
The Dataset is about fake news and can be found on kaggle:
https://www.kaggle.com/c/fake-news/data

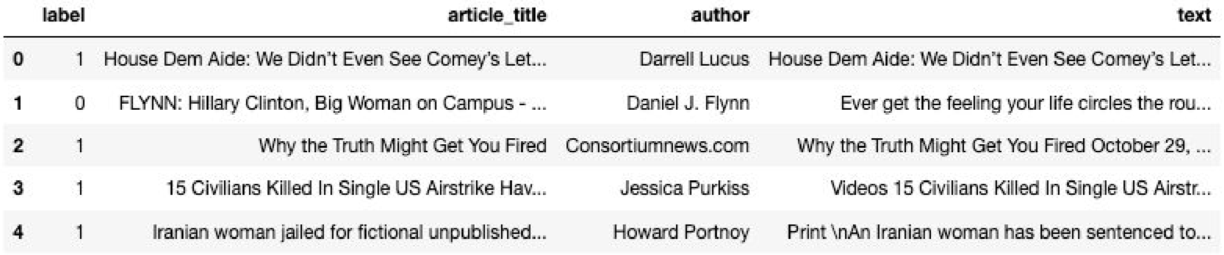
Introduction:

Social Media companies need to have satisfied customers for it to be successful and this report looks at our ability to predict fake news articles on a social media platform and flag them as fake or real. From a business perspective, the social media giant is the client and they are concerned about potentially losing current users from their platform. This is because current users have complained of fake news articles on the platform that are eroding their trust in the platform. The client is interested in predicting which news articles circulating on their platform are real or fake, and to flag the fake articles so users on the platform are aware of whether the article they are reading is real or fake. This project is time-sensitive because of the increase in misinformation being spread on the client's platform and increasing pressure from regulatory authorities to contain and identify fake news on its platform.

Data Wrangling:

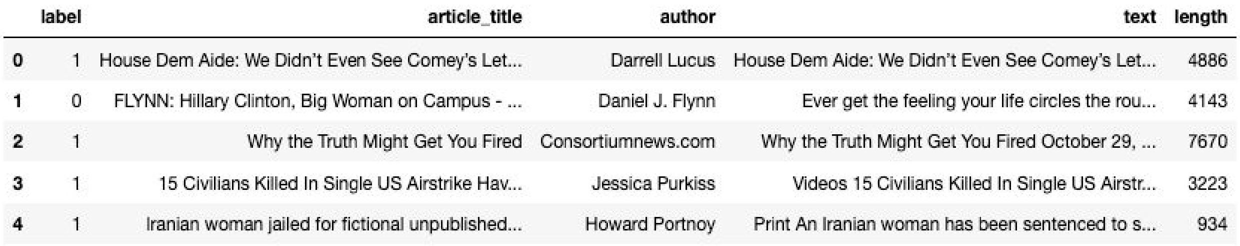
The dataset for news articles containing real and fake news articles was obtained from Kaggle in a csv format. The shape of the dataset was 20800 rows and 5 columns. There were 10413 real articles and 10387 fake articles in the dataset indicating that the dataset is balanced. The column names of the dataset were checked and **'title'** was converted to **'article_title'.** The **'id'** column was dropped as it was redundant in the dataset.

For visual purposes, the response variable, in this case 'label', was moved to the left side of the table. Also, it makes the dataset splitting into train/test set easier later on.



	label	article_title	author	text
0	1	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...
1	0	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...
2	1	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...
3	1	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...
4	1	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...

To help with text preprocessing, a helper function was created. This helper function removes line breaks, new lines, hyperlinks, ampersand, greater than sign, less than sign, non breaking space, emails, new line characters and distracting single quotes. A new column **'length'** was created that contains the length of the article text.



