Kavya Mavuru
11413627 | Spring 2021

Covid-19 Fake News Detection

Introduction and Background of the Project:

From the report's cases in Wuhan, China, researchers have worked nonstop to assemble data about this new COVID. Yet, with new data comes deception. There have been numerous conceivably hazardous hypotheses identified with COVID, going from the new COVID being human-made to the possibility that infusing blanch or different disinfectants could ensure against disease. This sort of fake news makes numerous individuals alarm everywhere in the world. It is an extremely broad issue that even the most driving media once in a while gets into the snare of fake News. So as to gear these issues, a supported model will be developed using Machine learning and NLP concepts.

This project mainly focuses on Fake News Detection on COVID 19 and for this when we give any news of COVID 19 as input then our model will tell that news is fake or not. To perform these, we will use NLP concepts and Machine Learning Algorithms like MultonomialNB Classifier, SVM Classifier and Passive Aggressive Classifier.

Statement of the Project Problem:

Fake News is one of the major concerns right now. It is a very wide spread issue that even the most leading media sometimes gets into the trap of Fake News. The proposed project, need for investigation is to determine the COVID news generated daily is real or fake. Initially, the data collected from news headlines is cleaned and composed for analysis. Then the data is analyzed using Natural Language Processing (NLP) and Machine Learning (ML) which results to predict from the information if the news is fake or real. Later on, this model is tuned and integrated into the User Interface where an input of news is given then the result fake or real will be exhibited. The aim is to call out public attention to the problems of COVID-19-related misinformation and work together to help develop accurate detection and deterrence of such misinformation.

Review of Literature:

The construction of fake news dataset aims to extract useful features that can better distinguish fake news from true ones. Researchers have proposed a series of methods to extract news and verify the truthfulness of the news Huge research has been performed on COVID 19 Fake News Detection in the past and one of them is Tweet Analysis for COVID-19 Fake News Detection, tweets from Twitter has been collected and these tweets are determined if they project fake or real news about COVID. In these six baseline machine learning models like Passive-Aggressive learning models are LSTM, RNN, and BERT. Out of all these models it is concluded that the BERT model performs better than others [1]. Other Research work performed on this topic is Infoveillance Study on Twitter and Instagram. This research has been conducted in two phases, one data has been scrapped from Twitter and Instagram and in a second phase, NLP and Deep Learning to identify potential sellers that were then manually annotated for characteristics of interest [2].

Objectives of the Study:

As the pandemic continues to create uncertainty there is an urgent need to take powerful action to manage wild spread of fake news headlines on COVID. To provide a solution to this, an Initiative is taken to design a model that it can be helpful for researchers to combat COVID health misinformation.

Data Collection:

Dataset was taken from Dataset consist of 10202 Rows and 2 columns. One column represents Headline, we can find various news headlines on COVID 19 and in other, target variable there are 2 variables. 0 and 1, where 0 represents Fake and 1 represents it is true.

Dataset URL: https://zenodo.org/record/4282522#.YIYEUlVKg2z

	headlines	outcome
0	A post claims compulsory vacination violates t...	0
1	A photo claims that this person is a doctor wh...	0
2	Post about a video claims that it is a protest...	0
3	All deaths by respiratory failure and pneumoni...	0
4	The dean of the College of Biologists of Euska...	0
...	...	...
10196	A Chinese market caused the new coronavirus (v...	0
10197	The peak of the new coronavirus will happen in...	0
10198	Stores and supermarkets in Veracruz (Mexico) w...	0
10199	A chain message circulated on Tuesday, Jan. 14...	0
10200	Photo shows Muslims in Tamil Nadu state of Ind...	0

