

Uppsala University

Data Mining Project

NewsVerifier

Patel Hetvi Piyushbhai Vinay Mehra

Group 6

october 2024

**Abstract**

The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for in- formation dissemination that has never been witnessed in human history before. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality. Automated classification of a text article as misinformation or disinformation is a challenging task. Even an expert in a particular domain has to explore multiple aspects before giving a verdict on the truthfulness of an article. In this work, we propose to use a machine learning ensemble approach for the automated classification of news articles. Our study explores different textual properties that can be used to distinguish fake content from real. By using those properties, we train a combination of different machine learning algorithms using various ensemble methods and evaluate their performance on real-world datasets. The experimental evaluation confirms the superior performance of our proposed ensemble learner approach in comparison to individual learners.

**Contents**

1. [Introduction 4](#_TOC_250025)
   1. [What is Fake news 4](#_TOC_250024)
   2. [RANDOM FOREST CLASSIFIER 5](#_TOC_250023)
   3. [Prediction and Evaluation 5](#_TOC_250022)
   4. [Feature Extraction 5](#_TOC_250021)
2. [DATA SETS 6](#_TOC_250020)
   1. [Importing True.csv file 6](#_TOC_250019)
   2. [Importing Fake.csv file 7](#_TOC_250018)
3. [Data prepossessing 7](#_TOC_250017)
   1. [Adding labels and merging datasets 7](#_TOC_250016)
   2. [Defining the filtering function 8](#_TOC_250015)
   3. [VISUALIZATION 10](#_TOC_250014)
      1. [Distribution of Articles by Subject 10](#_TOC_250013)
      2. [Scatter Plot 10](#_TOC_250012)
      3. [Line Plot 11](#_TOC_250011)
   4. [Identification of Issues 12](#_TOC_250010)
4. [DATA PREPARATION 13](#_TOC_250009)
   1. [SPLITTING DATA 13](#_TOC_250008)
5. [Modeling 14](#_TOC_250007)
   1. [RandomForest 14](#_TOC_250006)
   2. [Decision trees 16](#_TOC_250005)
   3. [LogisticRegression 18](#_TOC_250004)
6. [VERIFICATION 18](#_TOC_250003)
   1. [Model testing with Manual Testing 19](#_TOC_250002)
7. [Result 19](#_TOC_250001)
8. [CONCLUSION 20](#_TOC_250000)

## Introduction

The Internet has become compulsory in our life. It is now very easy to access the internet than it was before. There is no doubt that many young people prefer the internet to get their news rather than the newspaper, radio, etc. The internet provides many opportunities for us, we can search for anything on the internet to clear our doubts and for research purposes also. Simply saying, we can’t even imagine our life without the internet. The overwhelming growth of social media and news websites has made the news much more accessible, interactive and vivid than before. Further, the social networking platforms (like Twitter, Instagram and Facebook) serve as popular micro blogs for people to share or even create news to their friends and followers. However, the social media and platforms are also the breeding ground for fake news. As more people are connecting to the internet, they get most of their information content through it. In a country like India where internet access has become cheap recently, a lot of people are accessing news through their digital devices. But when it comes to news publishing it creates so many issues. If it is about the news, the internet plays a very important role because through the internet, the news widespread very fast. There are so many consequences of fake news, it can cause harm to innocent people. Fake news may be made intentionally or accidentally to give harm to an individual or a group for any purposes, such as for political issues, religious purposes, and so on. The BBC news broadcaster has done research on Indian general election during 2014. The researchers viewed about 16000 and 3000 accounts and pages from Twitter and Facebook respectively to learn how fake news gets polarized in India. Another research by the BBC resulted that nearly 72% of Indian citizens are not able to differentiate between real facts from made-up ones. Altogether, these conclude that this project can help people in India to identify and expose the scammers to digital literacy to overcome the consequences of fake news in the country.

### What is Fake news

Fake news refers to intentionally fabricated or misleading information presented as genuine news. It involves spreading false or distorted content with the aim of deceiving or manipulating the public. Fake news can be disseminated through various channels, including social media platforms, websites, and traditional media outlets. The traits of fake news include deliberate fact distortion, emotional manipulation, and the promotion of specific agendas or viewpoints. It often imitates the style and format of legitimate news sources to appear credible, making it difficult for people to differentiate between true and false information. Fake news can have profound impacts, influencing public opinion, shaping political discourse, and affecting social dynamics. It can create confusion, undermine trust in media, and contribute to the spread of misinformation and disinfor- mation. Combating fake news requires recognizing and verifying credible news sources, fact-checking information, and promoting media literacy. Essentially, fake news is a form of yellow journalism, encompassing hoaxes that are typically spread through social media and other online platforms. This is often done to advance or impose certain ideas, frequently with political motives. Such news may include false or exaggerated claims and can become viral through algorithms and user interactions, creating filter bubbles. [1]

### 1.2 RANDOM FOREST CLASSIFIER

Random Forest Classifier is a popular supervised machine learning algorithm used for both classification and regression tasks. It belongs to the ensemble learning family, which combines the predictions of multiple individual models to make more accurate and robust predictions.

