University of London

BSc in Computer Science Final Project

De-Fake My News

By: Andres Salvador

Dataset

My data set of choice is the 'Fake and real news' dataset available from the Kaggle official website (downloaded Feb 24, 2024): https://www.kaggle.com/datasets/bhavikjikadara/fake-news-detection

This dataset has 2 csv files, True.csv and False.csv

Each csv file has 4 columns:

1. title: The title of the article
2. text: The text of the article
3. subject: The subject of the article
4. date: The date that this article was posted at

This dataset was chosen for the clarity and contents of the files.

Categorical Attributes

```
subject: News, politics, left-news, Government News, US_News, Middle-east
```

Continuous Attributes

```
date: continuous
```

Initial Data Manipulation

Library imports and file manipulation

```
# Standard library imports
import os
import re
import shutil
import string
import warnings
```

```
# Third-party imports
from google.colab import drive
import matplotlib.pyplot as plt
import nltk
from nltk.sentiment import SentimentIntensityAnalyzer
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import numpy as np
import pandas as pd
from PIL import Image
import seaborn as sns
from sklearn.metrics import accuracy_score, confusion_matrix,
f1_score, precision_score, recall_score
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
import tensorflow as tf
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, Conv1D,
GlobalMaxPooling1D, Dense, MaxPooling1D, Bidirectional, LSTM
from sklearn.preprocessing import LabelEncoder
from wordcloud import STOPWORDS, ImageColorGenerator, WordCloud

# NLTK downloads
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('vader_lexicon')

# Warnings configuration
warnings.filterwarnings("ignore")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
```

```
# folder manipulation
dir = '/content/drive/MyDrive/NLP/extracts'

if os.path.exists(dir):
    shutil.rmtree(dir)

!mkdir dir

# Unzipping the files, set to override for easy re-runs of the entire
notebook
!unzip -o '/content/drive/MyDrive/NLP/Fake.csv.zip' -d
```

```
'/content/drive/MyDrive/NLP/extracts'
!unzip -o '/content/drive/MyDrive/NLP/True.csv.zip' -d
'/content/drive/MyDrive/NLP/extracts'

Archive:  /content/drive/MyDrive/NLP/Fake.csv.zip
  inflating: /content/drive/MyDrive/NLP/extracts/Fake.csv
Archive:  /content/drive/MyDrive/NLP/True.csv.zip
  inflating: /content/drive/MyDrive/NLP/extracts/True.csv

# making sure the files are there
for dirname, _, filenames in
os.walk('/content/drive/MyDrive/NLP/extracts'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

/content/drive/MyDrive/NLP/extracts/Fake.csv
/content/drive/MyDrive/NLP/extracts/True.csv

# assigning the files to datasets
fake=pd.read_csv('/content/drive/MyDrive/NLP/extracts/Fake.csv')
real=pd.read_csv('/content/drive/MyDrive/NLP/extracts/True.csv')
```

Quick look into the data

```
# rows / columns
fake.shape

(23481, 4)

# rows / columns
real.shape

(21417, 4)

# quick look
fake.head()

{"summary":"{\n  \"name\": \"fake\",\n  \"rows\": 23481,\n
\"fields\": [\n    {\n      \"column\": \"title\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 17903,\n        \"samples\": [\n          \"
Fox News Mocked Into Oblivion After This F*cking STUPID Attempt To
Make Steve Bannon Look Sane (TWEETS)\",\n          \"BREAKING: FL GOV
RICK SCOTT Calls for FBI Director to Resign\",\n          \" WATCH:
Mike Pence\\u2019s Photo Op With Puerto Rico Survivors Just Went
TERRIBLY Wrong (VIDEO)\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"text\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 17455,\n
\"samples\": [\n          \"The moral decay continues The Kapiolani
Medical Center for Women and Children at the University of Hawaii is
```

currently recruiting pregnant girls and women to participate in second-trimester abortions to measure their bleeding during the operation, with and without antihemorrhagic drugs. According to the Clinical Trials website, run by the National Institutes of Health, participants must be at least 14 years old and 18-24 weeks pregnant.The controversial study, led by Bliss Kaneshiro, MD and Kate Whitehouse, DO, will monitor bleeding during D&E abortions to determine the effects of the drug oxytocin, commonly used to minimize blood loss and decrease the risk of hemorrhage.The clinical trial, called Effects of Oxytocin on Bleeding Outcomes during Dilation and Evacuation began in October 2014 and is a collaboration between UH, Society of Family Planning and the University of Washington.The Society of Family Planning funds a number of similar research projects, such as experimenting with the dosage of Misoprostol, a uterine contracting agent, prior to surgical abortions at 13-18 weeks and exploring umbilical cord injections to produce fetal death prior to late-term abortions.In the UH study, researchers will carry out a randomized, double-blinded, placebo-controlled trials, to determine the effect of oxytocin s use on uterine bleeding, meaning that they will either provide or deny intravenous oxytocin to the women.Reports suggest that some doctors are concerned that withholding oxytocin during surgery may put patients, especially teen girls, at risk. This study is reminiscent of Nazi concentration camp experiments. I pity the poor women who are being treated like lab rats, especially those who are denied the drug to reduce hemorrhaging, said Troy Newman, President of Operation Rescue.Dilation and evacuation abortions are surgical procedures that involve dismembering the pre-born baby with forceps, scraping the inside of the uterus with a curette to remove any residuals and finally suctioning out the womb to make sure the contents are completely removed.After the abortion, the corpse of the fetus is reassembled and examined to ensure everything was successfully removed and that the abortion was complete.The study is hoping to attract up to 166 test subjects and is expected to conclude in July 2015.Via: Breitbart News\",\n   \"CNN was quick to scoop up Corey Lewandowski after Donald Trump kicked him out of his role as campaign manager, but his first week on the job is going pretty much exactly how you would expect it to go   terribly.Not only has Lewandowski proven himself to be pretty much like a paid spokesman for Trump, but his defense of the disgraced GOP candidate isn t being received well. Earlier this week, Lewandowski revealed that he was under contract and couldn t criticize The Donald, even after being fired from the campaign. Today, Lewandowski got called out by Hillary Clinton surrogate Christine Quinn for hyping Trump up to be an expert on the Brexit decision   a suggestion that was clearly false.On Monday s edition of CNN s New Day, Lewandowski made another pathetic defense of Trump by trying to reframe the candidate s disgusting reaction to Brexit, where he mostly spoke about how much the decision would be good for his Scotland golf resort. Lewandowski s defense was: Obviously the U.S. dollar has become much stronger now against the

British pound. If you re going to spend money in Europe, now would actually be a good time to go with the fall of the pound.What you have is a world view, so what you have is someone who is saying, Let s look at this from the U.S. perspective. If you want to go and travel overseas   just from a monetary perspective   now is the right time to do that because what you re getting is more for your dollar. Quinn wasn t having it. She ripped into Lewandowski, firing back, Donald Trump is not running to be travel agent of the world, he s running to be president of the United States. She continued: What he said wasn t a commentary on international markets, it was, When the pound goes down, more people will come to my golf course. Donald Trump s main concern isn t the international markets, it isn t the impact that Brexit will have on hard working Americans 401Ks, it s himself. How can he make more money, how can he put more money in his bank account? Lewandowski compared the Brexit decision to Trump s rise in the GOP, and Quinn once again called him out and put him back in his place. She said: Trump touted that he saw this coming. That s ridiculous because when he was first asked about Brexit by the press, he didn t appear to know what it was. Lewandowski tried to counter by insisting that People are too smart, they are tired of being told what to do. He then tried to commend Trump for being a selfish moron: You know what Donald Trump said about Brexit? What he said was, you don t have to listen to me because it s not my decision. He didn t weigh in like Hillary Clinton did, like Barack Obama did, saying that you can t do this. Quinn fought back, Because he didn t know what it was. Lewandowski was fighting a losing battle. Trump s reaction to Brexit was just as terrifying as it was humorous   it truly proved that Trump knows nothing about foreign affairs, and hasn t spent any time educating himself since the beginning of his presidential candidacy. If only some of the hours he spent getting into fights on Twitter were being used for learning about how the world works. But instead, he once again exposed himself as an unfit choice for President. And when people like Lewandowski try to make sense of his idiocy, they only make themselves look equally foolish.You can watch the embarrassing video below:Featured image via screen capture\",\n   \"A Michigan woman decided to defend against tyranny? when she and another shopper couldn t agree over who got to buy the last notebook on the shelf at the Novi Towne Center store.According to ABC 13, the brawl yes, brawl   involved two Farmington Hills residents, ages 46 and 32, and a mother and daughter from South Lyon, ages 51 and 20. In other words, these were all grown adults who should have known better but hey   there was only one notebook on the shelf, and we ve all seen what happens in those post-apocalyptic movies when a store is down to the last gallon of milk, right?Two of the women, one of whom was the unnamed 20-year-old, reached for the notebook at the same time. The 46 and 32-year-olds apparently decided that she wasn t getting their goddamn notebook and began pulling her hair.Then, because this had almost hit peak trailer park, the 20-year-old s mother decided to go for bonus points by pulling out her gun. Fortunately, someone pushed

her aside before she could do any harm.This is one of the NRA s responsible gun owners (conservatives can t dismiss this one, as it is confirmed that she is a concealed carry permit holder)   ready to leap into action at the most minor sign of danger and make things worse by turning the situation potentially deadly.Watch it happen below:Featured image via screengrab\"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"subject\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 6,\n   \"samples\": [\n   \"News\",\n \"politics\",\n   \"Middle-east\"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"date\",\n   \"properties\": {\n \"dtype\": \"category\",\n   \"num_unique_values\": 1681,\n \"samples\": [\n   \"Jun 5, 2015\",\n   \"August 28, 2016\",\n   \"June 3, 2017\"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe","variable_name":"fake"}

```
# quick look
real.head()
```

{"summary":"{\n \"name\": \"real\",\n \"rows\": 21417,\n \"fields\": [\n   {\n   \"column\": \"title\",\n \"properties\": {\n   \"dtype\": \"string\",\n \"num_unique_values\": 20826,\n   \"samples\": [\n \"German, Turkish foreign ministers meet after detainee released\",\n \"Kremlin calls North Korea's latest missile launch another 'provocation'\",\n   \"Transgender soldiers, veterans shaken by Trump's ban on their service\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n \"column\": \"text\",\n   \"properties\": {\n   \"dtype\": \"string\",\n   \"num_unique_values\": 21192,\n \"samples\": [\n   \"WASHINGTON (Reuters) - A majority of the U.S. Senate on Tuesday backed a new round of disaster aid to help Puerto Rico and several states recover from damage from hurricanes and wildfires. The legislation would provide $36.5 billion in emergency relief as Puerto Rico in particular struggles to regain electricity and other basic services following destructive hurricanes. The House of Representatives approved the bill earlier this month. The Trump administration already has indicated it will seek another round of emergency relief from Congress. \",\n   \"BUDAPEST (Reuters) - About a thousand Hungarians protested on Friday against a crackdown on the main opposition party Jobbik which has been threatened by a record political campaign fine that the party leader describes as a death sentence for democracy. Despite the gloomy rhetoric and Jobbik saying it was fighting for survival, support for the demonstration was well down on other similar rallies over the past year. Hungarians will vote for a new parliament in April and Prime Minister Viktor Orban s conservative, anti-migrant Fidesz party is far ahead in the polls,

with Jobbik its nearest rival. Jobbik, once on the far right, has turned toward the center in a bid to attract more support and is now campaigning nationwide against Orban, depicting him as the leader of a criminal gang. Orban, rejecting the charges, says his financial standing is an open book . Last week the state audit office (ASZ) ruled Jobbik had bought political posters far below market prices, breaching rules on political funding, then it slapped a 663 million forint ($2.5 million) penalty on the party. The protesters, waving Jobbik flags and posters deriding the ruling elite, gathered outside the headquarters of Orban s Fidesz party. What we see unfolding is not an audit office investigation. It is not an official penalty. This is a death sentence with Jobbik s name on it. But in reality, it is a death sentence for Hungarian democracy, Jobbiik leader Gabor Vona told the crowd. A government spokesman could not comment immediately on his remarks. ASZ chairman Laszlo Domokos is a former Fidesz lawmaker, whom Jobbik and other critics accuse of making decisions in favor of Orban. The audit office denies that. On Friday, ASZ again called on Jobbik to submit information that would challenge its findings, saying it acted fully within its rights throughout the probe. The ruling Fidesz party and the government have denied any involvement in the ASZ probe. This case has nothing to do with the election campaign, Orban aide Janos Lazar said on Thursday. For over a year Fidesz has targeted Jobbik, whose move to the center could upend the longstanding status quo of a dominant Fidesz with weaker opponents to its left and its right, said analyst Zoltan Novak at the Centre for Fair Political Analysis. Gyorgy Illes, a 67-year-old pensioner attending the rally, said he used to be a Socialist supporter but got disillusioned as the party struggled to overcome its internal divisions. This ASZ probe is a clear sign that Orban is way past any remedy. It is a ruthless attack on everything we hold dear. Democracy, the rule of law, equality, you name it, he said. \",\n \"BEIJING/TAIPEI (Reuters) - China accused the United States on Thursday of interfering in its internal affairs and said it had lodged a complaint after U.S. President Donald Trump signed into law an act laying the groundwork for possible U.S. navy visits to self-ruled Taiwan. Tensions have risen in recent days after a senior Chinese diplomat threatened China would invade Taiwan if any U.S. warships made port visits to the island which China claims as its own territory. On Monday, Chinese jets carried out island encirclement patrols around Taiwan, with state media showing pictures of bombers with cruise missiles slung under their wings as they carried out the exercise. On Tuesday, Trump signed into law the National Defense Authorization Act for the 2018 fiscal year, which authorizes the possibility of mutual visits by navy vessels between Taiwan and the United States. Such visits would be the first since the United States ended formal diplomatic relations with Taiwan in 1979 and established ties with Beijing. Chinese Foreign Ministry spokesman Lu Kang said while the Taiwan sections of the law were not legally binding, they seriously violate the One China policy and constitute an

interference in China s internal affairs . China is resolutely opposed to this, and we have already lodged stern representations with the U.S. government, Lu told a daily news briefing. China is firmly opposed to any official exchanges, military contact, or arms sales between Taiwan and the United States, he added. Proudly democratic Taiwan has become increasingly concerned with the ramped up Chinese military presence, that has included several rounds of Chinese air force drills around the island in recent months. Taiwan is confident of its defenses and responded quickly to the Chinese air force drills this week, its government said, denouncing the rise in China s military deployments as irresponsible. Taiwan presidential spokesman Alex Huang, speaking to Taiwan media in comments reported late on Wednesday, said the defense ministry had kept a close watch on the patrols and responded immediately and properly. Taiwan can ensure there are no concerns at all about national security, and people can rest assured , Huang said. Both sides of the narrow Taiwan Strait, which separates Taiwan from its giant neighbor, have a responsibility to protect peace and stability, he added. Such a raised military posture that may impact upon and harm regional peace and stability and cross-strait ties does not give a feeling of responsibility, and the international community does not look favorably upon this, Huang was quoted as saying. Relations have soured considerably since Tsai Ing-wen, who leads Taiwan s independence-leaning Democratic Progressive Party, won presidential elections last year. China suspects Tsai wants to declare the island s formal independence, a red line for Beijing. Tsai says she wants to maintain peace with China but will defend Taiwan s security. Taiwan is well equipped with mostly U.S. weapons but has been pressing for more advanced equipment to deal with what it sees as a rising threat from China. The United States is bound by law to provide the island with the means to defend itself. China has never renounced the use of force to bring Taiwan under its control. \"\n],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n }\n   },\n   {\n   \"column\": \"subject\",\n \"properties\": {\n   \"dtype\": \"category\",\n \"num_unique_values\": 2,\n   \"samples\": [\n \"worldnews\",\n   \"politicsNews\"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\ n   },\n   {\n   \"column\": \"date\",\n   \"properties\": {\n \"dtype\": \"object\",\n   \"num_unique_values\": 716,\n \"samples\": [\n   \"September 2, 2017 \",\n \"February 2, 2017 \"\n   ],\n   \"semantic_type\": \"\",\n \"description\": \"\"\n   }\n   }\n]\ n}","type":"dataframe","variable_name":"real"}

```
# Now, the fake news has this subjects and their count is
fake["subject"].value_counts()
```

```
News                9050
politics            6841
left-news           4459
```

```
Government News    1570
US_News             783
Middle-east         778
Name: subject, dtype: int64
```

```
# Now, the real news has this subjects and their count is
real["subject"].value_counts()
```

```
politicsNews    11272
worldnews       10145
Name: subject, dtype: int64
```

```
# How many different dates are there in the fake dataset
fake["date"].value_counts()
```

```
May 10, 2017
46
May 6, 2016
44
May 5, 2016
44
May 26, 2016
44
May 11, 2016
43


..
https://100percentfedup.com/video-hillary-asked-about-trump-i-just-
want-to-eat-some-pie/
1
November 19, 2017
1
November 20, 2017
1
https://100percentfedup.com/12-yr-old-black-conservative-whose-video-
to-obama-went-viral-do-you-really-love-america-receives-death-threats-
from-left/     1
December 4, 2017
1
Name: date, Length: 1681, dtype: int64
```

```
# How many different dates are there in the real dataset
real["date"].value_counts()
```

```
December 20, 2017      182
December 6, 2017       166
November 30, 2017      162
November 9, 2017       158
October 13, 2017       155
                      ...
September 11, 2016       1
```

```
May 28, 2016              1
May 30, 2016              1
December 30, 2017         1
January 24, 2016          1
Name: date, Length: 716, dtype: int64
```

Setting up Fake news

```
# Lets add a label to the fake datasets, with 0 (as in how much we
trust that row)
fake["label"]=0
```

Do we need to clean the dataset?

```
# Of the fake dataset, how many rows have null values?
fake.isnull().sum()
```

```
title      0
text       0
subject    0
date       0
label      0
dtype: int64
```

No need to clean the fake data! But if needed to clean this dataset, we would have eliminated the empty rows, since this is text we cant replace the contents with mid/median values

```
# Another quick look into the fake news, now with the label=0
fake.head()
```


```
{"summary":"{\n  \"name\": \"fake\",\n  \"rows\": 23481,\n
\"fields\": [\n    {\n      \"column\": \"title\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 17903,\n        \"samples\": [\n          \"
Fox News Mocked Into Oblivion After This F*cking STUPID Attempt To
Make Steve Bannon Look Sane (TWEETS)\",\n          \"BREAKING: FL GOV
RICK SCOTT Calls for FBI Director to Resign\",\n          \" WATCH:
Mike Pence\\u2019s Photo Op With Puerto Rico Survivors Just Went
TERRIBLY Wrong (VIDEO)\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"text\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 17455,\n
\"samples\": [\n          \"The moral decay continues The Kapiolani
Medical Center for Women and Children at the University of Hawaii is
currently recruiting pregnant girls and women to participate in
second-trimester abortions to measure their bleeding during the
operation, with and without antihemorrhagic drugs. According to the
Clinical Trials website, run by the National Institutes of Health,
participants must be at least 14 years old and 18-24 weeks
```

pregnant.The controversial study, led by Bliss Kaneshiro, MD and Kate Whitehouse, DO, will monitor bleeding during D&E abortions to determine the effects of the drug oxytocin, commonly used to minimize blood loss and decrease the risk of hemorrhage.The clinical trial, called Effects of Oxytocin on Bleeding Outcomes during Dilation and Evacuation began in October 2014 and is a collaboration between UH, Society of Family Planning and the University of Washington.The Society of Family Planning funds a number of similar research projects, such as experimenting with the dosage of Misoprostol, a uterine contracting agent, prior to surgical abortions at 13-18 weeks and exploring umbilical cord injections to produce fetal death prior to late-term abortions.In the UH study, researchers will carry out a randomized, double-blinded, placebo-controlled trials, to determine the effect of oxytocin s use on uterine bleeding, meaning that they will either provide or deny intravenous oxytocin to the women.Reports suggest that some doctors are concerned that withholding oxytocin during surgery may put patients, especially teen girls, at risk. This study is reminiscent of Nazi concentration camp experiments. I pity the poor women who are being treated like lab rats, especially those who are denied the drug to reduce hemorrhaging, said Troy Newman, President of Operation Rescue.Dilation and evacuation abortions are surgical procedures that involve dismembering the pre-born baby with forceps, scraping the inside of the uterus with a curette to remove any residuals and finally suctioning out the womb to make sure the contents are completely removed.After the abortion, the corpse of the fetus is reassembled and examined to ensure everything was successfully removed and that the abortion was complete.The study is hoping to attract up to 166 test subjects and is expected to conclude in July 2015.Via: Breitbart News\",\n   \"CNN was quick to scoop up Corey Lewandowski after Donald Trump kicked him out of his role as campaign manager, but his first week on the job is going pretty much exactly how you would expect it to go   terribly.Not only has Lewandowski proven himself to be pretty much like a paid spokesman for Trump, but his defense of the disgraced GOP candidate isn t being received well. Earlier this week, Lewandowski revealed that he was under contract and couldn t criticize The Donald, even after being fired from the campaign. Today, Lewandowski got called out by Hillary Clinton surrogate Christine Quinn for hyping Trump up to be an expert on the Brexit decision   a suggestion that was clearly false.On Monday s edition of CNN s New Day, Lewandowski made another pathetic defense of Trump by trying to reframe the candidate s disgusting reaction to Brexit, where he mostly spoke about how much the decision would be good for his Scotland golf resort. Lewandowski s defense was: Obviously the U.S. dollar has become much stronger now against the British pound. If you re going to spend money in Europe, now would actually be a good time to go with the fall of the pound.What you have is a world view, so what you have is someone who is saying, Let s look at this from the U.S. perspective. If you want to go and travel overseas   just from a monetary perspective   now is the right time to

do that because what you re getting is more for your dollar. Quinn wasn t having it. She ripped into Lewandowski, firing back, Donald Trump is not running to be travel agent of the world, he s running to be president of the United States. She continued: What he said wasn t a commentary on international markets, it was, When the pound goes down, more people will come to my golf course. Donald Trump s main concern isn t the international markets, it isn t the impact that Brexit will have on hard working Americans 401Ks, it s himself. How can he make more money, how can he put more money in his bank account? Lewandowski compared the Brexit decision to Trump s rise in the GOP, and Quinn once again called him out and put him back in his place. She said: Trump touted that he saw this coming. That s ridiculous because when he was first asked about Brexit by the press, he didn t appear to know what it was. Lewandowski tried to counter by insisting that People are too smart, they are tired of being told what to do. He then tried to commend Trump for being a selfish moron: You know what Donald Trump said about Brexit? What he said was, you don t have to listen to me because it s not my decision. He didn t weigh in like Hillary Clinton did, like Barack Obama did, saying that you can t do this. Quinn fought back, Because he didn t know what it was. Lewandowski was fighting a losing battle. Trump s reaction to Brexit was just as terrifying as it was humorous   it truly proved that Trump knows nothing about foreign affairs, and hasn t spent any time educating himself since the beginning of his presidential candidacy. If only some of the hours he spent getting into fights on Twitter were being used for learning about how the world works. But instead, he once again exposed himself as an unfit choice for President. And when people like Lewandowski try to make sense of his idiocy, they only make themselves look equally foolish.You can watch the embarrassing video below:Featured image via screen capture\",\n   \"A Michigan woman decided to defend against tyranny? when she and another shopper couldn t agree over who got to buy the last notebook on the shelf at the Novi Towne Center store.According to ABC 13, the brawl yes, brawl   involved two Farmington Hills residents, ages 46 and 32, and a mother and daughter from South Lyon, ages 51 and 20. In other words, these were all grown adults who should have known better but hey   there was only one notebook on the shelf, and we ve all seen what happens in those post-apocalyptic movies when a store is down to the last gallon of milk, right?Two of the women, one of whom was the unnamed 20-year-old, reached for the notebook at the same time. The 46 and 32-year-olds apparently decided that she wasn t getting their goddamn notebook and began pulling her hair.Then, because this had almost hit peak trailer park, the 20-year-old s mother decided to go for bonus points by pulling out her gun. Fortunately, someone pushed her aside before she could do any harm.This is one of the NRA s responsible gun owners (conservatives can t dismiss this one, as it is confirmed that she is a concealed carry permit holder)   ready to leap into action at the most minor sign of danger and make things worse by turning the situation potentially deadly.Watch it happen

```
below:Featured image via screengrab\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"subject\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
6,\n        \"samples\": [\n          \"News\",\n
\"politics\",\n          \"Middle-east\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"date\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 1681,\n
\"samples\": [\n          \"Jun 5, 2015\",\n          \"August 28,
2016\",\n          \"June 3, 2017\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"label\",\n      \"properties\": {\
n        \"dtype\": \"number\",\n        \"std\": 0,\n        \"min\":
0,\n        \"max\": 0,\n        \"num_unique_values\": 1,\n
\"samples\": [\n          0\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"fake"}
```


Setting up Real News

```
# Now lets work on the real dataset, lets add that label
real["label"]=1
```

Do we need to clean the dataset?

```
# Of the real dataset, how many rows have null values?
real.isnull().sum()
```

```
title      0
text       0
subject    0
date       0
label      0
dtype: int64
```

No need to clean the real data! But if needed to clean this dataset, we would have eliminated the empty rows, since this is text we cant replace the contents with mid/median values

```
# Another quick look into the real news, now with the label=1
real.head()
```


```
{"summary":"{\n  \"name\": \"real\",\n  \"rows\": 21417,\n
\"fields\": [\n    {\n      \"column\": \"title\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 20826,\n        \"samples\": [\n
\"German, Turkish foreign ministers meet after detainee released\",\n
\"Kremlin calls North Korea's latest missile launch another
'provocation'\",\n          \"Transgender soldiers, veterans shaken by
```

Trump's ban on their service\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"text\",\n   \"properties\": {\n   \"dtype\": \"string\",\n   \"num_unique_values\": 21192,\n   \"samples\": [\n   \"WASHINGTON (Reuters) - A majority of the U.S. Senate on Tuesday backed a new round of disaster aid to help Puerto Rico and several states recover from damage from hurricanes and wildfires. The legislation would provide $36.5 billion in emergency relief as Puerto Rico in particular struggles to regain electricity and other basic services following destructive hurricanes. The House of Representatives approved the bill earlier this month. The Trump administration already has indicated it will seek another round of emergency relief from Congress. \",\n   \"BUDAPEST (Reuters) - About a thousand Hungarians protested on Friday against a crackdown on the main opposition party Jobbik which has been threatened by a record political campaign fine that the party leader describes as a death sentence for democracy. Despite the gloomy rhetoric and Jobbik saying it was fighting for survival, support for the demonstration was well down on other similar rallies over the past year. Hungarians will vote for a new parliament in April and Prime Minister Viktor Orban s conservative, anti-migrant Fidesz party is far ahead in the polls, with Jobbik its nearest rival. Jobbik, once on the far right, has turned toward the center in a bid to attract more support and is now campaigning nationwide against Orban, depicting him as the leader of a criminal gang. Orban, rejecting the charges, says his financial standing is an open book . Last week the state audit office (ASZ) ruled Jobbik had bought political posters far below market prices, breaching rules on political funding, then it slapped a 663 million forint ($2.5 million) penalty on the party. The protesters, waving Jobbik flags and posters deriding the ruling elite, gathered outside the headquarters of Orban s Fidesz party. What we see unfolding is not an audit office investigation. It is not an official penalty. This is a death sentence with Jobbik s name on it. But in reality, it is a death sentence for Hungarian democracy, Jobbiik leader Gabor Vona told the crowd. A government spokesman could not comment immediately on his remarks. ASZ chairman Laszlo Domokos is a former Fidesz lawmaker, whom Jobbik and other critics accuse of making decisions in favor of Orban. The audit office denies that. On Friday, ASZ again called on Jobbik to submit information that would challenge its findings, saying it acted fully within its rights throughout the probe. The ruling Fidesz party and the government have denied any involvement in the ASZ probe. This case has nothing to do with the election campaign, Orban aide Janos Lazar said on Thursday. For over a year Fidesz has targeted Jobbik, whose move to the center could upend the longstanding status quo of a dominant Fidesz with weaker opponents to its left and its right, said analyst Zoltan Novak at the Centre for Fair Political Analysis. Gyorgy Illes, a 67-year-old pensioner attending the rally, said he used to be a Socialist supporter but got disillusioned as the party struggled to overcome its

