Analysis Report

Arnav Gupta, 200968030, Batch 3

Problem statement – Fake News Detection

Often people perceive whatever conveyed in the news to be true. There were circumstances where even the news channels acknowledged that their news is not true as they wrote. It is important to identify the fake news from the real true news. Fake news denotes a type of yellow press which intentionally presents misinformation or hoaxes spreading through both traditional print news media and recent online social media. With deceptive words, online social network users can get infected by this online fake news easily, which has brought about tremendous effects on the offline society already. Headlines tend to be more focused on attracting the reader's attention and going viral because of this, despite the lack of veracity within the information in the body text, thus leading to misinformation through false facts. Headlines are misleading and incongruent.

What is the meaning of "Fake News"

News which are fake; fabricated stories/news
Intentionally designed to mislead the readers (clickbait)

Objective

To classify news from dataset as being real or fake.
To train and predict, from our model, particular piece of news.
To apply and deploy baseline models, sequential models and deep learning models like LSTM to achieve a high accuracy.

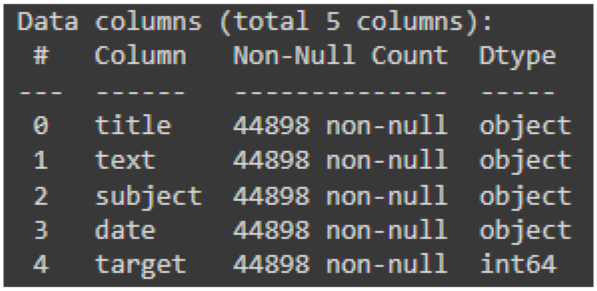
Dataset Analysis

Metadata

Two datasets: - Fake.csv, True.csv
Fake and real news data are given in two separate data sets, with each data set consisting of approximately 20000 articles.

Description of columns in the file:

- title – signifies news headlines
- text- signifies news content/article
- subject- signifies type of news
- date- signifies date the news was published

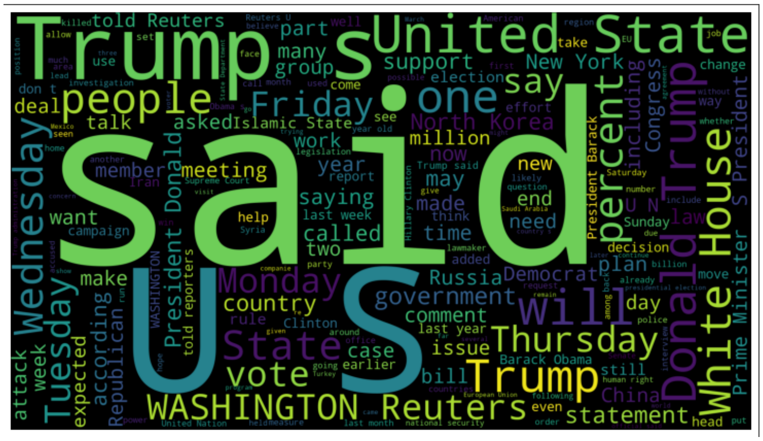


```
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    44898 non-null  object
 1   text     44898 non-null  object
 2   subject  44898 non-null  object
 3   date     44898 non-null  object
 4   target   44898 non-null  int64
```

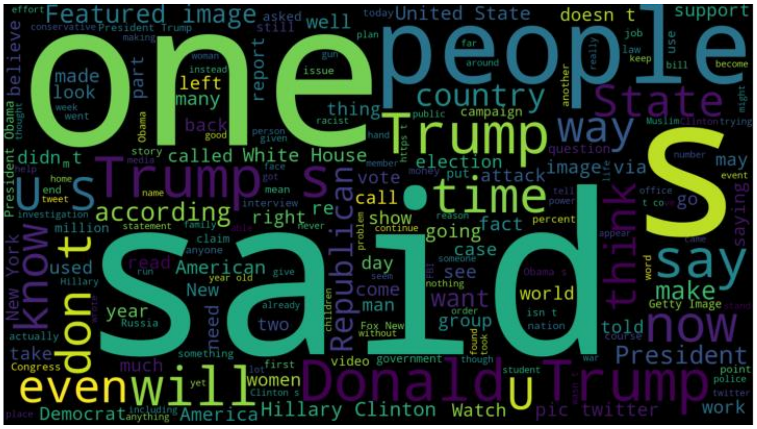
Here we have concatenated the datasets, that is why it is showing 40k values.