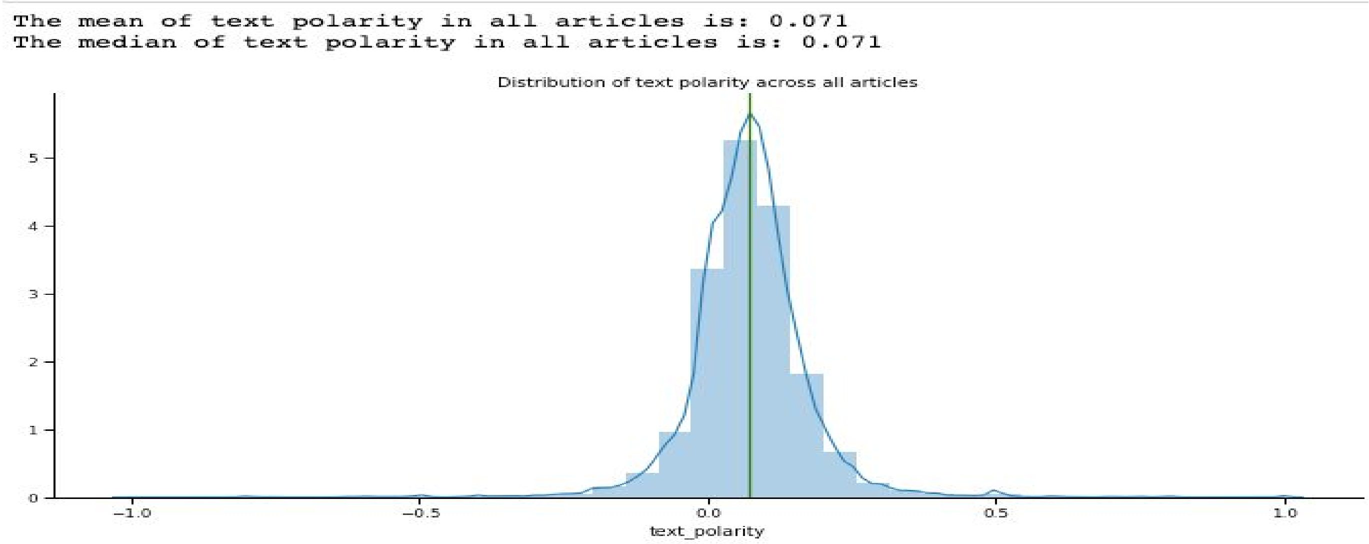
	label	article_title	author	text	length
0	1	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	4886
1	0	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	4143
2	1	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	7670
3	1	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	3223
4	1	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print An Iranian woman has been sentenced to s...	934

Next, null values were checked in the dataset. The columns ‘author’ and ‘article_title’ contained 1957 and 558 null values respectively. The null values in these columns were replaced by ‘Unknown’. To remove articles that are not really articles as they contain really few characters, I picked an arbitrary number of 50 characters. The length of text in articles that was less than 50 were dropped. The new shape of the dataset was 20554 rows and 5 columns.

Exploring the Data

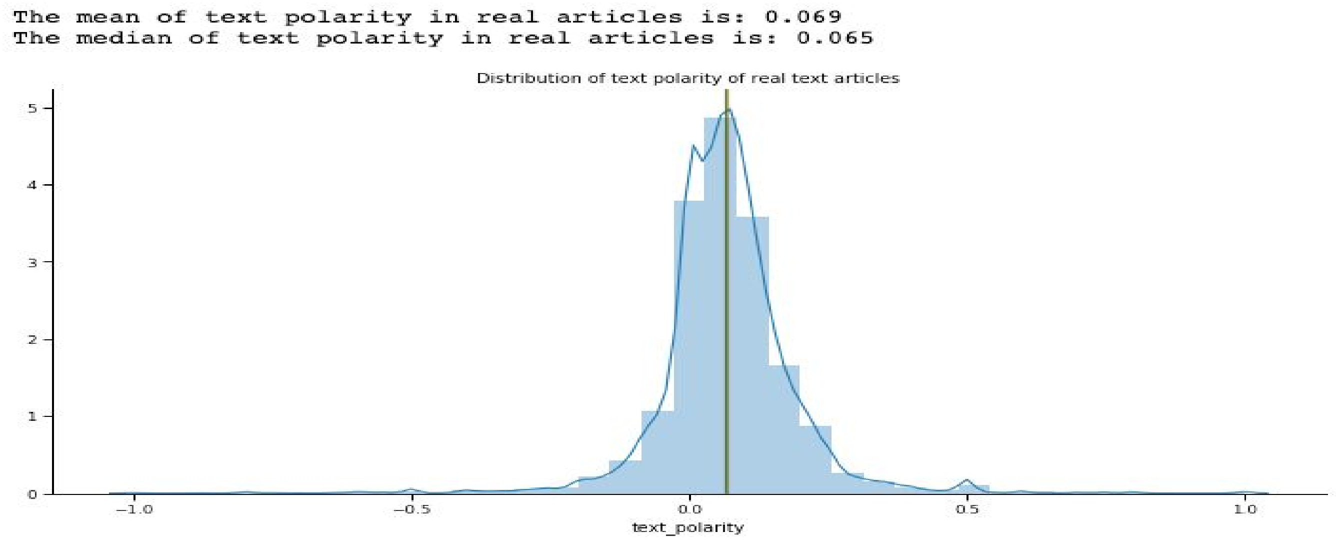
Before making any visualizations, another column called ‘text_polarity’ was added to the dataset. Using the TextBlob library’s sentiment polarity function on the ‘text’ column returned polarity of text on scale -1 to 1 indicating negative to positive.

It would be interesting to see how the distributions of text polarity in all articles, real articles and fake articles differ. The distribution of text polarity across all articles seems to be evenly distributed with some articles being completely negative and completely positive.