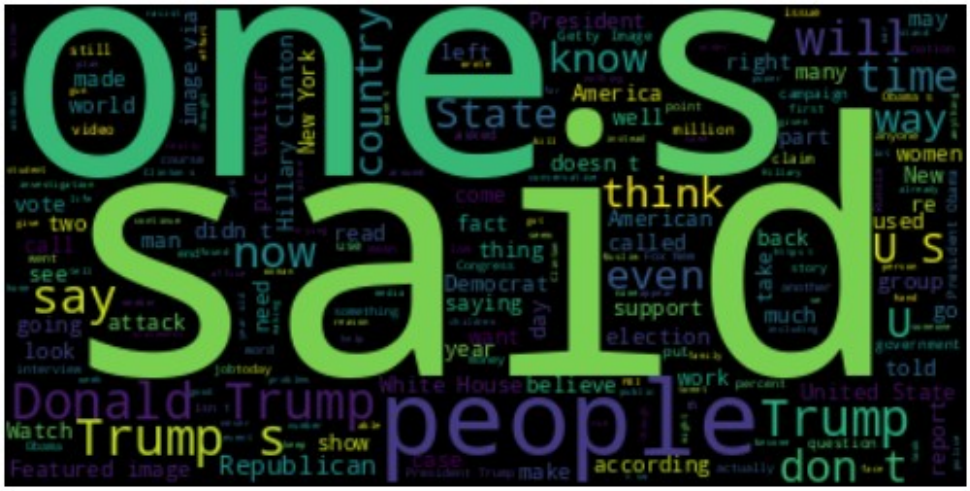
internal divisions. This ASZ probe is a clear sign that Orban is way past any remedy. It is a ruthless attack on everything we hold dear. Democracy, the rule of law, equality, you name it, he said. \",\n \"BEIJING/TAIPEI (Reuters) - China accused the United States on Thursday of interfering in its internal affairs and said it had lodged a complaint after U.S. President Donald Trump signed into law an act laying the groundwork for possible U.S. navy visits to self-ruled Taiwan. Tensions have risen in recent days after a senior Chinese diplomat threatened China would invade Taiwan if any U.S. warships made port visits to the island which China claims as its own territory. On Monday, Chinese jets carried out island encirclement patrols around Taiwan, with state media showing pictures of bombers with cruise missiles slung under their wings as they carried out the exercise. On Tuesday, Trump signed into law the National Defense Authorization Act for the 2018 fiscal year, which authorizes the possibility of mutual visits by navy vessels between Taiwan and the United States. Such visits would be the first since the United States ended formal diplomatic relations with Taiwan in 1979 and established ties with Beijing. Chinese Foreign Ministry spokesman Lu Kang said while the Taiwan sections of the law were not legally binding, they seriously violate the One China policy and constitute an interference in China s internal affairs . China is resolutely opposed to this, and we have already lodged stern representations with the U.S. government, Lu told a daily news briefing. China is firmly opposed to any official exchanges, military contact, or arms sales between Taiwan and the United States, he added. Proudly democratic Taiwan has become increasingly concerned with the ramped up Chinese military presence, that has included several rounds of Chinese air force drills around the island in recent months. Taiwan is confident of its defenses and responded quickly to the Chinese air force drills this week, its government said, denouncing the rise in China s military deployments as irresponsible. Taiwan presidential spokesman Alex Huang, speaking to Taiwan media in comments reported late on Wednesday, said the defense ministry had kept a close watch on the patrols and responded immediately and properly. Taiwan can ensure there are no concerns at all about national security, and people can rest assured , Huang said. Both sides of the narrow Taiwan Strait, which separates Taiwan from its giant neighbor, have a responsibility to protect peace and stability, he added. Such a raised military posture that may impact upon and harm regional peace and stability and cross-strait ties does not give a feeling of responsibility, and the international community does not look favorably upon this, Huang was quoted as saying. Relations have soured considerably since Tsai Ing-wen, who leads Taiwan s independence-leaning Democratic Progressive Party, won presidential elections last year. China suspects Tsai wants to declare the island s formal independence, a red line for Beijing. Tsai says she wants to maintain peace with China but will defend Taiwan s security. Taiwan is well equipped with mostly U.S. weapons but has been pressing for more advanced equipment to deal with what it

```
sees as a rising threat from China. The United States is bound by law
to provide the island with the means to defend itself. China has never
renounced the use of force to bring Taiwan under its control. \"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"subject\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n
\"worldnews\",\n          \"politicsNews\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"date\",\n      \"properties\": {\n
\"dtype\": \"object\",\n        \"num_unique_values\": 716,\n
\"samples\": [\n          \"September 2, 2017 \",\n
\"February 2, 2017 \"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"label\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0,\n        \"min\": 1,\n        \"max\": 1,\n
\"num_unique_values\": 1,\n        \"samples\": [\n          1\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    }\n  ]\n}","type":"dataframe","variable_name":"real"}
```


Individual Word cloud visual representation

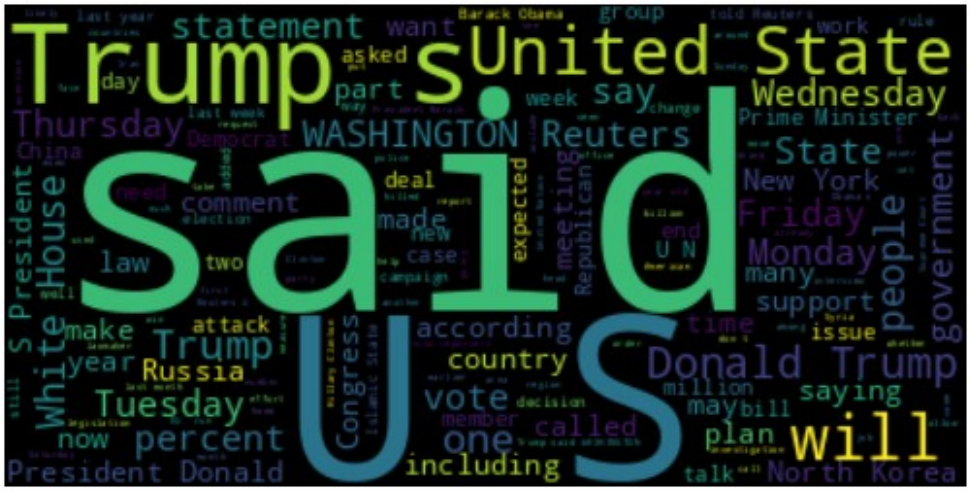
```
# Now lets do it for the whole fake dataset
text_fake_all = " ".join(review for review in fake.text)

# The text might be too long to display, so lets just get the word
count
print("The word count for all the true news text is: ",
len(text_fake_all.split()))

# Lets generate the work cloud
wordcloud = WordCloud().generate(text_fake_all)

# Lets display the word cloud
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
The word count for all the true news text is:  9937110
```




```
# Now lets do it for the whole real dataset
text_real_all = " ".join(review for review in real.text)

# The text might be too long to display, so lets just get the word
count
print("The word count for all the true news text is: ",
len(text_real_all.split()))

# Lets generate the work cloud
wordcloud = WordCloud().generate(text_real_all)

# Lets display the word cloud
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
The word count for all the true news text is:  8259254
```



Unifying the datasets

```
# Now lets join the datasets into 1
data=pd.concat([fake,real],ignore_index=True)
data.head()
```


```
{"summary":"{\n  \"name\": \"data\",\n  \"rows\": 44898,\n
\"fields\": [\n    {\n      \"column\": \"title\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 38729,\n        \"samples\": [\n
\"Supreme Court Justice Ginsburg 'regrets' Trump criticisms\",\n
\" DOZENS Of GOP Foreign Policy Experts Pledge To Stop Trump From
Winning Nomination\",\n          \" REPORT: Trump Laughed After Woman
Was Grabbed \\u2018By The P*ssy\\u2019 On Apprentice Set\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"text\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 38646,\n        \"samples\": [\n          \"
(This September 29 has been corrected to fix date of election in
paragraph 3) NAIROBI (Reuters) - A Kenyan government watchdog said on
Friday it was investigating whether police had assaulted students
during protests this week at the University of Nairobi over the
detention of an opposition lawmaker.   Police fired tear gas on
Thursday at the protesting students. Video footage posted on social
media later in the day showed uniformed officers outside dormitories
and inside classrooms using batons to hit people who did not appear to
be involved in the campus protests.  It was the latest crackdown by
police on protests since an Aug. 8 presidential election that was
later annulled by the Supreme Court. A re-run of the vote has been set
for Oct. 26. The Independent Police Oversight Authority (IPOA)  this
morning noted from social media reports of an incident in which
members of the National Police Service allegedly stormed the
University of Nairobi and assaulted students at the institution,  the
authority said on its Twitter feed.  It requested that any member of
the public come forward to provide information to aid the
investigation.  The students had been protesting against the re-arrest
of a lawmaker, Paul Ongili Owino, on Wednesday, shortly after he was
released on bail on charges of subversion for calling President Uhuru
Kenyatta a  son of a dog  at a campaign rally. Kenya is a key Western
ally in a region often roiled by violence. Preparations for the re-run
of the election are being closely monitored for signs of instability,
after at least 28 people were killed in unrest following the Aug. 8
vote. The IPOA watchdog was created in 2011, after police came under
severe criticism for the number of protesters killed during
demonstrations against disputed elections in 2007. The violence then
killed around 1,2000 people. But local and international rights groups
say the institution is struggling to fulfill its mandate to
investigate allegations of police brutality in a country rife with
reports from civilians of extrajudicial killings by security forces.
The IPOA has secured two convictions of police officers in the four
```

years it has been operational. The watchdog pledged last month to fast-track investigations into high-profile deaths such as that of a baby allegedly killed by police in the violence after the election. [L4N1L34FC] But people familiar with the status of these investigations say the police are not cooperating with them, and that senior officials in the police force and the interior ministry insist officers killed only thieves and thugs. \",\n
\"WASHINGTON/CAIRO (Reuters) - Five Iraqi passengers and one Yemeni were barred from boarding an EgyptAir flight from Cairo to New York on Saturday after President Donald Trump halted the entry of citizens from seven Muslim-majority countries, sources at Cairo airport said. The passengers, arriving in transit to Cairo airport, were stopped and re-directed to flights headed for their home countries despite holding valid visas, the sources said. Trump on Friday put a four-month hold on allowing refugees into the United States and temporarily barred travelers from Syria and six other Muslim-majority countries, saying the moves would help protect Americans from terrorist attacks. He said his most sweeping use of his presidential powers since taking office a week ago, barring travelers from the seven nations for at least 90 days, would give his administration time to develop more stringent screening procedures for refugees, immigrants and visitors. \\u201cI\\u2019m establishing new vetting measures to keep radical Islamic terrorists out of the United States of America. Don\\u2019t want them here,\\u201d Trump said earlier on Friday at the Pentagon. \\u201cWe only want to admit those into our country who will support our country and love deeply our people,\\u201d he said. The bans, though temporary, took effect immediately, causing havoc and confusion for would-be travelers with passports from Iran, Iraq, Libya, Somalia, Sudan, Syria and Yemen. Besides Cairo it was not immediately clear whether other airports of countries listed by Trump had swiftly implemented the ban. Arab officials of the listed countries would not comment on the matter. The order seeks to prioritize refugees fleeing religious persecution, a move Trump separately said was aimed at helping Christians in Syria. That led some legal experts to question whether the order was constitutional. One group said it would announce a court challenge on Monday. The Council on American-Islamic Relations said the order targets Muslims because of their faith, contravening the U.S. Constitutional right to freedom of religion. \\u201cPresident Trump has cloaked what is a discriminatory ban against nationals of Muslim countries under the banner of national security,\\u201d said Greg Chen of the American Immigration Lawyers Association. Trump has long pledged to take this kind of action, making it a prominent feature of his campaign for the Nov. 8 election. But people who work with Muslim immigrants and refugees were scrambling to determine the scope of the order. Even legal permanent residents - people with \\u201cgreen cards\\u201d allowing them to live and work in the United States - were being advised to consult immigration lawyers before traveling outside the country, or trying to return, according to Muslim Advocates, a civil rights group in Washington. On Friday

evening, Abed Ayoub of the American-Arab Anti-Discrimination Committee said he had fielded about 100 queries from people anxious about the order, which he said he believed could affect traveling green card holders, students, people coming to the United States for medical care and others. \\u201cIt\\u2019s chaos,\\u201d Ayoub said. During his campaign, Trump tapped into American fears about Islamic State militants and the flood of migrants into Europe from Syria\\u2019s civil war, saying refugees could be a \\u201cTrojan horse\\u201d that allowed attackers to enter the United States. In December 2015, he called for a ban on all Muslims entering the United States, drawing fire for suggesting a religious test for immigrants that critics said would violate the U.S. Constitution. His idea later evolved into a proposal for \\u201cextreme vetting.\\u201d Trump\\u2019s order also suspends the Syrian refugee program until further notice, and will eventually give priority to minority religious groups fleeing persecution. Trump said in an interview with the Christian Broadcasting Network that the exception would help Syrian Christians fleeing the civil war there. Legal experts were divided on whether this order would be constitutional. \\u201cIf they are thinking about an exception for Christians, in almost any other legal context discriminating in favor of one religion and against another religion could violate the constitution,\\u201d said Stephen Legomsky, a former chief counsel at U.S. Citizenship and Immigration Services in the Obama administration. But Peter Spiro, a professor at Temple University Beasley School of Law, said Trump\\u2019s action would likely be constitutional because the president and Congress are allowed considerable deference when it comes to asylum decisions. \\u201cIt\\u2019s a completely plausible prioritization, to the extent this group is actually being persecuted,\\u201d Spiro said. The order may also affect special refugee programs for Iraqis who worked for the U.S. government as translators after the 2003 invasion of Iraq. It is already affecting refugees and their families, said Jen Smyers of the Church World Service, a Protestant faith-based group that works with migrants. Smyers said she spoke to an Iraqi mother whose twin daughters remain in Iraq due to processing delays. \\u201cThose two 18-year-old daughters won\\u2019t be able to join their mother in the U.S.,\\u201d she said. Democrats on Friday were quick to condemn Trump\\u2019s order as un-American, saying it would tarnish the reputation of the United States as a land that welcomes immigrants. \\u201cToday\\u2019s executive order from President Trump is more about extreme xenophobia than extreme vetting,\\u201d said Democratic Senator Edward Markey in a statement. Some Republicans praised the move. Representative Bob Goodlatte, chairman of the House of Representatives Judiciary Committee, said Islamic State has threatened to use the U.S. immigration system, making it important to do more screening. \\u201cI am pleased that President Trump is using the tools granted to him by Congress and the power granted by the Constitution to help keep America safe and ensure we know who is entering the United States,\\u201d Goodlatte said in a statement. Without naming

Trump, Iranian President Hassan Rouhani said on Saturday it was no time to build walls between nations and criticized steps towards cancelling world trade agreements. Trump on Wednesday ordered the construction of a U.S.-Mexican border wall, a major promise during his election campaign, as part of a package of measures to curb illegal immigration. \\u201cToday is not the time to erect walls between nations. They have forgotten that the Berlin wall fell years ago,\\u201d Rouhani said in a speech carried live on Iranian state television. He made no direct reference to Trump\\u2019s order regarding refugees and travelers from the seven mainly Muslim states. Rouhani, a pragmatist elected in 2013, thawed Iran\\u2019s relations with world powers after years of confrontation and engineered its 2015 deal with them under which it curbed its nuclear program in exchange for relief from sanctions. Rouhani said earlier this month that Trump could not unilaterally cancel the nuclear deal and that talk of renegotiating it was \\u201cmeaningless\\u201d. France and Germany voiced disquiet on Saturday over Trump\\u2019s new restrictions on immigration. \\u201cWelcoming refugees who flee war and oppression is part of our duty,\\u201d French Foreign Minister Jean-Marc Ayrault said at a joint news conference with German counterpart Sigmar Gabriel. \\u201cThe United States is a country where Christian traditions have an important meaning. Loving your neighbor is a major Christian value, and that includes helping people,\\u201d said Gabriel. \\u201cI think that is what unites us in the West, and I think that is what we want to make clear to the Americans.\\u201d \",\n   \"PRISTINA (Reuters) - Kosovo s center-right coalition led by the Democratic Party of Kosovo signed a deal on Monday with the small New Alliance for Kosovo party to form a government, ending nearly three months of political deadlock after an election on June 11.   Finally Kosovo has started to move ... we had some big delays and our institutions now will be formed, said Ramush Haradinaj, from the center-right coalition of parties made up of former guerrillas who fought the 1998-99 war against Serb forces. Under the deal, the parties along with ethnic minorities will secure 63 seats in the 120-seat parliament. President Hashim Thaci is expected to give Haradinaj a mandate to form the government within days. A source who asked not to be named told Reuters the parliament session to elect the parliament speaker would be held this week. Haradinaj, who twice stood trial before the United Nations war crimes court for war crimes and was acquitted, briefly held the post of prime minister in 2005. The smaller New Alliance for Kosovo party is led by Behgjet Pacolli, who is dubbed by media the richest Kosovar. Pacolli, who also holds a Swiss passport, won many contracts from the Russian government to rebuild state buildings in Moscow in the 90s but a decade ago he moved his business from Moscow to Kazakhstan. It is unclear what post Pacolli will hold in the new government. The new government will have to tackle unemployment running at 30 percent and improve relations with Kosovo s neighbors, especially Serbia, a precondition for both countries to move forward in the European Union accession process. It

must also reform health and education and the tax administration system as well as include representatives of some 120,000 Kosovo Serbs who do not recognize independence. Kosovo declared independence from Serbia in 2008, almost a decade after NATO air strikes drove out Serbian forces accused of expelling and killing ethnic Albanian civilians in a two-year counter-insurgency. \"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"subject\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 8,\n   \"samples\": [\n   \"politics\",\n \"Middle-east\",\n   \"News\"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"date\",\n   \"properties\": {\n \"dtype\": \"category\",\n   \"num_unique_values\": 2397,\n \"samples\": [\n   \"October 6, 2016\",\n   \"June 10, 2017\",\n   \"Sep 13, 2015\"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"label\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0,\n   \"min\": 0,\n   \"max\": 1,\n   \"num_unique_values\": 2,\n \"samples\": [\n   1,\n   0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe","variable_name":"data"}

```
# Lets take a quick look at the joined dataset
data.head()
```

{"summary":"{\n \"name\": \"data\",\n \"rows\": 44898,\n \"fields\": [\n   {\n   \"column\": \"title\",\n \"properties\": {\n   \"dtype\": \"string\",\n \"num_unique_values\": 38729,\n   \"samples\": [\n \"Supreme Court Justice Ginsburg 'regrets' Trump criticisms\",\n \" DOZENS Of GOP Foreign Policy Experts Pledge To Stop Trump From Winning Nomination\",\n   \" REPORT: Trump Laughed After Woman Was Grabbed \\u2018By The P*ssy\\u2019 On Apprentice Set\"\n   ],\n   \"semantic_type\": \"\",\n \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"text\",\n   \"properties\": {\n   \"dtype\": \"string\",\n \"num_unique_values\": 38646,\n   \"samples\": [\n   \" (This September 29 has been corrected to fix date of election in paragraph 3) NAIROBI (Reuters) - A Kenyan government watchdog said on Friday it was investigating whether police had assaulted students during protests this week at the University of Nairobi over the detention of an opposition lawmaker.   Police fired tear gas on Thursday at the protesting students. Video footage posted on social media later in the day showed uniformed officers outside dormitories and inside classrooms using batons to hit people who did not appear to be involved in the campus protests. It was the latest crackdown by police on protests since an Aug. 8 presidential election that was later annulled by the Supreme Court. A re-run of the vote has been set

for Oct. 26. The Independent Police Oversight Authority (IPOA) this morning noted from social media reports of an incident in which members of the National Police Service allegedly stormed the University of Nairobi and assaulted students at the institution, the authority said on its Twitter feed. It requested that any member of the public come forward to provide information to aid the investigation. The students had been protesting against the re-arrest of a lawmaker, Paul Ongili Owino, on Wednesday, shortly after he was released on bail on charges of subversion for calling President Uhuru Kenyatta a son of a dog at a campaign rally. Kenya is a key Western ally in a region often roiled by violence. Preparations for the re-run of the election are being closely monitored for signs of instability, after at least 28 people were killed in unrest following the Aug. 8 vote. The IPOA watchdog was created in 2011, after police came under severe criticism for the number of protesters killed during demonstrations against disputed elections in 2007. The violence then killed around 1,2000 people. But local and international rights groups say the institution is struggling to fulfill its mandate to investigate allegations of police brutality in a country rife with reports from civilians of extrajudicial killings by security forces. The IPOA has secured two convictions of police officers in the four years it has been operational. The watchdog pledged last month to fast-track investigations into high-profile deaths such as that of a baby allegedly killed by police in the violence after the election. [L4N1L34FC] But people familiar with the status of these investigations say the police are not cooperating with them, and that senior officials in the police force and the interior ministry insist officers killed only thieves and thugs. \",\n\"WASHINGTON/CAIRO (Reuters) - Five Iraqi passengers and one Yemeni were barred from boarding an EgyptAir flight from Cairo to New York on Saturday after President Donald Trump halted the entry of citizens from seven Muslim-majority countries, sources at Cairo airport said. The passengers, arriving in transit to Cairo airport, were stopped and re-directed to flights headed for their home countries despite holding valid visas, the sources said. Trump on Friday put a four-month hold on allowing refugees into the United States and temporarily barred travelers from Syria and six other Muslim-majority countries, saying the moves would help protect Americans from terrorist attacks. He said his most sweeping use of his presidential powers since taking office a week ago, barring travelers from the seven nations for at least 90 days, would give his administration time to develop more stringent screening procedures for refugees, immigrants and visitors. \\u201cI\\u2019m establishing new vetting measures to keep radical Islamic terrorists out of the United States of America. Don\\u2019t want them here,\\u201d Trump said earlier on Friday at the Pentagon. \\u201cWe only want to admit those into our country who will support our country and love deeply our people,\\u201d he said. The bans, though temporary, took effect immediately, causing havoc and confusion for would-be travelers with passports from Iran, Iraq, Libya, Somalia,

Sudan, Syria and Yemen. Besides Cairo it was not immediately clear whether other airports of countries listed by Trump had swiftly implemented the ban. Arab officials of the listed countries would not comment on the matter. The order seeks to prioritize refugees fleeing religious persecution, a move Trump separately said was aimed at helping Christians in Syria. That led some legal experts to question whether the order was constitutional. One group said it would announce a court challenge on Monday. The Council on American-Islamic Relations said the order targets Muslims because of their faith, contravening the U.S. Constitutional right to freedom of religion. \\u201cPresident Trump has cloaked what is a discriminatory ban against nationals of Muslim countries under the banner of national security,\\u201d said Greg Chen of the American Immigration Lawyers Association. Trump has long pledged to take this kind of action, making it a prominent feature of his campaign for the Nov. 8 election. But people who work with Muslim immigrants and refugees were scrambling to determine the scope of the order. Even legal permanent residents - people with \\ u201cgreen cards\\u201d allowing them to live and work in the United States - were being advised to consult immigration lawyers before traveling outside the country, or trying to return, according to Muslim Advocates, a civil rights group in Washington. On Friday evening, Abed Ayoub of the American-Arab Anti-Discrimination Committee said he had fielded about 100 queries from people anxious about the order, which he said he believed could affect traveling green card holders, students, people coming to the United States for medical care and others. \\u201cIt\\u2019s chaos,\\u201d Ayoub said. During his campaign, Trump tapped into American fears about Islamic State militants and the flood of migrants into Europe from Syria\\u2019s civil war, saying refugees could be a \\u201cTrojan horse\\u201d that allowed attackers to enter the United States. In December 2015, he called for a ban on all Muslims entering the United States, drawing fire for suggesting a religious test for immigrants that critics said would violate the U.S. Constitution. His idea later evolved into a proposal for \\u201cextreme vetting.\\u201d Trump\\u2019s order also suspends the Syrian refugee program until further notice, and will eventually give priority to minority religious groups fleeing persecution. Trump said in an interview with the Christian Broadcasting Network that the exception would help Syrian Christians fleeing the civil war there. Legal experts were divided on whether this order would be constitutional. \\u201cIf they are thinking about an exception for Christians, in almost any other legal context discriminating in favor of one religion and against another religion could violate the constitution,\\u201d said Stephen Legomsky, a former chief counsel at U.S. Citizenship and Immigration Services in the Obama administration. But Peter Spiro, a professor at Temple University Beasley School of Law, said Trump\\u2019s action would likely be constitutional because the president and Congress are allowed considerable deference when it comes to asylum decisions. \\ u201cIt\\u2019s a completely plausible prioritization, to the extent

this group is actually being persecuted,\\u201d Spiro said. The order may also affect special refugee programs for Iraqis who worked for the U.S. government as translators after the 2003 invasion of Iraq. It is already affecting refugees and their families, said Jen Smyers of the Church World Service, a Protestant faith-based group that works with migrants. Smyers said she spoke to an Iraqi mother whose twin daughters remain in Iraq due to processing delays. \\u201cThose two 18-year-old daughters won\\u2019t be able to join their mother in the U.S.,\\u201d she said. Democrats on Friday were quick to condemn Trump\\u2019s order as un-American, saying it would tarnish the reputation of the United States as a land that welcomes immigrants. \\u201cToday\\u2019s executive order from President Trump is more about extreme xenophobia than extreme vetting,\\u201d said Democratic Senator Edward Markey in a statement. Some Republicans praised the move. Representative Bob Goodlatte, chairman of the House of Representatives Judiciary Committee, said Islamic State has threatened to use the U.S. immigration system, making it important to do more screening. \\u201cI am pleased that President Trump is using the tools granted to him by Congress and the power granted by the Constitution to help keep America safe and ensure we know who is entering the United States,\\u201d Goodlatte said in a statement. Without naming Trump, Iranian President Hassan Rouhani said on Saturday it was no time to build walls between nations and criticized steps towards cancelling world trade agreements. Trump on Wednesday ordered the construction of a U.S.-Mexican border wall, a major promise during his election campaign, as part of a package of measures to curb illegal immigration. \\u201cToday is not the time to erect walls between nations. They have forgotten that the Berlin wall fell years ago,\\u201d Rouhani said in a speech carried live on Iranian state television. He made no direct reference to Trump\\u2019s order regarding refugees and travelers from the seven mainly Muslim states. Rouhani, a pragmatist elected in 2013, thawed Iran\\u2019s relations with world powers after years of confrontation and engineered its 2015 deal with them under which it curbed its nuclear program in exchange for relief from sanctions. Rouhani said earlier this month that Trump could not unilaterally cancel the nuclear deal and that talk of renegotiating it was \\u201cmeaningless\\u201d. France and Germany voiced disquiet on Saturday over Trump\\u2019s new restrictions on immigration. \\u201cWelcoming refugees who flee war and oppression is part of our duty,\\u201d French Foreign Minister Jean-Marc Ayrault said at a joint news conference with German counterpart Sigmar Gabriel. \\u201cThe United States is a country where Christian traditions have an important meaning. Loving your neighbor is a major Christian value, and that includes helping people,\\u201d said Gabriel. \\u201cI think that is what unites us in the West, and I think that is what we want to make clear to the Americans.\\u201d \",\n   \"PRISTINA (Reuters) - Kosovo s center-right coalition led by the Democratic Party of Kosovo signed a deal on Monday with the small New Alliance for Kosovo party to form a government, ending

nearly three months of political deadlock after an election on June 11. Finally Kosovo has started to move ... we had some big delays and our institutions now will be formed, said Ramush Haradinaj, from the center-right coalition of parties made up of former guerrillas who fought the 1998-99 war against Serb forces. Under the deal, the parties along with ethnic minorities will secure 63 seats in the 120-seat parliament. President Hashim Thaci is expected to give Haradinaj a mandate to form the government within days. A source who asked not to be named told Reuters the parliament session to elect the parliament speaker would be held this week. Haradinaj, who twice stood trial before the United Nations war crimes court for war crimes and was acquitted, briefly held the post of prime minister in 2005. The smaller New Alliance for Kosovo party is led by Behgjet Pacolli, who is dubbed by media the richest Kosovar. Pacolli, who also holds a Swiss passport, won many contracts from the Russian government to rebuild state buildings in Moscow in the 90s but a decade ago he moved his business from Moscow to Kazakhstan. It is unclear what post Pacolli will hold in the new government. The new government will have to tackle unemployment running at 30 percent and improve relations with Kosovo s neighbors, especially Serbia, a precondition for both countries to move forward in the European Union accession process. It must also reform health and education and the tax administration system as well as include representatives of some 120,000 Kosovo Serbs who do not recognize independence. Kosovo declared independence from Serbia in 2008, almost a decade after NATO air strikes drove out Serbian forces accused of expelling and killing ethnic Albanian civilians in a two-year counter-insurgency. \"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"subject\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 8,\n   \"samples\": [\n   \"politics\",\n \"Middle-east\",\n   \"News\"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"date\",\n   \"properties\": {\n \"dtype\": \"category\",\n   \"num_unique_values\": 2397,\n \"samples\": [\n   \"October 6, 2016\",\n   \"June 10, 2017\",\n   \"Sep 13, 2015\"\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"label\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0,\n   \"min\": 0,\n   \"max\": 1,\n   \"num_unique_values\": 2,\n \"samples\": [\n   1,\n   0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe","variable_name":"data"}

```
# The shape should only change in the rows
data.shape
```

```
(44898, 5)
```

