Capstone Project -2

Table of Content –

Abstracts – Recently there has been a topic of fake news detection on social media, where lots of posts are getting published by many companies and daily basis and in order to identify if there is a fake news or not its not very easy, so with help of Machine learning, we will develop a solution which can identify if this is a fake news or not.

Business Problem Description – In this era, where social media has so much dominance on knowledge and information across the globe, it is very important to identify if it is a fake or a genuine article, so that the knowledge and information is valuable and can a real education for the society.

1. With help of NLP (Natural Language processing), we will create a corpus of words from real and fake news articles. This corpus will be used to create a classifier model, which can predict the news/ article to be fake or real. With this model we can focus on the source of these articles and classify them with high confidence that the news or article coming from the source is real or fake.

Dataset Details –

There are 20800 and 5 attributes. Key features from the dataset are as below from the training dataset

Columns	Description
id	Identified/ Unique Id for a news articles
title	Title of a news articles
author	Author/ Source of the news articles
text	It is the text of the article; could be incomplete
label	Label that marks the article as potentially unreliable

Reference data source –

- https://www.kaggle.com/c/fake-news/data

Data Wrangling –

We have downloaded the dataset provided on the Kaggle; and with our analysis of the data, there are 20800 records in the training dataset and 5200 records in the test dataset. This dataset set has the Author, Title, text and label as the attributes in the dataset.

Due to Indexing already available, it looks like ID column is duplicate column, hence it been dropped from the dataset, as shown in The sample with positive values are the of same positive sentiments. The sample with positive values are the of same positive sentiments.1,2.

	id	title	author	text	label
0	0	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
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0

figure1

	id	title	author	text
0	20800	Specter of Trump Loosens Tongues, if Not Purse...	David Streitfeld	PALO ALTO, Calif. — After years of scorning...
1	20801	Russian warships ready to strike terrorists ne...	NaN	Russian warships ready to strike terrorists ne...

figure2

There are 10413 records, which are labeled as real / valid news and 10387 records are labeled as fake news. There are records, with null for Author, Title and Text columns.

In Training Dataset

```
title      558
author    1957
text        39
label        0
```

In Test Dataset

```
title      122
author     503
text         7
```

Data Preprocessing, Below in figure4, steps will be performing for the text – attribute, Remove Line Breaks element, remove new Line element, remove Hyperlink element, remove ampersand, remove greater than sign, remove less than sign, remove non-breaking space, remove Emails, remove new line characters, remove distracting single quotes.

	label	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...

figure4

In figure4.1, using these preprocessed text, we created the length attribute of the words in the text.

	label	title	author	text	textlen
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

figure4.1

Below in figure5, plotting the bar graph to see if check the null records in Author, Title and Text -

- Around 1957 records are null for Author
- Around 558 records are null for Title

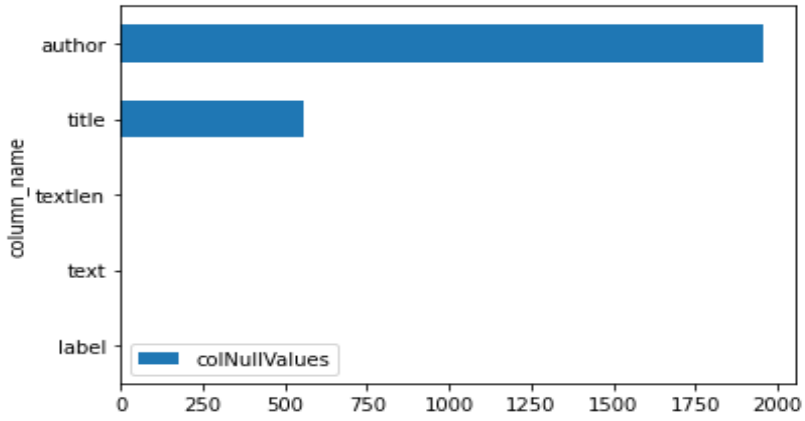

figure5

To handle the Null values, replaced Author and Title column as NA and dropped all the NA records. Also removed any text, which has less than 45 characters. Finally we are having 20563 records and 5 attributes.

