CS418-PROGRESS REPORT

Project Name : How Twitter drives your Portfolio

Project Team Name : The Insightful

Project Team Members :

Sri Sai Kiran Reddy Gorla

Venkata Pratheek Reddy Chavva

Richa Reddy Edulakanti

Project Introduction

Social media has been a huge influence on day to day activities in most people's lives. We want to analyze its effect on financial instruments. We have chosen the stock market and crypto market for our analysis. Our goal is to find the relation between ”Public Sentiment” and ”Market Sentiment” of Stock market and Crypto market using twitter and predict the price changes. To extract the data, we planned to use Twitter API, Yahoo finance API and Binance API to extract twitter data, stock market and crypto market information. Our hypothesis is that average Tweet sentiment is correlated with the change in stock and crypto prices

Any changes

According to the plan, we couldn’t extract the required twitter data using the API. API has access to retrieve a week’s data. This doesn’t suffice our requirement. So, we will be using a data set from kaggle which has tweets related to 5 stocks during 2015-2019. Thus, our scope changes to only these 5 stocks. Our hypothesis has changed to average Tweet sentiment is correlated with the change in stock and prices.

Data cleaning

We have two different types of data sets, one for stocks and the other for tweets.There isn't much to clean up with regard to the stock data. However, when it comes to the tweet data, we started by eliminating the writer column because it is useless and contains a lot of null values. We then eliminated all the null values from the other columns as well.Automated tweets were eliminated, followed by the @user Twitter handles, punctuation, special characters.Filtering ,tokenization and removal of stop words is also done.Finally, we retrieved every tweet based on its ticker symbol.

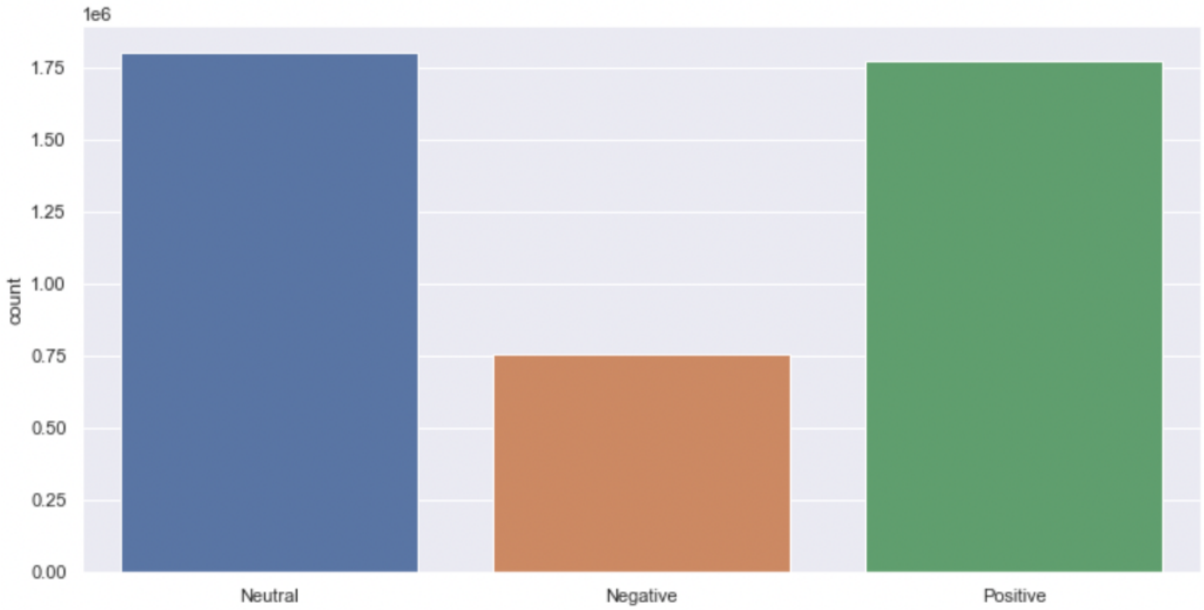
Sentimental Analysis

Sentiment analysis is contextual mining of text which identifies and extracts subjective information. To classify the sentiments of the given tweets to determine which tweets are showing a negative, positive or neutral sentiment, we used VADER-Sentiment-Analysis. A sentiment lexicon, which is a collection of lexical features (such as words) that are often classified as either positive or negative depending on their semantic orientation, is used in conjunction with VADER. VADER not only informs us of the positivity and negativity scores, but also of the sentimentality of each score.