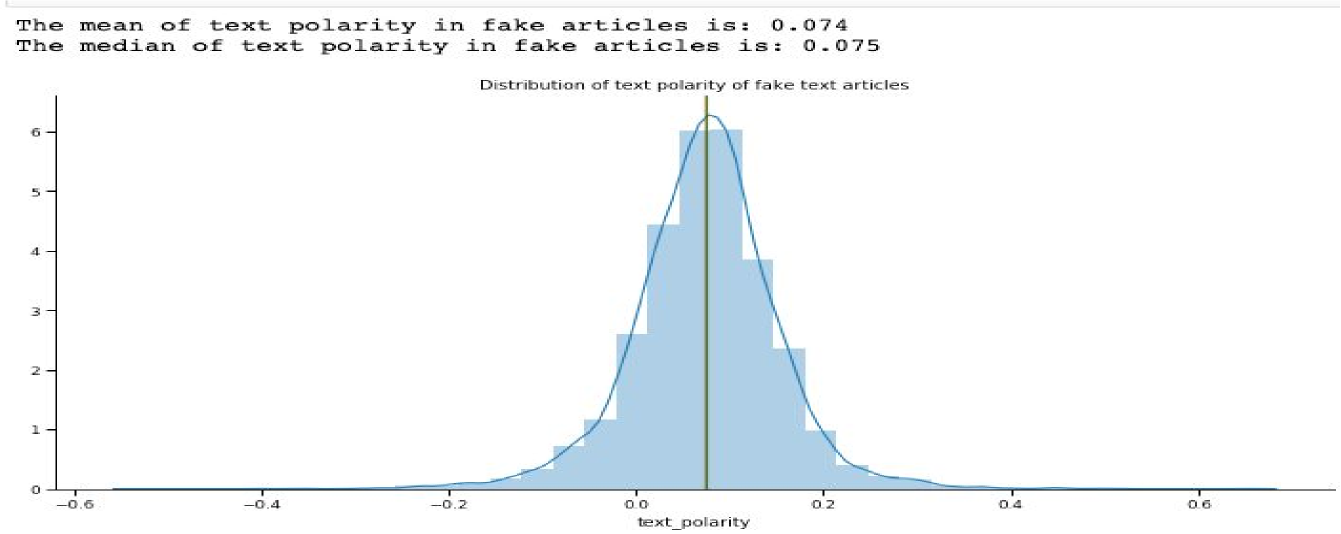
```
The mean of text polarity in all articles is: 0.071
The median of text polarity in all articles is: 0.071
```

The distribution of text_polarity of real text articles again seems to be evenly distributed. It also contains articles that have completely negative or positive polarity.



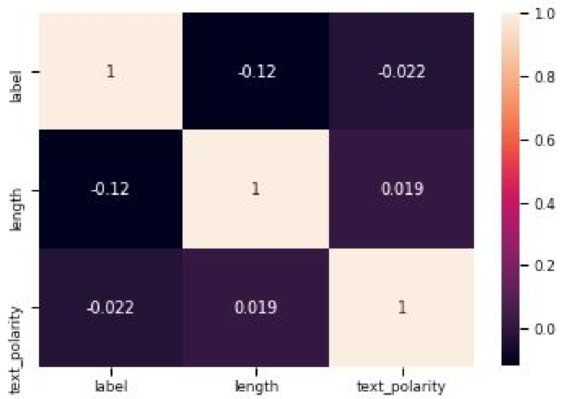
```
The mean of text polarity in real articles is: 0.069
The median of text polarity in real articles is: 0.065
```

The distribution of text polarity of fake news articles seems normally distributed. An interesting thing to note is that there are no perfectly negative or positive text polarity (i.e. -1 and +1) in the fake news articles.

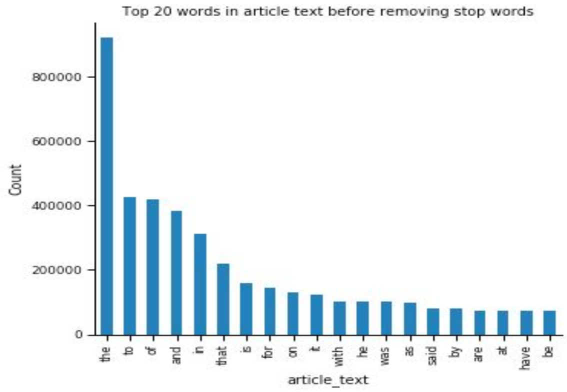


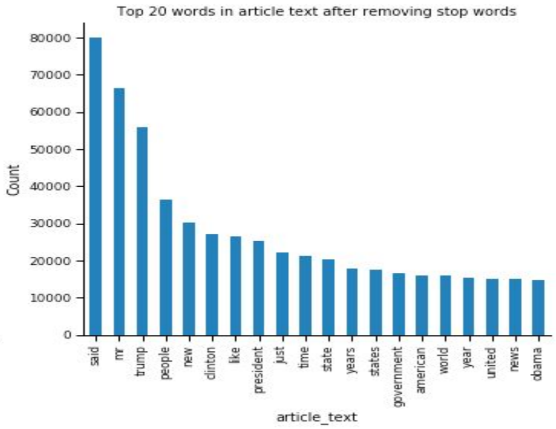
```
The mean of text polarity in fake articles is: 0.074
The median of text polarity in fake articles is: 0.075
```

To check for any correlations between the different features in the dataset, a heatmap was used. There did not seem to be any strong correlations in the dataset. The greatest correlation value in the dataset was -0.12 between length and label, indicating a very weak negative correlation.

