DATA 690 NLP Project Report

Stock and Cryptocurrency Price Prediction Using Sentiment Analysis

Chetan Basawanthraya Desai

Johnny Morgan

Tirusew Wube

Yuksel Baris Dokuzoglu

ABSTRACT

The purpose of this study is to determine whether there is sufficient correlation between sentiment on social media, and real prices on the stock market in the U.S (New York Stock Exchange) to support stock price prediction. Social media data was gathered from Twitter and Reddit. Stocks used in this study are Apple, Tesla, Amazon, Microsoft, Google, Gamestop. Crypto currencies analyzed are Bitcoin, Ethereum, and Dogecoin. Sentiment scores were produced using the Vader Sentiment analysis. Model selection was performed using the Lazy Predict: LazyClassifier and LazyRegressor models. Results from the Lazy Predict model were analyzed and price prediction was performed using the RandomForestRegressor model from sklearn.ensemble. Results of the study reveal that the calculated mean absolute errors are low, the predicted prices are close to the actual prices, and there is a positive correlation between the sentiment and the price.

INTRODUCTION

Although forecasting stock prices is difficult. Generally, there are two main viewpoints in stock trading: these are fundamental and technical analysis. In fundamental analysis, the financial conditions of the company and the financial indicators are utilized to evaluate a company's general status to foresee its stock price. On the other hand, technical analysis uses repetitive patterns in stock price. It analyzes the stock's historical price and tries to predict future fluctuations according to these patterns (Derakhshan & Beigy, 2019). Some researchers used only the price action, which is another common method for traders. They only focus on price and slightly use other indicators. According to some research by Patel et al. (2015), time-series analysis and Auto Regressive models can be used. They focused on patterns in the history of the price and implemented their method accordingly. Some studies are based on a random walk model. They assume that the stock market reflects according to events and news. Since it is impossible to predict them, they imply that it is also impossible to predict the changes in stock prices (Walczak, 2001). However, none of the researchers discussed so far used sentiment analysis as a tool. As data science is improving, research methodology is also changing. Since information is readily available to all nowadays, reaction time is getting shorter. There is more research coming out that proves the opposing notions, especially with the growth of data science.

According to research by the California Institute of Technology, while many economists claim that the reflection of the market is rational, it is affected by people’s behavior. Even though it may not be the only factor, people's psychology affects the market (*Psychology Influences Markets*, 2013). A new study suggests that people who are more advanced at processing information have an advantage in stock market investment (Pedersen, 2015). One of the best ways to process this kind of information is sentiment analysis.

Sentiment analysis is a branch of Natural Language Processing (NLP), and it is the newest and optimal way to analyze people’s opinion in real time. Analysts, scholars, and practitioners are aware of the importance of sentiment analysis in decision making for investment. Therefore, there are several studies which use the combination of historical price data and textual data on the internet, particularly social media platforms. Most of the studies had their focus on binary classification problems on stock prices questioning whether they will go up or down.

One of the main sources of information is the news headlines. FinViz is an online news platform that is used in this study. However, the understanding of news can vary, thus it is not a sufficient source alone to analyze the sentiment. For that reason, to enhance our analysis we also used data from Twitter and Reddit in this study. By doing that, we were able to get the sentiment from various types of investors from different channels. Nonetheless, this kind of analysis is not easy because people's reviews on Twitter are short, and they contain idioms and sarcasm. Therefore, using different sources for data and using thousands of elements of data is very crucial to make a good analysis.

DATASET - SOCIAL MEDIA DATA ACCESS

Data was collected from Alpaca, FinViz, Reddit, and Twitter through API calls. The data acquisition was difficult and time consuming as these APIs are slow, and the available data is severely limited in terms of historic data. Several attempts were made to acquire Twitter, Reddit, and FinViz data going back six months. The stocks chosen here are some of the biggest tech companies in the world and they are following an upgoing trend in the last 10 years except Gamestop. Gamestop is chosen because of the fluctuation and public movement that happened towards it in the last year. We selected Bitcoin and Ethereum as the leading top two coins in the crypto market. Dogecoin is chosen for similar reasons as Gamestop. In the last year, Dogecoin was promoted by Elon Musk and some other groups on Reddit. Due to these manipulations, its price was one of the most unstable in the crypto market.

Twitter

In this study, Twitter REST API is used to gather data. The type of the API used was Search API. Twitter data only allows free account access to 7 days of data. Twitter offers an Academic Research Access account for qualified master's and Ph.D. students working on research projects. Our credentials as master's degree students could potentially qualify us. The process was difficult as there was a credential process that was daunting at the time under a time crunch. We have since completed the application and hope to have this access for capstone. There were a few difficulties such as providing a url to our google scholar profile. We didn't have one, and it was unclear how to get one since we have not published any papers publicly. There was a work around and we were able to get past that part. Then there was a problem that the url to our profile was too long. So, we made a short url to the profile in bitly and were able to process the application. This was too late to benefit this project, but future projects will benefit from this extended access to historical data. For purposes of this project the Twitter data was limited to 7 days of data.