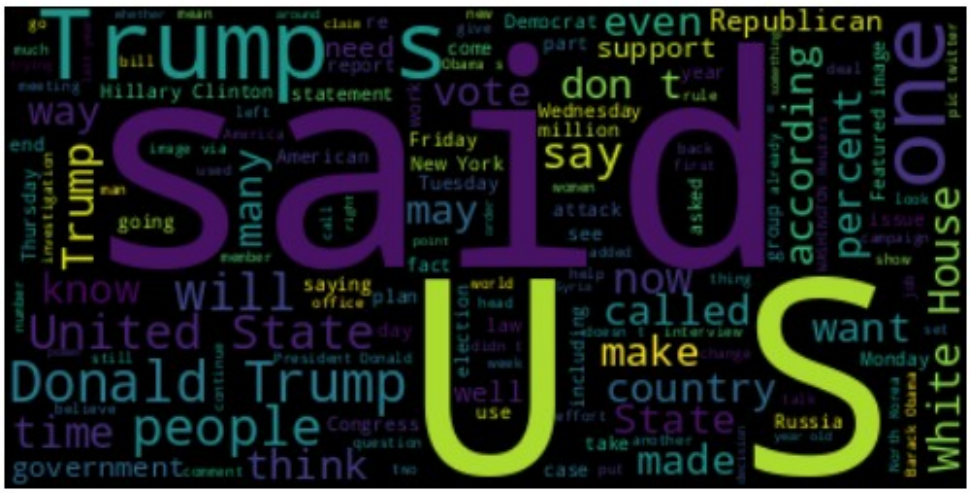
```
# Lets describe the dataset to get a better overview of the dataset
data.describe(include = 'all')
```


```
{"summary":"{\n  \"name\": \"data\",\n  \"rows\": 11,\n  \"fields\":
[\n    {\n      \"column\": \"title\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 4,\n
\"samples\": [\n          38729,\n          \"14\",\n
\"44898\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"text\",\n      \"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 4,\n        \"samples\": [\n          38646,\n
\"627\",\n          \"44898\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"subject\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 4,\n        \"samples\":
[\n          8,\n          \"11272\",\n          \"44898\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"date\",\n      \"properties\": {\n        \"dtype\": \"date\",\n
\"min\": \"1970-01-01 00:00:00.000000182\",\n        \"max\": \"2017-
12-20 00:00:00\",\n        \"num_unique_values\": 4,\n
\"samples\": [\n          2397,\n          \"182\",\n
\"44898\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"label\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 15873.689800783677,\n        \"min\": 0.0,\n        \"max\":
44898.0,\n        \"num_unique_values\": 5,\n        \"samples\": [\n
0.47701456635039424,\n          1.0,\n          0.49947695279473303\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    }\n  ]\n}","type":"dataframe"}
```


Unified Dataset Word cloud visual representation

```
# Now lets do it for the whole dataset
text_data = " ".join(review for review in data.text)

# The text might be too long to display, so lets just get the word
count
print("The word count for all the news text is: ",
len(text_data.split()))

# Lets generate the work cloud
wordcloud = WordCloud().generate(text_data)

# Lets display the word cloud
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
The word count for all the news text is:  18196364
```



Feature engineering

Unsorting the data

```
data.head()
```


```
{"summary":"{\n  \"name\": \"data\",\n  \"rows\": 44898,\n
\"fields\": [\n    {\n      \"column\": \"title\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 38729,\n        \"samples\": [\n
\"Supreme Court Justice Ginsburg 'regrets' Trump criticisms\",\n
\" DOZENS Of GOP Foreign Policy Experts Pledge To Stop Trump From
Winning Nomination\",\n          \" REPORT: Trump Laughed After Woman
Was Grabbed \\u2018By The P*ssy\\u2019 On Apprentice Set\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"text\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 38646,\n        \"samples\": [\n          \"
(This September 29 has been corrected to fix date of election in
paragraph 3) NAIROBI (Reuters) - A Kenyan government watchdog said on
Friday it was investigating whether police had assaulted students
during protests this week at the University of Nairobi over the
detention of an opposition lawmaker.   Police fired tear gas on
Thursday at the protesting students. Video footage posted on social
media later in the day showed uniformed officers outside dormitories
and inside classrooms using batons to hit people who did not appear to
be involved in the campus protests.  It was the latest crackdown by
police on protests since an Aug. 8 presidential election that was
later annulled by the Supreme Court. A re-run of the vote has been set
for Oct. 26. The Independent Police Oversight Authority (IPOA)  this
morning noted from social media reports of an incident in which
members of the National Police Service allegedly stormed the
```

University of Nairobi and assaulted students at the institution, the authority said on its Twitter feed. It requested that any member of the public come forward to provide information to aid the investigation. The students had been protesting against the re-arrest of a lawmaker, Paul Ongili Owino, on Wednesday, shortly after he was released on bail on charges of subversion for calling President Uhuru Kenyatta a son of a dog at a campaign rally. Kenya is a key Western ally in a region often roiled by violence. Preparations for the re-run of the election are being closely monitored for signs of instability, after at least 28 people were killed in unrest following the Aug. 8 vote. The IPOA watchdog was created in 2011, after police came under severe criticism for the number of protesters killed during demonstrations against disputed elections in 2007. The violence then killed around 1,2000 people. But local and international rights groups say the institution is struggling to fulfill its mandate to investigate allegations of police brutality in a country rife with reports from civilians of extrajudicial killings by security forces. The IPOA has secured two convictions of police officers in the four years it has been operational. The watchdog pledged last month to fast-track investigations into high-profile deaths such as that of a baby allegedly killed by police in the violence after the election. [L4N1L34FC] But people familiar with the status of these investigations say the police are not cooperating with them, and that senior officials in the police force and the interior ministry insist officers killed only thieves and thugs. \",\n\"WASHINGTON/CAIRO (Reuters) - Five Iraqi passengers and one Yemeni were barred from boarding an EgyptAir flight from Cairo to New York on Saturday after President Donald Trump halted the entry of citizens from seven Muslim-majority countries, sources at Cairo airport said. The passengers, arriving in transit to Cairo airport, were stopped and re-directed to flights headed for their home countries despite holding valid visas, the sources said. Trump on Friday put a four-month hold on allowing refugees into the United States and temporarily barred travelers from Syria and six other Muslim-majority countries, saying the moves would help protect Americans from terrorist attacks. He said his most sweeping use of his presidential powers since taking office a week ago, barring travelers from the seven nations for at least 90 days, would give his administration time to develop more stringent screening procedures for refugees, immigrants and visitors. \\u201cI\\u2019m establishing new vetting measures to keep radical Islamic terrorists out of the United States of America. Don\\u2019t want them here,\\u201d Trump said earlier on Friday at the Pentagon. \\u201cWe only want to admit those into our country who will support our country and love deeply our people,\\u201d he said. The bans, though temporary, took effect immediately, causing havoc and confusion for would-be travelers with passports from Iran, Iraq, Libya, Somalia, Sudan, Syria and Yemen. Besides Cairo it was not immediately clear whether other airports of countries listed by Trump had swiftly implemented the ban. Arab officials of the listed countries would not

comment on the matter. The order seeks to prioritize refugees fleeing religious persecution, a move Trump separately said was aimed at helping Christians in Syria. That led some legal experts to question whether the order was constitutional. One group said it would announce a court challenge on Monday. The Council on American-Islamic Relations said the order targets Muslims because of their faith, contravening the U.S. Constitutional right to freedom of religion. \\u201cPresident Trump has cloaked what is a discriminatory ban against nationals of Muslim countries under the banner of national security,\\u201d said Greg Chen of the American Immigration Lawyers Association. Trump has long pledged to take this kind of action, making it a prominent feature of his campaign for the Nov. 8 election. But people who work with Muslim immigrants and refugees were scrambling to determine the scope of the order. Even legal permanent residents - people with \\ u201cgreen cards\\u201d allowing them to live and work in the United States - were being advised to consult immigration lawyers before traveling outside the country, or trying to return, according to Muslim Advocates, a civil rights group in Washington. On Friday evening, Abed Ayoub of the American-Arab Anti-Discrimination Committee said he had fielded about 100 queries from people anxious about the order, which he said he believed could affect traveling green card holders, students, people coming to the United States for medical care and others. \\u201cIt\\u2019s chaos,\\u201d Ayoub said. During his campaign, Trump tapped into American fears about Islamic State militants and the flood of migrants into Europe from Syria\\u2019s civil war, saying refugees could be a \\u201cTrojan horse\\u201d that allowed attackers to enter the United States. In December 2015, he called for a ban on all Muslims entering the United States, drawing fire for suggesting a religious test for immigrants that critics said would violate the U.S. Constitution. His idea later evolved into a proposal for \\u201cextreme vetting.\\u201d Trump\\u2019s order also suspends the Syrian refugee program until further notice, and will eventually give priority to minority religious groups fleeing persecution. Trump said in an interview with the Christian Broadcasting Network that the exception would help Syrian Christians fleeing the civil war there. Legal experts were divided on whether this order would be constitutional. \\u201cIf they are thinking about an exception for Christians, in almost any other legal context discriminating in favor of one religion and against another religion could violate the constitution,\\u201d said Stephen Legomsky, a former chief counsel at U.S. Citizenship and Immigration Services in the Obama administration. But Peter Spiro, a professor at Temple University Beasley School of Law, said Trump\\u2019s action would likely be constitutional because the president and Congress are allowed considerable deference when it comes to asylum decisions. \\ u201cIt\\u2019s a completely plausible prioritization, to the extent this group is actually being persecuted,\\u201d Spiro said. The order may also affect special refugee programs for Iraqis who worked for the U.S. government as translators after the 2003 invasion of Iraq. It is

already affecting refugees and their families, said Jen Smyers of the Church World Service, a Protestant faith-based group that works with migrants. Smyers said she spoke to an Iraqi mother whose twin daughters remain in Iraq due to processing delays. \\u201cThose two 18-year-old daughters won\\u2019t be able to join their mother in the U.S.,\\u201d she said. Democrats on Friday were quick to condemn Trump\\u2019s order as un-American, saying it would tarnish the reputation of the United States as a land that welcomes immigrants. \\u201cToday\\u2019s executive order from President Trump is more about extreme xenophobia than extreme vetting,\\u201d said Democratic Senator Edward Markey in a statement. Some Republicans praised the move. Representative Bob Goodlatte, chairman of the House of Representatives Judiciary Committee, said Islamic State has threatened to use the U.S. immigration system, making it important to do more screening. \\u201cI am pleased that President Trump is using the tools granted to him by Congress and the power granted by the Constitution to help keep America safe and ensure we know who is entering the United States,\\u201d Goodlatte said in a statement. Without naming Trump, Iranian President Hassan Rouhani said on Saturday it was no time to build walls between nations and criticized steps towards cancelling world trade agreements. Trump on Wednesday ordered the construction of a U.S.-Mexican border wall, a major promise during his election campaign, as part of a package of measures to curb illegal immigration. \\u201cToday is not the time to erect walls between nations. They have forgotten that the Berlin wall fell years ago,\\u201d Rouhani said in a speech carried live on Iranian state television. He made no direct reference to Trump\\u2019s order regarding refugees and travelers from the seven mainly Muslim states. Rouhani, a pragmatist elected in 2013, thawed Iran\\u2019s relations with world powers after years of confrontation and engineered its 2015 deal with them under which it curbed its nuclear program in exchange for relief from sanctions. Rouhani said earlier this month that Trump could not unilaterally cancel the nuclear deal and that talk of renegotiating it was \\u201cmeaningless\\u201d. France and Germany voiced disquiet on Saturday over Trump\\u2019s new restrictions on immigration. \\u201cWelcoming refugees who flee war and oppression is part of our duty,\\u201d French Foreign Minister Jean-Marc Ayrault said at a joint news conference with German counterpart Sigmar Gabriel. \\u201cThe United States is a country where Christian traditions have an important meaning. Loving your neighbor is a major Christian value, and that includes helping people,\\u201d said Gabriel. \\u201cI think that is what unites us in the West, and I think that is what we want to make clear to the Americans.\\u201d \",\n   \"PRISTINA (Reuters) - Kosovo s center-right coalition led by the Democratic Party of Kosovo signed a deal on Monday with the small New Alliance for Kosovo party to form a government, ending nearly three months of political deadlock after an election on June 11.   Finally Kosovo has started to move ... we had some big delays and our institutions now will be formed, said Ramush Haradinaj, from

```
the center-right coalition of parties made up of former guerrillas who fought the 1998-99 war against Serb forces. Under the deal, the parties along with ethnic minorities will secure 63 seats in the 120-seat parliament.  President Hashim Thaci is expected to give Haradinaj a mandate to form the government within days. A source who asked not to be named told Reuters the parliament session to elect the parliament speaker would be held this week. Haradinaj, who twice stood trial before the United Nations war crimes court for war crimes and was acquitted, briefly held the post of prime minister in 2005.  The smaller New Alliance for Kosovo party is led by Behgjet Pacolli, who is dubbed by media the richest Kosovar.  Pacolli, who also holds a Swiss passport, won many contracts from the Russian government to rebuild state buildings in Moscow in the  90s but a decade ago he moved his business from Moscow to Kazakhstan.  It is unclear what post Pacolli will hold in the new government.  The new government will have to tackle unemployment running at 30 percent and improve relations with Kosovo s neighbors, especially Serbia, a precondition for both countries to move forward in the European Union accession process. It must also reform health and education and the tax administration system as well as include representatives of some 120,000 Kosovo Serbs who do not recognize independence. Kosovo declared independence from Serbia in 2008, almost a decade after NATO air strikes drove out Serbian forces accused of expelling and killing ethnic Albanian civilians in a two-year counter-insurgency.  \"\n        ],\n \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"subject\",\n      \"properties\": {\n        \"dtype\": \"category\",\n        \"num_unique_values\": 8,\n        \"samples\": [\n          \"politics\",\n \"Middle-east\",\n          \"News\"\n        ],\n \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"date\",\n      \"properties\": {\n \"dtype\": \"category\",\n        \"num_unique_values\": 2397,\n \"samples\": [\n          \"October 6, 2016\",\n          \"June 10, 2017\",\n          \"Sep 13, 2015\"\n        ],\n \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"label\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 0,\n        \"min\": 0,\n        \"max\": 1,\n        \"num_unique_values\": 2,\n \"samples\": [\n          1,\n          0\n        ],\n \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe","variable_name":"data"}
```

```
data.tail()
```

```
{"repr_error":"0","type":"dataframe"}
```

After joining the datasets (real and fake), lets unsort them

```
data = data.sample(frac=1).reset_index(drop=True)
```

```
data.head()
```


```
{"summary":"{\n  \"name\": \"data\",\n  \"rows\": 44898,\n
\"fields\": [\n    {\n      \"column\": \"title\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 38729,\n        \"samples\": [\n
\"Alabama governor could face charges after ethics panel ruling\",\n
\"BITTER RADICAL ERIC HOLDER Goes After \\u2018ORANGE MAN\\u2019
President Trump In Scorching Interview: \\u2018We want the America of
Barack Obama\\u2019\",\n          \"WATCH HILLARY LAUGH When Trump
Mentions Gays Who Are Thrown Off Buildings By Muslims In Countries Who
Fund Her Campaign\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"text\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 38646,\n        \"samples\": [\n          \"NEW
YORK (Reuters) - Preet Bharara, the top federal prosecutor in
Manhattan known for pursuing a series of cases targeting public
corruption and crime on Wall Street, said on Wednesday he has agreed
to remain in his post after Donald Trump becomes U.S. president.
Bharara, appointed to his position by Democratic President Barack
Obama in 2009, told reporters following a meeting with the Republican
president-elect at Trump Tower in Manhattan that Trump asked him to
stay on during his administration and he accepted. Trump takes office
on Jan. 20. \\u201cWe had a good meeting,\\u201d Bharara said. \\
u201cI said I would absolutely consider staying on. I agreed to stay
on.\\u201d The announcement\\u2019s timing, when Trump has not yet
finished filling all Cabinet-level positions, was unusual. But some
former prosecutors who served under Bharara said they were not
surprised their former boss would be willing to remain as the U.S.
Attorney for the Southern District of New York. \\u201cI think Preet
is an independent, law enforcement-minded prosecutor who loves his job
and is clearly talented in it,\\u201d said Arlo Devlin-Brown, a former
chief of Bharara\\u2019s public corruption unit who is now a partner
at the law firm Covington & Burling. U.S. Senator Charles Schumer of
New York, the incoming Senate Democratic leader who Bharara previously
worked for as chief counsel, said Trump called him last week to ask
what he thought about Bharara staying in his job. \\u201cI am glad
they met and am glad Preet is staying on,\\u201d Schumer said in a
statement. \\u201cHe\\u2019s been one of the best U.S. Attorneys New
York has ever seen.\\u201d Bharara\\u2019s office has pursued an
aggressive push against corruption in state and city politics, an
agenda that could fit with Trump\\u2019s vow to \\u201cdrain the
swamp\\u201d in Washington. Those political investigations led last
year to the convictions of former New York Assembly Speaker Sheldon
Silver, a Democrat, and former New York Senate Majority Leader Dean
Skelos, a Republican, in separate corruption trials. Bharara also
brought dozens of successful cases against insider traders and was on
the cover of Time magazine in 2012 with the headline \\u201cThis man
is busting Wall St.\\u201d Those cases include the 2011 conviction of
Galleon Group founder Raj Rajaratnam, who is serving an 11-year prison
```

term, and a $1.8 billion settlement and plea deal in 2013 with hedge fund SAC Capital Advisors LP. His 227-lawyer office also secured corporate settlements with companies including General Motors Co and JPMorgan Chase & Co; won several convictions and guilty pleas of former employees of Ponzi scheme operator Bernard Madoff; and prosecuted Suleiman Abu Ghaith, a son-in-law of the late al Qaeda leader Osama bin Laden. Bharara\\u2019s office\\u2019s priorities have often matched those set by Obama\\u2019s Justice Department. Amid an increase in civil rights investigations nationally, for example, Bharara\\u2019s office joined a lawsuit that led to a settlement in 2015 aimed at reducing violence in New York City\\u2019s Rikers Island jail complex. How priorities set by the Justice Department under Trump\\u2019s pick for attorney general, Republican Senator Jeff Sessions of Alabama, affects the cases Bharara\\u2019s office pursues remains unclear. \\u201cObviously there is likely be some changes in priorities from Main Justice (the department\\u2019s Washington headquarters), and the office will have to adjust to those,\\u201d said Richard Zabel, who previously served as Bharara\\u2019s deputy before becoming hedge fund Elliott Management\\u2019s general counsel. Bharara said his office had for the past seven years pursued its work \\u201cindependently, without fear or favor.\\u201d Former prosecutors they expect that to stay the same. \\u201cHe would have only taken it on if he were 100 percent confident that that independence could be preserved,\\u201d said Matthew Schwartz, partner at the law firm Boies, Schiller & Flexner and a former prosecutor under Bharara. \",\n\"WASHINGTON (Reuters) - The U.S. House of Representatives on Thursday approved a $68 billion increase in military spending next year with legislation that also provides money to start construction of President Donald Trump\\u2019s Mexican border wall. The bill increased spending on the U.S. capability to defend itself from foreign missile attacks amid growing concerns about North Korea\\u2019s increasing capacity to hit the United States with a nuclear-tipped missile after it successfully tested an intercontinental ballistic missile in July. The money for the wall is dwarfed by the $658.1 billion the bill would provide for the Defense Department, an increase of $68.1 billion above the fiscal year 2017 enacted level and $18.4 billion above Trump\\u2019s budget request. The House voted 235-192 for the fiscal 2018 spending bill that would provide $1.6 billion for initial construction of a wall on the U.S.-Mexico border, which was a centerpiece of Trump\\u2019s 2016 presidential campaign. Democrats repeatedly have referred to any money for the wall as a \\u201cpoison pill\\u201d and are likely to try to kill it in the Senate. Congress is up against an Oct. 1 deadline - the start of a new fiscal year - for either passing spending bills or temporarily extending funding at current-year levels to give negotiators more time to come to agreements. Funding for the wall was tucked into a wide-ranging national security appropriations bill at the last minute by Republican leadership, knowing that many House members who oppose the wall would not sink defense spending with a \\u201cno\\u201d vote. Trump has

argued that a \\u201cbig beautiful wall\\u201d was needed along the entire southwestern U.S. border and that Mexico would ultimately pay for its construction. Mexico has flatly refused to pay and in recent weeks Trump indicated that there could be portions of the border that are not conducive to a wall. Democrats and many Republicans in Congress have questioned the feasibility and effectiveness of a border wall, with immigration advocacy groups arguing that it would not stem the flow of illegal border crossings and would hurt U.S.-Mexico relations. In interviews in recent weeks with more than a half-dozen Republican senators from states that voted for Trump for president last November, only Ted Cruz of Texas embraced building the wall. Similarly, House Republicans representing districts along the U.S.-Mexico border have expressed opposition to the barrier, which could end up costing well over $21 billion. Representative Nita Lowey, the senior Democrat on the House Appropriations Committee, called the wall a \\u201cwaste ... that experts confirm is unneeded and ineffective and cuts against our values as Americans.\\u201d Furthermore, Lowey noted that Pentagon funding would run into a technical problem as it breaches a cap on defense spending by $72 billion. If the bill became law, she said, it actually would \\u201ctrigger across-the-board cuts of 13 percent to every defense account\\u201d in order to stay within the cap. The beefed up defense spending would allow the Pentagon to continue military activities in Iraq, Afghanistan and other trouble spots and hire more troops while providing soldiers with a 2.4 percent pay raise. It also would allow the Pentagon to undertake a shopping spree with money to buy ships and submarines, aircraft, tanks and other big-ticket items. The House-passed bill also includes an increase for America\\u2019s nuclear weapons stockpile managed by the Department of Energy, as well as for U.S. Capitol Police following a June 14 shooting that gravely wounded Republican Representative Steve Scalise. A $825 million increase for the Missile Defense Agency to more than $8.6 billion is more than Trump asked for and includes additional boosters and missile silos for the main system that would defend against an ICBM attack, a program run by Boeing Co. Missile defense would also gain 14 more THAAD interceptors made by Lockheed Martin Co. \",\n   \"BOSSASO, SOMALIA S (Reuters) - A wheelbarrow exploded outside a police station in Bossaso, a port city in Somalia s semi-autonomous region of Puntland, on Tuesday, killing the lone man pushing it, a police officer told Reuters. Mohammed Abdi, a police officer, told Reuters security personnel had stopped a man outside a police station s checkpoint and then suddenly his wheelbarrow exploded. Only the wheelbarrow man died. A Reuters witness saw the debris of the wooden wheelbarrow and the dead porter. Abdi said police did not know what type of bomb had been used or who was behind the attack. \"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"subject\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 8,\n   \"samples\": [\n   \"Government News\",\n   \"politicsNews\",\n

```
\"politics\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"date\",\n      \"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 2397,\n        \"samples\": [\n
\"September 18, 2016\",\n          \"Dec 15, 2016\",\n
\"December 22, 2016 \"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"label\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0,\n        \"min\": 0,\n        \"max\": 1,\n
\"num_unique_values\": 2,\n        \"samples\": [\n          1,\n
0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"data"}

data.tail()

{"repr_error":"0","type":"dataframe"}
```

Removing punctuation from the text

By taking out punctuation, you cut down on the extra stuff in the data. This makes it simpler for the algorithms to pay attention to the main words and what they mean.

```
# Lets remove the puntiation from the text
# from https://stackoverflow.com/questions/53664775/how-to-remove-
punctuation-in-python
# and https://www.geeksforgeeks.org/python-remove-punctuation-from-
string/

# Procedure to remove punctuation
def clearPunctuation(text: str) -> str:
    incommingText = text.encode("utf8").decode("ascii",'ignore')
    listOfCharactersWithoutPunctuation = [ch for ch in incommingText
if ch not in string.punctuation]
    stringWithoutPunctuation =
''.join(listOfCharactersWithoutPunctuation)
    # Remove trailing spaces
    stringWithoutPunctuation = stringWithoutPunctuation.strip()

    return stringWithoutPunctuation

# adding a column to my dataset, a no-punctuation column
data['no_puntuation_text'] = data['text'].apply(clearPunctuation)
```

Taking a look at the original data

```
original_text = data['text']
original_text.head()
```

```
0    West Virginia has been devastated by a loss of...
1    THE SOCIALIST PRESIDENT WAS THIS GUY S TEACHER...
2    The Ferguson #BlackLivesMatter protesters are ...
3    Hysterical With all the evidence available to ...
4    CNN Money has released a report saying that Do...
Name: text, dtype: object
```

Taking a look at the cleaned data

```
cleaned_text = data['no_puntuation_text']
cleaned_text.head()
```

```
0    West Virginia has been devastated by a loss of...
1    THE SOCIALIST PRESIDENT WAS THIS GUY S TEACHER...
2    The Ferguson BlackLivesMatter protesters are s...
3    Hysterical With all the evidence available to ...
4    CNN Money has released a report saying that Do...
Name: no_puntuation_text, dtype: object
```

Lowercasing the text

```
# Seeting up the column with lowercase text
data['lowercase_text'] = data['no_puntuation_text'].str.lower()
```

What we had before

```
data['no_puntuation_text'].head()
```

```
0    West Virginia has been devastated by a loss of...
1    THE SOCIALIST PRESIDENT WAS THIS GUY S TEACHER...
2    The Ferguson BlackLivesMatter protesters are s...
3    Hysterical With all the evidence available to ...
4    CNN Money has released a report saying that Do...
Name: no_puntuation_text, dtype: object
```

What we have now

```
data['lowercase_text'].head()
```

```
0    west virginia has been devastated by a loss of...
1    the socialist president was this guy s teacher...
2    the ferguson blacklivesmatter protesters are s...
3    hysterical with all the evidence available to ...
4    cnn money has released a report saying that do...
Name: lowercase_text, dtype: object
```

Removing the double spaces

The other steps left double spaces in the text column, lets remove them

```
# taken and modified from:
https://stackoverflow.com/questions/1546226/is-there-a-simple-way-to-
remove-multiple-spaces-in-a-string
def remove_double_spaces(text):
    return re.sub(r'\s+', ' ', text)

# Lets add another column, now one without doble empty spaces
data['no_double_spaces'] =
data['lowercase_text'].apply(remove_double_spaces)

data['lowercase_text'].head()
```

```
0    west virginia has been devastated by a loss of...
1    the socialist president was this guy s teacher...
2    the ferguson blacklivesmatter protesters are s...
3    hysterical with all the evidence available to ...
4    cnn money has released a report saying that do...
Name: lowercase_text, dtype: object
```

```
data['no_double_spaces'].head()
```

```
0    west virginia has been devastated by a loss of...
1    the socialist president was this guy s teacher...
2    the ferguson blacklivesmatter protesters are s...
3    hysterical with all the evidence available to ...
4    cnn money has released a report saying that do...
Name: no_double_spaces, dtype: object
```

Removing Stopwords

Now, lets remove the stopwords from the text

```
# taken and modified from: https://www.geeksforgeeks.org/removing-
stop-words-nltk-python/
def remove_stopwords(text):
    stop_words = set(stopwords.words('english'))
    word_tokens = word_tokenize(text)
    filtered_sentence = [word for word in word_tokens if word.lower()
not in stop_words]
    return ' '.join(filtered_sentence)

# lets add another column, now one with no stopwords
data['no_stopwords'] =
data['no_double_spaces'].apply(remove_stopwords)

data['no_stopwords'].head()
```

```
0    west virginia devastated loss 10000 jobs due o...
1    socialist president guy teacher great encounte...
2    ferguson blacklivesmatter protesters spilling ...
3    hysterical evidence available american voters ...
4    cnn money released report saying donald trump ...
Name: no_stopwords, dtype: object
```

It is clean now!