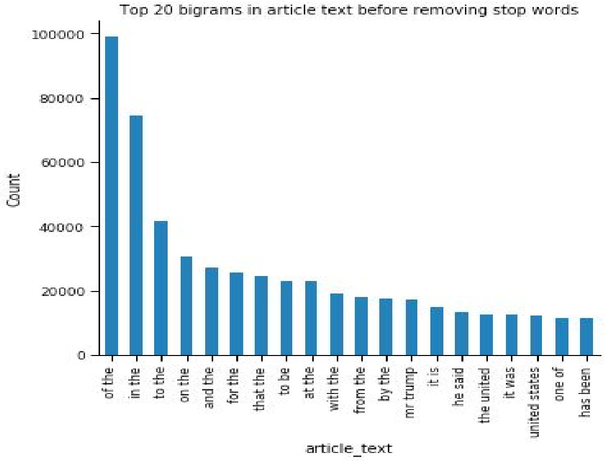


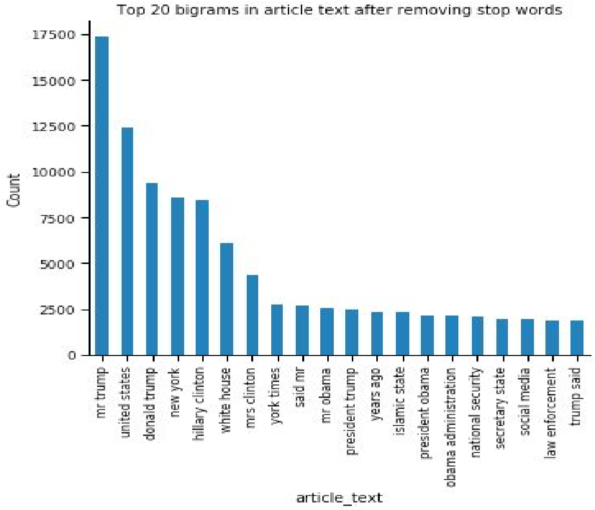
For all news articles text, the top 20 unigrams, bigrams and trigrams with and without removal of stop words was plotted. Without removal of stop words, the top word that occured in all articles was ‘the’. With the removal of stop words, the top word that occured was ‘said’.



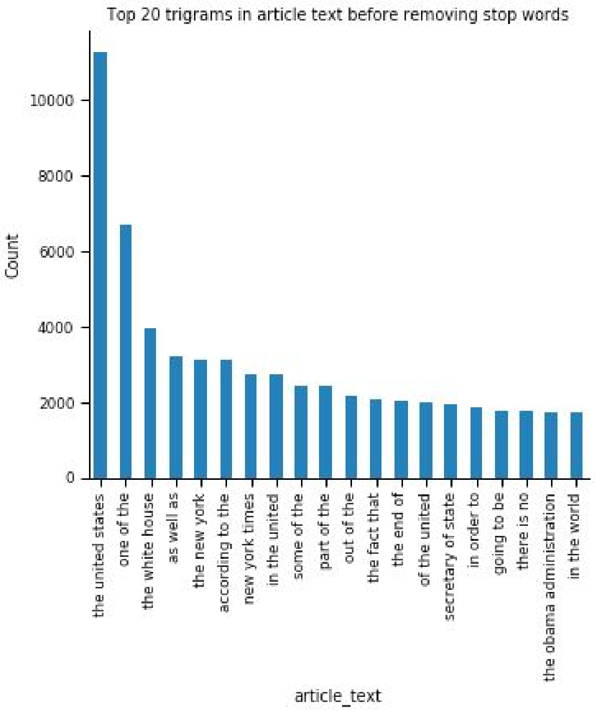


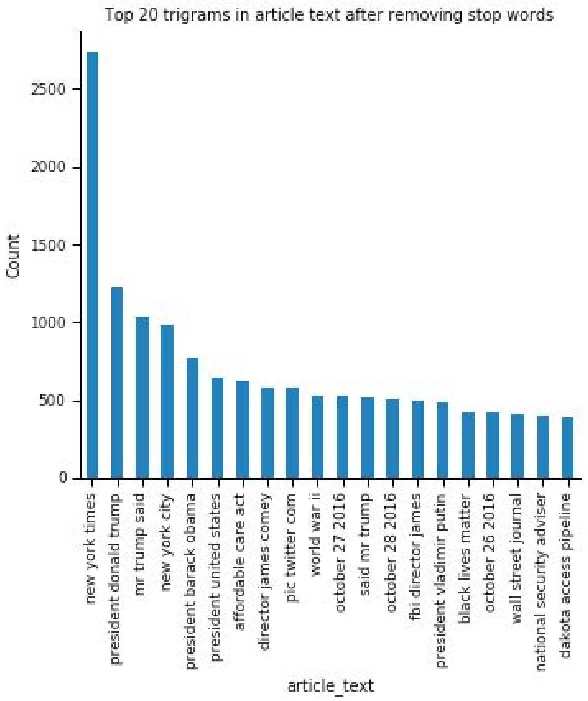
Without removal of stop words, the top bigram that occured in all articles was ‘of the’. With the removal of stop words, the top bigram that occured was ‘mr trump’.