10201 rows × 2 columns

Exploratory data analysis:

a) Checking Missing values, datatypes, shape of the dataset and column names in the dataset

```
[5]  df2.shape

     (10201, 2)

[6]  df2.columns

     Index(['headlines', 'outcome'], dtype='object')

[7]  df2.dtypes

     headlines    object
     outcome       int64
     dtype: object

[12] df2.isnull().sum()

     headlines    0
     outcome      0
     dtype: int64
```

b) Fake Headlines observation

```
df2[df2["outcome"]==0]
```

	headlines	outcome
0	A post claims compulsory vacination violates t...	0
1	A photo claims that this person is a doctor wh...	0
2	Post about a video claims that it is a protest...	0
3	All deaths by respiratory failure and pneumoni...	0
4	The dean of the College of Biologists of Euska...	0
...	...	...
10196	A Chinese market caused the new coronavirus (v...	0
10197	The peak of the new coronavirus will happen in...	0
10198	Stores and supermarkets in Veracruz (Mexico) w...	0
10199	A chain message circulated on Tuesday, Jan. 14...	0
10200	Photo shows Muslims in Tamil Nadu state of Ind...	0

9727 rows × 2 columns

c) True Headlines observation

```
df2[df2["outcome"]==1]
```

	headlines	outcome
1893	“3.8% of Wisconsin’s coronavirus funding has b...	1
1897	There’s a “direct correlation” between North C...	1
1898	“There have been five randomized controlled, p...	1
1900	“Five veterinary labs have their CLIA certific...	1
1910	Say Wisconsin Republican lawmakers have done n...	1
...	...	...
8670	Post says Harvard scientists say the coronavir...	1
8748	Says 80% of novel coronavirus cases are â€œmil...	1
8915	â€œWith regard to the cost, let me be very cle...	1
9019	The United States is â€œactually screening few...	1
9795	“You’re more likely to die of influenza right ...	1

474 rows × 2 columns

d) Data Pre-Processing

Steps for preprocessing my data.

- Tokenization
- Stemming:
- Lemmatization
- Stop Words Removal

- Rejoining Tweets

Once above steps are performed then we will clean our headlines by removing punctuations, Unicode, Numbers.

```
df2["headlines"]=df2["headlines"].str.lower()
df2["headlines"]=df2["headlines"].str.replace("[^a-zA-Z]"," ") #Removing Punctuations
df2["headlines"]=df2["headlines"].str.encode('ascii', 'ignore').str.decode('ascii') #Removing Unicodes
df2["headlines"]=df2["headlines"].str.replace('\d+','')#Removing numbers
df2['headlines']=df2.apply(identify_tokens, axis=1)
df2['headlines']=df2.apply(stem_list, axis=1)
df2["headlines"]=df2.apply(lem_list, axis=1)
df2['headlines']=df2.apply(remove_stops, axis=1)
df2['headlines']=df2.apply(rejoin_words, axis=1)
```

Cleaned Dataset:

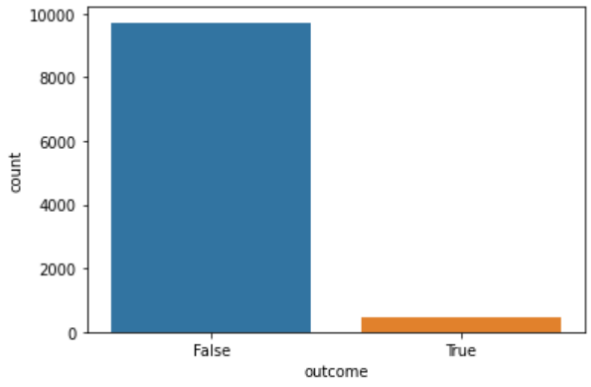
	headlines	outcome
0	post claim compulsori vacin violat principl bi...	0
1	photo claim thi person doctor die attend mani ...	0
2	post video claim protest confin town aranda de...	0
3	death respiratori failur pneumonia regist covi...	0
4	dean colleg biologist euskadi state lot pcr fa...	0
...	...	...
10196	chine market caus new coronaviru video	0
10197	peak new coronaviru happen two week jan dure t...	0
10198	store supermarket veracruz mexico close due ne...	0
10199	chain messag circul tuesday jan warn peopl avo...	0
10200	photo show muslim tamil nadu state india float...	0

