Capstone

Sentiment Analysis of Stocks from Twitter and News

Nikhil Valse

Department of Business, Long Island University

Gurpreet Singh

25th April 2023

Contents:

Background:

The financial market is highly unpredictable, and the stock prices of companies fluctuate rapidly based on various factors such as economic conditions, company performance, and market sentiment. The sentiment of the market, which is the collective opinion of investors and traders, is a crucial factor in determining the direction of the stock prices. With the rise of social media platforms and online news, sentiment analysis has become a popular tool in predicting the stock market's direction.

This project aims to perform sentiment analysis for stocks using news and Twitter data. The objective is to analyze the sentiment of news articles and tweets related to a particular company or industry and predict the potential impact on the stock prices. The project will use machine learning techniques such as Natural Language Processing (NLP) and Machine Learning based sentiment analysis algorithms to analyze the textual data.

The project will use two sources of data for sentiment analysis: news articles and tweets. News articles will be collected from various online news sources, and tweets will be collected using the Twitter API. The data will be pre-processed as required.

The sentiment analysis will be performed using machine learning algorithms such as Roberta twitter sentiment analyser. The trained models will be used to predict the sentiment of new news articles and tweets related to the company or industry.

Objective:

The project's objective will be to study the effect of the sentiment of news articles and tweets related to a particular company or industry and their potential impact on the stock prices. This can be used to create a tool in future to be used by investors and traders to make informed decisions in the stock market.

It also demonstrates how web scraping can be used to mine the important data related to our purpose and how it can be used to our advantage to improve upon the decision making using various analytical tools in hand.

Disclaimer:

The analysis performed in this Project is intended to increase your information about market trends to make decision depending on lot of other factors.

The information provided through Twitter stock sentiment analysis is for informational purposes only and should not be considered as investment advice or recommendations to buy, sell or hold stocks. The analysis is based on public sentiment expressed on Twitter, which may not reflect the opinions of all investors and may not be a reliable indicator of market trends. Investors should always conduct their own research and analysis before making any investment decisions. The accuracy and reliability of sentiment analysis can be influenced by various factors, including but not limited to the quality and quantity of data, the algorithm used for analysis, and the language and tone of tweets. Therefore, investors should exercise caution and carefully evaluate the results of sentiment analysis before making investment decisions.

Data Extraction:

We are using requests and beautiful soup libraries to scrape the news related to specific tickers on the Finviz website for last four days. For the tweets related to the stocks, we are trying to use Twitter API, but since it has very limited abilities for fetching the tweets for our purpose, we have taken already mined tweets related to top NASDAQ companies of years 2021 and 2022.

The Tweets data we have acquired from the following source:

https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction

Apart from that we are also fetching the data from Yfinance for time series data about the stock to analyze the effect of sentiment on stock,

Data Variables:

Finviz:

Ticker: Name of the ticker ex. AMZN, MSFT

Date: YYYY-MM-DD

Time: HH:MM: SS

Tweets:

Date: Timestamp

Tweet: Actual Tweet about ticker

Stock Name: Ticker name

Company name

Yfinance:

Datetime: Timestamp

Open: Numeric

High: Numeric

Low: Numeric

Close: Numeric

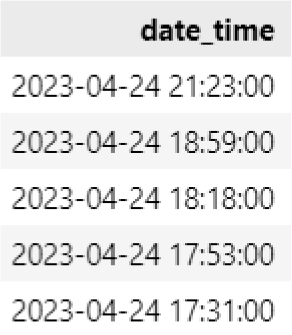
Adj Close: Numeric

Volume: Numeric

Data Exploration/Pre-processing/Visualizations/Analysis:

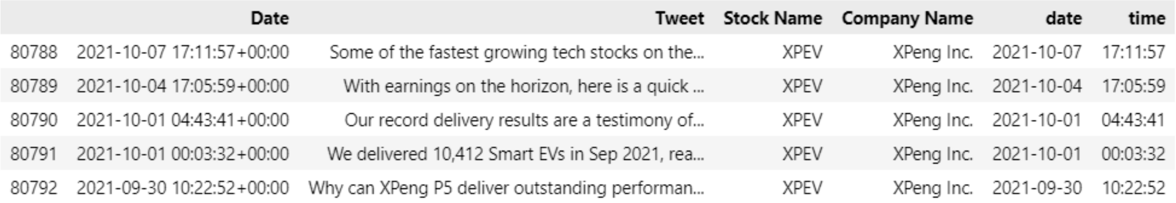
As we move through data, we can see that there are date and time separate columns, in order to plot or analyze the data we need to create a column which incorporated date and time simultaneously to avoid conflicts.

First, we convert the date and time columns into a datetime object using the pandas to_datetime() method for further use of only using the dates. Then we combine these columns as strings and convert whole timestamp string into panda's date-time format i.e., timestamp.



date_time
2023-04-24 21:23:00
2023-04-24 18:59:00
2023-04-24 18:18:00
2023-04-24 17:53:00
2023-04-24 17:31:00

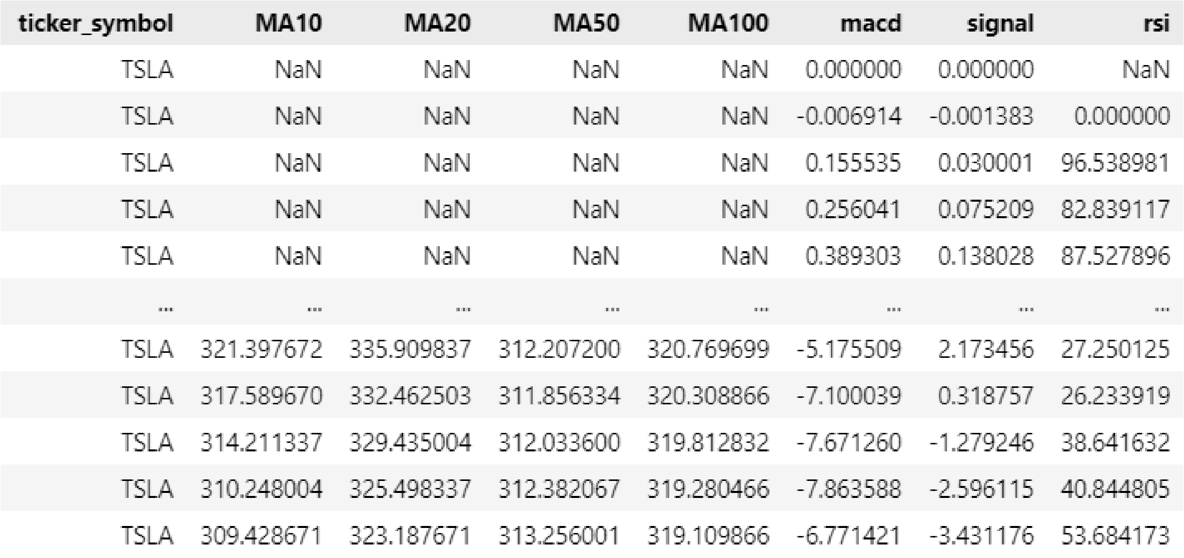
This will avoid the order mismatch when plotting the time series plots. When we go into the tweets data, we have exactly opposite scenario, we divide the date-time time-stamp into separate date and time columns for comparing just the dates while filtering.