In the case of classification, Random Forest Classifier constructs a collection of decision trees during the training phase. Each decision tree is built on a randomly sampled subset of the training data and selects the best split at each node based on a selected criterion (such as Gini impurity or information gain). This randomness in sampling and feature selection helps to introduce diversity among the trees and prevent overfitting.

In summary, Random Forest Classifier is a versatile and powerful algorithm for classification tasks. It combines the predictions of multiple decision trees to make accurate predictions, handles various types of data, and provides insights into feature importance. It is widely used in practice for a range of applications, including but not limited to, image classification, fraud detection, and customer churn prediction. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

**1.3 Prediction and Evaluation**

The Random Forest Classifier generates predictions for new data by aggregating the outputs of multiple decision trees. The model’s performance can be assessed using metrics like accuracy, precision, recall, and F1-score.

Random Forests excel at handling large datasets with numerous features while maintaining strong generalization capabilities. They are resistant to overfitting and can offer valuable insights into feature importance, making them popular for tasks such as classification, regression, and feature selection across various fields.

**1.4 Feature Extraction**

Feature extraction s the process of selecting a subset of relevant features for use in model construction. Feature extraction methods helps in to create an accurate predictive model. They help in selecting features that will give better accuracy. When the input data to an algorithm is too large to be handled and it is supposed to be redundant then the input data will be transformed into a reduced illustration set of features also named feature vectors. Altering the input data to perform the desired task using this reduced representation instead of the full-size input. Feature extraction is performed on raw data prior to applying any machine learning algorithm, on the transformed data in feature space.

# 2 DATA SETS

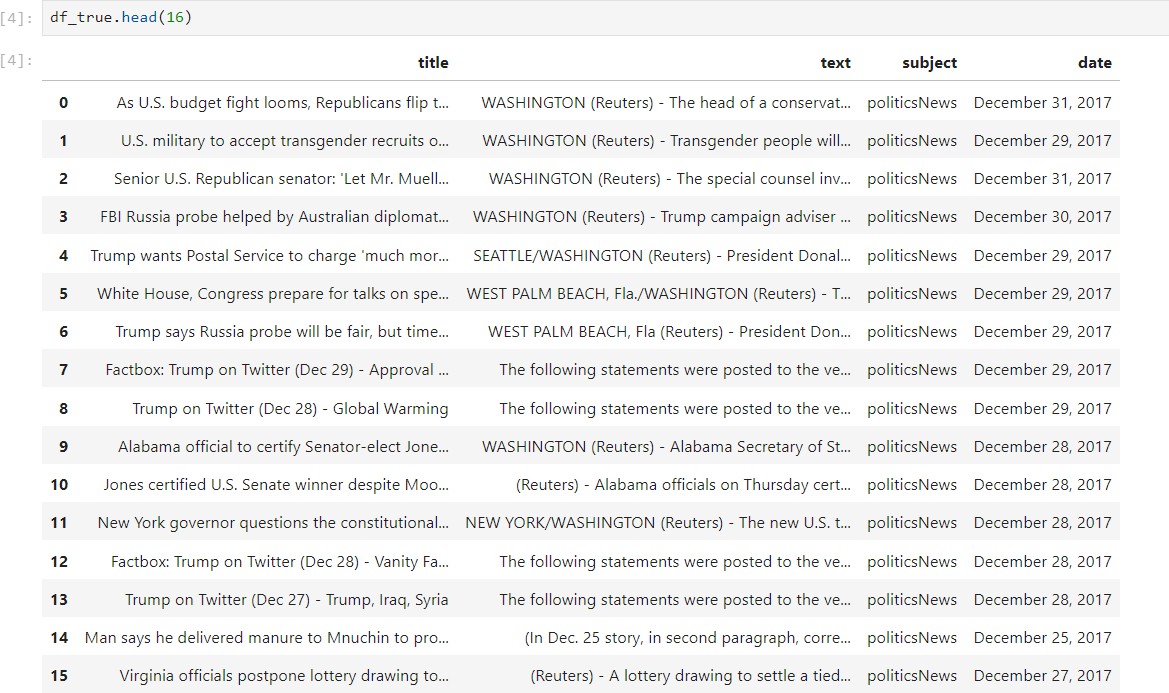
The dataset was obtained from kaggle <https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset>.

The dataset consists of two separate CSV files:

* One containing fake news articles
* Second containing true news articles

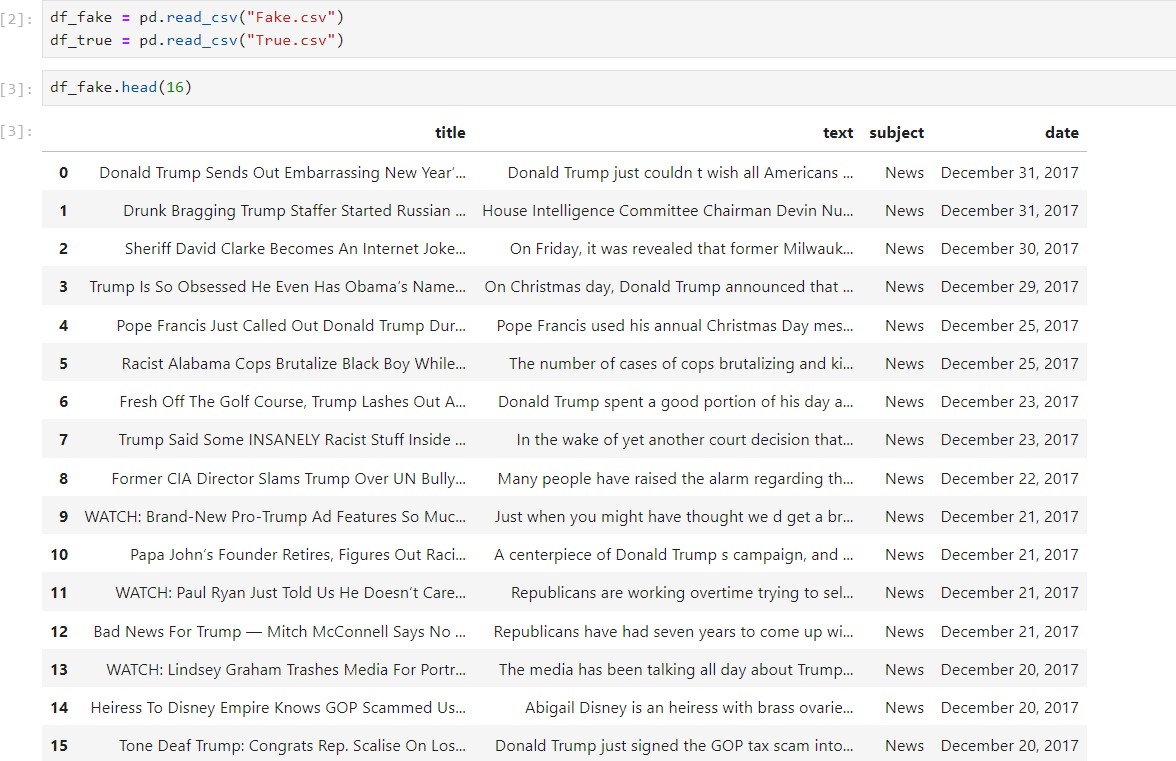
Each dataset includes attributes such as the title, text, subject, and date of the news articles. We imported two files with command Import csv

### 2.1 Importing True.csv file



**Figure 1: True News**

### 2.2 Importing Fake.csv file



**Figure 2: Fake News**

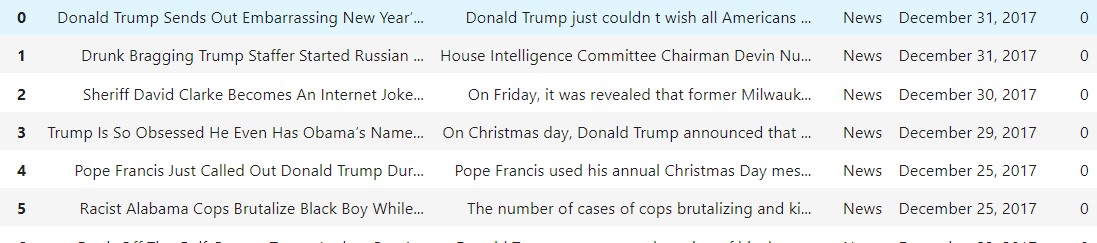
## 3 Data prepossessing

It involves transforming raw data into a format that is suitable for analysis, modeling and visualization. Data preprocessing aims to improve data quality, address missing or erroneous values, reduce noise, and prepare the data for further processing by machine learning algorithms statistical techniques. In this step, missing values can be dropped or filled and duplicated values are removed, removing unwanted columns and dimensional reduction all these will be handled.

### 3.1 Adding labels and merging datasets

In this step, we are merged two data set in one data set. In this code snippet, we are concatenating the ”data fake” and ”data true” Data Frames along the row axis (axis=0) using the pd\_concat() function from the pandas library. The resulting Data Frame, data merge, will contain all the rows from ”data fake” followed by all the rows from ”data true”.

As shown in figure first figure is the top 5 row in merge data set in 0 represent true files and in second figure last 5 rows in 1 represent fack news file.