This paper uses stocks' ticker names and coins' full names for keywords. In the dataset, there is the full tweet and date columns. The language of the tweets is limited to only English. After gathering data, the datasets are saved in ".pkl" format. The number of tweets is limited to 20 thousand because otherwise it takes too long the code to accomplish the task. The run time usually just stops, and the task is not complete. The sentiment scores from tweets and reddit for each stock between 11.13.2021 and 11.20.2021 are shown in the Figures(Fig.01 and Fig.02) below.

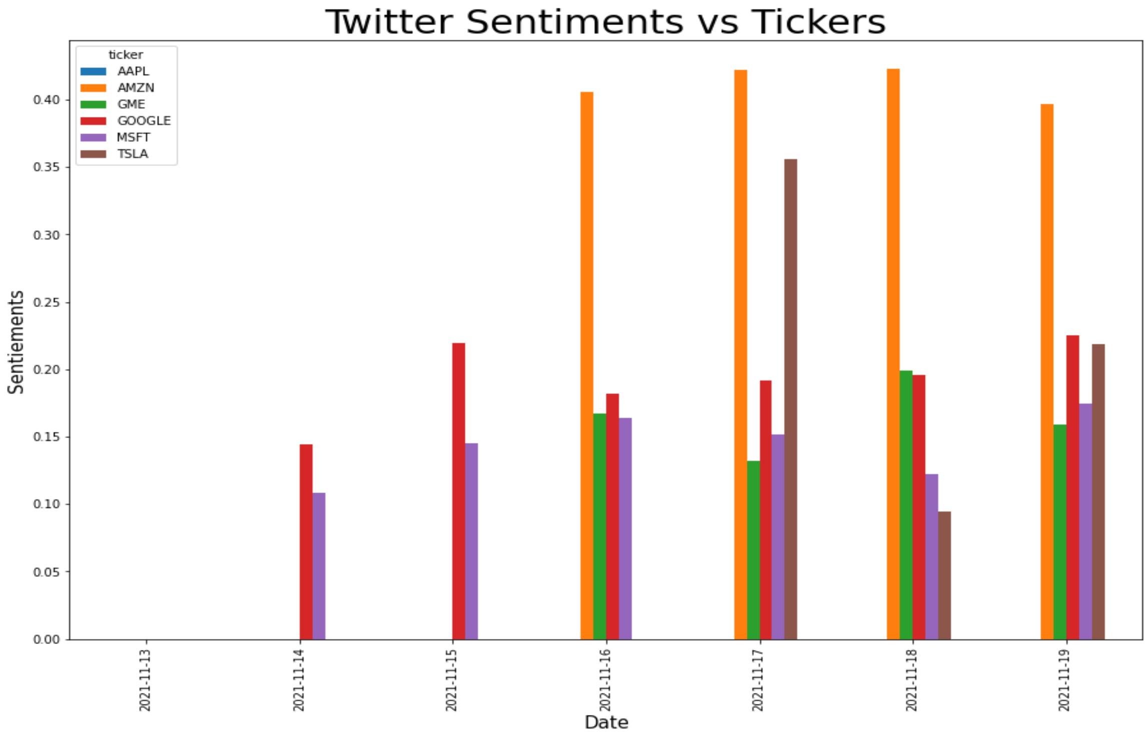


Fig. 01

FinViz

FinViz is a stock screener platform and a trading tool used for creating financial displays. In this study, FinViz is used to get news headlines about the stock market and crypto market. The data is gathered about the same stocks and coins. The dataset is saved in ".pkl" format and contains text and date columns.

Reddit

Reddit is one of the most important social media platforms for investors. The effect of users and communities on Reddit to the stock market is undeniable. In 2021, a movement started by Reddit users to support Gamestop company, and its stock price increased 2700%. Similar kind of impact happened on meme coins like Dogecoin and Shiba coin. There are hundreds of subreddits for each topic. The data format is the same as others and contains text and date columns. The time and text columns are used in analysis from the Reddit dataset.