As the dataset is ready, we have split the data between training and test data with 70:30 ratio. Created the Count Vector Training and Test dataset. Created the TFIDF train and test dataset for later in modeling section.

Data Visualization -

A new additional attribute is created to capture the sentiments from the text, used the sentiment polarity API to calculate the values. The values are calculated to -1 to 1, being 1 as positive sentiments and -1 as negative sentiments, as shown in figure6.

	label	title	author	text	textlen	sentiment
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	0.001796
1	0	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	4143	0.100880
2	1	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	7670	0.056258
3	1	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	3223	0.017497
4	1	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print An Iranian woman has been sentenced to s...	934	-0.012500

figure6

In figure7, Plotted the distribution of the sentiments score, it has close to normal distribution, as it seems, it has both positive and negative sentiments almost equally.

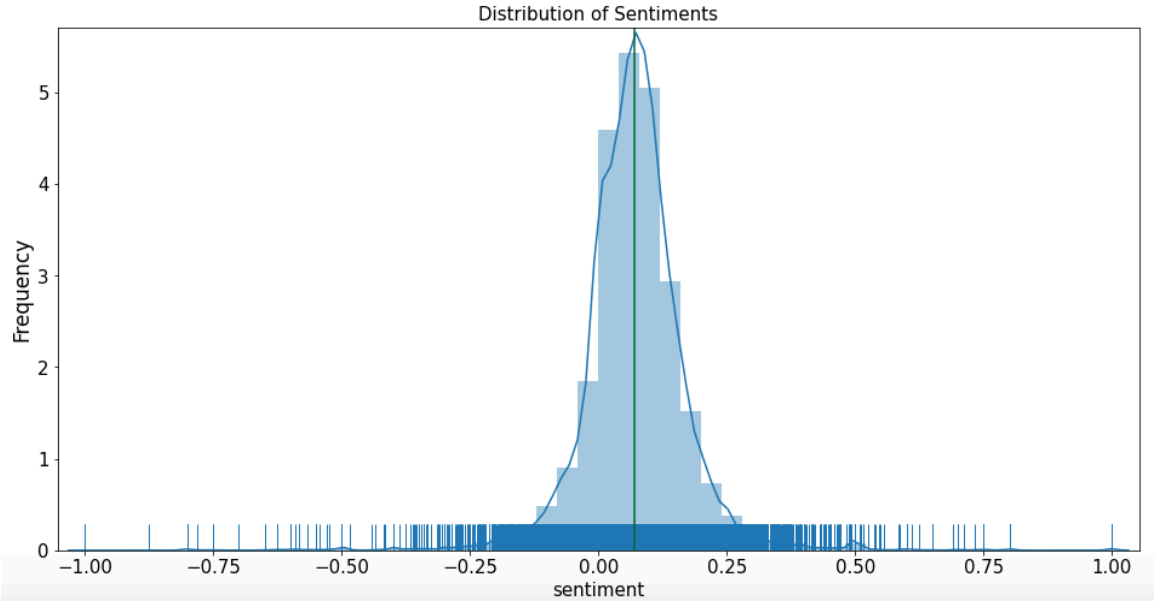

figure7

In figure8, Below is the PIE chart for fake and real news; it is almost same number records classified as fake and real news.

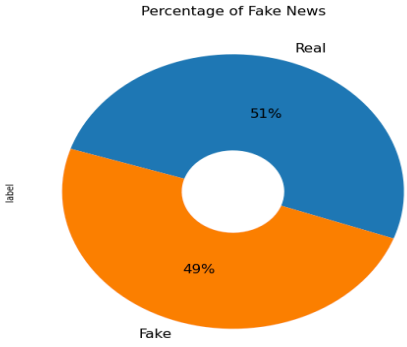

figure8

Next in figure9, is the correlation matrix, describing the relation between the attributes, the values of the correlation are between -1 and 1, showing positive and negative correlation. There is not strong correlation between any attributes, but there is a negative correlation of -0.12 between length and label.

