ILLINOIS INSTITUTE OF TECHNOLOGY

Fake News Classification

Praveen Surendran
A20496784
(s5@hawk.iit.edu)

Niveditha Mangala Venkatesha
A20466182
(nmangalavenkatsha@hawk.iit.edu)

December 2nd 2022

Prof. Yan Yan
Machine Learning - CS 584

Abstract

Fake news is a social issue that has an impact on democratic countries by influencing and manipulating public opinion. The term "fake news" refers to a wide spectrum of material that appears to be true. They may, for example, contain outright false assertions, certain portions of the truth, or even convey the facts in a distorted manner to disqualify a particular figure or a specific group. The intolerable influence of fake news necessitates the development of tools to assist professional journalists, on the one hand, and the general public, on the other, in detecting fake news. In this project, we describe a methodology for distinguishing between fake and authentic news in a data set comprising hundreds of human-curated journalistic items. The concept depends on shallow classifiers and is based on a bag-of-words method. We wanted to test the notion that authentic and fraudulent news may be distinguished using basic assumptions.

Introduction

Fake news is information that is untrue, partially true or altered to offer the public a distorted impression of a certain process, public actor, or social phenomenon. Fake news is a huge source of confusion in all types of societies, hence massive attempts are underway, primarily from public organizations, to uncover not only fake news but also the players behind it. The dissemination of fake news is typically faster than that of factual news. Because many steps must be performed to combat erroneous information, the difference in dissemination rate amplifies the damages caused by misleading information.

So far, the human aspect in the detection of bogus news is the best alternative. The detection of questionable news is a difficult process. To recognize bogus news, the reader must go through a sequence of processes. First, she or he should determine the subject of the essay, then, in the second stage, try to contextualize what is already known, and, last, validate that information in other reliable sites or media. He/she could also confirm the information with an expert. The mainstream media is not immune to the spread of fake news. Indeed, some big newspapers and television firms are often involved, consciously or unknowingly, in the spread of fake news.

A classifier can provide the public with an automatic tool that can aid in the identification of suspicious news. The challenge of automatic classification is divided into two parts. The first is a dataset, which typically consists of a sizable sample with an appropriate number of instances belonging to a relevant class or label. The second part is an algorithm that can distinguish the appropriate classes based on specific characterizations of examples in the dataset. The classification in the situation at hand, the determination of whether a journalistic piece is true or untrue (fake), is a dichotomy, meaning that it either belongs to one class or the other. We describe our efforts to categorize journalistic works as true or false in this project.

The Dataset (Fake News Classification)

Link: https://www.kaggle.com/datasets/ruchi798/source-based-news-classification

This dataset, which has 2095 entries of news data, comprises 12 different variables that describe various information related to the news.

The first step to the working of the project is to collect the dataset and analyze the obtained data set.

1. author: The author of the reported news
2. published: Date on which the article was published
3. title: Title of the content
4. text: Article text
5. language: Language written in
6. site_url: Link to the article
7. main_img_url: Link to the image in the article
8. type: Article type
9. label: Fake or real (response variable)
10. title_without_stopwords: Title without stopwords
11. text_without_stopwords: Article text without stopwords
12. hasImage: Whether the article has an image or not

Almost all of the variables are categorical with only the 'hasImage' variable being numerical.

We shall have a detailed look at the dataset further down in the Data Cleaning step.

Proposed Methodology

Data Preprocessing

Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, we must preprocess our data before feeding it into our model. [3] Let us see the steps to be followed for successfully analyzing the data:

Dealing with Missing Values:

There are only a few NaN values in our dataset. Since the count of null values is small, we can just drop the row corresponding to the missing value. A total of 104 missing values were removed from the dataset and now the dataset is free from null values.

***Feature*: published**

The published feature has time data which is of a different format from which the model can understand. So, to convert the feature to an understandable format for the model, we first split the time data into 'time' and 'date'. We then split 'date' into 3 subparts 'year', 'month', and 'day'.

