TWITTER SENTIMENT ANALYSIS USING NLP

- **Problem:**

Recently, the world is facing a severe public health emergency due to the ongoing COVID-19 global pandemic. As many people around the world are affected by coronavirus so analysing the public sentiment is important to investigate people's opinions about this pandemic outbreak? The aim of this project is to develop the classification model to classify the coronavirus tweets to positive, negative or neutral by analysing the sentiments expressed in the tweets.

- **Background**:

Over the past few years, researchers have focused lately on sentiment analysis more than before. By using this method, researchers can extract people's opinions that are related to a crisis,events, brands, and products. In order to achieve a reflection of the public to this virus, Twitter can be considered as a popular online social-networking source which allows people to share their feelings about any subject quickly. As the sentiment-classifier system has many applications from business to social sciences, researchers use the natural language processing and machine learning techniques to create them in order to explore the polarity of the text easily [2].

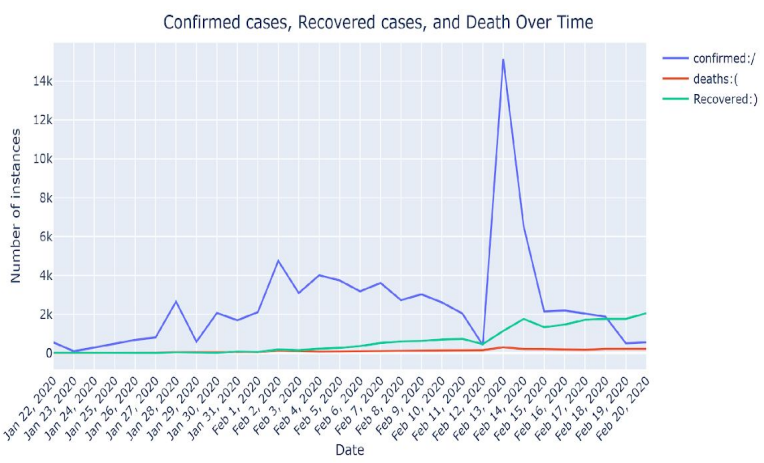
- **Data Acquisition:**

In order to explore some stats about coronavirus, this dataset has been taken in a CSV file format from this link https://github.com/CSSEGISandData/COVID-19. This dataset has daily level information from the number of confirmed cases, deaths, and recovery of Covid-19. To train the sentiment classifier, the pre-labeled tweet datasets has been provided by Kaggle in the following link:[https://www.kaggle.com/shashank1558/preprocessed -twitter-tweets]. Finally, Tweepy is a library of Twitter API for extracting the unlabeled 20000 tweets directly from Twitter then saved into CSV files. This dataset was provided by Dr. Kin Lee from EE department at University of Victoria.

- **EDA (Exploratory Data Analysis):**

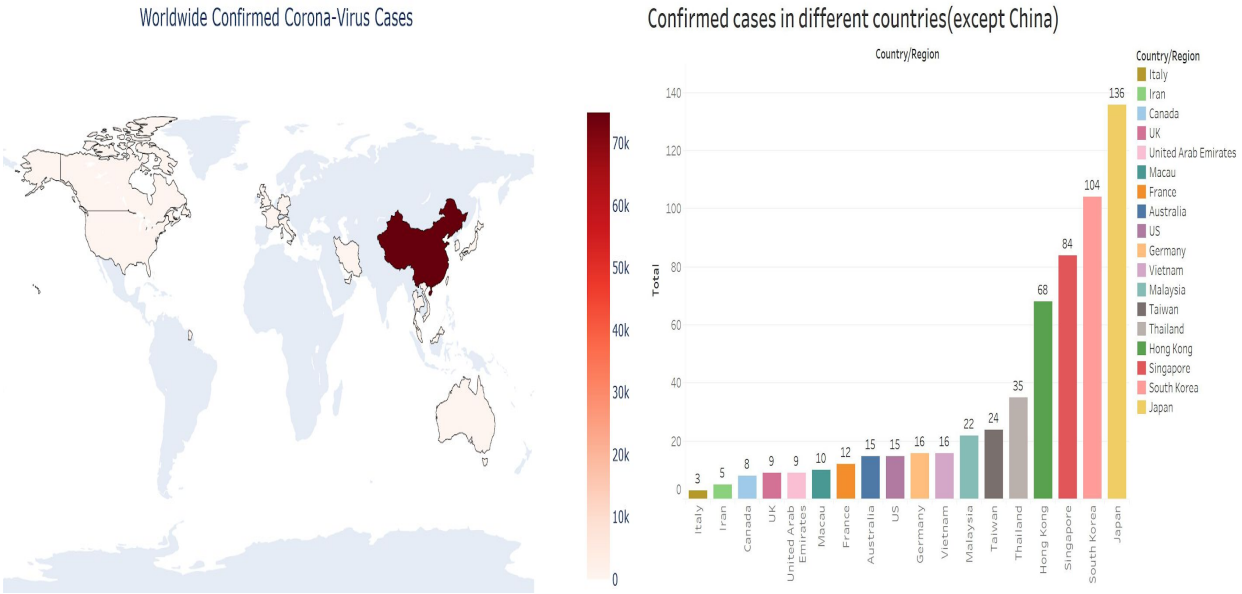
COVID-19 spread dataset can provide sufficient information regarding the growth of the outbreak all over the globe. Here, some EDA has been made to show the spread of the number of confirmed, death, and recovered cases in the world between 22 January 2020 and 20 February 2020. As it's shown, there is a huge increase in the number of confirmed cases on Feb 13 2020. After checking the news for that day, I found that it had been an announcement from the health organizations in china that they decided to train medical professionals regarding classifying a suspected case of Covid-19 as a confirmed case based on results in chest imaging and a doctor's

analysis not just checking the symptoms. I think this might be the same reason for the recent sudden jump in the number of the affected people for some countries like the US, Iran as well.



- **Data Visualization:**

Some data visualizations using Plotly have been made to demonstrate the spread of this virus in different countries outside of China.



The left side map plot displays the spread of the coronavirus worldwide by using a significant number of affected peoples over different countries . In order to exhibit the number of confirmed cases outside China, Tableau was used to make the right side plot.

- **Data Preparing:**

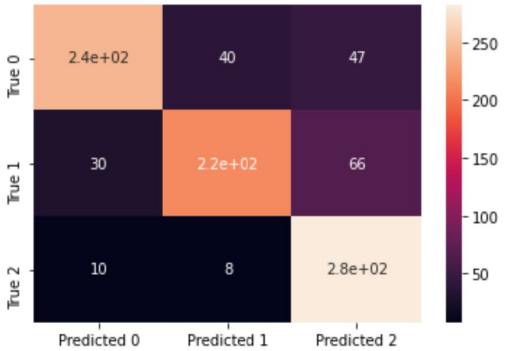
Now it is time to preprocess and clean the tweets because all these modifications will directly affect the classifier's performance. There are some text formatting techniques which will help us in feature extraction including removing urls and user references ("http" and "@"), punctuation marks/digits , and also stop-words. In addition, the implementation of lemmatization words using NLTK can be workable to maximize the performance. Tokenization is the last step to break tweets up into words and meaningful tokens.

- **Machine Learning models:**

In order to train the classifier, we use the pre-labeled dataset to create the best model to predict the unlabeled tweets sentiment. First, we divided tweets in 80:20 split for train and test data. By implementing the sklearn library, we can use TF_IDF vectorizing to find the weighted words that occur more frequently in the document that leads to creation of the bag of words model. I've implemented different models with different score shown in following table:

Classifier	Train Score	Test score
Logistic	0.88	0.786
Random Forest	0.94	0.782
XGBoost	0.88	0.76
Naive Bayes	0.87	0.779

There are several metrics to evaluate our classifiers include: Precision, Recall, Accuracy, F1-measure. In order to increase the test score, the grid search has been implemented to optimize our model by varying the hyperparameters. Because of high accuracy, the logistic classifier has been considered as our best model. A confusion table for our model is given below:

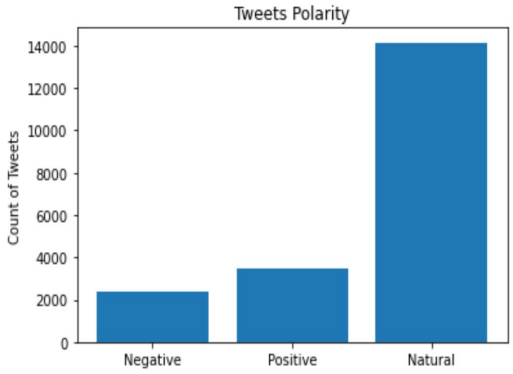


	precision	recall	f1-score
0	0.859155	0.737160	0.793496
1	0.817490	0.691318	0.749129
2	0.713924	0.940000	0.811511
accuracy	0.786624	0.786624	0.786624

The only challenge that I've faced was about balancing my train dataset in terms of having the equal number of tweets for all three classes. So I used another dataset to add some positive negative tweets to my initial tweets dataset. According to the recall values, we can conclude that our model with %94 recall for neutral class can predict them more relevantly.

- **Tweet Label Prediction:**

In this stage, after creating the best classifier's model, now we can use it to predict the sentiment of the coronavirus tweets. We got the 20000 tweets created on Feb 13 2020 to see people 's reaction to the number of the affected people. First, the tweets need to get pre-processed and cleaned and then vectorized. Finally, the tweets can be classified by logistic model shown in below:



According to this classification, it can be concluded that people usually try to share information about this coronavirus rather than talking positive or negatively. Additionally, some word clouds were created for negative tweets(with black background) and positive tweets(white background) shown in appendix. We can see some interesting words like "quarantined" in positive tweets. People talk about politics and governments with more negative sentiments.

- **Conclusion and Recommendation:**

Nowadays, sentiment analysis is a popular topic in machine learning to examine the psychological characteristics of the public about any specific crises. In this project, by using the NLP methods, the logistic classifier has been built in order to classify the coronavirus tweets to positive, negative, and neutral based on their polarity. Initial assumption about the sentiment of these coronavirus tweets was negative which means people usually think and talk about this virus negatively. But according to our classification, the sentiment of the total tweets in percentage is 17% positive ,13% negative, and 70% natural. In this case, the misinformation about this virus can be a serious threat to public health. So I think all these social networking platforms now are

responsible to restrict any coronavirus misinformation. On the other hand, they must try to provide more trustworthy and reliable resources about this pandemic virus for people. For future steps, I will try to find any correlation between user location or countries with the polarity of tweets and also to find any pattern in tweets classification for the most affected countries.

- **REFERENCES**

1. https://www.elsevier.com/connect/coronavirus-information-center 2020/27

2. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., ... & Belyaeva, J. Sentiment analysis in the news. arXiv preprint arXiv:1309.6202(2013).

- **Appendix:**