FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY

BITI 2513

INTRODUCTION TO DATA SCIENCE

CAPSTONE PROJECT

TASK 2

PROJECT TITLE : FAKE AND REAL NEWS CLASSIFIER

GROUP : DS GENG

Name:	No Matric :
Muhammad Naim Syahmi bin Roslan	B031810312
Muhamed Hussain Bin Hithayatullah	B031810392
Muhammad Fikrun Amin	B031810404
Ramanan Gobalakrishnan	B031810334

Lecturer Name : AP DR SHARIFAH SAKINAH SYED AHMAD

Data Management Plan

Project, experiment, and data description

This project purpose to solve fake news problem by creating a classifier that is able to label a given news as either fake or real based on their attributes such as the text and subject of a tweet of Twitter. It involves natural language processing and text analysis procedures .And, this project is based on ideas put forward in a paper "Fake News Detection Using Sentiment Analysis by B. Bhutani" [1].

The data that has been collected comes from tweets of Twitter users using data crawling technique where we retrieve tweets data by accessing Twitter through the API that has been provided by Twitter and export the data in ".csv" file. We collect data that contains content about "Covid" and "Corona" especially in English, totaling 5000 data tweets.

There are some format of raw data obtained from Twitter, such as in the form of strings in the tweet content itself, the author's name, language and so on. There is also a time format in the data when the tweet was posted, and also in numerical form in the data that contains the number of likes in the tweet.

In this project, we only use data that is the content of tweets, where the data format is a data string. As the purpose that we have proposed, namely creating a fake news classifier using sentiment analysis, which in this analysis only analyzes the sentiment of the content of the tweet because only that data allows it to be analyzed. By this, we agree to only use the content of the tweet data and not use data that is not related to it for example, the time when the tweet was posted, the number of likes, the name of the authors, location, etc.

After collecting 5000 tweet content of "Corona", we are in the process of cleaning up the tweet which is "Data Pre-processing". Before we process a data, we must ensure that the data is completely clean and ready to be processed. "Clean" is meant here is to discard anything that is not related and also does not have a significant effect on the results at the end. For example, removing stop words, emoji's, emoticons, and punctuations.

After successfully getting 5000 clean data text and ready to be processed, we start the process of the data itself, in this case we try to calculate the sentiment value of 5000 data. In this process, we do not do calculate manually. But, we do it with a sentiment analysis approach. in this case we use a python library that is TextBlob to get the value of polarity and also the subjectivity value of the text data. In this project we only use data that has positive and negative polarity and discard data that has polarity = 0.

After getting the polarity and subjectivity values from the text data, we do the labeling of the data by using the assumptions that written in the paper that we refer to, namely labeling all data that have negative polarity values as fake news. Otherwise, the label is real news.

At this point, we can analyze the data using several analytical methods such as word cloud, TFIDF score, visualization distribution method in the form of a histogram chart. More details are in the Exploratory Data Analysis section.

The next plan after this is the process of modeling the data using several classification methods such as Random Forest, Naive Bayes and Logistic Regression. We will compare and analyze the results obtained from the 3 algorithms. And choose which algorithm is most suitable to be used to produce a model in this case or project.

And after the model is successfully formed, we will use it to classify and predict the label of remaining 4000 data of the initial 5000 data to see how precise and accurate a model that we have produced and can ultimately be used in further projects related to this project.

The data that we have obtained is not permitted to be accessed or shared publicly due to the terms and policies of Twitter itself. However, the model that will be produced can be accessed and used for other purposes publicly.

