COVID-19 Vaccine Understanding and Aftereffects: Twitter User Mental Health Analysis and Pfizer Stock Predictions for Major Regions

MSc Research Project
Data Analytics

Deeksha Chaudhry
Student ID: 20154267

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Deeksha Chaudhry
Student ID:	20154267
Programme:	Msc in Data Analytics
Year:	2021
Module:	Data Analytics
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	16/12/2021
Project Title:	COVID-19 Vaccine Understanding and Aftereffects: Twitter User Mental Health Analysis and Pfizer Stock Predictions for Major Regions
Word Count:	XXX
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Deeksha Chaudhry
Date:	16th December 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

COVID-19 Vaccine Understanding and Aftereffects: Twitter User Mental Health Analysis and Pfizer Stock Predictions for Major Regions

Deeksha Chaudhry
20154267

Abstract

These discussions can help governments, organizations, and individuals respond to the pandemic. Covid-19 has resulted in unprecedented duration of power outages, loneliness, insecurity, and agony. The impact of Covid-19 is investigated utilizing sentiment analysis and emotion categorization throughout this study. The dataset was constructed by collecting tweets from the last two months in The United Kingdom, The United States, Canada, Singapore, and Ireland. These tweets are then analyzed to determine the mental health of those who use Twitter. Following that, these tweets are utilized to forecast stock prices for the Pfizer corporation, as this is the vaccine is widely used in the aforementioned nations. Various machine learning models were applied and analyzed to get the desired and more accurate results.

1 Introduction

Novel Corona Viruses, generally known as COVID-19, have proliferated across the globe. This virus spreads quickly and is lethal to humans. As of December 3 (*WHO Coronavirus (COVID-19) Dashboard*; n.d.), 2021, there have already been 263 million documented cases of COVID-19 registered to WHO, comprising more than 5 million fatalities. Several almost 8 billion vaccination shots have already been delivered as of Dec 3, 2021. Since so many places have implemented lockdown, the epidemic scenario impacts all areas, particularly the economy. Certain human activities across the world have been affected by the lockdown.

Several people share their thoughts about vaccines on social media sites. This COVID-19 epidemic has demonstrated how and why the spread of misinformation, helped by media platforms and other current technologies, is presenting as much of a threat to global public health and safety as virus infections themselves. Advances in social networks and the internet provide possibilities to keep the masses secure, updated, and interconnected. Unfortunately, the same construction allows and magnify the present information, which threatens to impair worldwide reaction and risk epidemic control efforts.

Whereas youth are least likely to get serious illness through Coronavirus and they indeed are a crucial population in relation with this outbreak and have the communal obligation to assist us prevent spreading. Individuals are perhaps the most engaged digitally, connecting with mostly used five digital platforms including such Facebook, Twitter, Snapchat, and Instagram everyday. An interdisciplinary study was taken to better comprehend how emerging adults are interacting with technologies amid this worldwide communication issue, involving roughly 23K respondents which were aged between 18-40 years from 24 different nations across 5 continents.

Social networks are employed since it offers adaptability, engagement, and low cost. Social media platform programs such as Twitter have grown in popularity not just for advertising as well as for the interchange of knowledge and the formation of subjective opinions. Twitter is expanding rapidly as well as gaining popularity all around the globe Tariq Soomro et al. (2020). Several individuals are using the Twitter platform to aid various opinions, such as either a platform for fighting, governmental agendas, and dissemination of data and information, and it really is playing an important role in societal growth.

This project focuses on analysing tweets retrieved from Twitter utilizing developer access. The tweets connected to Covid 19 are taken into account and utilized to extract distinct emotions and perform mental health analysis and doing sentiment analysis by comparing and classifying the tweets made by users from various countries. The second portion would investigate the Pfizer vaccine and attempt to forecast stock prices for the Pfizer company using data from five key consumers of this vaccine: the United States, Canada, Singapore, Ireland, and the United Kingdom.

A combined examination of machine learning-based approaches will be performed in this paper. The research is divided into parts. The first part will do the sentiment analysis using a list of emotions. This second phase will take up three algorithms, and we will undertake a comparative analysis on each one. Neural Network, LSTM, and Bi-LSTM models are utilized as the machine learning models. The primary goal of this research is to categorize tweets and develop the best algorithm for predicting stock prices for Pfizer vaccine. The third section will go over the findings and discussions, and the last section will go over the conclusions and future scope.

1.1 Research Questions

- Mental health analysis by emotions recognition by scraping twitter's data in pandemic situation
- Predicting Pfizer stock prices by analysing tweets from Twitter for countries that rely heavily on Pfizer with help of machine learning models.

2 Related Work

Several papers have now been published in relation to studying the Twitter dataset on various themes during the COVID-19. In (Villavicencio et al.; 2021) it is seen that quite a few research have focused on Twitter data associated with COVID-19 immunization. During COVID-19 epidemic, Glowacki et al. (Glowacki et al.; 2020) used text mining to discover addiction problems. They collected public tweets combining the two terms

”addiction” ”covid” and identified 14 common subjects, as well as providing debate upon these issues. The data they used contains 3301 tweets. They wish to identify public talks about addictions on Twitter during COVID epidemic, however due to the outbreak, they haven’t concentrated on sentiment analysis concerning addiction.

The purpose of this article Sattar and Arifuzzaman (2021) is to examine public perception of COVID-19 immunization and the consequences of vaccine in relation to health safety precautions. We collect tweets depending on multiple keywords connected to vaccinations and health security concerns following vaccination to interpret public response and policymakers anticipate vaccination campaigns and health and wellbeing initiatives. Using Twitter textual analysis, health scientists and politicians may learn well about public’s attitude to vaccine during the outbreak of coronavirus.

The research Abhishek Akshay Chaudhri (n.d.) is limited to tweets from the Philippines, but we gather tweets from all around the world. As a result, we have almost 1.2 million tweets, and that they only evaluated 993 of them. This study likewise used the Nave Bayes method to forecast categorization, whereas tweets are classified using a lexicon-based classification and the freely accessible tools VADER and Textblob.

The current study (Wang, Ahorsu, Lin, Chen, Yen, Kuo, Griffiths and Pakpour; 2021) discovered that the component of the health of COVID-19 were positively linked with willingness to get vaccinated against it. Various COVID-19 vaccine data sources were strongly linked with PMT constituents and awareness of the process of COVID-19 immunization, however the orientations of these connections differed.

In this work Rahul et al. (2021), NLP was used to analyze the tweets/retweets. The null hypothesis in specified columns have been replaced with 0 such that the research can proceed. By default, the ’user id’ field received inputs in scientific notation, which were subsequently converted to a standardized format for ease of study. A separate data frame was built for the number of tweets generated by different user ids, as well as the unique hashtag used. After that, the column ’tweets’ was transferred into some other dataset for sentiment classification. ’symbol’, ’@mentions, URLs, and RTs were then deleted from the collected tweets. In **?** the stopwords are eliminated, and the tweets are tokenized. The tokens are lemmatized, subsequently untokenized and put to the dataframe as cleansed tweets. This dataframe was downloaded in.csv form for evaluation. Python packages including such ”NumPy,” , ”pandas”, ”nltk” and ”re,” and were utilized for data preprocessing.

