STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Identification of the sentiments and emotions of stressful tweets on Twitter during the COVID-19 pandemic

Web Mining (BIA-660) FALL 2021
Final Report
Instructor: Jingyi Sun

Abdulrazzak Moulvi, Deepanshu Negi, Shreeya Kokate, Rishi Singh
Group 11

Introduction:

The COVID-19 pandemic has brought uncertainty and great panic into our lives making people adjust their daily routines as they attempt to return to their usual lives. People are still practicing better public hygiene and self-care along with following Government guidelines. The novel Coronavirus outbreak has taken many lives over the past three years and has been classified as a pandemic that has been a 'significant increase in the time average time spent by users on social media during this pandemic (Department, 2021).

The pandemic has also seen a rise in the mental health problems people have been facing as '4 in 10 adults in the US may face problems related to depression and anxiety (Panchal, Kamal, Cox, & Garfield, 2021). People tended to spend the majority of their time on social media while they were under lockdown or secluded at home. Twitter is one of the sites that has seen a significant increase as people spent about 32 minutes per day on Twitter on average, which is greater than in the pre-covid years, and this increased time is projected to continue in the post-covid years (IANS, 2021).

In addition, the pandemic has resulted in an increase in mental health and issues related to stress. People on Twitter share a lot of information and openly express themselves by sending tweets about how they feel, what they have been up to, their mental status, health difficulties, stress, and so on, which is why we chose Twitter as a venue for our project.
In this paper, we aim to find out the emotions and sentiments behind the tweets related to stress using unsupervised sentimental analysis and emotion analysis. The paper also aims to find classify the source of tweets using by creating a binary classification machine learning model to classify if the source is a user using their technology devices to tweet or is the source a Twitter bot/third party user.

Literature Review:

Several studies have been conducted regarding sentimental analysis and Twitter tweets on various domains, some of these domains include health care (Gohil, Vuik, & Darzi, 2018), politics (Wang et al. 2012), finance (Smailović, Grčar, Lavrač, & Žnidaršič, 2013), movie reviews (Jain, 2013) and others. Sentimental analysis has been conducted on these three domains specifically showing the strength of using sentimental analysis as a technique to predict the sentiment of textual data. Gohil et al use sentimental analysis on twitter health care research to explore and better understand the ability of future research conducted in this field. Smailovic et al 'adapted a Support Vector Machine classification mechanism' on Twitter data to categorize tweets into three categories that improved the predictive power in the field of the Stock Market.

Similarly, in the field of politics as well sentimental analysis has been used for predicting the outcome of particular elections. Tumasjan et al. focused in their research on the German federal election in 2009. Their investigation identified that Twitter became a platform for analyzing sentiment and also predicting the outcome of the elections. They examined around one lakh political tweets identifying either a politician or a political party. They concluded that the number of tweets is directly proportional to the chances of winning the election. Wang et al. presented a real-time sentiment application system for the U.S.

presidential election of 2012 based on political tweets extracted from Twitter. Jain explores tweets to predict different aspects of a movie's popularity using sentimental analysis; due to the wide popularity of sentimental analysis, this method would be ideal to measure the sentiment behind the tweets related to stress.

Another technique that has been identified is emotional analysis in which multiple emotion categories. Prior studies have been conducted using emotional analysis on Twitter data for example the work carried out by Mathur et al who look at a wide distribution of Twitter data and categorize these tweets into emotions such as anger, anticipation, disgust, fear, and joy. This analysis helps to understand the mental health of people on Twitter (Mathur, Kubde, & Vaidya, 2020). Another study that supports the use of emotional analysis is seen by Cabezas et al who detect the emotional evolution of tweets during the COVID-19 pandemic. The paper looked at Twitter data collected from Spanish-speaking countries and analyses the evolution of emotion in textual data during the COVID-19 pandemic.

Prior work has also been conducted in the classification of social media bots this can be seen by the work of Wang et al who use the k-Nearest Neighbor to detect social media bots. This research was able to output an area under the curve score of 98% and effectively distinguish between users and bots (Wang, et al., 2021). A review by Rodriguez-Ruis also supports the research already conducted in the classification of bots and users from Twitter data using machine learning techniques (Rodríguez-Ruiza, Mata-Sánchez, Monroy, & Loyola-González, 2020).

Therefore, from the literature review conducted there is prior work that has been conducted on Twitter data using sentimental and emotional analysis on general tweets over the COVID-19 pandemic period however this paper proposes to extend this research and carry out sentimental and emotional analysis of Twitter data specifically related to stress and sampled during the COVID-19 period to identify future Tweets related to stress. Along with implementing a binary classification on the source of Tweets collected from Twitter to help researchers in the future who might not have a labeled dataset with the source of a tweet.

Research Question:

Research 1 – To identify key trends between the use of tweets related to stress during the COVID-19 pandemic. The benefits of understanding these key trends allow future researchers to gain an insight into the Twitter data that has been scraped in this research and identify features that can be used in future machine learning algorithms. This can be done by conducting data analytics techniques on the corpus of Twitter by analyzing the frequency-specific words, length of tweets, sources from where these tweets have been tweeted from.

Research objective 2 – Two pivotal research questions are asked within this objective: can sentimental analysis be used to analyze the sentiments behind tweets related to stress? Can emotional analysis be implemented to investigate different emotions behind tweets related to stress? The paper proposes to use sentimental and emotional analysis to understand the sentiments and emotions behind tweets related to stress. This will help to differentiate between people who might be going through a stressful time during the COVID-19 pandemic

and others who are just talking about the topic of stress. This research will collect data from the beginning of the pandemic in late 2019 till October 2021. From this research, it will be pivotal to understand the sentiments and emotions behind tweets related to stress so support can be provided to owners of tweets who show a negative sentiment and have a sad emotion. This will help social media companies like Twitter provide support to their users in future pandemics.