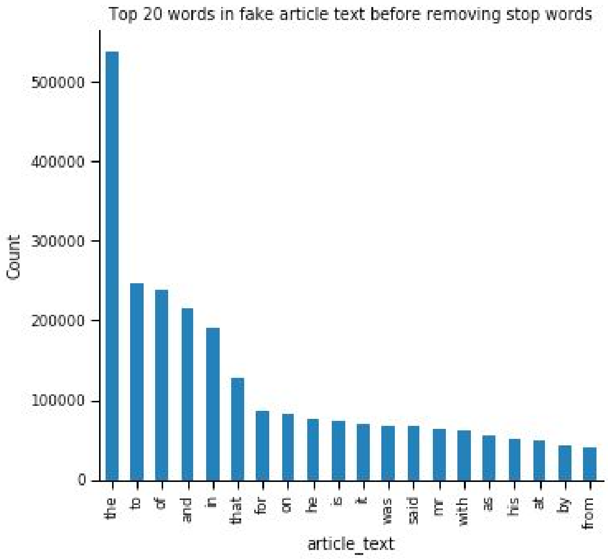


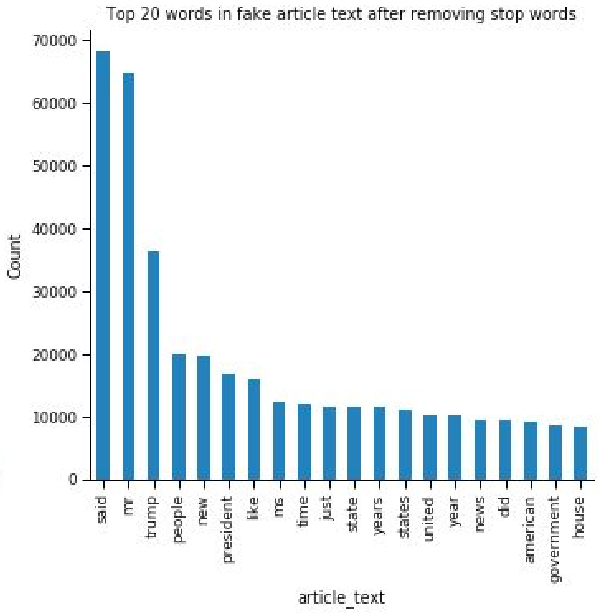
Without removal of stop words, the top trigram that occured in all articles was ‘the united states’. With the removal of stop words, the top trigram that occured was ‘new york times’.