In Pramod and Pm (2021) Stock value forecasting is a hard undertaking that need a solid computational foundation in order to calculate relatively long share values. Because stock prices are connected within the structure of the market, predicting expenses will be challenging. In Sayavong et al. (2019) Deep learning has emerged recently as a category of current techniques ideal for autonomous extracting the features and predictions. Deep learning approaches have been demonstrated in several fields, including computer vision and NLP, to be capable of progressively constructing meaningful complex features from the data or smaller characteristics.IIt is mentioned in Wang, Wang, Cao, Li, Sun and Wang (2021), that because stock market activity is complicated, irregular, and noisy, it appears that feature extraction that are meaningful enough to predict things is a major difficulty, and machine learning appears to be a potential solution to this.

The data utilized in the research J et al. (2021) comes from Yahoo Finance library, S andP 500 constituent shares. This provides the open, high, low, close, and volume figures of the 500 largest firms. The suggested approach helps in predicting close price using a closing price. This section goes through how our system works. This technique is divided into various phases.

2.1 Literature Review Summary

According to our results and research, a significant amount of effort has been done in this subject. Various writers and researchers have presented various approaches for sentiment classification and stock forecasting using Twitter data, but none have achieved both objectives combined together. The method consists of various dataset pre processing techniques, feature extraction techniques, selection of features, generating new types of models, using online algorithms, and so many more. However, in the majority of the research, we discovered that tests were conducted on a limited portion of the Twitter dataset. Using a machine learning method, we found the potential for improvement in this study. In the next sections, we will go through the recommended strategy, our research, and our findings, as well as go over the performance evaluation in detail.

3 Methodology

The contemporary digital landscape of the modern period has witnessed a growth in internet consumption. With the rising benefit of the net, the downside has grown in counterpart. This study article examines several procedures such as data gathering, pre-processing, visualizations, extraction of features, and others. The Tweets data was originally gathered by utilizing Twitter's API. The dataset is then refined to prepare it for the following phases, which turns the twitter posts to a readable format.

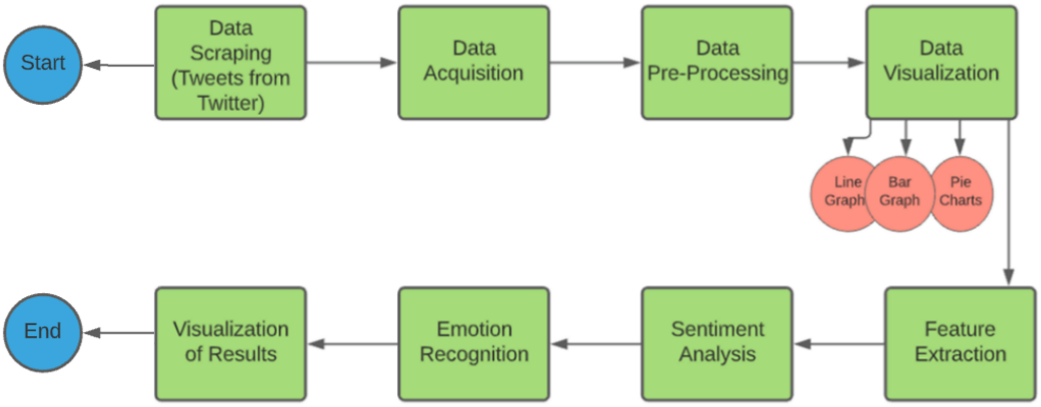


Figure 1: Proposed Framework for Mental Health Analysis through Image Recognition

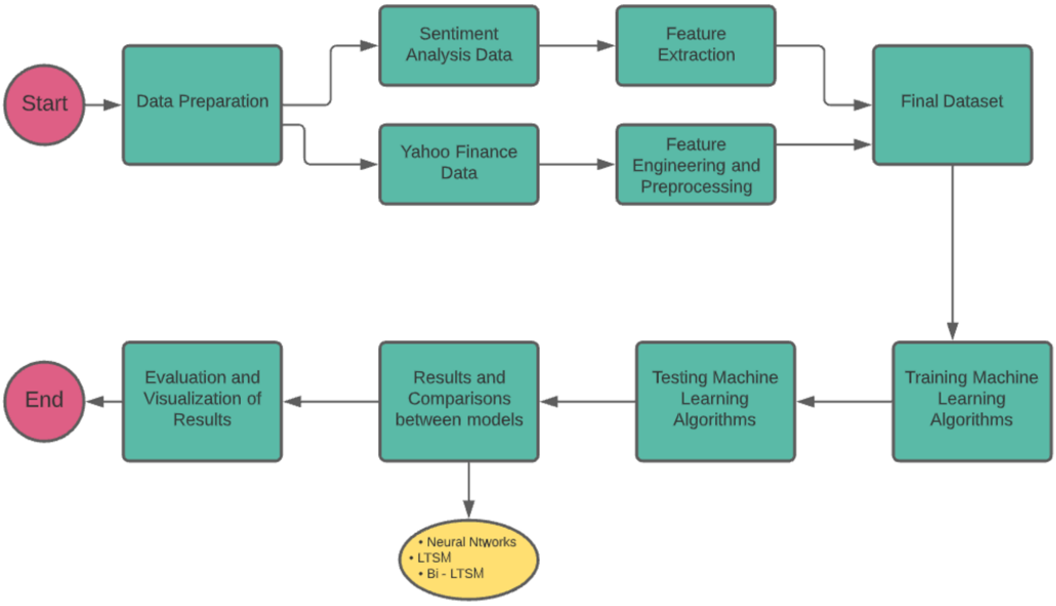
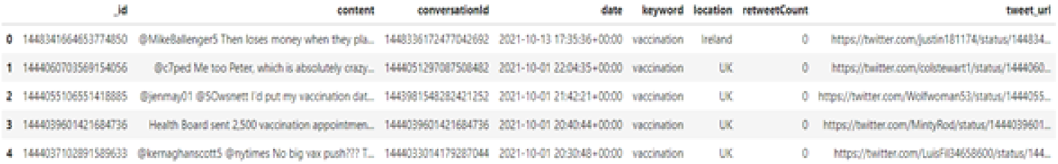

Start
Data Preparation
Sentiment Analysis Data
Yahoo Finance Data
Feature Extraction
Feature Engineering and Preprocessing
Final Dataset
Training Machine Learning Algorithms
Testing Machine Learning Algorithms
Results and Comparisons between models
Evaluation and Visualization of Results
End
• Neural Ntworks
• LTSM
• Bi - LTSM

Figure 2: Proposed Framework for Pfizer stock prediction

3.1 Data Acquisition

Data for this investigation was obtained from Twitter over the last two months while the vaccination rate was high in five major countries: the United States, the United Kingdom, Canada, Singapore, and Ireland. First, we asked Twitter for developer access. Once access was allowed, the credentials were utilized to acquire raw data and tweets based on the Covid-19. To filter out the tweets, we used filters including location, dates, and all Covid vaccination related terms. We obtained almost 146489 tweets from user across the mentioned nations. We also utilized Yahoo Finance Library to collect Pfizer stock data for the last two months, which included the open, high, low, and closing rates for each date for the previous three months.

