Wall Street Bets Subreddit Sentiment Analysis

An Examination of the Association between Sentiment and Stock Financial Data on the Top S&P 500 Companies

Executive Summary

In January of 2021, the ‘wallstreetbets’ subreddit made waves in the stock market. While the Gamestop stock price was expected to tank by traditional traders and hedge funds, a collective of retail investors in the ‘wallstreetbets’ subreddit community were able to shift the momentum completely - raising the share price from $3 before the beginning of the year to over $300 before the month of January had come to an end. It all started with a single post on the forum by Keith Gill, a financial analyst that posts regularly on the forum, convincing his fellow members to purchase gamestop stocks. As many followed suit, the movement became one that caused many eyes to turn toward the power of social media sentiment even in a traditional field such as the U.S. stock market.

Established institutions such as investment banks, mutual funds, and institutional investors have been forced to take public opinion into account. Despite the unconventional manner in which many social media posts relay sentiment regarding this topic, this specific case proves that these seemingly insignificant memes and posts should not be overlooked.

One question then comes to mind - was this narrative a one-time spectacle or a recurring occurrence in the stock market? This project aims to provide an answer to this very query by uncovering whether there exists a relationship between the sentiment on the ‘wallstreetbets’ subreddit and the financial data on the top companies in the stock market. A multiple linear regression model will be built to then study whether the monthly returns of a company can be predicted using financial ratios and sentiment scores as independent variables.

This analysis will reveal whether there is an association between the variables to be forecasted and the independent variables if the predicted values have a significant enough match to the actual values. To accomplish this task the project entailed the scrapping of Yahoo Finance for the financial data and Reddit for the ‘wallstreetbets’ subreddit posts from the start of October 2021 to the end of October 2022 for Apple (AAPL), Microsoft (MSFT), Google (GOOG), Amazon (AMZN), Meta (META), AMD (AMD), Nvidia Corporation (NVDA), PepsiCo (PEP), Tesla (TSLA), and Netflix (NFLX). Additionally, the reddit posts were then cleaned and organized by the month it was posted and the average monthly sentiment scores were calculated for each company. The monthly sentiment scores and a selection of financial ratios were chosen as variables to be used by the model to predict the monthly returns (a second model was built as well to predict the sentiment score given the monthly returns and the earnings ratios). After the multiple linear regression were built for each company, a comparative analysis was then conducted between the different regression models per company. This gives insight into whether the impact of the ‘wallstreetbets’ subreddit sentiment spans across other top S&P 500 companies or ends with the gamestop incident.

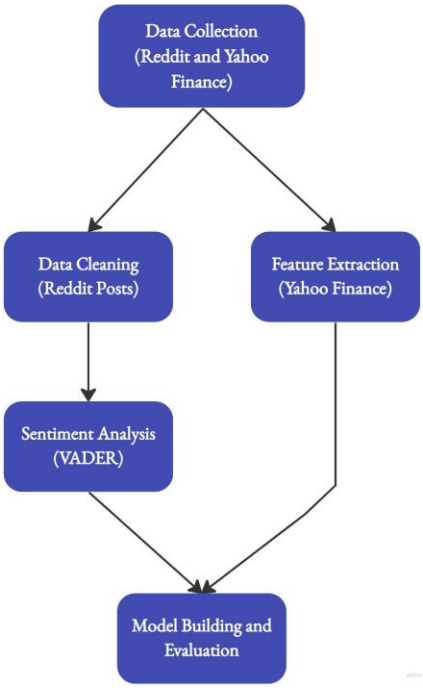
Business Goal

The goal of this analysis is to understand the relationship between the Wall Street Bets subreddit sentiment, the Earning Ratios and Monthly Returns of selected companies. Essentially, the project aims to determine whether social media sentiment has predictive power in determining the monthly return of a company's stock.

Web Analytics is needed for this project because features needed for the model implementation are scraped from the web. APIs will be used to make data extraction more efficient and python modules such as pandas and sklearn will be used to collate the data in dataframes and build the regression models, respectively.