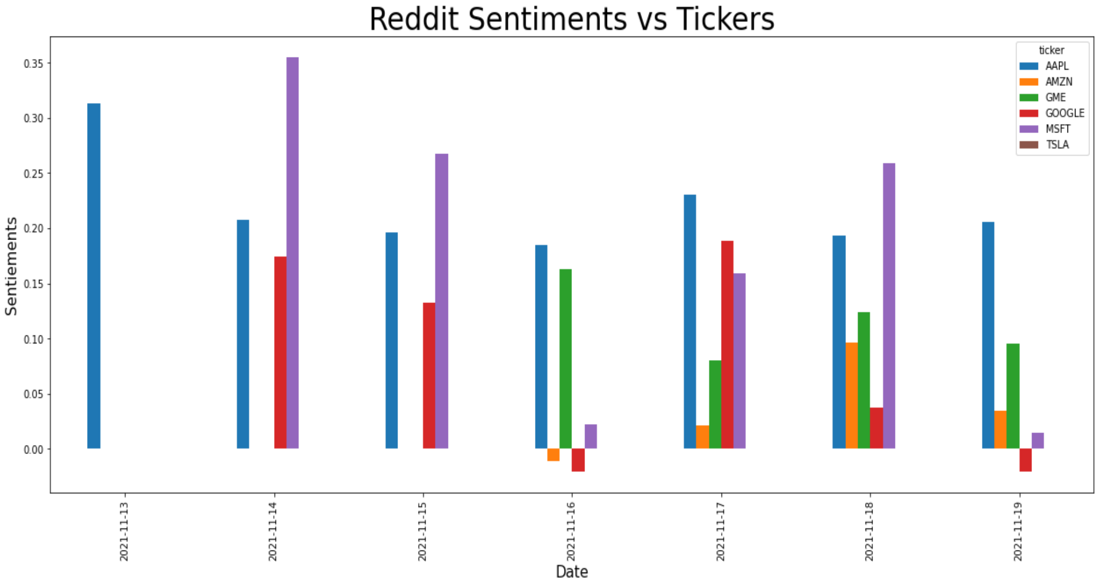
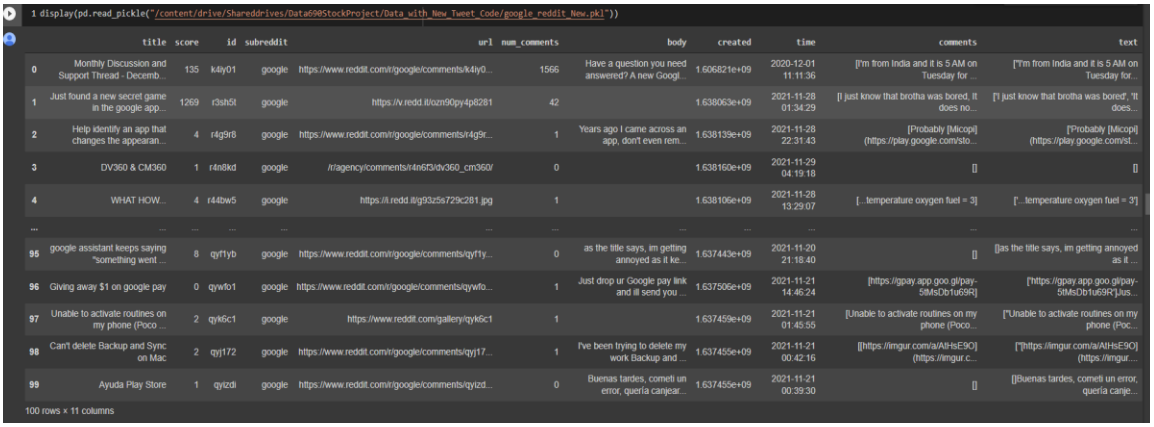


Fig. 02



```
display(pd.read_pickle("/content/drive/Shareddrives/Data690StockProject/Data_with_New_Tweet_Code/google_reddit_New.pkl"))
```

