CS6301 Special Topics in Computer Science

Project

Fake News Detection in social media

Summer2023

`

Team Members

Akhila Petnikota(AXP210228)

Pavan Thota(PXT210026)

Vinay Sugasi(VXS210076)

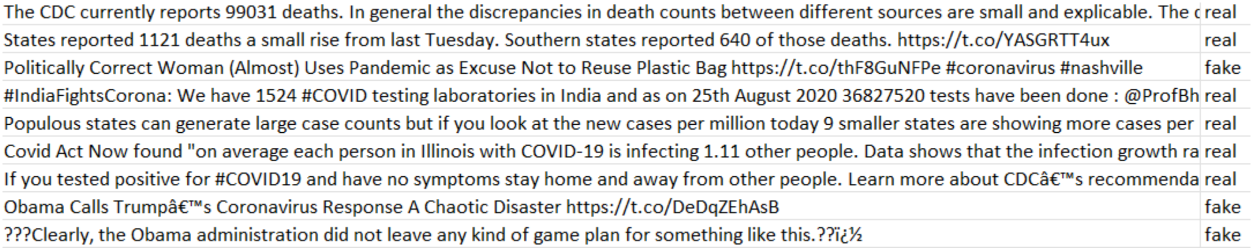
Problem Description:

There are various news spreading in social media every day. We are not sure about the authenticity of the news spreading i.e., whether the news is real or fake. If a fake news is spread, we are not sure how much it might affect the users around the world. So, we developed this problem of detecting fake news and tried to solve the problem using machine learning and Natural language processing. So, the problem is developed into a classic text classification problem where each text can be classified as real one or not.

System Design:

Data Collection:

To solve this problem, we took data from Kaggle. We took a data set containing tweets about Covid 19, The data is collected from various social media sites like Facebook Instagram and Twitter. The data has two columns, one is the tweet column and other is the label columns which classifies that tweet as either real one or fake one.



The CDC currently reports 99031 deaths. In general the discrepancies in death counts between different sources are small and explicable. The c	real
States reported 1121 deaths a small rise from last Tuesday. Southern states reported 640 of those deaths. https://t.co/YASGRTT4ux	real
Politically Correct Woman (Almost) Uses Pandemic as Excuse Not to Reuse Plastic Bag https://t.co/thF8GuNFPe #coronavirus #nashville	fake
#IndiaFightsCorona: We have 1524 #COVID testing laboratories in India and as on 25th August 2020 36827520 tests have been done : @ProfBh	real
Populous states can generate large case counts but if you look at the new cases per million today 9 smaller states are showing more cases per	real
Covid Act Now found "on average each person in Illinois with COVID-19 is infecting 1.11 other people. Data shows that the infection growth ra	real
If you tested positive for #COVID19 and have no symptoms stay home and away from other people. Learn more about CDCâ€™s recommenda	real
Obama Calls Trumpâ€™s Coronavirus Response A Chaotic Disaster https://t.co/DeDqZEhAsB	fake
???Clearly, the Obama administration did not leave any kind of game plan for something like this.??ï¿½	fake

The above screenshot shows how the data looks like.

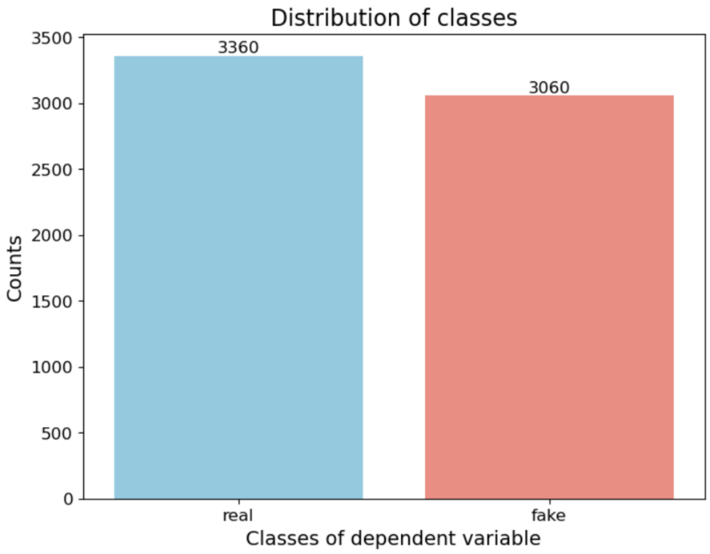
Once the data is collected our first task was to preprocess the data, make some conclusions from analyzing the data and split the data into train and test and train the machine learning model on train data and test the accuracy on the test data.

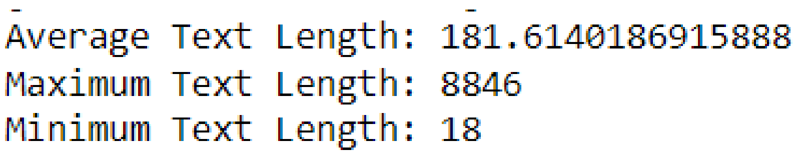
Data Preprocessing:

Before training the model on the data our first step in implementing machine learning algorithm is to preprocess the data as a part of preprocessing, we have removed stop words, removed punctuations removed special characters and converted the entire text into lowercase. Stemming and lemmatization was also performed for improving the accuracy of the model. We also checked if there are any missing values and duplicate rows.