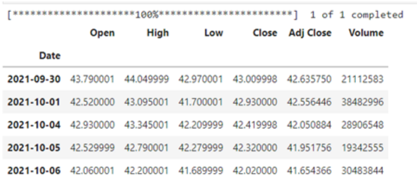


	_id	content	conversationId	date	keyword	location	retweetCount	tweet_url
0	1448341664653774850	@MikeBallenger5 Then loses money when they pla...	1448336172477042692	2021-10-13 17:35:36+00:00	vaccination	Ireland	0	https://twitter.com/justin181174/status/144834...
1	1444060703569154056	@c7ped Me too Peter, which is absolutely crazy...	1444051297087508482	2021-10-01 22:04:35+00:00	vaccination	UK	0	https://twitter.com/colstewart1/status/1444060...
2	1444055106551418885	@jenmay01 @SOwsnett I'd put my vaccination dat...	1443981548282421252	2021-10-01 21:42:21+00:00	vaccination	UK	0	https://twitter.com/Wolfwoman53/status/1444055...
3	1444039601421684736	Health Board sent 2,500 vaccination appointmen...	1444039601421684736	2021-10-01 20:40:44+00:00	vaccination	UK	0	https://twitter.com/MintyRod/status/1444039601...
4	1444037102891589633	@kemaghanscott5 @nytimes No big vax push??? T...	1444033014179287044	2021-10-01 20:30:48+00:00	vaccination	UK	0	https://twitter.com/LuisFil34658600/status/144...

Figure 3: Data head of the tweets data scrapped from twitter

3.2 Data Visualisation

After importing the scraped data from Twitter and Yahoo Finance for the Pfizer vaccination, the information about the feature must be displayed to acquire a better understanding of the dataset. We depicted numerous characteristics of the twitter dataset using line



```
[*********************100%***********************]  1 of 1 completed
                Open       High        Low      Close  Adj Close    Volume
      Date
2021-09-30  43.790001  44.049999  42.970001  43.009998  42.635750  21112583
2021-10-01  42.520000  43.095001  41.700001  42.930000  42.556446  38482996
2021-10-04  42.930000  43.345001  42.209999  42.419998  42.050884  28906548
2021-10-05  42.529999  42.790001  42.279999  42.320000  41.951756  19342555
2021-10-06  42.060001  42.200001  41.689999  42.020000  41.654366  30483844
```

Figure 4: Data head of the tweets data scrapped from twitter

graphs, bar graphs and pie charts. Some information about each dataset characteristic will be gathered to aid in modelling.

In the dataset insight visualization of line graph in figure 5, before October 10, when the corona instances were less, the number of tweets and retweets were fewer. When the number of instances increased in November, the number of tweets remained constant, but the number of retweets soared. The data shows that the most retweets were received on November 9th, with almost 17.43k.

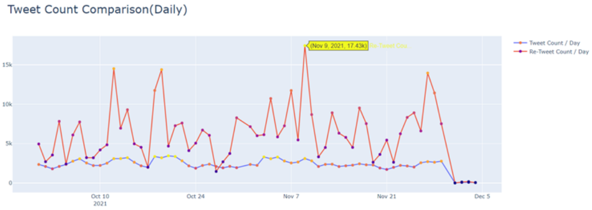


Figure 5: Data head of the tweets data scrapped from twitter

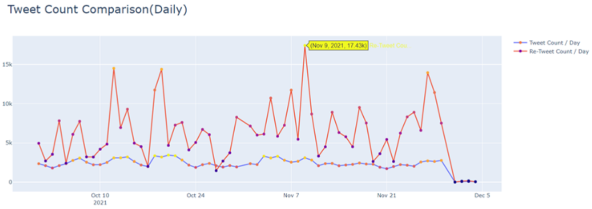


Figure 6: Data head of the tweets data scrapped from twitter

Figure 7 shows a line chart comparing the number of tweets and retweets hourly. It demonstrates how tweets remain consistent throughout the day whereas retweets change significantly during the day.

In another plot figure 8, we can observe the top ten most popular Twitter accounts user as indicated by their Twitter ID. The first Twitter ID identifies the most popular individual on Twitter in terms of tweets and retweets, with around 10k tweets done in the last two months discussing covid vaccine.

We can see five different large nations mentioning about covid vaccination on Twitter in the above Pie chart (figure 8). We could easily examine the percentage of tweets

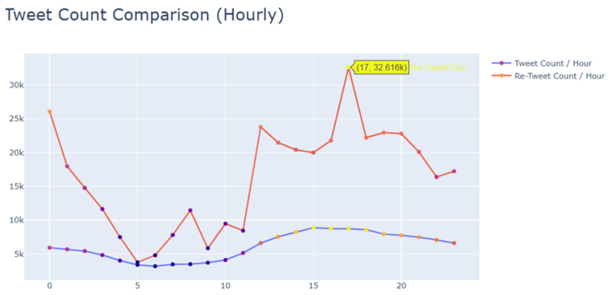


Figure 7: Comparison of tweet and retweet counts hourly

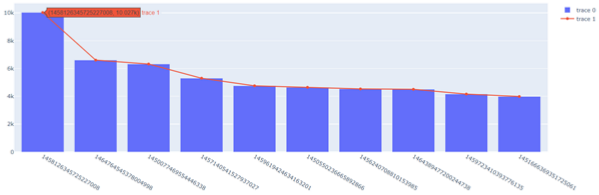


Figure 8: Popular user depiction using bar graph

contributed by each nation. We can observe that the United States alone contributed around 42.4 percent of the tweets, amounting to 62,191 tweets in the previous three months. Canada, on the other side, contributes the least, accounting for only 2.34 percent of total tweets collected which are 3,425.

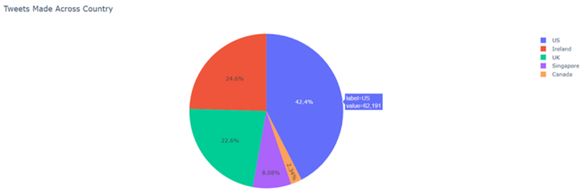


Figure 9: Labels distribution of tweets made in various locations

3.3 Stock Data Visualization

In figure 10, a graphical representation of stock prices fluctuation could be seen with the help of OHLC line graph. An open-high-low-close chart is a type of graph that is often used to depict price changes in a financial product over some time duration. Each horizontal line on the graph represents the pricing structure (highest and lowest prices) over an amount of time.

