Fraud Article Detector Based on Sources

Milestone: Final Project Report

Group 40
Navya Pragathi Molugaram
Sanjana Konte

617-318-8317
857-265-8414

molugarm.n@northeastern.edu
konte.s@northeastern.edu

Percentage of Effort Contributed by Student1:___50%_____

Percentage of Effort Contributed by Student2:____50%_____

Signature of Student 1: _____NavyaPragathiMolugaram_____

Signature of Student 2:_____SanjanaKonte_____

Submission Date:_____04-21-2023_____

I. Problem Setting

Social media forums have been a major contributor to the dissemination of false information and conspiracy theories. It offers many advantages for us and has developed into a forum for us to openly voice our thoughts, but it has also been used to propagate misleading content that could be dangerous. During the new coronavirus pandemic, a study performed in the MedLine, Virtual Health Library showed that the spread of incorrect information directly influences people's lives and an increased risk of emotional overload, distress, and anxiety in society. Politicians, Organisations, and also certain news media assist in propagating incorrect information to sway people's judgments as well as distort public events. Manually classifying the authenticity of such news is arduous, biased, and time-consuming. Therefore, a necessity for efficient detection and mitigation tools arises to tackle the rising problem.

II. Problem Definition

Our primary objective is to identify a practical method for classifying real-time data and to dissect the various aspects of machine learning utilized in the classification of fake articles. Our intention is to find the patterns included in the articles in the form of embedded spam URLs, unusual word frequency, lavish wording, etc which could give a gist of how the fake news is structured. Upon successful deployment, this model can be used as an additional check to find the legitimacy of any news posted online.

III. Data Sources

Among the various open source datasets available, we have taken the KaggleFN - Fake News Detection dataset consisting of articles from 244 websites representing 12,999 posts which were pulled using the webhose.io web crawler. The study summary used in the data description was referenced from a publication on springer. Research

IV. Data Description

The dataset consists of 12,999 records and 20 attributes including one output variable spam_score which indicates (0) for real and (1) for fake articles. Some of the predictors for classification would be Published Date - datetime format, Language, Type, Label - categorical type, Site_URL, Image_URL, Title, Text, hasImage, and Author - character format.

V. Data Preprocessing:

Text, photos, video, and other types of unprocessed, real-world data are disorganized. In addition to the possibility of inaccuracies and inconsistencies, it frequently lacks a regular, uniform format. Data must first undergo a series of preprocessing actions in order to be read.

- The shape of a data frame, as determined by its rows and columns, is known using the shape function.
  (12,999,20) denotes that there are 12999 rows and 20 columns in the data frame.
- The info function is used to determine the datatype, non-null values, range index, and memory use for each variable. The data types for the provided data frame are float, int, and object, all with non-null values.
- The duplicate function is utilized to look for duplicates in the data. We noticed that there are no duplicate values in the data.
- Using the IsNull function, we have now determined if the data contains any null or missing values. 13 out of 20 records are found to be missing values.

uuid	0
ord_in_thread	0
author	2424
published	0
title	680
text	46
language	0
crawled	0
site_url	0
country	176
domain_rank	4223
thread_title	12
spam_score	0
main_img_url	3643
replies_count	0
participants_count	0
likes	0
comments	0
shares	0
type	0

Table1 : Data count

- The text columns such as title, text, and thread_title were preprocessed by removing the stop words and converting them into lowercase. Stemming and tokenization was performed using the nltk library.

VI. Data Exploration:

Exploring a huge data set in an unstructured manner is part of the data exploration process in order to identify the first patterns, traits, and areas of interest. This procedure helps to build a broad picture of significant trends and key topics to investigate in more detail rather than revealing every piece of information that a dataset contains. Data exploration can combine human techniques with automated tools such as early reports, charts, and data visualizations.

Insight1:

- A heat map is plotted to find the relationship between each numeric column.