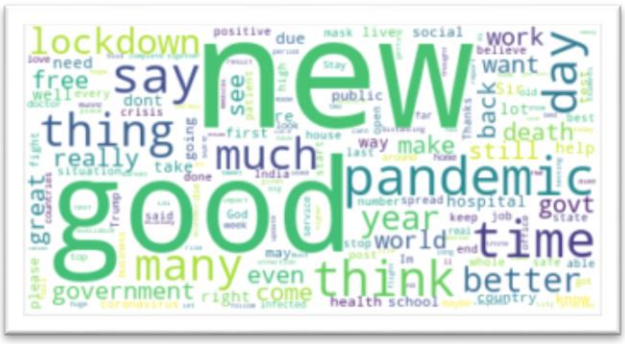
Exploratory Data Analysis (EDA): Findings and Hypotheses

For EDA purposes, we have prepared a few visualizations for the data. The visualizations will enable us to identify any special characteristics or pattern present in the data. These visualizations include word clouds, TfIdf score of words, and distributions of specific attributes within the data.

1. Word Clouds

The word clouds generated from the data is created via the Word Cloud function from the word cloud python package. For the visualization, we have created separate word clouds for real tweets and fake tweets.

1.1.Real Tweets Word Cloud



Word cloud of real tweets

From the real tweets word cloud, the largest words present are positive words such as 'new', 'good' and serious matters such as 'pandemic' and 'lockdown'.

From the word cloud we can see that the words with the biggest font are the words that are mostly used in the tweets. In this case, words such as ‘new’,’good’ which shows people are worried about the new developments about the coronavirus and they also interested about the positive side of things which are happening

1.2.Fake Tweets Word Cloud

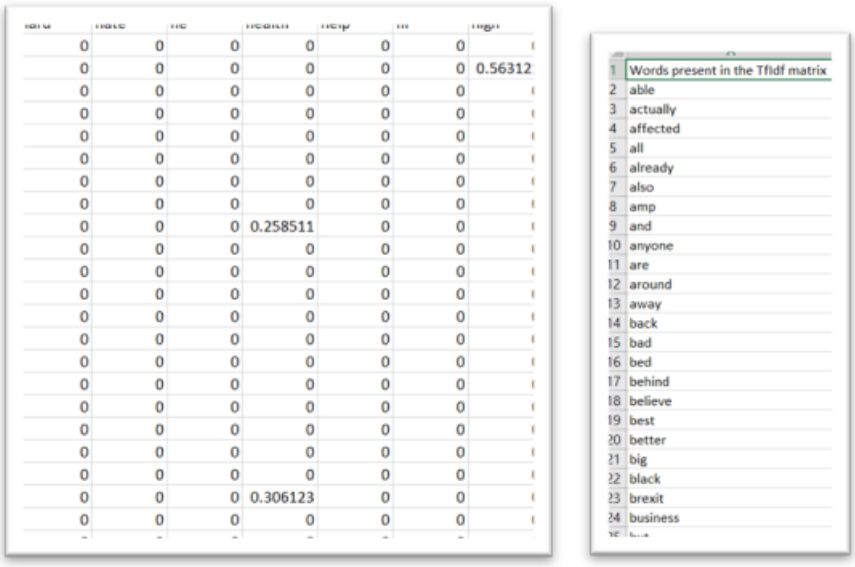
Word cloud of fake tweets

From the fake tweets word cloud, we can immediately see words such as 'worst', 'due', 'time' and 'death'.

In this word cloud, it is obvious that words such as ‘worst’, ‘say’ ,‘due’ and ‘time’ are mostly mentioned in tweets which shows that the public is concerned about the time taken from them due to the pandemic. People are mostly worried about the time taken for the countries all over the world heal from this pandemic.

2. TfIdf Score of Words

The TfIdf is used to find terms that are relevant to the documents in a corpus. Words that are frequently encountered is penalized and has a lower TfIdf score. For visualization purposes, the TfIdf has been adjusted to remove words that are encountered in 90% of the tweets and to remove words that appear in less than 10 tweets. It resulted in a matrix of 1000 rows and 276 columns.



	Words present in the Tfidf matrix
1	
2	able
3	actually
4	affected
5	all
6	already
7	also
8	amp
9	and
10	anyone
11	are
12	around
13	away
14	back
15	bad
16	bed
17	behind
18	believe
19	best
20	better
21	big
22	black
23	brexit
24	business

TfIdf matrix and a list of words present in the matrix imported to a CSV file.