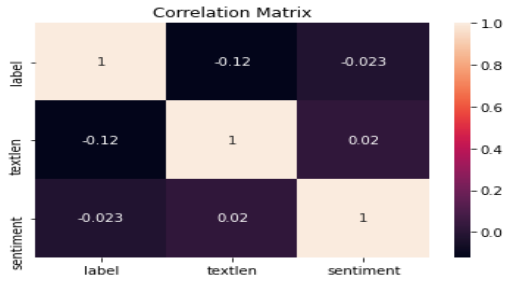

figure9

In fi10, WordCloud from Train Dataset, creating the word cloud of 50 most common words are “Obama”, followed by “Clinton” and “American”.

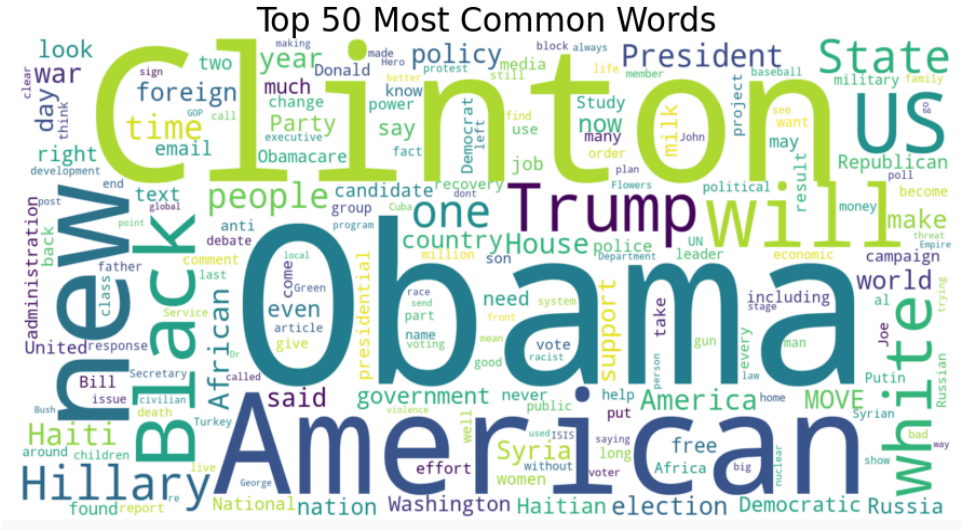

figure10

Creating the n-Gram plots for Unigram, Bigram and Trigram, in the unigram, the most common words after stopword are “said”, “mr” and “trump”. In the Bigram, we can see “mr trump”, “united states” and “donald trump”. In the trigram, we can see the common words are “new york times”, “president Donald trump ” and “mr trump said” as shown in figure11.