Figure 11 shows a line graph that depicts price swings over time, and a bar graph that depicts the number of tweets during that time period. We noticed that when booster doses were introduced in November, there was a jump in covid vaccine-related tweets as well as a rise in Pfizer share prices, which had previously been stable.

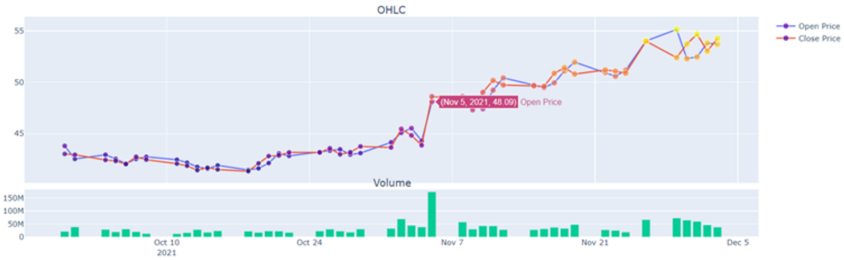


Figure 10: Open, high, low and close for Pfizer during last two months

In the second figure of OHLC, distinct values could be observed at the period when the volume for vaccine tweets was at its peak, with an open price of 48.09 and a close price of 48.61 on the same day (Nov 5th)

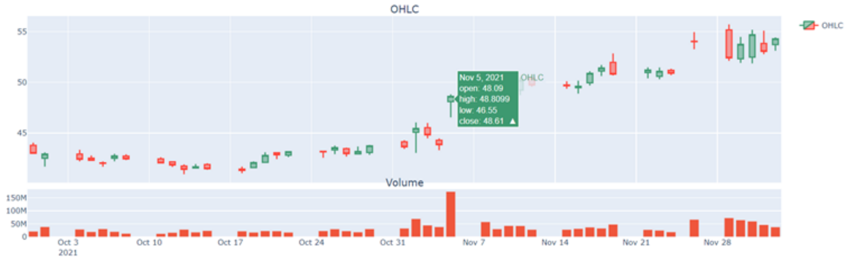


Figure 11: More detailed open, high, low and close for Pfizer during last two months.

3.4 Data Cleaning and Pre-Processing

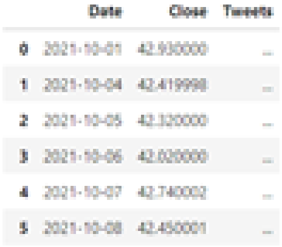
The dataset pre-processing step is crucial since it is where various data treatments are carried out. Identifying null values, missing values, generalizing data, balancing data, and other data pre-processing phases are all part of the data pre-processing process. For our data to be accurate we changed all the content on Twitter into lower case for our data.

There are some URLs that must be eliminated. As a result, for adequate data analysis, all URLs were eliminated.We also eliminated all the ” from hashtags, as well as the usernames. For a more balanced sample, we additionally eliminate punctuation and any repeated characters in text. For example, ’helllloooo’ gets converted to ’hello.’

Following that, we must eliminate the remaining stopwords. Stop words are terms that are commonly found in nature and are taken out before the text is processed. Stop words include all articles, pronouns, prepositions, and other words. As a result, we eliminated all of the stop words from our text. After, all the processing we obtained a total of 119378 words for further analysis.

3.5 Data Preparation For Stock Prediction

The data for stock prediction is organized and classified in such a way that all dates are mapped with the closing price on that specific date.



	Date	Close	Tweets
0	2021-10-01	42.930000	...
1	2021-10-04	42.419998	...
2	2021-10-05	42.320000	...
3	2021-10-06	42.020000	...
4	2021-10-07	42.740002	...
5	2021-10-08	42.450001	...

Figure 12: Stock Prediction Dataframe

4 Implementation

Our preliminary purpose in this research is to investigate the machine-based technique for improved analysis of tweets/retweets for mental health analysis and prediction of Pfizer stock prices. With new Tweets being sent on Twitter every day, the collecting of data with various patterns is becoming increasingly difficult. As a result, it is vital to substantially refine the data gathered before incorporating it into the Machine Learning model. We employed a variety of libraries for this goal, including pandas, sklearn, matplotlib, NLTK, Tensorflow, SentimenIntensityAnalyzer, Yfinance, and NumPy. D data was extracted that had been filtered and had only relevant characteristics. Matplotlib library is used for visualization of data, including such bar charts, pie charts, and line graphs, among other things. We also utilized the seaborn library to connect the pandas library to the matplotlib package for statistical visualizations. We also utilized SentimentIntensityAnalyzer to analyze the sentiment of the words collected from the tweets. In our model, this library is used for processing the words. We also utilized the warning library to notify a few of the program's components. Because our models were executed on a single computer, the specification for our suggested framework is as follows.

- Operating System: Windows 10
- Hard Disk: 1TB
- RAM: 8GB
- Programming Language: Python3
- Libraries: Pandas, NumPy, Matplotlib, Sklearn, ,yfinance, NLTK, SentimenIntensityAnalyzer, Tensorflow

4.1 Sentiment Analysis for Pfizer stock data processing

Sentiment Analysis is the method of detecting if a text is positive, negative, or neutral 'algorithmically' S et al. (2021). It is also called as opinion mining since it involves determining a speaker's viewpoint or perspective.When it comes to sentiment analysis, there are two main ways.

- Deep Machine learning or Supervised machine learning technique.
- Approaches based on unsupervised lexicons.

It can be used in three major sectors the business, politics and for public actions analysis. Organizations makes use of it in business to establish strategies, analyse consumers' attitudes about brands and products related to them, how the individual react to promotions or new product introductions, and why other items are not purchased. Sentiment Analysis is used in politics to keep a record of political viewpoints and to find uniformity and discrepancy between governmental statements and behaviour.*Python: Sentiment Analysis using VADER* (2021) It is sometimes used to forecast results of the election! Sentiment classification can also assist to track and evaluate sociological trends, identifying potentially harmful circumstances, and evaluating the overall atmosphere of the internet

In this, we conducted a sentiment analysis using Twitter COVID-19 dataset for the purposes of this research using VADER Sentiment Analysis Tariq Soomro et al. (2020). To begin, it is critical to grasp regarding sentiment analysis, interpretation and categorization of emotions within textual information using text analytical. VADER, or Valence Aware Dictionary and Sentiment Reasoner, is a social media-aware lexicon and rule-based analyzer.VADER uses a combination of lexical properties (such as words) that are classified as positive or negative depending on the underlying sentiment polarity Singh 08/12/2020 (2021) . VADER displays not just the Positivity and Negativity scores, but also the degree to which an emotion is positive or negative.