	title	score	id	subreddit	url	num_comments	body	created	time	comments	text
0	Monthly Discussion and Support Thread - Decemb...	135	k4iy01	google	https://www.reddit.com/r/google/comments/k4iy0...	1566	Have a question you need answered? A new Googl...	1.606821e+09	2020-12-01 11:11:36	[I'm from India and it is 5 AM on Tuesday for ...	["I'm from India and it is 5 AM on Tuesday for...
1	Just found a new secret game in the google app...	1269	r3sh5t	google	https://v.redd.it/ozn90py4p8281	42		1.638063e+09	2021-11-28 01:34:29	[I just know that brotha was bored, It does no...	['I just know that brotha was bored', 'It does...
2	Help identify an app that changes the appearan...	4	r4g9r8	google	https://www.reddit.com/r/google/comments/r4g9r...	1	Years ago I came across an app, don't even rem...	1.638139e+09	2021-11-28 22:31:43	[Probably [Micopi] (https://play.google.com/sto...	['Probably [Micopi] (https://play.google.com/st...
3	DV360 & CM360	1	r4n8kd	google	/r/agency/comments/r4n6f3/dv360_cm360/	0		1.638160e+09	2021-11-29 04:19:18	[]	[]
4	WHAT HOW...	4	r44bw5	google	https://i.redd.it/g93z5s729c281.jpg	1		1.638106e+09	2021-11-28 13:29:07	[...temperature oxygen fuel = 3]	['...temperature oxygen fuel = 3']
...	...	...	...	...	...	...	...	...	...	...	...
95	google assistant keeps saying "something went ...	8	qyf1yb	google	https://www.reddit.com/r/google/comments/qyf1y...	0	as the title says, im getting annoyed as it ke...	1.637443e+09	2021-11-20 21:18:40	[]	[]as the title says, im getting annoyed as it ...
96	Giving away $1 on google pay	0	qywfo1	google	https://www.reddit.com/r/google/comments/qywfo...	1	Just drop ur Google pay link and ill send you ...	1.637506e+09	2021-11-21 14:46:24	[https://gpay.app.goo.gl/pay-5tMsDb1u69R]	['https://gpay.app.goo.gl/pay-5tMsDb1u69R']Jus...
97	Unable to activate routines on my phone (Poco ...	2	qyk6c1	google	https://www.reddit.com/gallery/qyk6c1	1		1.637459e+09	2021-11-21 01:45:55	[Unable to activate routines on my phone (Poco...	["Unable to activate routines on my phone (Poc...
98	Can't delete Backup and Sync on Mac	2	qyj172	google	https://www.reddit.com/r/google/comments/qyj17...	1	I've been trying to delete my work Backup and ...	1.637455e+09	2021-11-21 00:42:16	[[https://imgur.com/a/AtHsE9O] (https://imgur.c...	["[https://imgur.com/a/AtHsE9O] (https://imgur....
99	Ayuda Play Store	1	qyizdi	google	https://www.reddit.com/r/google/comments/qyizd...	0	Buenas tardes, cometi un error, quería canjear...	1.637455e+09	2021-11-21 00:39:30	[]	[]Buenas tardes, cometi un error, quería canje...

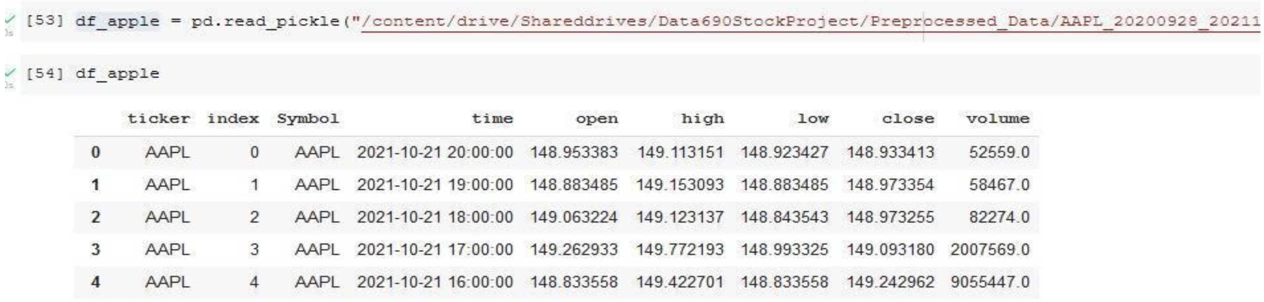
100 rows × 11 columns

Fig. 03

Figure (Fig.03) Shows the dataset for the Reddit entries obtained for Google, including the subreddit, creation date, and time. Reddit API calls are available through a few python libraries. However, documentation is sparse, and the API access is extremely time consuming. A lot of time is spent on researching API accesses in the various platforms. Reddit API access has a popular library called praw and a newer API called pmaw (Podolak, 2021). The benefit of pmaw is that it is optimized to take advantage of multithreading. Additionally, API calls for posts return a record for each post, but comments to the post require additional API calls. These calls are not intuitive and resulting comment trees are often so deep that the API call seems to stall and run for hours.

Stock/Coin Price Dataset

Price dataset is gathered using Yahoo Finance. The data is gathered for every 60 minutes containing stock's open, close, high, and low price. The same timeframe is used with the other datasets. The closing price is used to do prediction and analysis. Crypto market data is also gathered from Yahoo Finance for 60 minutes intervals. The difference between the crypto market and stock market is that the crypto market is open everyday 24 hours. For that reason, the number of rows in the crypto dataset is more than the stock price dataset. These datasets are also saved in ".pkl" format.The price dataset for Apple stock is displayed at one-hour intervals, along with the date and time below(Fig. 02).



```
[53] df_apple = pd.read_pickle("/content/drive/Shareddrives/Data690StockProject/Preprocessed_Data/AAPL_20200928_20211
```

```
[54] df_apple
```

```
  ticker  index  Symbol                 time        open        high         low       close     volume
0   AAPL      0    AAPL  2021-10-21 20:00:00  148.953383  149.113151  148.923427  148.933413    52559.0
1   AAPL      1    AAPL  2021-10-21 19:00:00  148.883485  149.153093  148.883485  148.973354    58467.0
2   AAPL      2    AAPL  2021-10-21 18:00:00  149.063224  149.123137  148.843543  148.973255    82274.0
3   AAPL      3    AAPL  2021-10-21 17:00:00  149.262933  149.772193  148.993325  149.093180  2007569.0
4   AAPL      4    AAPL  2021-10-21 16:00:00  148.833558  149.422701  148.833558  149.242962  9055447.0
```

Fig. 02

SENTIMENT ANALYSIS

Sentiment analysis makes it possible to classify the opinions. There are different kinds of classifications to determine the polarity of the opinions. Some can be more basic and classified into 3 groups like positive, negative, and neutral. Some models can be more detailed, and it gives a polarity rate from 1 to 10 where 1 is the most negative and 10 stands for the most positive. Another example of this is classification into mood states like alert, vital, sad, angry, happy, calm, etc. These models differ depending on the needs of analysts.

There are multiple resources to gather data about finance like news articles. While there is data from FinViz in this study, there is more data used from social media platforms like Twitter, and Reddit. Investors share their opinion in real time on these platforms. This makes it easier for analysts to understand the sentiment, rather than just using news headlines which are limited compared to social media channels. While there are limited numbers of news, there are millions of different opinions on Twitter and Reddit. However, what can be gathered from any source of information is raw data. These datasets need to be cleaned and manipulated in a certain way to do analysis. Vadersentiment is chosen for this study to do sentiment analysis. Vadersentiment reads the clean text column and creates a sentiment score for positive, negative, neutral, and compound. These scores are generated and shown in four new columns.

STRUCTURE OF PROCESSING

Data collected and processed were managed in pandas data frames and each step in the process was saved as dataframes in pickle files. Twitter and Reddit data was collected and stored in separate files for each stock/coin from each social media platform.The text column containing the social media post text was cleaned to support sentiment processing and the cleaned text was placed in a new column. Sentiment processing was performed on each social media file and resulting sentiment columns were added to the dataframe and stored in a preprocessed directory. The preprocessed files were processed into groupby files where rows of sentiment were grouped by time into hourly groups and the numeric data was averaged. This data was stored as groupby files in a group by directory. The non-numeric data does not propagate to the groupby files. This data can still be accessed in the preprocessed files directory. The groupby files are then merged into a single file where the price data and sentiment data from both social media sources are merged into one file per stock/coin. These merged files were stored in the merged files directory. The resulting data was still extremely limited in counts and by time. The team was given a file collected from Reddit from a subreddit called Wallstreetbets. Whereas the previous data was limited to 7 days from Twitter and amounted to 20k records at most, Reddit was only a little better in that it had 70k rows it still only spanned 30 days. By contrast the Wallstreetbets data spanned a year and held 5.2million rows. This data had to be processed and merged into the same data structure. Filters were developed to extract data associated with each stock/coin since this data was not from a subreddit associated with a single stock. Wallstreetbets subreddit is a forum where people discuss any number of stocks. It was well worth the extra effort since the additional data significantly improved the modeling accuracy.

PRICE PREDICTION COMPARED WITH SENTIMENT

We have trained two models, Random Forest Classifier and Random Forest Regressor. We used a random forest classifier to see the trend of stock price with corresponding sentiments from social media. We took the price difference from the previous day and changed it to categorical value as fall, rise and no change to get the stock price trend over time. We trained the Random Forest Classifier model using the sentiments of positive, negative and neutral values as X variable and the stock price trend as target variable. The model would predict stock price as it fell, raised or did not change when the sentiments changed. The model prediction performed well as its performance accuracy scored 94%, we can conclude our model could predict 94 percent accurate.

The second model we used was Random Forest Regressor to predict stock price based on the sentiments of social media. Lazy Predict is used to determine which machine learning model will be used for prediction. Apple stock data is analyzed with Lazy Predict. According to the analysis, Random Forest Regressor was the model with the highest R-squared score 95% and RMSE 4.04, so it is used for prediction. Table 01 shows the models analyzed with Lazy Predict.

Model	**Adjusted R-Squared**	**R-Squared**	**RMSE**	**Time Taken**
RandomForestRegressor	0.95	0.95	4.04	1.81
DecisionTreeRegressor	0.94	0.94	4.11	0.07
BaggingRegressor	0.94	0.94	4.25	0.23
XGBRegressor	0.93	0.93	4.66	0.96
ExtraTreesRegres	0.92	0.92	4.90	1.22

sor				
ExtraTreeRegressor	0.87	0.87	6.35	0.06
LGBMRegressor	0.85	0.85	6.62	0.22
KNeighborsRegressor	0.85	0.85	6.78	0.35

Table. 01

The data is split to 20% test and 80% train. This process is done for each stock and coin. The R-squared score, mean absolute error, mean squared error, and root mean square error are what we used to determine how well the model functioned(Table. 01). Mean absolute error measures error between paired observations expressing the same phenomenon. Mean squared error measures the average of the squares of the errors. It is the average squared difference between the predicted values and the actual values. Root-mean-square error is used to measure the differences between the predicted values and the actual values. Equations for R-squared score, MAE, MSE, and RMSE are shown below:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Reference: Chugh, A. (2020, December 11)

At this point of the study we had a problem with the Ethereum data set. Contrary to our solution trials, we could not use the data because of the format problems. For that reason, there is no result below for the Ethereum coin data.The values for the stocks and coins from the Evaluation metrics are listed below.

Evaluation Metrics for the Regression Model.

Stocks and Coins	**MAE(Mean Absolute Error)**	**MSE(Mean Squared Error)**	**RMSE(Root Mean Square Error)**	**R2(coefficient of determination)**
APPL	**10.18**	**161.35**	**12.7**	**0.46**
AMZN	**93.21**	**19379.34**	**139.21**	**0.6**
TSLA	**124.88**	**28232.77**	**168.02**	**0.43**
MSFT	**24.74**	**1269.80**	**35.63**	**0.47**
GOOGL	**260.4**	**135582.28**	**368.21**	**0.47**
GME	**76.49**	**7874.17**	**88.73**	**0.06**
BTC	**4010.90**	**42665383.75**	**6531.87**	**0.16**
DOGE	**0.034**	**0.00575**	**0.075**	**0.033**

Table. 02

Lower error matrix shows better performance of the model. However, we should take the stock price into consideration since it can affect the error matrices. For example, the Mean Absolute Error, Mean Squared Error, or Root Mean Squared Error of DOGE is extremely small since the stock price of DOGE is extremely small, so it does not mean the model performance is better for DOGE. Hence, to evaluate the model performance and accuracy in a better way, we should also consider R^2 score that shows how well our model is fitted to the data by comparing it to the average of the output (stock price). If the score is closer to 1, then it indicates that our model performs well but if the score is farther from 1, then it indicates that our model does not perform so well.

Based on the values of the error metrics for regression and R^2 score in Table 3, the model performance for AMZN stock is better, that means the model worked well to predict Amazon stock. The model performance for APPL, TSLA, or MSFT is also good since both the error metrics for regression and the R^2 score are good for these stocks although their MSEs are higher that might be caused by outliers. Based on the Error metrics and R^2 score, the model performance is not bad for GOOGL. However, for BTC and DOGE, the model performance is bad as we can see the error metrics and R^2 score. Hence, we can conclude that the model works well for APPL, AMZN, TSLA, MSFT, and GOOGL.

Based on this model and data set, the correlation value of Apple stock price and public sentiment was 46%. For Amazon the correlation was 60%. For Google and Microsoft stocks this score was 47%, while Tesla's score was about 43%. The rest of the stocks had low correlation values with respect to their public sentiments. From this study we observed that because of the volatile nature of some stocks, the prediction model performance might be affected.

According to the calculated mean absolute errors and as explained above, we can say that our model worked well for the prediction. The predicted price for stocks and coins are not too far out and it is close to their actual price. Below section includes the line charts of comparison between predicted price and actual price for Apple, Bitcoin, and Tesla(Fig 05,Fig 06,Fig 07) respectively.