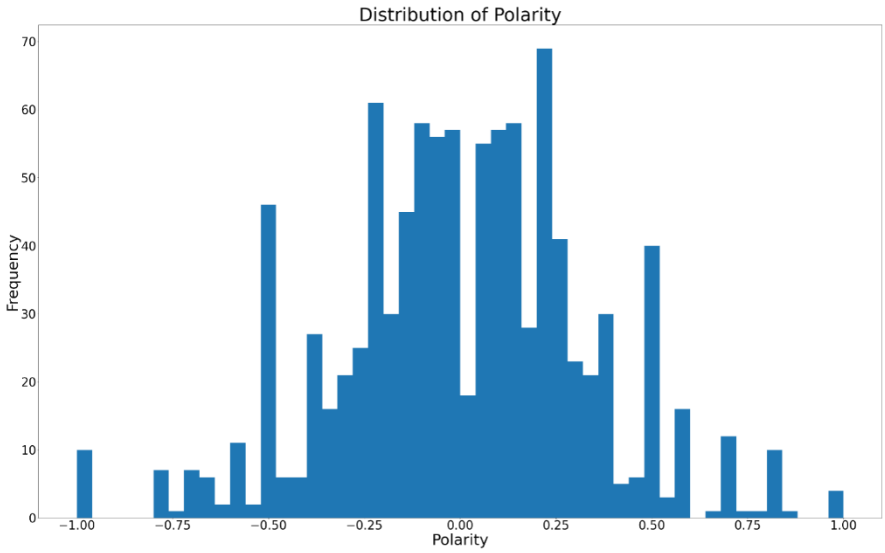
The TfIdf matrix is stored in a CSV file so it can be explored in more detail. From the left figure we can see that the word 'health' has a TfIdf greater than 0, this indicates that the word is relevant in that particular tweet. The right figure lists all 276 words that are present in the TfIdf matrix.

We can see that from the score of words which is where if the score is closer to 1 that that word is more relevant in that tweet, that words such as ‘corona’ which is present in a lot tweets but have very little relevancy to the tweet. Words such as ‘pandemic’ are also present in most tweets mostly because COVID-19 has become one the worst pandemics in human history. In general form the score of words it is visible that the general public is worried about the coronavirus disease.

3. Distribution of Attributes and Properties

For the final few visualization tasks, a few histograms were prepared using the python matplotlib package.