Exploratory Data Analysis

Most frequent words occurring in True.csv



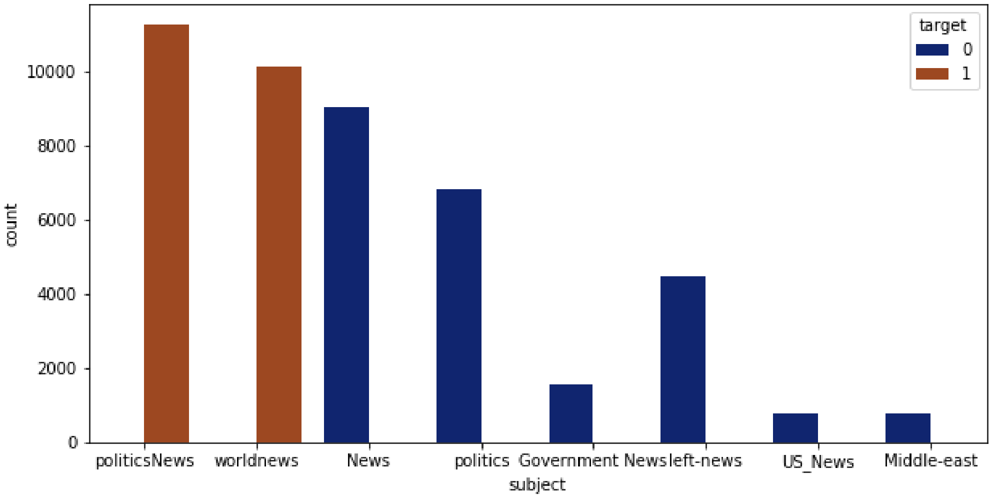
Most frequent words occurring in Fake.csv



Some texts are tweets from Twitter
Real news has source of publication which fake news doesn't have
"said", "Donald Trump", "US" are commonly used

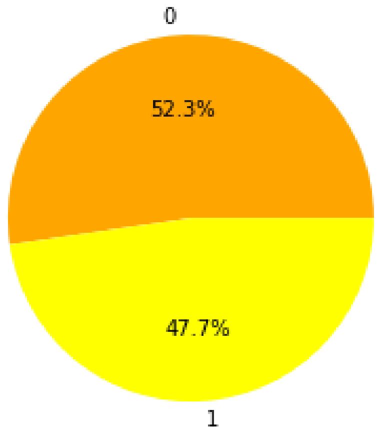
Types of news in dataset (subject) with values: -

- politicsNews 11272
- worldnews 10145
- News 9050
- politics 6841
- left-news 4459
- Government News 1570
- US_News 783
- Middle-east 778



Number of fake and real news data

Fake News – 23481 (52.3%)
Real News – 21417 (47.7%)



Preprocessing

Text processing is essential for building an efficient model in which I will make text lowercase, remove punctuations and hyperlinks, remove unnecessary brackets and numbers. All the above, doesn't affect our model in any way so it is better to remove them. This will be achieved via Natural Language Processing techniques.
Remove stopwords like - "the", "a", "an", "in". We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words.
NLTK (Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages

Stemming - process of reducing infected words to their stem, e.g. 'worked' becomes 'work'.
Lemmatization - the drawbacks of stemming. In stemming, for some words, it may not give may not give meaningful representation. Here, lemmatization comes into picture as it gives meaningful word.
Tokenization - breaks the text into small chunks for easing the computation.
Performing Unigram, Bigram, Trigram Analysis on the dataset.