There were two random error-producing values, and we drop the corresponding rows, which are the 848th and 1838th rows.

Creating Dummy Variables:

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups. In the simplest case, we would use a 0,1 dummy variable where a person is given a value of 0 if they are in the control group or a 1 if they are in the treated group. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. This means that we don't need to write out separate equation models for each subgroup. [18]

***Feature*: label**

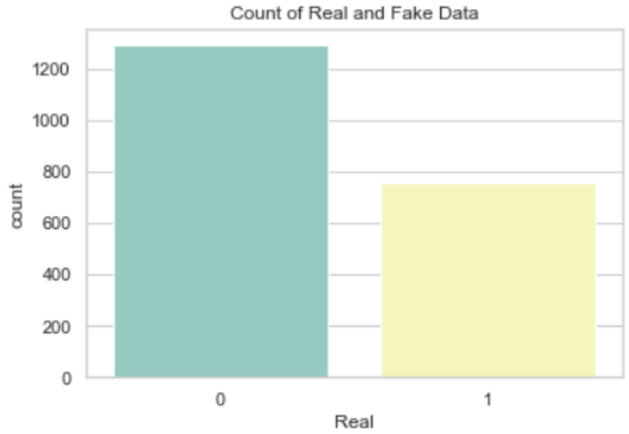
We create dummy variables for the feature 'label' to represent Real as 1 and Fake as 0.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis is the crucial process of doing preliminary investigations on data to uncover patterns, spot anomalies, test hypotheses, and validate assumptions using summary statistics and graphical representations. [4]

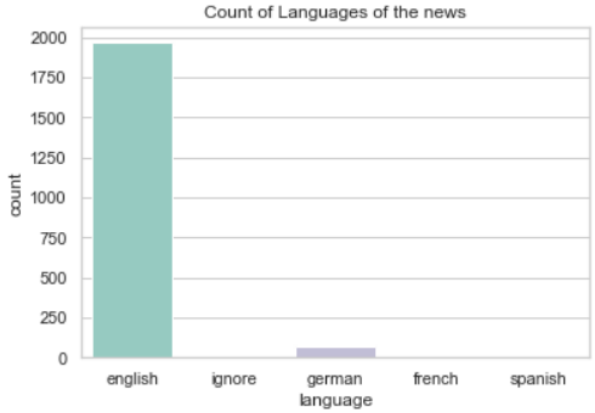
Feature: Real

Real is now our predictor/response variable. Upon plotting the count plot, we can see that there is almost 2 the amount of news that is classified as fake than those that are classified as true.



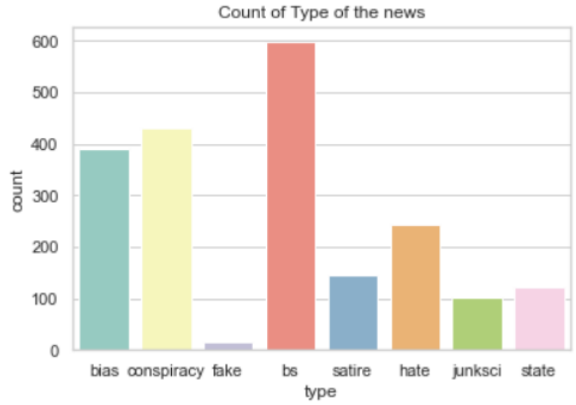
Feature: language

By plotting the count plot for the language feature, we can see that the vast majority is English. There are a total of 5 unique values in this feature: English, German language, French, Spanish, and ignore.



