkvisa.net/10.1007/s42452-020-2326-y)  2020 | Ensemble Voting Classifier | Buzz Feed News | - stretched out the proposed structure  - Use Deep learning techniques | 94.50% |
| [[2]](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9378748&isnumber=9378701)  2021 | Ensemble Voting Classifier | Kaggle Site |  | 89.20% |
| [[3]](https://ieeexplore.ieee.org/document/8978436)  2019 | Logistics regression | Kaggle | * Large datasets * Hindering analysis * Better classification * Border generalizability | 72% |
| [[4]](https://ieeexplore.ieee.org/document/9071056)  2019 | CNIRI-FS (Contextual Negation Handling and Inherent Relation Identification for Enhanced Feature Selection) | FEVER 1.0 [25] dataset | * Exploits the genetic algorithm | 76% |
| [[5]](https://ieeexplore.ieee.org/abstract/document/8977659)  2019 | Natural Language Processing. (NLP), Support Vector Machine (SVM) | Google news, feedly, news360 | use hybrid approaches | 93.60% |
| [[6]](https://ieeexplore.ieee.org/document/9395766)  2021 | SGD (Stochastic Gradient Descent) Classifier | Dataflair | Decision tree classifier gives less accuracy | 93.29% |
| [[7]](https://ieeexplore.ieee.org/document/8100379)  2017 | Naïve Bayes Classifier | BuzzFeed News | Use larger data sets  Use stemming  Remove stop words  Treat rare words separately.  Use group of words | 74% |
| [[8]](https://ieeexplore.ieee.org/document/9358890)  2020 | Support Vector Machine(SVM), Naïve Bayes, Passive Aggressive Classifier | News dataset |  | 95.05% |
| [[9]](https://ieeexplore.ieee.org/document/9198610)  2020 | Naive Bayes, Passive Aggressive Classifier and Deep Neural Networks | FND -JRU  POLITIFACT  PONTES ROUTE 1  PONTES ROUTE 2  CLAIMS KG  KAGGLE COMPETITION  LIAR  NEWS FILES  SUPERSET | Pre-trained word embeddings such as word2Vec and GloVe could be used |  |
| [[10]](https://ieeexplore.ieee.org/document/9137915)  2020 | SVM  Linear  Logistic Regression Decision Tree  Random Forest  XG-Boost Gradient Boosting Neural Network | Kaggle |  | 94% |
| [[11]](https://ieeexplore.ieee.org/document/9411638)  2020 | KNN  Random Forest -RF  Support Vector Machine - SVM XGBoost as XGB | Kaggle | -Data normalisation  dimensionality reduction techniques were not considered.  -use of deep learning approaches for fake news detection. | 99.4% |
| [**[12]**](https://ieeexplore.ieee.org/document/8977659)  2019 | Naïve Bayes classifier,  Support Vector Machines  NLP |  | - hybrid approaches  - enhance the user interface  -combing attribution feature extraction | 93.50% |
| [[13]](https://ieeexplore.ieee.org/document/8973059)  2019 | LSTM and GRU | Google News API (real),  Web scraping (fake) | - expects decent size article  - article is not large to form 10 lines of summary  - cant function to classify tweets  - similar  grammatical constructs as real news is failed to classify | 80.2% |
| [[14]](https://ieeexplore.ieee.org/document/8960044) | Tech-Decision Trees, Naive Bayes, Logistic Regression, Support Vector Machines and Neural Networks | Kaggle | -use of multiple attributes  - logistic regression and SVM model gives the same accuracy | 91% |

**Table 1**

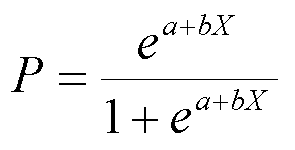
**1.3 MODEL DESCRIPTION**

**1.3.1 Logistic Regression**

Logistic regression is a basic and widely used machine learning algorithm with better results compared to other complicated ML algorithms.

It comes under supervised learning, and is used to solve classification problems and to predict the discrete values.  This model delivers a binary outcome like yes/no,0/1 or true or false.

Some of the use cases include, Fraud detection in online transactions, disease prediction, churn prediction, Rain forecasting etc.



Where

     x=input value

     y=predicted output

     a=bias or intercept term

     b=coefficient for input i.e. x

**1.3.2 Decision Tree Classifier**

A very effective classifier and supervised learning algorithm, the decision tree classifier. It can perform both classification and regression problems. All the possible

Solutions to a decision are graphically represented.

The logic behind the use of decision trees is as it shows tree-like structure. The data is broken into smaller parts and the decision tree is built. A decision tree can contain categorical data i.e. yes/no and numeric data as well.

Some of the practical use of this algorithm is seen in fault diagnosis, sentiment analysis etc.

**1.3.3 Random Forest Classifier**

Random forest is a classifier belonging to supervised learning, which is used for classification and regression problems in ML.

This algorithm is based on the concept of ensemble learning in which multiple classifiers combine to solve complex problems and finally improves the accuracy of the model.

It Contains several decision trees on various subsets of the randomly selected training data set. Here the outcome is predicted by taking the mean of output from different trees and more the number of trees more is the precision of the outcomes.

This algorithm is used in a lot of different fields: finance,medicines,banking etc.

**1.3.4 Naive Bayes Classifier**

Based on the Bayes theorem, the Naive Bayes algorithm is a supervised learning method for classification problems. Most of the time, it is used to classify text using a large training dataset. It's one of the easiest and most effective categorization algorithms, assisting in the creation of quick machine learning models that can make quick predictions. As a probabilistic classifier, it predicts the likelihood of an event occurring. The Naive Bayes algorithm can be used for a variety of tasks, including spam filtration, sentiment analysis, and article classification.