Adding a Sentiment analysis to the dataset

```
# Taken and modified from:
https://stackoverflow.com/questions/57803412/applying-
sentimentintensityanalyzer-function-on-each-row-of-the-dataframe-prov

# Initialize the Sentiment Intensity Analyzer
sia = SentimentIntensityAnalyzer()

# Function to get the compound sentiment score
def get_sentiment(text):
    return sia.polarity_scores(text)['compound']

# Lets add another column, but now with the sentiment analysis
data['text_sentiment'] = data['no_stopwords'].apply(get_sentiment)

data.head()
```


```
{"summary":"{\n  \"name\": \"data\",\n  \"rows\": 44898,\n
\"fields\": [\n    {\n      \"column\": \"title\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 38729,\n        \"samples\": [\n
\"Alabama governor could face charges after ethics panel ruling\",\n
\"BITTER RADICAL ERIC HOLDER Goes After \\u2018ORANGE MAN\\u2019
President Trump In Scorching Interview: \\u2018We want the America of
Barack Obama\\u2019\",\n          \"WATCH HILLARY LAUGH When Trump
Mentions Gays Who Are Thrown Off Buildings By Muslims In Countries Who
Fund Her Campaign\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"text\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 38646,\n        \"samples\": [\n          \"NEW
YORK (Reuters) - Preet Bharara, the top federal prosecutor in
Manhattan known for pursuing a series of cases targeting public
corruption and crime on Wall Street, said on Wednesday he has agreed
to remain in his post after Donald Trump becomes U.S. president.
Bharara, appointed to his position by Democratic President Barack
Obama in 2009, told reporters following a meeting with the Republican
president-elect at Trump Tower in Manhattan that Trump asked him to
stay on during his administration and he accepted. Trump takes office
on Jan. 20. \\u201cWe had a good meeting,\\u201d Bharara said. \\
u201cI said I would absolutely consider staying on. I agreed to stay
```

on.\\u201d The announcement\\u2019s timing, when Trump has not yet finished filling all Cabinet-level positions, was unusual. But some former prosecutors who served under Bharara said they were not surprised their former boss would be willing to remain as the U.S. Attorney for the Southern District of New York. \\u201cI think Preet is an independent, law enforcement-minded prosecutor who loves his job and is clearly talented in it,\\u201d said Arlo Devlin-Brown, a former chief of Bharara\\u2019s public corruption unit who is now a partner at the law firm Covington & Burling. U.S. Senator Charles Schumer of New York, the incoming Senate Democratic leader who Bharara previously worked for as chief counsel, said Trump called him last week to ask what he thought about Bharara staying in his job. \\u201cI am glad they met and am glad Preet is staying on,\\u201d Schumer said in a statement. \\u201cHe\\u2019s been one of the best U.S. Attorneys New York has ever seen.\\u201d Bharara\\u2019s office has pursued an aggressive push against corruption in state and city politics, an agenda that could fit with Trump\\u2019s vow to \\u201cdrain the swamp\\u201d in Washington. Those political investigations led last year to the convictions of former New York Assembly Speaker Sheldon Silver, a Democrat, and former New York Senate Majority Leader Dean Skelos, a Republican, in separate corruption trials. Bharara also brought dozens of successful cases against insider traders and was on the cover of Time magazine in 2012 with the headline \\u201cThis man is busting Wall St.\\u201d Those cases include the 2011 conviction of Galleon Group founder Raj Rajaratnam, who is serving an 11-year prison term, and a $1.8 billion settlement and plea deal in 2013 with hedge fund SAC Capital Advisors LP. His 227-lawyer office also secured corporate settlements with companies including General Motors Co and JPMorgan Chase & Co; won several convictions and guilty pleas of former employees of Ponzi scheme operator Bernard Madoff; and prosecuted Suleiman Abu Ghaith, a son-in-law of the late al Qaeda leader Osama bin Laden. Bharara\\u2019s office\\u2019s priorities have often matched those set by Obama\\u2019s Justice Department. Amid an increase in civil rights investigations nationally, for example, Bharara\\u2019s office joined a lawsuit that led to a settlement in 2015 aimed at reducing violence in New York City\\u2019s Rikers Island jail complex. How priorities set by the Justice Department under Trump\\u2019s pick for attorney general, Republican Senator Jeff Sessions of Alabama, affects the cases Bharara\\u2019s office pursues remains unclear. \\u201cObviously there is likely be some changes in priorities from Main Justice (the department\\u2019s Washington headquarters), and the office will have to adjust to those,\\u201d said Richard Zabel, who previously served as Bharara\\u2019s deputy before becoming hedge fund Elliott Management\\u2019s general counsel. Bharara said his office had for the past seven years pursued its work \\u201cindependently, without fear or favor.\\u201d Former prosecutors they expect that to stay the same. \\u201cHe would have only taken it on if he were 100 percent confident that that independence could be preserved,\\u201d said Matthew Schwartz, partner at the law firm

Boies, Schiller & Flexner and a former prosecutor under Bharara. \",\n\"WASHINGTON (Reuters) - The U.S. House of Representatives on Thursday approved a $68 billion increase in military spending next year with legislation that also provides money to start construction of President Donald Trump\\u2019s Mexican border wall. The bill increased spending on the U.S. capability to defend itself from foreign missile attacks amid growing concerns about North Korea\\u2019s increasing capacity to hit the United States with a nuclear-tipped missile after it successfully tested an intercontinental ballistic missile in July. The money for the wall is dwarfed by the $658.1 billion the bill would provide for the Defense Department, an increase of $68.1 billion above the fiscal year 2017 enacted level and $18.4 billion above Trump\\u2019s budget request. The House voted 235-192 for the fiscal 2018 spending bill that would provide $1.6 billion for initial construction of a wall on the U.S.-Mexico border, which was a centerpiece of Trump\\u2019s 2016 presidential campaign. Democrats repeatedly have referred to any money for the wall as a \\u201cpoison pill\\u201d and are likely to try to kill it in the Senate. Congress is up against an Oct. 1 deadline - the start of a new fiscal year - for either passing spending bills or temporarily extending funding at current-year levels to give negotiators more time to come to agreements. Funding for the wall was tucked into a wide-ranging national security appropriations bill at the last minute by Republican leadership, knowing that many House members who oppose the wall would not sink defense spending with a \\u201cno\\u201d vote. Trump has argued that a \\u201cbig beautiful wall\\u201d was needed along the entire southwestern U.S. border and that Mexico would ultimately pay for its construction. Mexico has flatly refused to pay and in recent weeks Trump indicated that there could be portions of the border that are not conducive to a wall. Democrats and many Republicans in Congress have questioned the feasibility and effectiveness of a border wall, with immigration advocacy groups arguing that it would not stem the flow of illegal border crossings and would hurt U.S.-Mexico relations. In interviews in recent weeks with more than a half-dozen Republican senators from states that voted for Trump for president last November, only Ted Cruz of Texas embraced building the wall. Similarly, House Republicans representing districts along the U.S.-Mexico border have expressed opposition to the barrier, which could end up costing well over $21 billion. Representative Nita Lowey, the senior Democrat on the House Appropriations Committee, called the wall a \\u201cwaste ... that experts confirm is unneeded and ineffective and cuts against our values as Americans.\\u201d Furthermore, Lowey noted that Pentagon funding would run into a technical problem as it breaches a cap on defense spending by $72 billion. If the bill became law, she said, it actually would \\u201ctrigger across-the-board cuts of 13 percent to every defense account\\u201d in order to stay within the cap. The beefed up defense spending would allow the Pentagon to continue military activities in Iraq, Afghanistan and other trouble spots and hire more troops while providing soldiers with a 2.4 percent

pay raise. It also would allow the Pentagon to undertake a shopping spree with money to buy ships and submarines, aircraft, tanks and other big-ticket items. The House-passed bill also includes an increase for America\\u2019s nuclear weapons stockpile managed by the Department of Energy, as well as for U.S. Capitol Police following a June 14 shooting that gravely wounded Republican Representative Steve Scalise. A $825 million increase for the Missile Defense Agency to more than $8.6 billion is more than Trump asked for and includes additional boosters and missile silos for the main system that would defend against an ICBM attack, a program run by Boeing Co. Missile defense would also gain 14 more THAAD interceptors made by Lockheed Martin Co. \",\n \"BOSSASO, SOMALIA S (Reuters) - A wheelbarrow exploded outside a police station in Bossaso, a port city in Somalia s semi-autonomous region of Puntland, on Tuesday, killing the lone man pushing it, a police officer told Reuters. Mohammed Abdi, a police officer, told Reuters security personnel had stopped a man outside a police station s checkpoint and then suddenly his wheelbarrow exploded. Only the wheelbarrow man died. A Reuters witness saw the debris of the wooden wheelbarrow and the dead porter. Abdi said police did not know what type of bomb had been used or who was behind the attack. \"\n],\n \"semantic_type\": \"\",\n \"description\": \"\"\n }\n },\n {\n \"column\": \"subject\",\n \"properties\": {\n \"dtype\": \"category\",\n \"num_unique_values\": 8,\n \"samples\": [\n \"Government News\",\n \"politicsNews\",\n \"politics\"\n],\n \"semantic_type\": \"\",\n \"description\": \"\"\n }\n },\n {\n \"column\": \"date\",\n \"properties\": {\n \"dtype\": \"category\",\n \"num_unique_values\": 2397,\n \"samples\": [\n \"September 18, 2016\",\n \"Dec 15, 2016\",\n \"December 22, 2016 \"\n],\n \"semantic_type\": \"\",\n \"description\": \"\"\n }\n },\n {\n \"column\": \"label\",\n \"properties\": {\n \"dtype\": \"number\",\n \"std\": 0,\n \"min\": 0,\n \"max\": 1,\n \"num_unique_values\": 2,\n \"samples\": [\n 1,\n 0\n],\n \"semantic_type\": \"\",\n \"description\": \"\"\n }\n },\n {\n \"column\": \"no_puntuation_text\",\n \"properties\": {\n \"dtype\": \"string\",\n \"num_unique_values\": 38639,\n \"samples\": [\n \"NEW YORK Reuters Preet Bharara the top federal prosecutor in Manhattan known for pursuing a series of cases targeting public corruption and crime on Wall Street said on Wednesday he has agreed to remain in his post after Donald Trump becomes US president Bharara appointed to his position by Democratic President Barack Obama in 2009 told reporters following a meeting with the Republican presidentelect at Trump Tower in Manhattan that Trump asked him to stay on during his administration and he accepted Trump takes office on Jan 20 We had a good meeting Bharara said I said I would absolutely consider staying on I agreed to stay on The announcements

timing when Trump has not yet finished filling all Cabinetlevel positions was unusual But some former prosecutors who served under Bharara said they were not surprised their former boss would be willing to remain as the US Attorney for the Southern District of New York I think Preet is an independent law enforcementminded prosecutor who loves his job and is clearly talented in it said Arlo DevlinBrown a former chief of Bhararas public corruption unit who is now a partner at the law firm Covington Burling US Senator Charles Schumer of New York the incoming Senate Democratic leader who Bharara previously worked for as chief counsel said Trump called him last week to ask what he thought about Bharara staying in his job I am glad they met and am glad Preet is staying on Schumer said in a statement Hes been one of the best US Attorneys New York has ever seen Bhararas office has pursued an aggressive push against corruption in state and city politics an agenda that could fit with Trumps vow to drain the swamp in Washington Those political investigations led last year to the convictions of former New York Assembly Speaker Sheldon Silver a Democrat and former New York Senate Majority Leader Dean Skelos a Republican in separate corruption trials Bharara also brought dozens of successful cases against insider traders and was on the cover of Time magazine in 2012 with the headline This man is busting Wall St Those cases include the 2011 conviction of Galleon Group founder Raj Rajaratnam who is serving an 11year prison term and a 18 billion settlement and plea deal in 2013 with hedge fund SAC Capital Advisors LP His 227lawyer office also secured corporate settlements with companies including General Motors Co and JPMorgan Chase Co won several convictions and guilty pleas of former employees of Ponzi scheme operator Bernard Madoff and prosecuted Suleiman Abu Ghaith a soninlaw of the late al Qaeda leader Osama bin Laden Bhararas offices priorities have often matched those set by Obamas Justice Department Amid an increase in civil rights investigations nationally for example Bhararas office joined a lawsuit that led to a settlement in 2015 aimed at reducing violence in New York Citys Rikers Island jail complex How priorities set by the Justice Department under Trumps pick for attorney general Republican Senator Jeff Sessions of Alabama affects the cases Bhararas office pursues remains unclear Obviously there is likely be some changes in priorities from Main Justice the departments Washington headquarters and the office will have to adjust to those said Richard Zabel who previously served as Bhararas deputy before becoming hedge fund Elliott Managements general counsel Bharara said his office had for the past seven years pursued its work independently without fear or favor Former prosecutors they expect that to stay the same He would have only taken it on if he were 100 percent confident that that independence could be preserved said Matthew Schwartz partner at the law firm Boies Schiller Flexner and a former prosecutor under Bharara\",\n   \"Reuters London s Angel underground station is closed while authorities respond to a security alert outside the station Transport For London said in a tweet on Wednesday\"\n   ],\n   \"semantic_type\": \"\",\n

\"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"lowercase_text\",\n   \"properties\": {\n   \"dtype\": \"string\",\n   \"num_unique_values\": 38638,\n   \"samples\": [\n   \"new york reuters preet bharara the top federal prosecutor in manhattan known for pursuing a series of cases targeting public corruption and crime on wall street said on wednesday he has agreed to remain in his post after donald trump becomes us president bharara appointed to his position by democratic president barack obama in 2009 told reporters following a meeting with the republican presidentelect at trump tower in manhattan that trump asked him to stay on during his administration and he accepted trump takes office on jan 20 we had a good meeting bharara said i said i would absolutely consider staying on i agreed to stay on the announcements timing when trump has not yet finished filling all cabinetlevel positions was unusual but some former prosecutors who served under bharara said they were not surprised their former boss would be willing to remain as the us attorney for the southern district of new york i think preet is an independent law enforcementminded prosecutor who loves his job and is clearly talented in it said arlo devlinbrown a former chief of bhararas public corruption unit who is now a partner at the law firm covington burling us senator charles schumer of new york the incoming senate democratic leader who bharara previously worked for as chief counsel said trump called him last week to ask what he thought about bharara staying in his job i am glad they met and am glad preet is staying on schumer said in a statement hes been one of the best us attorneys new york has ever seen bhararas office has pursued an aggressive push against corruption in state and city politics an agenda that could fit with trumps vow to drain the swamp in washington those political investigations led last year to the convictions of former new york assembly speaker sheldon silver a democrat and former new york senate majority leader dean skelos a republican in separate corruption trials bharara also brought dozens of successful cases against insider traders and was on the cover of time magazine in 2012 with the headline this man is busting wall st those cases include the 2011 conviction of galleon group founder raj rajaratnam who is serving an 11year prison term and a 18 billion settlement and plea deal in 2013 with hedge fund sac capital advisors lp his 227lawyer office also secured corporate settlements with companies including general motors co and jpmorgan chase co won several convictions and guilty pleas of former employees of ponzi scheme operator bernard madoff and prosecuted suleiman abu ghaith a soninlaw of the late al qaeda leader osama bin laden bhararas offices priorities have often matched those set by obamas justice department amid an increase in civil rights investigations nationally for example bhararas office joined a lawsuit that led to a settlement in 2015 aimed at reducing violence in new york citys rikers island jail complex how priorities set by the justice department under trumps pick for attorney general republican senator jeff sessions of alabama affects the cases bhararas office pursues remains unclear obviously

there is likely be some changes in priorities from main justice the departments washington headquarters and the office will have to adjust to those said richard zabel who previously served as bhararas deputy before becoming hedge fund elliott managements general counsel bharara said his office had for the past seven years pursued its work independently without fear or favor former prosecutors they expect that to stay the same he would have only taken it on if he were 100 percent confident that that independence could be preserved said matthew schwartz partner at the law firm boies schiller flexner and a former prosecutor under bharara\",\n   \"how much more depraved does a socalled pastor have to get before people acknowledge that he is not a real christianbecause one would think that conservative christians of all people would be absolutely opposed to donald trump after his lewd comments about groping women came to light on fridaybut evangelical leaders are willing to abandon their own supposed values and continue supporting trump as long as it means hillary clinton does not become presidentfamily research council head tony perkins made that clear on saturday when he still voiced support for trump despite his grab them by the pssy remark my personal support for donald trump has never been based upon shared values it is based upon shared concerns about issues such as justices on the supreme court that ignore the constitution america s continued vulnerability to islamic terrorists and the systematic attack on religious liberty that we ve seen in the last 7 12 years perkins saidralph reed of faith and freedom coalition basically said the same thing voters of faith are voting on issues like who will protect unborn life defend religious freedom create jobs and oppose the iran nuclear deal tenyearold tapes of private conversation with a television talk show host rank very low on their hierarchy of concerns in other words trump could sexually assault all the women he wants and still be supported by conservative christian leadersand that includes 700 club preacher pat robertson who called trump s behavior macho while continuing to endorse him for president a guy does something 11 years ago it was a conversation in hollywood where he s trying to look like he s macho robertson said on monday and 11 years after that they surface it from the washington post or whatever bring it out within 30 days or so of the election and this is supposed to be the death blow and everybody writes him off okay he s dead now you ve got to get out of the way and let mike pence run the campaign robertson proceeded to declare trump the winner of sunday night s debate and called him a phoenix rising from the ashes of an imploded campaignhere s the video via youtubethese hypocrites have been whining for years about how america needs a president who shares their socalled biblical values but now they couldn t care less about those values as long as they can put a puppet in the white house who will rubber stamp their extreme religious agenda and force it upon the nation even if the candidate they want is totally unfit for the office these people are not christians they merely used religion to gain power and wealth and if conservatives actually had spines they would toss these pretenders out on their asses for disgracing the

churchfeatured image screenshot\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"no_double_spaces\",\n
\"properties\": {\n   \"dtype\": \"string\",\n
\"num_unique_values\": 38620,\n   \"samples\": [\n   \"if
the way donald trump reportedly ran the maralago resort in palm beach
is any indication if he wins he will be taking full advantage of every
spy agency in his charge just for his own agendaaccording to six
former employees at the resort trump had a telephone console in his
room that allowed him to tap into every single phone in the place he
often listened in on staff calls but he could also listen in on guest
calls although the gop frontrunner s spokeswoman denies itaccording to
buzzfeed who spoke to six former employees four of whom spoke under
the condition of anonymity because of nondisclosure agreements this is
indeed true those four said he listened in from the private room he
keeps during the mid2000sthey said he listened in on calls between
club employees or in some cases between staff and guests none of them
knew of trump eavesdropping on guests or members talking on private
calls with people who were not employees of maralago they also said
that trump could eavesdrop only on calls made on the club s landlines
and not on calls made from guests cell phoneseach of these four
sources said they personally saw the telephone console which some
referred to as a switchboard in trump s bedroomnone of the four
supports trump s bid for president all said they enjoyed their time
working at maralagothe two other employees did not wish to remain
anonymous they are trump s former butler the one who wants obama
killed and trump s former security director both of whom support trump
s run for president said that the console was only to make phone
callstrump the presumptive republican nominee is running at a time
when americans are increasingly concerned about surveillance both by
the government and by their employers some of his own campaign staff
feared that their offices in trump tower in new york might be bugged
the new york times reported last month trump has backed the nsa s bulk
collection of metadata telling conservative radio host hugh hewitt
that i tend to err on the side of security trump added i assume when i
pick up my telephone people are listening to my conversations anyway
if you want to know the truth if trump was listening in on guests
calls that was certainly a violation of florida law the law is less
clear on calls involving employees but since both parties in florida
need to be notified if the second party is not an employee that is
legally murky at bestthis accusation is frightening i have always been
of the opinion that trump is running for the job of dictator hence his
admiration of putin and kim jong un this does not bode well all i can
say is i hope hillary clinton s offices are locked down tight or we ll
have another watergate on our handsfeatured image via john moore at
getty images\",\n   \"a few weeks ago pepsi decided to try to
bridge the racial divide with what someone must have thought was a
great idea they ran an ad featuring kendall jenner who in the ad tried
to show police and africanamericans that their lives matter by letting

a cop drink pepsihere it isthe internet hated it which might have been the pointwithin 48 hours the video got nearly 16 million views on youtube five times as many downvotes as upvotes and twitter and facebook lit up with people pointing out just how gauche the whole thing was activist deray mckesson called it trash adding if i had carried pepsi i guess i never would ve gotten arrested who knew people made memes some even reaching back and evoking pepper spray cop and rightfully many folks pointed out that using protest imagery in order to peddle soda particularly images that evoked the photo of ieshia evans facing down police in baton rouge louisiana last year was pretty tasteless it was one of the few times the internet ever agreed on anythingsource wiredwhether in response to pepsi s ad that never should have been or whether this began organically heineken beer had their own version of an ad that addresses the political divide but they got it rightin an experiment they had complete strangers who would hate each other under normal circumstances and put them in situations that created bonding guess how it ends here it isit won the internetheineken just dropped an ad that actually brings people together pepsi take note openyourworld httpstcoe72swzqqp6 hayley jones meetmissjoness april 27 2017brilliant heineken httpstco9zae83adox pictwittercomgiwy1440gn navid mokhberi navidmg april 27 2017watch heineken school pepsi on how to advertise to gen z it s a lesson for every brand httpstcobhkn7rv8i3 pictwittercomoryr4q9otg denkyuu media denkyuumedia april 27 2017omg can we be a heineken school instead of a pepsi school stretched thin domznoriega april 27 2017well done heineken can we put a bar inside congress please httpstcobkqs3cwi94 ignacio contreras ignaciotechie april 27 2017it even got the attention of some celebritiesthis is brilliant this is what it s all about go heineken httpstcoedyp9opkmr sarah silverman sarahksilverman april 26 2017not everyone loved it thoughthat heineken ad never ok to out a trans person we could literally be assaulted murdered raped lose jobs etc bc you opened your mouth theblackdoriangray queeringpsych april 27 2017even with a few dissenters we found one the ad did what it set out to do bridge dividesfeatured image via screen capture from embedded video\"\n   ],\n   \"semantic_type\": \"\",\n \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"no_stopwords\",\n   \"properties\": {\n   \"dtype\": \"string\",\n   \"num_unique_values\": 38616,\n \"samples\": [\n   \"brussels reuters european council president donald tusk wednesday noted promising progress brexit talks said would propose 27 eu leaders open internal preparations second phase negotiations ties london bloc britain leaves march 2019\",\n \"weeks ago pepsi decided try bridge racial divide someone must thought great idea ran ad featuring kendall jenner ad tried show police africanamericans lives matter letting cop drink pepsihere isthe internet hated might pointwithin 48 hours video got nearly 16 million views youtube five times many downvotes upvotes twitter facebook lit people pointing gauche whole thing activist deray mckesson called trash adding carried pepsi guess never would gotten arrested knew

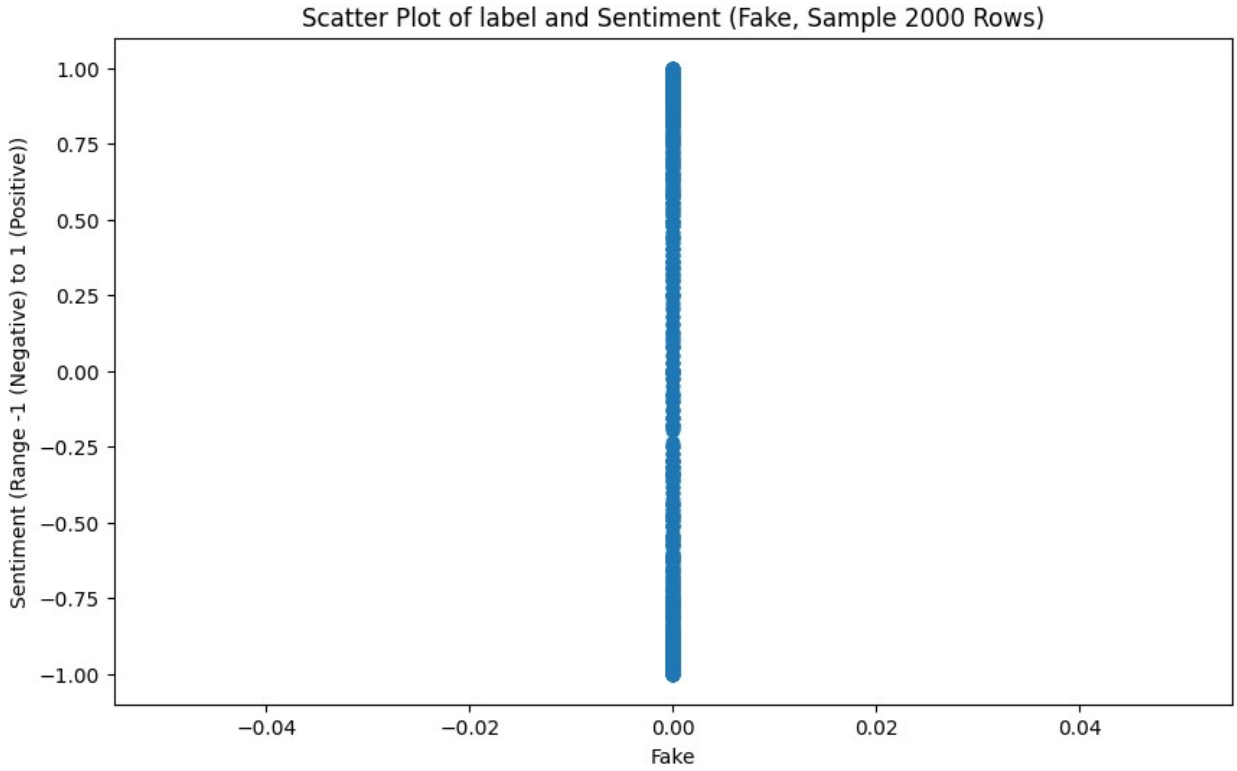
```
people made memes even reaching back evoking pepper spray cop
rightfully many folks pointed using protest imagery order peddle soda
particularly images evoked photo ieshia evans facing police baton
rouge louisiana last year pretty tasteless one times internet ever
agreed anythingsource wiredwhether response pepsi ad never whether
began organically heineken beer version ad addresses political divide
got rightin experiment complete strangers would hate normal
circumstances put situations created bonding guess ends isit
internetheineken dropped ad actually brings people together pepsi take
note openyourworld httpstcoe72swzqqp6 hayley jones meetmissjoness
april 27 2017brilliant heineken httpstco9zae83adox
pictwittercomgiwy1440gn navid mokhberi navidmg april 27 2017watch
heineken school pepsi advertise gen z lesson every brand
httpstcobhkn7rv8i3 pictwittercomoryr4q9otg denkyuu media denkyuumedia
april 27 2017omg heineken school instead pepsi school stretched thin
domznoriega april 27 2017well done heineken put bar inside congress
please httpstcobkqs3cwi94 ignacio contreras ignaciotechie april 27
2017it even got attention celebritiesthis brilliant go heineken
httpstcoedyp9opkmr sarah silverman sarahksilverman april 26 2017not
everyone loved thoughthat heineken ad never ok trans person could
literally assaulted murdered raped lose jobs etc bc opened mouth
theblackdoriangray queeringpsych april 27 2017even dissenters found
one ad set bridge dividesfeatured image via screen capture embedded
video\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"text_sentiment\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0.831801516422388,\n        \"min\": -
1.0,\n        \"max\": 0.9999,\n        \"num_unique_values\": 5367,\n
\"samples\": [\n          0.9029,\n          -0.8153\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"data"}
```


Interesting, values are between -1 for bad/sad news and +1 for good/happy news

Sentiment Analysis on Fake News

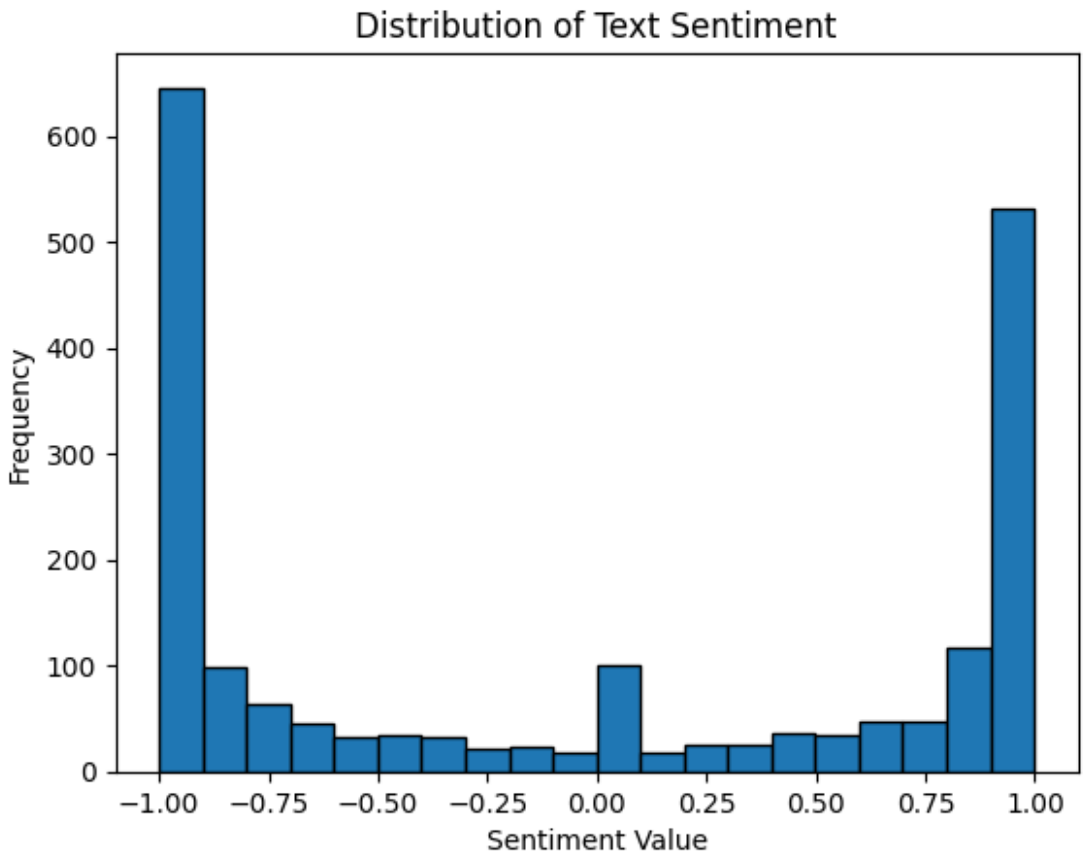
```
# Let's get a sample of Real news and test it against a Sentiment
analysis
df_filtered_fake = data[data['label'] == 0]

#Lets limit that to 2000 values, just to graph a sample
df_limited_fake = df_filtered_fake.sample(2000)

# Plotting
plt.figure(figsize=(10, 6))
plt.scatter(df_limited_fake['label'],
df_limited_fake['text_sentiment'])

# Adding titles and labels
plt.title('Scatter Plot of label and Sentiment (Fake, Sample 2000
```

```
Rows)')
plt.xlabel('Fake')
plt.ylabel('Sentiment (Range -1 (Negative) to 1 (Positive))')

# Show the plot
plt.show()
```



This seems well distributed, no actual skewness.

```
plt.hist(df_limited_fake['text_sentiment'], bins=20,
edgecolor='black')

plt.title('Distribution of Text Sentiment')
plt.xlabel('Sentiment Value')
plt.ylabel('Frequency')
plt.show()
```



Now we can see that the sample data leans more to the negative/sad/bad news, but not by much.