	Date	Tweet	Stock Name	Company Name	date	time
80788	2021-10-07 17:11:57+00:00	Some of the fastest growing tech stocks on the...	XPEV	XPeng Inc.	2021-10-07	17:11:57
80789	2021-10-04 17:05:59+00:00	With earnings on the horizon, here is a quick ...	XPEV	XPeng Inc.	2021-10-04	17:05:59
80790	2021-10-01 04:43:41+00:00	Our record delivery results are a testimony of...	XPEV	XPeng Inc.	2021-10-01	04:43:41
80791	2021-10-01 00:03:32+00:00	We delivered 10,412 Smart EVs in Sep 2021, rea...	XPEV	XPeng Inc.	2021-10-01	00:03:32
80792	2021-09-30 10:22:52+00:00	Why can XPeng P5 deliver outstanding performan...	XPEV	XPeng Inc.	2021-09-30	10:22:52

As we go further to analyze the specific stock, we have chosen TESLA ‘TSLA’ since there is large number of tweets (37000) available for this ticker and hence, we create a data frame for Tesla’s time series data from yfinance.

Also, for checking correlation, we create different indicators for time series data of tesla.

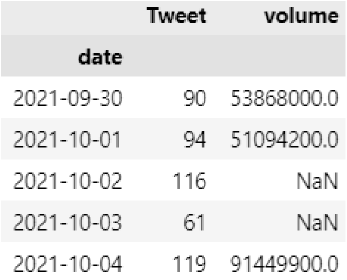


ticker_symbol	MA10	MA20	MA50	MA100	macd	signal	rsi
TSLA	NaN	NaN	NaN	NaN	0.000000	0.000000	NaN
TSLA	NaN	NaN	NaN	NaN	-0.006914	-0.001383	0.000000
TSLA	NaN	NaN	NaN	NaN	0.155535	0.030001	96.538981
TSLA	NaN	NaN	NaN	NaN	0.256041	0.075209	82.839117
TSLA	NaN	NaN	NaN	NaN	0.389303	0.138028	87.527896
...	...	...	...	...	...	...	...
TSLA	321.397672	335.909837	312.207200	320.769699	-5.175509	2.173456	27.250125
TSLA	317.589670	332.462503	311.856334	320.308866	-7.100039	0.318757	26.233919
TSLA	314.211337	329.435004	312.033600	319.812832	-7.671260	-1.279246	38.641632
TSLA	310.248004	325.498337	312.382067	319.280466	-7.863588	-2.596115	40.844805
TSLA	309.428671	323.187671	313.256001	319.109866	-6.771421	-3.431176	53.684173

We now merge the data-frame of stock’s tweets and yfinance data with outer join to make our first visualization,

We merge data on “date” column so we can get date-wise tweets and statistical data. We then group the data by date and aggregate numerical columns by either count or mean, for tweets, we aggregate by count, so that we can measure the volume of tweets on that day and trading volume by mean since the volume is same for all the rows for one day.

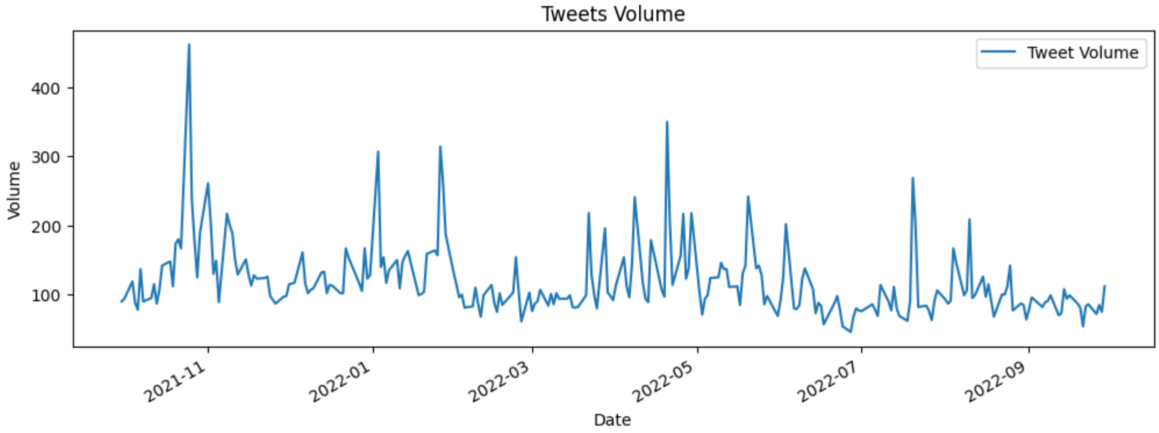
We don’t drop the duplicate tweets because they can be re-tweets and they have significance with respect to the stock.