**1.3.5 XGBoost Classifier**

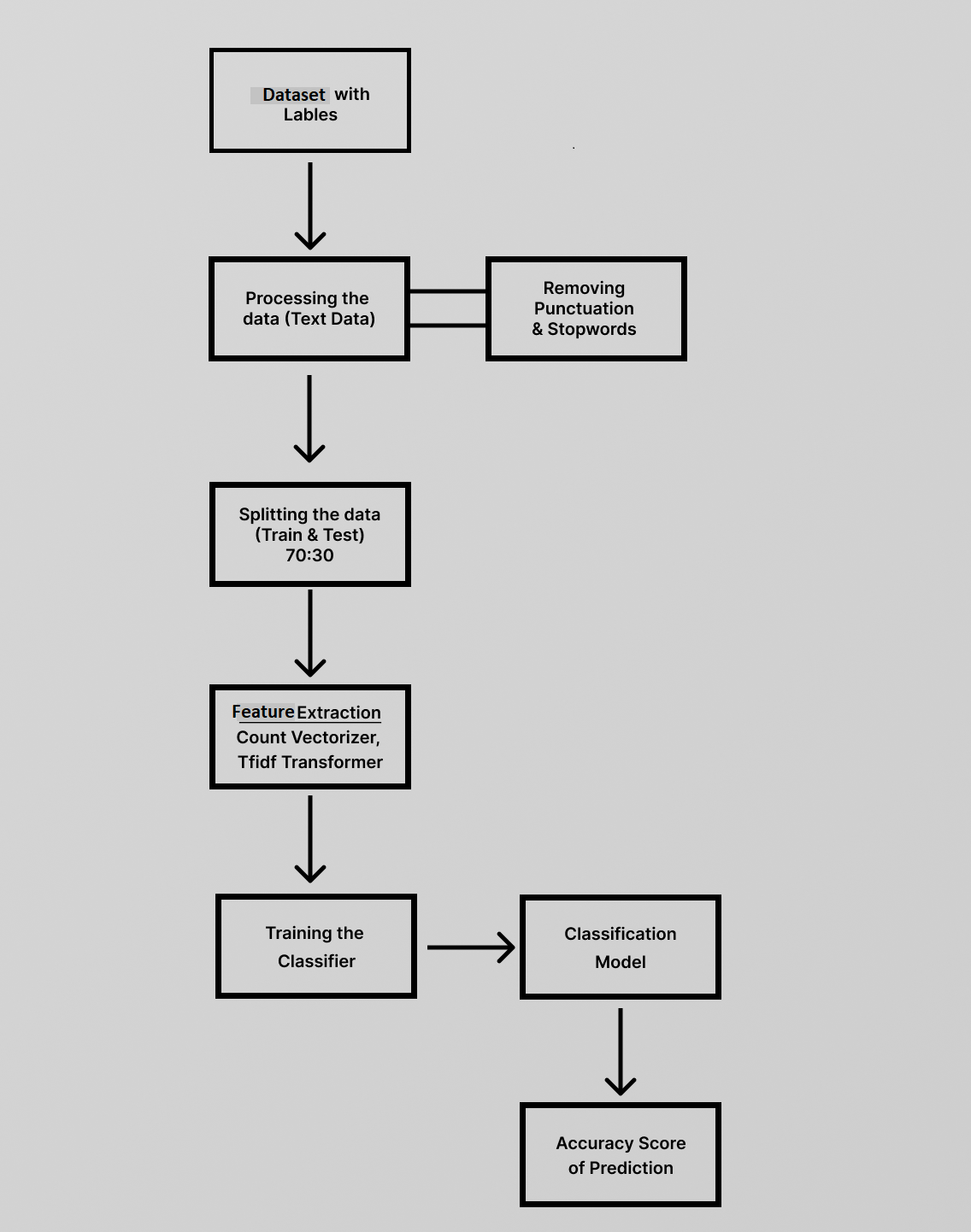
An open-source gradient boosting framework for C++, Java, Python, R, and Julia is provided by the XGBoost software library. It is utilized for classification and regression supervised learning issues and stands for "eXtreme Gradient Boosting." The gradient boosting algorithm, a machine learning method for regression and classification issues, is implemented in XGBoost. It creates a model out of a collection of weak models, usually decision trees. Gradient boosting is used to train a model that can make predictions that are better than those of any of the weak models individually. The accuracy, efficiency, and performance of XGBoost are well-known. It has been widely used in the industry and used to win numerous data science competitions.

**1.3.6 PassiveAggresive Classifier**

The passive aggressive classifier algorithm is a type of online learning algorithm that can handle large datasets and adapts its model to every new case it encounters. Because it is an online learning algorithm, the passive aggressive algorithm may alter its weights when new information is received. At each iteration, the passive aggressive classifier examines a new instance to see if it was correctly classified and adjusts its weights accordingly. If the case is properly classified, there is no change in its weight. The passive aggressive algorithm, on the other hand, adjusts its weights to better classify subsequent occurrences if an instance is incorrectly classified.