In our research we have classified tweets using SentimentIntensityAnalyzer as an object whereas polarity-scores act as a method which gives us the score as positive, negative, neutral and compound. The total of Positive, Negative and Neutral ratings is known as Compound Score. The score is then adjusted between +1 to -1 as most strongly positive to most severe negative respectively Singh 08/12/2020 (2021) . The closer the compound score is to 1, more positive the text is.

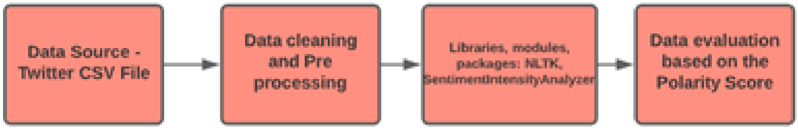
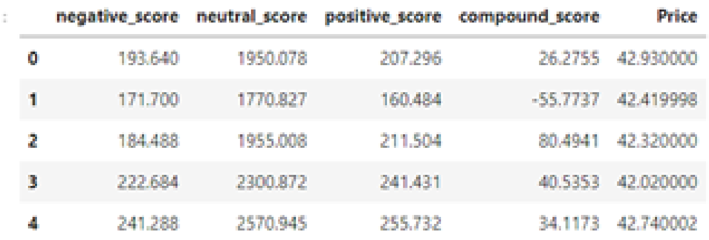

Data Source - Twitter CSV File
Data cleaning and Pre processing
Libraries, modules, packages: NLTK, SentimentIntensityAnalyzer
Data evaluation based on the Polarity Score

Figure 13: Sentiment Analysis using VADER

The graph depicts how the compound score percentage for each tweet in our dataset is calculated. We can simply demonstrate that an increase in negative score leads to a fall in pricing. This means that the bad tweets are causing Pfizer's stock price to decrease.



	negative_score	neutral_score	positive_score	compound_score	Price
0	193.640	1950.078	207.296	26.2755	42.930000
1	171.700	1770.827	160.484	-55.7737	42.419998
2	184.488	1955.008	211.504	80.4941	42.320000
3	222.684	2300.872	241.431	40.5353	42.020000
4	241.288	2570.945	255.732	34.1173	42.740002

Figure 14: Sentiments Score evaluation for tweets using VADER

4.2 Model Approach

In order to forecast Pfizer stock prices, an effective methodology must be used. Taking the majority of the techniques into consideration, it is concluded that Machine Learning is by far the most appropriate option. As a result, we compared three techniques in our paper: Neural Network, LTSM, and Bi-LTSM. We'll go through a quick summary of each approach we'll be using down below. We first created the train and test data to train our models and then perform testing for our models. We have used iloc create the final data for training and testing purposes.

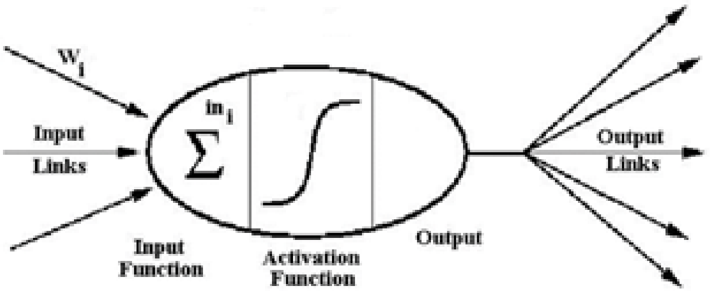


Figure 15: Neural Network

4.2.1 Neural Network

Neural networks, also known as simulated neural networks (SNNs) or artificial neural networks (ANNs), are indeed a subfield of machine learning which provide the basis of deep learning approaches. Their terminology and shape are comparable to that of a human mind, and they mimic how genuine neurons interact with each other Menon et al. (2019). The neuron receives a lot of signals (inputs), all of which have a weight (wi) related with it. That's then multiplied by a threshold value which is b then passed as an argument to the activation function. These computation units, which serve as the various sections, can be organized in layers.

A Neural network needs data as input [number of data points, time steps, attributes], where data points are the samples, time steps is the count of time-dependent stages in a single sample, and attributes are the count of variables for the associated actual value.

The first layer in Neural Network is layer with 64 memory units. This return sequences which ensures that the next layer receives the sequence and not the random data which is scattered. The next layer is with 128 memory units. We also used relu as our activation function for this model which generates output only if the input will be positive else it will output zero.

The model has been fitted across 50 epochs. To address sparse gradients in a noisy area, we employed the Adam optimizer in our model. We used Keras to create a set of metrics to track while the model is training. At the conclusion of each epoch, the metrics value will be assessed. We calculated Mean squared error and Mean absolute error for our data at the conclusion of each epoch.

4.2.2 LTSM

A basic LSTM network is composed of several blocks of memory referred as cells. The two main components that are now passed to the next cell are the cell state and the concealed state.Memory blocks are in charge of memorizing, and modifications to this information are made via three key process which is known as gates. These gates are input gate, output gate and forget gate *Long Short Term Memory: Architecture Of LSTM* (2020a). We utilized the Keras library, an elevated API for neural-networks that extend the functionality of TensorFlow.

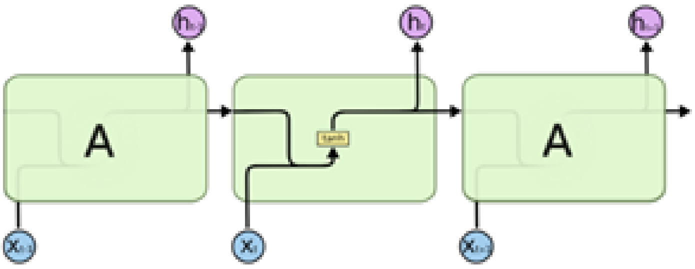


Figure 16: LTSM

We have used the reshape function to train and test data from 2D to 3D. A LSTM network needs data as input [number of data points, time steps, attributes], where data points are the samples, time steps is the count of time-dependent stages in a single sample, and attributes are the count of variables for the associated actual value. In this model, we have used Tanh as the activation function. The Tanh function is another name for the hyperbolic tangent activation function. It is quite close to, and also has the identical S-shape as, sigmoid activation function. The method accepts any actual value as input and returns values ranging from -1 to 1. We have used the same compile function with 50 epochs and Adam optimizer in the model.

4.2.3 Bi-LTSM

Bi-LTSM or Bidirectional long-short term memory is the technique of allowing any neural network to store sequence data in both ways, either backwards or forwards i:e from future to past and vice versa. We may enable the inputs flow across both ways in bi-LTSM to retain both past and future information Verma 17/07/2021 (2021). We can observe the information flow flows back and forth in the diagram.