3.1 Distribution of Polarity

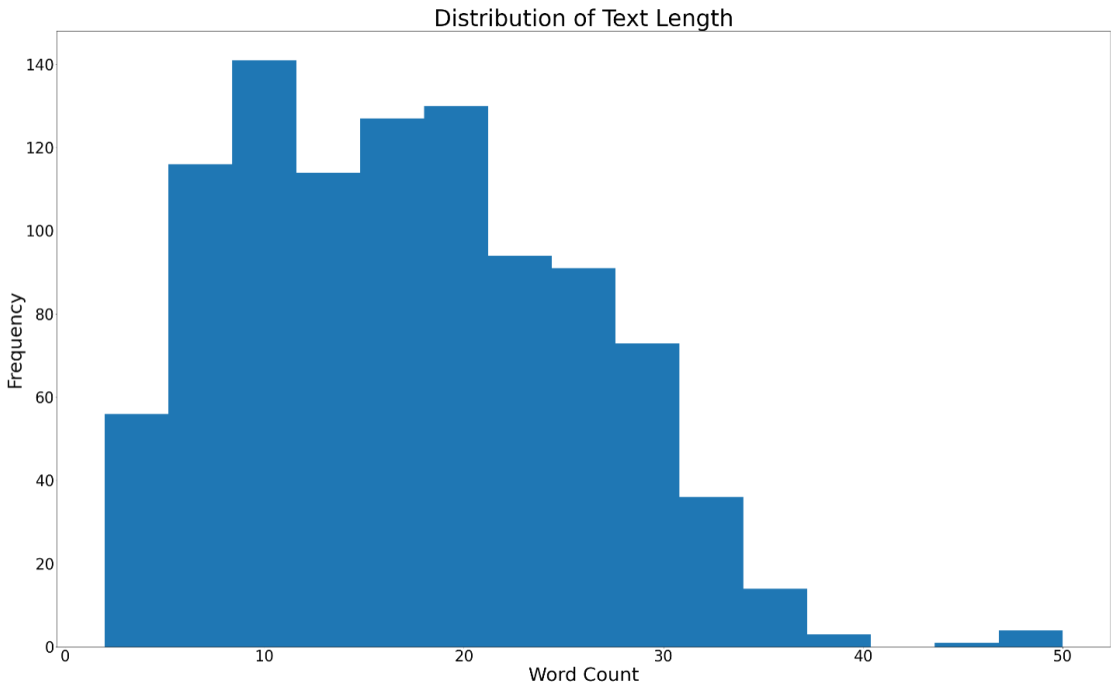


Distribution of Polarity

The distribution of polarity is shown in the above histogram. Surprisingly, the highest bin is in the area of positive sentiment (polarity>0).

In terms of polarity, most tweets are around the middle which shows that most tweets are neutral which means that most people are not very pessimistic, and they are also not very optimistic. There are a few spikes in the polarity but mostly they are neutral.