By collecting and analyzing the data made available on the web, insights regarding the relationship between Wallstreet Subreddit forum and financial data on specific companies can be garnered. Relevant points may include whether monthly returns of certain companies can be predicted with the help of sentiment analysis on the wallstreetbets subreddit forum, how relevant sentiment from wallstreet bets is across different companies in bulding a multiple linear regression model to predict monthly returns, amongst others.

System Design



Data Collection and Understanding

The main components that the predictive model will use to garner a forecast require the following data: closing value, financial ratios (Trailing P/E, Forward P/E, PEG Ratio (5 yr expected), Price/Sales (ttm), Price/Book (mrq), Enterprise Value/Revenue and Enterprise Value/EBITDA) of the chosen companies, as well as social media sentiment on the wallstreetbets subreddit on these specific companies. The sources used to collect the necessary data consist of Yahoo Finance and Reddit. Both 'PMAW' and 'YFinance' APIs were used to gather all the Reddit posts on wallstreetbets and the financial data, respectively from October 2021 to October 2022.

Data Cleaning and Sentiment Analysis

After scraping the Reddit Posts, the following were removed to prepare the text data for Sentiment Analysis:

- Links from the text posts
- Duplicate posts
- Posts that have no significant text data (i.e. posts that only have links to youtube videos, posts that only have images, etc.)

After the text data had been preprocessed with python, the VADER sentiment analysis tool was used to get the compound sentiment scores for each individual post. The compound sentiment score ranges from -1 to 1 (with -1 corresponding to a negative sentiment, 0 to a neutral sentiment, and 1 to a positive sentiment). The average sentiment score for the month was then used as an input feature for the predictive model. The VADER sentiment analysis python package tool was chosen as it is able to take into account qualities distinctive to social media posts (i.e. capitalization, exclamation marks, emojis, etc.) that also impact the degree of sentiment. A post with exclamation points and smiley face emojis, for example, will be regarded as one with a more positive sentiment compared to a post with the exact same text sans said exclamation points and additional emojis.

Feature Extraction and Selection

For the first model, the current monthly returns, which is the assigned target variable, was computed based on the closing price at the start and the end of the month, which had both been extracted from Yahoo Finance. In addition to this extracted feature, the Earning Ratios will be used as input features for the predictive model alongside the average sentiment score for the month. As for the second model, the monthly average sentiment score is to be used as the target variable while the earning ratios and the monthly returns are to be used as input features.

Model Building and Evaluation

The final steps of the System Design include the model building using SkLearn and Evaluation. When building the model, the monthly sentiment scores were first adjusted to handle the insufficient number of monthly reddit posts. All 'NaN' values in sentiment were replaced to 0 (neutral sentiment). Afterwhich, the data was ready to be inputted into the two MLR models for each company. A thorough breakdown of all the models built will be elaborated under the System Implementation Segment.

As per the Evaluation, the following insights were observed:

1. The companies with the most posts usually had the best performing models
2. Within the span of a month, it is possible sentiments may have very contrasting values - which may cancel out when averaged. This was a limitation of the model which made companies with sentiments that aligned perform better

Dataset Description

Financial Data

The Financial Data has been acquired from the website Yahoo Finance. The Earning Ratio dated from October 2021 to October 2022 of selected companies was scraped using the YFinance API. The Earning Ratios are on a quarterly basis. Moreover, we scraped data of the Closing Stock Price for each company to be able to calculate the Current Daily Return with the following formula:

$$Current\ Month\ Return\ =\ ln\,e\left(\ \frac{Closing\ Price\ Current\ Month}{Closing\ Price\ Previous\ Month}\right)$$