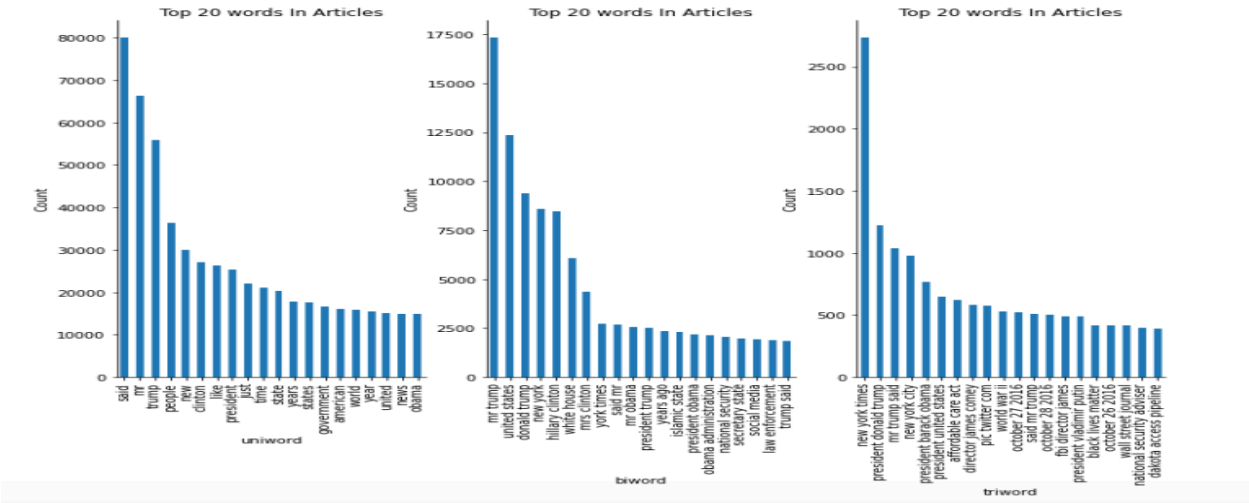

figure11

Creating the n-Gram plots for Unigram, Bigram and Trigram, in the unigram, the most common words after stopword and updating stopword, are “trump”, “will” and “one”. In the Bigram, we can see “united states” and “donald trump” and “new york”. In the trigram, we can see the common words are “new york times”, “president Donald trump ” and “new york city”, as shown in figure12.

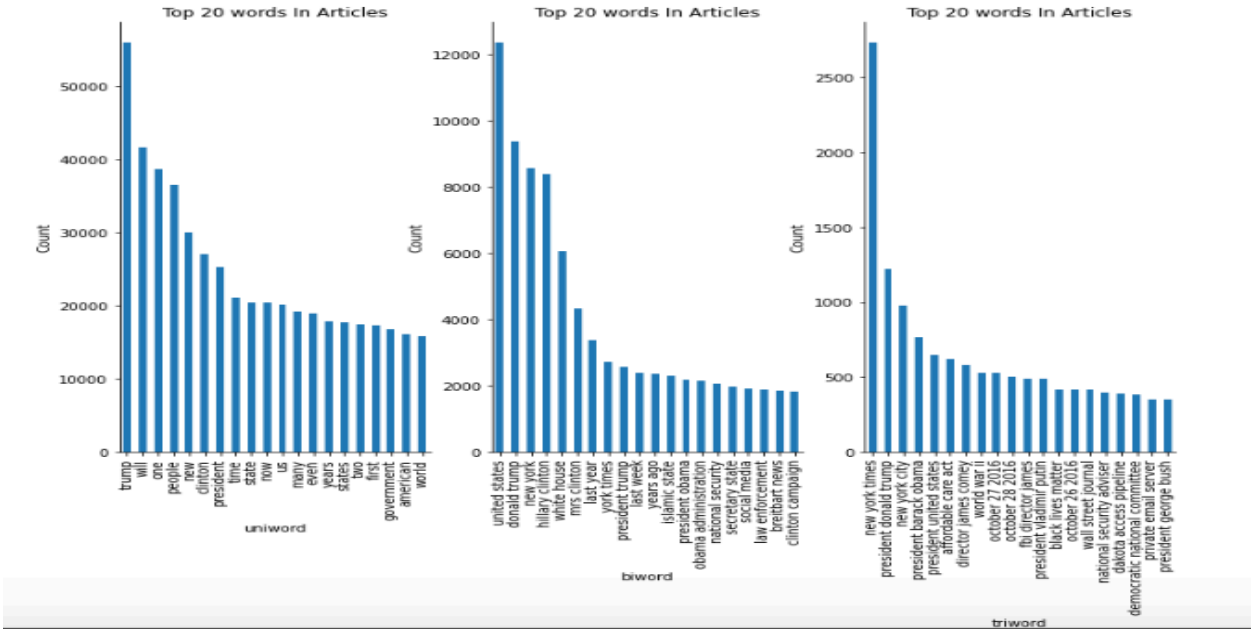

figure12

Inferential statistics techniques –

The dataset has 10385 fake news articles and legit/ valid news articles are 10178, out of total articles of 20563 records.

Calculate the T-Statistics of two independent **sentiments** sample from the population of fake news articles and real news articles. We have the below hypothesis –

Null Hypothesis – Both of the sample are same and equal, there is no difference in their sentiments analysis.

Alternate Hypothesis – Both of the samples are different and not equal and have difference in their sentiments analysis.