Research objective 3 – From the labeled data can the source of tweets related to stress be classified into two classes which represent users tweeting from the technological devices or Twitter bots/other third party sources.

Methodology:

The focus was to scrap the data from Twitter mostly for the period of peak COVID. The data includes tweets from 17th November 2019 to October 2021. Modules and packages like snscrape, pandas, intertools CSV, datetime have been used to scrape the data using Twitter API, conducting emotion analysis, and using the intended time. Focusing on stress-related tweets, tweets with the 'stress' were obtained. Sentimental analysis and emotional analysis have been used to understand the sentiments and emotions behind the tweets.
The strategy is to treat part of this project as a Machine Learning binary Classification problem by feeding training data in the form of words to predict the source of a tweet. Sentimental Analysis and Emotional Analysis will help us determine the extent of negativity from a tweet like how severe the matter must be with respect to that tweet. The project will also be using topic modeling and LDA to discover abstract topics that occur in the collection of tweets. The combination of these four components will provide a deep understanding of the sentiments and emotions of the tweets collected along with topics being discussed in these tweets.

Data Crawling

The successful scrapping of 30,000 tweets in US English language was completed within the peak COVID time span with all the information formulated in a data frame which included the URL, date, the content of the tweet, user, reply count, retweets, number of likes, mentioned user, etc. Later, data cleaning was done by removing the unwanted columns from the data frame output to obtain a formulated and desired data frame with the information needed. A total of 15 columns were eliminated in this process living us with the 30,000 tweets and 14 desired columns.

For this project the snscrape library was used, it is important to note that the developer version of the snscrape library was as this allowed more columns to be fetched. The figure below shows the 11 columns that have been kept from the initial 27 columns that were generated. Some of the columns that were removed included 'place', 'cashtag', 'hashtag',

'coordinates', and others that were not very relevant to the research question that is being explored.

	url	date	content	renderedContent	id	user	replyCount	retweetCount	likeCount	conversationId	lang	sourceLabel	coordinates	place
0	https://twitter.com/kinanabb/status/1266519684...	2020-05-29 23:59:58+00:00	cover letters stress me out :)	cover letters stress me out :)	1266519684695969792	{'username': 'kinanabb', 'id': 747911479282794...	1	1	5	1266519684695969792	en	Twitter for iPhone	NaN	NaN
1	https://twitter.com/bsprntr2b/status/126651967...	2020-05-29 23:59:56+00:00	@StuBishop_LPD Thanks and prayers to you and y...	@StuBishop_LPD Thanks and prayers to you and y...	1266519678769344512	{'username': 'bsprntr2b', 'id': 31551973, 'dis...	0	0	1	1266514781311205376	en	Twitter for iPhone	NaN	NaN
2	https://twitter.com/Lisaandbarry777/status/126...	2020-05-29 23:59:52+00:00	@JoyceWhiteVance @ShellyL27525853 Let's see\n-...	@JoyceWhiteVance @ShellyL27525853 Let's see\n-...	1266519659039395840	{'username': 'Lisaandbarry777', 'id': 91679958...	0	1	2	1266355238845534213	en	Twitter for Android	NaN	NaN
3	https://twitter.com/uitlingzi/status/126651965...	2020-05-29 23:59:51+00:00	@cezkiethh Im trying like im really stressed 😭	@cezkiethh Im trying like im really stressed 😭	1266519656841568258	{'username': 'uitlingzi', 'id': 12592257299480...	1	0	1	1266517273092796417	en	Twitter for iPhone	NaN	NaN
4	https://twitter.com/fichuntie/status/126651965...	2020-05-29 23:59:51+00:00	@WynterStorm24 take care of yourself! this str...	@WynterStorm24 take care of yourself! this str...	1266519654773776384	{'username': 'fichuntie', 'id': 10802272365093...	0	0	0	1266454440174931969	en	Twitter Web App	NaN	NaN

Figure 1: Data crawling initial data

Preliminary Data Description

Further performing data visualization, the word and the character count of the tweets were analyzed. Using the matplotlib library, the histogram was obtained in the output representing the word length and the character length of the tweet. Lastly, a pie plot of the source label and the waveform representing the word length of the tweet is obtained. The sentimental and emotional analysis will be conducted on the tweets that have been gathered along with further explanatory data analysis to visualize the types of words being used in a tweet along with the most used words in a tweet.

Also, it is important to highlight other variables that are kept within the data set, the date is an important piece of information that has not been removed from the dataset as it can be used later analyzing the trend of when specific tweets have been tweeted. The content, id, user, word, and character length are important variables because they can help direct tweets back to the main source and can help in analyzing the frequency of a tweet by a user. The replyCount, retweetCount, and likeCount variables are helpful to analyze the popularity of certain tweets.

Figure 2 represents histograms created to measure a range of word length found in the tweets collected and the character lengths. This shows that a lot of the tweets collected contained a word length of 0 to 20 words and 0 to 150 characters. This can be helpful in the further development of the project if a correlation is to be made between the sentiment or emotion and the length of a tweet a tweet contains the word 'stress'. The average word and character length were also calculated which can be seen in the EDA.