Data Analysis:

The initial data analysis showed that there are 6420 tweets related to covid 19. The distribution of number of fake and real tweets is shown in below figure. From the below figure it is clear that the data is not biased towards any class as both real and fake tweets are almost of equal number .

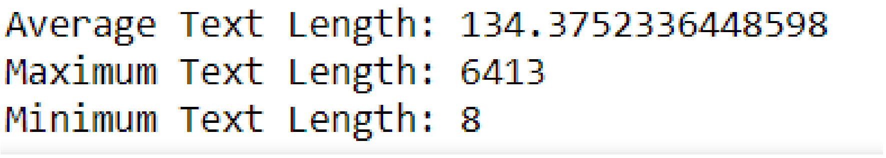




```
Average Text Length: 181.6140186915888
Maximum Text Length: 8846
Minimum Text Length: 18
```

Fig1.Text lengths before preprocessing

The above image shows the text lengths before preprocessing the data.



```
Average Text Length: 134.3752336448598
Maximum Text Length: 6413
Minimum Text Length: 8
```

Fig2.Text lengths after preprocessing

From fig1 and fig2 we can observe that text lengths after preprocessing have reduced significantly. So the Machine learning model would perform better after preprocessing the data.

Feature Extraction:

As we are dealing with text data in order to implement any machine learning algorithm, we need to convert the text data into numerical form so that the machine can understand it. For converting text data into vector of numbers we have used Total frequency and inverse document frequency where TFIDF is given by the product of Total frequency and Inverse document frequency,

TF(t, d) (Term Frequency) is the number of times term 't' appears in document 'd’.

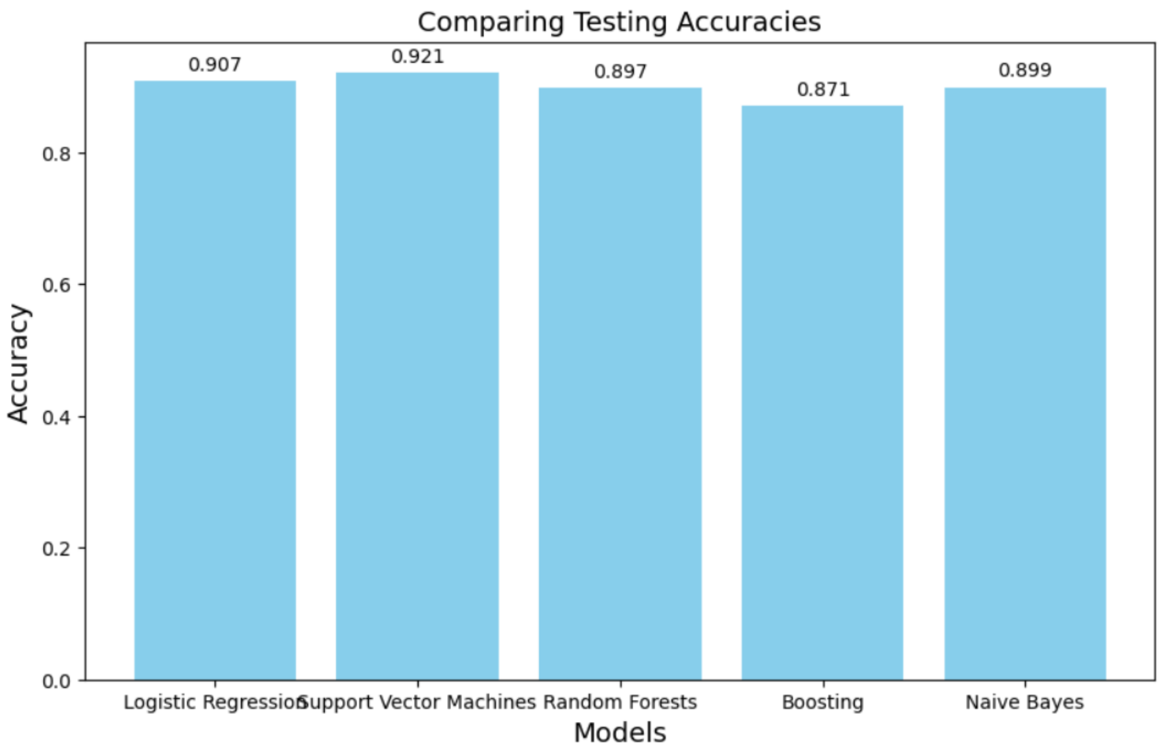
IDF(t) (Inverse Document Frequency) is the logarithmically scaled inverse fraction of the number of documents that contain term 't' out of the total number of documents in the collection.

Splitting into train and Test:

The data was split into training and test using scikit learn and machine learning models were applied on train data and the accuracy of both train and test is calculated.

Implementation:

We used five machine learning models to train our data as we are not sure which model would perform better. It has been observed that support vector machines has performed better as compared to other machine learning algorithms. The below image shows the comparison of different models and their accuracy.



Challenges Encountered:

As we are dealing with text data converting text data into vectors has become difficult as texts are of different length in each tweet. Tweets are also from different users around the world due to difference in dialects the meaning would have been different which would have affected our model.

Future Work and Directions:

1. Implementing neural networks and see how accuracy is obtained.
2. Different feature extraction techniques like word2vect and count vectorizer can be used to convert the text data into vector form.
3. As Data we are dealing with is of small size we are not sure if our model is overfitting so taking data of large size would give us accurate results.