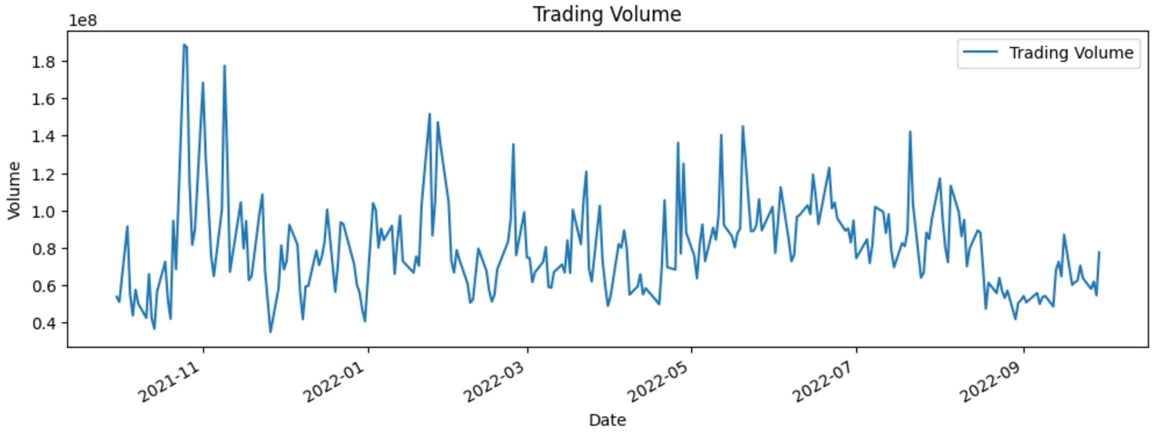


	Tweet	volume
date		
2021-09-30	90	53868000.0
2021-10-01	94	51094200.0
2021-10-02	116	NaN
2021-10-03	61	NaN
2021-10-04	119	91449900.0

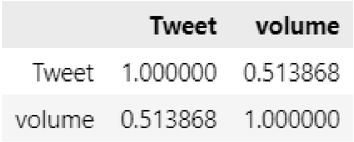
Note: NaN's are the Saturdays and Sundays, we excluded them by using dropna while visualizing.

On the basis of above grouped data frame, we now create a time series tweet volume and trading volume chart to see if the volume of tweets for that day affect the trading volume.





As we can see these graphs are quite similar in shape, so we can conclude that the number of tweets on a specific day largely affects the number of trades done on that day.



	Tweet	volume
Tweet	1.000000	0.513868
volume	0.513868	1.000000

There seems to be more than 50% correlation between those two variables.

Cleaning:

We now proceed to clean tweets data, for that we use python's in-built functions along with regex. We start by looking at the Hyperlinks which are useless and slow down our pretrained model's performance, also we take care of lot of emoji and some symbols like &amp; which means "and" to clean our tweets.

For applying sentiment analysis, I researched on internet and found the best machine learning model for checking the sentiment of a tweet. I also found that machine learning models perform much better than rule-based models especially when there are slangs and other complex phrases are present in the text.

We then import the "twitter-roberta-base-sentiment" pretrained model and Auto-tokenizer from transformers library, this model is based on BERT language model and is trained on 160 GBs of text, developed by Facebook researchers. This model automatically tokenizes the words in the text and there is no need to pre-process the text.

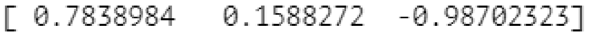
After applying this model to cleaned tweets, we can see the result in the three particular values, which are positive, neutral and negative, on which we apply the SoftMax function to get probability of each of the three possibilities.

For example,

example = "Mainstream media has done an amazing job at brainwashing people. Today at work, we were asked what companies we believe in and I said @Tesla because they make the safest cars and EVERYONE disagreed with me because they heardâ€œthey catch on fire and the batteries cost 20k to replaceâ€• "

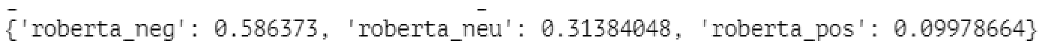
Note that '@' and emoji symbols are ignored readily since this is ML model, removing @ will infer different meaning, as if he is talking to Tesla.

Before SoftMax,



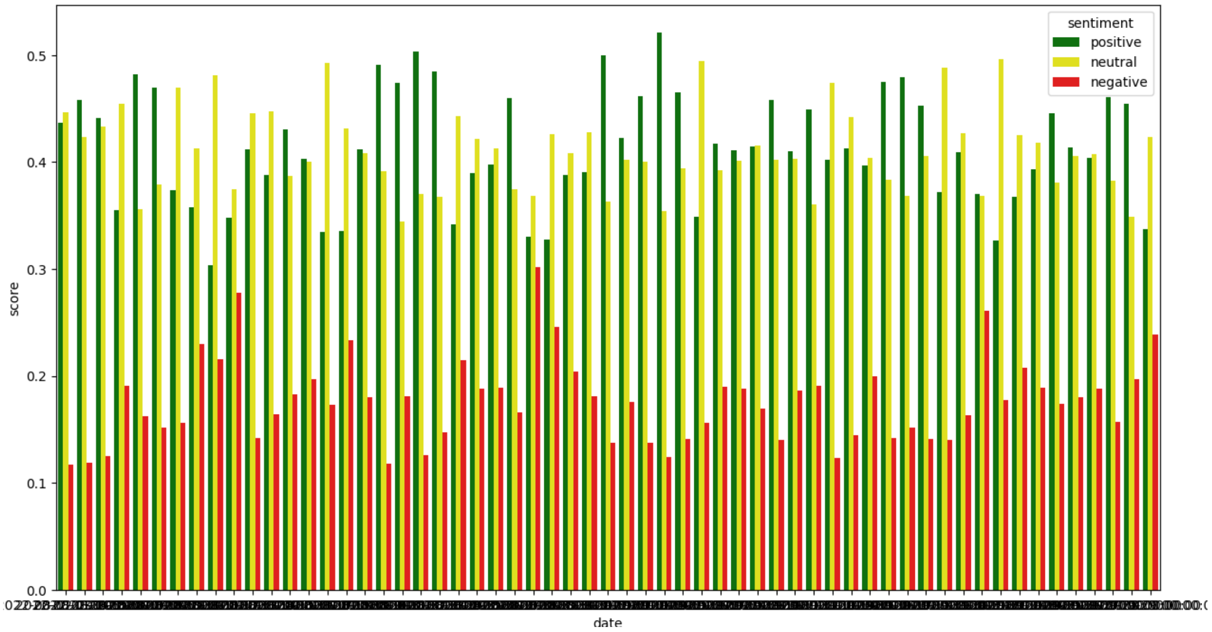
```
[ 0.7838984   0.1588272  -0.98702323]
```