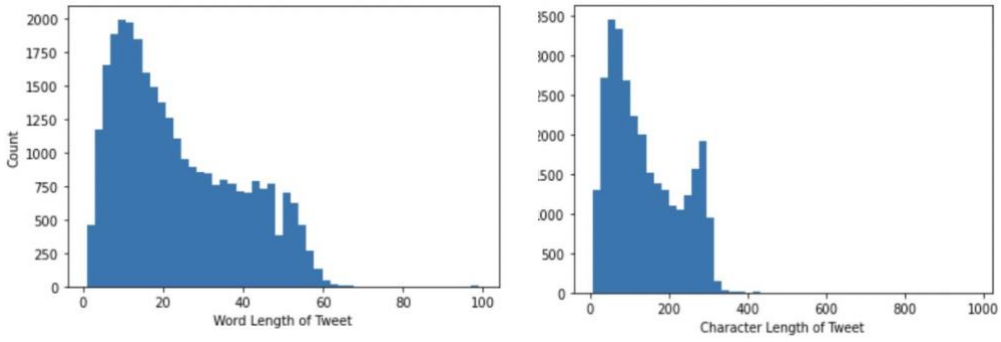


Figure 2: Data description using word and character count

The pie chart in figure 3 shows a representation of the top five sources used by tweeter users to tweet. The sources column was grouped and then sorted into ascending order where the tail of the data frame was kept, and a pie chart was created. Out of the 30000 tweets that were gathered around 1885 tweets came from different sources that are not represented in the pie chart. The main sources that were used to tweet are mobile devices which are split up into Android devices and iPhones. This information is useful as tweets can be differentiated by sources and it can be identified if a tweet was created by a user or by a bot or other entities.

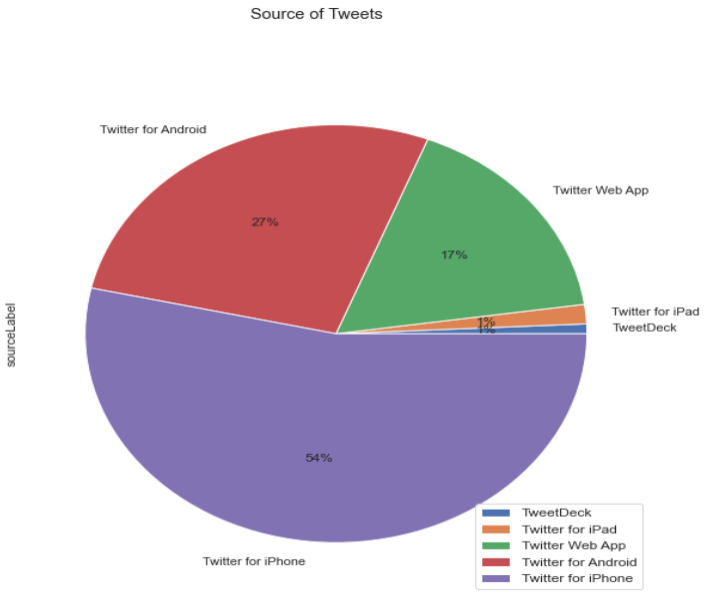


Figure 3: Pie chart showing the top 5 tweet sources

Figure 4 below represents the length of words in the tweets from other sources that have not been represented in the pie chart above. This figure is also important as it shows the distribution of the length of tweets from differing sources and their importance. This is to highlight the importance of the source column regarding where the tweets are coming from and the range of the lengths of the tweets.

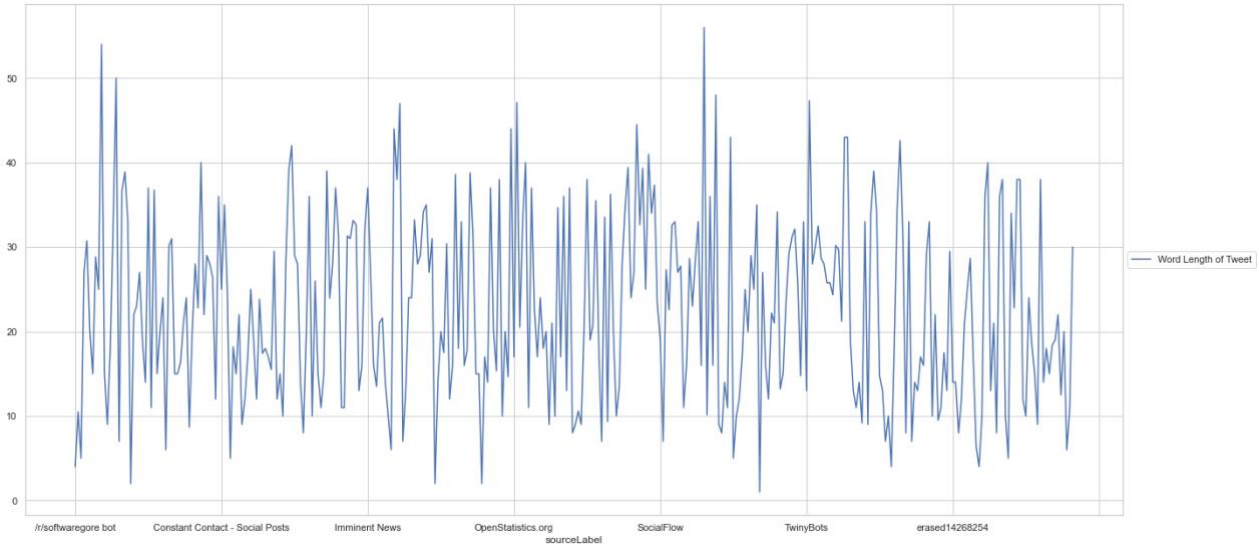


Figure 4: Frequency diagram showing different number of sources

Analytical Strategy

Unsupervised Sentimental Analysis

Sentimental Analysis is the collection of people's perspectives on any real-life occurrence. It's a branch of NLP that looks at how people's opinions are expressed in unstructured text. We have done some more pre-processing on these extracted tweets and converted them in a structured manner. There are two ways to sentiment analysis: rule-based and machine learning-based. We've concentrated on the Rule-based Sentiment Analysis method. This is a practical method for analyzing text that does not require any training or the use of machine learning models. This method yields a set of principles based on which the text is classified as positive, negative, or neutral. Lexicons are another name for these rules. As a result, the Rule-based approach is also known as the Lexicon-based approach. Cleaning the text, tokenization, enrichment – POS tagging, stop words removal, and obtaining the stem words are all conducted in data pre-processing.