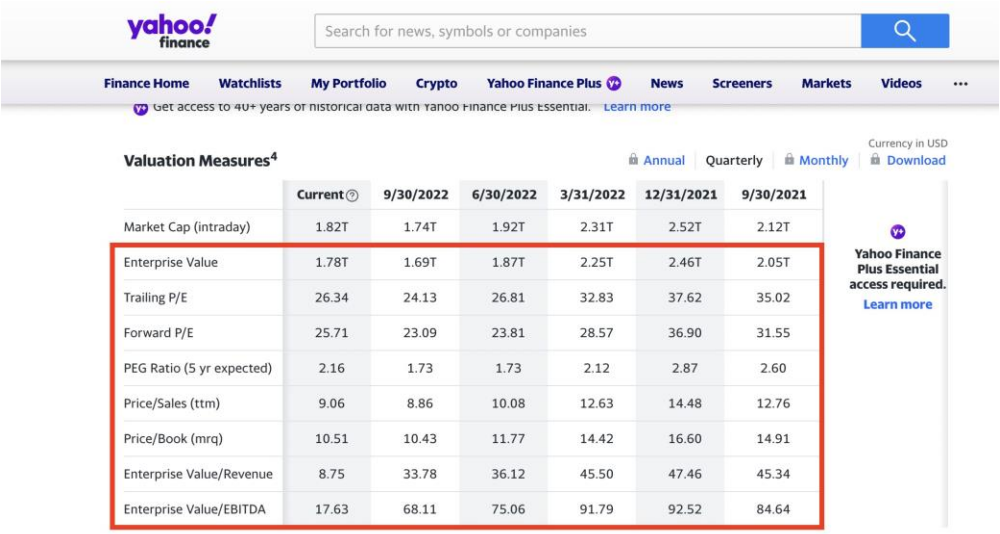
Closing price Current Month/Closing Price Previous Month = Closing price on the last trading day of the respective months

Company Names	Financial Ratio for each Company
• Apple (AAPL) • Microsoft (MSFT) • Google (GOOG) • Amazon (AMZN) • Meta (META) • AMD (AMD) • Nvidia Corporation (NVDA) • PepsiCo (PEP)	• Trailing P/E • Forward P/E (EPS for next 12 Months) • PEG Ratio (5 yr expected) • Price/Sales (ttm) • Price/Book (mrq) • Enterprise Value/Revenue • Enterprise Value/EBITDA

• Tesla (TSLA) • Netflix (NFLX)	