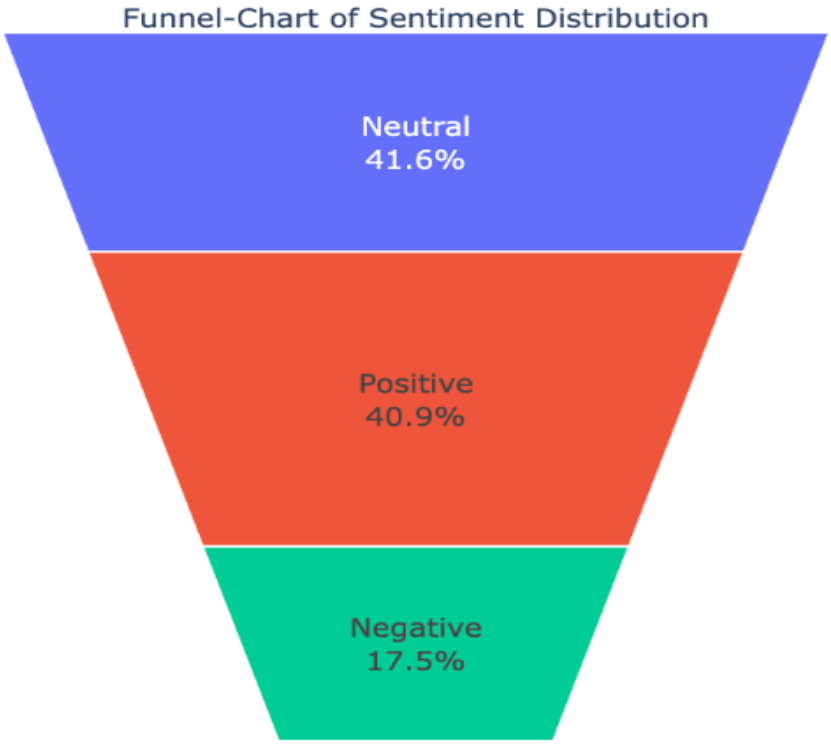
Exploratory Data analysis

We started with the funnel-chart of sentiment distribution and then plotted the bar graph of common words in the tweets.We analyzed the stock data by plot between timestamp and closing price.Then the historical view of the closing price and total volume of stock being traded each day.Lastly a plot between total volume of stock and days high of the stock.

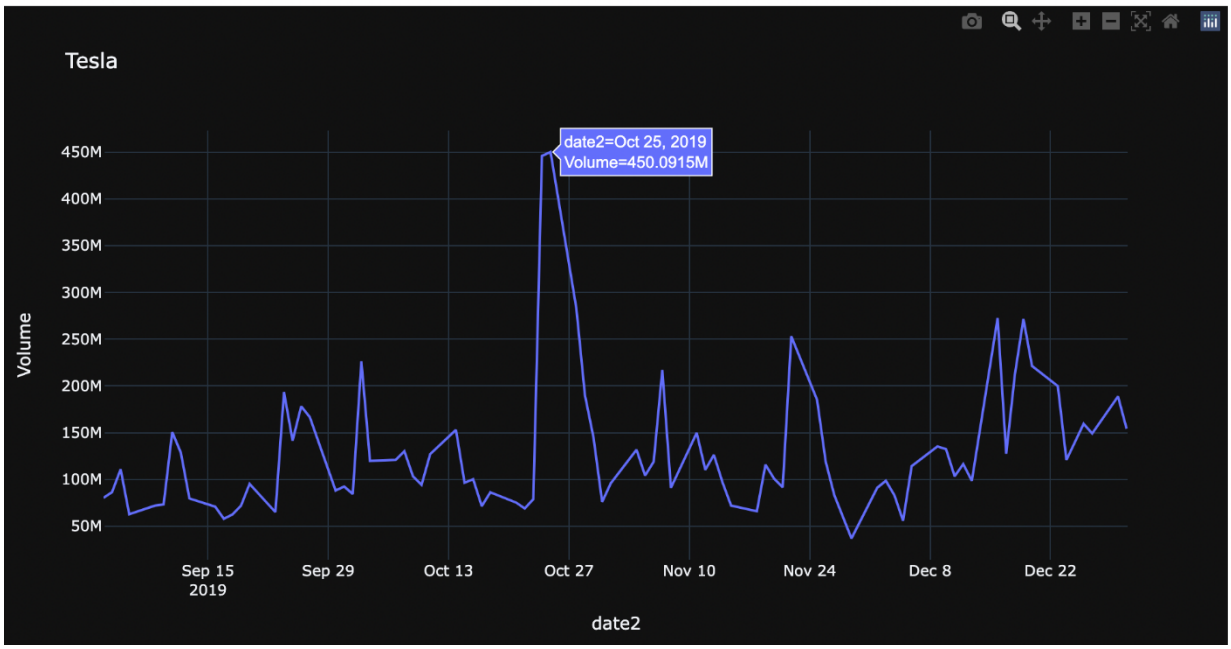
Bar Graph of tweets based on their sentiment.We can observe that the count of neutral tweets is high.