10201 rows × 2 columns

Checking Target variable value_counts:

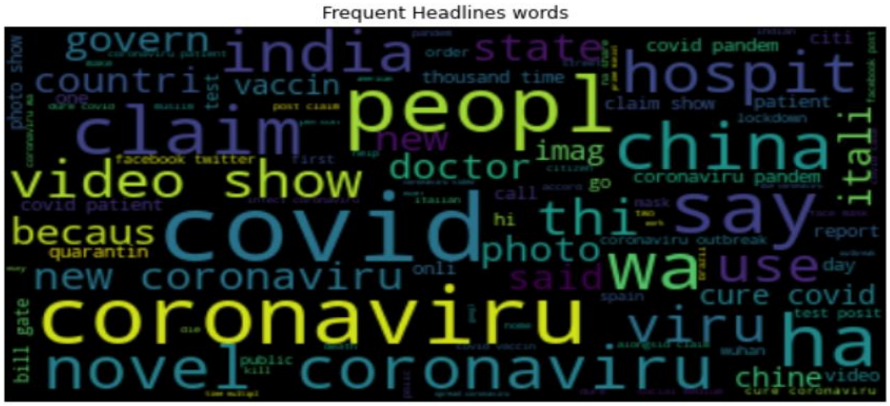
```
ax = sns.countplot(x="outcome", data=df2)
ax.set_xticklabels(["False","True"])
```

```
[Text(0, 0, 'False'), Text(0, 0, 'True')]
```



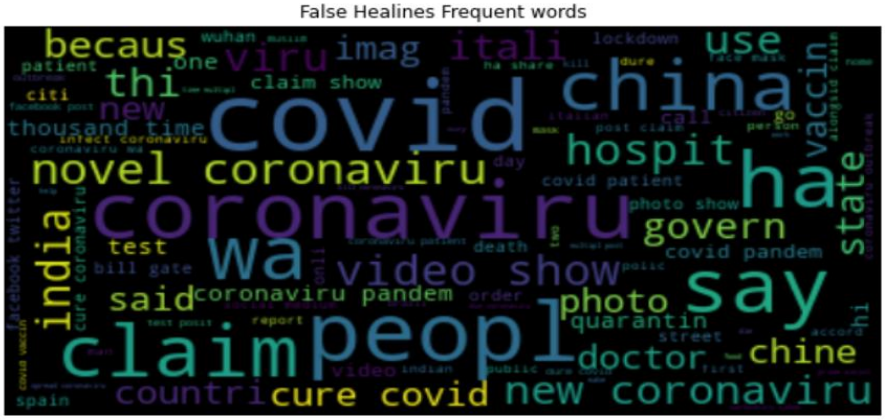
In this we can see False value are high and True values are less. Clearly, we can say data is imbalanced.

Frequent Headline Words:



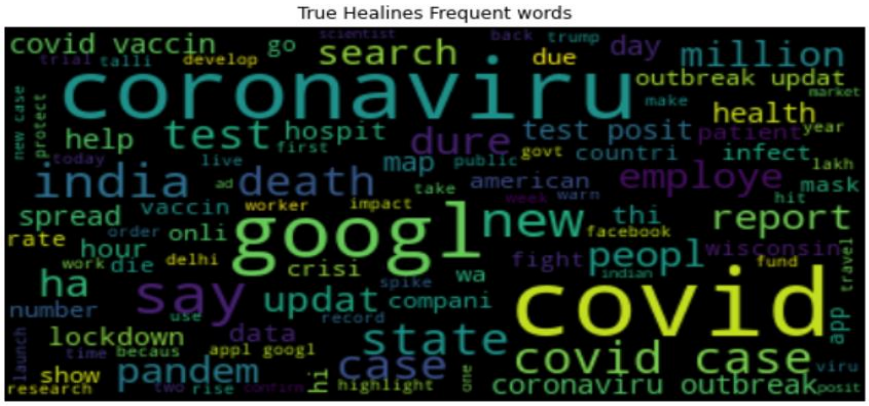
From the above Word cloud we can see that in most of the headlines words like Covid, Claim, India, novel, people are repeated. By this we can think that most headlines are related to these words.