First, we deleted the text's special letters and numerals, "clean" is a function that accepts a string as input and returns it without any punctuation or digits. It was applied to the 'content' column, and a new column called 'Cleaned content' was created with the cleaned text. Then, the columns which were not useful were excluded. The content column and the source label column were kept unchanged as they were crucial for the further process. We used the nltk tokenize function word_tokenize() to tokenize the text at the word level to divide it into smaller pieces. With this, we introduced a column called Tokenized Content, which tokenized all the tweets. We removed the words in English which carried little useful information such as ‘I’, ‘me’, ‘myself’, etc. The nltk library has a list of stop words of every language. Furthermore, we used POS tagging to turn each token into a tuple of the form (word, tag) and created a column named 'POS tagged'. POS tagging is required for Lemmatization and to maintain the context of the word. The nltk pos_tag function can be used to accomplish this.

Further, there are two prominent ways for obtaining all the stem words: lemmatization and stemming. The issue with stemming is that it produces nonsensical root words because it just removes certain characters at the end. For example, if we use Stemming on Studies, it becomes studi, and if we use it on Computer, it becomes comput. Lemmatization, on the other hand, provides meaningful root words, which is why we utilized it. Lemmatize is a function that takes pos tag tuples and returns the Lemma for each word in the tag-based on its pos. We used it on the 'POS tagged' column and generated the 'Lemma' column to hold the results.

For sentimental analysis, we used the TextBlob Python library. We could have gone with VADAR or SentiWordNet, but our study shows that TextBlob outperforms VADAR and SentiWordNet when it comes to textual data. TextBlob offers a unified API for common natural language processing (NLP) operations like part-of-speech tagging, noun phrase extraction, sentiment analysis, and more. Polarity and Subjectivity were the two measures we utilized to analyze sentiments. Polarity is measured on a scale of -1 to 1. If it's close to 1, the tweet is positive; if it's close to -1, the tweet is negative. Subjectivity is rated on a scale of 0 to 1. If it's close to 1, the tweet is subjective, implying that it has an opinion. If it's close

to 0, it's objective, or factual. We performed polarity on column 'Lemma' and produced a new column named 'Polarity.' We also performed Subjectivity on column 'Lemma' and established a new column titled 'Subjectivity.' to measure the opinionated tweets.

Finally, using the en_core_sci_md package, we extracted biomedical words from all the tweet's text. This is a spaCy pipeline with 50k word vectors for biological data. However, we did not conduct sentimental analysis on these words to concentrate on public opinion.

Emotional Analysis

In some cases, the sentiment analysis might not enough understand what the user feels. A lot of industry experts regard emotional analysis as a sort of higher, evolved form of sentiment analysis. Sentiment analysis is limited by only dividing data points by whether they reflect a negative or positive feeling, but that is it. This is far from being the whole picture. Emotional analytics, on the other hand, is a more involved, deeper analysis of consumer emotions that tries to drill down into the psychology of different user behaviors. Emotion analysis is the process of identifying and analyzing the underlying emotions expressed in textual data.

It can be easily done based on the types of feelings expressed in the text such as fear, anger, happiness, sadness, love, inspiring, or neutral. Emotion analytics can extract text data from multiple sources to analyze subjective information and understand the emotions behind it. Various organizations can benefit from Emotion analysis as it improves user experience and monitors reputation.

Text2Emotion is the python package that will assist you to pull out the emotions from the content. It processes any textual data, recognizes the emotion embedded in it, and provides the output in the form of a dictionary. Well suited with 5 basic emotion categories such as Happy, Angry, Sad, Surprise, and Fear.

Following are the features of the library we have used:

1. TEXT PRE-PROCESSING

At first, we have the major goal to perform data cleaning and make the content suitable for emotion analysis.

- Remove the unwanted textual part from the message.
- Perform the natural language processing techniques.
- Bring out the well-pre-processed text from the text pre-processing.

2. EMOTION INVESTIGATION

Detect emotion from every word that we got from pre-processed text and take a count of it for further analytical process.

- Find the appropriate words that express emotions or feelings.
- Check the emotion category of each word.
- Store the count of emotions relevant to the words found.

3. EMOTION ANALYSIS

After emotion investigation, there is the time of getting the significant output for the textual message we input earlier.

- The output will be in the form of a dictionary.
- There will be keys as emotion categories and values as emotion scores.
- Higher the score of a particular emotion category, we can conclude that the message belongs to that category.

After installing this library, import it into the work environment which will be required to build your model. Now, call the get_emotion() function using the defined column parameter. Here, we got the output in terms of the dictionary where we have emotion categories along with the respective score. Further, we find the dominant emotion category for each tweet entry.

Topic Modelling LDA

Topic modeling is a branch of unsupervised natural language processing which is used to represent a text document with the help of several topics, that can best explain the underlying information in a particular document. This can be thought of in terms of clustering, but with a difference. Now, instead of numerical features, we have a collection of words that we want to group together in such a way that each group represents a topic in a document. Okay, so now the question arises why do we need topic modeling? If we look around, we can see a huge amount of textual data lying around us in an unstructured format in the form of news articles, research papers, social media posts, etc. and we need a way to understand, organize and label this data to make informed decisions. Topic modeling is used in various applications like finding questions on stack overflow that are similar to each other, news flow aggregation and analysis, recommender systems, etc. All of these focus on finding the hidden thematic structure in the text, as it is believed that every text that we write be it a tweet, post, or a research paper is composed of themes like sports, physics, aerospace, etc.