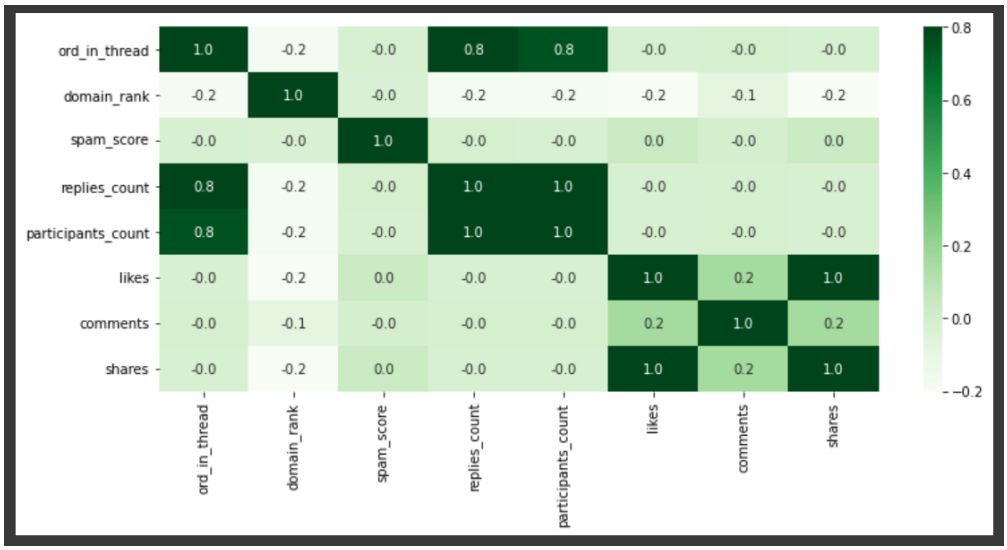


Fig 1: Correlation plot

Insight2:

- For the column country, a count plot is used to show the count using bins. As the majority are in the US, filling the null values with it.

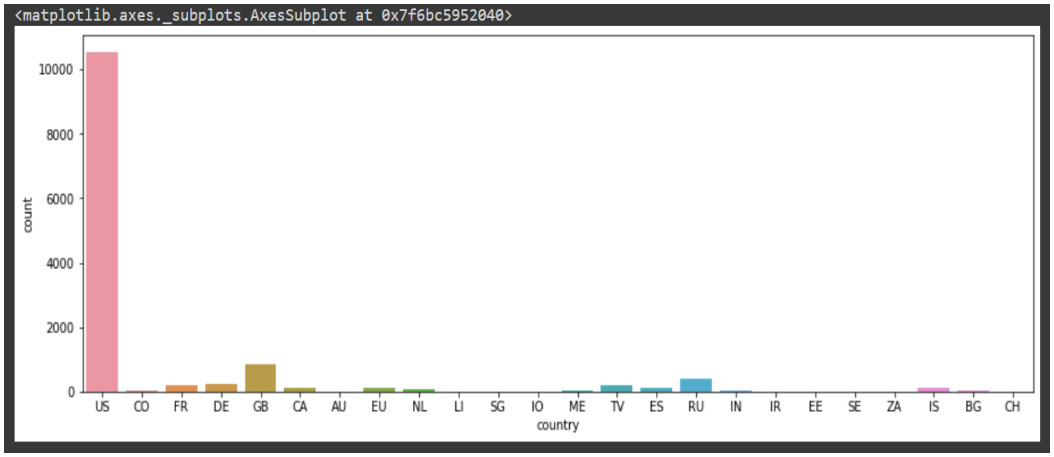


Fig 2: Distribution of country

- There is a varied distribution of countries but only a few records are present for some countries. Hence removing the articles from countries having less than 20 records.
- To view the count of languages, the count plot has been plotted. It is observed that relatively the number of English articles is very large so limiting the scope.

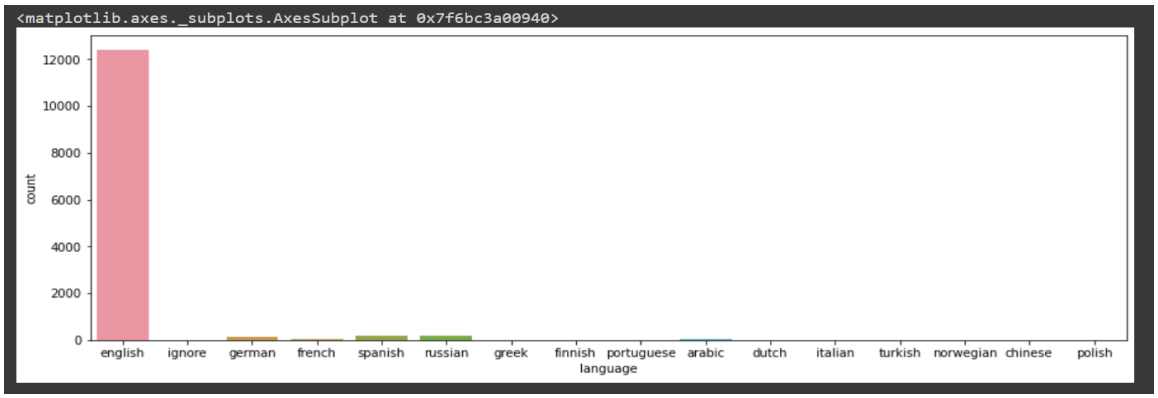


Fig 3: Distribution of language

Insight3:

- The replies_count, likes, shares, comments, and participants_count have been visualized and it is observed that there are a lot of null values. Hence, these columns can be considered redundant.

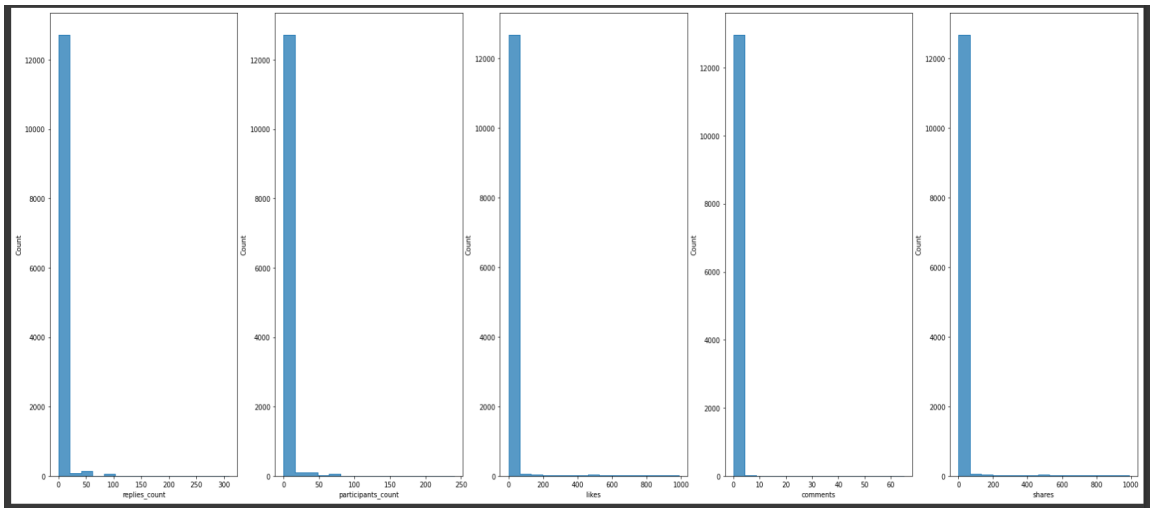


Fig 4: Distribution of customer reviews based predictors

Insight4:

- The below graph shows that the data published and crawled has significantly decreased across the time