3.2 Distribution of Text Length

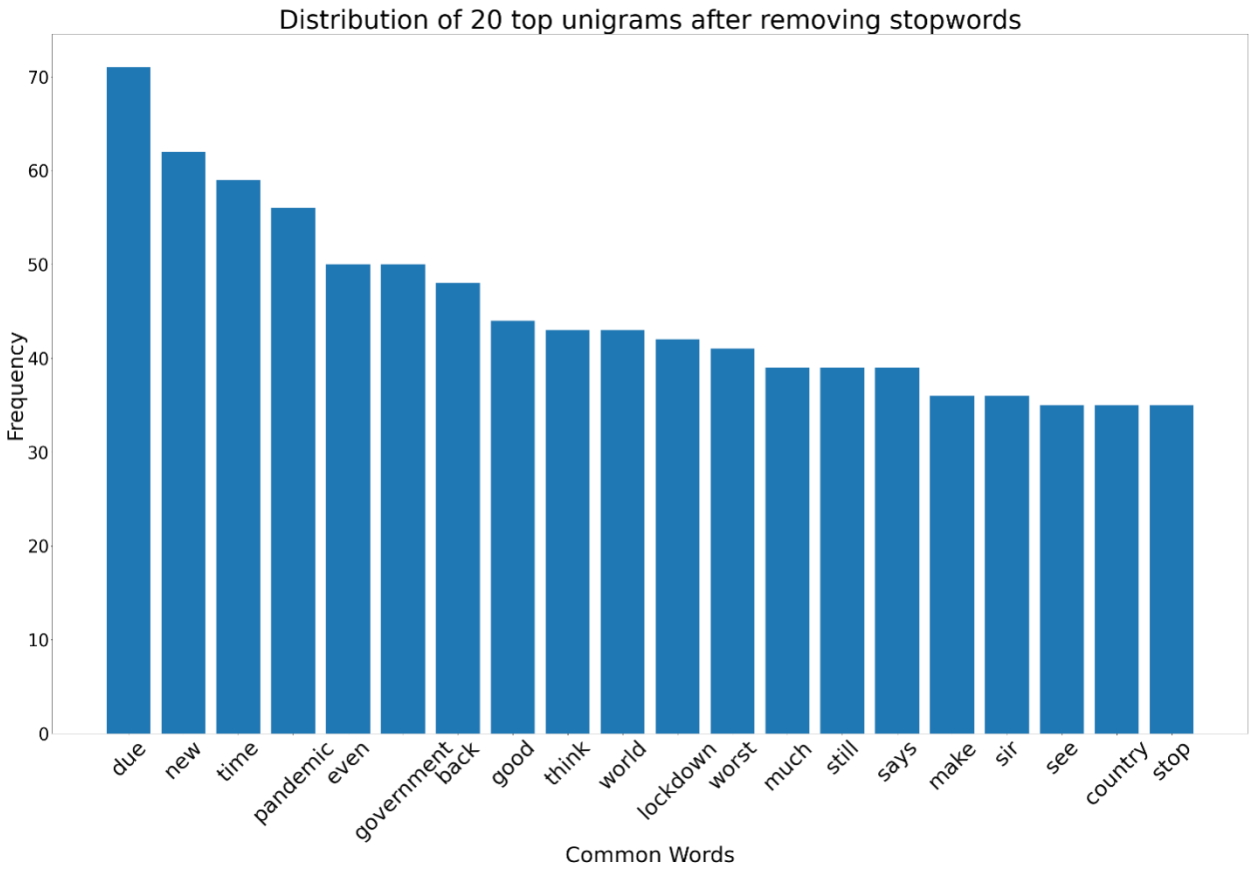


Distribution of Text Length

The above figure shows the distribution of text length throughout the tweets. As expected, most of the tweets are not too lengthy and falls under 10-20 characters.

This visualization shows the length of each tweet against the frequency of the length. It is obvious that most of tweets are from 10-20 words which is higher than the average tweet length which shows the concern of the public as they are worried of the current situation with the current pandemic. Thus, twitter being one of the platforms for people to post their opinions, the tweets subsequently also get longer.