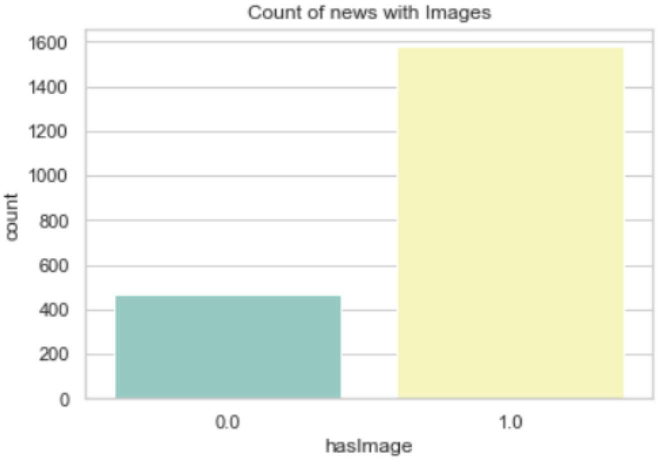
Feature: type

From the count plot, we can see that there are 8 different types and the vast majority is bs.



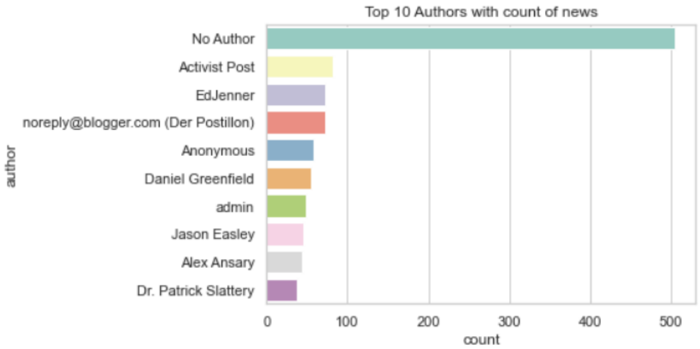
***Feature*: hasImage**

Similarly, plotting the count plot for the hasImage feature, we can see that there are two unique values for the feature, 0 and 1. We can notice that there are more 1's than 0's.



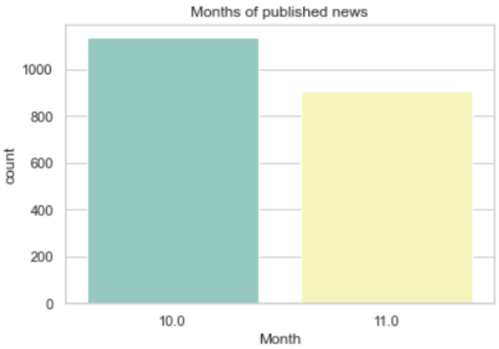
***Feature*: author**

Now, visualizing the count plot of the Top 10 authors in the author feature, we can see that 'No Author' has the majority count.



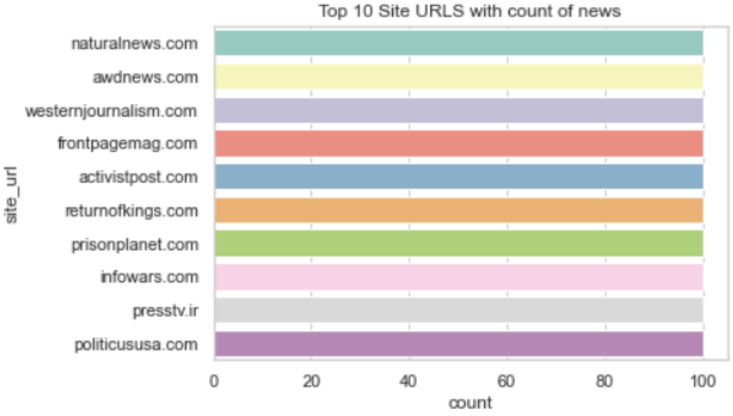
***Feature*: month**

By just viewing the data, we can see that there are only 2 values, 10 and 11. This means that the news data we have is just from the month of October and November.