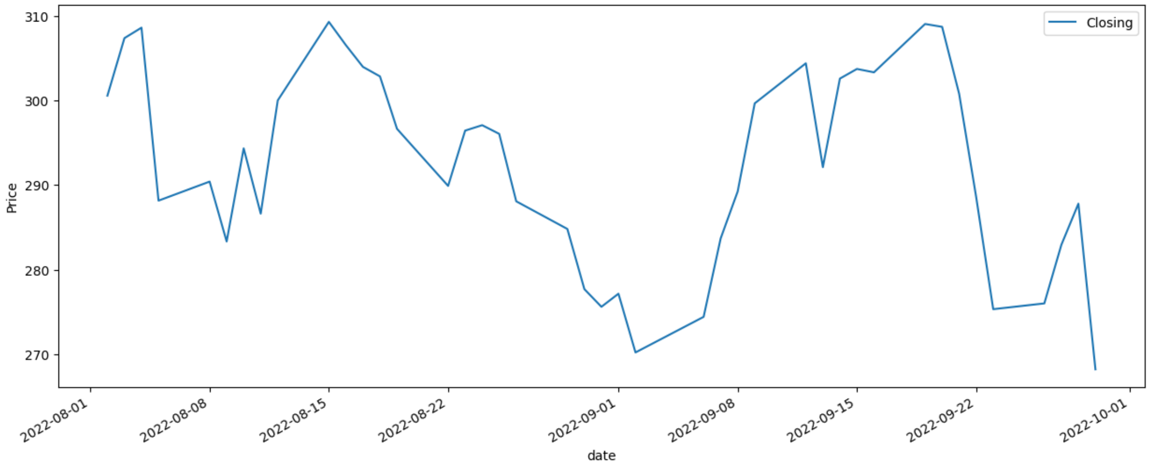
After SoftMax, we can see that roberta_neg (0.586373) is highest and hence sentence above is inclined towards negativity.



```
{'roberta_neg': 0.586373, 'roberta_neu': 0.31384048, 'roberta_pos': 0.09978664}
```

After that, we add all the scores to the data frame to visualize our selected stock Tesla’s sentiment scores for two months from 1st of August to 30th of September, we use clustered column chart to visualize the parameters of sentiment.

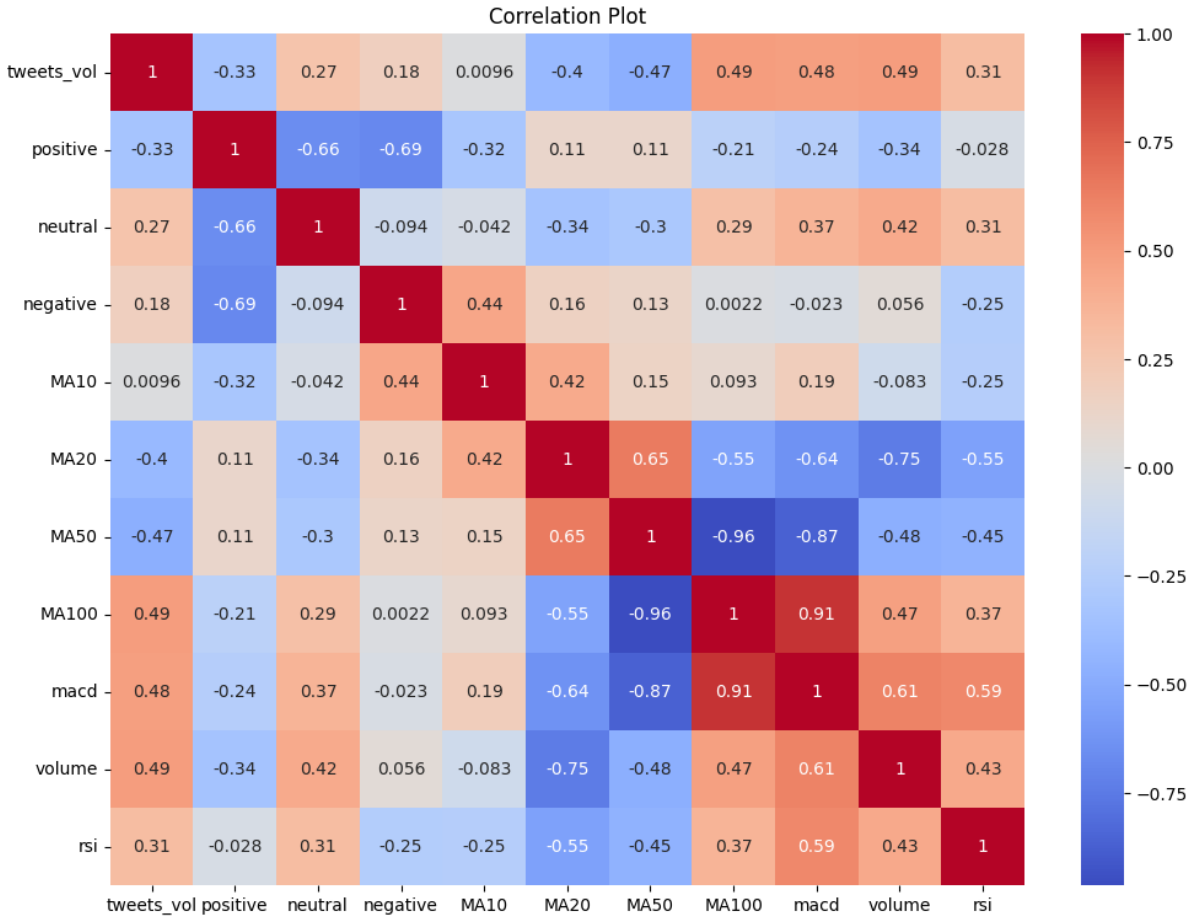




In the above two graphs, we can see that the negative sentiments are less dominant for this stock, because generally the news are little biased towards positivity, but we can see that whenever there is spike in the negative bars, stock's closing price is in down trend.