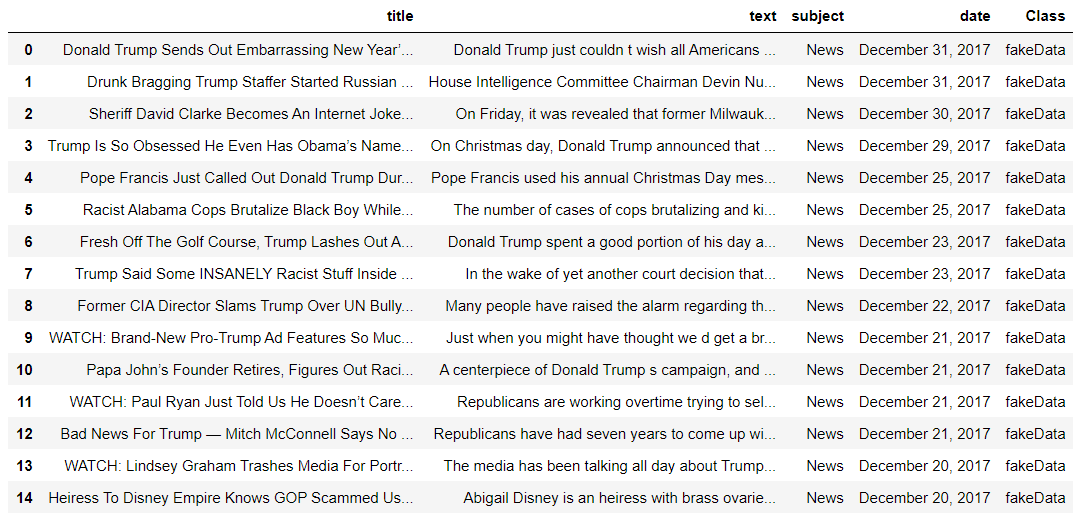
**2. MATERIALS AND METHODS**

**2.1 METHODOLOGY AND WORKFLOW**

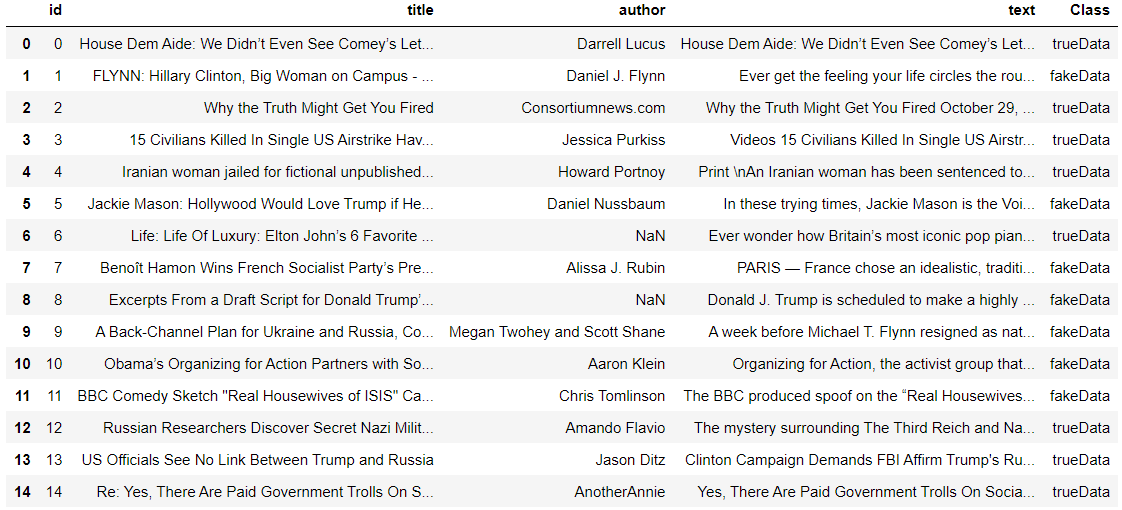
****

**Fig. 1. Workflow Diagram**

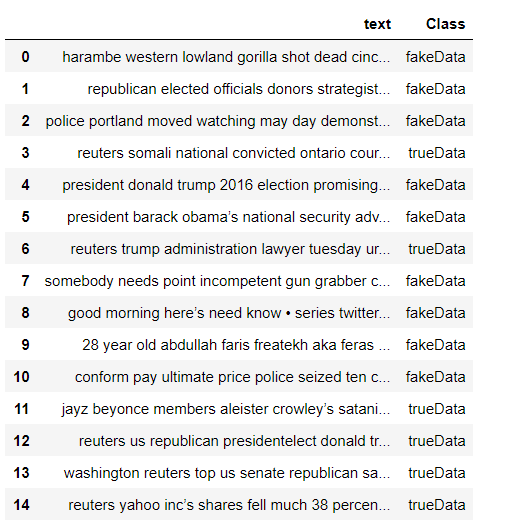
**2.2 DATASET**

****

**Fig.2** - **Dataset 1**

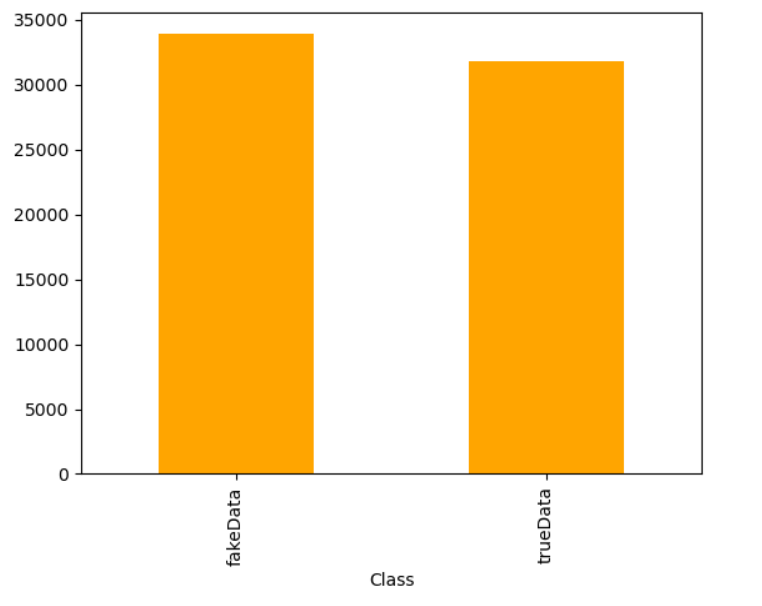
****

**Fig.3 - Dataset 2**

****

**Fig.4** - **Final Dataset (Dataset1 + Dataset2)**

**Classification of Final dataset into True and Fake Data**

****

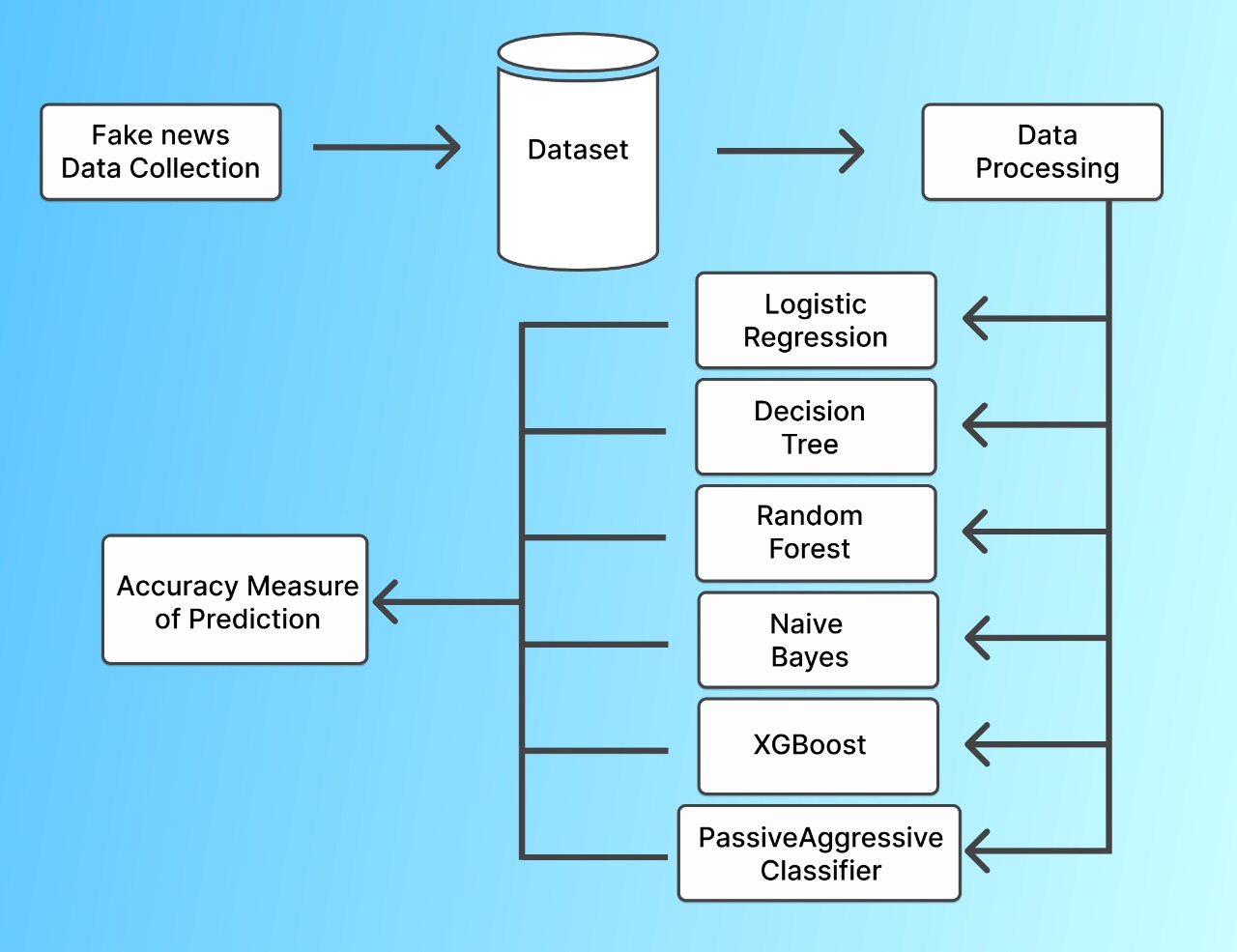
**Fig.5. Classification of final dataset Diagram**

**3. EXPERIMENTAL PROCEDURE**

**3.1 PROPOSED WORK**

Fake news detection is a subtask of text classification and is often defined as the task of classifying news as real or fake. There is a need for a way to detect these fake news stories. We have used machine learning classifiers, which are used for detecting fake news. The classifiers are first trained with a data set called the training data set. After that, these classifiers can automatically detect fake news.