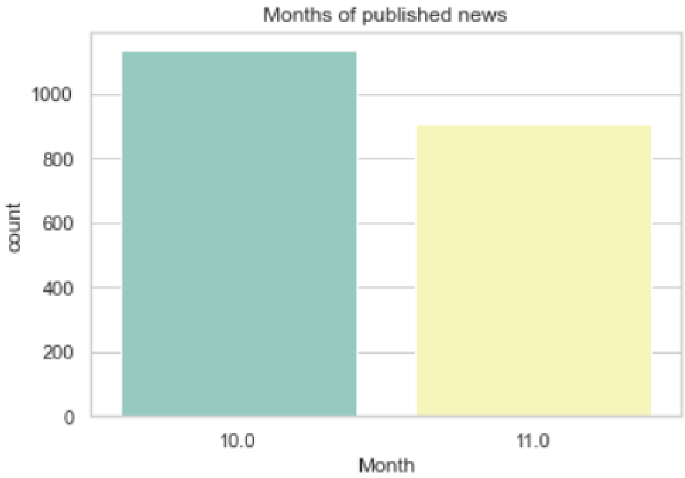
***Feature*: site_url**

We can see from the count plot of the top 10 site urls that all of them have an equal count value of 100.



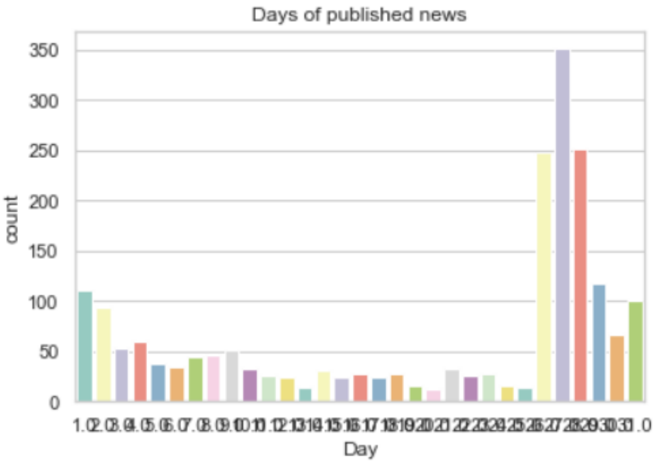
***Feature*: month**

By just viewing the data, we can see that there are only 2 values, 10 and 11. This means that the news data we have is just from the month of October and November.



***Feature*: day**

From the count plot, we can see that the majority of the days are on the 26, 27, and 28.



Findings from EDA

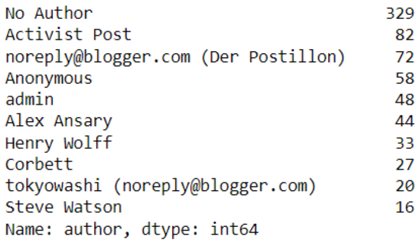
1) The dataset is not biased because we have both real and fake data.
2) Because the majority of news data are written in English, we will employ NLP and other approaches designed with the English language in mind.
3) The type of news data is beneficial for classification. Furthermore, bs stands for Bullshit; all items with no type are categorized as bs.
4) The majority of news stories include an image.
5) The most frequently occurring site addresses and author names have been gathered. We can take advantage of the occurrence of similar sites.
6) We have news data just from the month of October and November.
7) Toward the end of each month, we have a lot of articles.

Exploring the Necessary Features:

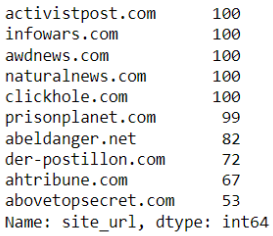
We will be using author, title_without_stopwords, text_without_stopwords, and site_urls features to train our model. Let us explore these features in more detail.

Let us look at Fake Author's data and Fake URL site data.