We can see on 22nd August 2022, there was little upward trend in the stock, but as we see right after that 'red marker' of sentiments went up and stock price went down with that. Same happened on 8th August and at the end of the September.

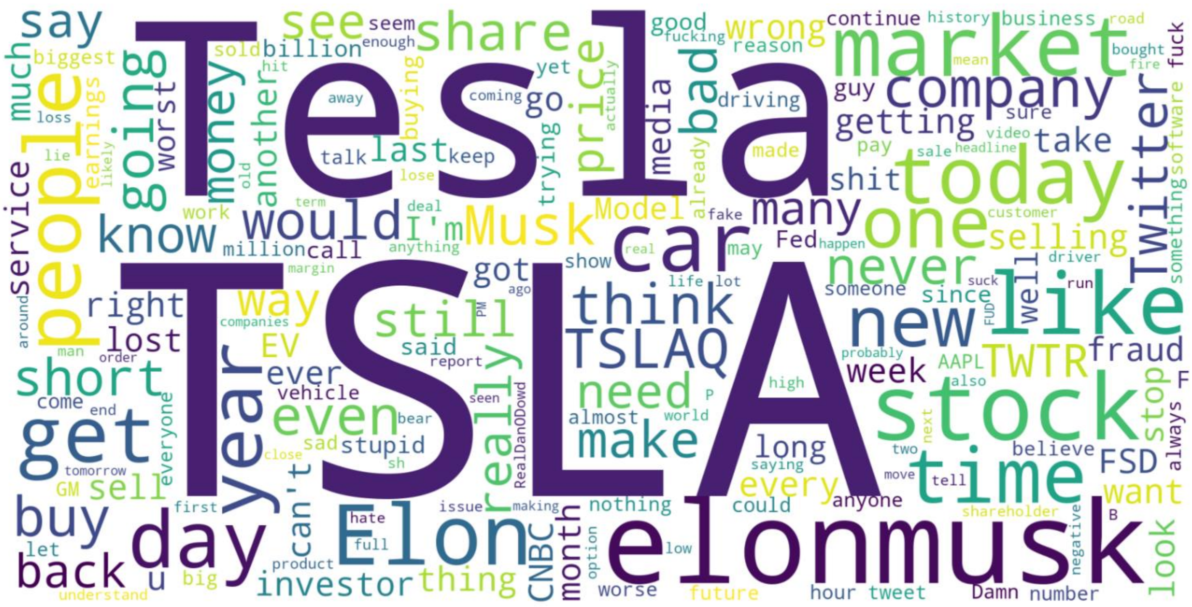
Now we make some more visualizations, such as correlation plot, for just last two months of same stock,



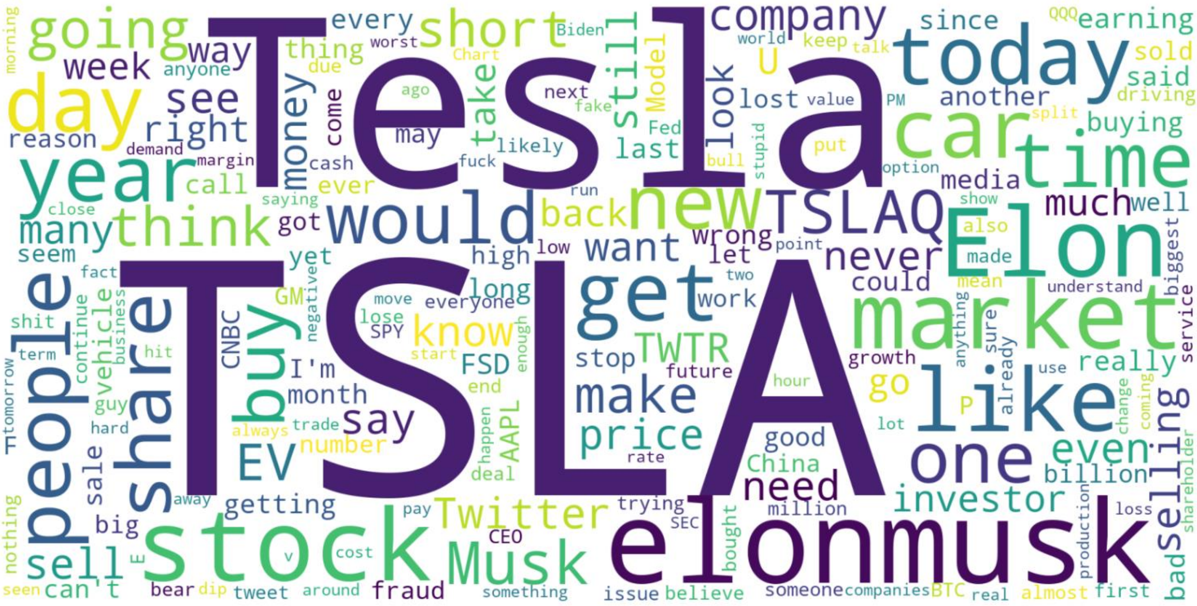
We already seen that tweet volume is largely correlated to trading volume but it is also correlated to macd indicator and moving averages of 100 days.

As we know that MACD is also used as bearish and bullish signal which fluctuates above and below the zero line, and helps traders to analyze the stock. It is possible that the tweets volume is directly proportional to MACD bullish signals.