Topic modeling is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of a topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions. A tool and technique for Topic Modeling, Latent Dirichlet Allocation (LDA) classifies or categorizes the text into a document and the words per topic, these are modeled based on the Dirichlet distributions and processes.

The LDA makes two key assumptions:

1. Documents are a mixture of topics, and
2. Topics are a mixture of tokens (or words)

And, these topics using the probability distribution generate the words. In statistical language, the documents are known as the probability density (or distribution) of topics and the topics are the probability density (or distribution) of words. Latent Dirichlet Allocation (LDA) does two tasks: it finds the topics from the corpus, and at the same time, assigns these topics to the document present within the same corpus.

Binary Classification using SVM & Logistic Regression

During the explanatory data analysis stage, the sourceLabel column was a very interesting data column that held several different sources of the origin of tweets. Here the strategy is to group together the top five most frequent sources which can be seen in Figure 3 as these sources represent users who use technological devices such as iPad, iPhone, Android devices, web app, and TweetDeck and then group together all the other sources which represent third party sources or bots. By doing this a binary classification problem can be created and we aim to create a binary classification model using a Support Vector Machine (SVM) algorithm and Logistic Regression. These two models were chosen due to their popularity and ease of implementation along with being used in prior research such as the work conducted by Foysal et al in the classification of AI social bots on Twitter using SVM (Foysal, Islam, & Rahaman, 2019).

The use of Logistic Regression can also be supported by the work of Efthimion et al who use both algorithms in their work to identify social Twitter bots. Hence due to this the strategy for this project was to implement both SVM and Logistic Regression and compare the results of both algorithms. The input data in this case will be the tweets and using the TfidfVectorizer function the tweets will be taken processed before the fitting of the model. A crucial part of this strategy is also to balance the dataset as after grouping the datasets together there was a significant imbalance hence the Synthetic Minority Oversampling Technique (SMOTE) function will be used to balance the dataset so better results can be obtained along with a better representation of the models. The SMOTE function randomly duplicates points in the minority class and balances the dataset by adding duplicate values of the data.

Results

Unsupervised Sentimental Analysis

We then applied polarity and subjectivity to the column Lemma. Polarity outcomes were positive, neutral, or negative, whereas subjectivity results were zero or negative. When it comes to polarity, a tweet is positive if it has a polarity larger than zero. The tweet is neutral if the polarity is zero, and negative if the polarity is less than zero. When it comes to subjectivity, if the tweet's subjectivity is closer to zero, it's objective or factual, however, if it's closer to one, it's subjective, meaning that it has an opinion. The subjectivity and polarity graph is shown in the image below, with 59.04 percent objective tweets and 40.06 percent subjective tweets. From this, we can observe that most of the tweets were objective.

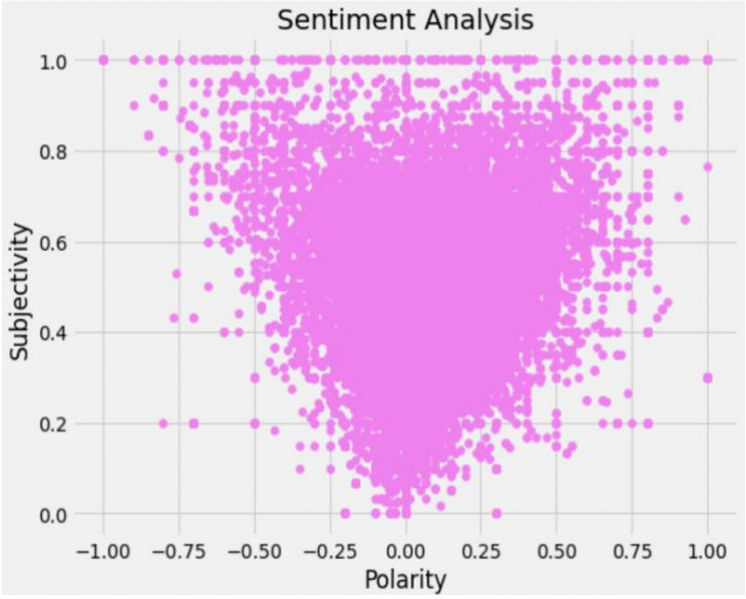


Figure 5: Sentimental analysis subjectivity and polarity

The sentiment analysis of the tweets was visualized using a bar graph represented below. We can see that we have over 12000 positive tweets, nearly 10000 neutral tweets, and 7000 negative tweets, as can be seen from it. To be specific, 43.09 percent of tweets were good, 32.09 percent were neutral, and 23.02 percent were negative.
This outcome surprised us because we expected most of the tweets to be negative owing to the word ‘stress,’ but it was the other way round.