3.3. Distribution of 20 top unigrams after removing stopwords.

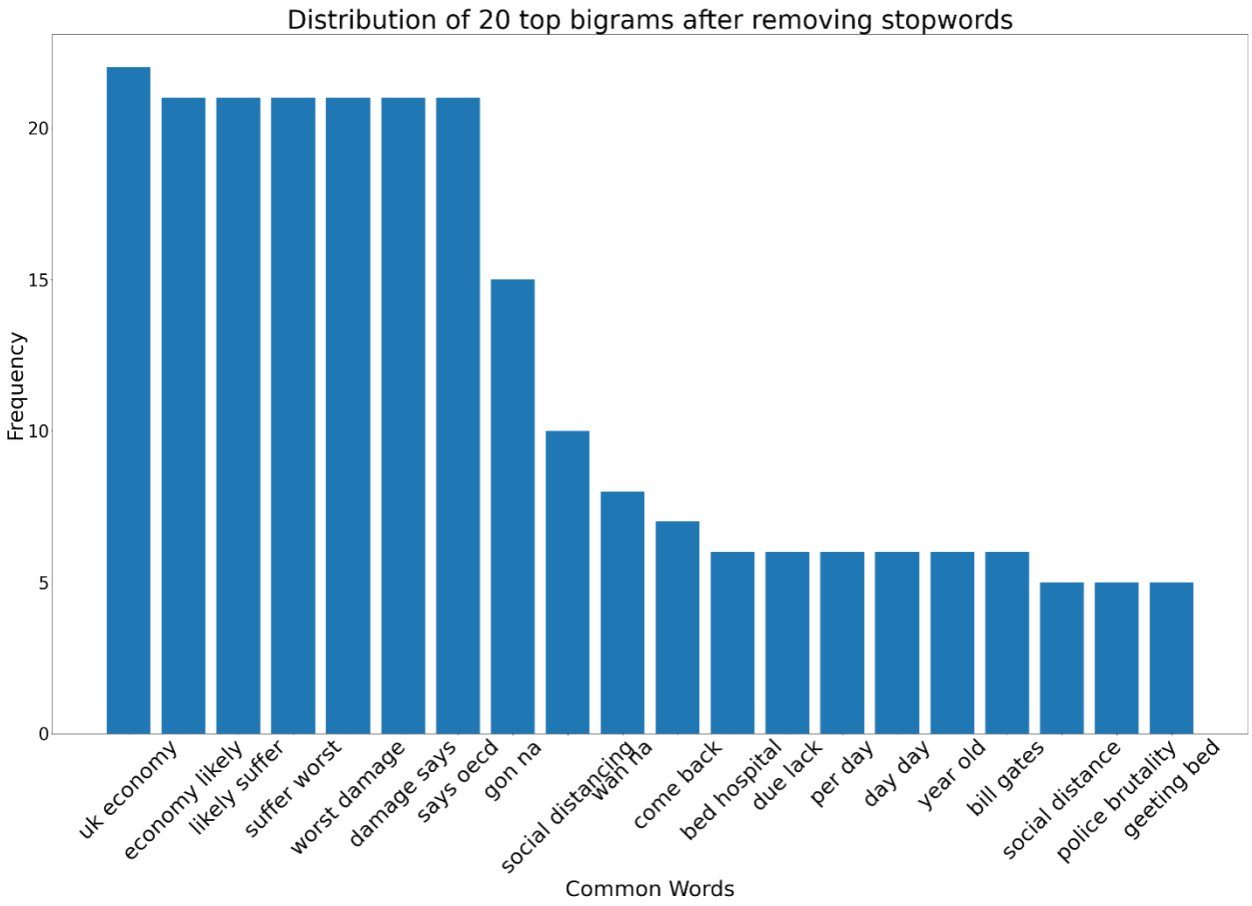


Bar chart of top 20 unigrams after removing stopwords.

The figure above shows the distribution of 20 top unigrams after stopwords are removed. The unigram with the highest frequency is 'due'.

From this visualization we are analysing individual words from each tweets, we can see that the words ‘due’, ‘new’ and ‘time’ are all the words that have the most frequency in tweets which shows that most people are worried about period and the time taken from them from the current pandemic. Followed by words such as ‘pandemic’, ‘even’ and ‘government’ which shows people are concerned about the actions taken by the government during this pandemic.