False Headlines Frequent Words:



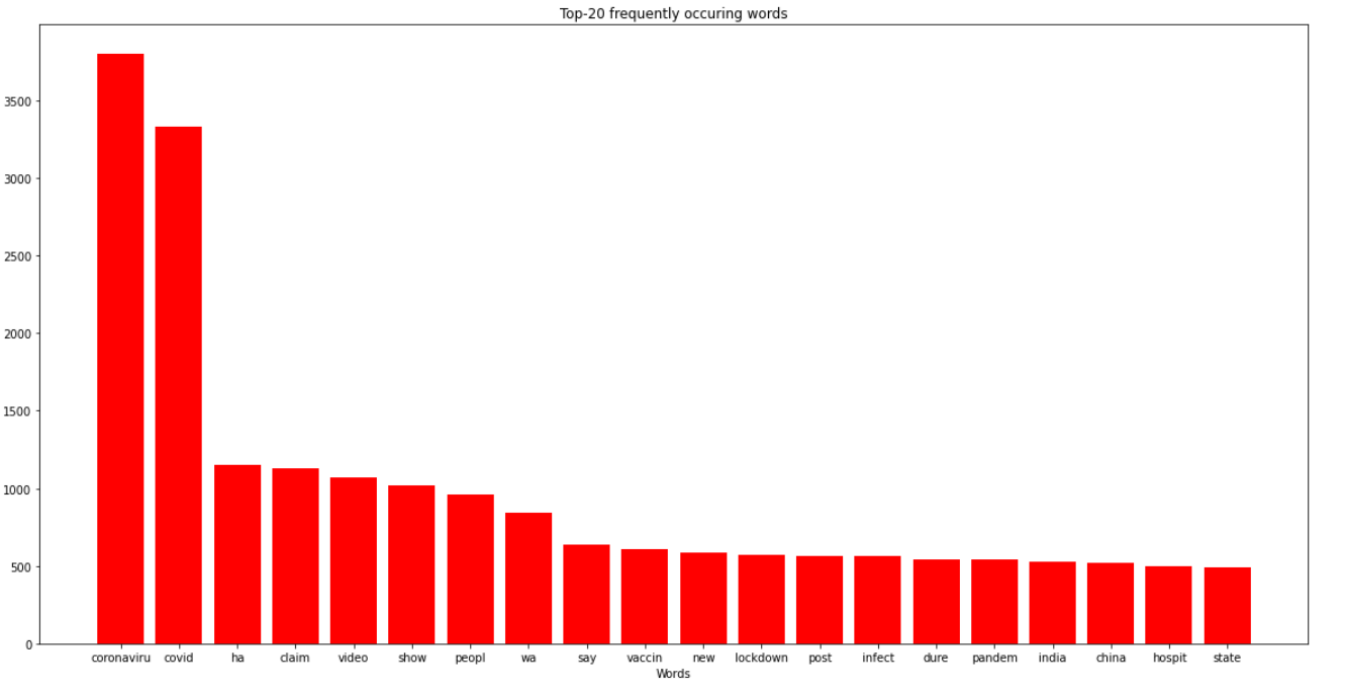
From the above Word cloud we can see that in most of the headlines words like Covid, China, use, claim, coronaviru are repeated. By this we can think that most False headlines are related to these words.