Now we do some visualization related to the words and main terms included in the negative and positive tweets, for this, we collect different levels of positivity's and negativity's. First, we see the word cloud for negative sentiment tweets with probability more than 0.7,

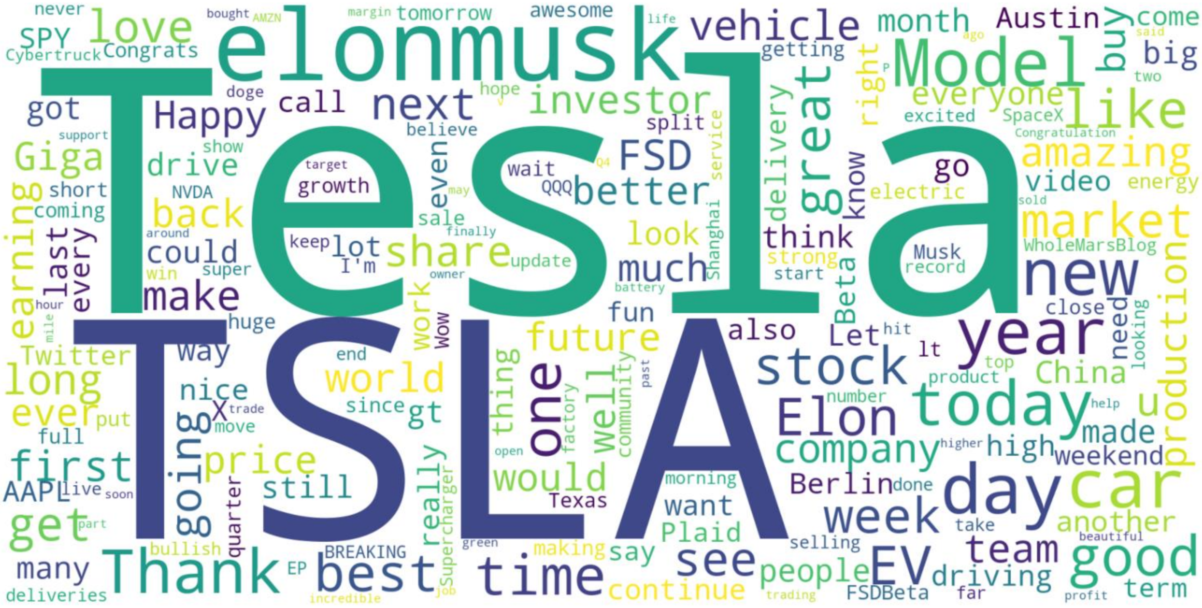


Now, probability greater than 0.5, so that more tweets are included and we may get new words.