Let's examine the whole of the fake news.

```
# Calculate min, median, max, and mean (average) of Column2
min_val = df_filtered_fake['text_sentiment'].min()
median_val = df_filtered_fake['text_sentiment'].median()
max_val = df_filtered_fake['text_sentiment'].max()
mean_val = df_filtered_fake['text_sentiment'].mean()

(min_val, median_val, max_val, mean_val)
```

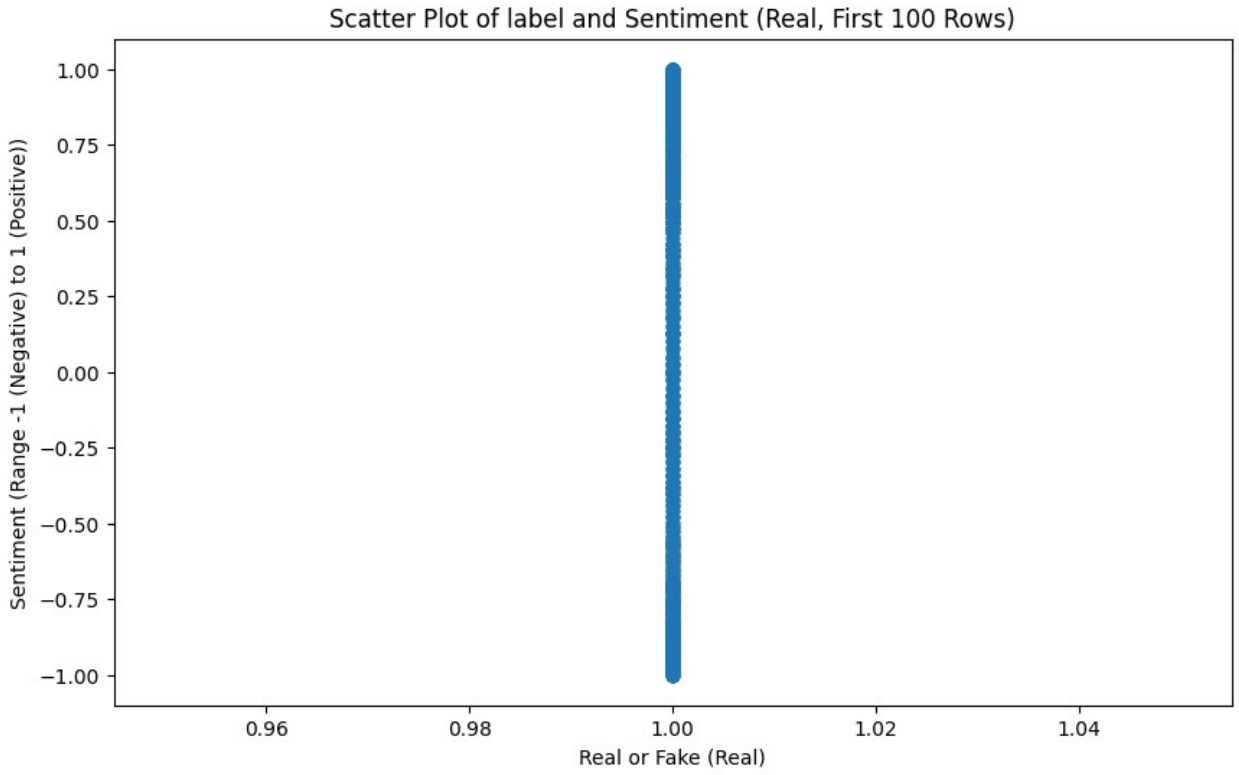
```
(-1.0, -0.0772, 0.9999, -0.05652798858651677)
```

This suggests that in this dataset, Fake News tends more to the negative side, but by a little bit.

Sentiment Analysis on Real News

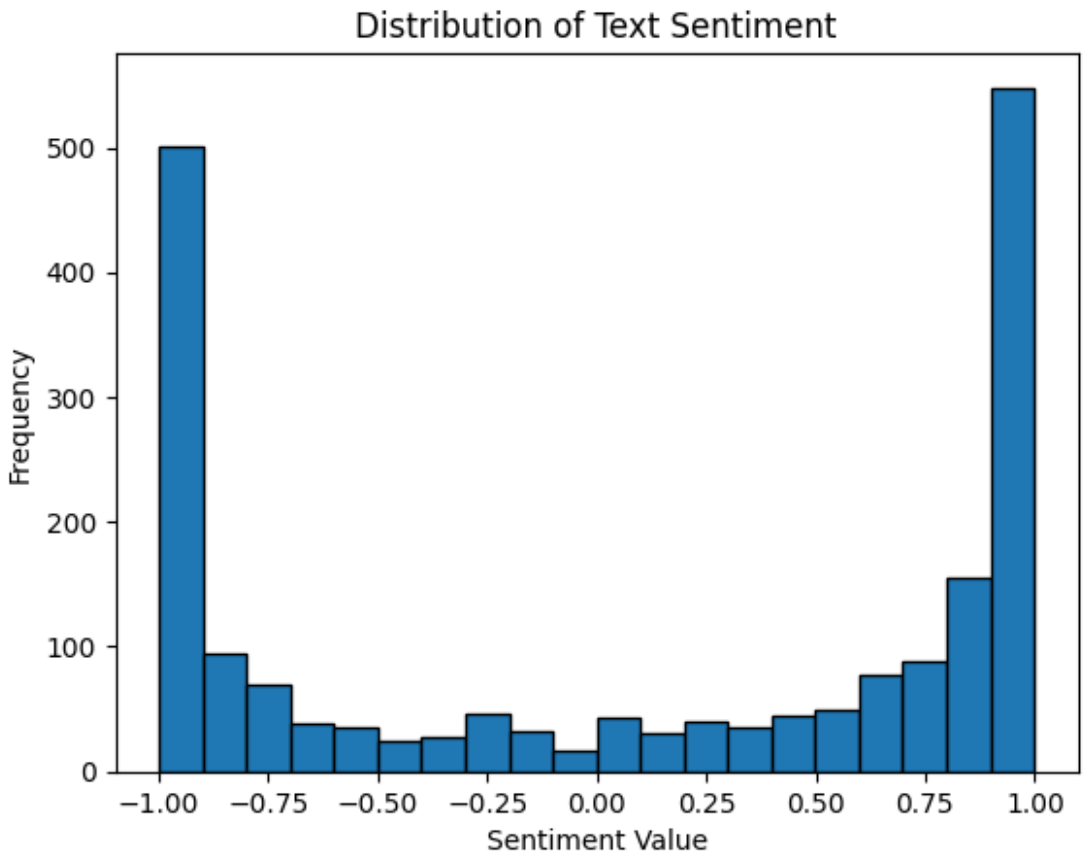
```
# Let's get a sample of Real news and test it against a Sentiment
analysis
df_filtered_real = data[data['label'] == 1]

#Lets limit that to 200 values, just to graph
```

```
df_limited_real = df_filtered_real.sample(2000)

# Plotting
plt.figure(figsize=(10, 6))
plt.scatter(df_limited_real['label'],
df_limited_real['text_sentiment'])

# Adding titles and labels
plt.title('Scatter Plot of label and Sentiment (Real, First 100
Rows)')
plt.xlabel('Real or Fake (Real)')
plt.ylabel('Sentiment (Range -1 (Negative) to 1 (Positive))')

# Show the plot
plt.show()
```



This seems well distributed too, no actual skewness.

```
plt.hist(df_limited_real['text_sentiment'], bins=20,
edgecolor='black')

plt.title('Distribution of Text Sentiment')
plt.xlabel('Sentiment Value')
plt.ylabel('Frequency')
plt.show()
```



Now we can see that the sample data leans more to the positive news, but not by much.

Let's examine whether Real News tends to be more positive overall.

```
# Calculate min, median, max, and mean (average) of Column2
min_val = df_filtered_real['text_sentiment'].min()
median_val = df_filtered_real['text_sentiment'].median()
max_val = df_filtered_real['text_sentiment'].max()
mean_val = df_filtered_real['text_sentiment'].mean()

(min_val, median_val, max_val, mean_val)
```

```
(-0.9997, 0.3182, 0.9999, 0.08324935798664612)
```

This indicates that in this dataset, Real News is generally more positive.

Sentiment Analysis Conclusion

Let's note the caveat that this observation is specific to this dataset and should not be used as a standard to compare real versus fake news across the entire news landscape. That being said, there is no distinct or big enough correlation between the positivity or negativity of content and its classification as real or fake news in this dataset. Real news looked more positive and fake news more negative, but the margins are really small. However, there is a slight trend: positive news leans towards being positive, while fake news tends to be more negative.

Adding Character Level Features to the dataset

```
#Taken and modified from: https://www.geeksforgeeks.org/count-uppercase-lowercase-special-character-numeric-values/

# Feature Extraction Functions

def count_characters(text):
    return len(text)

def count_digits(text):
    return sum(c.isdigit() for c in text)

def count_uppercase(text):
    return sum(c.isupper() for c in text)

# Applying the functions to the DataFrame
data['char_count'] = data['no_stopwords'].apply(count_characters)
data['digit_count'] = data['no_stopwords'].apply(count_digits)
data['uppercase_count'] = data['no_stopwords'].apply(count_uppercase)

# Display the char_count
print(data['char_count'])
```

```
0         171
1          89
2        1270
3         377
4        1342
         ...
44893    3216
44894    2675
44895    1804
44896    1913
44897    1137
Name: char_count, Length: 44898, dtype: int64
```

```
# Calculate min, median, max, and mean (average) of char_count
min_val = data['char_count'].min()
median_val = data['char_count'].median()
max_val = data['char_count'].max()
mean_val = data['char_count'].mean()

(min_val, median_val, max_val, mean_val)
```

```
(0, 1531.0, 38910, 1751.3913760078399)
```

```
# # Display the digit_count
print(data['digit_count'])
```

```
0          5
1          0
2         13
3          0
4          0
         ..
44893      6
44894     91
44895     29
44896     14
44897      2
Name: digit_count, Length: 44898, dtype: int64
```

```
# Calculate min, median, max, and mean (average) of Column2
min_val = data['digit_count'].min()
median_val = data['digit_count'].median()
max_val = data['digit_count'].max()
mean_val = data['digit_count'].mean()

(min_val, median_val, max_val, mean_val)
```

```
(0, 9.0, 1396, 16.14978395474186)
```

```
# Display the uppercase_count
print(data['uppercase_count'])
```

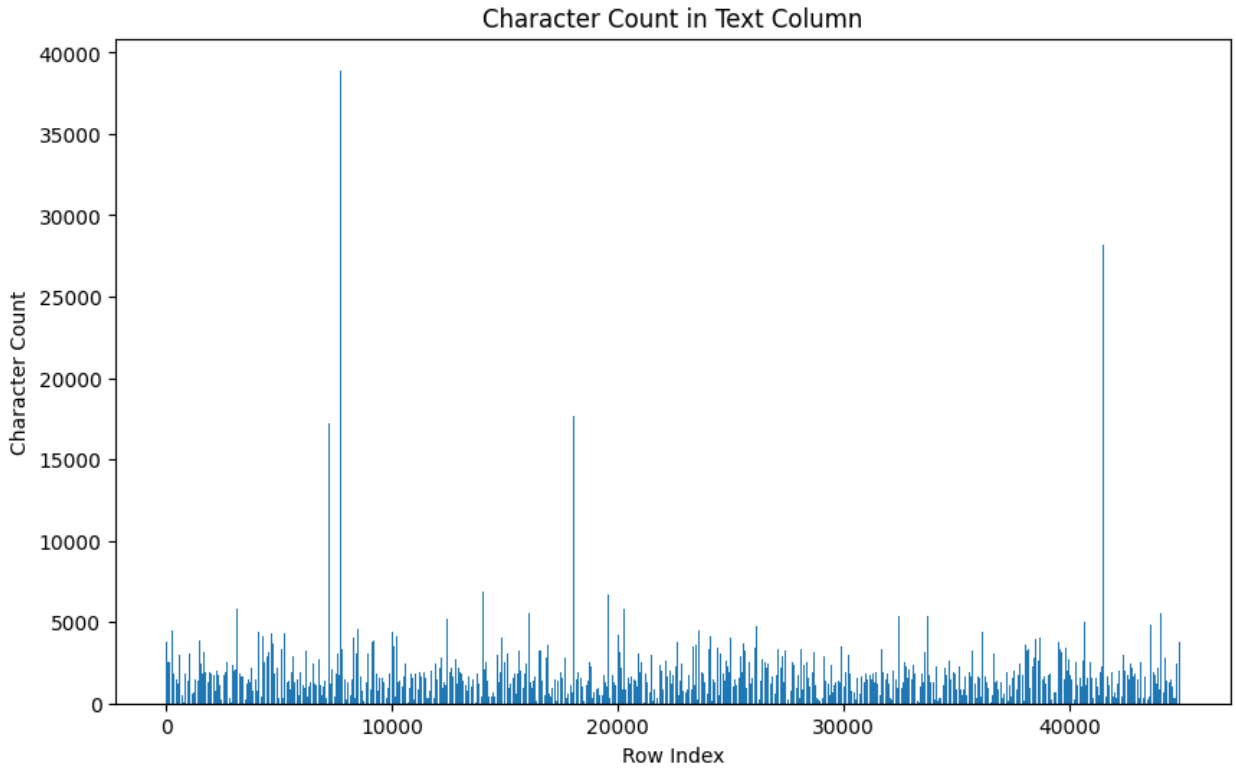
```
0         0
1         0
2         0
3         0
4         0
         ..
44893     0
44894     0
44895     0
44896     0
44897     0
Name: uppercase_count, Length: 44898, dtype: int64
```

```
# Calculate min, median, max, and mean (average) of Column2
min_val = data['uppercase_count'].min()
median_val = data['uppercase_count'].median()
max_val = data['uppercase_count'].max()
mean_val = data['uppercase_count'].mean()

(min_val, median_val, max_val, mean_val)
```

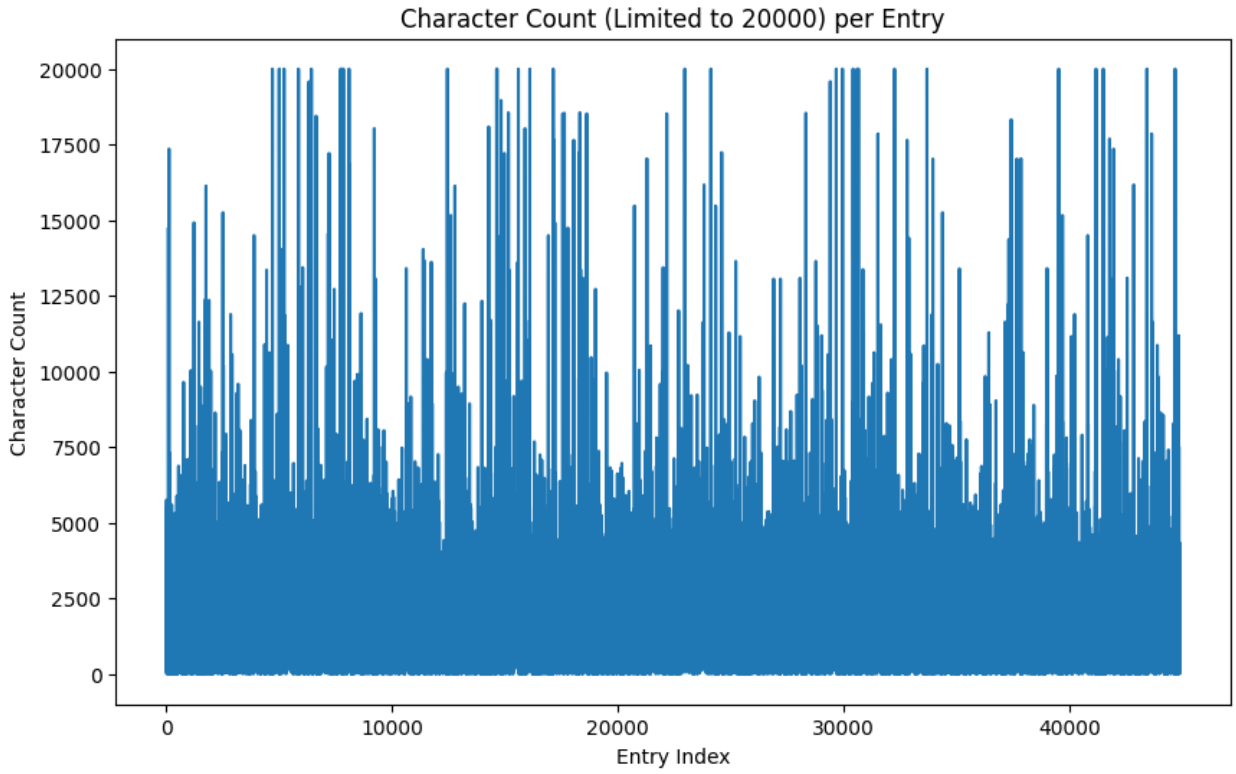
```
(0, 0.0, 0, 0.0)
```

```
# Plotting
plt.figure(figsize=(10,6))
```

```
plt.bar(data.index, data['char_count'])
plt.xlabel('Row Index')
plt.ylabel('Character Count')
plt.title('Character Count in Text Column')
plt.show()
```




```
# Limiting the character count to 20000
data['char_count_limited'] = data['char_count'].clip(upper=20000)

# Plotting
plt.figure(figsize=(10, 6))
plt.plot(data.index, data['char_count_limited'])
plt.title('Character Count (Limited to 20000) per Entry')
plt.xlabel('Entry Index')
plt.ylabel('Character Count')
plt.show()
```



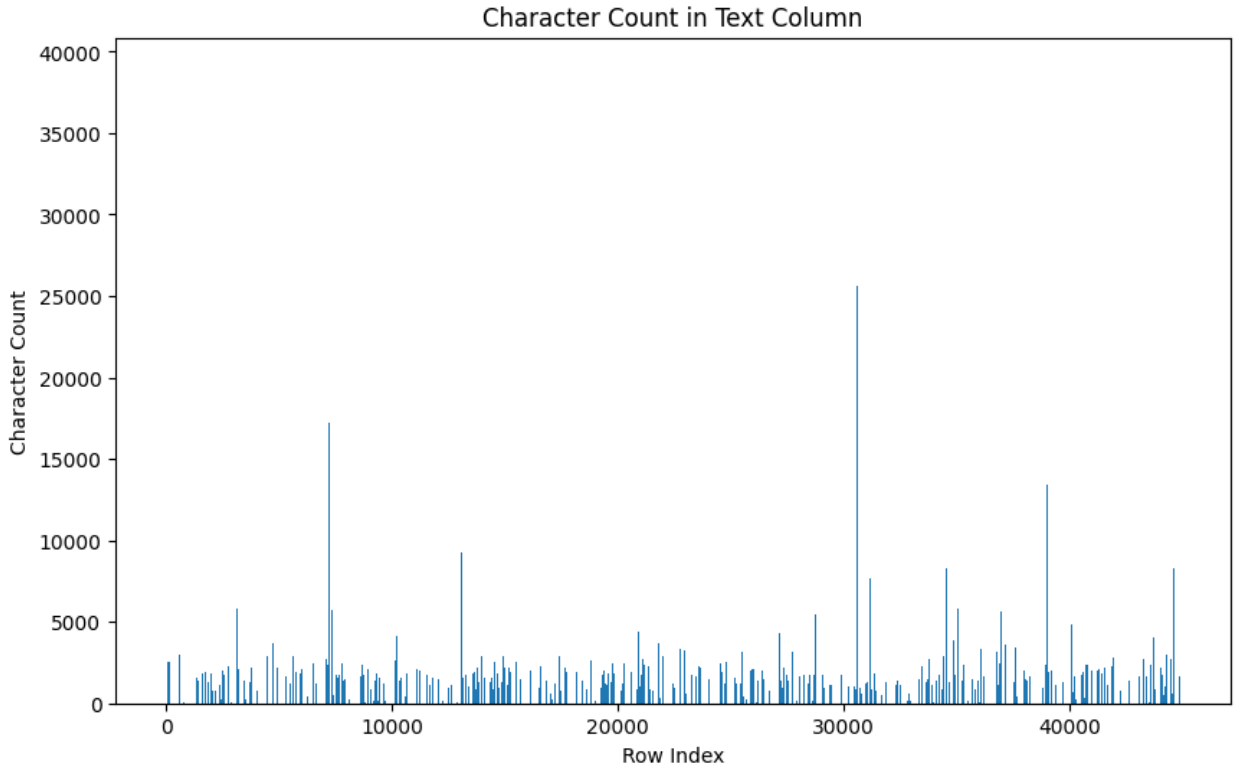
CLF on Fake news

```
# Taken and adapted from
https://stackoverflow.com/questions/55249360/count-the-number-of-
digits-in-a-dataframe-column

# Now just for Fake news
df_filtered_fake = data[data['label'] == 0]
# Applying the functions to the DataFrame
data['char_count'] =
df_filtered_fake['no_puntuation_text'].apply(count_characters)
data['digit_count'] =
df_filtered_fake['no_puntuation_text'].apply(count_digits)
data['uppercase_count'] =
df_filtered_fake['no_puntuation_text'].apply(count_uppercase)

# Display the DataFrame
print(df_filtered_fake['char_count'])
```

```
0          171
1           89
2         1270
3          377
4         1342
          ...
44888     1871
```

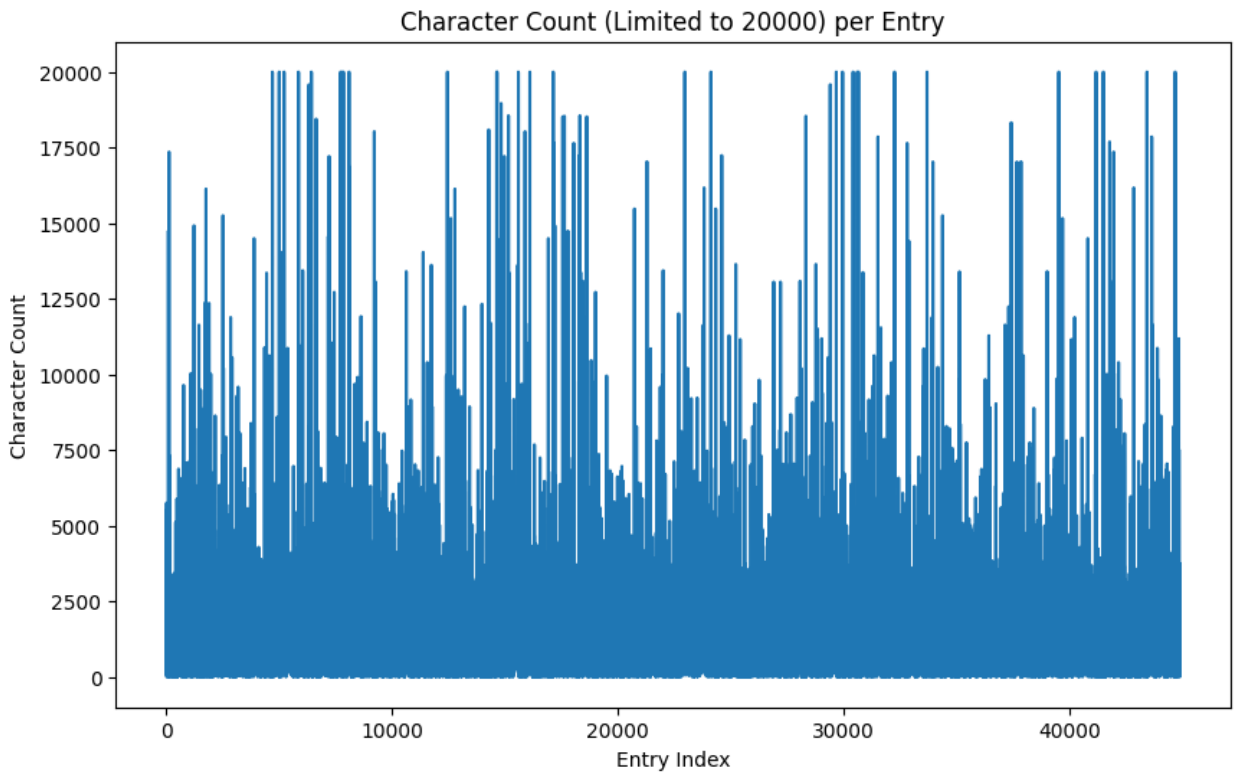
```
44889    3680
44890     960
44892    1376
44894    2675
Name: char_count, Length: 23481, dtype: int64
```

```
# Plotting
plt.figure(figsize=(10,6))
plt.bar(df_filtered_fake.index, df_filtered_fake['char_count'])
plt.xlabel('Row Index')
plt.ylabel('Character Count')
plt.title('Character Count in Text Column')
plt.show()
```




```
# Limiting the character count to 20000
df_filtered_fake['char_count_limited'] =
df_filtered_fake['char_count'].clip(upper=20000)

# Plotting
plt.figure(figsize=(10, 6))
plt.plot(df_filtered_fake.index,
df_filtered_fake['char_count_limited'])
plt.title('Character Count (Limited to 20000) per Entry')
plt.xlabel('Entry Index')
```

```
plt.ylabel('Character Count')
plt.show()
```