True Headlines Frequent words:

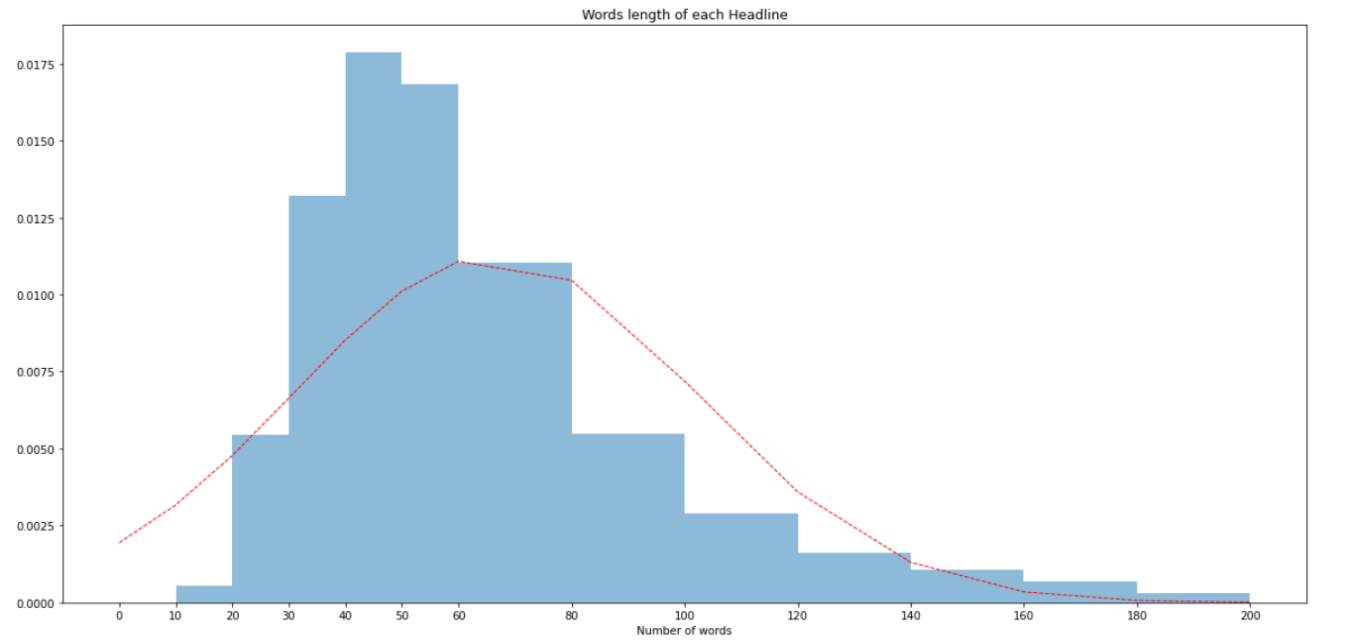


From the above Word cloud we can say that in most of the headlines words like Covid, search, googl, case, coronaviru are repeated. By this we can think that most True headlines are related to these words.We cannot say that because of this words there is impact on the headlines. In further we will analyze more.

Top 20 frequently occurring words:



Word Length of Each Headline



When we observe above distribution graph, in most of the headlines the words length is between 40 to 60 and we can also see there are many outliers fall on the right side of the curve. We can also think like if we have many True words in a headline then definitely that headline will goes to True headlines category and wise-versa.

We can consider some other issues like if any particular headline is false but due to high number of True words that headline will goes to True category and because of this many people will consider this news as True headline.

Hypothesis:

Null Hypothesis(H0) : Words in each headline will decide headline is spam or ham.

Alternative Hypothesis (H1): Words in each headline will not decide headline is spam or ham.

Research Design and Methodology:

The required data COVID news headlines are from online resources. Further the data is cleaned using python techniques and NLP methods are used to analyze. Machine Learning MultonomialNB Classifier, SVM Classifier and Passive Aggressive Classifier are used to create fake news detector engine. At the final stage the model will integrate with UI.

Data Analytics:

Before training models on the train data, we need to convert all out headlines into vectors so that machine can understand well.

For converting words of each headlines into vectors we use TF-IDF model.

TF-IDF is feature extraction technique which stands for Term Frequency-Inverse Document Frequency.

It is very useful for information retrieval to represent how important a specific word or phrase is to a given document.