Fake,

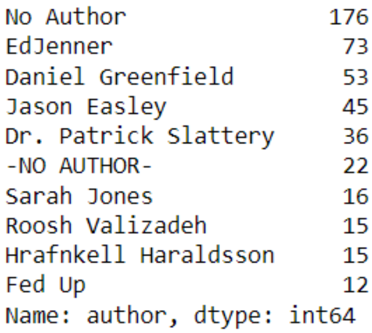


```
No Author                                  329
Activist Post                               82
noreply@blogger.com (Der Postillon)         72
Anonymous                                   58
admin                                       48
Alex Ansary                                 44
Henry Wolff                                 33
Corbett                                     27
tokyowashi (noreply@blogger.com)            20
Steve Watson                                16
Name: author, dtype: int64
```

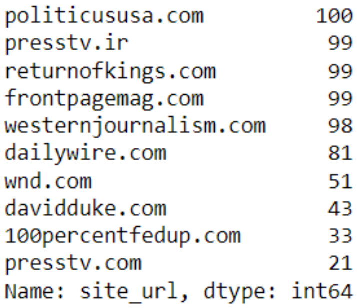


```
activistpost.com      100
infowars.com          100
awdnews.com           100
naturalnews.com       100
clickhole.com         100
prisonplanet.com       99
abeldanger.net         82
der-postillon.com      72
ahtribune.com          67
abovetopsecret.com     53
Name: site_url, dtype: int64
```

Real,

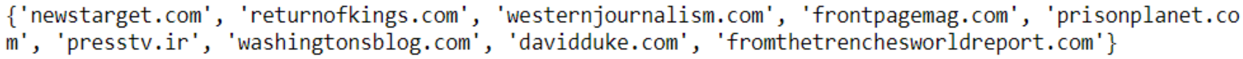


```
No Author                176
EdJenner                  73
Daniel Greenfield         53
Jason Easley              45
Dr. Patrick Slattery      36
-NO AUTHOR-               22
Sarah Jones               16
Roosh Valizadeh           15
Hrafnkell Haraldsson      15
Fed Up                    12
Name: author, dtype: int64
```



```
politicususa.com         100
presstv.ir                99
returnofkings.com         99
frontpagemag.com          99
westernjournalism.com     98
dailywire.com             81
wnd.com                   51
davidduke.com             43
100percentfedup.com       33
presstv.com               21
Name: site_url, dtype: int64
```

Finding the intersection of Real and Fake,



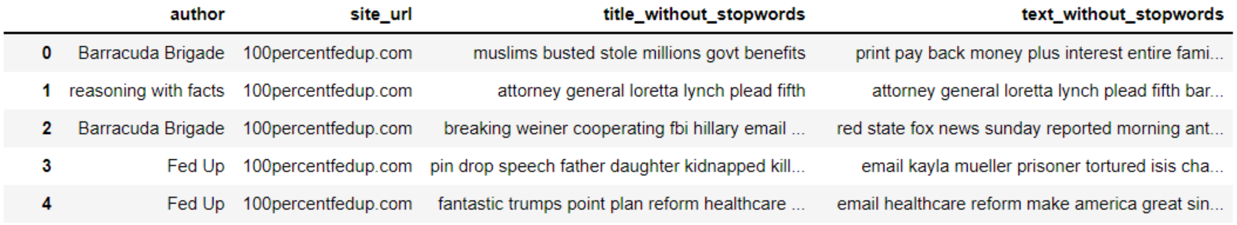
```
{'newstarget.com', 'returnofkings.com', 'westernjournalism.com', 'frontpagemag.com', 'prisonplanet.co
m', 'presstv.ir', 'washingtonsblog.com', 'davidduke.com', 'fromthetrenchesworldreport.com'}
```

The Authors and Sites for real and fake news are distinct here. That is, the same source is not delivering both true and false news. Some sites provide both types of news, but they predominantly produce one kind. So, in addition to the main text, site URLs and author names may aid in prediction.