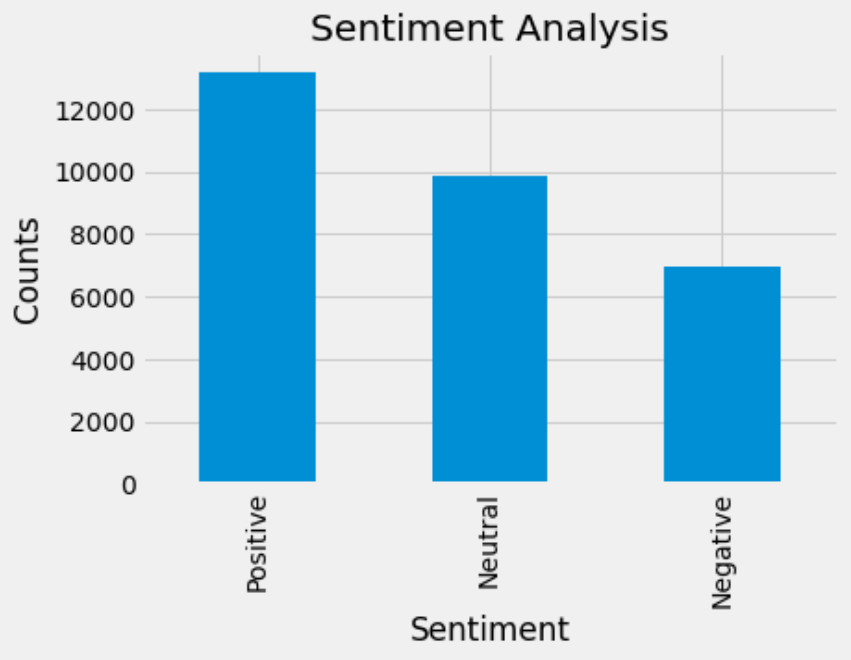
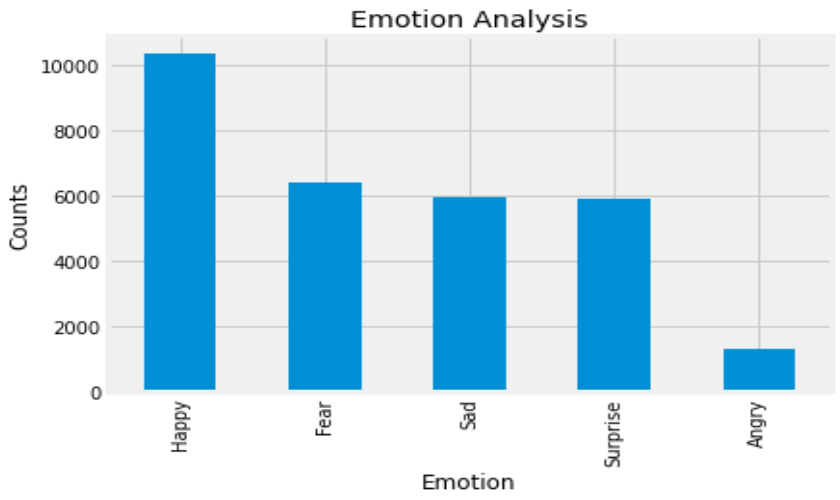


Figure 6: Visualization of Sentiments

Emotional Analysis

Here we study the frequency of different dominant emotions with respect to the total count of tweets. Also, we have constructed word clouds for each emotion to understand the most commonly used words which determine the emotion. In these results, it is evident that the most tweets have an emotion of which is quite surprising as stress is a negatively connotated word however if we look at these results closely the count two heavily negative emotions put together such as fear and sad show that most of the tweets have negative emotions behind them and this answers our research question. From this analysis, we can visualize the number of tweets that have fear and sad emotions behind them which can help social media companies to make decisions in regards to helping individuals who might be going through mental health problems such as stress. This visualization is a key fundamental step in the decision for social media companies to make changes to their policies.



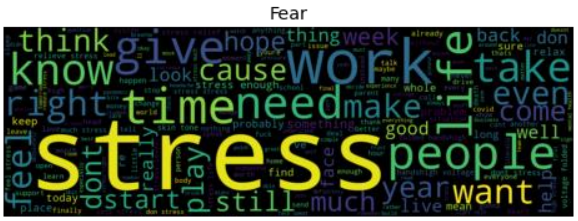


Figure 7: Emotional Analysis visualization

Topic Modelling

From the results, the given text corpus was categorized into 10 different topics which had some relevance in them. This was a good insight as this result shows the variance of words used to tweet and the categorization of these words into different topics. From the results, we can the variation in the words categorized in different topics and also can see the number of negative words used in one specific topic and the number of positive words used. For example, there is a greater number of positive words in topic 9 than in topic 0.

topic 0	topic 1	topic 2	topic 3	topic 4
stress	face	tear	enough	tear
face	tear	face	fuck	joyface
enough	like	joyface	tear	dont
like	much	loudly	cant	like
much	dont	faceloudly	face	feel
dont	feel	smile	joyface	know
fuck	loudly	stress	loudly	time
feel	joyface	weary	faceloudly	much
time	enough	sweat	hate	people
make	time	grin	actual	make
tear	need	joyloudly	smile	take
need	make	upside-down	weary	want
want	know	plead	vote	think
really	really	faceface	faceface	good
think	smile	stressface	police	things
know	take	woozy	joyloudly	game
good	want	pensive	hell	work
take	think	heart-eyes	acab	give
people	work	heart-eyessmilingupside-down		would
work	people	beam	trump	dream
even	fuck	sunglasses	stressface	stressface
life	faceloudly	slightly	donald	youre
cant	good	faceweary	absolute	life
watch	even	abeg	bitch	faceface
loudly	love	stressedface	bastards	well
love	please	heartssmiling	stressedface	better
game	life	tearsmiling	holy	shit
help	cant	relieve	importance	care
dream	would	frown	donate	something
please	help	persevere	woozy	sure
cause	heart	stressedloudly	redtape	keep
would	cause	sweatgrinning	faceweary	enough
never	watch	joyrolling	america	every
heart	come	nigga	kante	could
look	look	hearts	monkey	arsenal
right	right	unamused	mcdonald	also
give	game	chale	communication	thats
level	back	downcast	ridiculous	thing
come	things	facesmiling	beam	talk
free	never	disappoint	badly	help
today	give	eyesbeaming	brady	health
shit	still	roll	joyrolling	stressedface