It uses techniques like cosine similarity which works well on vectors for extracting features.

```
from sklearn.feature_extraction.text import TfidfVectorizer
Tfidf_vect = TfidfVectorizer(ngram_range=(1,2),max_features=1000)
Tfidf_vect.fit(df2['headlines'])
Train_X_Tfidf = Tfidf_vect.transform(Train_X).toarray()
Train_X_Tfidf_final_1=pd.DataFrame(Train_X_Tfidf,columns=Tfidf_vect.get_feature_names())
Train_X_Tfidf_final=pd.merge(df_TrainX_final,Train_X_Tfidf_final_1,left_index=True, right_index=True)
Train_X_Tfidf_final=pd.merge(Train_X_Tfidf_final_1,Train_Y,left_index=True, right_index=True)
Test_X_Tfidf = Tfidf_vect.transform(Test_X).toarray()
Test_X_Tfidf_final_1=pd.DataFrame(Test_X_Tfidf,columns=Tfidf_vect.get_feature_names())
Test_X_Tfidf_final=pd.merge(df_TestX_final,Test_X_Tfidf_final_1,left_index=True, right_index=True)
Test_X_Tfidf_final=pd.merge(Test_X_Tfidf_final_1,Test_Y,left_index=True, right_index=True)
```

In the below screenshot we can observe that all headlines are converted into vectors now we can consider all these vectors as dependent variables and outcome as out target variable. Now, we will try some machine learning models to train on this vectors and then predict outcome on test data.

```
[60] Test_X_Tfidf

     array([[0., 0., 0., ..., 0., 0., 0.],
            [0., 0., 0., ..., 0., 0., 0.],
            [0., 0., 0., ..., 0., 0., 0.],
            ...,
            [0., 0., 0., ..., 0., 0., 0.],
            [0., 0., 0., ..., 0., 0., 0.],
            [0., 0., 0., ..., 0., 0., 0.]])
```

Classifying Real and Fake News:

TF-IDF classifies below words as real words**.**

```
names=Tfidf_vect.get_feature_names()
sorted(zip(classifier.coef_[0],names),reverse=True)[:10] #Real Words
```

```
[(19.825038800959575, 'googl'),
 (10.461330426283663, 'life'),
 (10.251839241792236, 'mobil'),
 (9.816297119185133, 'launch'),
 (9.123647033977917, 'vote'),
 (9.086644026066129, 'updat'),
 (8.971635087746005, 'app'),
 (8.82109033437075, 'social medium'),
 (8.74254862520873, 'long'),
 (8.224829763285877, 'ten')]
```

TF-IDF below words as fake words.

```
sorted(zip(classifier.coef_[0],names))[:10] #Fake Words
```

```
[(-13.249813369325823, 'everi'),
 (-10.223862745985786, 'due covid'),
 (-9.947718708355081, 'coronaviru crisi'),
 (-9.931870014762588, 'fight coronaviru'),
 (-9.210943081437224, 'govern'),
 (-8.608017170535305, 'governor'),
 (-8.591967166120963, 'minist health'),
 (-8.291112697013212, 'video'),
 (-7.995755899845641, 'thousand'),
 (-7.602248251374526, 'medic')]
```

Procedure for model training and predicting:

- In general, we will split the dataset into training and testing and then we will train our model on training data and then we will test it on test data.
- I have considered 3 models like MultinomialNB Classifier, SVM, Passive Aggressive model for this project
- Now, we will fit our models on train data with suitable parameters.
- Once model got trained then we will predict outcome on test data.

```
[54] from sklearn import model_selection, naive_bayes
     Train_X, Test_X, Train_Y, Test_Y = model_selection.train_test_split(df2['headlines'],df2['outcome'],test_size=0.3) #30 percent test
```