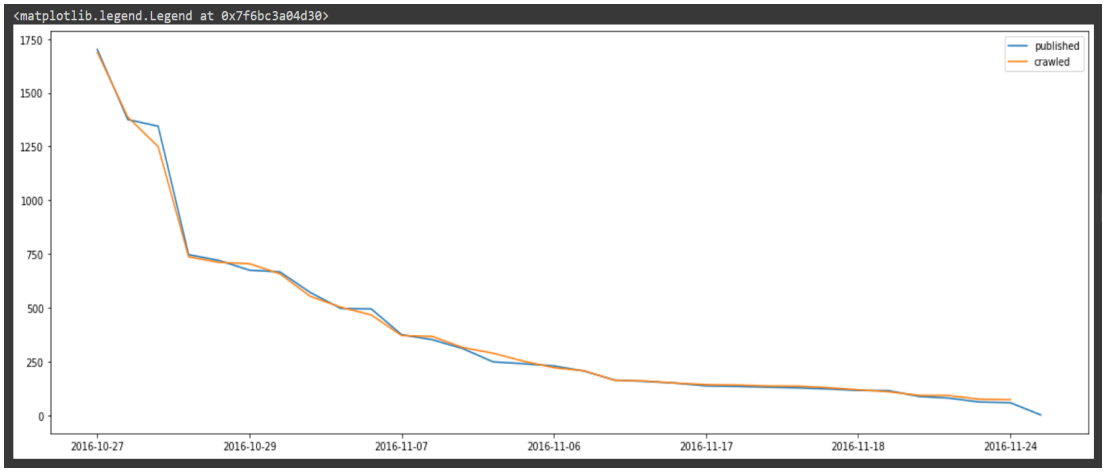
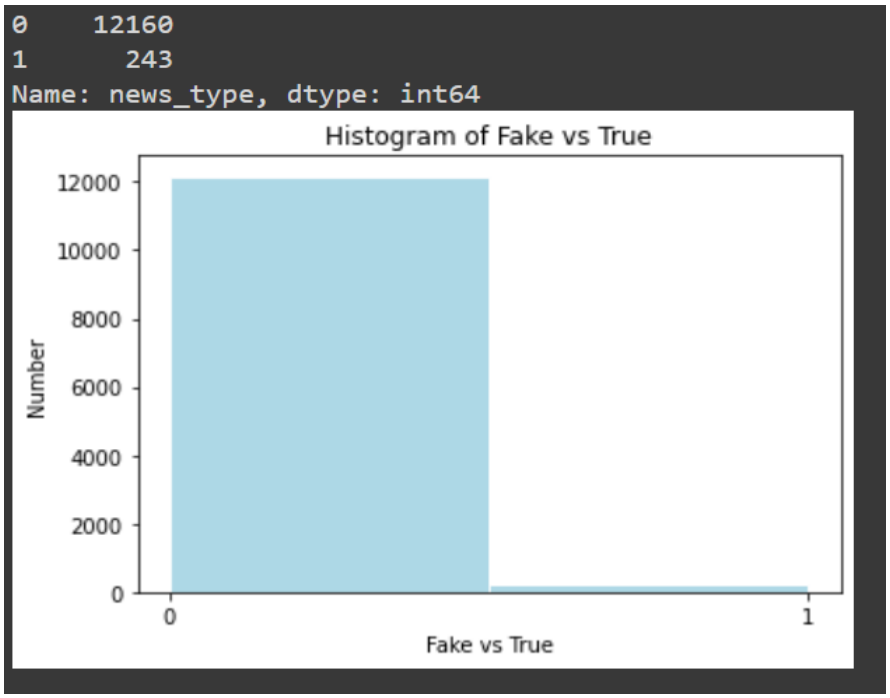


Fig 5: Crawled and published date of articles

Insight5

- From the plot below it can be seen that the percentage of fake data is very less compared to the non-spam data.