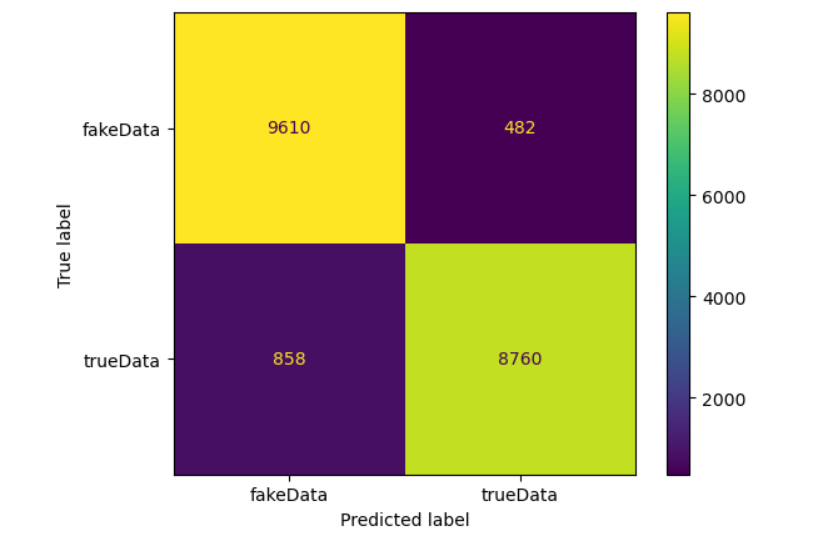
**3.2 ARCHITECTURE DESIGN**

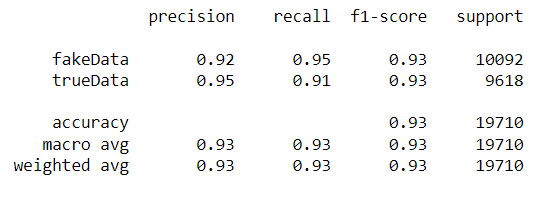
** Fig.6 Architecture Design Diagram**

**3.3 PERFORMANCE**

**3.3.1 Confusion Matrix and Classification Report**

A. Logistic Regression

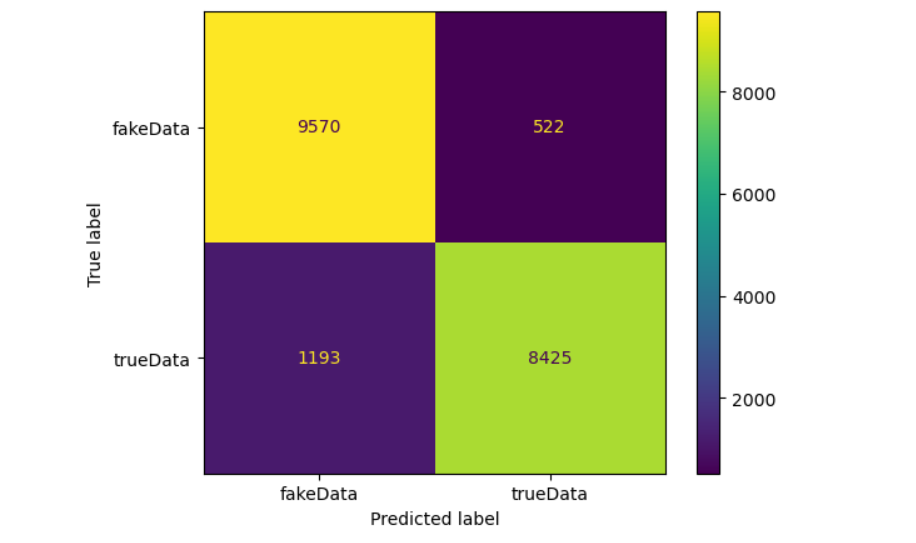


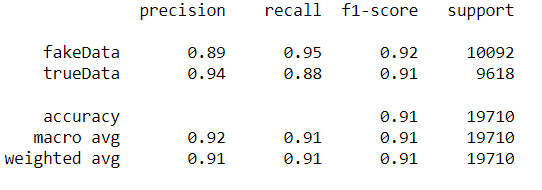


**Fig.7 and Fig.8.**

**Confusion Matrix and Classification report of Logistic Regression Model**

B. Decision Tree

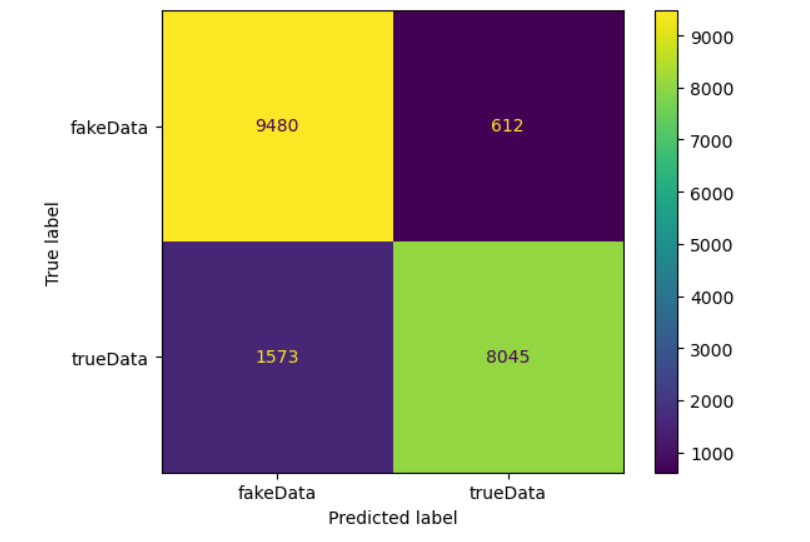


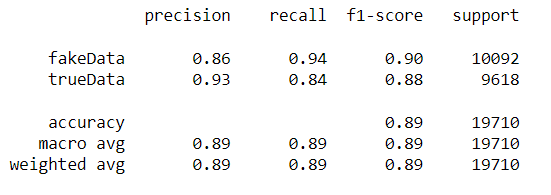


**Fig.9 and Fig.10**

**Confusion Matrix and Classification report of Decision Tree Model**

C. Random Forest

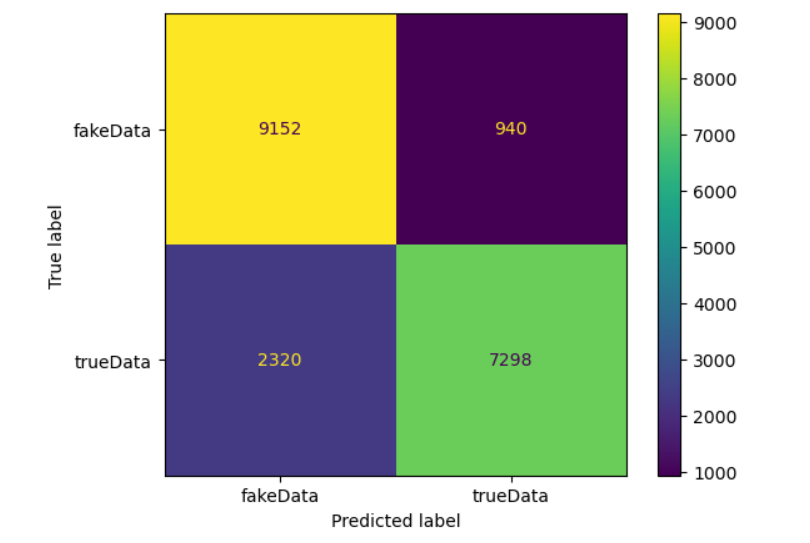


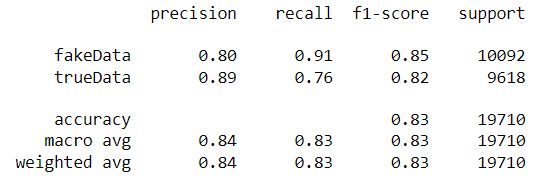


**Fig.11 and Fig.12**

**Confusion Matrix and Classification report of Random Forest Classifier**

D. Naive Bayes

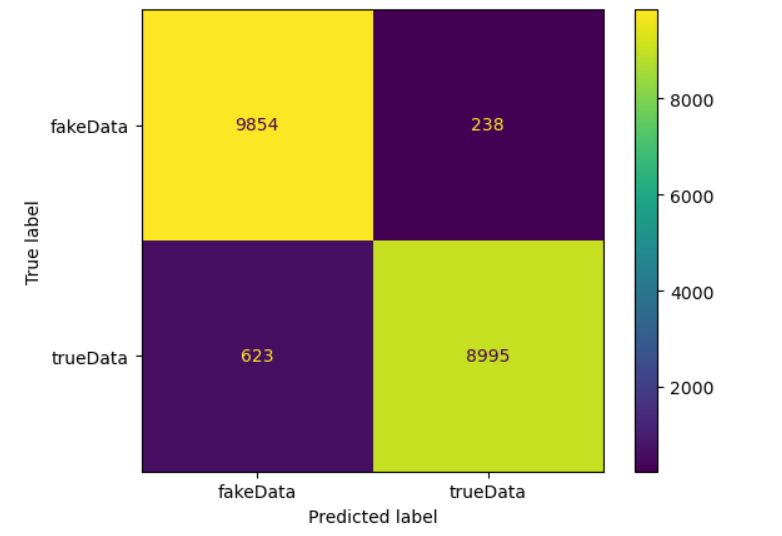


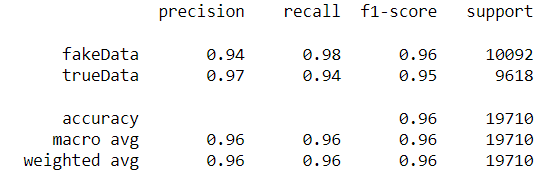