Feature Engineering:

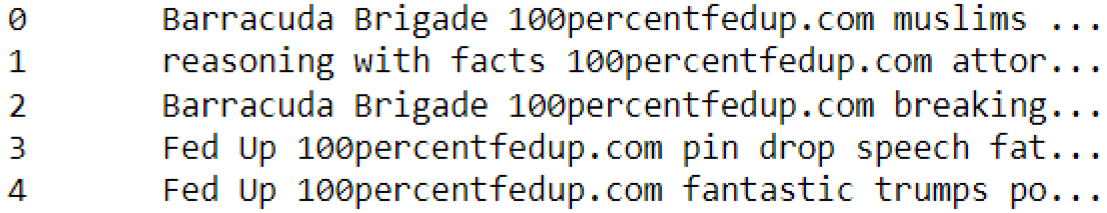
Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, to **simplify and speed up data transformations** while also **enhancing model accuracy**. Feature engineering is required when working with machine learning models. Regardless of the data or architecture, a terrible feature will have a direct impact on your model. [5]

We set the target variable and the predictor variables. The target variable is just the feature ‘Real’ and the predictor variables will be 'author', 'site_url', 'title_without_stopwords', and 'text_without_stopwords'.



	author	site_url	title_without_stopwords	text_without_stopwords
0	Barracuda Brigade	100percentfedup.com	muslims busted stole millions govt benefits	print pay back money plus interest entire fami...
1	reasoning with facts	100percentfedup.com	attorney general loretta lynch plead fifth	attorney general loretta lynch plead fifth bar...
2	Barracuda Brigade	100percentfedup.com	breaking weiner cooperating fbi hillary email ...	red state fox news sunday reported morning ant...
3	Fed Up	100percentfedup.com	pin drop speech father daughter kidnapped kill...	email kayla mueller prisoner tortured isis cha...
4	Fed Up	100percentfedup.com	fantastic trumps point plan reform healthcare ...	email healthcare reform make america great sin...

Now since all the predictor variables are text data, we shall join all the variables together into a single predictor variable. This will help us use text-mining techniques.



```
0        Barracuda Brigade 100percentfedup.com muslims ...
1        reasoning with facts 100percentfedup.com attor...
2        Barracuda Brigade 100percentfedup.com breaking...
3        Fed Up 100percentfedup.com pin drop speech fat...
4        Fed Up 100percentfedup.com fantastic trumps po...
```

Now, let us explore the similarities and dissimilarities between real and fake data.

TDIF Vectorizer

TF-IDF stands for *term frequency-inverse document frequency* and it is a measure, used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document amongst a collection of documents (also known as a corpus). [6]

Term Frequency: This summarizes how frequently a specific word appears in a document.
Inverse Document Frequency: This downscales words that appear often across documents.
TF-IDF: Gives word frequency scores that attempt to highlight more interesting words.

Forming Two Clusters using K-Means

K-means clustering is an unsupervised learning technique to classify unlabeled data by grouping them by features, rather than pre-defined categories. The variable K represents the number of groups or categories created. The goal is to split the data into K different clusters and report the location of the center of mass for each cluster. Then, a new data point can be assigned a cluster (class) based on the closed center of mass. [7]