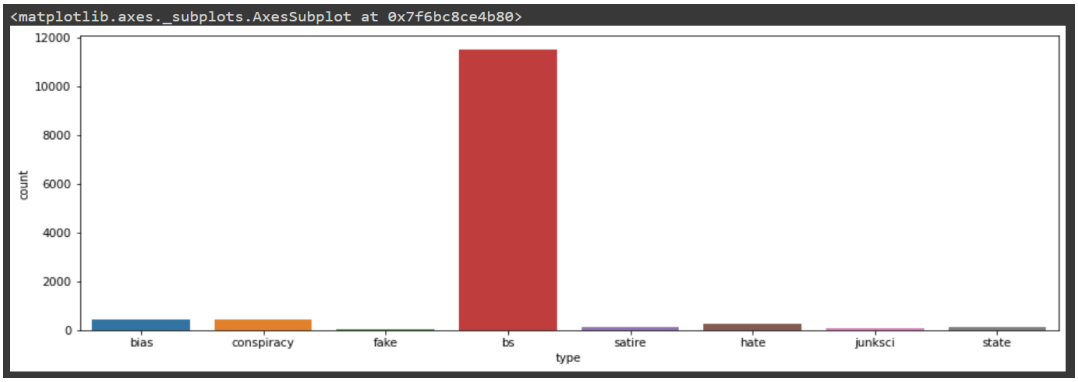
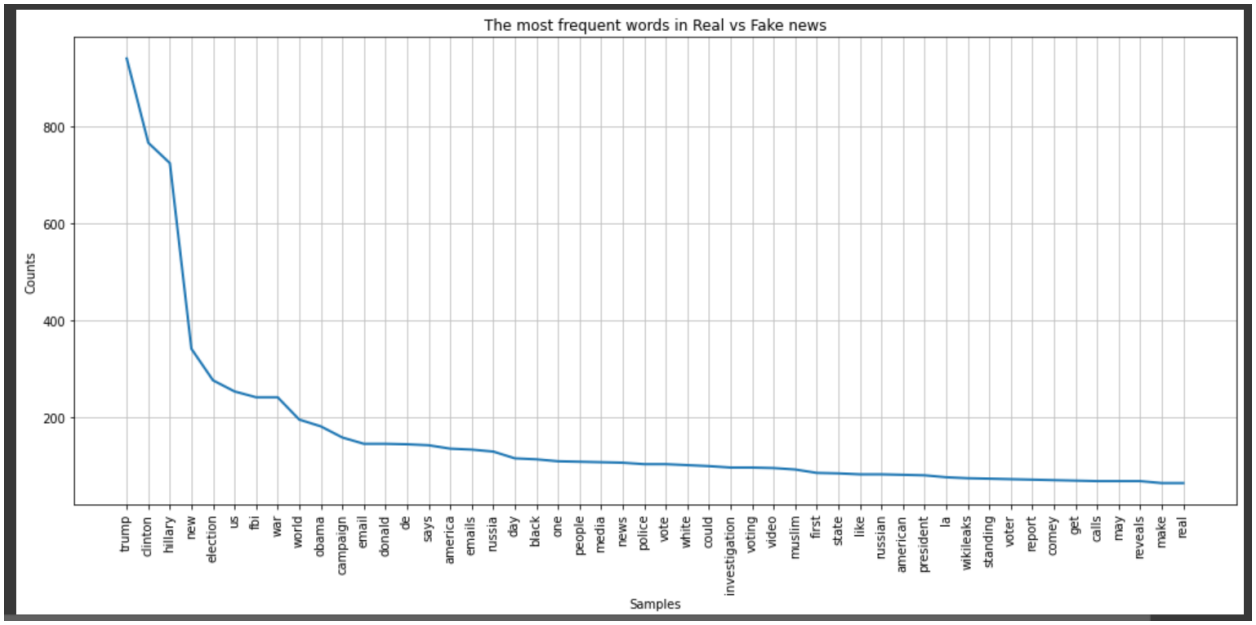


Fig 6: Distribution of classes of the target variable

Insight6:

- It is observed that the word trump, Clinton, Hillary, and new are the most frequent words used in the title of the books.



- A word cloud was used to visualize the most frequent words in real news and fake news

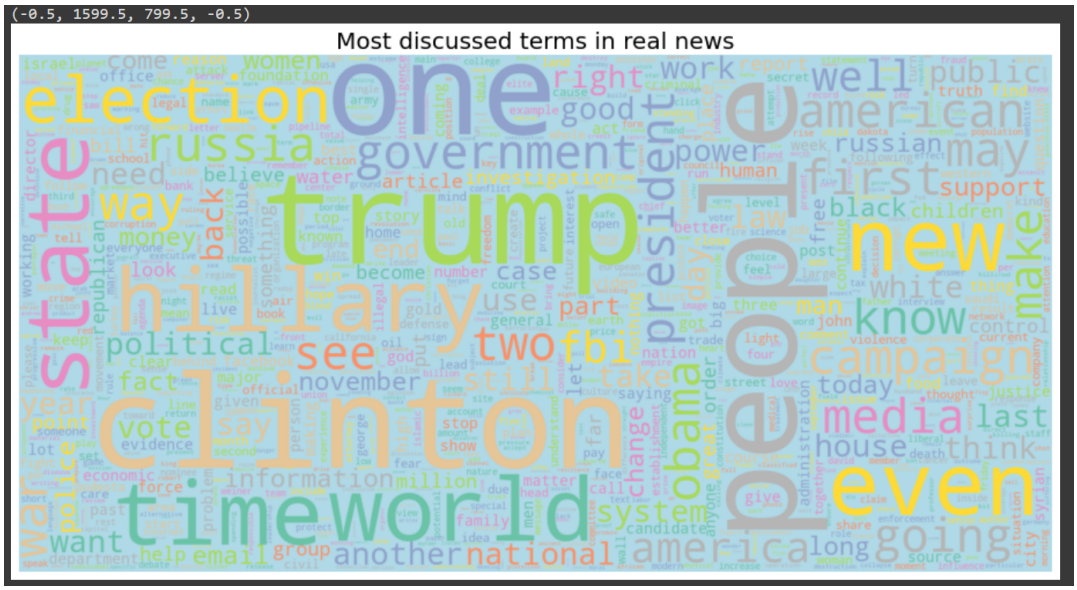


Fig 7: Word Cloud of terms used in real and fake news

VII. Dimension Reduction:

- From the above analysis, the columns such as uuid, crawled, domain_rnk, replies_count, participants_count, likes, comments, and shares look redundant. So these columns have been dropped for further analysis in modeling.
- For categorical variables, pd.getdummies method was used to convert the columns to integer variables

- Further, Principal component analysis was performed to transform the variables to principal components and feature selection was performed on the dataset to extract the minimum variables that can capture maximum variance.

```
Explained Variance:
 [0.13886743 0.11070234 0.03126648 0.0271487  0.02512234 0.01186883
 0.01132249 0.01052286]
Proportion Variance:
 [0.34697623 0.27660252 0.07812288 0.06783415 0.06277104 0.02965564
 0.02829054 0.02629257]
Cummulative proportion of variance:
 [0.34697623 0.62357876 0.70170164 0.7695358  0.83230684 0.86196247
 0.89025301 0.91654558]
```

Table 2: PCA Analysis

- Thus it can be observed that 91% of the variance can be captured by eight components.

VIII. Data Partitioning

The predictor variables are represented using 'X' and the target variable 'news-type' is represented using the variable 'Y'. The train_test_split method from sklearn is used to split the data into training and testing dataset. Using the holdout method, 70% of the dataset is used for training the classification data and the remaining 30% of the dataset is used to evaluate the model performance. x_train consists of 6367 records and 24 variables, while x_test consists of 3137 records and 24 variables. Y_train consists of 6367 records, while Y_test consists of 3137 records.

IX. Data Mining Models/Methods

The following model was built using the training data mentioned above

1. Logistic Regression

The algorithm estimates the probability of classifying an event as fake or real, based on a given dataset of independent variables. It uses a threshold cutoff on the probabilities for classifying into either of the classes. The outcome function called logit is used to model a linear function of predictors

Advantages :

- ➔ It is easier to implement, interpret, and very efficient to train
- ➔ It makes no assumptions about distributions of classes in feature space
- ➔ It can interpret model coefficients as indicators of feature importance

Disadvantages :

- ➔ It constructs linear boundaries
- ➔ It assumes linearity between the dependent variable and independent variables.
- ➔ It requires average or no multicollinearity between independent variables.

Implementation :

The base logistic regression model was executed resulting in an accuracy score of 0.88

	precision	recall	f1-score	support
bias	1.00	0.14	0.25	97
conspiracy	1.00	0.11	0.19	75
fake	0.00	0.00	0.00	4
bs	0.88	1.00	0.94	2165
satire	1.00	0.09	0.16	23
hate	1.00	0.05	0.10	55
junksci	0.00	0.00	0.00	31
state	1.00	0.03	0.05	36
accuracy			0.88	2486
macro avg	0.74	0.18	0.21	2486
weighted avg	0.88	0.88	0.84	2486

Table 3: Performance Evaluation for Logistic Regression

2. Support Vector Classifier (SVCs)

It is a type of machine learning algorithm that can be used for both classification and regression analysis. The SVM would learn to identify patterns and features in the text that distinguish between fraudulent and non-fraudulent articles. When a new article is presented to the SVM, the algorithm would then uses these learned patterns and features to predict whether the article is fraudulent or not. The SVM outputs a binary classification, indicating whether the article is likely fraudulent or not.

Advantages :

- ➔ Effective in handling high-dimensional feature spaces.
- ➔ Good at separating complex, non-linearly separable data.
- ➔ Computationally efficient and can be trained on large datasets with relatively few computational resources.

Disadvantages :

- ➔ support of big datasets and several dimensions at once is not possible.
- ➔ Easily prone to overfitting when the number of features is large.
- ➔ Computationally expensive and time-consuming.

Implementation :

The base SVC model was executed resulting in an accuracy score of 0.91

	precision	recall	f1-score	support
bias	0.94	0.32	0.47	104
conspiracy	1.00	0.33	0.50	84
fake	1.00	0.14	0.25	7
bs	0.90	1.00	0.94	2141
satire	1.00	0.15	0.26	34
hate	0.88	0.17	0.29	40
junksci	0.00	0.00	0.00	34
state	0.96	0.62	0.75	42
accuracy			0.90	2486
macro avg	0.83	0.34	0.43	2486
weighted avg	0.89	0.90	0.87	2486

Table 4: Performance Evaluation for SVC

3. Random Forest

The algorithm would then learn to identify patterns and features in the text that distinguish between fraudulent and non-fraudulent articles. When a new article is presented to the random forest, the algorithm would use each decision tree to predict whether the article is fraudulent or not. The final output of the random forest is then determined by aggregating the predictions of all the decision trees.