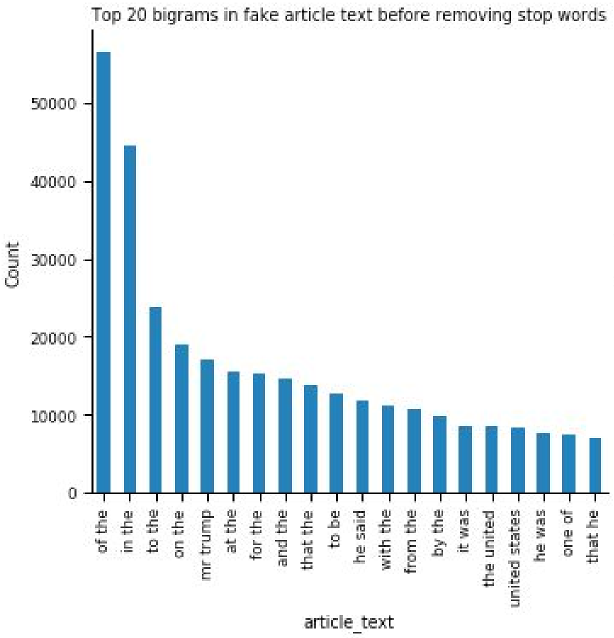


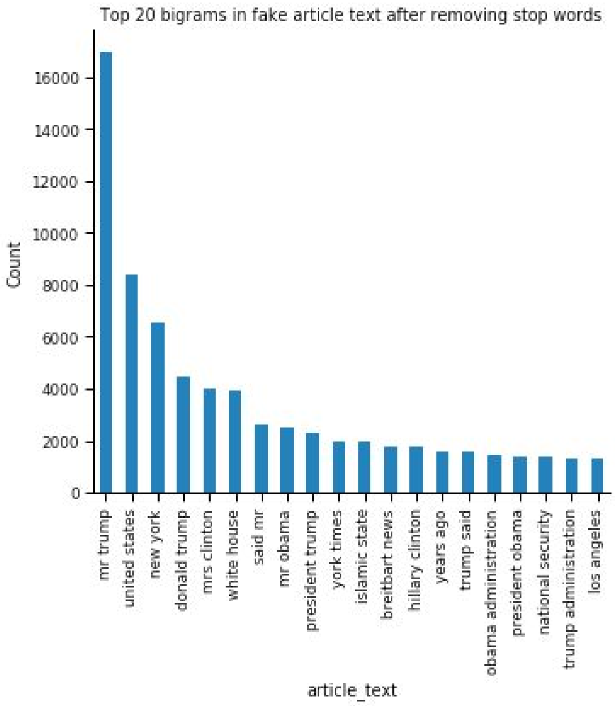
Next, for all fake articles text, the top 20 unigrams, bigrams and trigrams with and without removal of stop words were plotted. Without removal of stop words, the top word that occured in fake articles was ‘the’. With the removal of stop words, the top trigram that occured was ‘said’.



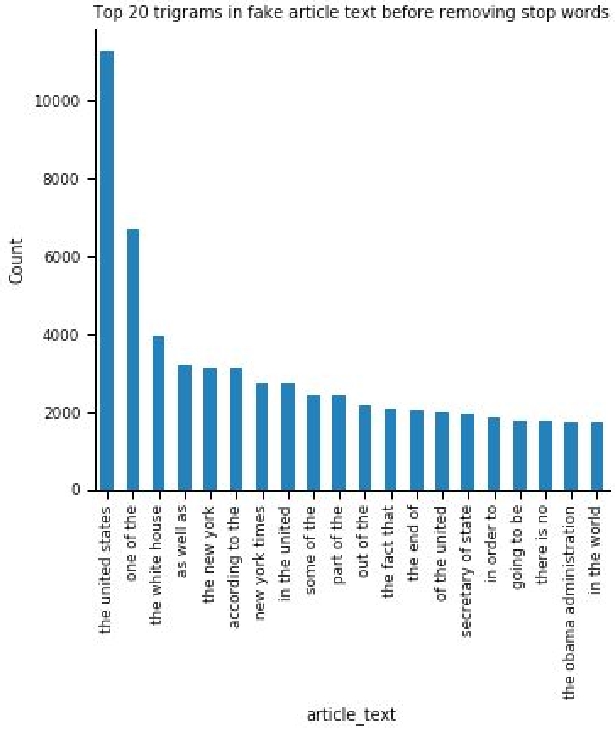


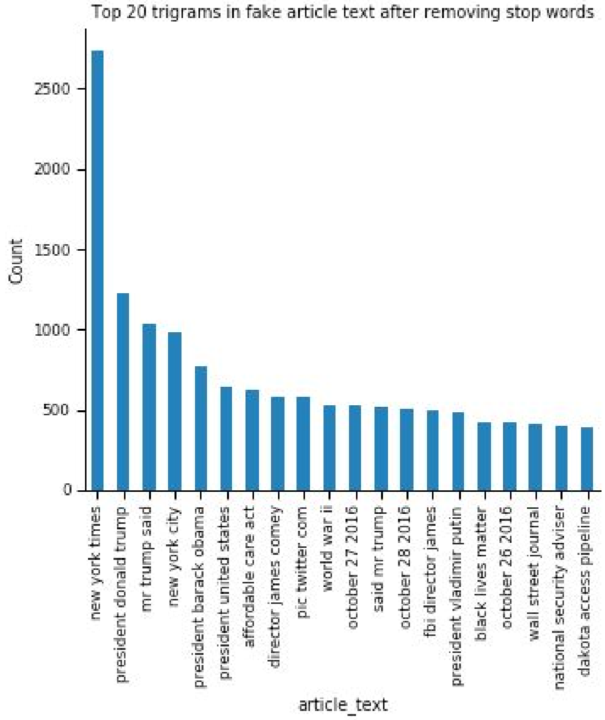
Without removal of stop words, the top bigram that occured in fake articles was ‘of the’. With the removal of stop words, the top bigram that occured was ‘mr trump’.



