Introduction

Prediction of the stock market and crypto market performance is one of the most difficult things to do. There are so many aspects influencing the prediction – physical factors vs. psychological, behaviour, sentiments, etc. All these factors combine to make share prices volatile and very difficult to predict with a high degree of accuracy.

In this project, we use sentiment analysis and machine learning to establish the correlation between "public sentiment" and "market sentiment". We use data from twitter to predict public mood and use the predicted mood to forecast the stock market and crypto market movements. We access historical tweets using snscrape, along with daily historical stock data, which we use to incorporate while training the model to make it more robust.

Problem Identified

Researchers for a long time have been working on volatile market prediction. The Efficient Market Hypothesis (EMH) says that stock market prices are largely influenced by trending information and follow a random walk pattern. Though this hypothesis is widely accepted by the research community as a central paradigm governing the markets in general, several people have attempted to extract patterns in the way stock markets behave and respond to external stimuli.

Financial markets consist of a lot of investment opportunities. Trading on Indian stock exchanges(NSE & BSE) is done by a lot of people in India as well as abroad. In recent years, investment activities in cryptocurrencies have spiked due to the rise of bitcoin and ethereum based coins. Social media networks, particularly twitter, allows people to interact with each other in relation to particular stocks or coins. Any financial instrument generally indicates the sentiment of investors, and any market is affected by the dominant sentiment. Hence this sentiment can be used to find effect on the market with respect to particular coins and stocks.

Methodology [Stocks]

Evolving Process

The tweet data will be used to categorise tweets into positive and negative tweets. The volume of negative and positive tweets will determine a polarity index through which we will classify the event as a negative, neutral or positive event. To cover a broad scale of the market, we are taking stocks from all the different sectors like telecom, pharma etc. Further the polarity data and current twitter prices will be fed into different machine learning models.

Data Collection

In addition to data collected using Twitter api v2, we used snscrape, which can scrape tweet data from beyond one week. Present data includes tweet data from Dec 7,2021 to April 30,2022 and daily stock price data from Dec 7,2021 to April 30,2022.

Methodology

Step 1	Step 2	Step 3
First collect all the tweets and then remove special characters from the tweet, to make the sentiment analysis more uniform. **Then**, sort all the tweets datewise and collect them in a dataframe.	**We** import daily stock price data for the stock, and add it in the datewise tweet dataframe. **For** days, where we do not have the markets open, we put in the price as mean closing price of all the days	**We** perform sentiment analysis on the tweet content clubbed together using nltk library. **Sentiments** are calculated and a compound score is calculated which tells the sentiment based on its value.

	Date	Tweets	Prices	Comp	Negative	Neutral	Positive
0	2022-04-14	May god shower mercy amp forgiveness on all th...	3189	0.9661	0.0	0.689	0.9661
1	2022-04-13	18042022 Nifty50 BankNifty Views with Asianpa...	3080	0.9797	0.03	0.843	0.9797
2	2022-04-12	RajkumarNKL asianpaint Intraday levels for 12...	3131	0.8074	0.043	0.821	0.8074
3	2022-04-11	TOMMOROW FOCUS CASH amp FampO stocksRENUKASUG...	3157	0.9501	0.019	0.871	0.9501
4	2022-04-10	Paul TorresCohen ASIANPAINTBO chartmojo MY FA...	3189	0.9738	0.032	0.78	0.9738
...	...	...	...	...	...	...	...
127	2021-12-06	Exit all qty ASIANPAINT 2900 DEC PUT at avg 3...	3038	0.1025	0.078	0.832	0.1025
128	2021-12-05	ASIANPAINT Asian Paints TradingView httpst...	3189	0.7003	0.027	0.865	0.7003
129	2021-12-04	Videography ASIANPAINT With Quick entries and...	3189	0.5267	0.0	0.891	0.5267
130	2021-12-03	I think even promoter realises overvalued ASI...	3110	0.9944	0.018	0.861	0.9944
131	2021-12-02	POONAWALA FincorpMetropolis 63Moonsasianpaint...	3183	0.8658	0.0	0.918	0.8658

Results

We have used a complex random forest model. The dataset is split into 4:1, after experimenting with 7:3 and 9:1 train:test ratio, with 4:1 giving the most optimum results. Sentiment of tweets is analysed and processed in order to relate sentiment score to the prices of a particular stock. The model is able to mimic the general pattern but is unable to fit to an exact level of price. This means that we can take its help to analyse the general sentiment to get an understanding if the stock price can rise or fall accordingly.
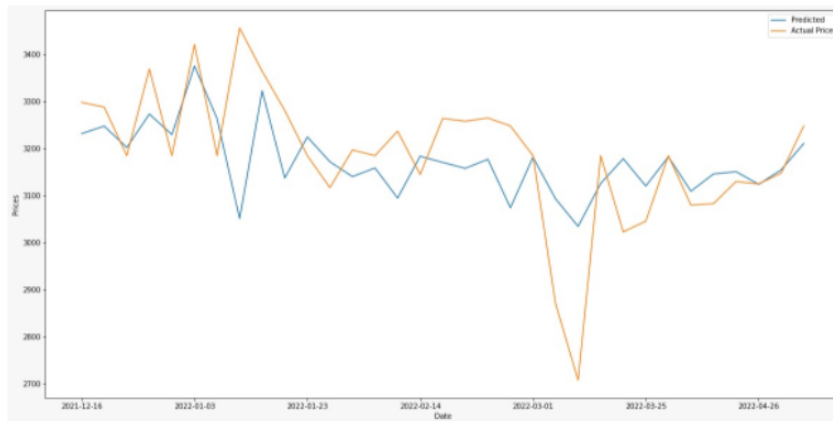
Asian Paints

model_category	Regression
scoring_time	1651579582041
predictions	·
MSE	14987.477261
RMSE	122.423353
nobs	33
custom_metric_name	·
custom_metric_value	0
r2	0.254758
mean_residual_deviance	14987.477261
mae	84.648164
rmsle	0.039417

Methodology [Cryptocurrency]

Data Collection

Using Twitter api v2, we used tweepy to collect tweet volumes. We have used coingecko api to collect price values for our listed crypto-coins which includes daily tweet volume from March 31st,2021 to April 15,2022.

```
     index        Date  Hour         Price   value
0        0  2022-03-31    20   3285.488038    7548
1        1  2022-03-31    21   3302.292471    6297
2        2  2022-03-31    22   3295.810537    4550
3        3  2022-03-31    23   3289.411453    4545
4        4  2022-04-01     0   3283.302843    4627
..     ...         ...   ...           ...     ...
162    162  2022-04-07    14   3228.910849   11614
163    163  2022-04-07    15   3193.383845   10710
164    164  2022-04-07    16   3199.321845    9980
165    165  2022-04-07    17   3211.435520    9054
166    166  2022-04-07    18   3224.677431    9215

[167 rows x 5 columns]
```

Methodology

The tweet data has been used to classify tweets into positive and negative categories. The volume of negative and positive tweets help us in determining a polarity index through which we will classify the event as a negative, neutral or positive event. To cover a broad scale of the market, we have taken stocks from all the different sectors like telecom, pharma etc. Further we feed the polarity data and current twitter prices into different machine learning models.

Results

Conclusion

We used sentiment analysis on tweets to forecast stock and crypto market movements. Our model is able to predict a general movement pattern in stocks but fails to fit in exact price ranges. This means a general sentiment analysis of tweets can forecast rise and fall in Indian stock markets. Oppositely such forecasts cannot be done in the case of cryptocurrencies. This implies that cryptocurrency market movements are feebly linked with tweet sentiments. But using certain variations of the LSTM model we were able to predict rise-fall movements in cryptocurrencies. This can be a topic of research with future scope.

Future Work

Finally, it's worthy to mention that our analysis doesn't take into account many aspects. Firstly, our dataset only considers twitter using English speaking people and hence doesn't really map the real public sentiment. A higher correlation can possibly occur if the actual mood is studied. It may be said that people's mood definitely influences their investment decisions, hence the correlation. But in that case, people who invest in stocks are not directly correlated to those who use twitter more often. Though there certainly is an indirect correlation because investment decisions of people may be affected by the moods of people around them, ie. the general public sentiment. All these remain as areas of future research.