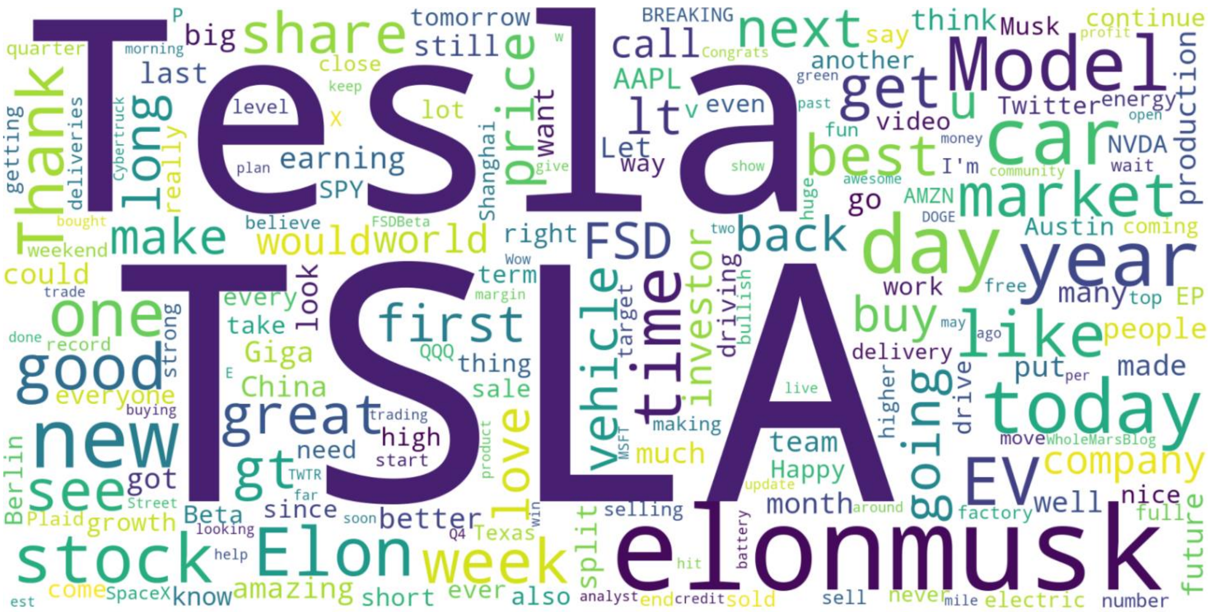


We do same with the positive tweets,

Greater than 0.7,

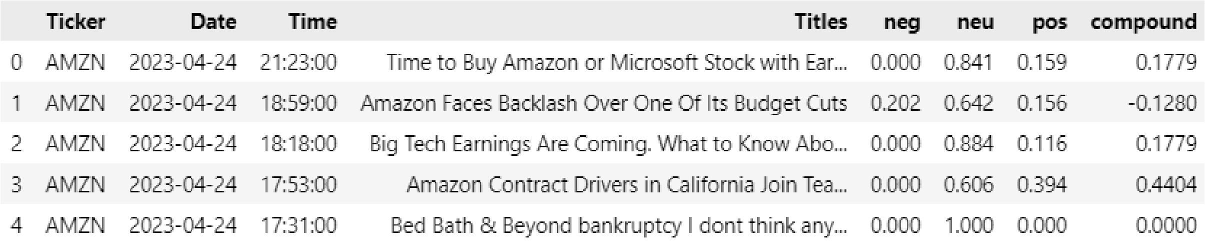


Greater than 0.5,



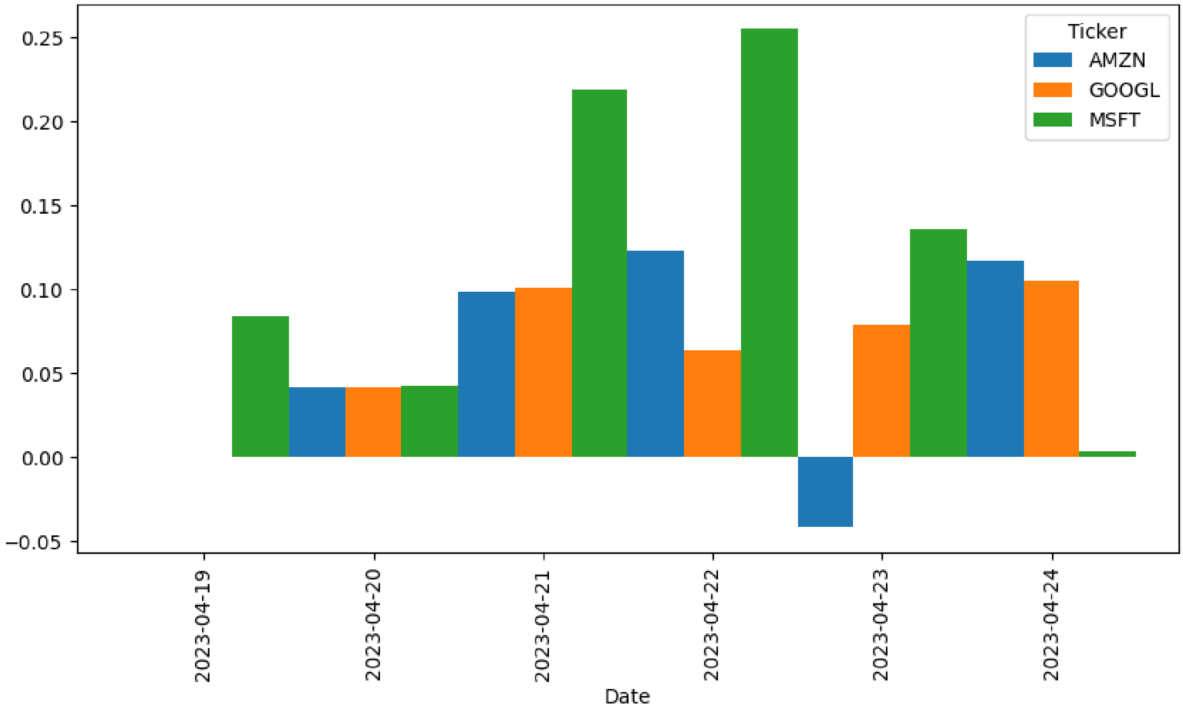
We can see many positive terms in the graphs above as well as negative terms in negative sentiment graphs on the previous page, that means that the sentiment classification done by the Roberta model is pretty good.

Now we head towards, news data fetched from last 5 days, we are fetching news data of total 3 stocks which are, Amazon, Google and Microsoft, we get the data from 20th April to 24th April. We apply NLTK's Vader sentiment analyzer for giving the sentiment scores to the news data fetched from Finviz. This algorithm is rule based and hence can not comprehend complex phrases that much effectively as ML based analyzers. Advantage of this algorithm is that again we don't need preprocessing of the text.

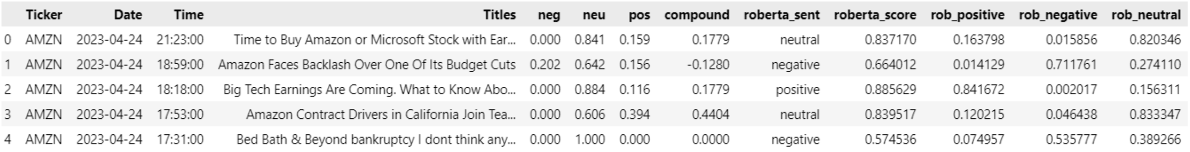


	Ticker	Date	Time	Titles	neg	neu	pos	compound
0	AMZN	2023-04-24	21:23:00	Time to Buy Amazon or Microsoft Stock with Ear...	0.000	0.841	0.159	0.1779
1	AMZN	2023-04-24	18:59:00	Amazon Faces Backlash Over One Of Its Budget Cuts	0.202	0.642	0.156	-0.1280
2	AMZN	2023-04-24	18:18:00	Big Tech Earnings Are Coming. What to Know Abo...	0.000	0.884	0.116	0.1779
3	AMZN	2023-04-24	17:53:00	Amazon Contract Drivers in California Join Tea...	0.000	0.606	0.394	0.4404
4	AMZN	2023-04-24	17:31:00	Bed Bath & Beyond bankruptcy I dont think any...	0.000	1.000	0.000	0.0000

We plot them with clustered column chart for compound sentiment scores given by the algorithm.