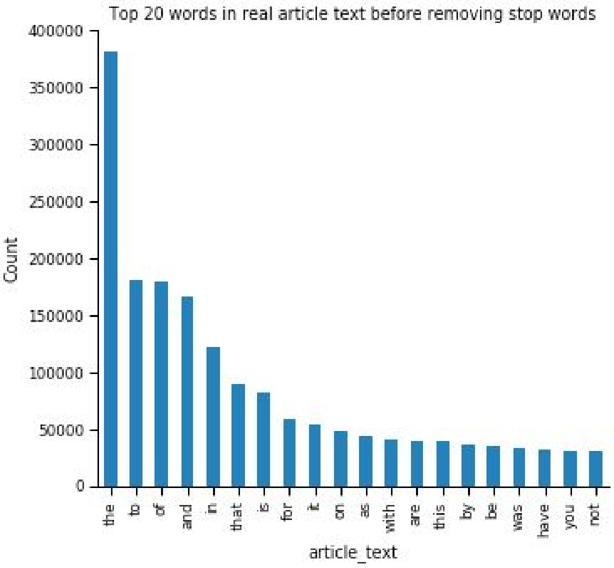


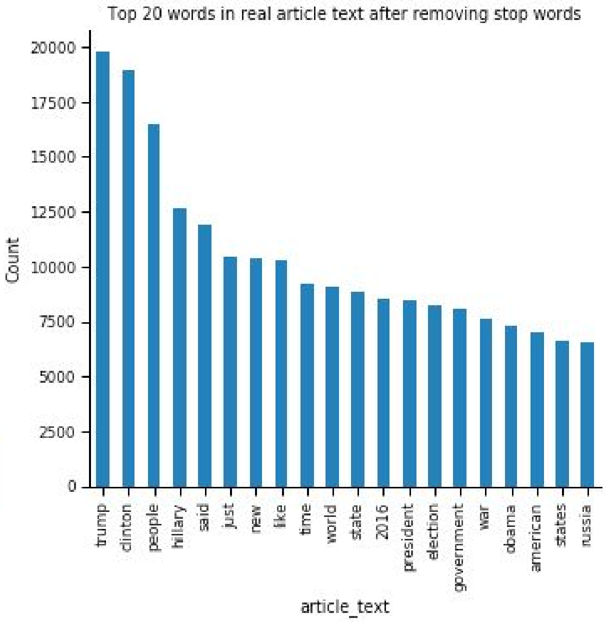
Without removal of stop words, the top trigram that occured in fake articles was ‘the united states’. With the removal of stop words, the top trigram that occured was ‘new york times’.



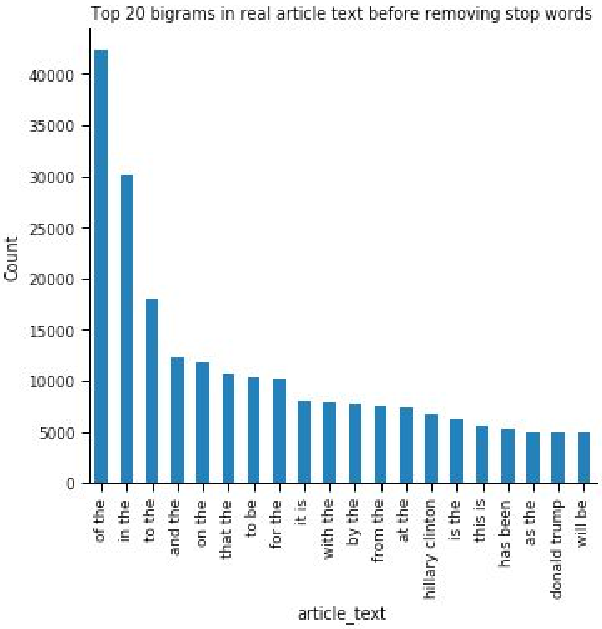


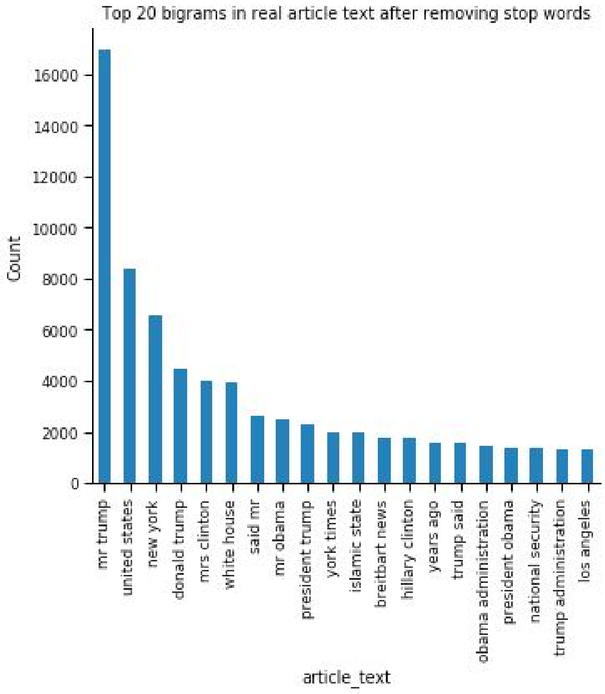
Finally, for all real articles text, the top 20 unigrams, bigrams and trigrams with and without removal of stop words were plotted. Without removal of stop words, the top word that occured in real articles was ‘the’. With the removal of stop words, the top word that occured was ‘trump’.