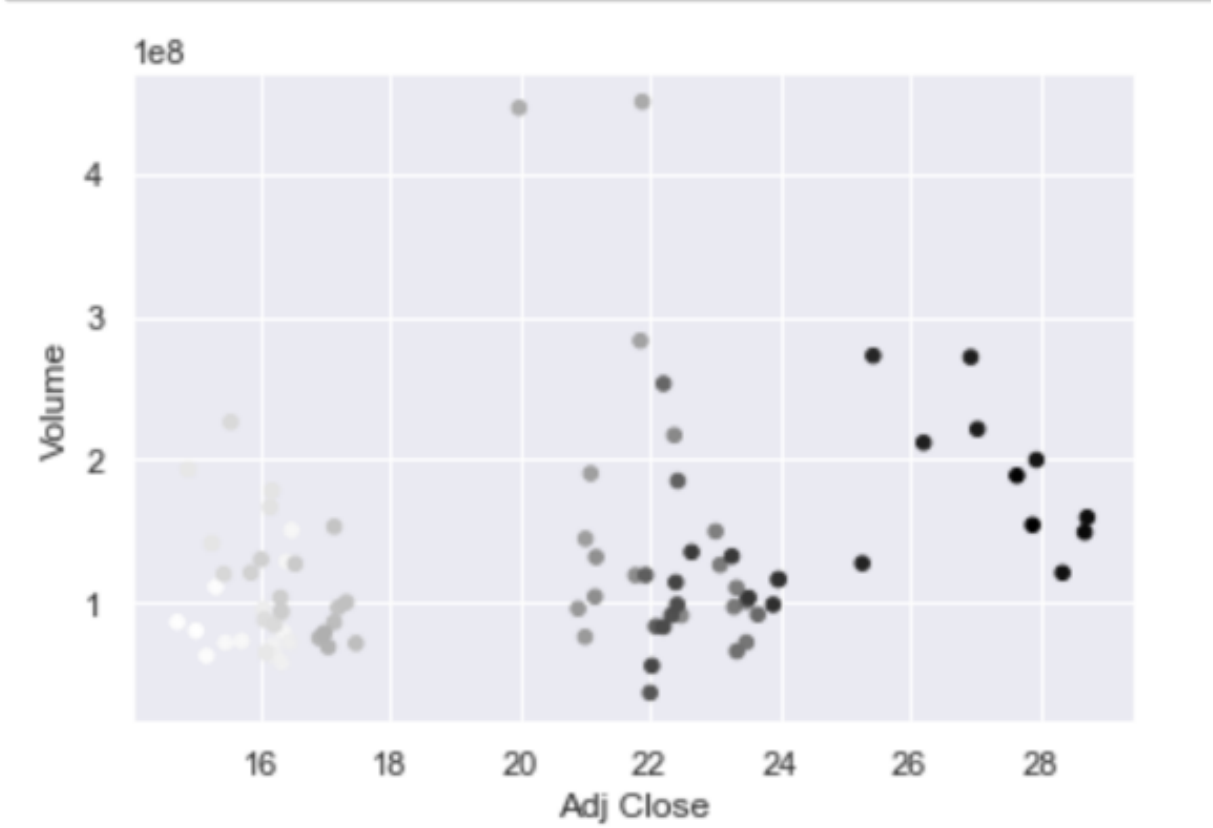
Funnel chart of Sentimental Distribution on tweets,using a funnel chart we can actually see the percentage of each tweet based on their sentiment.