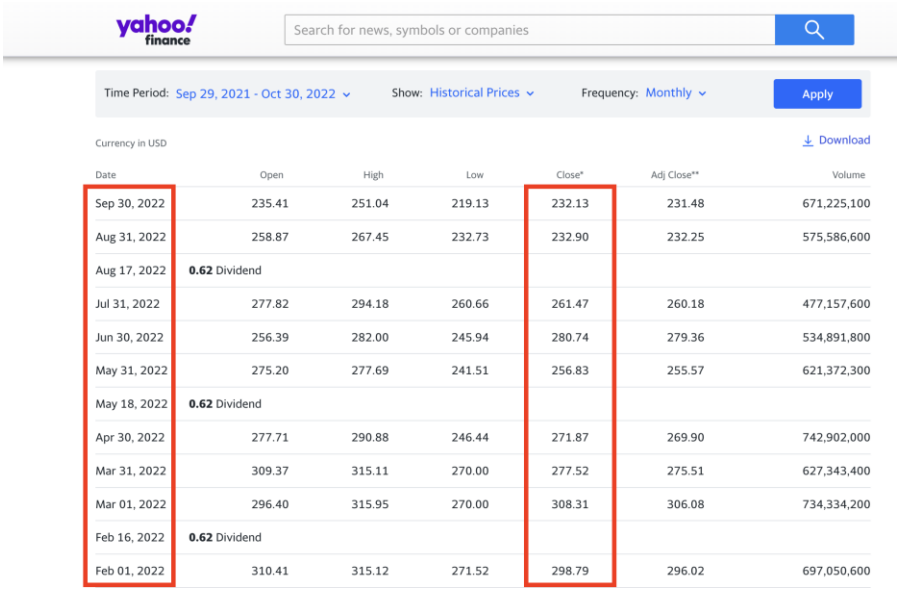
Yahoo Finance Data Scraped

Financial Ratios



	Current	9/30/2022	6/30/2022	3/31/2022	12/31/2021	9/30/2021
Market Cap (intraday)	1.82T	1.74T	1.92T	2.31T	2.52T	2.12T
Enterprise Value	1.78T	1.69T	1.87T	2.25T	2.46T	2.05T
Trailing P/E	26.34	24.13	26.81	32.83	37.62	35.02
Forward P/E	25.71	23.09	23.81	28.57	36.90	31.55
PEG Ratio (5 yr expected)	2.16	1.73	1.73	2.12	2.87	2.60
Price/Sales (ttm)	9.06	8.86	10.08	12.63	14.48	12.76
Price/Book (mrq)	10.51	10.43	11.77	14.42	16.60	14.91
Enterprise Value/Revenue	8.75	33.78	36.12	45.50	47.46	45.34
Enterprise Value/EBITDA	17.63	68.11	75.06	91.79	92.52	84.64

Monthly Closing Prices



Date	Open	High	Low	Close*	Adj Close**	Volume
Sep 30, 2022	235.41	251.04	219.13	232.13	231.48	671,225,100
Aug 31, 2022	258.87	267.45	232.73	232.90	232.25	575,586,600
Aug 17, 2022	**0.62** Dividend					
Jul 31, 2022	277.82	294.18	260.66	261.47	260.18	477,157,600
Jun 30, 2022	256.39	282.00	245.94	280.74	279.36	534,891,800
May 31, 2022	275.20	277.69	241.51	256.83	255.57	621,372,300
May 18, 2022	**0.62** Dividend					
Apr 30, 2022	277.71	290.88	246.44	271.87	269.90	742,902,000
Mar 31, 2022	309.37	315.11	270.00	277.52	275.51	627,343,400
Mar 01, 2022	296.40	315.95	270.00	308.31	306.08	734,334,200
Feb 16, 2022	**0.62** Dividend					
Feb 01, 2022	310.41	315.12	271.52	298.79	296.02	697,050,600

Sentiment Data

Social Media data on which we based the Sentiment Analysis has been obtained from the social media subreddit forum Wall Street Bets. We used the PMAW API to scrape posts from Wall Street Bets forum (r/wallstreetbets). As discussed earlier, after acquiring the data we cleaned and organized it using Python. Afterwards, with the cleaned data, we performed the Sentiment Analysis using Vader.

Subreddit Posts collected per Company

Company	**Total Count of Subreddit Posted and Scrapped** (October 2021- October 2022)
Apple (AAPL)	104
Microsoft (MSFT)	55
Google (GOOG)	22
Amazon (AMZN)	103
Meta (META)	175
AMD (AMD)	152
Nvidia Corporation (NVDA)	110
PepsiCo (PEP)	5
Tesla (TSLA)	501
Netflix (NFLX)	63

System Implementation

Tools Used

The following are the tools we implemented to conduct our analysis:

1. API's to extract web data (PMAW and YFinance).
2. VADER Sentiment Analysis Module on Python to get sentiment scores for Reddit Posts.
3. Multiple Linear Regression Model from Sklearn Python Module for model implementation
4. MatplotLib and Pandas to Create Charts and Organize data into dataframes

Multiple Linear Regression Models Built

MLR 1: Predicting Monthly Returns Using Earning Ratios and Average Monthly Sentiment Score