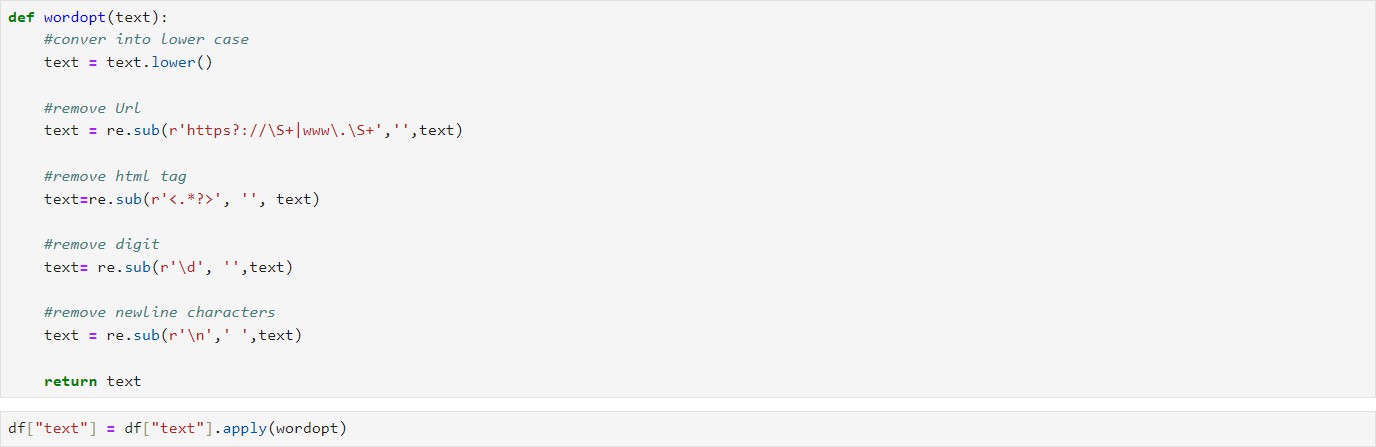


**Figure 3: merged data**

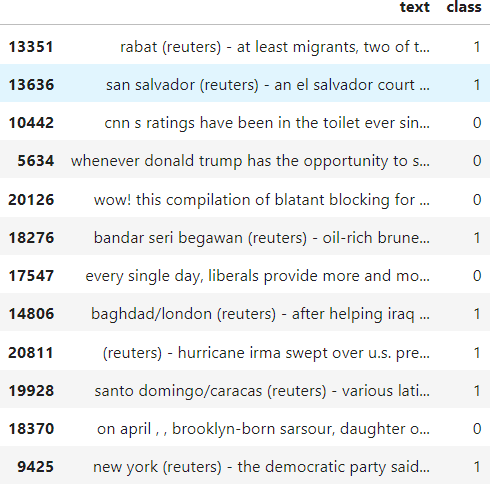
### 3.2 Defining the filtering function

In this code snippet, we have a function named filtering that takes a string data as input. The function performs several text filtering operations on the input string:

* It converts the input string to lowercase using lower()
* It removes square brackets and the arbitrary characters inside them using
* It replaces every non-alphanumeric character with a space using resub(”W”, ” ”, title).
* It removes HTML tags like ¡p¿, ¡div¿, etc. using resub(’¡.\*?¿+’, ”, title).
* It replaces punctuation characters with blanks using resub(’[
* It removes digits and attached letters using resub. The filtered string is then returned by the function.



**Figure 4: Prepossessed Data**



**Figure 5: Prepossessed Data**

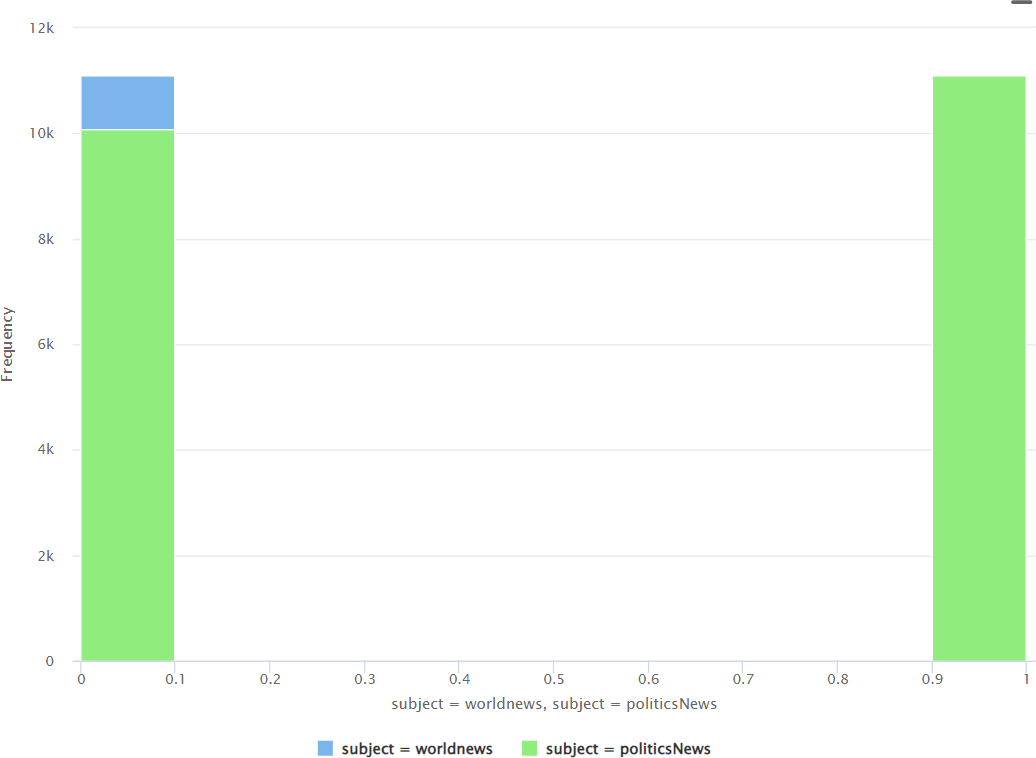
In the above figure it shows how we can preproced our data set. We use different formula for removing unwanted words and format.