T-Statistics helps explain if the means from two samples are different from each other, by calculating the stand error in difference between two means. The critical value is calculated using degree of freedom and significance level with percent point function (PPF), if the critical value is greater than t-statistics, we reject the null hypothesis, else we accept the null hypothesis.

Another method is to calculate the p-value from the cumulative distribution function (CDF) from t-distribution, this p-values is compared to the alpha (significance level). If it is more than alpha, we accept the null hypothesis and if it is less, then reject the null hypothesis.

Values calculated from the t-distribution as –

The t-distribution left quartile range is: -1.9600793684470008. The t-distribution right quartile range is: 1.9600793684470004, as shown in figure14

- T-stats =3.249, degree of freedom=20561, cv=1.645, p=0.001, alpha = 0.05.
- Comparing the critical values to the t-stat, reject the null hypothesis that the means are equal.
- Comparing the p-value to alpha, reject the null hypothesis that the means are equal.

```
Values of t=3.249, df=20561, cv=1.645, p=0.001
Reject the null hypothesis that the means are equal.
Reject the null hypothesis that the means are equal.
The t-distribution left quartile range is: -1.9600793684470008
The t-distribution right quartile range is: 1.9600793684470004
Values of using API is t=3.264, p=0.001
```

figure14

Based, on the above details, the sentiments of fake and real news are different from each other. The p-v alue is less than 5% chance that both sentiments sample are same, so reject the null hypothesis.

Correlation between 'textlen' and 'sentiment':
Calculating for high correlation data from dataset, for 'textlen', 'sentiment', we find that the coeff values is 0.01971846321212139 and p-value is 0.004688487215723314. There is very less correlation (0.0197) between the text length and sentiments calculated. As seen in figure 14.a

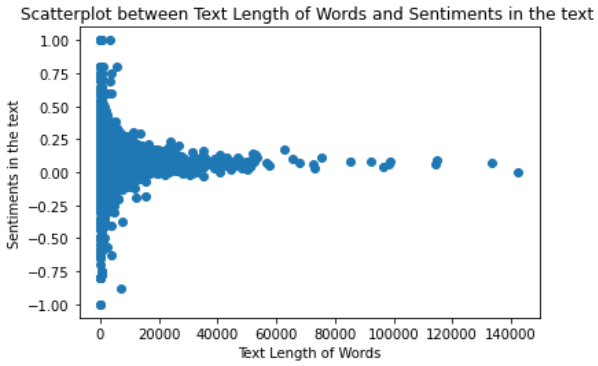

figure14.a

Detail on feature influencing the sentiments –

From figure 14.b, Coefficient table (middle table). We can interpret for textlen, coefficient (`3.965e-07`) first noticing that the p-value (under P>|t|) is 0.005, which is small. This means that the textlen is a stati stically significant predictor of sentiments.

The confidence interval of textlen gives us a range of plausible values for this average change, about (`1.22e-07, 6.71e-07`)

`R^2 is only 0.00`, F-Statistic is 7.998 and the probability for this statistic is 0.004.

OLS Regression Results

Dep. Variable:	sentiment	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	7.998
Date:	Sun, 02 Aug 2020	Prob (F-statistic):	0.00469
Time:	23:43:58	Log-Likelihood:	17608.
No. Observations:	20563	AIC:	−3.521e+04
Df Residuals:	20561	BIC:	−3.520e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0694	0.001	72.253	0.000	0.068	0.071
textlen	3.965e-07	1.4e-07	2.828	0.005	1.22e-07	6.71e-07

Omnibus:	4061.787	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	136432.375
Skew:	−0.048	Prob(JB):	0.00
Kurtosis:	15.619	Cond. No.	9.19e+03

figure14.b

Chi-Squared Test Statistics:

A chi-square is goodness of fit validation to find out of the sample data matches the population. It tests for independence of two variables in a contingency table to find out if they are related. It determines weather distribution of the categorical variables differ from each other.

A small value of chi-square stats means that there is a relationship and observed data fits the expected data very well. A large value of chi-square test stats doesn't have any relationship and observed data doesn't fit the expected data.