**Fig.13 and Fig.14**

**Confusion Matrix and Classification report of Naive Bayes Classifier**

E. XGBoost

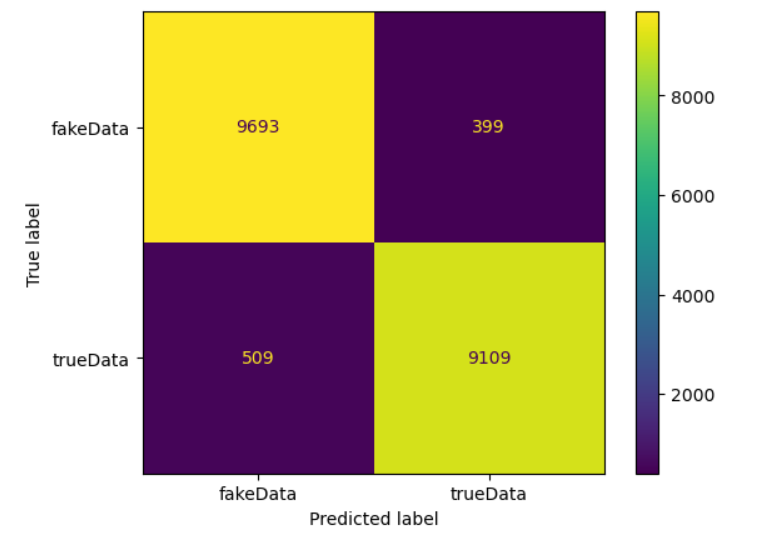


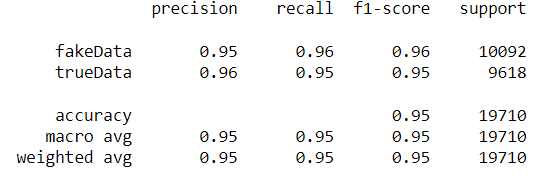


**Fig.15 and Fig.16**

**Confusion Matrix and Classification report of XGBoost Classifier**

F. PassiveAggressive Classifier





**Fig.17 and Fig.18**

**Confusion Matrix and Classification report of PassiveAggressive Classifier**

**4. RESULT AND DISCUSSIONS**

**4.1 COMPARATIVE ANALYSIS**

|  |  |  |  |
| --- | --- | --- | --- |
| **ML Models** | **Dataset 1** | **Dataset 2** | **Final Dataset**  **(Dataset 1 + Dataset 2)** |
| **Logistic Regression** | 98.75 % | 98.75 % | 93.2 % |
| **Decision Tree** | 99.65 % | 99.65 % | 91.3 % |
| **Random Forest** | 99.18 % | 99.12 % | 88.91 % |
| **Naive Bayes** | 94.95 % | 94.95 % | 83.46 % |
| **XGBoost** | 99.73 % | 99.73 % | 95.63 % |
| **PassiveAggressive**  **Classifier** | 99.6 % | 99.61 % | 95.39 % |