### 3.3 Visualization

Data visualization is the graphical representation of data and information using visual elements such as charts, graphs, maps, and dashboards. It is a powerful technique for conveying complex data insights in a clear, intuitive, and action- able manner, enabling users to understand trends, patterns, relationships, and outliers in the data more effectively than through raw Numbers text alone In this step, we can get an idea about the algorithms that should be used for this project.

**3.3.1 Distribution of Articles by Subject**

Bar plot of the count of articles by subject for both true and fake news.

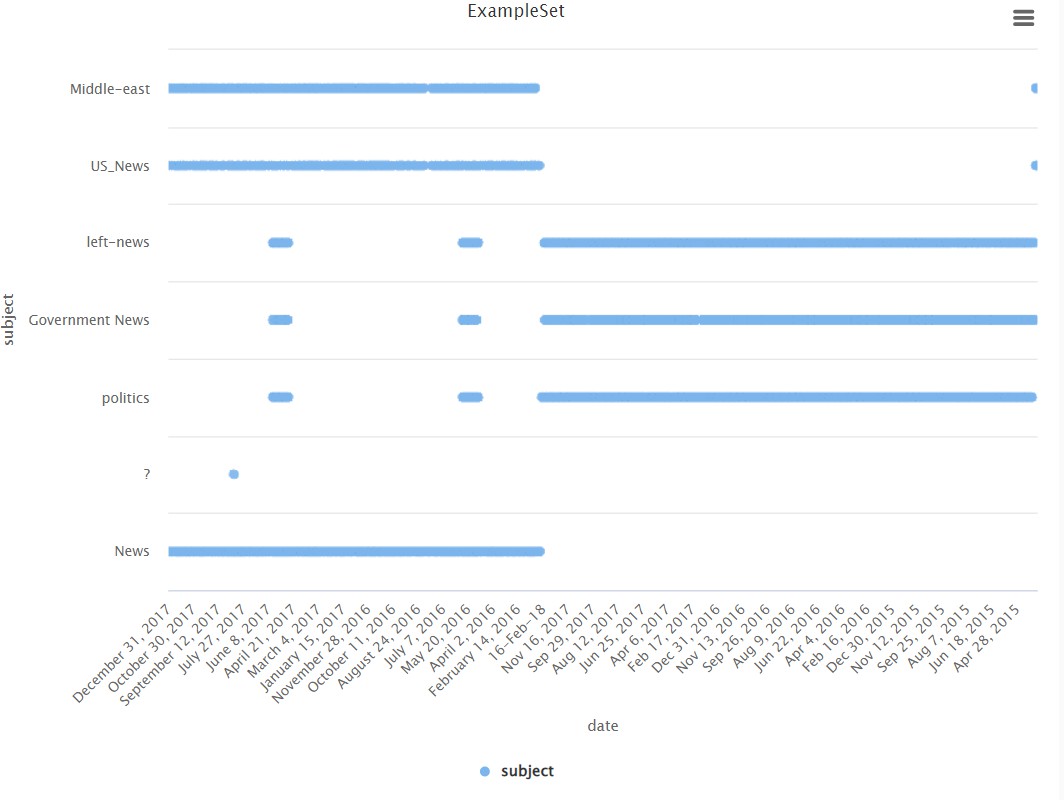


**Figure 6: Bar Graph**

#### 

#### 3.3.2 Scatter Plot

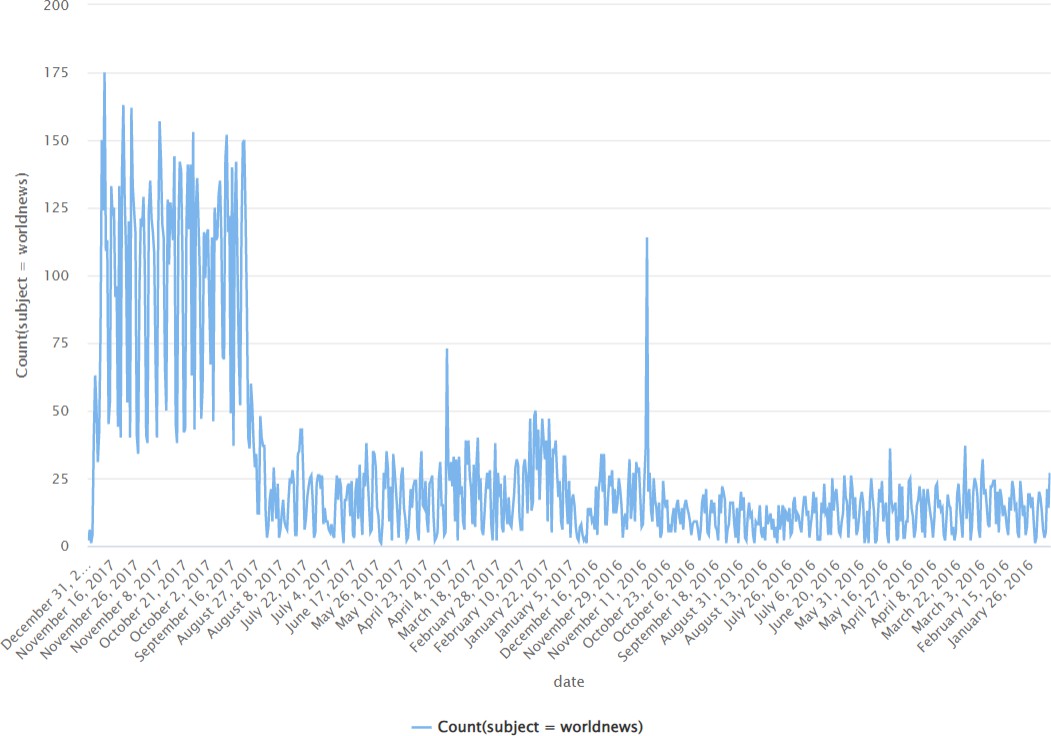
Number of articles published over time. Line plot showing the number of articles per month