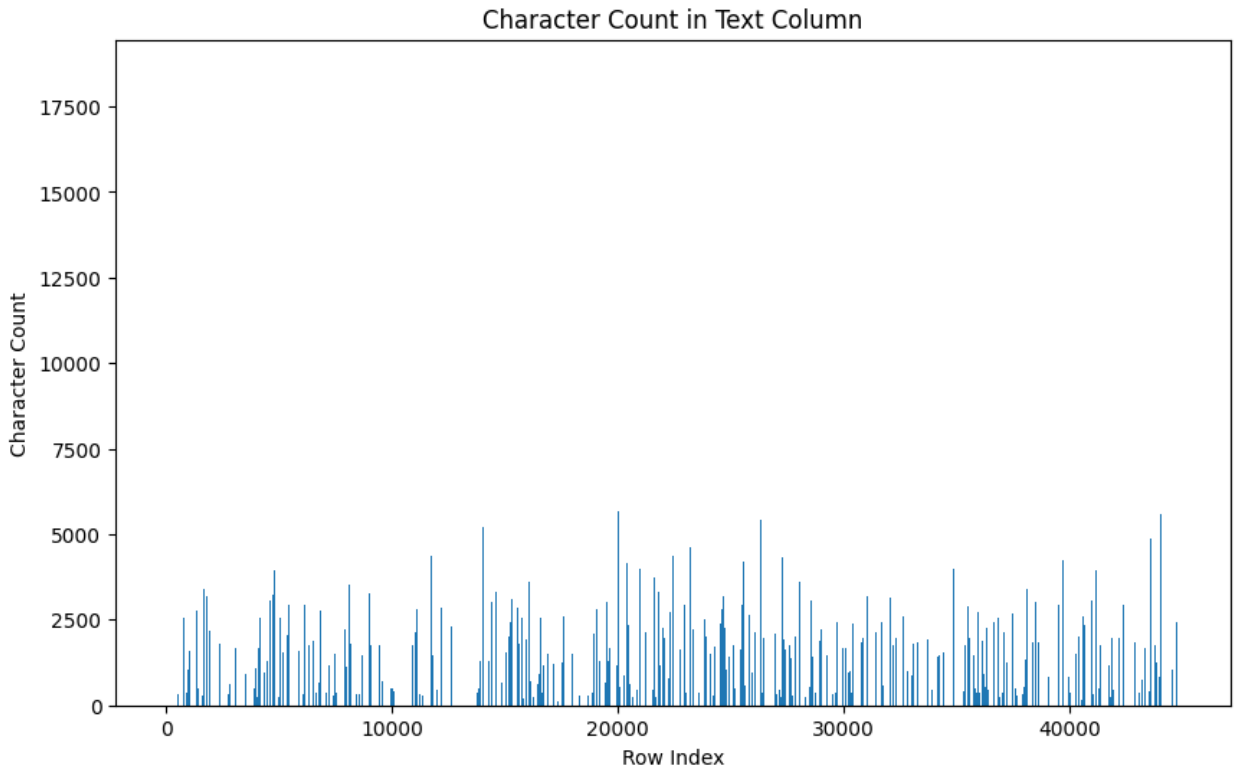
CLF on Real news

```
# Now just for Real news
df_filtered_true = data[data['label'] == 1]
# Applying the functions to the DataFrame
df_filtered_true['char_count'] =
df_filtered_true['no_stopwords'].apply(count_characters)
df_filtered_true['digit_count'] =
df_filtered_true['no_stopwords'].apply(count_digits)
df_filtered_true['uppercase_count'] =
df_filtered_true['no_stopwords'].apply(count_uppercase)

# Display the DataFrame
print(df_filtered_true['char_count'])
```

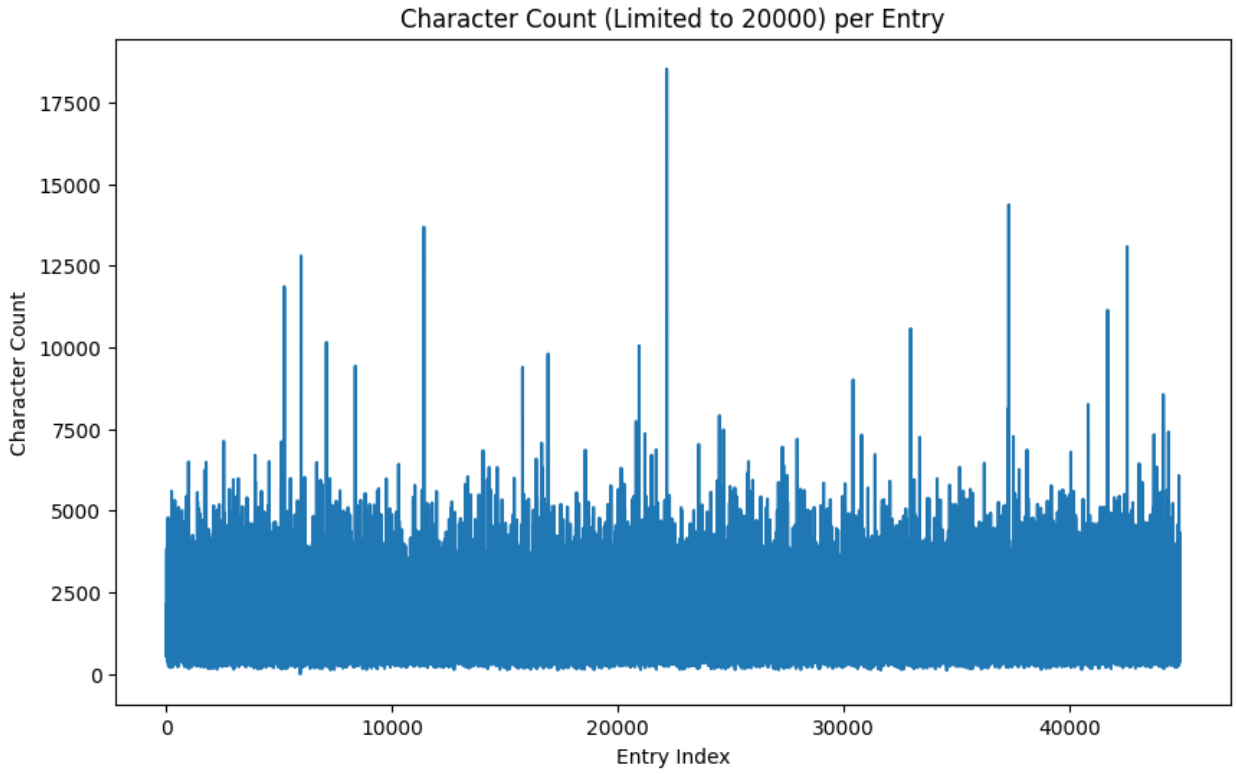
```
7         3796
10        1854
11        2416
12        2233
16         526
          ...
44891     3311
44893     3216
```

```
44895    1804
44896    1913
44897    1137
Name: char_count, Length: 21417, dtype: int64
```

```
# Plotting
plt.figure(figsize=(10,6))
plt.bar(df_filtered_true.index, df_filtered_true['char_count'])
plt.xlabel('Row Index')
plt.ylabel('Character Count')
plt.title('Character Count in Text Column')
plt.show()
```




```
# Limiting the character count to 20000
df_filtered_true['char_count_limited'] =
df_filtered_true['char_count'].clip(upper=20000)

# Plotting
plt.figure(figsize=(10, 6))
plt.plot(df_filtered_true.index,
df_filtered_true['char_count_limited'])
plt.title('Character Count (Limited to 20000) per Entry')
plt.xlabel('Entry Index')
plt.ylabel('Character Count')
plt.show()
```



CLF Conclusion

In this dataset, it seems that Fake News exhibits a more even distribution of character counts.

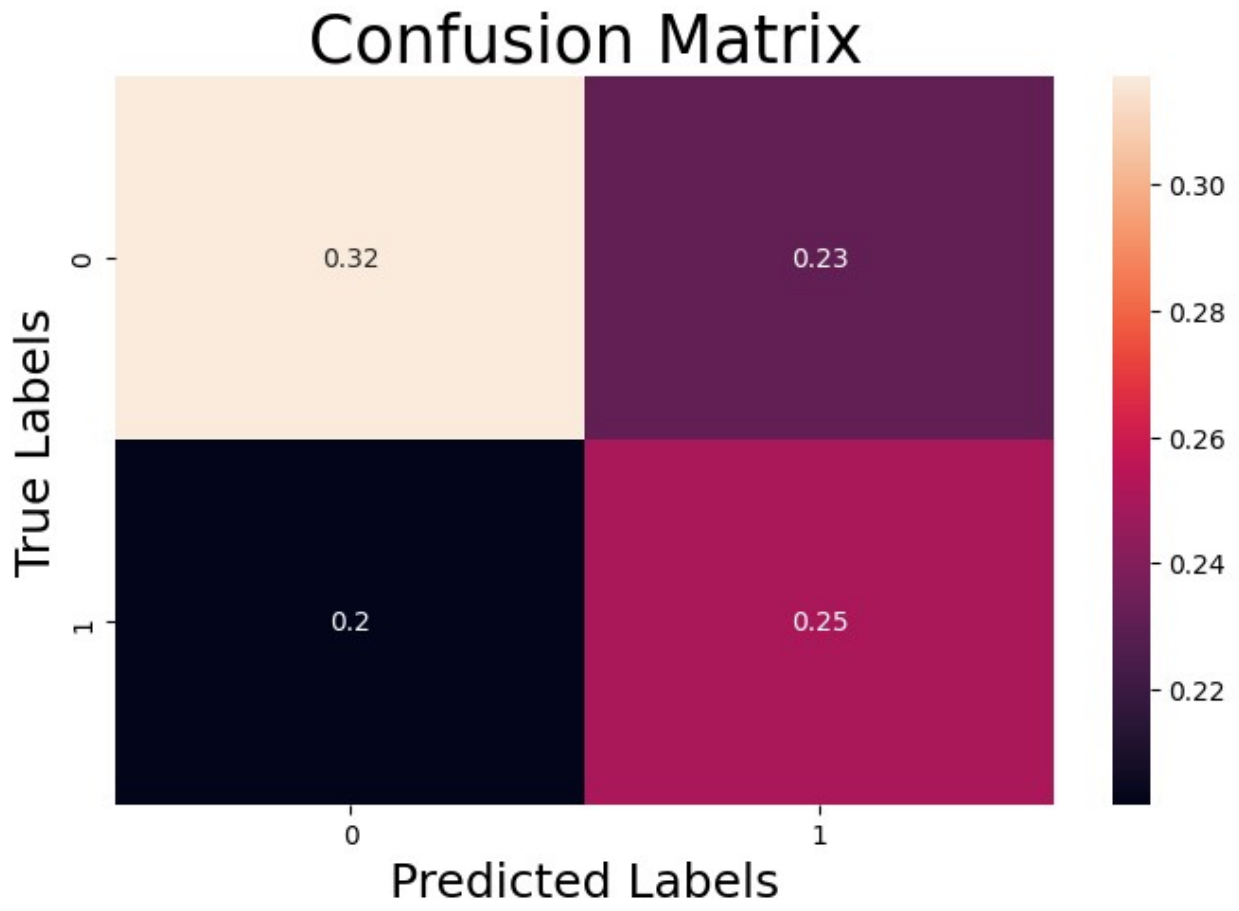
NLP Model Training

Bayes on raw data

```
# Taken and adapted from: https://www.geeksforgeeks.org/multinomial-naive-bayes/

# Let's get everything ready for using the multinomial Naive Bayes classifier
# (Bayes for text)
# Features are the text of the news
features_raw = data['text']
# Targets are the values 0 for fake and 1 for true
targets_raw = data['label']

# Let's split the data into train and test datasets
X_train, X_test, y_train, y_test = train_test_split(features_raw,
targets_raw,
                                                    test_size=0.30,
                                                    random_state=13)
```

```
# Since we are want to predict using words, rather than numbers, we
need to
# limit the word vocabulary and we need to set the tokenizer to limit
the max
# amount of the vocabulary
max_vocabulary = 10000
tokenizer = Tokenizer(num_words=max_vocabulary)
tokenizer.fit_on_texts(X_train)

# Now let's use the tokenizer to turn text into lists
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)

# This pads the sequences to make them the same length, we need this
for processing the data
X_train = tf.keras.preprocessing.sequence.pad_sequences(X_train,
padding='post', maxlen=256)
X_test = tf.keras.preprocessing.sequence.pad_sequences(X_test,
padding='post', maxlen=256)

# Creating the model, in this case Bayes
bayes_model = MultinomialNB()

# Fitting the model with the train data
bayes_model.fit(X_train, y_train)

# Predicting
predicted_bayes_model = bayes_model.predict(X_test)

# Checking values
print("The model Accuracy is: ", accuracy_score(predicted_bayes_model,
y_test))
print("The model F1 is: ", f1_score(predicted_bayes_model, y_test))
print("The model Precision is: ",
precision_score(predicted_bayes_model, y_test))
print("The model Recall is: ", recall_score(predicted_bayes_model,
y_test))

# Now let's try to graph that model

# This is where I'll store the predictions
binary_predictions = []

# Let's place the predicted values into buckets and then into my
storage
for i in predicted_bayes_model:
    # If it was higher/equal to 0.5, it's a 1 (true)
    if i >= 0.5:
        binary_predictions.append(1)
```

```
    # It's false
    else:
        binary_predictions.append(0)

# Let's fill out the confusion matrix with values
matrix_of_confusion = confusion_matrix(binary_predictions, y_test,
normalize='all')

# Let's plot it (standard 8 by 5)
plt.figure(figsize=(8, 5))

matrix_graph = plt.subplot()
# Adding a heat map
sns.heatmap(matrix_of_confusion, annot=True, ax = matrix_graph)
# Labeling and prettying it up
matrix_graph.set_xlabel('Predicted Labels', size=18)
matrix_graph.set_ylabel('True Labels', size=18)
matrix_graph.set_title('Confusion Matrix', size=25)
matrix_graph.xaxis.set_ticklabels([0,1], size=10)
matrix_graph.yaxis.set_ticklabels([0,1], size=10)


The model Accuracy is:  0.5672605790645879
The model F1 is:  0.5360923199363311
The model Precision is:  0.519833307609199
The model Recall is:  0.5534012487676635

[Text(0, 0.5, '0'), Text(0, 1.5, '1')]
```



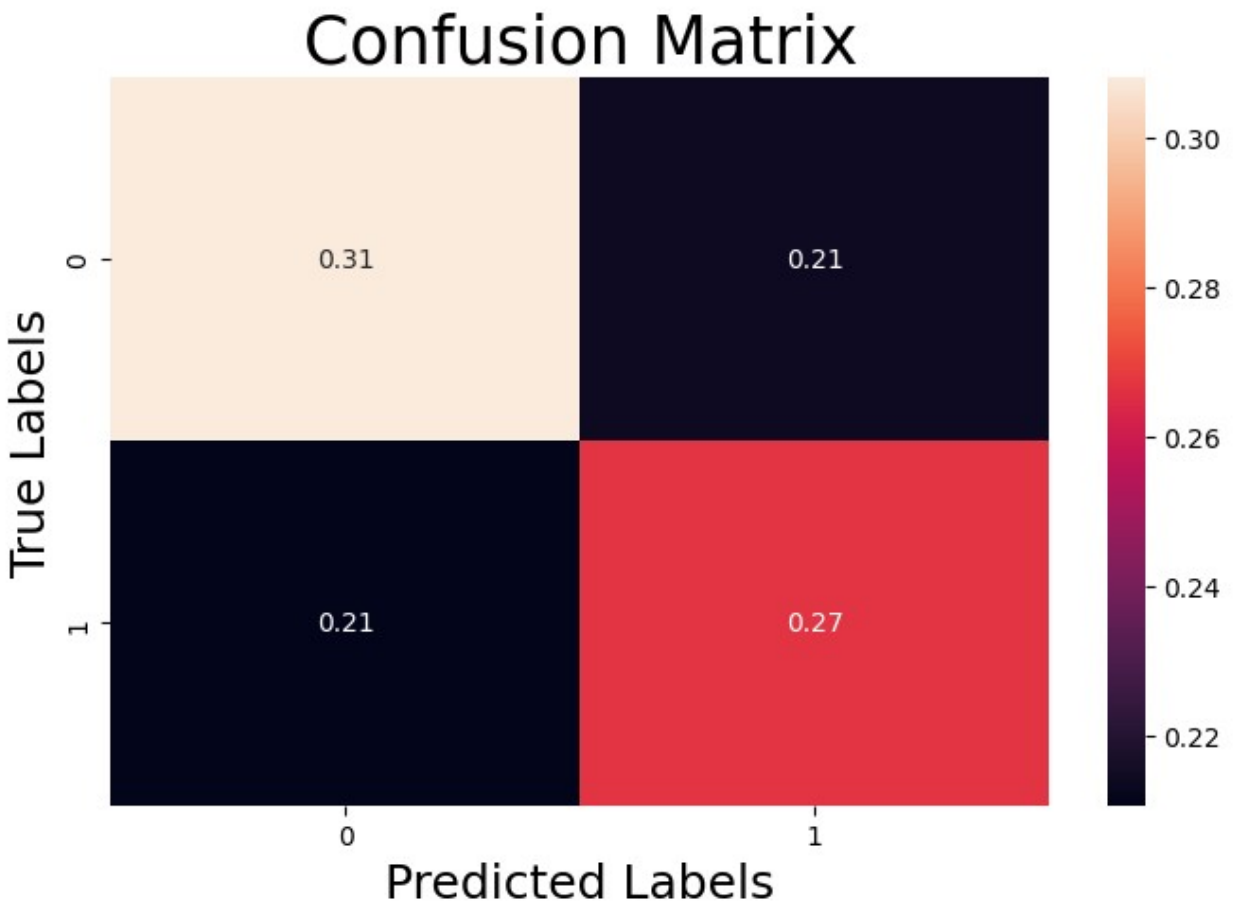
Bayes on cleaned data

```
# Let's get everything ready for using the multinomial Naive Bayes
classifier
# (Bayes for text)
# Features are the text of the news
features = data['no_puntuation_text']
# Targets are the values 0 for fake and 1 for true
targets = data['label']

# Let's split the data into train and test datasets
X_train, X_test, y_train, y_test = train_test_split(features, targets,
                                                    test_size=0.30,
                                                    random_state=13)

# Since we are want to predict using words, rather than numbers, we
need to
# limit the word vocabulary
# and we need to set the tokenizer to limit the max amount of the
vocabulary
max_vocabulary = 10000
tokenizer = Tokenizer(num_words=max_vocabulary)
```

```
tokenizer.fit_on_texts(X_train)

# Now let's use the tokenizer to turn text into lists
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)

# This pads the sequences to make them the same length, we need this
for
# processing the data
X_train = tf.keras.preprocessing.sequence.pad_sequences(X_train,
padding='post',
                                                        maxlen=256)
X_test = tf.keras.preprocessing.sequence.pad_sequences(X_test,
padding='post',
                                                       maxlen=256)

# Creating the model, in this case Bayes
bayes_model = MultinomialNB()

# Fitting the model with the train data
bayes_model.fit(X_train, y_train)

# Predicting
predicted_bayes_model = bayes_model.predict(X_test)

# Checking values
print("The model Accuracy is: ", accuracy_score(predicted_bayes_model,
y_test))
print("The model F1 is: ", f1_score(predicted_bayes_model, y_test))
print("The model Precision is: ",
precision_score(predicted_bayes_model, y_test))
print("The model Recall is: ", recall_score(predicted_bayes_model,
y_test))

# Now let's try to graph that model

# This is where I'll store the predictions
binary_predictions = []

# Let's place the predicted values into buckets and then into my
storage
for i in predicted_bayes_model:
    # If it was higher/equal to 0.5, it's a 1 (true)
    if i >= 0.5:
        binary_predictions.append(1)
    # It's false
    else:
        binary_predictions.append(0)

# Let's fill out the confusion matrix with values
```

```
matrix_of_confusion = confusion_matrix(binary_predictions, y_test,
                                       normalize='all')

# Let's plot it (standard 8 by 5)
plt.figure(figsize=(8, 5))

matrix_graph = plt.subplot()
# Adding a heat map
sns.heatmap(matrix_of_confusion, annot=True, ax = matrix_graph)
# Labeling and prettying it up
matrix_graph.set_xlabel('Predicted Labels', size=18)
matrix_graph.set_ylabel('True Labels', size=18)
matrix_graph.set_title('Confusion Matrix', size=25)
matrix_graph.xaxis.set_ticklabels([0,1], size=10)
matrix_graph.yaxis.set_ticklabels([0,1], size=10)
```

```
The model Accuracy is:  0.5752041573867854
The model F1 is:  0.5569149759950442
The model Precision is:  0.5550239234449761
The model Recall is:  0.5588189588189588

[Text(0, 0.5, '0'), Text(0, 1.5, '1')]
```



Conclusion on Bayesian Analysis

It appears that there is a slight improvement in the accuracy of True predictions when using data without punctuation, while the rest of the results remain largely unchanged.

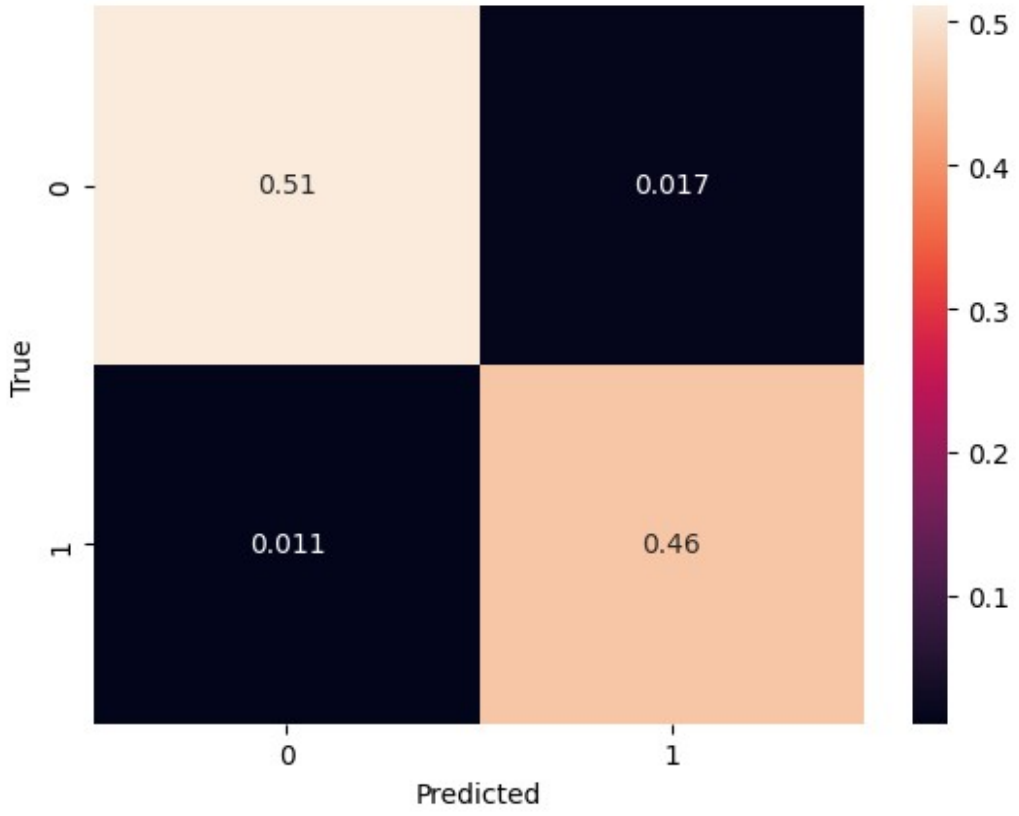
SimpleRNN

On raw data

```
# Taken and adapted from:
https://subscription.packtpub.com/book/data/9781788292061/7/ch07lvl1sec57/simple-rnn-with-keras

# Features are the text of the news
features_raw = data['text']
# Targets are the labels
targets_raw = data['label']

# Split the data into train and test datasets
X_train, X_test, y_train, y_test = train_test_split(features_raw,
targets_raw, test_size=0.30)

# Basic tokenization and conversion to sequences, limit to 100 words
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X_train[:100])
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)

# Basic padding of sequences
X_train = pad_sequences(X_train, maxlen=100)
X_test = pad_sequences(X_test, maxlen=100)

# Create a model
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(len(tokenizer.word_index) + 1, 32),
    tf.keras.layers.SimpleRNN(10),
    tf.keras.layers.Dense(1, activation='sigmoid')
])

# Compile and fit the model
model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
model.fit(X_train, y_train, epochs=1, batch_size=128,
validation_split=0.1)

# Evaluate the model
model.evaluate(X_test, y_test)

# Predict and generate binary predictions
predicted = model.predict(X_test)
```

```
binary_predictions = [1 if x >= 0.5 else 0 for x in predicted]

# Generate a confusion matrix and plot it
conf_matrix = confusion_matrix(y_test, binary_predictions,
normalize='all')
sns.heatmap(conf_matrix, annot=True)
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
221/221 [==============================] - 45s 189ms/step - loss:
0.2845 - accuracy: 0.9256 - val_loss: 0.1161 - val_accuracy: 0.9758
421/421 [==============================] - 4s 11ms/step - loss: 0.1209
- accuracy: 0.9722
421/421 [==============================] - 5s 12ms/step
```



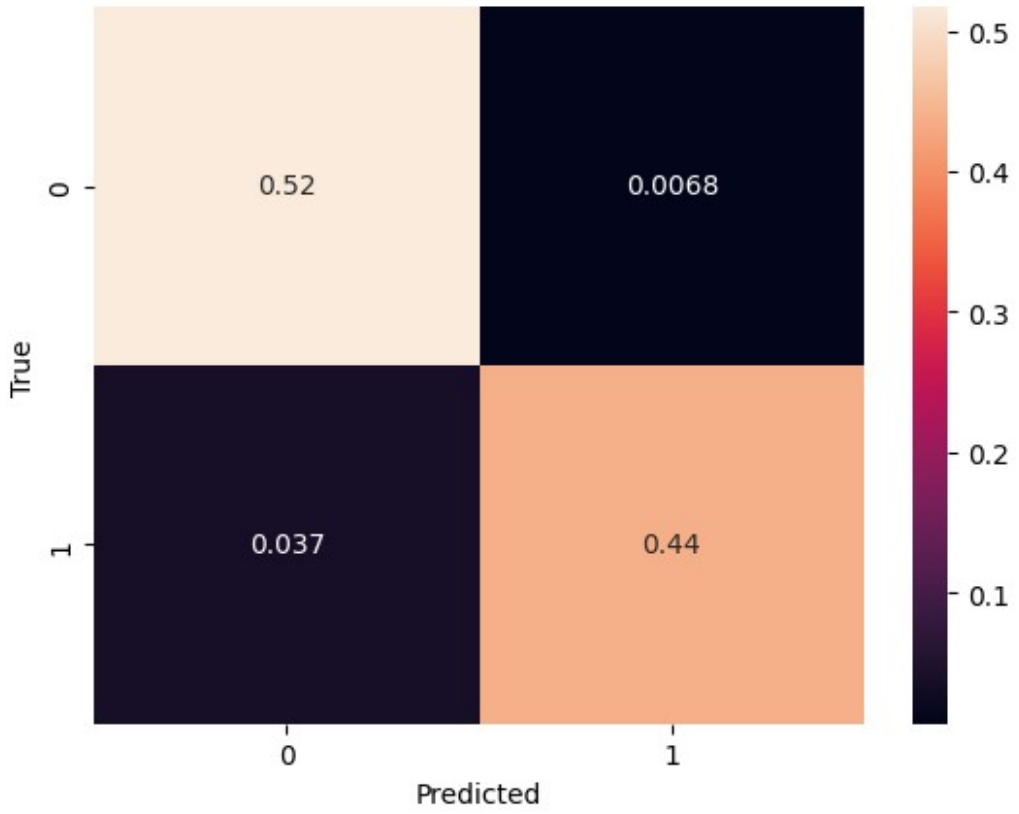

```
421/421 [==============================] - 6s 13ms/step - loss: 0.1209
- accuracy: 0.9722
Test Accuracy: 97.22%
```

On cleaned data

```
# Features are the text of the news
features_clean = data['no_puntuation_text']
# Targets are labels
targets_clean = data['label']

# Split the data into train and test datasets
X_train, X_test, y_train, y_test = train_test_split(features_clean,
targets_clean, test_size=0.30)

# Basic tokenization and conversion to sequences, limit to 100 words
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X_train[:100])
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)

# Basic padding of sequences
X_train = pad_sequences(X_train, maxlen=100)
X_test = pad_sequences(X_test, maxlen=100)

# Create a model
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(len(tokenizer.word_index) + 1, 32),
    tf.keras.layers.SimpleRNN(10),
    tf.keras.layers.Dense(1, activation='sigmoid')
])

# Compile and fit the model
model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
model.fit(X_train, y_train, epochs=1, batch_size=128,
validation_split=0.1)

# Evaluate the model
model.evaluate(X_test, y_test)

# Predict and generate binary predictions
predicted = model.predict(X_test)
binary_predictions = [1 if x >= 0.5 else 0 for x in predicted]

# Generate a confusion matrix and plot it
conf_matrix = confusion_matrix(y_test, binary_predictions,
normalize='all')
sns.heatmap(conf_matrix, annot=True)
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()

# Evaluate the model
```

```
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
221/221 [==============================] - 44s 192ms/step - loss:
0.2959 - accuracy: 0.8957 - val_loss: 0.1538 - val_accuracy: 0.9570
421/421 [==============================] - 6s 13ms/step - loss: 0.1542
- accuracy: 0.9564
421/421 [==============================] - 4s 10ms/step
```




```
421/421 [==============================] - 5s 13ms/step - loss: 0.1542
- accuracy: 0.9564
Test Accuracy: 95.64%
```

Conclusion on SimpleRNN

On raw data, we get an accuracy of: 97.22%

On cleaned data, we get an accuracy of: 95.64%

In SimpleRNN cleaning the data does affect the accurary of the trained models, in the wrong way!