**Table 2: - Accuracy Scores**

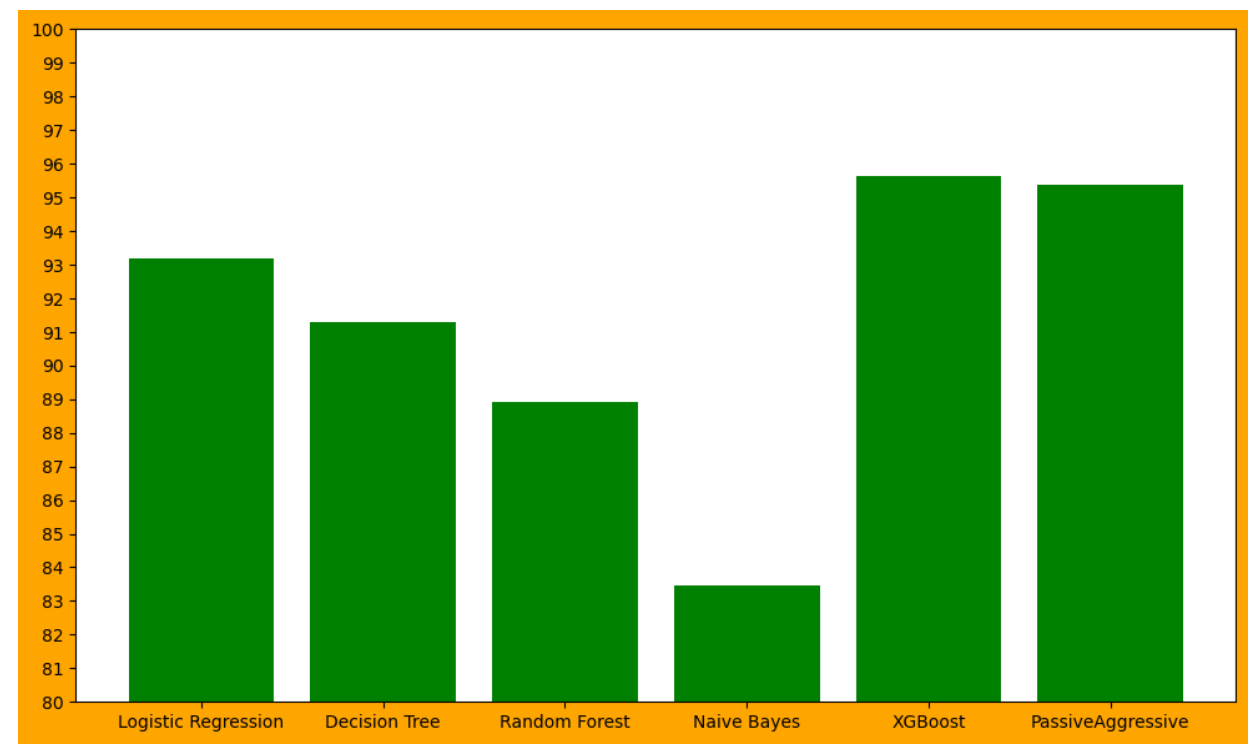
**Our Interpretation:**

The table above gives us the overall summary of the accuracy scores achieved by the 6 different Machine Learning Algorithms.

We can see that the performance of all the 6 ML Algorithms on Dataset 1 and Dataset 2 is somewhat very similar, but on the Final Dataset (which is basically made by merging Dataset 1 and Dataset 2), we can observe a drastic drop in accuracies of some models which were performing extremely good previously on the individual datasets.

Now, in our final observation, we found that “XGBoost Classifier” and “PassiveAggressive Classifier” were the only two models which performed extremely well on all the three datasets and were consistent throughout.

The difference between the two models i.e. “XGBoost Classifier” and “PassiveAggressive Classifier” was very minute, they both performed equally good, but the overall best performance with best accuracy throughout was achieved by “XGBoost Classifier” with accuracies 99.73%, 99.73% and 96.63% respectively.

****

**Fig.19 Comparison Bar graph of Final Dataset**

**5. CONCLUSION**

In this paper, we have discussed what is a fake news, how it spreads in this online era, how harmful it can be. In our project we tried to analyse how machine learning techniques can be used, how the ML models can be trained to identify the fake news spreading and floating on the web.

For our analysis we have collected the datasets from the Internet, one of the datasets contained news related to US politics and the other dataset contained General news related to any field. We have performed our analysis on these individual datasets also we have merged these two datasets together to form a bigger and a more diverse dataset to make it more challenging for the ML models to train on them and detect accurately and called it Final Dataset, of course the datasets were labelled.

After data collection, we pre-processed the data and did feature extraction and then trained the Machine Learning Models with parameter-tuning and performed our analysis, we found that the “XGBoost Classifier” performed the best overall comparatively.

**6**. **CONFLICTS OF INTEREST STATEMENT**

No conflicts of interest exist, according to the authors, with the publishing of this paper.

**7. DATA AVAILABILITY STATEMENT**

The fake news datasets are collected from [https://www.kaggle.com](http://www.kaggle.com/)

Dataset 1: -

https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

Dataset 2: -

<https://www.kaggle.com/competitions/fake-news/data>

**REFERENCES:**

[[1](https://link.springer.com/article/10.1007/s42452-020-2326-y#article-info)]  Mahabub, A. A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. *SN Appl. Sci.* **2**, 525 (2020). https://doi.org/10.1007/s42452-020-2326-y

[[2](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9378748&isnumber=9378701)] N. F. Baarir and A. Djeffal, "Fake News detection Using Machine Learning," 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), 2021, pp. 125-130, doi: 10.1109/IHSH51661.2021.9378748.