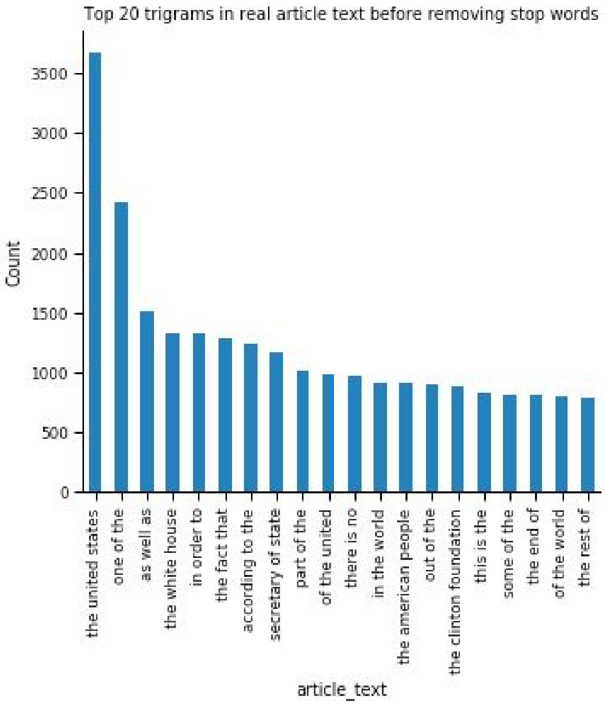


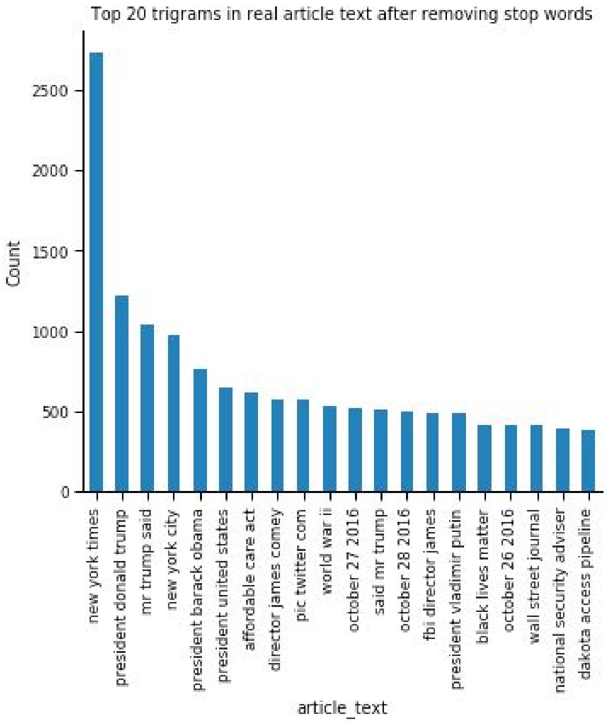
Without removal of stop words, the top bigram that occured in real articles was ‘of the’. With the removal of stop words, the top bigram that occured was ‘mr trump’.





Without removal of stop words, the top trigram that occured in real articles was ‘the united states’. With the removal of stop words, the top trigram that occured was ‘new york times’.





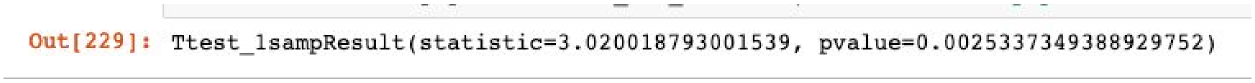
Next, a one-sample t-test was performed to test whether a population mean is significantly different from some hypothesized value. We are going to test to see whether the average text polarity of fake articles differs from the average text polarity of all articles.

Hypothesis Testing: Is there significant difference in the means of text polarity between articles that are fake and all text articles?

Null Hypothesis: The null hypothesis would be there there is no difference in text polarity between articles that are fake and all text articles.

Alternate Hypothesis: The alternative hypothesis would be that there is a difference in text polarity between fake articles and all articles.

The t-distribution left quartile range was -1.961129255575354 and right quartile range was 1.9611292557535396. The test result showed the test statistic 't' is equal to 3.0200. T is simply the calculated difference represented in units of standard error and tells us how much the sample mean deviates from the null hypothesis. The null hypothesis can be rejected if the t-statistic lies outside the quantiles of the t-distribution corresponding to the chosen confidence level and degrees of freedom.



```
Out[229]: Ttest_1sampResult(statistic=3.020018793001539, pvalue=0.0025337349388929752)
```

A p-value of 0.00253 means we'd expect to see data as extreme as our sample due to chance way less than 5% of the time if the null hypothesis was true. In this case, the p-value is lower than our significance level α (equal to 1-conf.level or 0.05) so we should reject the null hypothesis.

Based on the statistical analysis of a one sample t-test, there is significant difference between the mean text polarity of all articles and the mean text polarity of fake news articles. The low P-value of 0.0025 at a 5% confidence interval is a good indicator to reject the null hypothesis.