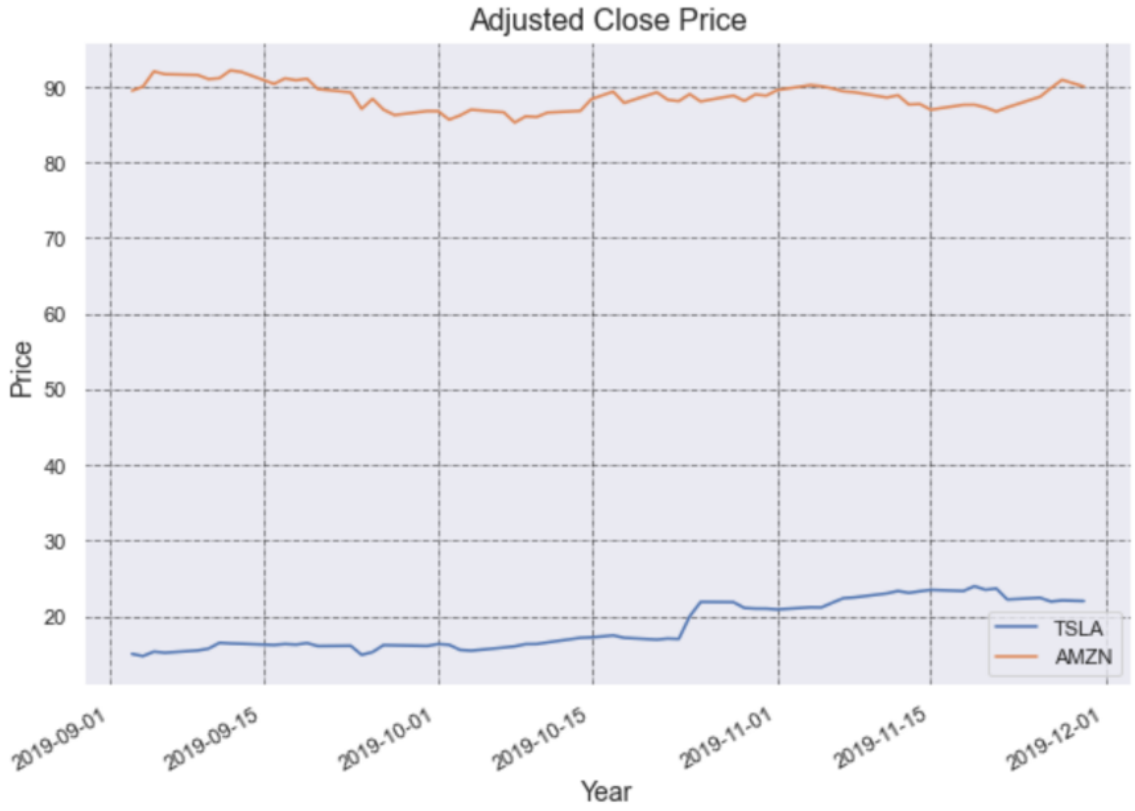
Plot between volume of stock traded with respect to the date.



Scatter plot of volume and Adjacent close price of tesla stock.