topic 5	topic 6	topic 7	topic 8	topic 9
fuck	dont	much	floor	like
like	much	would	laughingrolling	much
feel	fuck	cant	laugh	dont
shit	please	cause	roll	floor
tear	want	hate	need	feel
make	need	love	much	laughingrolling
even	know	arsenal	time	laugh
watch	love	fuck	take	roll
actual	take	thank	work	loudly
joyface	dream	take	watch	faceloudly
holy	heart	work	skin	face
lmao	youre	tear	make	dream
floor	give	humble	really	fuck
tire	care	better	would	sound
gon	loudly	school	game	look
never	deserve	deserve	tone	forget
people	well	joyface	back	hate
paypal	forget	people	give	seem
police	time	life	help	arsenal
actually	bill	less	love	humble
stop	okay	already	people	enough
yall	faceloudly	heart	level	felt
hell	urself	give	facerolling	lmao
literally	plead	never	right	facerolling
look	anymore	damn	thank	joyrolling
tell	sandraaat	literally	joyrolling	anymore
still	essay	mental	still	stress
skin	sorry	faceloudly	good	ball
work	jeffreeee	anxiety	heart	damn
dream	humble	year	away	theyre
annoy	arsenal	pretty	could	stressloudly
bitch	face	miss	life	funny
minutes	amaze	handle	arsenal	werner
anxious	purple	team	cause	pretty
right	future	honestly	free	grin
movie	youll	loudly	come	fine
sick	great	could	break	lady
game	yall	husband	never	weary
shut	holy	health	less	overthink
today	rest	lately	please	squeeze

Figure 8: Topic Modelling visualization

Binary Classification

First, the SVM model was implemented using the TFIDF Vectorizer and SMOTE function to convert the raw documents into a TF-IDF feature matrix and balance the dataset. The results showed that the SVM model performed better than the Logistic Regression algorithm this can be seen by the classification report in the figure below. The f1-score for the SVM algorithm after performing SMOTE showed an accuracy of 92% compared to the Logistic Regression model which showed an accuracy of 89% however when looked at the results closely it can be seen that both models perform poorly when classifying bots and much better when classifying users tweeting from iPads, iPhones and the other devices. One of the main reasons for this can be due to a significantly imbalanced dataset. To improve these algorithms and generate better results that can represent the models in a better way it is crucial to get more data.

The results show that the SVM algorithm had a precision of 0.69, a recall of 0.75, and an accuracy of 0.92. The Logistic regression on the other hand had a precision of 0.66, recall of 0.81, and accuracy of 0.89. The precision tells us when the SVM model predicts if a user with an iPad, iPhone, Android device, etc has tweeted it is correct 69% of the time and has a recall of 75%. The Logistic Regression model correctly predicts 66% of users with the top 5 technology devices and has a recall of 81%. Visual representations of the AUC curve and the precision curve can be seen in Appendix A.

```
Before SMOTE : Counter({1: 19702, 0: 1298})
After SMOTE : Counter({1: 19702, 0: 19702})
SVM Accuracy:  92
              precision    recall  f1-score   support

           0       0.40      0.56      0.47       587
           1       0.97      0.94      0.96      8413

    accuracy                           0.92      9000
   macro avg       0.69      0.75      0.71      9000
weighted avg       0.93      0.92      0.92      9000
```

```
Before SMOTE : Counter({1: 19702, 0: 1298})
After SMOTE : Counter({1: 19702, 0: 19702})
Logistic Regression Accuracy:  89
              precision    recall  f1-score   support

           0       0.34      0.72      0.46       587
           1       0.98      0.90      0.94      8413

    accuracy                           0.89      9000
   macro avg       0.66      0.81      0.70      9000
weighted avg       0.94      0.89      0.91      9000
```

Figure 9: Classification Report

Discussion:

This project was aimed to be conducted from a broad angle and examined the use of sentimental analysis then the project narrowed down looking at emotional analysis and topic modeling and finally finished with a binary classification of the sources of tweets found on Twitter during the COVID-19 pandemic. The results regarding the sentimental analysis and emotional surprised the team as when looking at the results of sentimental analysis there were a greater number of tweets that were positive compared to the other two classes of neutral and negative. This surprised the group but made it evident that to discover more about the data that had been collected it was important to conduct emotional analysis to gain an understanding of the real emotions of the tweets tweeted. Hence, we were able to see even though there were a high proportion of tweets that had a sentiment of positive a greater number of tweets had an emotion of fear and sad.

From this project, it can be seen that two models have been created to classify if a tweet has come from a user using one of the five technological devices shown in figure 3 or if the tweet has been tweeted by other third-party sources such as a bot. It is crucial to note that the dataset was heavily imbalanced thus the results are not surprising as both the Support vector Machine and the Logistic Regression model were both able to classify the five technological devices class correctly at a high accuracy but underperformed when classifying the third party/ bots sources. After visualizing both results of the algorithms it can be seen that both algorithms perform at a similar standard however based on accuracy the SVM model shows better results in the binary classification however the logistic regression model does show less accuracy but its precision and recall are in some cases better and par with the SVM model hence we believe the Logistic Regression model is preferred when predicting if the source is one of the five technological devices or a third-party source.

In conclusion, the team was successfully able to identify the sentiments and emotions of the Twitter data that had been collected along with performing topic modeling and implementing a binary classification model to classify the source of tweets. Going back to

our research question and business purpose the team can conclude that the project was able to gain deep insight into the Twitter data that was posted on Twitter during the COVID-19 pandemic and these insights can thoroughly help social media companies such as twitter understand their consumer base during the COVID-19 period much better. Social media companies can use the analysis conducted in this project to the sentiments and emotion behind the Twitter data evaluate if there is a need to provide extra support to their consumer base who might be struggling with stress in future pandemics. The other business need for this project was to help future researchers who are collecting data with no information of whether the tweet has originated from a Twitter user or a third-party source. After creating the two models highlighted in the previous section it can be said that from this project this business need is met however there are still business implications to this as the data used was highly imbalanced and, in the future, when performing unsupervised learning it can be suggested to use multiple search words to narrow down the content. The project can be improved by selecting multiple keywords when searching for tweets related to stress the team could have narrowed down the search significantly of the content that has been searched for. This would help get better results in the unsupervised sentimental analysis and emotional analysis. Another aspect of the project which requires further work is the binary classification which requires more tweets with labels of bots or third-party sources to balance the current dataset and avoid using the SMOTE function. Along with this gridsearch can be used to find the optimum parameters for this project and provide better results.