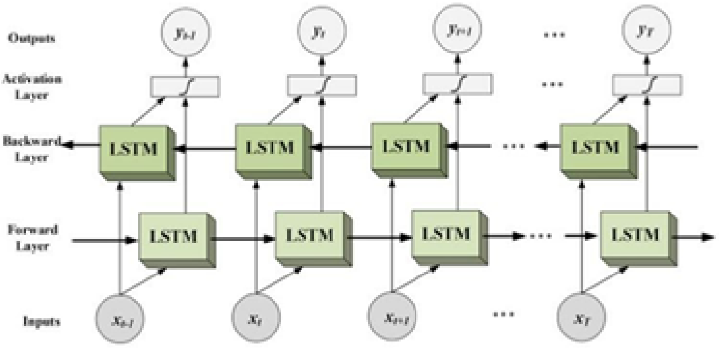


Figure 17: Bi-LTSM

BI-LSTM is typically used in jobs that need sequential matching. Text categorization, forecasting models and speech recognition can all benefit from this type of system. We have introduced a bi-LSTM layer to a conventional neural network utilising Keras. Tensorflow's Keras now has a new class [bi - directional] for generating bi-LSTM. We trained this model too using a training dataset with 50 epochs.

5 Evaluation

Our main objective in this research paper is to develop an effective algorithm to determine people's mental health for a given time period by analyzing tweets from five important nations, after which we forecasted Pfizer stock prices because these are the countries that use Pfizer vaccine more than any other. Sentiment Analysis, Neural Network, LTSM, and Bi-LTSM were among the methodologies we used to incorporate in research models.

We utilized criteria such as mean squared error and mean absolute error to select the best-performing model based on its performance and efficiency. With these metrics scores, we may consider developing an effective system to forecast Pfizer vaccine stock prices. We will compare the performance of every model based upon the metrics in this section.

The first finished bar graph represents the mental health of individuals who tweeted between 01-10-2021 and 03-12-2021. We extracted 120K words from the tweets and matched them to a set of emotions stored in a text file called emotions. As shown in recent research, there are up to 27 different types of emotions. These emotions were imported into the code and then measured. All the retrieved terms from tweets were compared to this file, which aided in identifying the users' moods and present state of mind. This storyline allows us to analyse how individuals feel and depict their mental wellbeing in their Twitter comments/gestures, which is the core objective of our study. It was evident that the emotions of sadness, fear, happiness, and anger had the greatest count. In comparison to this, the rest of the emotions had a far lower count.

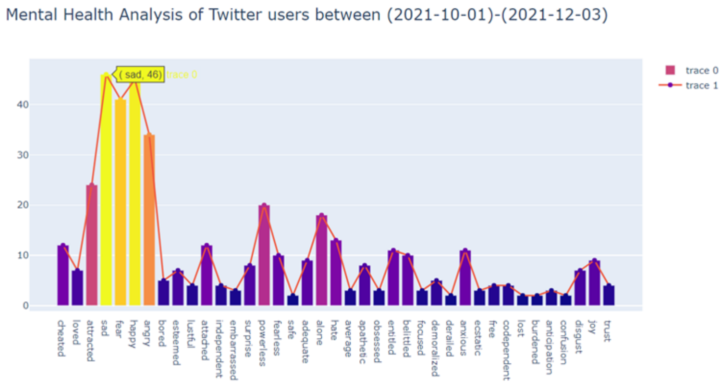


Figure 18: Mental Health Analysis Through Emotions Recognition

As a result, we may determine that the top 10 leading feeling in our Twitter dataset during the previous three months has been sadness, accounting for 17.4 percent of all feelings. Happy and fear are the second and third most common emotions, accounting for 17 percent and 15.5 percent, respectively. The covid 19 circumstance has a significant psychological impact on people, so we attempted to convey this as much as practicable.

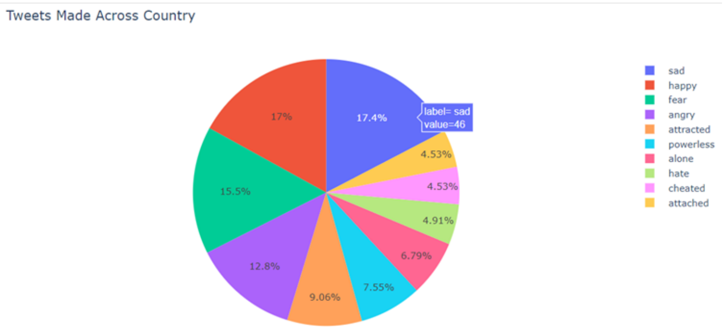


Figure 19: Tweets Percentage Made Across the Country

5.1 Machine Learning Models Performance

As during development and training of a ML model, the present state of the machine learning model may be assessed at each iteration of the training phase. It may be tested against the training data set to determine how effectively the model is learning. It could also be tested on a separate validation data set which is not included in the training sample. The validation data set is used to determine how effectively the model generalizes. In learning curves, the models exhibits three behaviors: underfit, goodfit, and overfit. The Mean Squared Error is computed, and it is the most generally used and successful programming metric *Long Short Term Memory: Architecture Of LSTM* (2020b). The equation is as follows.

$$\text{MSE}=\frac{1}{N}\sum_{p=1}^{P}\sum_{i=1}^{N}(t_{pi}-y_{pi})^2$$

Where, t_{pi}= Predicted value for data point i;
y_{pi} =Actual value for the data point i;
N = Total number of data points

Figure 20: Mean Squared Error

The underfit model can also be spotted by a falling training loss which continues to decline at the conclusion of the graph *Long Short Term Memory: Architecture Of LSTM* (2020b).Underfitting is indicated if the training loss continues flat irrespective of learning or continues to decline until the completion of training.

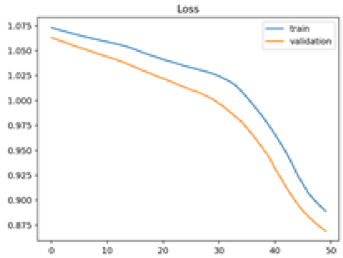


Figure 21: Training and Validation curves for Underfit Model

A model shows goodfit when that model has potential of more learning and prospective enhancements. A good-fitted plot is when the training loss graph lowers to a stable point or validation loss graph reaches a stable point and also has a narrow gap with the training loss plot.

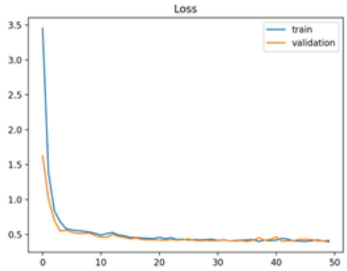


Figure 22: Training and Validation curves for Goodfit Model

Over fitting is considered as a model that really has learnt the training data set quite well, incorporating statistical uncertainty or random variations.