Advantages :

- ➔ Handles high-dimensional feature spaces.
- ➔ Good at identifying complex, non-linear relationships between features.
- ➔ Computationally efficient, especially when dealing with large datasets, as each decision tree can be trained in parallel.

Disadvantages :

- ➔ More trees slow down the model and are prone to overfitting.
- ➔ Involves combining the outputs of multiple decision trees, making it difficult to interpret.

Implementation :

The Random Forest model was executed resulting in an accuracy score of 0.95

	precision	recall	f1-score	support
bias	0.99	0.65	0.79	104
conspiracy	0.93	0.51	0.66	84
fake	1.00	0.71	0.83	7
bs	0.94	1.00	0.97	2141
satire	1.00	0.15	0.26	34
hate	1.00	0.82	0.90	40
junksci	1.00	0.35	0.52	34
state	1.00	0.98	0.99	42
accuracy			0.94	2486
macro avg	0.98	0.65	0.74	2486
weighted avg	0.94	0.94	0.93	2486

Table 5: Performance Evaluation for Random Forest

4. Decision Tree

The decision tree algorithm can be used to classify new articles as fraudulent or non-fraudulent by following the path of the tree from the root node to a leaf node. Each internal node represents a decision based on the values of a feature, and each leaf node represents a predicted class label. To classify a new article, the decision tree algorithm evaluates the feature values at each internal node and selects the appropriate branch until it reaches a leaf node with a predicted class label.

Advantages :

- ➔ Easy to understand and interpret.
- ➔ Handle both numerical and categorical data.
- ➔ Can handle missing values by using splits to fill in missing values or by ignoring missing values altogether.

Disadvantages :

- ➔ Prone to overfitting when the tree is too complex.
- ➔ Unstable and less dependable when data even slightly changes.
- ➔ Biased toward features with many categories or towards features with many levels.

Implementation :

The base Decision Tree model was executed resulting in an accuracy score of 0.99

	precision	recall	f1-score	support
bias	0.95	0.94	0.95	104
conspiracy	0.89	0.94	0.91	84
fake	1.00	1.00	1.00	7
bs	0.99	0.99	0.99	2141
satire	1.00	1.00	1.00	34
hate	1.00	1.00	1.00	40
junksci	0.97	0.97	0.97	34
state	0.98	1.00	0.99	42
accuracy			0.99	2486
macro avg	0.97	0.98	0.98	2486
weighted avg	0.99	0.99	0.99	2486

Table 6: Performance Evaluation for Decision Tree

Project Results

The below performance metrics were calculated to evaluate the model's classification performance on the holdout dataset.

Models	Accuracy	Precision	Sensitivity	F1score
Logistic Regression	0.884151	0.836734	0.386914	0.490960
SVM	0.913113	0.981095	0.671636	0.772724
Random Forest	0.950925	0.979923	0.964841	0.970916
Decision Tree	0.993162	0.836734	0.386914	0.490960

Table 7: Accuracy vs Precision vs F1 score

It is observed that a decision tree results in the highest accuracy but gives a lower sensitivity and F1 Score which implies that the model is not able to classify the spam news accurately (as the proportion of spam news is very low)

SVM gives a higher score as compared to Logistic Regression and Decision Tree but still fails to classify the priority class accurately. From the above metrics, it can be concluded that we can choose Random Forest Model for our classifying task and fine-tune the parameters further to improve the model's accuracy.

Future Scope

There is a critical need for comprehensive datasets such as KaggleFN (Getting Real With Fake News) that demand teamwork, curation, and crowdsourcing . Further, such technologies and expertise must be put to practical use in order to not only mitigate the spread of false information but also to educate people about the type of news they consume in an era where fake WhatsApp forwards and Tweets can influence

innocent minds. In addition , it is also concluded that the following steps demand for studies involving methods for analyzing texts for natural language processing as well as semantic and syntactic analysis. It is possible to improve and simplify the creation of useful applications that allow users to learn from the articles they consume, fact-checking websites, built-in plugins, and article parsers, and more importantly to create awareness .