References

Department, S. R. (2021). *Social media use during COVID-19 worldwide - statistics & facts*. Retrieved 10 27, 2021, from https://www.statista.com/topics/7863/social-media-use-during-coronavirus-covid-19-worldwide/#dossierKeyfigures

Foysal, A., Islam, S., & Rahaman, T. (2019). Classification of AI Powered Social Bots on Twitter by Sentiment Analysis and Data Mining through SVM. *International Journal of Computer Applications* , 13-19.

Gohil, S., Vuik, S., & Darzi, A. (2018). Sentiment Analysis of Health Care Tweets: Review of the Methods Used. JMIR Public Health Surveill.

IANS. (2021, April 30). At 199 million, Twitter logs 20% user growth as pandemic posts surge. San Francisco .

Jain, V. (2013). Prediction of Movie Success using Sentiment Analysis of Tweets. *The International Journal of Soft Computing and Software Engineering, 3*(3), 308-313.

Mathur, A., Kubde, P., & Vaidya, S. (2020). Emotional Analysis using Twitter Data during Pandemic Situation: COVID-19. IEEE.

Panchal, N., Kamal, R., Cox, C., & Garfield, R. (2021). The Implications of COVID-19 for Mental Health and Substance Use. Unkown: KFF.

Rodríguez-Ruiza, J., Mata-Sánchez, J. I., Monroy, R., & Loyola-González, O. (2020). A one-class classification approach for bot detection on Twitter. *Computers & Security*.

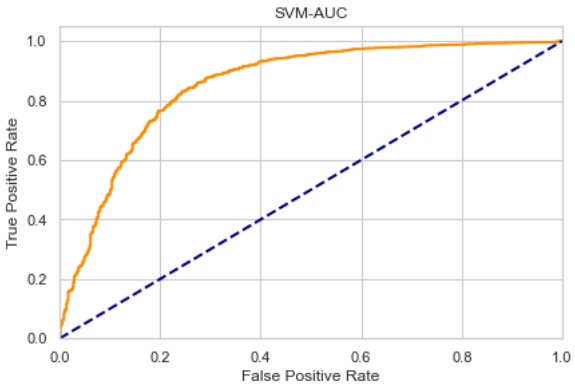
Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2013). Predictive Sentiment Analysis of Tweets: A Stock Market Application. *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, 7947*, 77-88.

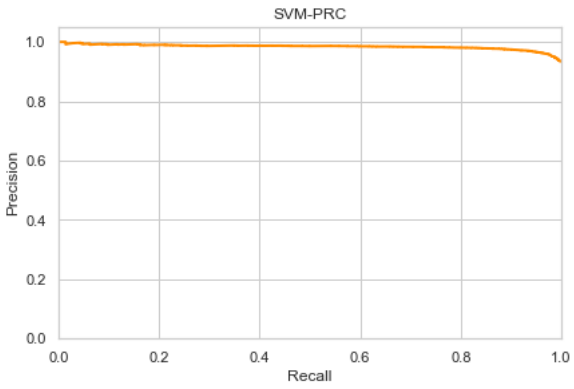
Wang, X., Zheng, Q., Zheng, K., Sui, Y., Cao, S., & Shi, Y. (2021). Detecting Social Media Bots with Variational AutoEncoder and k-Nearest Neighbor. *Applied Science*.

Bing Liu. *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, May 2012 from https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

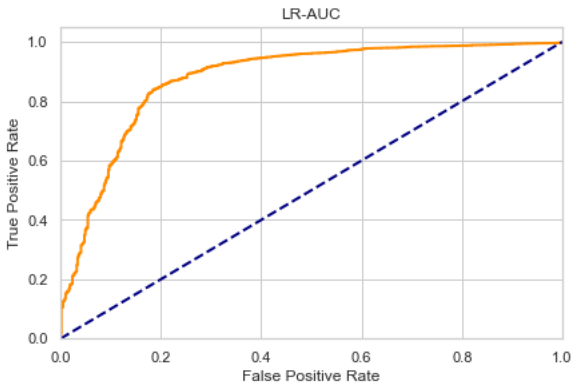
Appendix

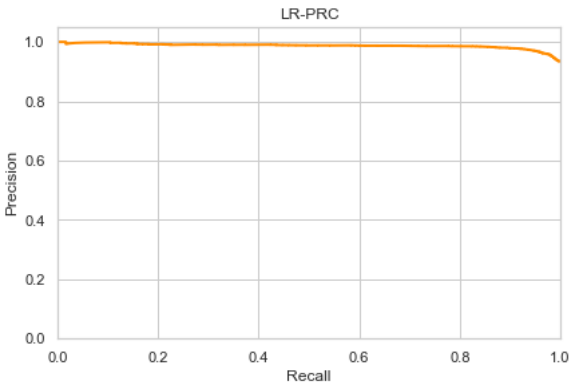
Appendix A – SVM





Appendix B – Logistic Regression





Appendix C – Link to code

https://colab.research.google.com/drive/11DPyDQetCm1lsmJ3D35YcPl_JYsTMTRR?usp=sharing