3.4. Distribution of 20 top bigrams after removing stopwords

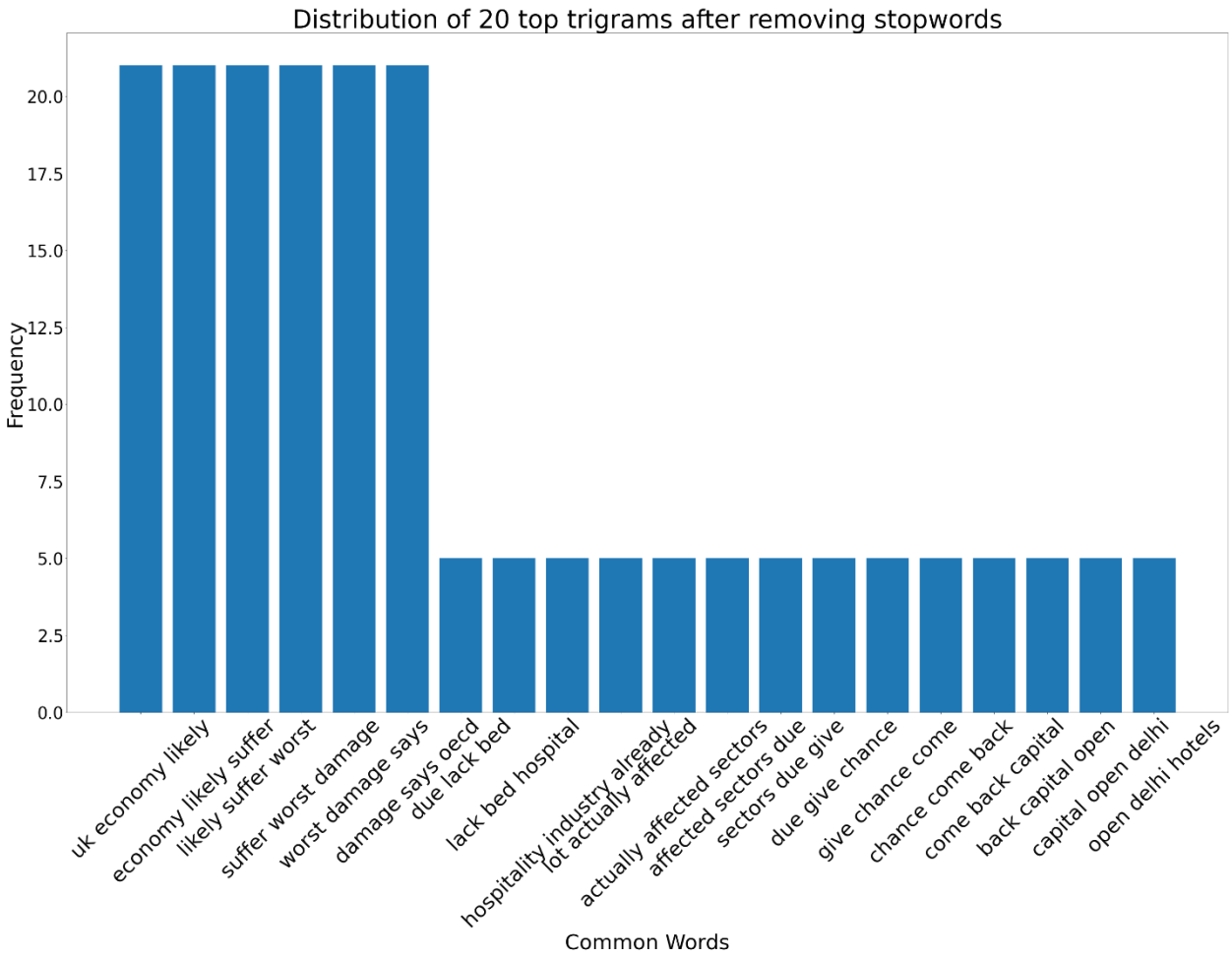


Bar chart of 20 top bigrams after removing stopwords

Next, we have the distribution of top 20 bigrams after removing stopwords. Bigrams are pairs of unigrams combined. Sometimes, a bigram can be a new term with a different meaning than its unigrams. Surprisingly, the top bigram encountered in the tweets after removing stopwords is the 'uk economy'.

From this visualization which shows the distribution of common words in tweets about the coronavirus we can see that ‘uk economy’ is the most used word which shows the concern of people in the fall of the economy amidst the coronavirus pandemic. Followed by the words ‘economy likely’ and ‘likely suffer’ which shows the concern of the public about the wellbeing of the economy.

3.5 Distribution of 20 top trigrams after removing stopwords



Bar chart of 20 top trigrams after removing stopwords

Similar to bigrams, trigrams are three words combined together to form terms. The figure above shows the top 20 trigrams in the tweet after stopwords are removed. The top trigrams are mostly about economy and negative sentiments such as 'suffer' and 'worst'.

From this current visualization we can are able to deduce the same hypotheses as the previous visualization as the same common words are seen to have a high frequency in the tweets of the public. The phrases ‘uk economy’, ‘economy likely suffer’ and ‘likely suffer worst’ are all testament to the current state of the economy and the public as many people are suffering with not being able to earn money and also the closure of many businesses.

4. Variables to Predict

From the data set that is currently used for the project, the main variable to predict is the ‘authenticity’ variable. For each text data, we aim to determine whether the contents of the text is real (authenticity = 1) or fake (authenticity = 0).

References

[1]. B. Bhutani, N. Rastogi, P. Sehgal and A. Purwar, "Fake News Detection Using Sentiment Analysis," *2019 Twelfth International Conference on Contemporary Computing (IC3)*, Noida, India, 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844880.

[2]. Extracting Twitter Data, Pre-Processing and Sentiment Analysis using Python 3.0 https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf