Stock Market Analysis Using Twitter

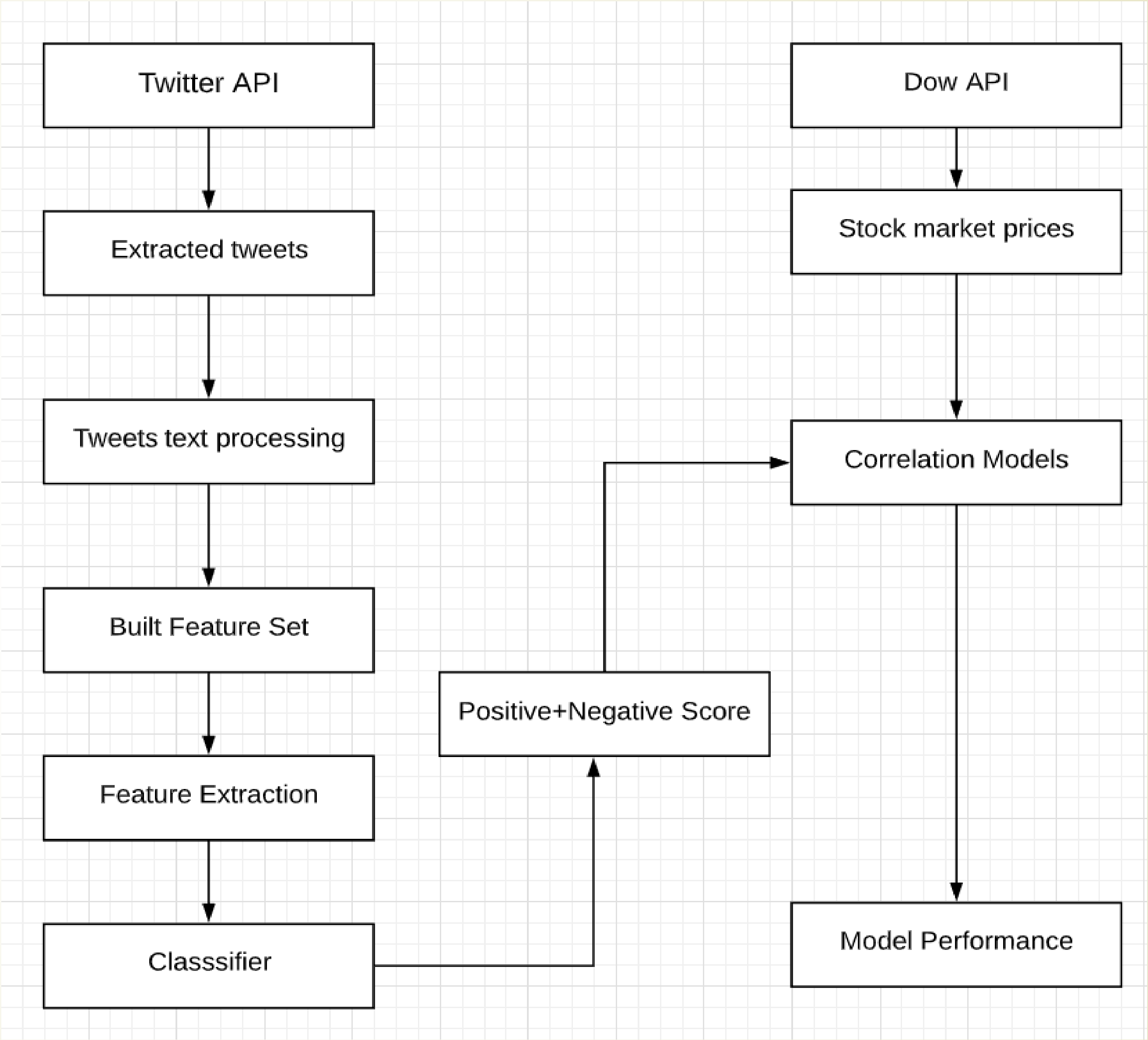
Presented By:
Abhisha Burande
Anshika Sharma
Maitreyee Das
Priyanka Kushwaha
Shweta Arande

Content

- Business Objective
- Web scrapping
- Data cleaning
- Word Cloud
- Unsupervised Sentiment Analysis Using K mean
- Loughran sentiment with Tf-idf
- Sentiment Analysis Using VADER
- Linear regression model
- Random Forest Regression Model
- Comparison of Random FR model and Linear Regression
- Future aspects
- Reference

Business Objective

- The main objective is to perform sentiment analysis and create a model for insights
- **Proposed Research Model:**



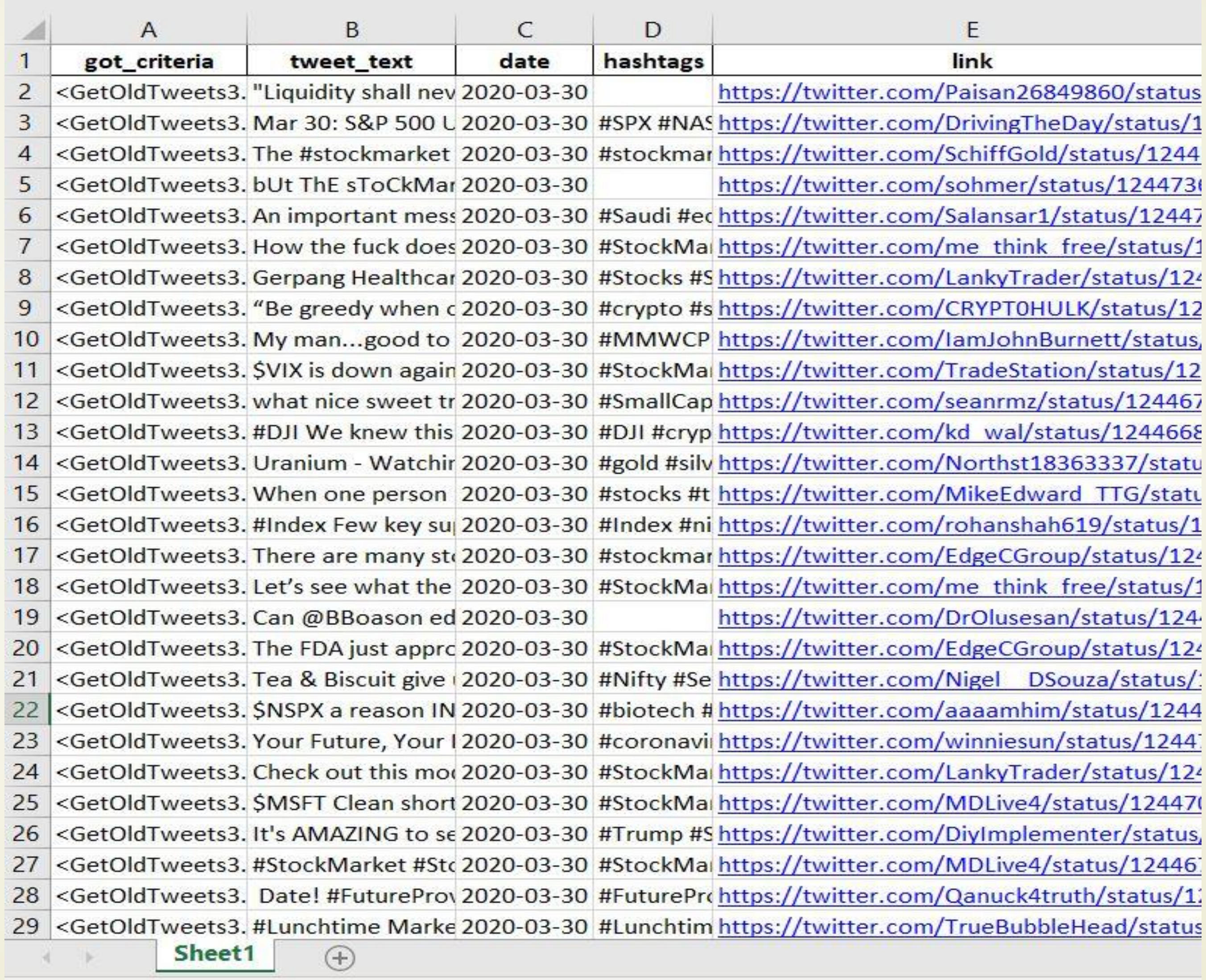
Web Scraping

USED GETOLDTWEETS LIBRARY TO EXTRACT DATA FROM TWITTER API AND DOW VALUES WERE EXTRACTED FROM YAHOO FINANCE

EXTRACTED DATA FROM JANUARY 1,2020 TO MARCH 30,2020.

ONLY TOP TWEETS FOR EVERY WEEK WERE EXTRACTED

Snapshot of Extracted Data from Twitter



	A	B	C	D	E
1	got_criteria	tweet_text	date	hashtags	link
2	<GetOldTweets3.	"Liquidity shall nev	2020-03-30		https://twitter.com/Paisan26849860/status
3	<GetOldTweets3.	Mar 30: S&P 500 U	2020-03-30	#SPX #NAS	https://twitter.com/DrivingTheDay/status/1
4	<GetOldTweets3.	The #stockmarket	2020-03-30	#stockmar	https://twitter.com/SchiffGold/status/1244
5	<GetOldTweets3.	bUt ThE sToCkMar	2020-03-30		https://twitter.com/sohmer/status/124473
6	<GetOldTweets3.	An important mess	2020-03-30	#Saudi #ec	https://twitter.com/Salansar1/status/12447
7	<GetOldTweets3.	How the fuck does	2020-03-30	#StockMa	https://twitter.com/me_think_free/status/1
8	<GetOldTweets3.	Gerpang Healthcar	2020-03-30	#Stocks #S	https://twitter.com/LankyTrader/status/124
9	<GetOldTweets3.	“Be greedy when c	2020-03-30	#crypto #s	https://twitter.com/CRYPT0HULK/status/12
10	<GetOldTweets3.	My man...good to	2020-03-30	#MMWCP	https://twitter.com/IamJohnBurnett/status/
11	<GetOldTweets3.	$VIX is down again	2020-03-30	#StockMa	https://twitter.com/TradeStation/status/12
12	<GetOldTweets3.	what nice sweet tr	2020-03-30	#SmallCap	https://twitter.com/seanrmz/status/124467
13	<GetOldTweets3.	#DJI We knew this	2020-03-30	#DJI #cryp	https://twitter.com/kd_wal/status/1244668
14	<GetOldTweets3.	Uranium - Watchir	2020-03-30	#gold #silv	https://twitter.com/Northst18363337/statu
15	<GetOldTweets3.	When one person	2020-03-30	#stocks #t	https://twitter.com/MikeEdward_TTG/statu
16	<GetOldTweets3.	#Index Few key su	2020-03-30	#Index #ni	https://twitter.com/rohanshah619/status/1
17	<GetOldTweets3.	There are many st	2020-03-30	#stockmar	https://twitter.com/EdgeCGroup/status/124
18	<GetOldTweets3.	Let’s see what the	2020-03-30	#StockMa	https://twitter.com/me_think_free/status/1
19	<GetOldTweets3.	Can @BBoason ed	2020-03-30		https://twitter.com/DrOlusesan/status/124
20	<GetOldTweets3.	The FDA just appr	2020-03-30	#StockMa	https://twitter.com/EdgeCGroup/status/124
21	<GetOldTweets3.	Tea & Biscuit give	2020-03-30	#Nifty #Se	https://twitter.com/Nigel__DSouza/status/
22	<GetOldTweets3.	$NSPX a reason IN	2020-03-30	#biotech #	https://twitter.com/aaaamhim/status/1244
23	<GetOldTweets3.	Your Future, Your	2020-03-30	#coronavi	https://twitter.com/winniesun/status/1244
24	<GetOldTweets3.	Check out this mo	2020-03-30	#StockMa	https://twitter.com/LankyTrader/status/124
25	<GetOldTweets3.	$MSFT Clean short	2020-03-30	#StockMa	https://twitter.com/MDLive4/status/12447
26	<GetOldTweets3.	It's AMAZING to se	2020-03-30	#Trump #S	https://twitter.com/DiyImplementer/status/
27	<GetOldTweets3.	#StockMarket #St	2020-03-30	#StockMa	https://twitter.com/MDLive4/status/12446
28	<GetOldTweets3.	Date! #FuturePro	2020-03-30	#FuturePr	https://twitter.com/Qanuck4truth/status/1
29	<GetOldTweets3.	#Lunchtime Marke	2020-03-30	#Lunchtim	https://twitter.com/TrueBubbleHead/status

Data Cleaning

- Used BeautifulSoup to clean HTML encoding
- Removed @mention as this information doesn’t add value
- Removed URL links starting with http and www as this can be ignored for sentiment analysis
- Negation words were split into two parts, and the ’t’ after the apostrophe vanished after cleaning
- Converted all text to lower-case
- Removed numbers and special characters
- Removed stop words
- Tokenizing and converting into stem words using nltk Porterstemmer.
- For Dow values, some of the weekend values were missing so we adopted the forward filling method to fill those values for further analysis.

Word Cloud

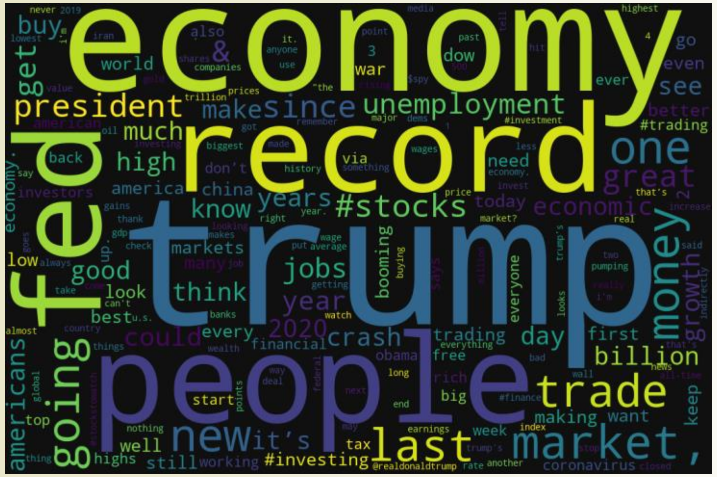


Fig1: January

Words like ‘money’, ‘high’, ‘billion’, ‘trade’, ‘booming’ are visible which means market was positive

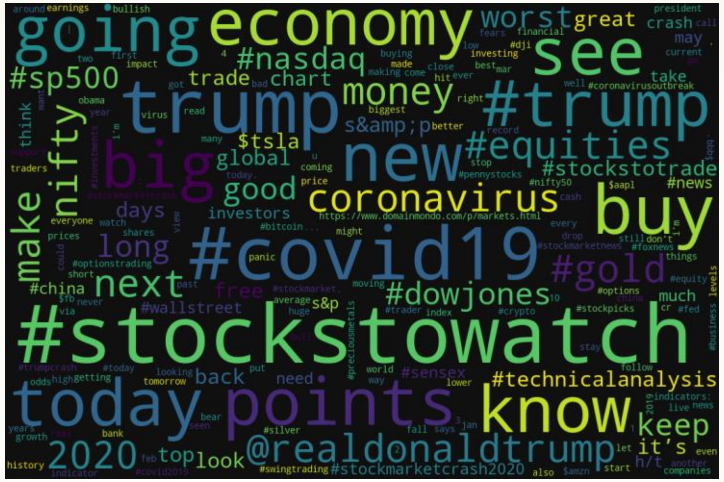


Fig 2: February

Words like ‘covid19’, ‘stocks’, ‘towatch’, ‘Trump’ , ‘points’ are visible which indicates the start of pandemic

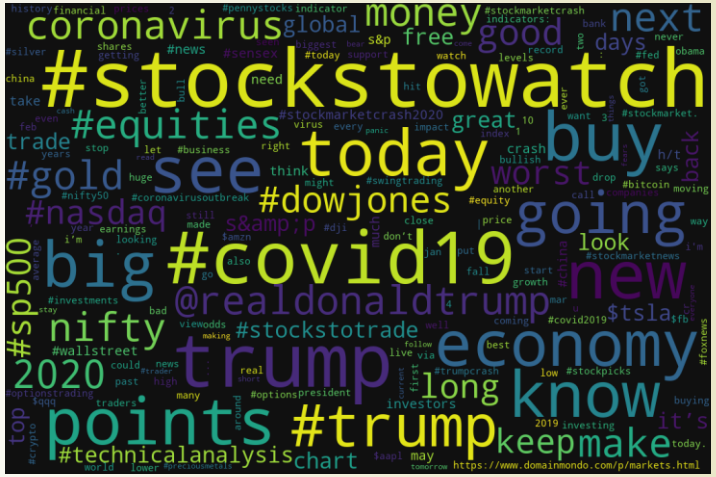
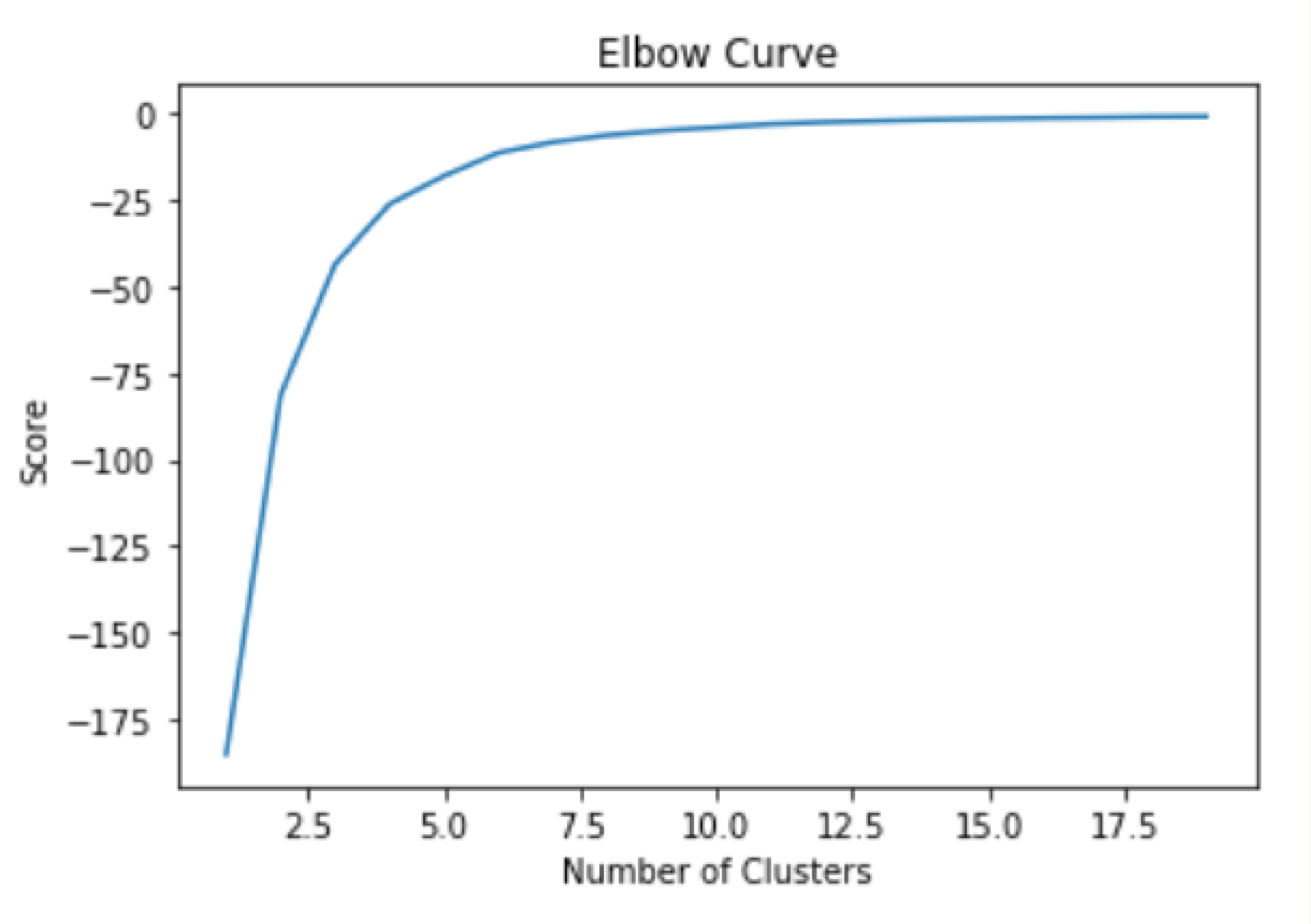


Fig 3: March

Words like ‘covid19’, ‘stocks’, ‘towatch’, ‘Trump’, ‘points’, ‘economy’, ’worst’, ‘crash’ indicates stock market was affected by pandemic

Unsupervised Sentiment Analysis Using K mean cluster

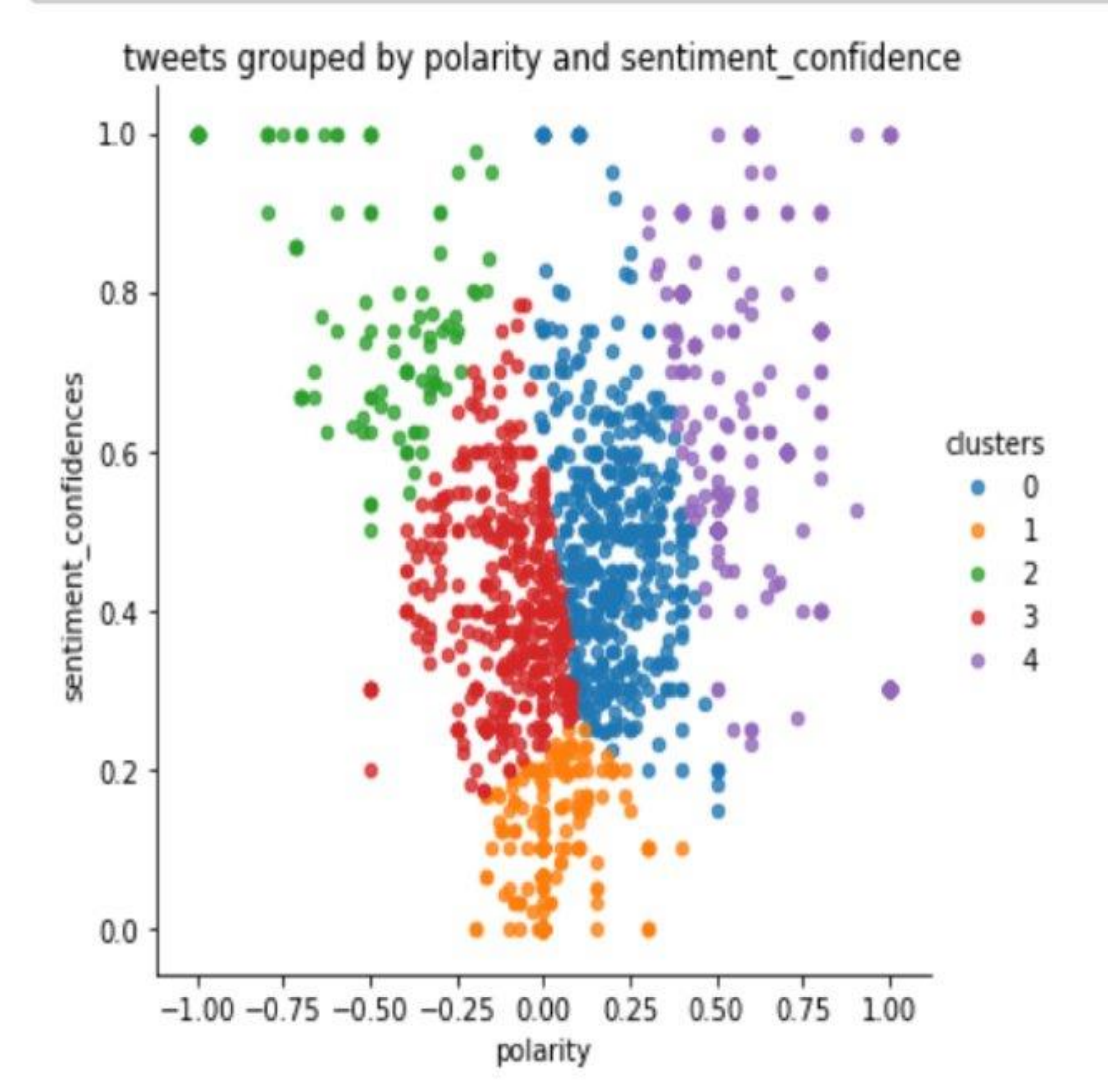
- Unsupervised sentiment analysis, where Elbow curve was used to decide number of clusters in the dataset.
- Used K-means clustering for evaluation about cluster formation
- From this curve, we found that the number of clusters was five.



ELBOW CURVE REPRESENTING THE NUMBER OF CLUSTERS AND SCORE

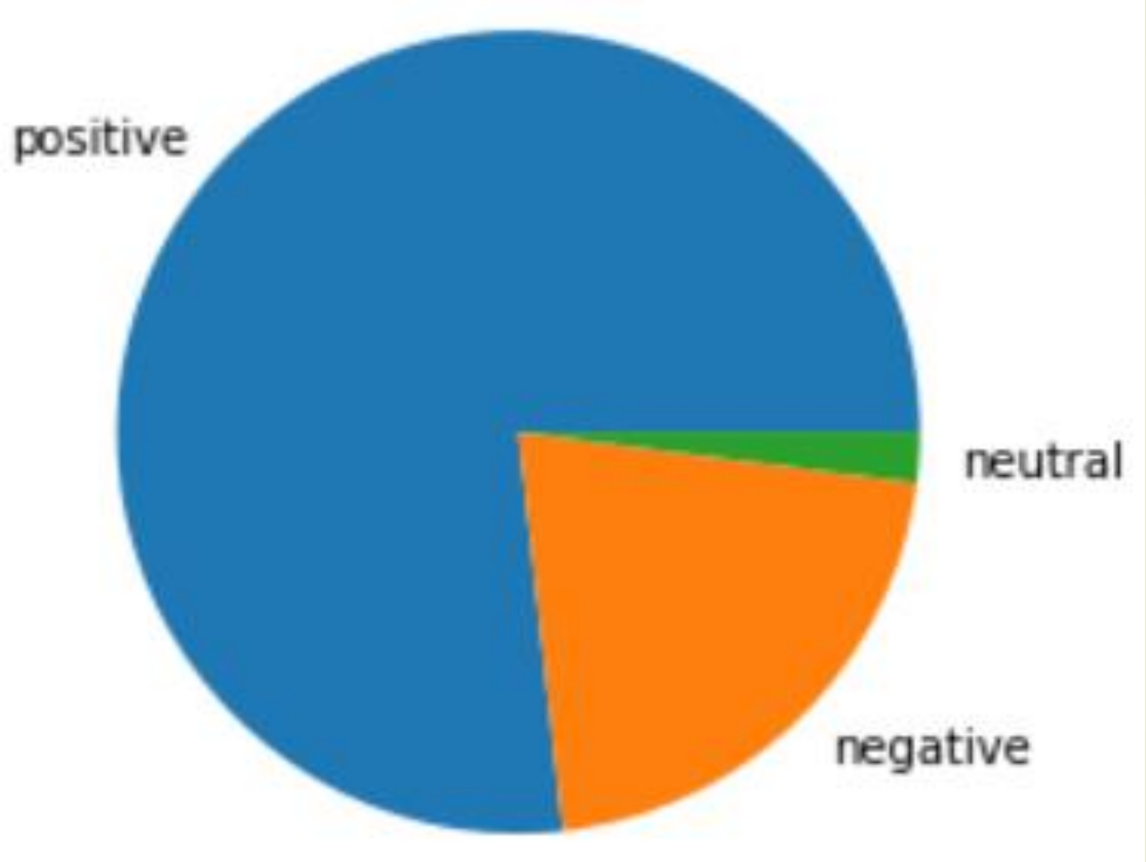
Clusters Based On Polarity And Sentiment Confidence

- Scatter plot for 5 clusters based on polarity and sentiment confidence.
- Dense clusters are observed between -0.25 to +0.25 polarity.
- Clusters beyond the -0.50 polarity with green color represent the negative sentiment.
- Clusters beyond the +0.50 polarity with purple color represent the positive sentiment.



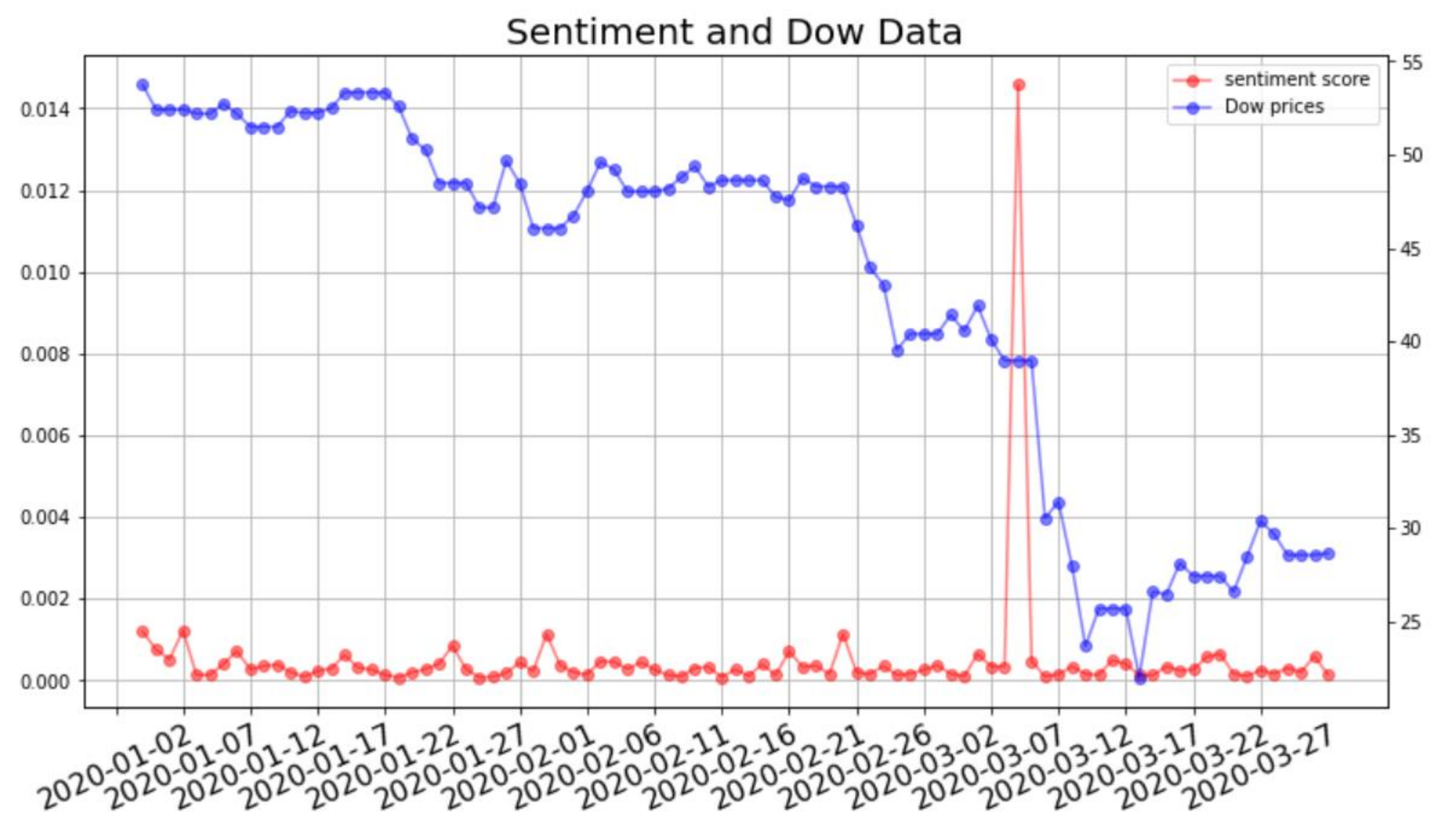
Sentiment Analysis Using VADER

- Sentiment Score were calculated using VADER sentiment analysis
- Positive sentiment was found to be 71 %, the negative sentiment to be 21% and the neutral sentiment was 2 %.
- We observed that positive sentiment were more because for first two months stock market was on rise and then started to decrease because of COVID-19.



Linear Regression model using Loughran Negative sentiment with Tf-idf

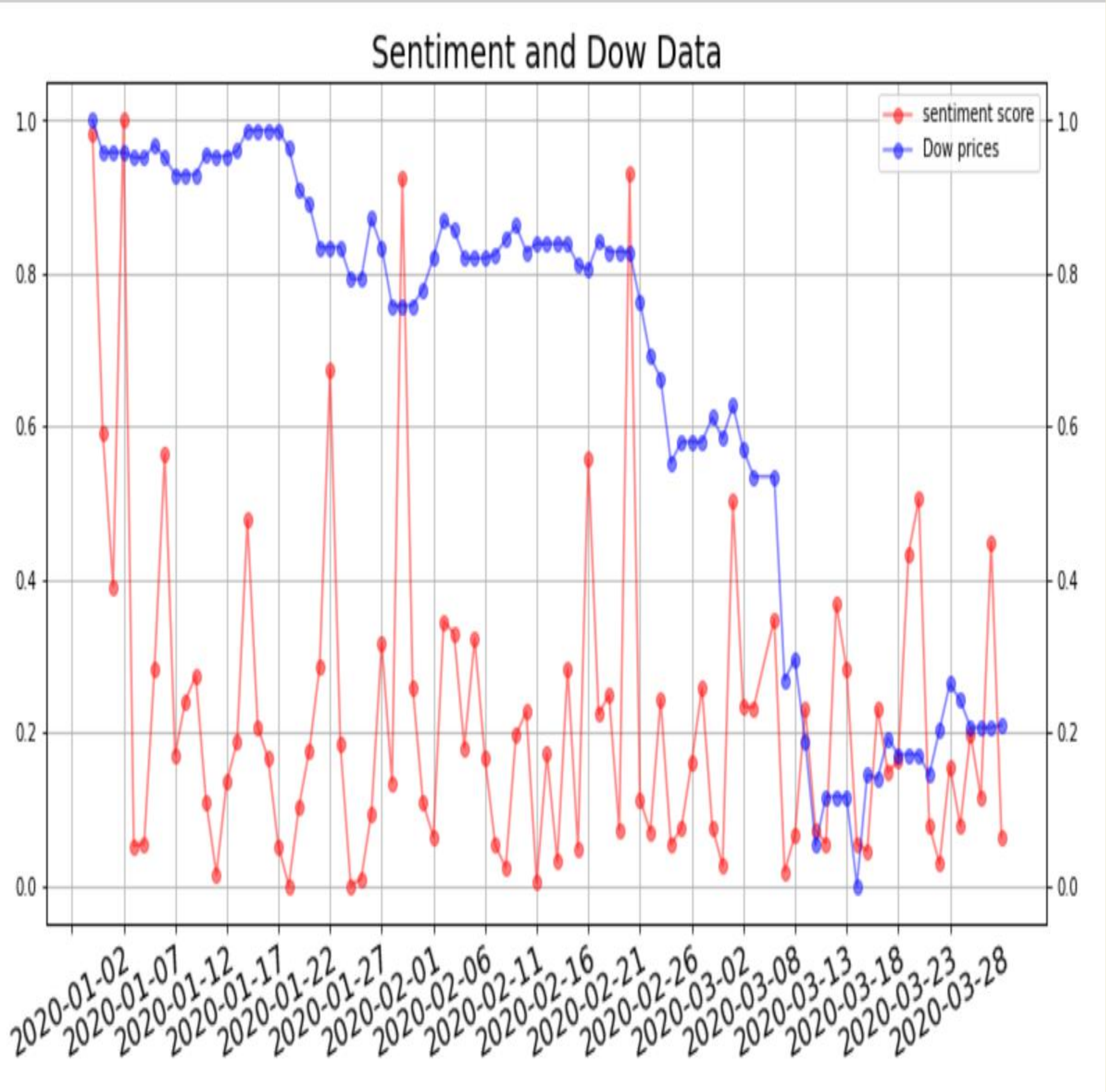
- In the plot the dow data and the normalized negative score is plotted with date on x-axis.
- It is clearly visible that there is a high negative score between the date range March 2nd 2020 and March 7th 2020 One of the sentiments with a high score.
- Stock price had decline between March 7th 2020 to March 12th 2020.The First official death due to Covid-19 was declared



LINE GRAPH REPRESENTING THE COMPARISON BETWEEN THE SENTIMENT SCORE AND DOW VALUES

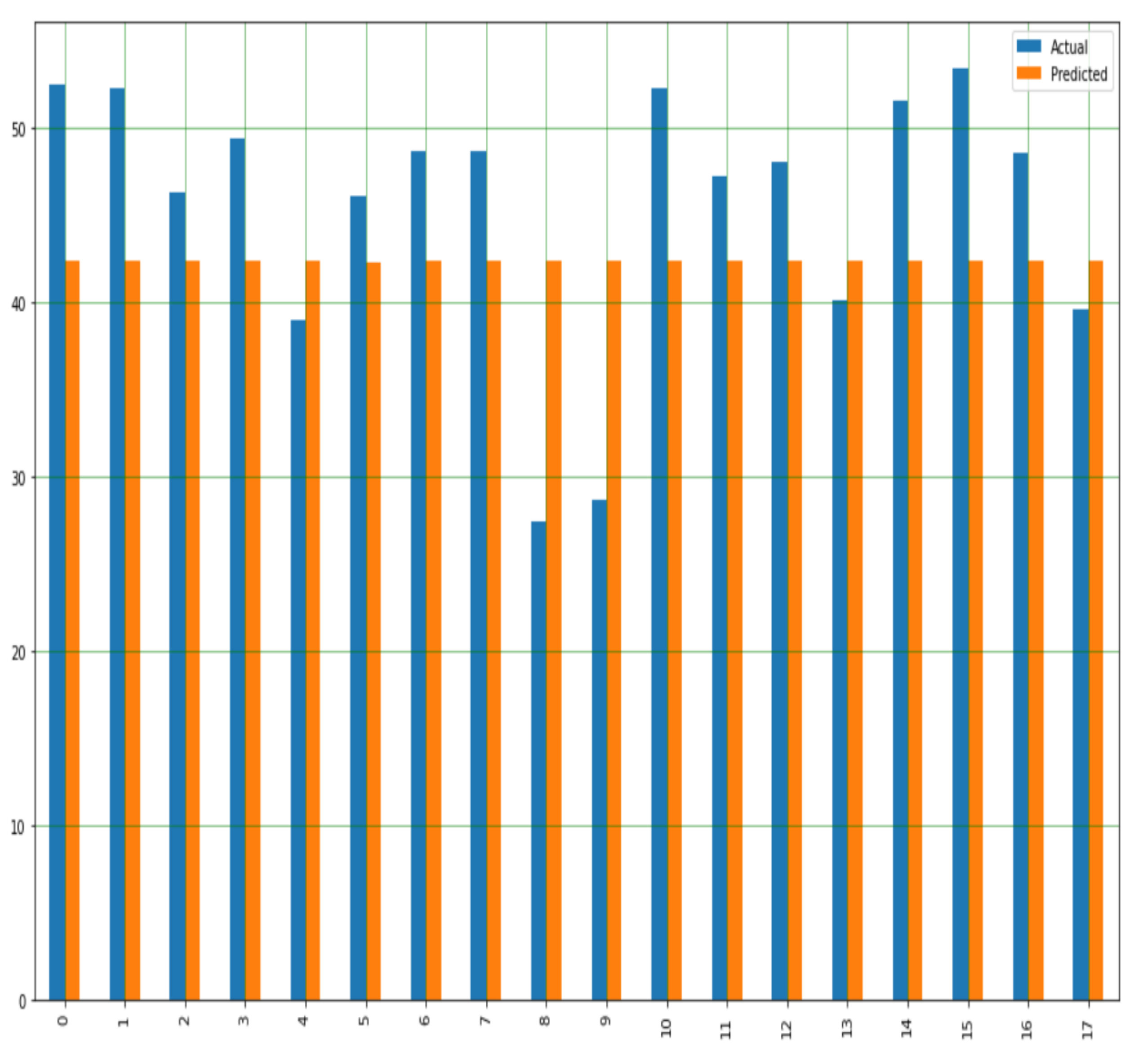
Line Graph Representing The Comparison Between The Sentiment Score And Dow Values Without An Outlier

- The Dow values and the normalized negative score is plotted with the date on X axis.
- It can be observed that when the negative score is high, there is a decline trend in DOW value not on the same day rather it can be visible in the next few days.
- We have observed that the sentiment normally has influence not on the market on the same day rather following day.



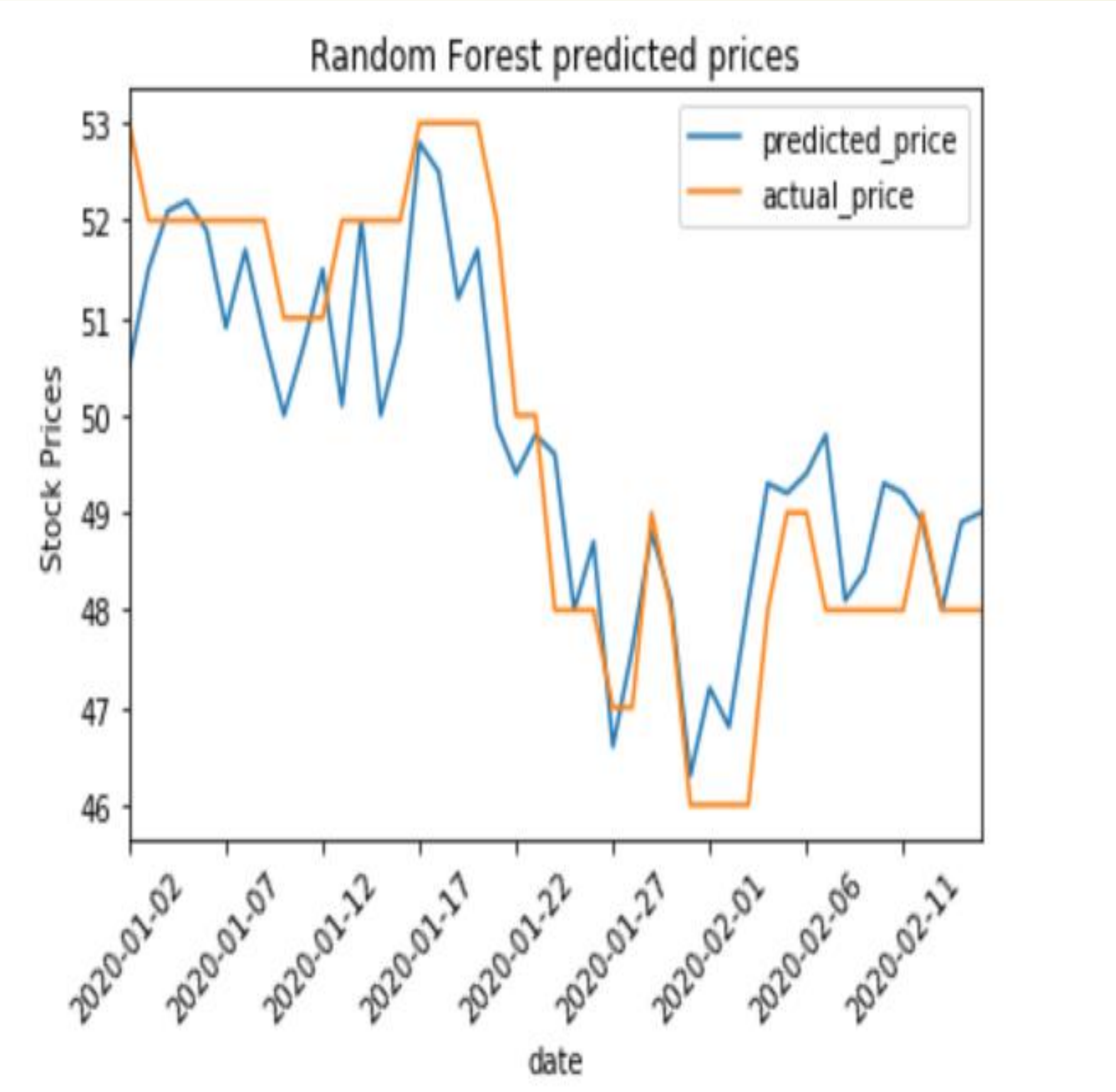
Prediction Based On Linear Regression Model

- Data set of 88 that contains the negative sentiment score and Dow Price records are randomly split into 80:20 train and Test Data.
- The figure shows the actual and predicted Dow prices for 17 days.
- For Prediction of Test data the MAE is around 7.27.



Prediction Based On Random Forest Regression Model using vader sentiment

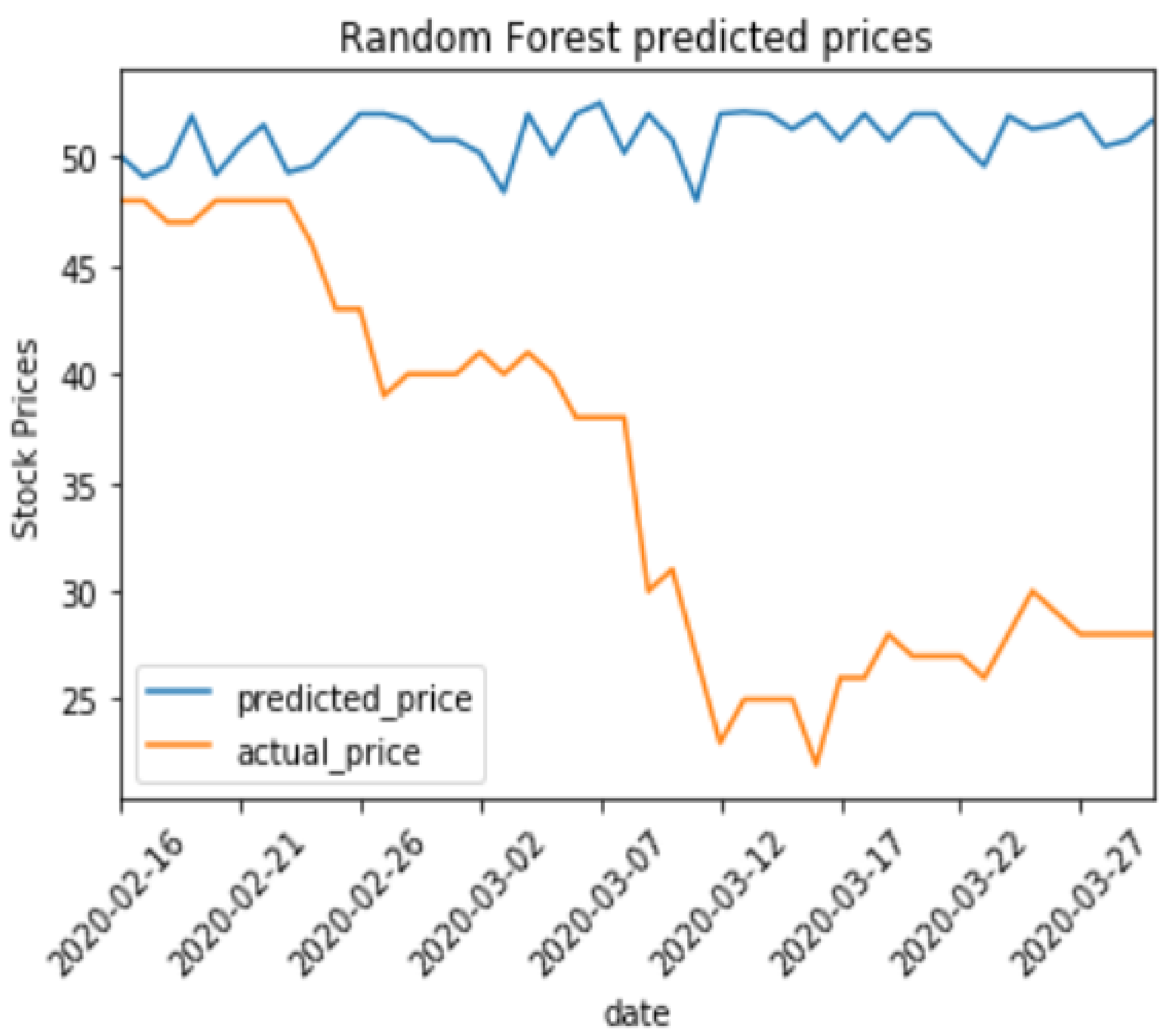
- For the training set we considered the sentiment values and stock prices from the start of January to mid-February
- We observed that the predicted prices were close to the actual price in our training set.
- According to the data, the stock market prices were less fluctuating and were at the higher side and the same prediction was made by our sentiment analysis.



TRAINING DATA SET

Prediction Based On Random Forest Regression Model

- For the testing set we considered the sentiment values and stock prices from mid-February to the end of March.
- We have seen that the predicted values were less fluctuating whereas the actual stock price is dropping rapidly.
- So, as per the training set provided, the stock market prices should not show high fluctuation and it should be at the high scale as according to January market prices.
- Due to the pandemic, there was a crash in the market at the end of February. So, from this graph it is clearly visible that there is a very huge difference between predicted and actual market prices.
- This may be because of the coronavirus outbreak all over the globe.
- The MAE for test data is around 15.53



TESTING DATA SET

Comparison Of Random Forest Regression And Linear Regression Model

- Mean Absolute error for Linear Regression model is 7.27
- Mean Absolute error for Random Forest Regression for testing data is 15.53 and for training data mean absolute error is 0.9.
- MAE for testing data of Random forest Regression is comparatively high which means average magnitude of the errors in a set of predictions is high.
- Hence, Linear Regression model is better model compared to Random Forest Regression.
- Therefore, Loughran with Tf-idf sentiment analysis found to be better for our stock market dataset for analysis.

Random Forest Regression	Linear Regression

MAE-Testing=15.53
MAE-Training= 0.9
	MAE=7.27

Accuracy:
Testing = 46.54
Training=98.17
	Mean Square Error= 66.02
	Root Mean Square Error = 8.12

Future Aspects

- More general tweets can be collected which is not specific to any tag to obtain a higher correlation between sentiment and stock market.
- The sentiment and Dow data can be treated as timeseries data and further analysis can be done.
- Correlation can be find between Investment decisions and general public sentiment.

References:

- Bollen, J., Mao, H., and Zing, X.J., “Twitter mood predicts the stock market predictor”.,*IEEE Computer,* 44(10), 2010, 91–94.
- Mittal, A. and Goel,A., “Stock prediction using twitter sentiment analysis”. 2012.
- Pak, A. and Paroubek, P., “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, *European Language Resources Association* (ELRA), 2010.
- Ranco, et.al.,“The Effects of Twitter Sentiment on Stock Price Returns”, *PLoS ONE,* 10(9), 2015.
- Sarkar, T.,“Clustering metrics better than the elbow method”, Towards data science, September 6, 2019, https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6

THANK YOU