Apple

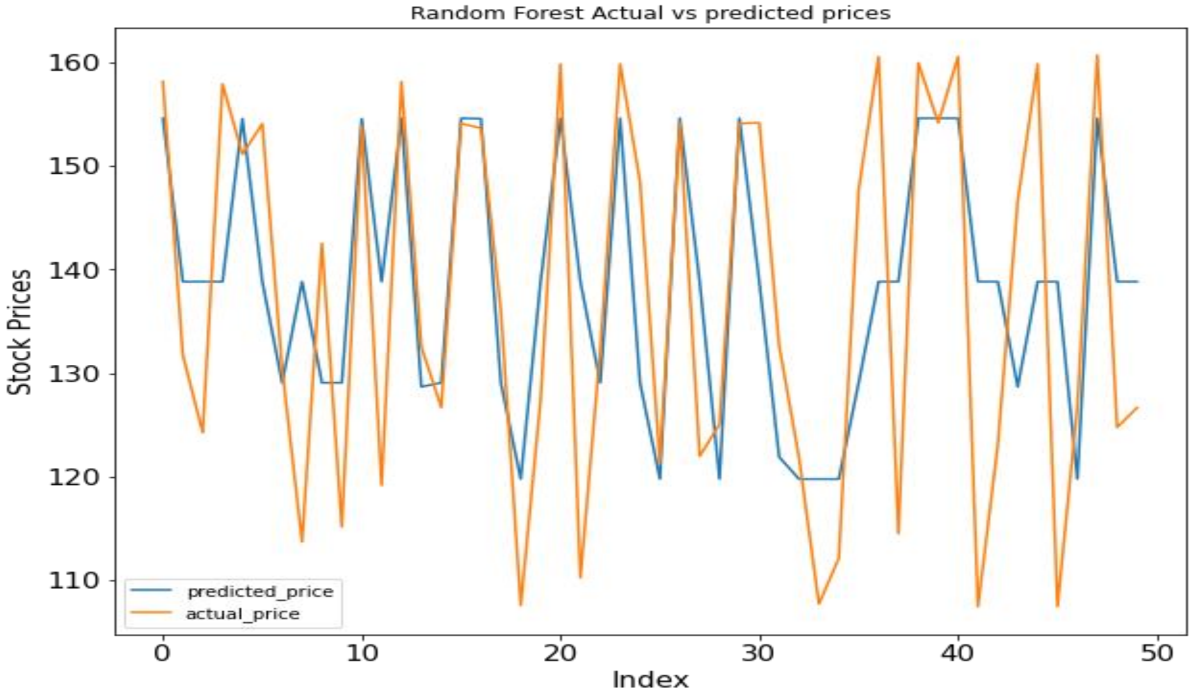


Fig. 05

Bitcoin

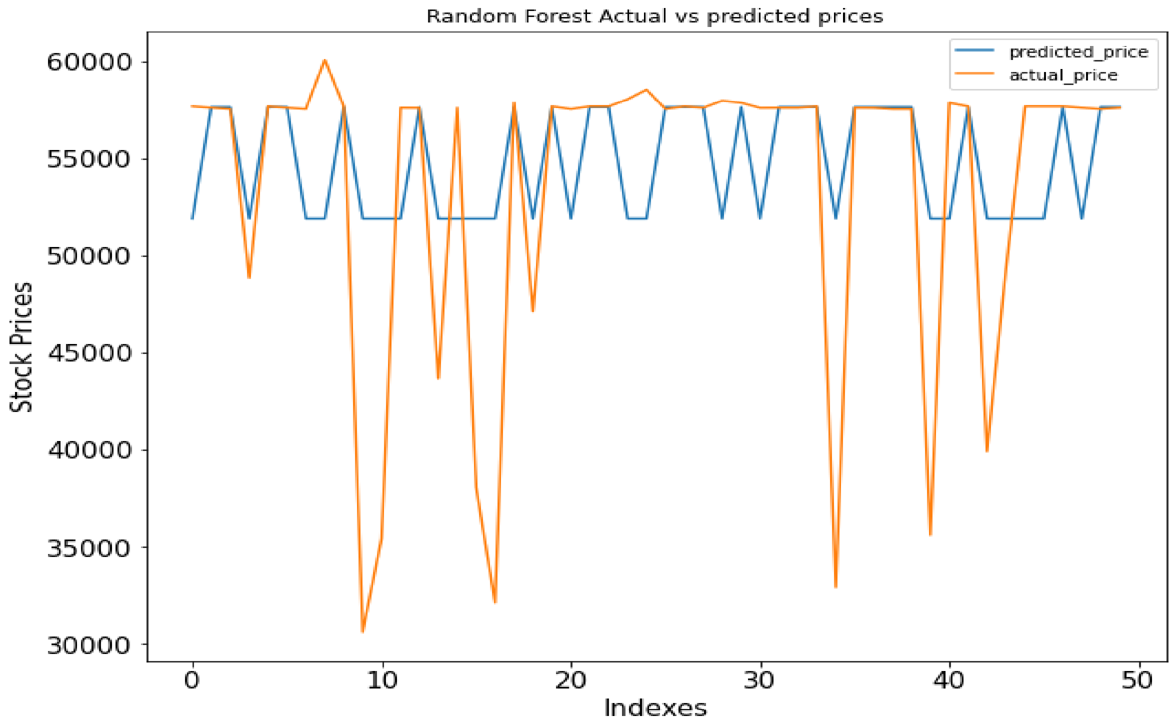


Fig. 06

Tesla

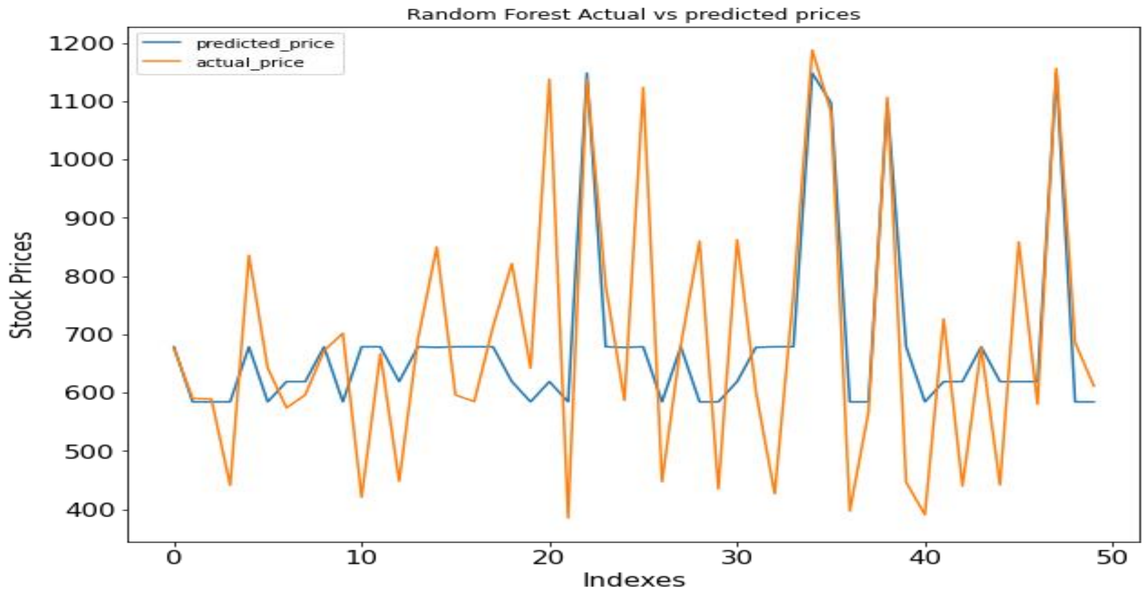


Fig. 07

CHALLENGES

Compute Environment

Difficulties in establishing connections to social media platforms and the resulting long running processes create difficulties in selecting compute resources that are appropriate for the data acquisition task. Accessing social media platform APIs is time consuming. Often processes will run for hours and even days. Resource timeouts are problematic and occur at different levels. Jupyter notebooks running on a home computer or laptop must contend with session timeouts of the computer as the computer account access often has a timeout. Google Colab environment session timeouts terminate running sessions. Even using the Colab pro environment, session timeouts still terminate long running processes.

CONCLUSION AND FUTURE WORK

In this project the model trained for 6 stocks' and 2 coins' data to observe how the public opinion affects the price of stocks. We investigated the potential of public sentiment attitudes extracted from web financial news, tweets and Reddit in predicting stock price movements, using Random Forest Regressor and Classifier. Our project data collected from different social media platforms for the time period between December 2020 and November 2021. The model prediction matched well with the actual stock price as can be viewed on fig (05,06,07). According to our model the correlation between the sentiments and the price of the stocks is positive (see Table 3). The outcome of this study is essential for financial advisors , as they add up to the understanding of the transmission instrument of the monetary policy.

A logical future step of stock price prediction beyond sentiment analysis could be to develop a predictive model to evaluate semantic word vector space distances for financial terms. This effort will provide a model library to provide text analysis classifiers to evaluate the relative nature of a corpus expressed as a set of distances