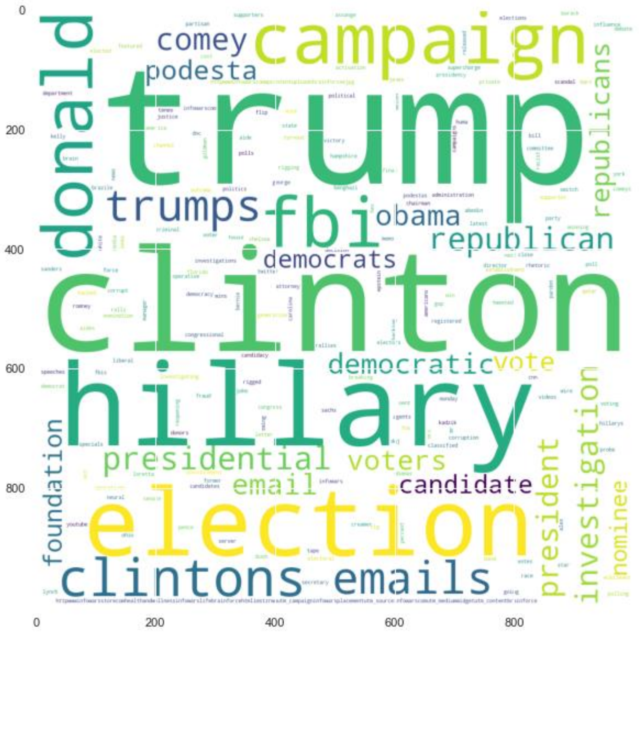
We use K-means to define two clusters from our data and the model converges at iteration 13. We then grouped the two clusters as Fake and True.

Natural Language Toolkit (NLTK)

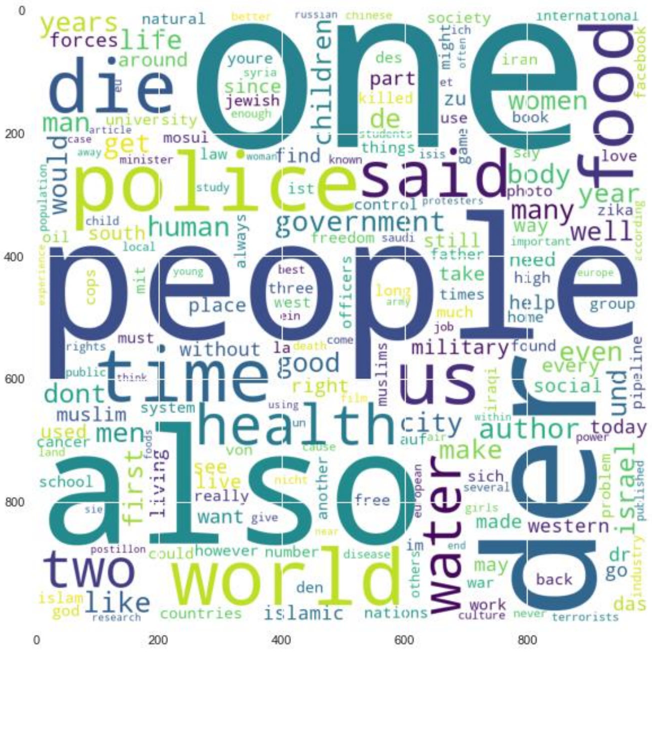
NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries. [8]

We use NLTK libraries to find keywords that appear in both clusters while also removing stop words since they usually have a high count. Using the above technique, we can generate the unique works that make up the Fake data and the True data.

Word cloud of **Fake Data**,



Word cloud of **True Data**,



Model Implementation and Analysis

To start with the model training, we must first split the data into training and testing. For this, we have used the 80-20 train-test split. Based on the selected features, let's apply 3 different Machine Learning Classification algorithms: K-Neighbors Classification, Random Forest Classification, and AdaBoost Classification, to the training data to check the model's accuracy. The table of accuracy is given below:

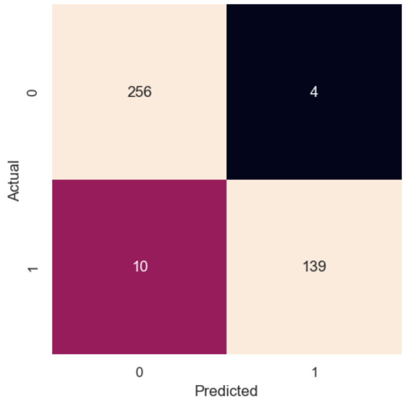
Method	**Accuracy**
K-Neighbors	73.11%
Random Forest	91.69%
AdaBoost	96.58%

Since the best model we get is AdaBoost, let's dive deeper to understand the model better.