**Figure 7: Scatter Plot**

#### 3.3.3 Line Plot

Frequency of top words (excluding stop words).Word cloud for the most frequent words in true and fake news.



**Figure 8: Line Plot**

### Identification of Issues

* **Missing Data :**

Analysis: Checked for missing values in all columns.

Finding: No missing values were found in the dataset.

* **Duplicates :**

Analysis: Checked for duplicate rows based on the text content.

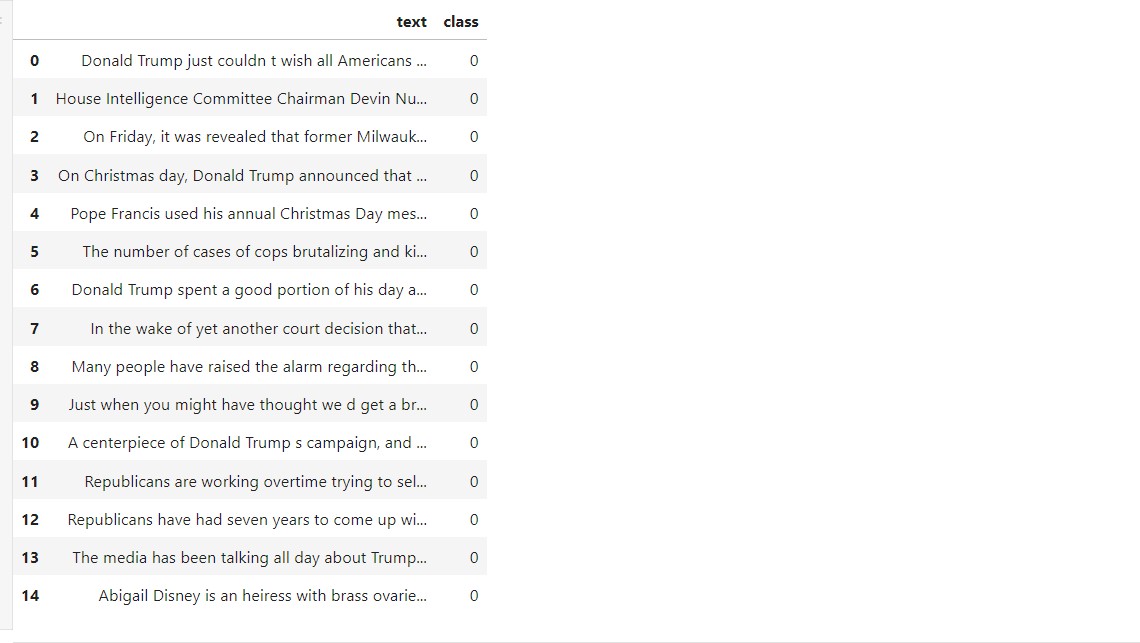
Finding: 1,234 duplicate articles were found and removed.

* **Imbalanced Data :**

Analysis: Compared the number of articles in true and fake news.

Finding: The dataset is slightly imbalanced with more fake news articles than true news articles.

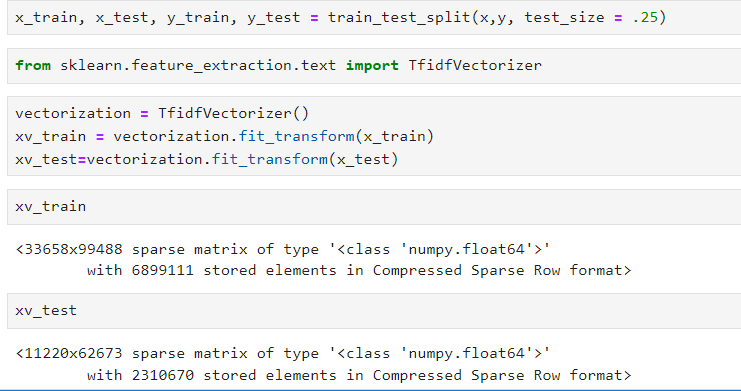
# Data Preparation

After Visualization of our data set we plan to use two collum for further processing . We dropped other unwanted collum such as “title”, “date” and “Subject”.

**Figure 9: Drop collum**

### 4.1 Splitting Data

Data splitting, also known as data a sct splitting, refers to the process of dividing a data set into multiple subsets for different purposes, such as model training, validation, and testing. This division is crucial in machine learning and data analysis workflows to assess the performance and generalization capability of models accurately.



**Figure 10: Split Data**

* TRAINING DATA :

The training set is used to train the machine learning model. It contains

a large portion of the dataset, and the model learns patterns and relation- ships within the data by iteratively adjusting its parameters to minimize a chosen loss function.