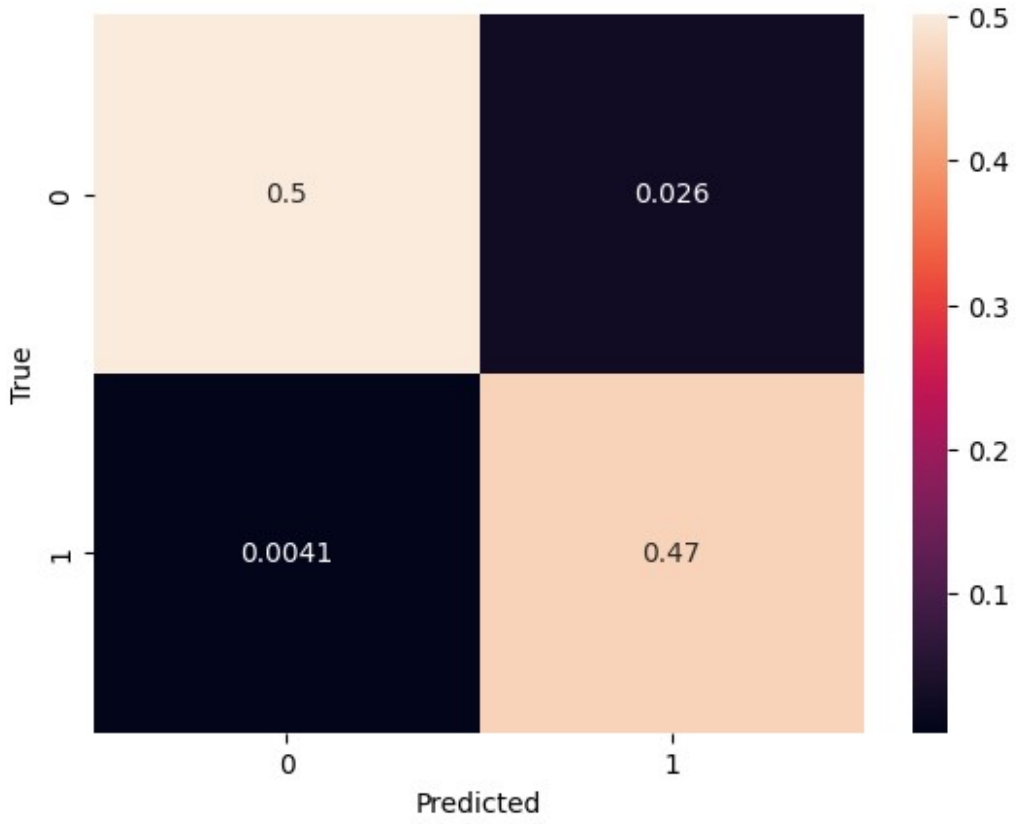
GRU

On raw data

```
# Taken and adapted from:
https://keras.io/api/layers/recurrent_layers/gru/

# Features are the text of the news
features_raw = data['text']
# Targets are the labels
targets_raw = data['label']

# Split the data into train and test datasets (simpler split)
X_train, X_test, y_train, y_test = train_test_split(features_raw,
targets_raw, test_size=0.30)

# Basic tokenization and conversion to sequences, limit to 100 words
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X_train[:100])
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)

# Basic padding of sequences
X_train = pad_sequences(X_train, maxlen=100)
X_test = pad_sequences(X_test, maxlen=100)

# Create a model
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(len(tokenizer.word_index) + 1, 32),
    tf.keras.layers.GRU(16),
    tf.keras.layers.Dense(1, activation='sigmoid')
])

# Compile and fit the model
model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
model.fit(X_train, y_train, epochs=1, batch_size=128,
validation_split=0.1)

# Evaluate the model
model.evaluate(X_test, y_test)

# Predict and generate binary predictions
predicted = model.predict(X_test)
binary_predictions = [1 if x >= 0.5 else 0 for x in predicted]

# Generate a confusion matrix and plot it
conf_matrix = confusion_matrix(y_test, binary_predictions,
normalize='all')
sns.heatmap(conf_matrix, annot=True)
```

```
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
221/221 [==============================] - 26s 103ms/step - loss:
0.2590 - accuracy: 0.9221 - val_loss: 0.0928 - val_accuracy: 0.9691
421/421 [==============================] - 2s 4ms/step - loss: 0.0890
- accuracy: 0.9699
421/421 [==============================] - 2s 3ms/step
```




```
421/421 [==============================] - 2s 4ms/step - loss: 0.0890
- accuracy: 0.9699
Test Accuracy: 96.99%
```

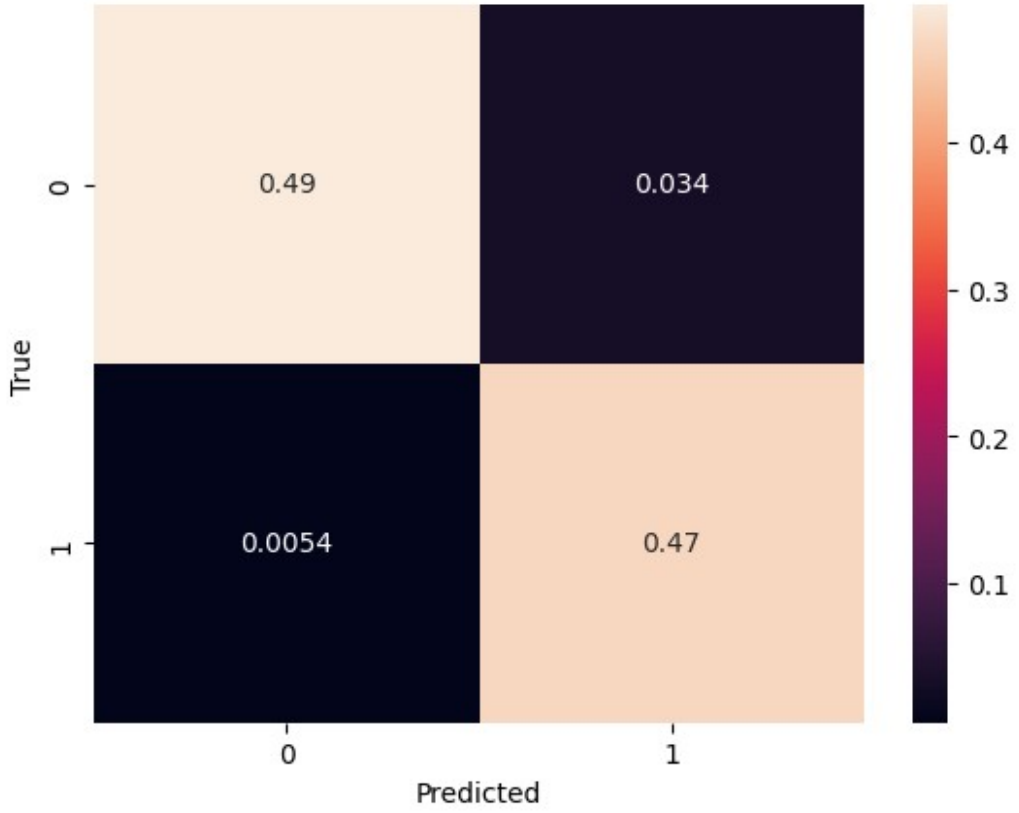
On cleaned data

```
# Features are the text of the news
features_clean = data['no_puntuation_text']
# Targets are the labels
```

```
targets_clean = data['label']

# Split the data into train and test datasets (simpler split)
X_train, X_test, y_train, y_test = train_test_split(features_clean,
targets_clean, test_size=0.30)

# Basic tokenization and conversion to sequences, limit to 100 words
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X_train[:100])
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)

# Basic padding of sequences
X_train = pad_sequences(X_train, maxlen=100)
X_test = pad_sequences(X_test, maxlen=100)

# Create a model
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(len(tokenizer.word_index) + 1, 32),
    tf.keras.layers.GRU(16),
    tf.keras.layers.Dense(1, activation='sigmoid')
])

# Compile and fit the model
model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
model.fit(X_train, y_train, epochs=1, batch_size=128,
validation_split=0.1)

# Evaluate the model
model.evaluate(X_test, y_test)

# Predict and generate binary predictions
predicted = model.predict(X_test)
binary_predictions = [1 if x >= 0.5 else 0 for x in predicted]

# Generate a confusion matrix and plot it
conf_matrix = confusion_matrix(y_test, binary_predictions,
normalize='all')
sns.heatmap(conf_matrix, annot=True)
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
221/221 [==============================] - 29s 123ms/step - loss:
0.2903 - accuracy: 0.9007 - val_loss: 0.1320 - val_accuracy: 0.9513
```

```
421/421 [==============================] - 2s 5ms/step - loss: 0.1163
- accuracy: 0.9606
421/421 [==============================] - 2s 3ms/step
```




```
421/421 [==============================] - 2s 4ms/step - loss: 0.1163
- accuracy: 0.9606
Test Accuracy: 96.06%
```

Conclusion on GRU

On raw data, we get an accuracy of: 96.99%

On cleaned data, we get an accuracy of: 96.06%

In GRU cleaning the data affect the accurary, it is more accurate to use the raw data!

LSTM

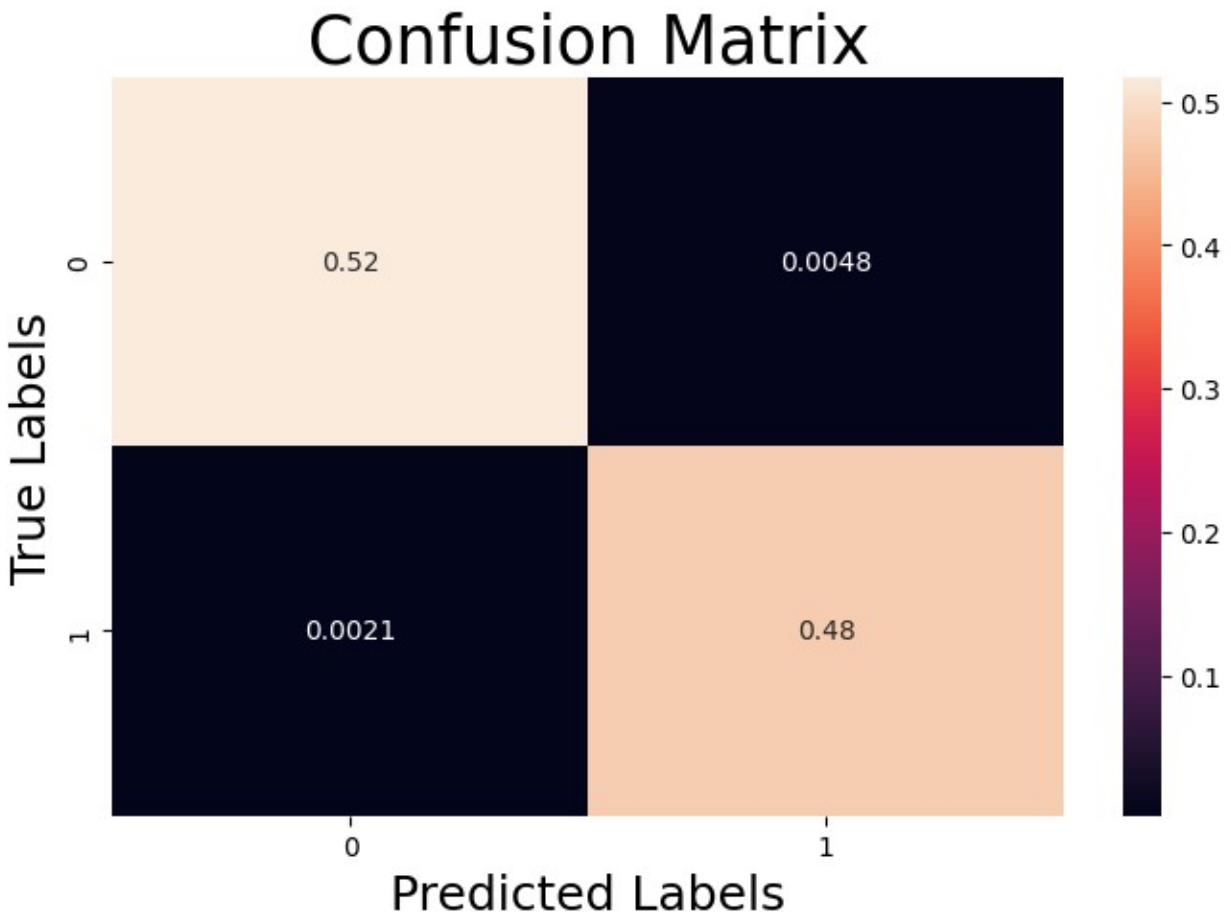
On raw data

```
# Taken and adapted from:
https://keras.io/examples/nlp/bidirectional_lstm_imdb/

# features are the text of the news
```

```
features_raw = data['text']
# targets are labels
targets_raw = data['label']

# Let's split the data into train and test datasets
X_train, X_test, y_train, y_test = train_test_split(features_raw,
targets_raw,
                                                    test_size=0.30,
random_state=7)

# Since we want to predict using words rather than numbers, we need to
limit the word vocabulary
# and we need to set the tokenizer to limit the max amount of the
vocabulary
max_vocabulary = 10000
tokenizer = Tokenizer(num_words=max_vocabulary)
tokenizer.fit_on_texts(X_train)

# Now, let's use the tokenizer to turn text into lists
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)

# This pads the sequences to make them the same length; we need this
for processing the data
X_train = tf.keras.preprocessing.sequence.pad_sequences(X_train,
padding='post',
                                                        maxlen=256)
X_test = tf.keras.preprocessing.sequence.pad_sequences(X_test,
padding='post',
                                                       maxlen=256)

# Let's create the model now!
model = tf.keras.Sequential([
    # This is the embedding layer, to convert into dense vectors for
better input
    tf.keras.layers.Embedding(max_vocabulary, 128),
    # This will process the input in both directions; it helps with
work context
    tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(16)),
    # This is technical, but it applies a non-linear transformation to
the inputs
    # using the ReLU activation function
    tf.keras.layers.Dense(64, activation='relu'),
    # This prevents overfitting by setting a fraction of the inputs as
0
    tf.keras.layers.Dropout(0.5),
    # This predicts the final value
    tf.keras.layers.Dense(1)
])
```

```
# Let's print a summary of the model
model.summary()

# Let's compile the model
model.compile(loss=tf.keras.losses.BinaryCrossentropy(from_logits=True
),
              optimizer=tf.keras.optimizers.Adam(1e-3),
              metrics=['accuracy'])

# Let's fit the model
model_fitter = model.fit(X_train, y_train, epochs=1,
validation_split=0.1,
                         batch_size=30, shuffle=True)


# Let's evaluate it
model.evaluate(X_test, y_test)

# Lets predict something
predicted_LSTM_model = model.predict(X_test)

# Now let's try to graph that model

# This is where I'll store the predictions
binary_predictions = []

# Let's place the predicted values into buckets and then into my
storage
for i in predicted_LSTM_model:
    # If it was higher/equal to 0.5, it's a 1 (true)
    if i >= 0.5:
        binary_predictions.append(1)
    # It's false
    else:
        binary_predictions.append(0)

# lets fill out the confusion matrix with values
matrix_of_confusion = confusion_matrix(binary_predictions, y_test,
                                       normalize='all')

# Lets plot it (standard eight by 5)
plt.figure(figsize=(8, 5))

matrix_graph = plt.subplot()
# Adding a heat map
sns.heatmap(matrix_of_confusion, annot=True, ax = matrix_graph)
# Labeling and prettying it up
matrix_graph.set_xlabel('Predicted Labels', size=18)
matrix_graph.set_ylabel('True Labels', size=18)
matrix_graph.set_title('Confusion Matrix', size=25)
```

```
matrix_graph.xaxis.set_ticklabels([0,1], size=10)
matrix_graph.yaxis.set_ticklabels([0,1], size=10)

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")

Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_5 (Embedding)     (None, None, 128)         1280000

 bidirectional_1 (Bidirecti  (None, 32)                18560
 onal)

 dense_6 (Dense)             (None, 64)                2112

 dropout_1 (Dropout)         (None, 64)                0

 dense_7 (Dense)             (None, 1)                 65

=================================================================
Total params: 1300737 (4.96 MB)
Trainable params: 1300737 (4.96 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
943/943 [==============================] - 63s 63ms/step - loss:
0.0804 - accuracy: 0.9584 - val_loss: 0.0268 - val_accuracy: 0.9908
421/421 [==============================] - 4s 10ms/step - loss: 0.0202
- accuracy: 0.9932
421/421 [==============================] - 4s 8ms/step
421/421 [==============================] - 5s 12ms/step - loss: 0.0202
- accuracy: 0.9932
Test Accuracy: 99.32%
```



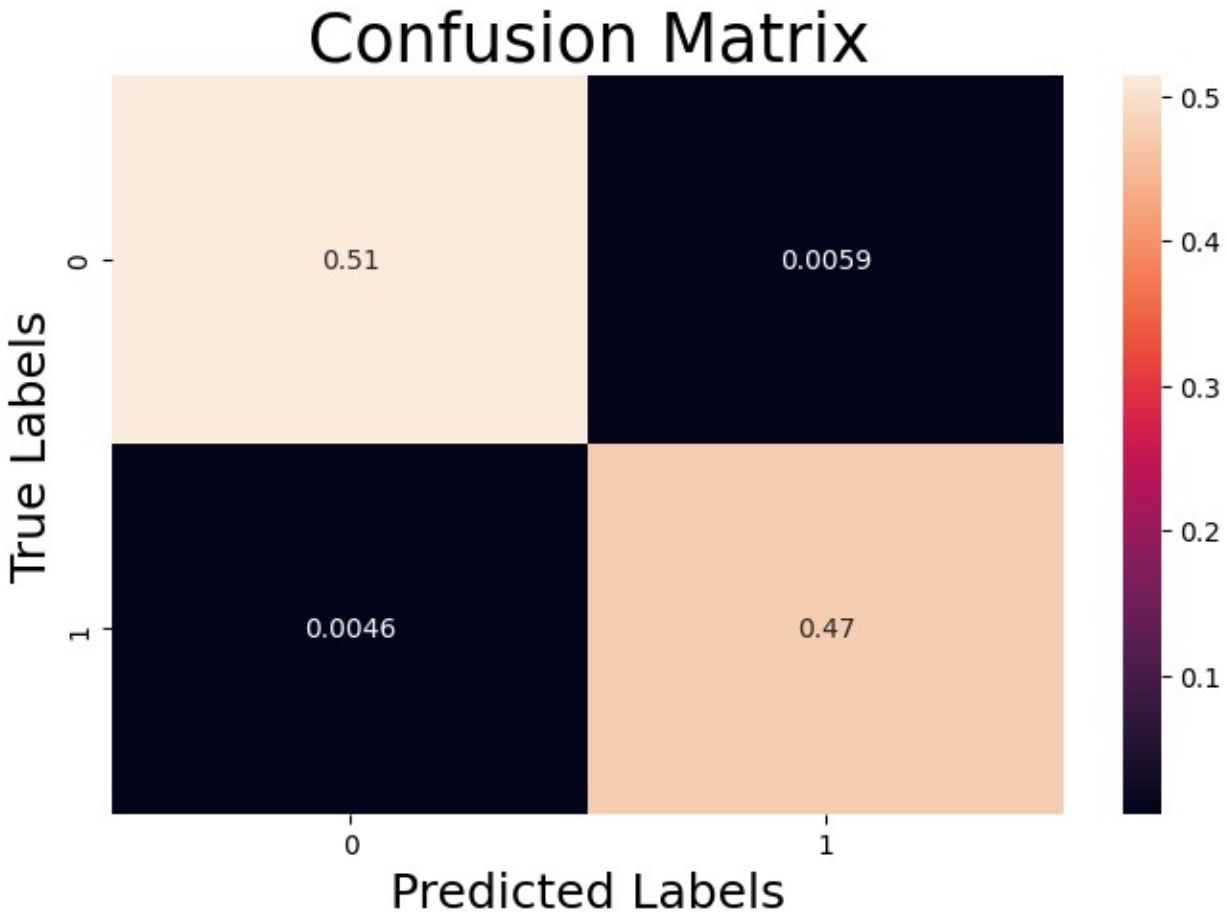
On cleaned data

```
# Let's get everything ready for using the multinomial Naive Bayes
classifier (Bayes for text)
# features are the text of the news
features_raw = data['no_puntuation_text']
# targets are the values 0 for fake and 1 for true
targets_raw = data['label']

# Let's split the data into train and test datasets
X_train, X_test, y_train, y_test = train_test_split(features, targets,
test_size=0.30, random_state=7)

# Since we want to predict using words rather than numbers, we need to
limit the word vocabulary
# and we need to set the tokenizer to limit the max amount of the
vocabulary
max_vocabulary = 10000
tokenizer = Tokenizer(num_words=max_vocabulary)
tokenizer.fit_on_texts(X_train)

# Now, let's use the tokenizer to turn text into lists
```

```
X_train = tokenizer.texts_to_sequences(X_train)
X_test = tokenizer.texts_to_sequences(X_test)

# This pads the sequences to make them the same length; we need this
for processing the data
X_train = tf.keras.preprocessing.sequence.pad_sequences(X_train,
padding='post', maxlen=256)
X_test = tf.keras.preprocessing.sequence.pad_sequences(X_test,
padding='post', maxlen=256)

# Features are the text of the news
features = data['no_puntuation_text']
# Targets are the values 0 for fake and 1 for true
targets = data['label']

# Let's create the model now!
model = tf.keras.Sequential([
    # This is the embedding layer, to convert into dense vectors for
better input
    tf.keras.layers.Embedding(max_vocabulary, 128),
    # This will process the input in both directions; it helps with
work context
    tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(16)),
    # This is technical, but it applies a non-linear transformation to
the inputs
    # using the ReLU activation function
    tf.keras.layers.Dense(64, activation='relu'),
    # This prevents overfitting by setting a fraction of the inputs as
0
    tf.keras.layers.Dropout(0.5),
    # This predicts the final value
    tf.keras.layers.Dense(1)
])

# Let's print a summary of the model
model.summary()

# Let's compile the model
model.compile(loss=tf.keras.losses.BinaryCrossentropy(from_logits=True
),
              optimizer=tf.keras.optimizers.Adam(1e-3),
              metrics=['accuracy'])

# Let's fit the model
model_fitter = model.fit(X_train, y_train, epochs=1,
validation_split=0.1,
                         batch_size=30, shuffle=True)


# Let's evaluate it
```

```
model.evaluate(X_test, y_test)

# Lets predict something
predicted_LSTM_model = model.predict(X_test)

# Now let's try to graph that model

# This is where I'll store the predictions
binary_predictions = []

# Let's place the predicted values into buckets and then into my
storage
for i in predicted_LSTM_model:
    # If it was higher/equal to 0.5, it's a 1 (true)
    if i >= 0.5:
        binary_predictions.append(1)
    # It's false
    else:
        binary_predictions.append(0)

# lets fill out the confusion matrix with values
matrix_of_confusion = confusion_matrix(binary_predictions, y_test,
                                       normalize='all')

# Lets plot it (standard eight by 5)
plt.figure(figsize=(8, 5))

matrix_graph = plt.subplot()
# Adding a heat map
sns.heatmap(matrix_of_confusion, annot=True, ax = matrix_graph)
# Labeling and prettying it up
matrix_graph.set_xlabel('Predicted Labels', size=18)
matrix_graph.set_ylabel('True Labels', size=18)
matrix_graph.set_title('Confusion Matrix', size=25)
matrix_graph.xaxis.set_ticklabels([0,1], size=10)
matrix_graph.yaxis.set_ticklabels([0,1], size=10)

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")

Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_6 (Embedding)     (None, None, 128)         1280000

 bidirectional_2 (Bidirecti  (None, 32)                18560
 onal)
```

```
 dense_8 (Dense)             (None, 64)                2112

 dropout_2 (Dropout)         (None, 64)                0

 dense_9 (Dense)             (None, 1)                 65

=================================================================
Total params: 1300737 (4.96 MB)
Trainable params: 1300737 (4.96 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
943/943 [==============================] - 64s 63ms/step - loss:
0.1044 - accuracy: 0.9517 - val_loss: 0.0365 - val_accuracy: 0.9873
421/421 [==============================] - 4s 9ms/step - loss: 0.0323
- accuracy: 0.9895
421/421 [==============================] - 5s 11ms/step
421/421 [==============================] - 4s 9ms/step - loss: 0.0323
- accuracy: 0.9895
Test Accuracy: 98.95%
```



Conclusion on LSTM

On raw data, we get an accuracy of: 99.32%

On cleaned data, we get an accuracy of: 98.95%

In LSTM cleaning the data affects the accurary, it is more accurate to use the raw data again, this can't be a coincidence!