The graph is a line graph that compares the loss, which is mean squared error, produced for all three models using the training dataset. This demonstrates how effectively

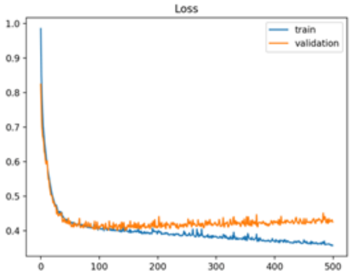


Figure 23: Training and Validation curves for Overfit Model

our model is learning. The mean square error function is the fundamental performance function that has a direct impact on the network Singh et al. (2014). The elimination of such errors will result in a more efficient system.

The model is thought to be superior to others since it has a lower mean squared error. If we look at the graph, we can see that the loss for the neural network model is significantly higher than that of the other two models, LTSM and Bi-LTSM. Both the Train loss graph and the Validation loss graph show that LTSM and Bi-LTSM meet all of the requirements for a goodfit model. Both graphs narrow to a steady point with very little difference in between. This suggests that the algorithm is learning fairly well and might be used to forecast stock prices.

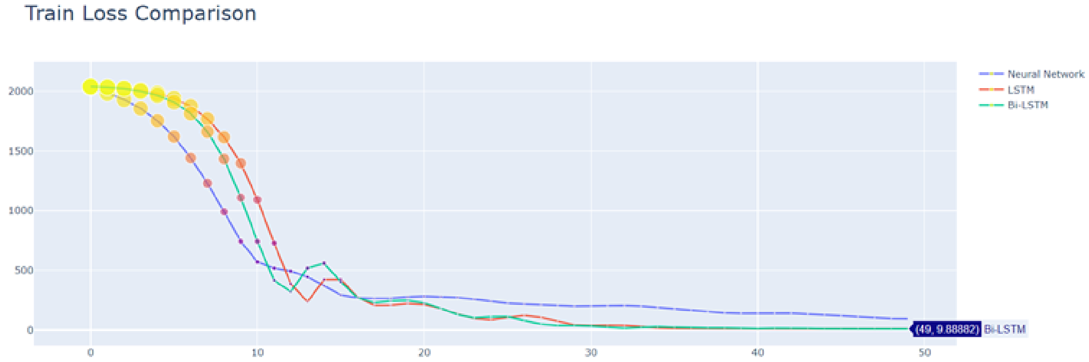


Figure 24: Train Loss Comparison

We also evaluted the mean absolute error for our models. The mean square error, abbreviated as MSE, is the predicted value of the square of the difference seen between predicted and actual values of a variable. TMAE is the arithmetic average errors. It can more accurately reflect real scenario of the error in predicted value.

There are substantial errors between Neural Network Model forecast values and the real data. The prediction values and the true data values for the LTSM model and Bi-LTSM are practically coincident, and the variation here between the prediction and the true value is very tiny, showing that the effectiveness of these two approaches is superior than Neural Network in the testing data.

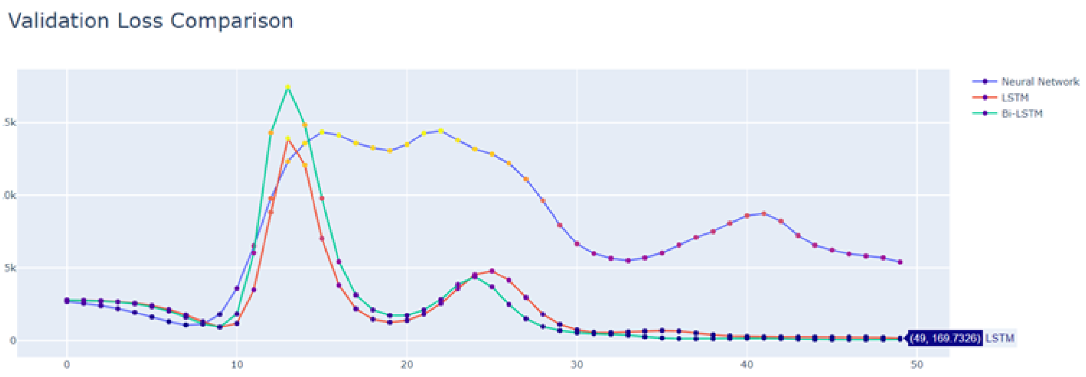


Figure 25: Validation Loss Comparison

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

Figure 26: Formula for Mean Absolute Value

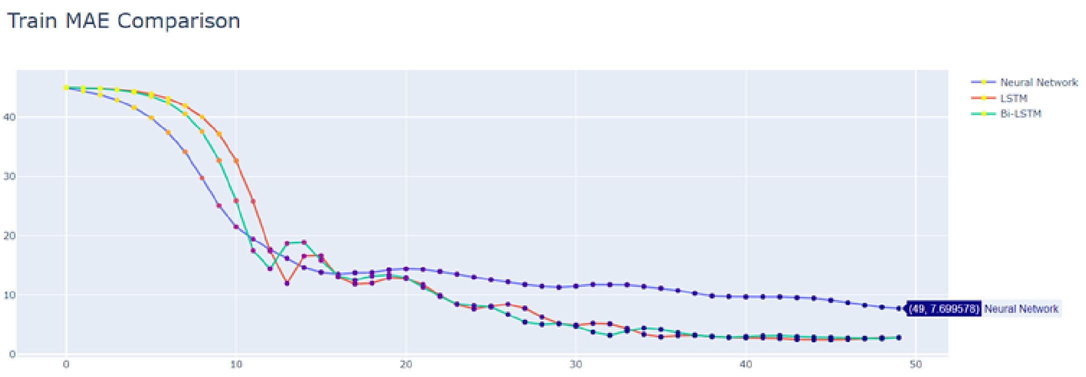


Figure 27: Formula for Mean Absolute Value

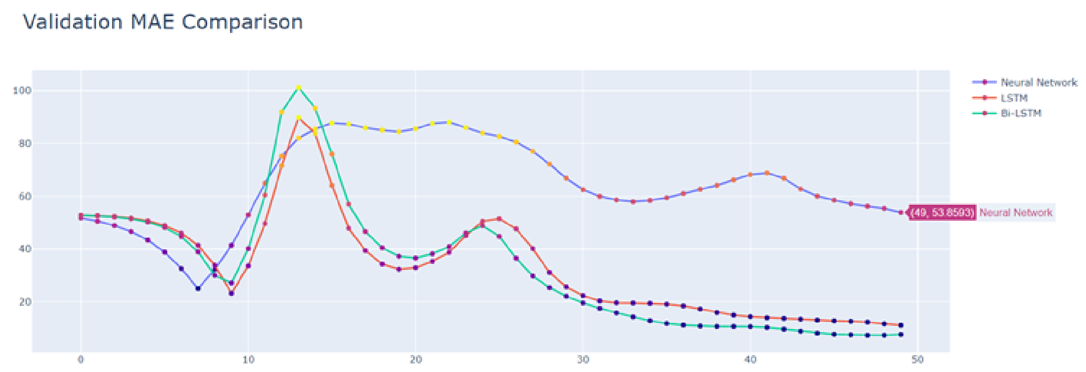
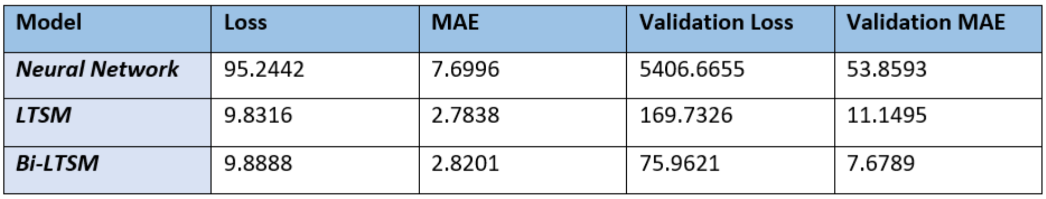