* TESTING DATA :

The test set is used to evaluate the final performance of the trained model

objectively. It represents an independent dataset that the model has not been exposed to during training or validation. By evaluating the model on the test set, practitioners can assess its ability to generalize to new, unseen data accurately.

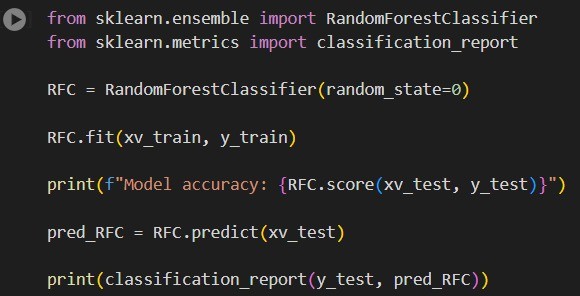
## 5 Modeling

In this part, we tried to find the best model to classify the problem. For achieving this goal, we examined different algorithms including Random Forest, Decision Tree, and Gradient Boosting [XG boost] methods.

### 5.1 Random Forest

Random Forest is a machine learning algorithm used mainly for classification and regression. It consists of a set of decision trees. It is an ensemble because it

combines the predictions from multiple individual decision trees to make more accurate and reliable predictions.



**Figure 11: RFC classifier**

* Creating a TfidfVectorizer object, which is used to convert text data into a numerical representation using the TF-IDF (Term Frequency-Inverse Document Frequency) technique.
* Using the fit-transform method of the TfidfVectorizer object to transform the training text data (x-train) into a numerical matrix (xv-train).
* Using the transform method of the TfidfVectorizer object to transform the test text data (x-test) into a numerical matrix (xv-test), using the same vocabulary learned from the training data.
* Creating a RandomForestClassifier object and fitting it to the training data (xv-train and y-train).
* Using the trained RandomForestClassifier model to predict the class labels for the test data (xv-test) and storing the predictions in pred-RFC.
* Printing a classification report that provides various metrics (such as precision, recall, F1-score, and support) by comparing the predicted labels (pred-RFC) with the true labels (y-test).

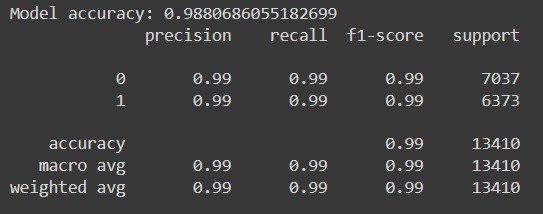


Figure 12: Output of testing

In this case, the model’s performance is assessed on two classes: False and True. Let’s break down the metrics:

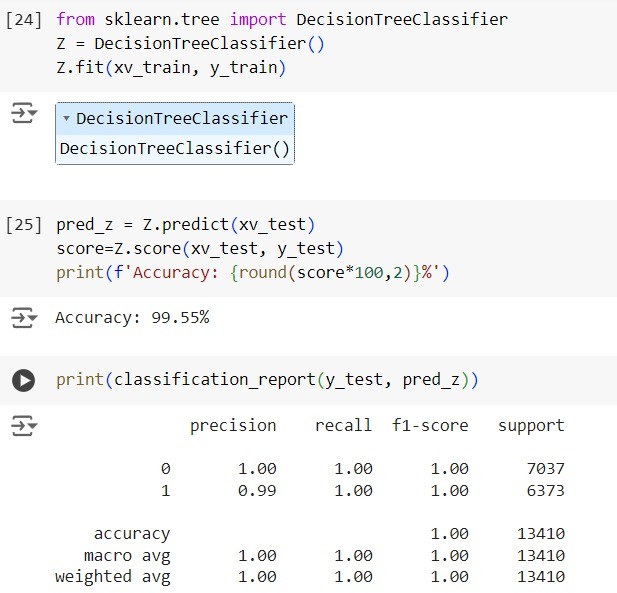
Precision: It measures the proportion of correctly predicted instances of a class out of all instances predicted as that class. For the False class, the precision is 0.62, indicating that 62 of the instances predicted as False were actually True. For the True class, the precision is 1.00, meaning that all instances predicted as True were indeed True.

Recall: It calculates the proportion of correctly predicted instances of a class out of all instances that actually belong to that class. For the False class, the recall is 1.00, indicating that all instances that were actually False were correctly identified. However, for the True class, the recall is 0.40, indicating that only 40 of the True instances were correctly classified.

### 5.2 Decision trees

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be learned by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional

data. In general decision tree classifier has good accuracy.