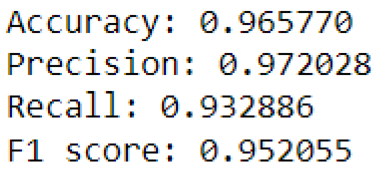
AdaBoost Classification

The AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as the variance for supervised learning. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boosting with a slight difference. [2]

The confusion matrix of the AdaBoost model is given below,



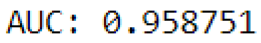
The summary of the AdaBoost Classification model is given below



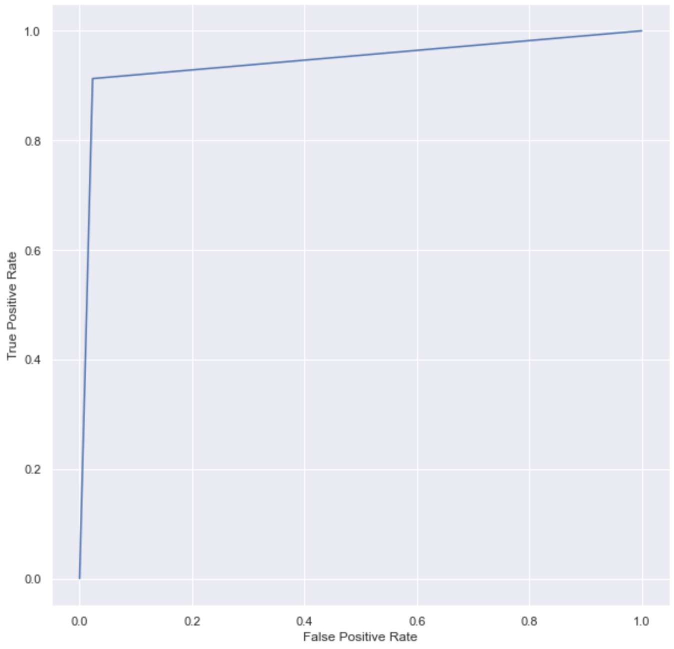
```
Accuracy: 0.965770
Precision: 0.972028
Recall: 0.932886
F1 score: 0.952055
```

We have got a classification rate of 96.57% and which is considered an excellent accuracy. Precision is about being precise, i.e., how accurate your model is. In our prediction case, the AdaBoost Classification model predicted that Fake & Real News 97.20% of the time. The recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. Our model has a Recall of 93.28%. The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems. In our case, the model predicted a score of 95.2%.

The ROC curve is given below



```
AUC: 0.958751
```



The receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity. An area Under Cover (AUC) score of 1 represents a perfect classifier, and 0.5 represents a worthless classifier. The AUC score for the model is 0.958. Which tells us that our classifier is almost perfect.

Conclusion

We discovered some interesting patterns and trends by analyzing Fake News with different important features. We were subsequently able to find a subset of the original features that are sufficient to explain our data using machine learning techniques. We used TDIF Vectorizer and NLTK libraries to further broaden our understanding of the dataset and prepare appropriate data to provide to the model for training. We next trained various classification algorithms and found that the AdaBoost Classification model was the best fit. On a test set that we kept separate during development, this model had an accuracy of 96.57% and an AUC score of 0.95.

Future Scope

Since political news is present in the majority of our training records, this model may be used best for political news. With a bigger and more diversified dataset, the same approach may be applied to all news categories.

References

[1] https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis#steps_involved_in_exploratory_data_analysis

[2] https://www.mygreatlearning.com/blog/adaboost-algorithm/

[3] https://rb.gy/mtewwr

[4] https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[5] https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

[6] https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/

[7] https://deepai.org/machine-learning-glossary-and-terms/k-means

[8] https://www.nltk.org/

GitHub Repository Link -
https://github.com/psurendran98/MachineLearning-Fall2022.git