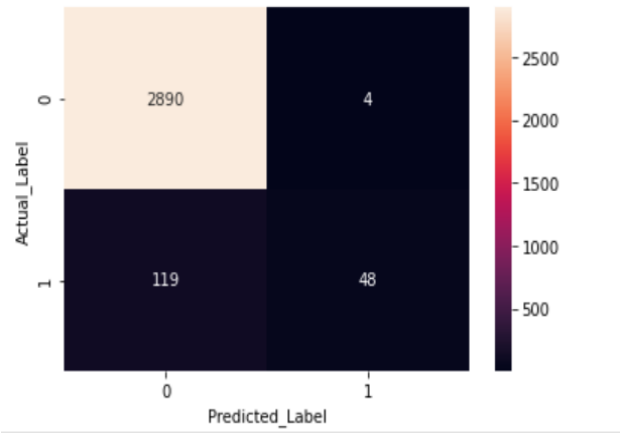
MultinomialNB Classifier:

Multinomial Naïve Bayes Classifier is computationally very efficient, and it is frequently used in text classification problems.

Multinomial Naïve Bayes consider a feature vector where a given term represents the number of times it appears or very often.

In this project I have considered alpha value as 0.5 and then trained my model on train data.

- Results:



- If our dataset is imbalanced, we need to consider precision and Recall as our classification metrics.

- When we observe confusion matrix False Negative(FN) values are very high we need to decrease False Negative values then only we can build powerful spam news detection engine.

	precision	recall	f1-score	support
0	0.96	1.00	0.98	2894
1	0.92	0.29	0.44	167
accuracy			0.96	3061
macro avg	0.94	0.64	0.71	3061
weighted avg	0.96	0.96	0.95	3061

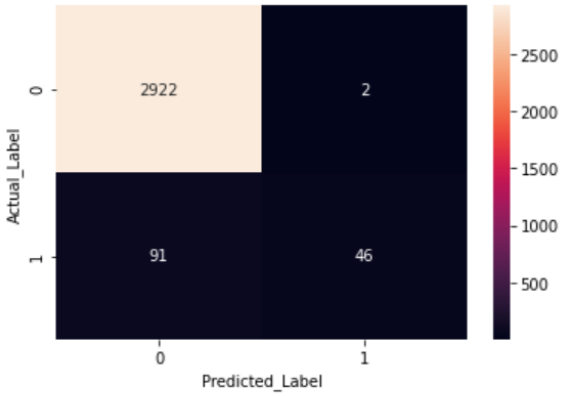
In this we can observe that how badly our model is working, my recall score is just 29%.

I need to improve my recall score for that I will try some other models like SVM and Passive Aggressive model.

Support Vector Machine:

SVM is a supervised machine learning algorithm, and it is used for both classification and regressions problems.

SVM performs classification by finding the hyper-plane that differentiate the classes we plotted in n-dimensional space.



When we observe confusion matrix False Negative(FN) values are very high we need to decrease False Negative values

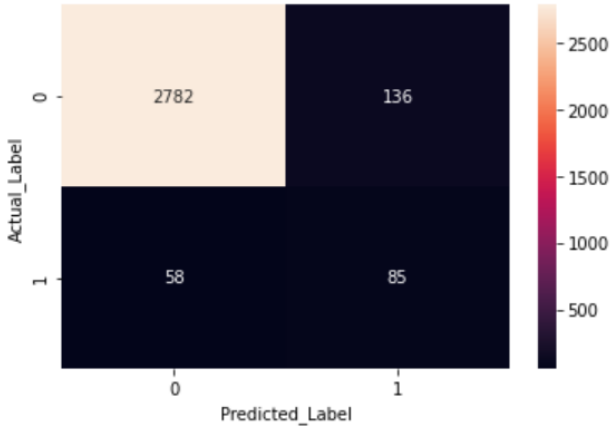
There is increase in model performance when we compare it with above model.

	precision	recall	f1-score	support
0	0.97	1.00	0.98	2918
1	0.94	0.36	0.52	143
accuracy			0.97	3061
macro avg	0.96	0.68	0.75	3061
weighted avg	0.97	0.97	0.96	3061

There is increase in recall score but still we will try to increase recall score by trying other models like Passive Aggressive model.