Null Hypothesis (H0): The sample with sentiment analysis and fake news are independent of each other.

Alternate Hypothesis (H1): The sample with sentiment analysis and fake news are not independent of each other.

Created the new column as positive sentiments in the dataset. Created a new view by grouping Sentiment and fake and real news, as shown in figure 14.c.

label	0	1
possen		
0	1429	2365
1	8956	7813

Figure 14.c

After applying the chi-square test, we found the below results

Stats =306.143 and pvalue=0.000 and dof=1

Expected
[[1916.09638671 1877.90361329]
[8468.90361329 8300.09638671]]

We can conclude that we can reject the Null Hypothesis of sentiments and fakes news are independent of each other. There is a clear deviation

The Fake news tends to have negative sentiments compared to real news which has most of positive sentiments.

Model Implementation –

In the data wrangling section, training dataset is split into train and validation/test datasets. The training dataset is created for COUNT vectors, TF-IDF vectors. In this section, we need to develop a model to predict fake or legit news based on the historical data collected in the training set with labels.

Two methods – CountVectorizer and TF-IDF Vectorizer, are used as Count Vectorizer provides the document term matrix, which is transposed tokens, or words in features with count of occurrence of each word.

TF-IDF (Term Frequency-Inverse Document Frequency), helps downgrade the weights of highly frequent words.

Model created for Logistics Regression with Count Vectors, Logistics Regression with TF-IDF Vectors, Multinomial Naïve Bayes classifier with Count Vectors with hyper parameter and Multinomial Naïve Bayes classifier with TF-IDF Vectors with hyper parameter. The hyper parameter tuning is done using "GridSearchCV". The Best parameters resulted as alpha = 0.1.

As shown in figure 15, Accuracy scores of each of the algorithms -

	ModelName	Accuracy Score
0	Logistic Regression with Count Vectorizer	0.954651
2	Multinomial NB with Count Vectorizer	0.909528
3	Multinomial NB with TF-IDF	0.909285
1	Logistic Regression with TF-IDF	0.875636

figure15

Best results from the above accuracy score is of – Logistic Regression with Count Vectors. Below we can see the ROC Accuracy is 0.95.

In the confusion matrix, as shown in figure16

1. True positive is 2944 and True Negative is 2945.
2. Type I – False negative is 120 and Type II – False positive is 160.

```
ROC Accuracy Score = 0.95 --
Accuracy Score = 0.95 --

confusion_matrix
[[2944  160]
 [ 120 2945]]

classification_report
              precision    recall  f1-score   support

           0       0.96      0.95      0.95      3104
           1       0.95      0.96      0.95      3065

   micro avg       0.95      0.95      0.95      6169
   macro avg       0.95      0.95      0.95      6169
weighted avg       0.95      0.95      0.95      6169
```

figure16

ROC (Receiver Operating Characteristics) curve with area under curve (AUC) is a measure of how well model is performing in predicting probability of classes. The False positive rate (FPR on x-axis) and True Positive Rate (TPR on y-axis) is plotted. Higher the True positive rate and the curve is more towards 1 on y-axis is considered to be best model.