Figure 28: Formula for Mean Absolute Value

Each model's actual values are shown in the table. It is clear that the Bi-LTSM model is the best model in this situation for predicting Pfizer stock prices since it has a highly consistent score and the discrepancy between the real and predicted values is considerably smaller than the other two models.



Model	Loss	MAE	Validation Loss	Validation MAE
Neural Network	95.2442	7.6996	5406.6655	53.8593
LTSM	9.8316	2.7838	169.7326	11.1495
Bi-LTSM	9.8888	2.8201	75.9621	7.6789

Figure 29: Metric Table

6 Conclusion and Future Work

The most known and widely debated issue is covid. This pandemic crisis has had far-reaching consequences all throughout the planet. The outbreak seems to have an economic impact and even a detrimental influence on lifestyle, food shortages, employability, and mental health. It is vital to avert more damage as a result of such a crisis throughout the world Khattar et al. (2020). People's sentiments in a certain place are directly tied to the growth of that region.

This research focuses on analyzing people's opinions and doing mental health analyses in a given location. The initiative will be tremendously valuable in terms of early knowledge and prevention of additional damage. Anxiety, worry, sleep loss, and sadness are already symptoms of the scenario. As a result of the circumstance, people have already been suffering tensions, anxiety, sleeplessness, and sadness. Among the most precise accurate predicting approaches is employed in this work, which assists investors, researchers, and persons prepared to invest by providing solid knowledge again for future condition of exchange in share market.

The research of the part is considered in this work, and that can be performed out for multiple interests in the hereafter. Prediction will be much more trustworthy if the system trains on a larger number of information sets with better processing capacity, a larger number of layers, or Long short - term memory module. Future improvements will include the incorporation of sentiment classification from social networks to comprehend whatever the market expects about the price changes for a specific share, which could be implemented by introducing additional Facebook and Twitter API to this program, as Facebook is indeed a popular social media platform with a large proportion of market data and reporting uploaded by people.

References

Abhishek Akshay Chaudhri, S. S. S. (n.d.). Implementation paper on analyzing covid-19 vaccines on twitter dataset using tweepy and text blob.
URL: *https://www.annalsofrscb.ro/index.php/journal/article/view/2381*

Glowacki, E., Wilcox, G. and Glowacki, J. (2020). Identifying addiction concerns on twitter during the covid-19 pandemic: A text mining analysis, *Substance abuse* **42**: 1–8.

J, K., E, H., Jacob, M. S. and R, D. (2021). Stock price prediction based on lstm deep learning model, *2021 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–4.

Khattar, A., Jain, P. R. and Quadri, S. M. K. (2020). Effects of the disastrous pandemic covid 19 on learning styles, activities and mental health of young indian students - a machine learning approach, *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1190–1195.

Long Short Term Memory: Architecture Of LSTM (2020a).
URL: *https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/*

Long Short Term Memory: Architecture Of LSTM (2020b).
URL: *https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/*

Menon, A., Singh, S. and Parekh, H. (2019). A review of stock market prediction using neural networks, *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–6.

Pramod and Pm, M. (2021). Stock price prediction using lstm, *Test Engineering and Management* **83**: 5246–5251.

Python: Sentiment Analysis using VADER (2021).
URL: *https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/*

Rahul, K., Jindal, B. R., Singh, K. and Meel, P. (2021). Analysing public sentiments regarding covid-19 vaccine on twitter, *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1, pp. 488–493.

S, R. B., Ezhilan, A., R, D., R, A. and R, S. (2021). Sentiment analysis and classification of covid-19 tweets, *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 821–828.

Sattar, N. S. and Arifuzzaman, S. (2021). Covid-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the usa, *Applied Sciences* **11**(13).
URL: *https://www.mdpi.com/2076-3417/11/13/6128*

Sayavong, L., Wu, Z. and Chalita, S. (2019). Research on stock price prediction method based on convolutional neural network, *2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*, pp. 173–176.

Singh 08/12/2020, F. (2021). Sentiment analysis made easy using vader.
URL: *https://analyticsindiamag.com/sentiment-analysis-made-easy-using-vader*

Singh, S., Singh, D. S. and Kumar, S. (2014). Modified mean square error algorithm with reduced cost of training and simulation time for character recognition in backpropagation neural network, *in* S. C. Satapathy, S. K. Udgata and B. N. Biswal (eds), *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013*, Springer International Publishing, Cham, pp. 137–145.

Tariq Soomro, Z., Waseem Ilyas, S. H. and Yaqub, U. (2020). Sentiment, count and cases: Analysis of twitter discussions during covid-19 pandemic, *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pp. 1–4.

Verma 17/07/2021, Y. (2021). Complete guide to bidirectional lstm (with python codes).
URL: *https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/*

Villavicencio, C., Macrohon, J. J., Inbaraj, X. A., Jeng, J.-H. and Hsieh, J.-G. (2021). Twitter sentiment analysis towards covid-19 vaccines in the philippines using naïve bayes, *Information* **12**(5).
URL: *https://www.mdpi.com/2078-2489/12/5/204*

Wang, H., Wang, J., Cao, L., Li, Y., Sun, Q. and Wang, J. (2021). A stock closing price prediction model based on cnn-bislstm.
URL: *https://www.hindawi.com/journals/complexity/2021/5360828/*

Wang, P.-W., Ahorsu, D. K., Lin, C.-Y., Chen, I.-H., Yen, C.-F., Kuo, Y.-J., Griffiths, M. D. and Pakpour, A. H. (2021). Motivation to have covid-19 vaccination explained using an extended protection motivation theory among university students in china: The role of information sources, *Vaccines* **9**(4).
URL: *https://www.mdpi.com/2076-393X/9/4/380*

WHO Coronavirus (COVID-19) Dashboard (n.d.).
URL: *https://covid19.who.int/*