Passive Aggressive Model:

Passive Aggressive model is online learning model. Algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting.



From confusion matrix we can say that FN value is still 58 but when we compare with above models we can say that this model performs well.

	precision	recall	f1-score	support
0	0.98	0.95	0.97	2918
1	0.38	0.59	0.47	143
accuracy			0.94	3061
macro avg	0.68	0.77	0.72	3061
weighted avg	0.95	0.94	0.94	3061

We clearly see that there is graduate increase in recall and f1-score when we compare this with above models.

Data Visualization and Result Reports:

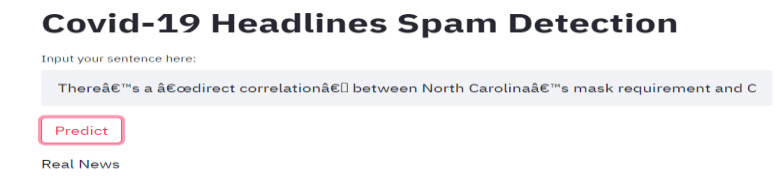
Model	Accuracy	Precision		Recall		F1-score	
		1	0	1	0	1	0
MultinomailNB	96%	92%	96%	29%	100%	44%	98%
SVM	97%	94%	97%	36%	100%	52%	98%
Passive Aggressive	94%	38%	98%	59%	95%	47%	97%

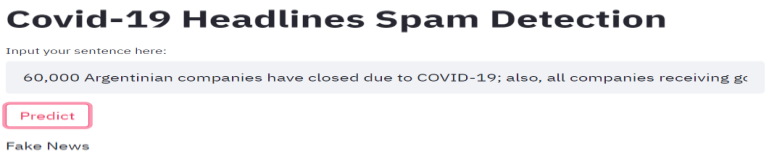
When we see above results clearly, we can see accuracy is very high for SVM and MultinomialNB when we compare with Passive Aggressive model but due to imbalance dataset, we cannot consider Accuracy metric for this problem.

For this problem we need to consider both Precision and Recall metrics so when we compare all 3 models then clearly, we can see Passive Aggressive model is best model for this problem.

Web Application:

I have created a web application using Streamlit package. With the help of this application any user can able to check whether Covid-19 headlines is Fake or Real.





Conclusion:

In this project, presenting a comprehensive COVID misinformation dataset, which has news headlines. Described how the data is collected, cleaned, analyzed. NLP and Machine Learning models like MultinomialNB, SVM, Passive Aggressive which results to predict from the information if the news is fake or real. This could help researchers to find useful for their research and together contribute to flatten the curve of COVID-19.

9. Bibliography:

[1]. Sahni, S. (2020, December 22). Tweet Analysis for COVID-19 Fake News Detection - Level Up Coding. Medium. https://levelup.gitconnected.com/tweet-analysis-for-covid-19-fake-news-detection-bf7cbea5c12d

[2] Mackey, T. K., Li, J., Purushothaman, V., Nali, M., Shah, N., Bardier, C., Cai, M., & Liang, B. (2020). Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales: Infoveillance Study on Twitter and Instagram. JMIR Public Health and Surveillance, 6(3), e20794. https://doi.org/10.2196/20794

[3] Vázquez, F. (2020, July 24). Detecting Fake News with and Without Code - Towards Data Science. Medium. https://towardsdatascience.com/detecting-fake-news-with-and-without-code-dd330ed449d9

[4] Sagar, R. (2020, April 24). Top ML Projects To Fight Fake News Fatigue During COVID-19. Analytics India Magazine. https://analyticsindiamag.com/fake-news-covid-19/

[5] George, J. A. (2020, December 4). Fake News Detection using NLP techniques - Analytics Vidhya. Medium. https://medium.com/analytics-vidhya/fake-news-detection-using-nlp-techniques-c2dc4be05f99

[6] Ahmad, I. (2020, October 17). Fake News Detection Using Machine Learning Ensemble Methods. Journal. https://www.hindawi.com/journals/complexity/2020/8885861