e.g. (Euclidean, Hamming, Manhattan, Minkowski) of a reference corpus to a target corpus. In the case of stock market text analysis, for example, a bull market reference vector would contain reference terms associated with a bull market, and another for a bear market. Target corpus would be evaluated as a distance from the reference corpus in very much the same way as a sentiment model. Thus, text would be classified as a 0 to 1 rating for a bull and a 0 to -1 for a bear.

REFERENCES

B. (2020, November 2). *Sentiment Analysis of Stocks from Financial News using Python*. Medium. https://towardsdatascience.com/sentiment-analysis-of-stocks-from-financial-news-using-python-82ebdcefb638

Briggs, J. (2021, September 2). *Sentiment Analysis for Stock Price Prediction in Python*. Medium. https://towardsdatascience.com/sentiment-analysis-for-stock-price-prediction-in-python-bed40c65d178

Chugh, A. (2020, December 11). *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?* Medium. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjuste d-r-squared-which-metric-is-better-cd0326a5697e

Derakhshan, A., & Beigy, H. (2019). Sentiment analysis on stock social media for stock price movement prediction. *Engineering Applications of Artificial Intelligence*, *85*, 569–578. https://doi.org/10.1016/j.engappai.2019.07.002

L. Pedersen (2015). *Efficiently inefficient: How Smart Money Invests and Market Prices are Determined.* Princeton University Press.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, *42*(4), 2162–2172. https://doi.org/10.1016/j.eswa.2014.10.031

Podolak, M. (2021, November 30). *GitHub - mattpodolak/pmaw: A multithread Pushshift.io API Wrapper for reddit.com comment and submission searches.* GitHub. Retrieved Dec 4, 2021, from https://github.com/mattpodolak/pmaw

Psychology influences markets. (2013). ScienceDaily. https://www.sciencedaily.com/releases/2013/07/130701151608.htm

Walczak, S. (2001). An Empirical Analysis of Data Requirements for Financial Forecasting with Neural Networks. *Journal of Management Information Systems*, *17*(4), 203–222. https://doi.org/10.1080/07421222.2001.11045659