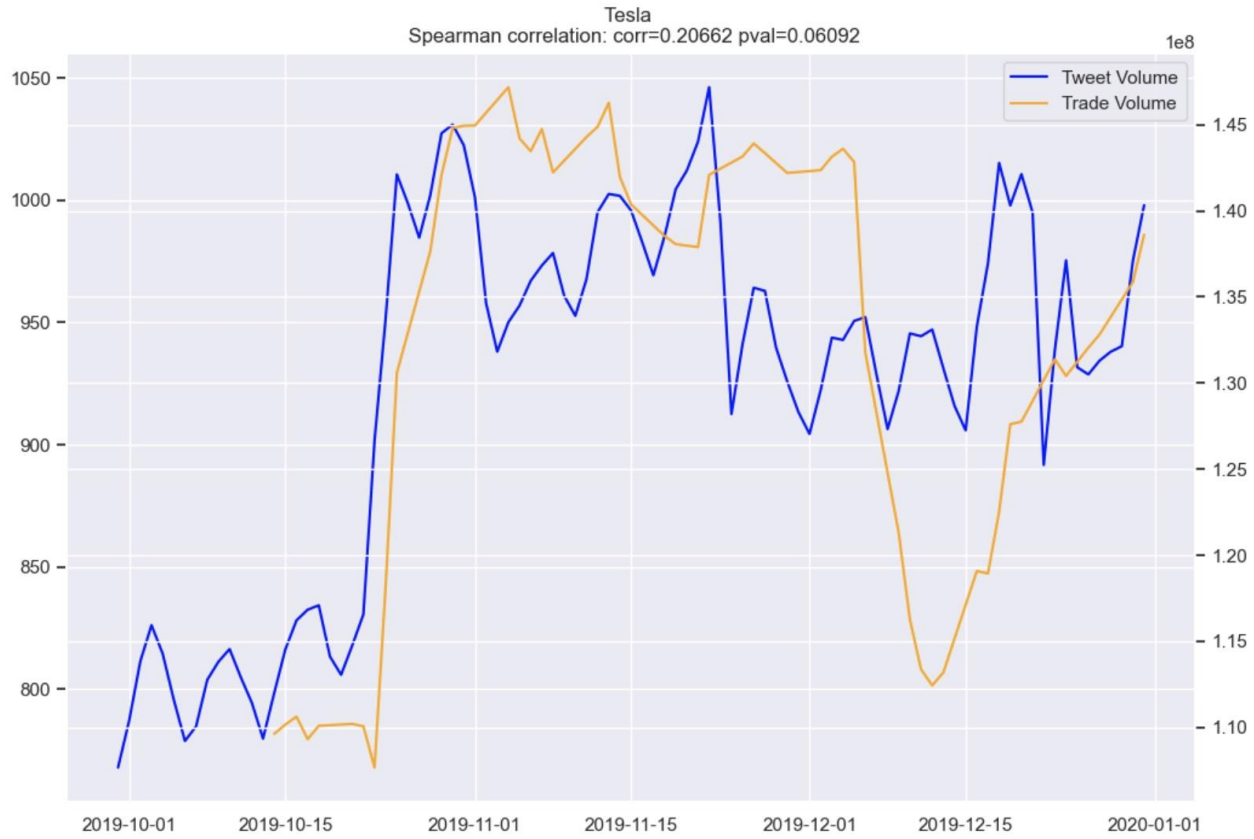


Data plot between Close price and date for both amazon and tesla.



At least one visualization

Spearman correlation between tweet volume and trade volume of amazon and tesla. Measures the strength and direction of association between two ranked variables. Comparing both the plot we can say that tesla has high correlation value when compared to amazon.