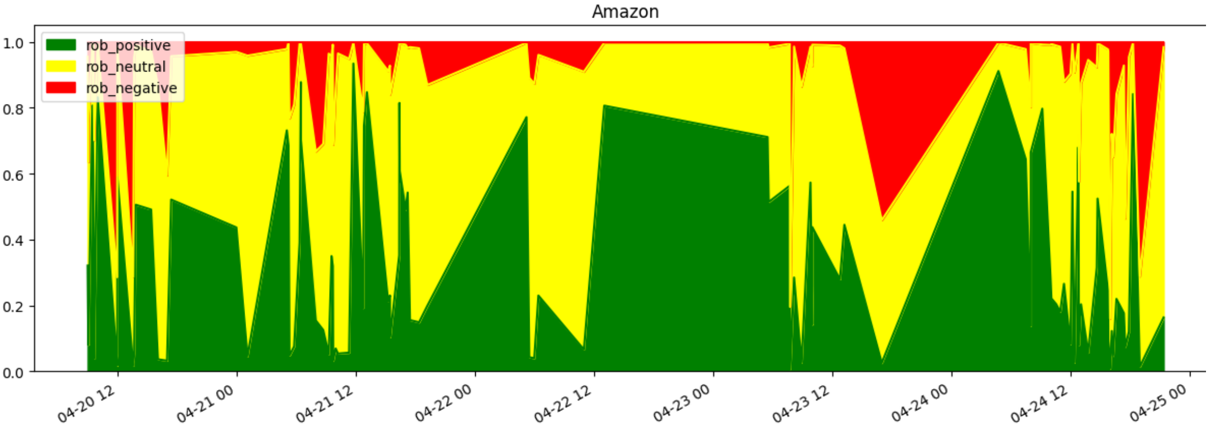


Now we assign Roberta model's score to the same data frame,

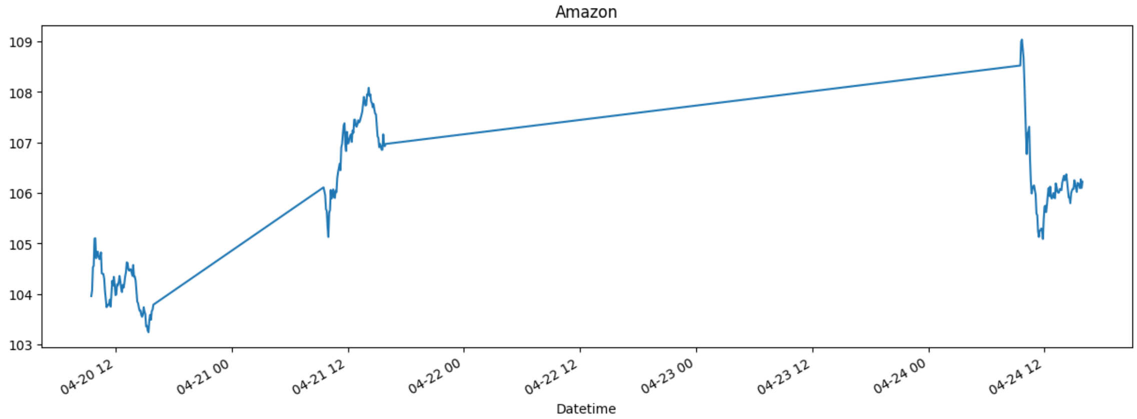


	Ticker	Date	Time	Titles	neg	neu	pos	compound	roberta_sent	roberta_score	rob_positive	rob_negative	rob_neutral
0	AMZN	2023-04-24	21:23:00	Time to Buy Amazon or Microsoft Stock with Ear...	0.000	0.841	0.159	0.1779	neutral	0.837170	0.163798	0.015856	0.820346
1	AMZN	2023-04-24	18:59:00	Amazon Faces Backlash Over One Of Its Budget Cuts	0.202	0.642	0.156	-0.1280	negative	0.664012	0.014129	0.711761	0.274110
2	AMZN	2023-04-24	18:18:00	Big Tech Earnings Are Coming. What to Know Abo...	0.000	0.884	0.116	0.1779	positive	0.885629	0.841672	0.002017	0.156311
3	AMZN	2023-04-24	17:53:00	Amazon Contract Drivers in California Join Tea...	0.000	0.606	0.394	0.4404	neutral	0.839517	0.120215	0.046438	0.833347
4	AMZN	2023-04-24	17:31:00	Bed Bath & Beyond bankruptcy I dont think any...	0.000	1.000	0.000	0.0000	negative	0.574536	0.074957	0.535777	0.389266

We can see the sentiment timeline, by plotting the Roberta’s three parameters as an area chart for each news in chronological order for AMZN ticker.

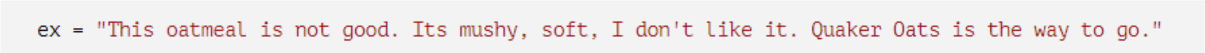


We can see there are some negative blobs near the start of the 24th April, so we expect the price to fall down on start of the 24th April, we can check it with the yfinance API, for AMZN ticker,



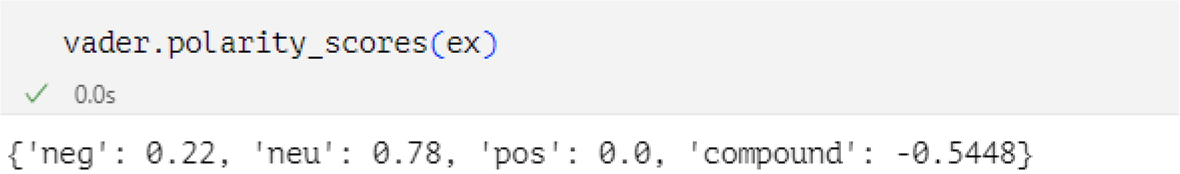
We can note that there was one positive news on 24th but after that, we got series and neutral and negative ones, so the stock is going downwards on 24th. As we seen previously that the news is possibly little bit biased towards giant companies, so we need to offset that effect a little bit. Collectively we can find fair correlation between stock movement and sentiment scores from all the above analyses.

Little demo of vader vs Roberta, we have a sentence which is not straight-forward and has negative inclination, we check scores on both models.



```
ex = "This oatmeal is not good. Its mushy, soft, I don't like it. Quaker Oats is the way to go."
```

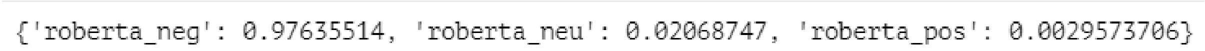
Vader:



```
vader.polarity_scores(ex)
```

```
{'neg': 0.22, 'neu': 0.78, 'pos': 0.0, 'compound': -0.5448}
```

Roberta:



```
{'roberta_neg': 0.97635514, 'roberta_neu': 0.02068747, 'roberta_pos': 0.0029573706}
```

Conclusion:

In summary, Twitter sentiment analysis can be a useful tool for investors looking to gain insights into market sentiment and make informed investment decisions, but it should be used alongside other forms of analyses and with caution.

We saw how ML based models for sentiment analysis are superior to their rules-based counterparts and we have seen fluctuations in stock price, trading volumes and indicators because of tweet volume and sentiment. We fetched data from news website, dumped that data into data frames and csv files, which can be reused and utilized for future analysis.

We surely can’t make the decisions based solely on the basis of sentiments but for sure, it will be a supporting factor for building the prediction models based on both the statistical data and sentiment analysis.

Github Link:

https://github.com/nvalse81/capstone_720.git

References:

Roberta ML model

https://huggingface.co/docs/transformers/model_doc/roberta

Beautiful Soup:

https://pypi.org/project/beautifulsoup4/

Rule based NLTK Vader sentiment analysis library

https://www.nltk.org/_modules/nltk/sentiment/vader.html

Roberta implementation:

https://www.kaggle.com/code/robikscube/sentiment-analysis-python-youtube-tutorial/notebook

Stock Indicators:

https://academy.binance.com/en/articles/5-essential-indicators-used-in-technical-analysis

Thank You!