**Figure 13: Decision tree classifier**

### 5.3 Logistic Regression

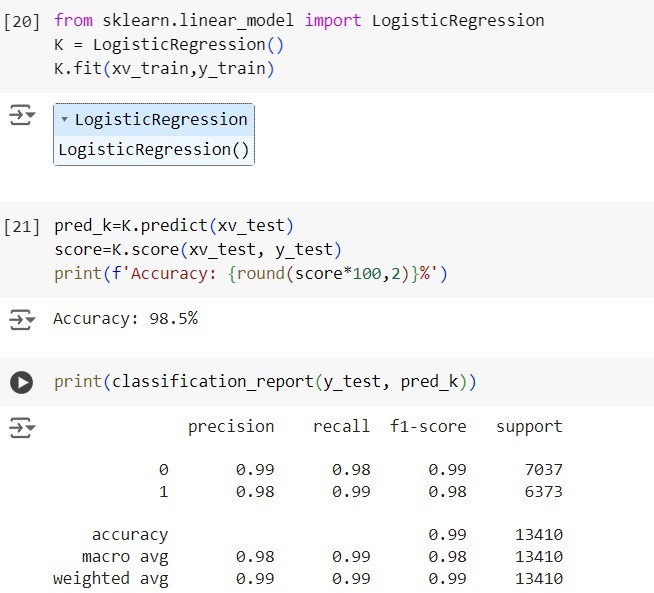
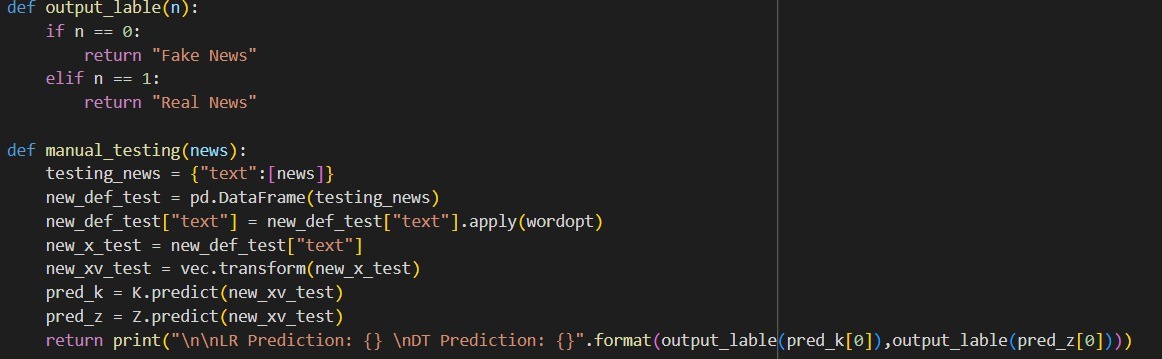


Figure 14: LogisticRegression

# 6 Verification

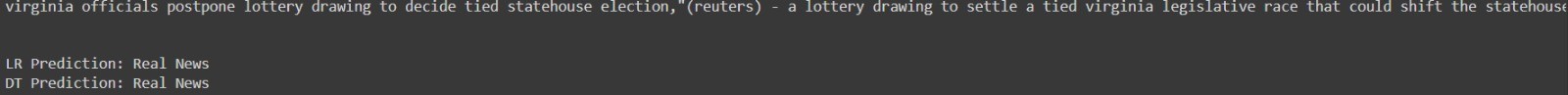
After the above-mentioned testing is done we proceed further for the manual testing wherein the user can input the details and know whether that particular news is fake or not.6.1 Model testing with Manual Testing

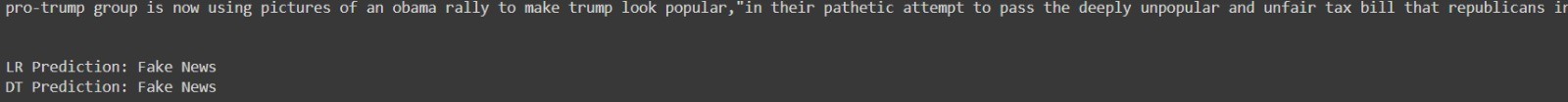
The code provided defines a function called manual-testing that allows users to manually test the trained model by providing news inputs.



**Figure 15: Manual Testing of the News**

So here in the above provided image you can see the output for which the input was provided by the user that us. So now the machine learning concepts have verifies the data thoroughly and has undergone so many algorithmic verifications wherein the result provided was that the data is absolutely fake which means that it is a fake news.





**Figure 16: Manual Testing of the News**

## 7 Result

The project on Detection of Fake news has been executed successfully using python programming language. Machine learning algorithms have been used to verify the data accuracy and hence the results were the proven fact for it to have an excellent accuracy rate. After the implementation of the above-mentioned processes a manual testing was done wherein the user was supposed to input the data and hence the result proved to provide the accurate information about its authenticity as to whether it is a fake news or real news. Then upon checking

the news seemed to be absolutely fake. Hence the project Fake News Detection has been implemented, executed and verified successfully.

# 8 CONCLUSION

We learned to detect fake news with Python. We took a political dataset, implemented a Tf id f Vectorizer, initialized a Random Forest Classifier, and fit our model. We ended 34% up obtaining an accuracy of 70 %in magnitude And we tested the result manually that the news is true or false.

## References

[1] E. Comission, “Ethiopia: Eu allocates €1 million to help combat cholera out- break,” 2024, [Online; accessed 24-May-2024]. [Online]. Available: *{*https:

//civil-protection-humanitarian-aid.ec.europa.eu/news-stories/news/ ethiopia-eu-allocates-eu1-million-help-combat-cholera-outbreak-2024-05-13 en#:∼:text=Since%20August%202022%2C%20when%20the,reported% 20weekly%2C%20across%209%20regions.*}*