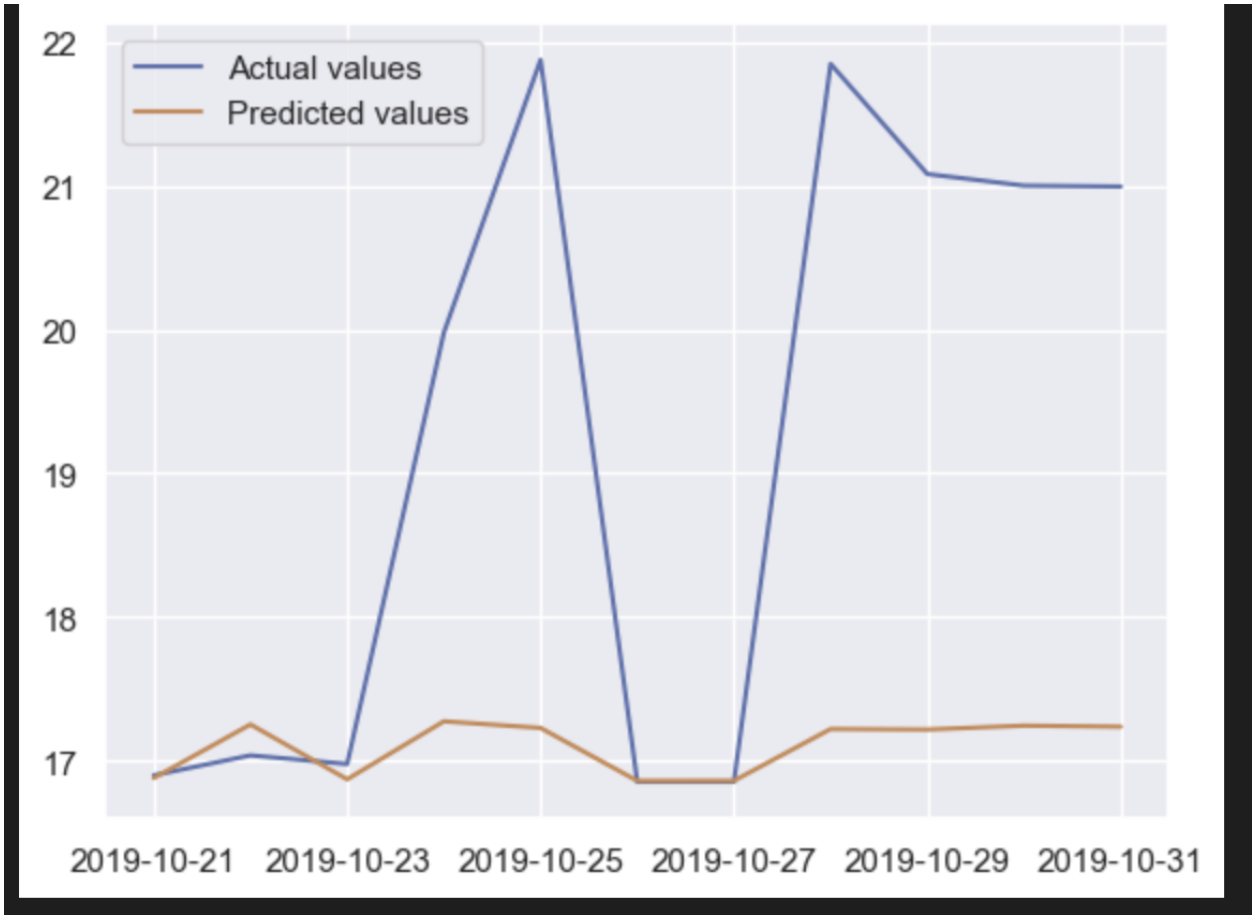



Amazon
Spearman correlation: corr=0.17048 pval=0.12335
1e7
Tweet Volume
Trade Volume
400
380
360
340
320
300
280
6.4
6.2
6.0
5.8
5.6
5.4
2019-10-01
2019-10-15
2019-11-01
2019-11-15
2019-12-01
2019-12-15
2020-01-01

At least one ML analysis

The prediction is made using the random forest method. Given that it is one of the most adaptable machine learning algorithms, it provides good prediction accuracy. Typically, this is applied to classification tasks.The core method of the random class classifier is to demand the decision aggregate of random subset decision trees and supply a final class result supported by the votes of the random subset of decision trees.We use a random forest classifier, which has the same hyperparameters as a decision tree, to predict the stock market. It essentially builds a group of decision trees that yields some result. Predicting is a difficult undertaking because of the stock market's low volatility.

Below is the plot for actual vs predicted values using Random Forest Regression model for two months of amazon stock data.



Reflection

- What is the hardest part of the project that you've encountered so far?
    - Our biggest challenge was to extract twitter data. We have tried several ways to do it but weren't successful. This has cut down our time working on the project.

- What are your initial insights?
    - As we are analyzing top 5 stocks in the stock market, they are immune to volatility. With very volatility, it is difficult to conclude any correlation between tweet sentiment and stock prices.
    - Also, it will be challenging to analyze the huge data set of over 4 million tweets considering our constraint of low computing power of our laptops.

- Are there any concrete results you can show at this point? If not, why not?
    - Initially, we started analyzing two years worth of data, but we couldn't achieve any results on our laptops. So, we reduced our dataset to one month. When we used Random forest regression model to predict the amazon stock price, we attained an accuracy of 89.81%. This is expected considering the low volatility. But surprisingly, the R-squared value and test score of the model are negative. This is due to extreme over-fitting due to an extremely small amount of data.

- Going forward, what are the current biggest problems you're facing.Do you think you are on track with your project? If not, what parts do you need to dedicate more time to?
    - Predicting is a difficult undertaking because of the stock market's low volatility. Our plan is to increase the dataset range to a year and filter the tweets based on bot tweets, tweet engagement. We also need to dedicate more time on EDA, explore if we can pick any intervals of high volatility of stock prices and see if twitter sentiments have any impact on it.

- Given your initial exploration of the data, is it worth proceeding with your project, why? If not, how are you going to change your project and why do you think it's better than your current results?
    - Yes, At present the accuracy of our project is very high for a short interval. So if we proceed with our project to the next level there are high chances to get better results even for a long interval of time.