[[3](https://ieeexplore.ieee.org/document/8978436)] K. Rajesh, A. Kumar and R. Kadu, "Fraudulent News Detection using Machine Learning Approaches," 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-5, doi: 10.1109/GCAT47503.2019.8978436.

[[4](https://ieeexplore.ieee.org/document/9071056)] V. Sabeeh, M. Zohdy and R. A. Bashaireh, "Enhancing the Fake News Detection by Applying Effective Feature Selection Based on Semantic Sources," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019, pp. 1365-1370, doi: 10.1109/CSCI49370.2019.00255.

[[5](https://ieeexplore.ieee.org/abstract/document/8977659)]  A. Jain, A. Shakya, H. Khatter and A. K. Gupta, "A smart System for Fake News Detection Using Machine Learning," 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2019, pp. 1-4, doi: 10.1109/ICICT46931.2019.8977659.

[[6](https://ieeexplore.ieee.org/document/9395766)] S. Gowri, J. Jenila, B. S. Reddy and M. A. Sheela, "Scrutinizing of Fake News using Machine Learning Techniques," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 223-227, doi: 10.1109/ICAIS50930.2021.9395766.

[[7](https://ieeexplore.ieee.org/document/8468301)] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), 2018, pp. 272-279, doi: 10.23919/FRUCT.2018.8468301.

[[8](https://ieeexplore.ieee.org/document/8100379)] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[[9](https://ieeexplore.ieee.org/document/9358890)] J. Shaikh and R. Patil, "Fake News Detection using Machine Learning," 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), 2020, pp. 1-5, doi: 10.1109/iSSSC50941.2020.9358890.

[[10](https://ieeexplore.ieee.org/document/9198610)] R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica and A. N. Krishna, "Identification of Fake News Using Machine Learning," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198610.

[[11](https://ieeexplore.ieee.org/document/9137915)] V. V. Hirlekar and A. Kumar, "Natural Language Processing based Online Fake News Detection Challenges – A Detailed Review," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 748-754, doi: 10.1109/ICCES48766.2020.9137915.

[[12](https://ieeexplore.ieee.org/document/9411638)] O. Ngada and B. Haskins, "Fake News Detection Using Content-Based Features and Machine Learning," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411638.

[[13](https://ieeexplore.ieee.org/document/8977659)] A. Jain, A. Shakya, H. Khatter and A. K. Gupta, "A smart System for Fake News Detection Using Machine Learning," 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2019, pp. 1-4, doi: 10.1109/ICICT46931.2019.8977659.

[[14](https://ieeexplore.ieee.org/document/8973059)] R. Barua, R. Maity, D. Minj, T. Barua and A. K. Layek, "F-NAD: An Application for Fake News Article Detection using Machine Learning Techniques," 2019 IEEE Bombay Section Signature Conference (IBSSC), 2019, pp. 1-6, doi: 10.1109/IBSSC47189.2019.8973059.

[[15](https://ieeexplore.ieee.org/document/8960044)] K. Poddar, G. B. Amali D. and K. S. Umadevi, "Comparison of Various Machine Learning Models for Accurate Detection of Fake News," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), 2019, pp. 1-5, doi: 10.1109/i-PACT44901.2019.8960044.

[[16](https://www.researchgate.net/publication/356267717_Fake_News_Prediction_in_Machine_Learning)] Sodhani, Hardik & Muchhala, Sanket & geedh, shreeram. (2021). Fake News Prediction in Machine Learning. 10.1729/Journal.28521.

[[17](https://www.ijert.org/fake-news-detection-using-machine-learning-algorithms)] Uma Sharma, Sidarth Saran, Shankar M. Patil, 2021, Fake News Detection using Machine Learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NTASU – 2020 (Volume 09 – Issue 03),

[[18](https://www.jetir.org/view?paper=JETIR2004387)] "Comparative Analysis of Fake News Detection using Machine Learning and Deep Learning Techniques.", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.7, Issue 4, page no. pp1379-1385, April-2020, Available at : <http://www.jetir.org/papers/JETIR2004387.pdf>

[[19](https://www.researchgate.net/publication/366612371_Detecting_Fake_News_in_Benchmark_English_News_Dataset_Using_Machine_Learning_Classifiers)] Bonny, Afrin Jaman & Bhowmik, Puja & Mahmud, Md. Shihab & Sattar, Abdus. (2022). Detecting Fake News in Benchmark English News Dataset Using Machine Learning Classifiers. 1-8. 10.1109/ICCCNT54827.2022.9984461.

[20] Kareem, Irfan and Shahid Mahmood Awan. “Pakistani Media Fake News Classification using Machine Learning Classifiers.” 2019 International Conference on Innovative Computing (ICIC) (2019): 1-6.

**LEGENDS**

**Tables**

Table 1: Related Works

Table 2: Comparative Analysis (Accuracy Scores)

**Figures**

Figure 1: Workflow Diagram

Figure 2: Dataset 1

Figure 3: Dataset 2

Figure 4: Final Dataset (Dataset 1 + Dataset 2)

Figure 5: Classification of Final Dataset into True and Fake data

Figure 6: Architecture Design

Figure 7 and 8: Confusion Matrix and Classification report of Logistic Regression Model

Figure 9 and 10: Confusion Matrix and Classification report of Decision Tree Model

Figure 11 and 12: Confusion Matrix and Classification report of Random Forest Model

Figure 13 and 14: Confusion Matrix and Classification report of Naive Bayes Classifier

Figure 15 and 16: Confusion Matrix and Classification report of XGBoost Classifier

Figure 17 and 18: Confusion Matrix and Classification report of PassiveAggressive Classifier

Figure 19: Comparison Bar graph of Final Dataset