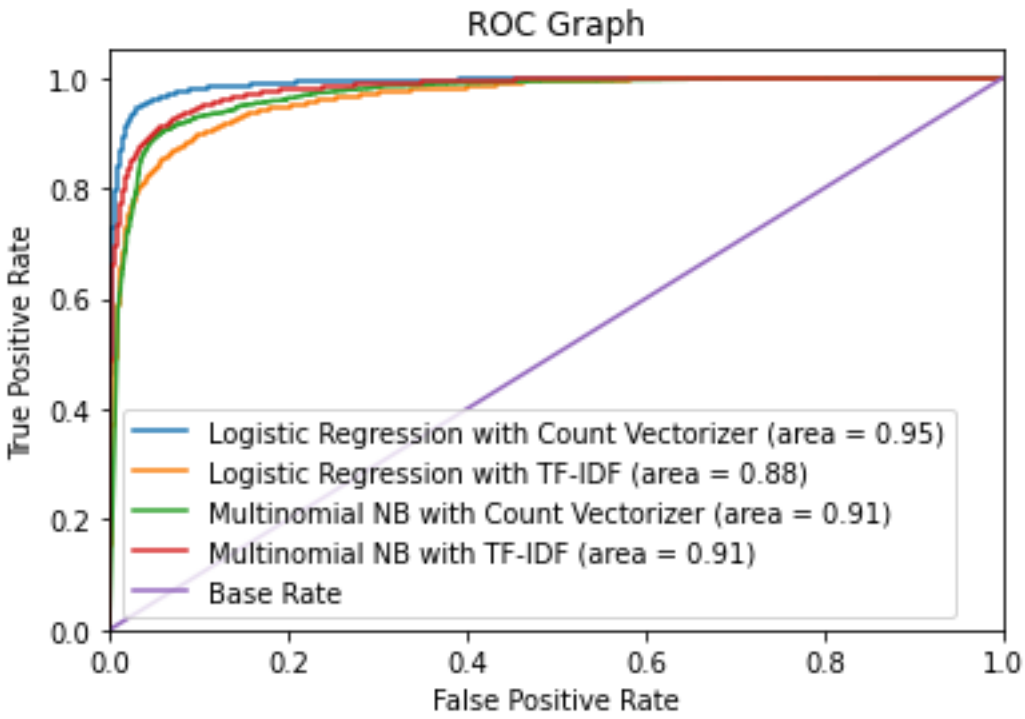

figure17

In Figure17, we can see that Logistic Regression with Count vector has highest AUC-ROC score of 95%.

Conclusion -

There are 20800 records in the training dataset and 5200 records in the test dataset. There are 10413 records, which are labeled as real / valid news and 10387 records are labeled as fake news.

Data Preprocessing, was performed on text – attribute with, remove Line Breaks element, remove new Line element, remove Hyperlink element, remove ampersand, remove greater than sign, remove less than sign, remove non-breaking space, remove Emails, remove new line characters, remove distracting single quotes.

Created the Count Vector Training and Test dataset with split of 70:30 ration for training and test dataset. Created the TFIDF train and test dataset for later in modeling section.

“n-Gram” plots for Unigram, Bigram and Trigram, in the unigram, the most common words after stopword and updating stopword, are “trump”, “will” and “one”. In the Bigram, we can see “united states” and “donald trump” and “new york”. In the trigram, we can see the common words are “new york times”, “president Donald trump ” and “new york city”.

T-Statistics of two independent **sentiments** sample from the population of fake news articles and real ne ws articles. T-stats =3.249, degree of freedom=20561, cv=1.645, p=0.001, alpha = 0.05. Comparing the c ritical values to the t-stat, reject the null hypothesis that the means are equal. Comparing the p-value to alpha, reject the null hypothesis that the means are equal.

Models created for Logistics Regression with Count Vectors, Logistics Regression with TF-IDF Vectors, Multinomial Naïve Bayes classifier with Count Vectors with hyper parameter and Multinomial Naïve Bayes classifier with TF-IDF Vectors with hyper parameter. The hyper parameter tuning is done using "GridSearchCV". The Best parameters resulted as alpha = 0.1.

Best results from the above accuracy score are of – Logistic Regression with Count Vectors with ROC Accuracy is 0.95 or AUC-ROC score of 95%.

Next Work –

We can create the weights and feature importance and use that to predict the fake and legit news. We can improve the AUC-ROC score using any ensemble algorithms.

Code –

https://github.com/arijitsinha80/Springboard/blob/master/Project2/CapstoneProject2-FakeNewsPrediction.ipynb

PPT –

https://github.com/arijitsinha80/Springboard/blob/master/Project2/Capstone%20Project%202_Final.pptx

Reference –

https://towardsdatascience.com/nlp-part-3-exploratory-data-analysis-of-text-data-1caa8ab3f79d

https://www.kaggle.com/aaroha33/fake-news-classifier-with-naive-bayes

https://machinelearningmastery.com/how-to-code-the-students-t-test-from-scratch-in-python/