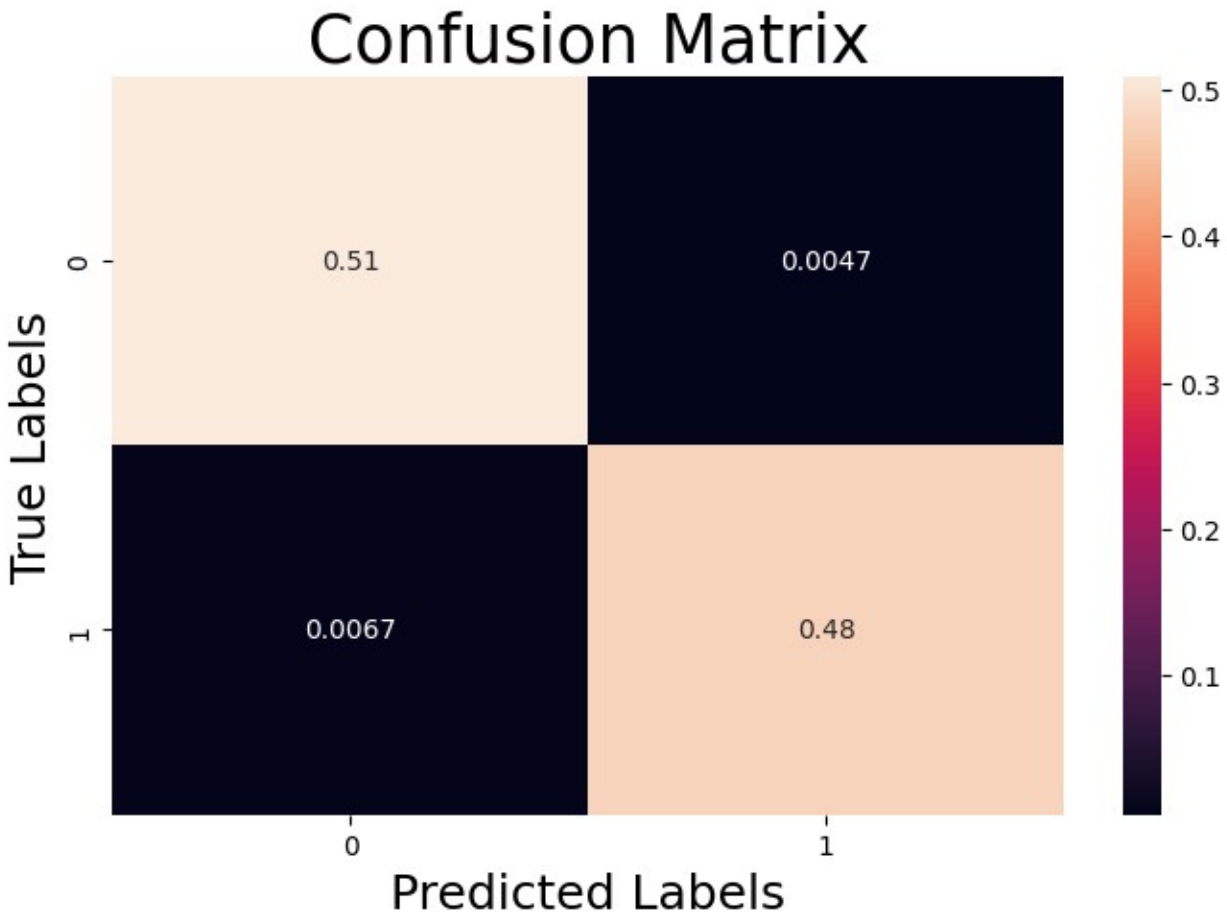
CNN

On raw data

```
# Taken and adapted from:
https://keras.io/api/layers/convolution_layers/convolution1d/

# Data Preparation
X = data['text']
y = data['label']

# Label encoding
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Text Vectorization
tokenizer = Tokenizer(num_words=5000)
tokenizer.fit_on_texts(X)
X_seq = tokenizer.texts_to_sequences(X)
X_pad = pad_sequences(X_seq, maxlen=100)

# Splitting the dataset
X_train, X_test, y_train, y_test = train_test_split(X_pad, y_encoded,
                                                    test_size=0.2,
                                                    random_state=42)

# Building the CNN Model
model = Sequential()
model.add(Embedding(input_dim=5000, output_dim=50, input_length=100))
model.add(Conv1D(filters=128, kernel_size=5, activation='relu'))
model.add(GlobalMaxPooling1D())
model.add(Dense(10, activation='relu'))
model.add(Dense(1, activation='sigmoid'))  # Use 'softmax' if you have
more than two classes

# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy',
metrics=['accuracy'])

# Train the model
model.fit(X_train, y_train, epochs=5, batch_size=32,
validation_data=(X_test, y_test))
```

```
# Let's evaluate it
model.evaluate(X_test, y_test)

# Lets predict something
predicted_CNN_model = model.predict(X_test)

# Now let's try to graph that model

# This is where I'll store the predictions
binary_predictions = []

# Let's place the predicted values into buckets and then into my
storage
for i in predicted_CNN_model:
    # If it was higher/equal to 0.5, it's a 1 (true)
    if i >= 0.5:
        binary_predictions.append(1)
    # It's false
    else:
        binary_predictions.append(0)

# lets fill out the confusion matrix with values
matrix_of_confusion = confusion_matrix(binary_predictions, y_test,
                                       normalize='all')

# Lets plot it (standard eight by 5)
plt.figure(figsize=(8, 5))

matrix_graph = plt.subplot()
# Adding a heat map
sns.heatmap(matrix_of_confusion, annot=True, ax = matrix_graph)
# Labeling and prettying it up
matrix_graph.set_xlabel('Predicted Labels', size=18)
matrix_graph.set_ylabel('True Labels', size=18)
matrix_graph.set_title('Confusion Matrix', size=25)
matrix_graph.xaxis.set_ticklabels([0,1], size=10)
matrix_graph.yaxis.set_ticklabels([0,1], size=10)

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")

Epoch 1/5
1123/1123 [==============================] - 42s 36ms/step - loss:
0.1101 - accuracy: 0.9583 - val_loss: 0.0465 - val_accuracy: 0.9849
Epoch 2/5
1123/1123 [==============================] - 9s 8ms/step - loss:
0.0172 - accuracy: 0.9946 - val_loss: 0.0430 - val_accuracy: 0.9872
Epoch 3/5
```

```
1123/1123 [==============================] - 7s 6ms/step - loss:
0.0037 - accuracy: 0.9991 - val_loss: 0.0495 - val_accuracy: 0.9889
Epoch 4/5
1123/1123 [==============================] - 8s 7ms/step - loss:
3.1831e-04 - accuracy: 1.0000 - val_loss: 0.0554 - val_accuracy:
0.9889
Epoch 5/5
1123/1123 [==============================] - 9s 8ms/step - loss:
6.2669e-05 - accuracy: 1.0000 - val_loss: 0.0591 - val_accuracy:
0.9886
281/281 [==============================] - 1s 3ms/step - loss: 0.0591
- accuracy: 0.9886
281/281 [==============================] - 1s 3ms/step
281/281 [==============================] - 1s 3ms/step - loss: 0.0591
- accuracy: 0.9886
Test Accuracy: 98.86%
```



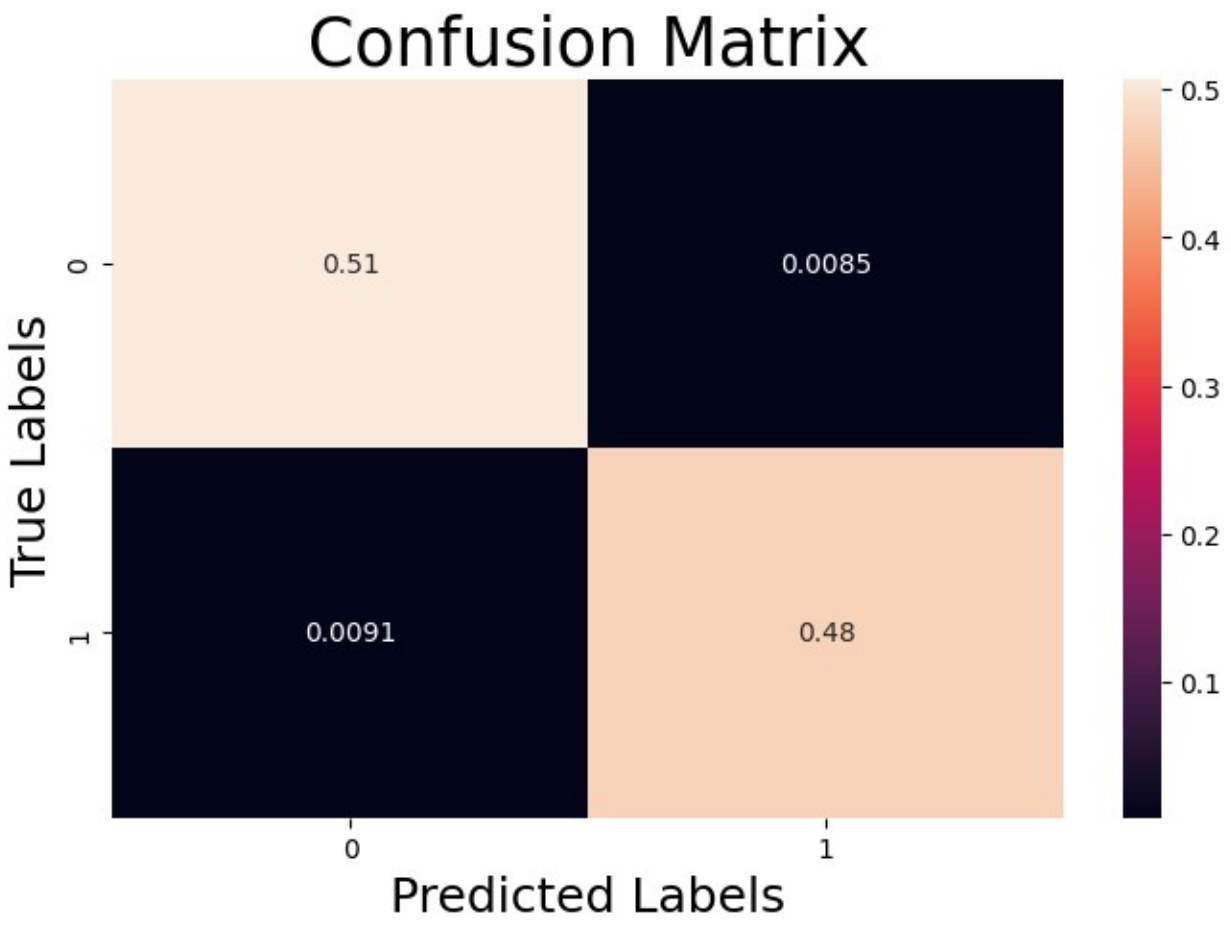
On cleaned data

```
# Data Preparation
X = data['no_stopwords']
```

```
y = data['label']

# Label encoding
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Text Vectorization
tokenizer = Tokenizer(num_words=5000)
tokenizer.fit_on_texts(X)
X_seq = tokenizer.texts_to_sequences(X)
X_pad = pad_sequences(X_seq, maxlen=100)

# Splitting the dataset
X_train, X_test, y_train, y_test = train_test_split(X_pad, y_encoded,
                                                    test_size=0.2,
random_state=42)

# Building the CNN Model
model = Sequential()
model.add(Embedding(input_dim=5000, output_dim=50, input_length=100))
model.add(Conv1D(filters=128, kernel_size=5, activation='relu'))
model.add(GlobalMaxPooling1D())
model.add(Dense(10, activation='relu'))
model.add(Dense(1, activation='sigmoid'))  # Use 'softmax' if you have
more than two classes

# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy',
metrics=['accuracy'])

# Train the model
model.fit(X_train, y_train, epochs=5, batch_size=32,
validation_data=(X_test, y_test))

# Let's evaluate it
model.evaluate(X_test, y_test)

# Lets predict something
predicted_CNN_model = model.predict(X_test)

# Now let's try to graph that model

# This is where I'll store the predictions
binary_predictions = []

# Let's place the predicted values into buckets and then into my
storage
for i in predicted_CNN_model:
    # If it was higher/equal to 0.5, it's a 1 (true)
    if i >= 0.5:
```

```
        binary_predictions.append(1)
    # It's false
    else:
        binary_predictions.append(0)

# lets fill out the confusion matrix with values
matrix_of_confusion = confusion_matrix(binary_predictions, y_test,
                                       normalize='all')

# Lets plot it (standard eight by 5)
plt.figure(figsize=(8, 5))

matrix_graph = plt.subplot()
# Adding a heat map
sns.heatmap(matrix_of_confusion, annot=True, ax = matrix_graph)
# Labeling and prettying it up
matrix_graph.set_xlabel('Predicted Labels', size=18)
matrix_graph.set_ylabel('True Labels', size=18)
matrix_graph.set_title('Confusion Matrix', size=25)
matrix_graph.xaxis.set_ticklabels([0,1], size=10)
matrix_graph.yaxis.set_ticklabels([0,1], size=10)

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
Epoch 1/5
1123/1123 [==============================] - 55s 48ms/step - loss:
0.1237 - accuracy: 0.9511 - val_loss: 0.0665 - val_accuracy: 0.9751
Epoch 2/5
1123/1123 [==============================] - 13s 12ms/step - loss:
0.0250 - accuracy: 0.9920 - val_loss: 0.0565 - val_accuracy: 0.9825
Epoch 3/5
1123/1123 [==============================] - 9s 8ms/step - loss:
0.0045 - accuracy: 0.9989 - val_loss: 0.0730 - val_accuracy: 0.9821
Epoch 4/5
1123/1123 [==============================] - 9s 8ms/step - loss:
0.0014 - accuracy: 0.9997 - val_loss: 0.0779 - val_accuracy: 0.9826
Epoch 5/5
1123/1123 [==============================] - 8s 7ms/step - loss:
3.3671e-04 - accuracy: 0.9999 - val_loss: 0.0854 - val_accuracy:
0.9824
281/281 [==============================] - 1s 2ms/step - loss: 0.0854
- accuracy: 0.9824
281/281 [==============================] - 1s 2ms/step
281/281 [==============================] - 1s 3ms/step - loss: 0.0854
- accuracy: 0.9824
Test Accuracy: 98.24%
```



Conclusion on CNN

On raw data, we get an accuracy of: 98.86%

On cleaned data, we get an accuracy of: 98.24%

In CNN cleaning the data affects the accurary, it is more accurate to use the raw data yet again!

Now, CNN + bidirectional LSTM

On raw data

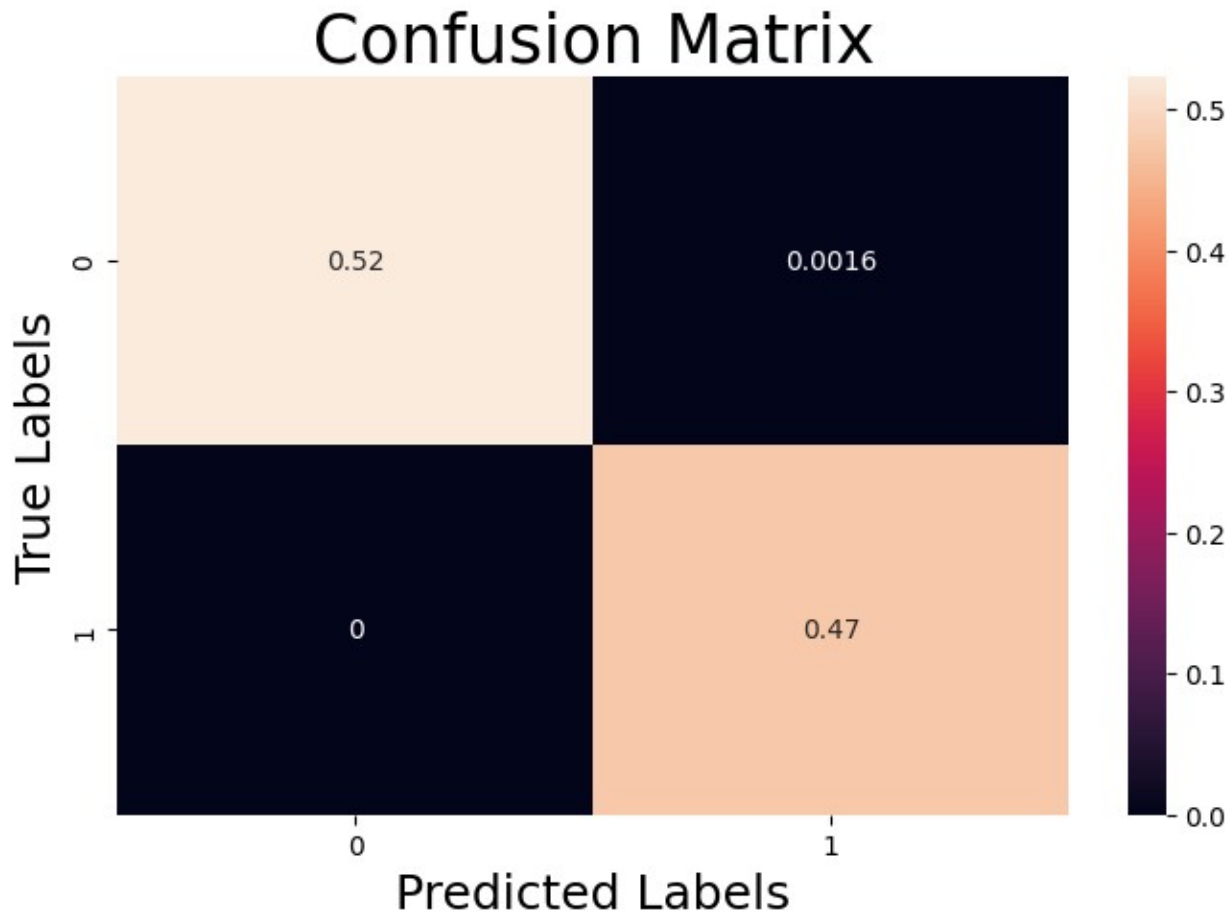
```
# Taken and adapted from:
https://stackoverflow.com/questions/64150587/combining-cnn-and-
bidirectional-lstm

# Parameters
vocab_size = 10000
embedding_dim = 64
max_length = 50
trunc_type='post'
padding_type='post'
```

```
oov_tok = "<OOV>"

# Tokenize the data
tokenizer = Tokenizer(num_words=vocab_size, oov_token=oov_tok)
tokenizer.fit_on_texts(data['text'])
word_index = tokenizer.word_index
sequences = tokenizer.texts_to_sequences(data['text'])
padded = pad_sequences(sequences, maxlen=max_length,
padding=padding_type,
                       truncating=trunc_type)

# Split data into training and testing (example split)
train_size = int(len(data) * 0.8)
train_sequences = padded[0:train_size]
train_labels = data['label'][0:train_size]
test_sequences = padded[train_size:]
test_labels = data['label'][train_size:]

# Building the model
model = Sequential([
    Embedding(vocab_size, embedding_dim, input_length=max_length),
    Conv1D(64, 5, activation='relu'),
    MaxPooling1D(pool_size=4),
    Bidirectional(LSTM(64)),
    Dense(1, activation='sigmoid')
])

# Compile the model
model.compile(loss='binary_crossentropy', optimizer='adam',
              metrics=['accuracy'])

# Train the model
model.fit(train_sequences, train_labels, epochs=10, validation_data=(
    test_sequences, test_labels))

# Let's evaluate it
model.evaluate(test_sequences, test_labels)

# Lets predict something
predicted_CNN_LSTM_model = model.predict(test_sequences)

# Now let's try to graph that model

# This is where I'll store the predictions
binary_predictions = []

# Let's place the predicted values into buckets and then into my
storage
for i in predicted_CNN_LSTM_model:
    # If it was higher/equal to 0.5, it's a 1 (true)
```

```
    if i >= 0.5:
        binary_predictions.append(1)
    # It's false
    else:
        binary_predictions.append(0)

# lets fill out the confusion matrix with values
matrix_of_confusion = confusion_matrix(binary_predictions,
test_labels,
                                       normalize='all')

# Lets plot it (standard eight by 5)
plt.figure(figsize=(8, 5))

matrix_graph = plt.subplot()
# Adding a heat map
sns.heatmap(matrix_of_confusion, annot=True, ax = matrix_graph)
# Labeling and prettying it up
matrix_graph.set_xlabel('Predicted Labels', size=18)
matrix_graph.set_ylabel('True Labels', size=18)
matrix_graph.set_title('Confusion Matrix', size=25)
matrix_graph.xaxis.set_ticklabels([0,1], size=10)
matrix_graph.yaxis.set_ticklabels([0,1], size=10)

# Evaluate the model
loss, accuracy = model.evaluate(test_sequences, test_labels)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
Epoch 1/10
1123/1123 [==============================] - 29s 22ms/step - loss:
0.0314 - accuracy: 0.9864 - val_loss: 0.0055 - val_accuracy: 0.9987
Epoch 2/10
1123/1123 [==============================] - 12s 11ms/step - loss:
0.0023 - accuracy: 0.9994 - val_loss: 0.0054 - val_accuracy: 0.9986
Epoch 3/10
1123/1123 [==============================] - 12s 10ms/step - loss:
0.0013 - accuracy: 0.9997 - val_loss: 0.0087 - val_accuracy: 0.9972
Epoch 4/10
1123/1123 [==============================] - 11s 10ms/step - loss:
5.2560e-04 - accuracy: 0.9999 - val_loss: 0.0061 - val_accuracy:
0.9986
Epoch 5/10
1123/1123 [==============================] - 10s 9ms/step - loss:
3.2503e-04 - accuracy: 1.0000 - val_loss: 0.0063 - val_accuracy:
0.9989
Epoch 6/10
1123/1123 [==============================] - 12s 11ms/step - loss:
3.2897e-04 - accuracy: 1.0000 - val_loss: 0.0076 - val_accuracy:
0.9984
Epoch 7/10
```

```
1123/1123 [==============================] - 11s 10ms/step - loss:
2.8137e-04 - accuracy: 1.0000 - val_loss: 0.0104 - val_accuracy:
0.9986
Epoch 8/10
1123/1123 [==============================] - 12s 10ms/step - loss:
3.0183e-04 - accuracy: 1.0000 - val_loss: 0.0109 - val_accuracy:
0.9984
Epoch 9/10
1123/1123 [==============================] - 11s 10ms/step - loss:
3.0923e-04 - accuracy: 1.0000 - val_loss: 0.0120 - val_accuracy:
0.9984
Epoch 10/10
1123/1123 [==============================] - 11s 10ms/step - loss:
2.5906e-04 - accuracy: 1.0000 - val_loss: 0.0127 - val_accuracy:
0.9984
281/281 [==============================] - 1s 4ms/step - loss: 0.0127
- accuracy: 0.9984
281/281 [==============================] - 2s 3ms/step
281/281 [==============================] - 1s 4ms/step - loss: 0.0127
- accuracy: 0.9984
Test Accuracy: 99.84%
```



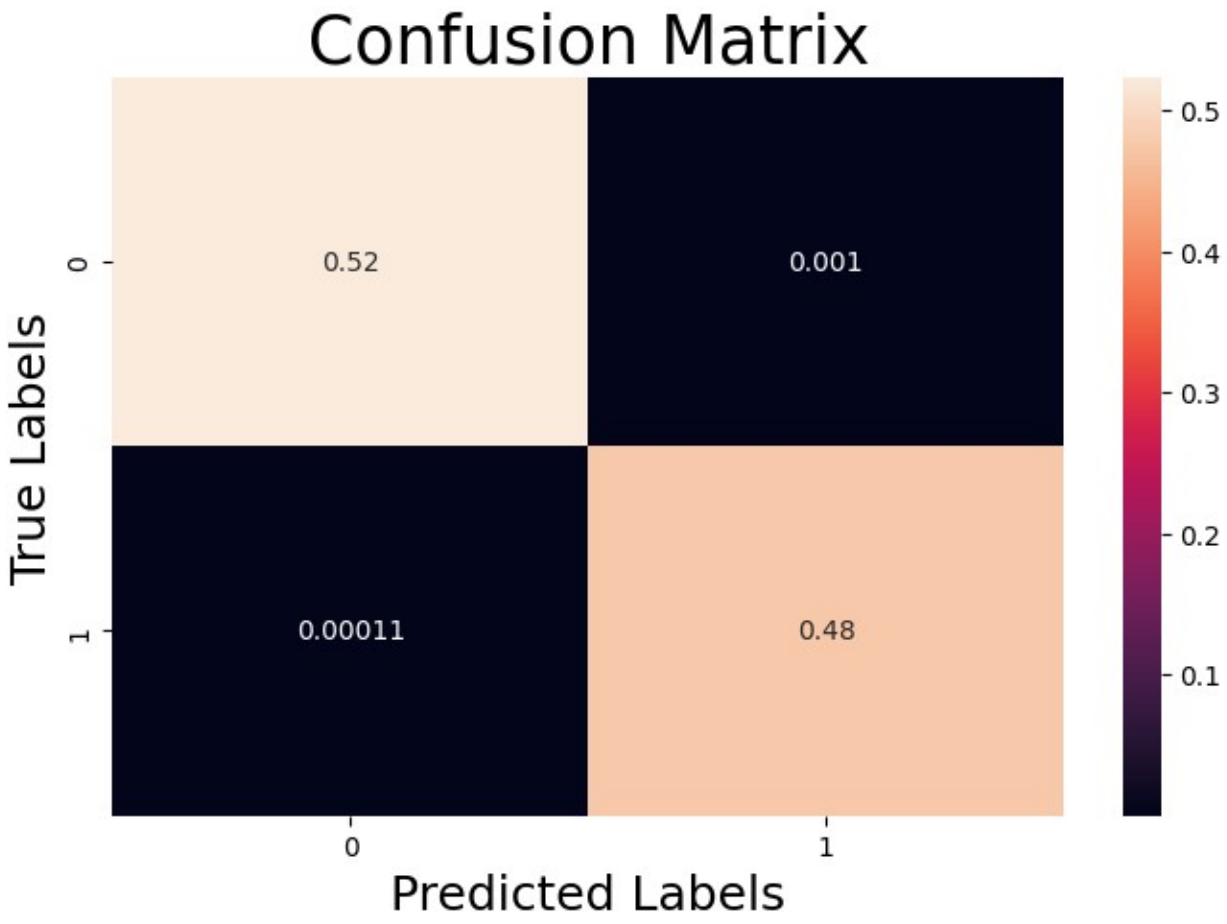
On cleaned data

```
# Parameters
vocab_size = 10000
embedding_dim = 64
max_length = 50
trunc_type='post'
padding_type='post'
oov_tok = "<OOV>"

# Tokenize the data
tokenizer = Tokenizer(num_words=vocab_size, oov_token=oov_tok)
tokenizer.fit_on_texts(data['no_stopwords'])
word_index = tokenizer.word_index
sequences = tokenizer.texts_to_sequences(data['no_stopwords'])
padded = pad_sequences(sequences, maxlen=max_length,
padding=padding_type,
                       truncating=trunc_type)

# Split data into training and testing (example split)
train_size = int(len(data) * 0.8)
train_sequences = padded[0:train_size]
train_labels = data['label'][0:train_size]
test_sequences = padded[train_size:]
test_labels = data['label'][train_size:]

# Building the model
model = Sequential([
    Embedding(vocab_size, embedding_dim, input_length=max_length),
    Conv1D(64, 5, activation='relu'),
    MaxPooling1D(pool_size=4),
    Bidirectional(LSTM(64)),
    Dense(1, activation='sigmoid')
])

# Compile the model
model.compile(loss='binary_crossentropy', optimizer='adam',
              metrics=['accuracy'])

# Train the model
model.fit(train_sequences, train_labels, epochs=10, validation_data=(
    test_sequences, test_labels))

# Let's evaluate it
model.evaluate(test_sequences, test_labels)

# Lets predict something
predicted_CNN_LSTM_model = model.predict(test_sequences)

# Now let's try to graph that model
```

```
# This is where I'll store the predictions
binary_predictions = []

# Let's place the predicted values into buckets and then into my
storage
for i in predicted_CNN_LSTM_model:
    # If it was higher/equal to 0.5, it's a 1 (true)
    if i >= 0.5:
        binary_predictions.append(1)
    # It's false
    else:
        binary_predictions.append(0)

# lets fill out the confusion matrix with values
matrix_of_confusion = confusion_matrix(binary_predictions,
test_labels,
                                       normalize='all')

# Lets plot it (standard eight by 5)
plt.figure(figsize=(8, 5))

matrix_graph = plt.subplot()
# Adding a heat map
sns.heatmap(matrix_of_confusion, annot=True, ax = matrix_graph)
# Labeling and prettying it up
matrix_graph.set_xlabel('Predicted Labels', size=18)
matrix_graph.set_ylabel('True Labels', size=18)
matrix_graph.set_title('Confusion Matrix', size=25)
matrix_graph.xaxis.set_ticklabels([0,1], size=10)
matrix_graph.yaxis.set_ticklabels([0,1], size=10)

# Evaluate the model
loss, accuracy = model.evaluate(test_sequences, test_labels)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
Epoch 1/10
1123/1123 [==============================] - 42s 34ms/step - loss:
0.0351 - accuracy: 0.9853 - val_loss: 0.0068 - val_accuracy: 0.9981
Epoch 2/10
1123/1123 [==============================] - 14s 12ms/step - loss:
0.0023 - accuracy: 0.9995 - val_loss: 0.0068 - val_accuracy: 0.9986
Epoch 3/10
1123/1123 [==============================] - 12s 10ms/step - loss:
7.4110e-04 - accuracy: 0.9999 - val_loss: 0.0082 - val_accuracy:
0.9983
Epoch 4/10
1123/1123 [==============================] - 11s 10ms/step - loss:
5.1769e-04 - accuracy: 0.9998 - val_loss: 0.0105 - val_accuracy:
0.9968
Epoch 5/10
```

```
1123/1123 [==============================] - 12s 11ms/step - loss:
9.2159e-04 - accuracy: 0.9998 - val_loss: 0.0069 - val_accuracy:
0.9989
Epoch 6/10
1123/1123 [==============================] - 13s 12ms/step - loss:
7.5634e-04 - accuracy: 0.9998 - val_loss: 0.0085 - val_accuracy:
0.9987
Epoch 7/10
1123/1123 [==============================] - 10s 9ms/step - loss:
2.8030e-04 - accuracy: 1.0000 - val_loss: 0.0073 - val_accuracy:
0.9989
Epoch 8/10
1123/1123 [==============================] - 11s 10ms/step - loss:
2.5416e-04 - accuracy: 1.0000 - val_loss: 0.0084 - val_accuracy:
0.9989
Epoch 9/10
1123/1123 [==============================] - 12s 10ms/step - loss:
2.4366e-04 - accuracy: 1.0000 - val_loss: 0.0092 - val_accuracy:
0.9989
Epoch 10/10
1123/1123 [==============================] - 11s 10ms/step - loss:
2.7968e-04 - accuracy: 1.0000 - val_loss: 0.0085 - val_accuracy:
0.9989
281/281 [==============================] - 1s 4ms/step - loss: 0.0085
- accuracy: 0.9989
281/281 [==============================] - 2s 3ms/step
281/281 [==============================] - 1s 4ms/step - loss: 0.0085
- accuracy: 0.9989
Test Accuracy: 99.89%
```



Conclusion on CNN + bidirectional LSTM

On raw data, we get an accuracy of: 99.84%

On cleaned data, we get an accuracy of: 99.89%

Now this makes more sense, with CNN + bidirectional LSTM cleaning the data affects the accurary, it is better to use clean data to get more accuracy.

99.84% is the best I've seen!

Summary and conclusions

Algoritms

Bayes was our baseline, simple and a low bar.

Simple RNN was a high bar, we selected it because it was mentioned more than once on the research papers.

GRU was a clear next bar, better than SimpleRNN but requiered more computing power.

LSTM, CNN and CNN+Bidirectional LSTM, was something we also read from the research papers, adding two layers was something new and required the most computational power i have ever done (this requiered that I pay for credits on Google Colab, otherwise it woudl have taken days to execute on the free version)

Conclusion

Better algorithms are heavier and requiere more computeing power, researching them and implementing them is not easy and the worse part is modifying the parameters to be able to run in hours and not days.

I was able to run everything in a few hours only becuase I was able to pay for Google Colab.

In my computer, it would have taken days (I have an old computer Intel i3 from 5 years ago)

Reaching over 99% was a dream, I took weeks of testing each algoritm to be able to run them in my computer, it was a shocking experience runing them in Google Colab after paying, it took minutes instead of hours and days.

I learned that in order to trully be able to do this, you need good computing power, everything else just consumes your patience and life.