Company	R-squared	Trailing P/E	Forward P/E (EPS for next 12 Months)	PEG Ratio (5 yr expected)	Price/Sales (ttm)	Price/Book (mrq)	Enterprise Value/Revenue	Enterprise Value/EBITDA	Average Monthly Sentiment Score
Apple (AAPL)	0.402	-8.163788 e+11	-1.226062e +12	-1.356535e +11	4.850303e+ 10	2.139960e+ 12	5.354123e+ 10	4.337362e+ 11	3.819529
Microsoft (MSFT)	0.436	1.486650e +12	6.496398e+ 12	1.875923e +11	-1.791533e +13	1.507282e+ 13	1.697353e+12	1.778856e+12	7.878697
Google (GOOG)	0.260	0.000678	0.020101	-0.1609	-0.014013	-0.051229	-0.111454	0.303017	-0.041064
Amazon (AMZN)	0.298	-0.62281	-0.059504	0.01714	-0.039681	-0.132044	0.003589	-0.067006	-2.379891
Meta (META)	0.493	5.559728e +12	-7.549611e +11	-2.390782e +12	-1.308447e +12	-1.493010e+ 12	9.955813e+12	3.460104e+12	30.532914
AMD (AMD)	0.263	2.823084e +13	3.933028e+ 12	4.873832e +12	-2.051624e +13	-3.315674e+ 13	-5.229554e+13	2.112753e+13	34.44013
Nvidia Corporation (NVDA)	0.352	1.025144	0.234631	0.29465	0.162482	0.216959	-0.610946	0.069346	-7.089265

PepsiCo (PEP)	0.455	-5.583534e+11	1.958870e+11	-1.634925e+10	-1.807271e+10	1.289948e+12	1.592565e+11	-2.944553e+10	6.881883
Tesla (TSLA)	0.541	0.018881	-0.263981	-0.01284	-0.058262	-0.10485	-0.114326	0.031453	76.613576
Netflix (NFLX)	0.511	5.533463	-4.929515	0.189189	0.048354	0.878898	0.391091	-2.233529	-54.005351

MLR 2: Predicting Monthly Average Sentiment Score Using Earning Ratios and Average Monthly Returns

Company	R-squared	Monthly Returns	Trailing P/E	Forward P/E (EPS for next 12 Months)	PEG Ratio (5 yr expected)	Price/Sales (ttm)	Price/Book (mrq)	Enterprise Value/Revenue	Enterprise Value/EBITDA
Apple (AAPL)	0.567	0.001063	-2.134399e+11	-2.899086e+11	-3.042066e+10	1.635141e+10	5.323631e+11	-1.276393e+10	1.163229e+11
Microsoft (MSFT)	0.291	0.020185	5.274095e+09	-1.932409e+11	-1.646264e+10	4.814096e+11	4.060109e+11	-7.573894e+10	-4.175703e+10
Amazon (AMZN)	0.092	-0.002412	-0.024273	-0.005428	0.00027	-0.00134	-0.004251	0.001172	-0.000476
Meta (META)	0.606	0.012586	1.195871e+11	-1.694768e+10	-5.647508e+10	-1.543883e+10	-2.788710e+10	-2.282683e+11	7.924014e+10
AMD (AMD)	0.339	0.002187	-4.624642e+11	-6.029523e+10	-9.424898e+10	3.424970e+11	5.572710e+11	8.579918e+11	-3.543621e+11
Nvidia Corporation (NVDA)	0.204	-0.002856	0.023502	0.003949	0.008091	0.003099	0.004534	-0.021261	0.001235
Tesla (TSLA)	0.637	0.004348	0.000179	-0.002092	-0.000147	-0.00044	-0.0008	-0.001084	0.002544

Netflix (NFLX)	0.808	-0.004356	0.068504	-0.024367	0.00111	0.000908	0.012602	-0.000892	-0.052146

**For the second MLR model, Pepsi and Google had been excluded as these lacked data for the sentiment scores.*

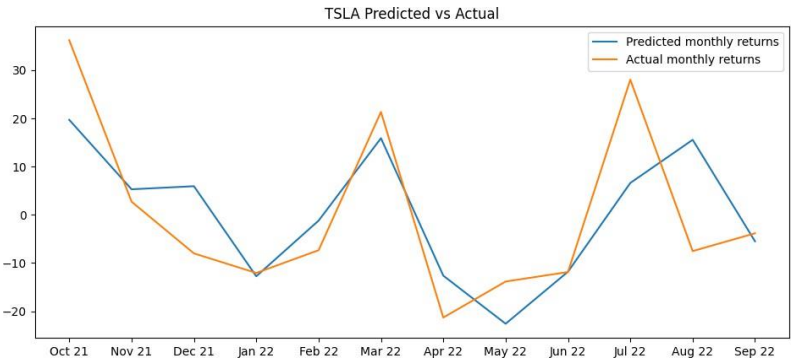
MLR 1: Significant Actual vs Predicted Value Plots

For an overview of all the plots, refer to the Appendix

Discussed below are the companies with multiple linear regression models that garnered the most significant results.

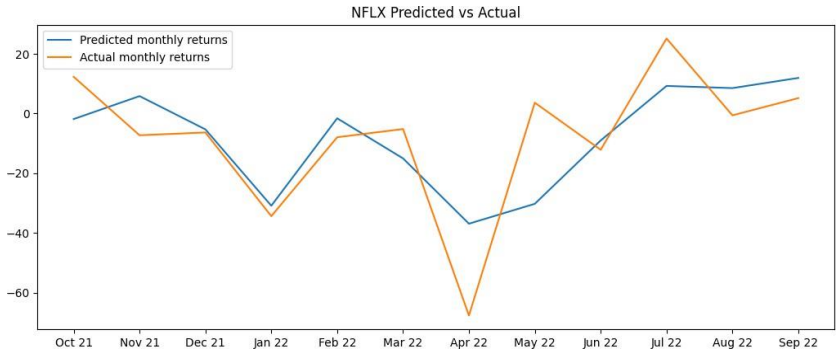
Tesla

- Highest R-squared at 0.541. This denotes that 54.1% of the change in the monthly returns is explained by the independent variables (Earning Ratios and Sentiment Scores)
- Significantly larger coefficient for Average Monthly Sentiment Scores compared to the other models. This indicates that the sentiment scores were influential in predicting the monthly returns
- The graph below shows the actual vs predicted values follow a similar trend line



Netflix

- Garnered an R-Squared value of 0.511, which is relatively close to Tesla's. Hence, 51.10% of the change in the monthly returns can be explained by the model
- The sentiment score also has a relatively large positive coefficient as compared to the rest of the models. Hence, it can be concluded that the Netflix monthly return prediction is heavily reliant on the sentiment scores as well
- Graph also shows a similar trend line for the predicted vs actuals

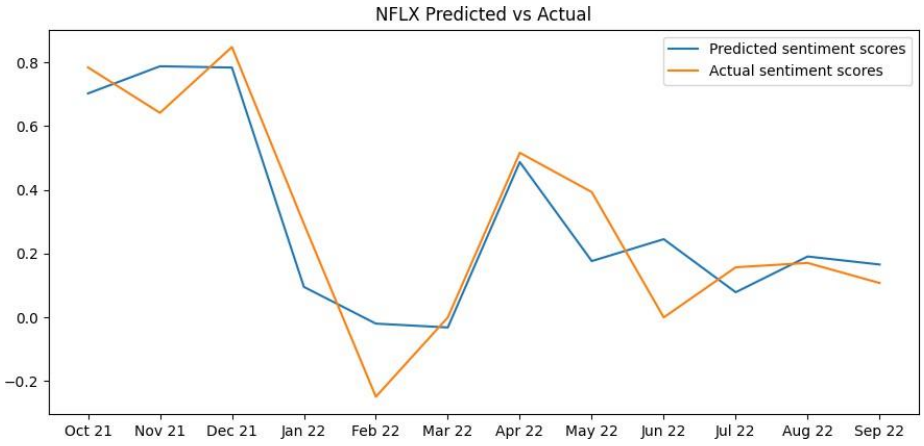


MLR 2: Significant Actual vs Predicted Value Plots

For an overview of all the plots, refer to the Appendix

Netflix

- Has the best R-Squared at a value of 0.808 - this shows that the sentiment scores can be explained by the change in the earnings ratios and the monthly return
- This provides insight as to whether Reddit posts on the wallstreetbets forum have a foundational financial data basis for it's sentiments
- The plot for the predicted vs actuals also follow a similar trend line



Tesla

- Produced a relatively high R-squared value as well at 0.637. Hence, 63.7% of the change in the sentiment scores can be explained by the input variables.
- The plot of the actual sentiment scores and the predicted sentiment scores also follow a similar trend line

- An indicator that affected the performance of both the Multiple Linear Regression Models of Tesla may be it's high volume in reddit posts. Additionally, the monthly sentiment scores may not have been highly contrasting - making the 'neutralization' of the scores through the monthly averaging minimal.

Evaluation

The business question initially posed was whether 1) the 'wallstreetbets' subreddit sentiment continues to impact the stock market and 2) whether the 'wallstreetbets' subreddit sentiment is founded on financial data. Based on the findings of this project, however, this assumption cannot be generalized across different companies. What the multiple linear regression models found is that there exists a relationship between monthly returns, earning ratios and sentiment from the 'wallstreetbet's subreddit in companies such as Tesla and Netflix, the same association may not be said for the rest of the top S&P companies. This may be due to factors such as lack of reddit posts on the forum, highly contrasting sentiments over the timeframe studied, or external factors that cannot be explained by the model. Hence, various features (i.e. ratios, sentiment from other forums/ platforms, etc) must be tested out to produce the best model for each individual company.

Given the performance of the models for the sentiment and monthly returns for Netflix and Tesla, it is recommended these companies, current investors, and concerned parties refrain from turning a blind eye to the sentiments on the 'wallstreetbets' subreddit. Alongside, a combination of solid understanding of financial stock data and awareness of the current accounts from traditional outlets, acknowledging social media sentiment (particularly on Reddit), may be value adding in understanding the overall value of these companies.

Future Direction

Conclusion

1. Collect more data

    Given the constraints of the PMAW API during the project period, there were limitations in collecting enough posts as it is currently being updated. However, when the API is once again fully functional, posts that go before the initial timeline set may be used.

    In addition to this, Yahoo Finance also has a paid Yahoo Finance Plus Add-On that allows users to collect more ratios. Hence, this can be used as well to increase the number of features to be used by the model. It is possible that these new features may improve the performance of the model.

2. Deploy Machine Learning Models and Add Features

    Machine Learning Models such as Artificial Neural Networks and CR&T (Classification and Regression Model) can be implemented as well as these work well with data that do not follow the assumptions for statistical data (i.e. normality, independence, similar variance). Hence, additional features that may be categorical can also be added to further improve the accuracy of the model predictions.

3. Model (Limitations)

The following point refer to the limitations of the current model's features and how these can be improved upon for future implementation:

- Missing data for certain months (Reddit Posts)

    For even some of the top S&P Companies, 'wallstreetbets' subreddit had an insufficient number of monthly posts. For future implementation, other subreddit forums may be scrapped as well. This may expand the study to analyze larger platforms and groups.

- Averaging the sentiment causes the highly contrasting sentiments to neutralize

    This may be improved by doing a daily analysis instead because sentiment may change drastically within a span of a month while it tends to stay similar within shorter timeframes.

- Using quarterly financial data to predict monthly values

    Due to limitations of the yfinance API, monthly and daily financial data is made available only to Yahoo Plus members. Hence, for future implementation, these more relevant data points may be used instead to

improve the model's prediction.

References

1. Python Packages Used
    a. Pandas
    b. Matplotlib
    c. Time
    d. Datetime
    e. RegEx
    f. NLTK VADER Sentiment
    g. Numpy
    h. SKLearn
2. API's Used:
    a. PMAW
    b. YFinance

Dataset

```
def mlr_sentiment_score(company): #this function will run an MLR with sentiment as the dependent variable

    ratios = pd.read_csv(fr'C:/Users/parth/Desktop/WA project/Historical/{company.upper()}_combined.csv', index_col=0)
    senti = pd.read_csv(fr'C:/Users/parth/Desktop/WA project/Sentiment files/{company}_monthly.csv', index_col=0).set_index(ratios.index)
```

Link to download datasets for the MLR:

Ratios:
https://drive.google.com/drive/folders/1go_3XDSDgZQ4XV94pFSAbcCv___-BgflY?usp=share_link

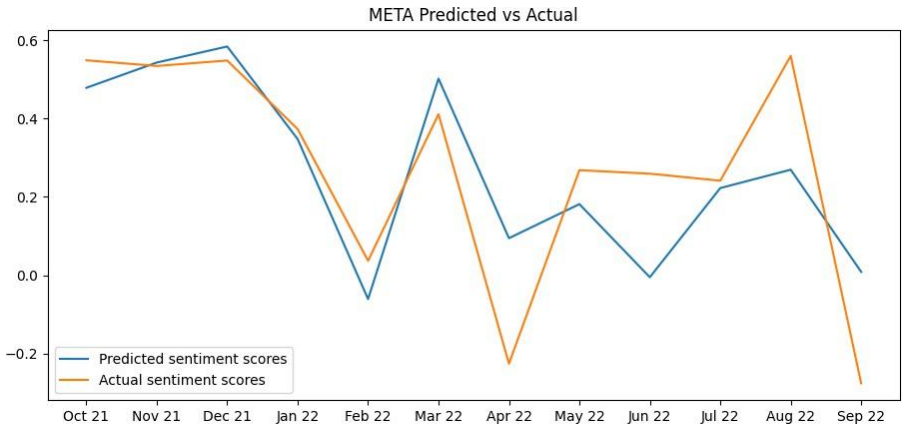
Sentiment:
https://drive.google.com/drive/folders/18Iye6yVEVJWGS___RNKD2E9-UtGzyyRKP?usp=share_link

Appendix

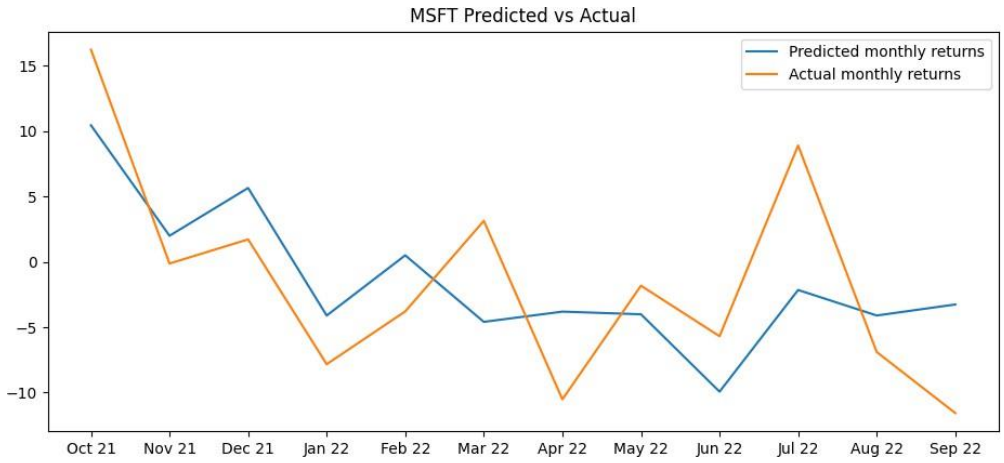
Charts: Actual vs Predicted Value Plots

MLR 1: Predicting Monthly Returns Using Earning Ratios and Average Monthly Sentiment Score

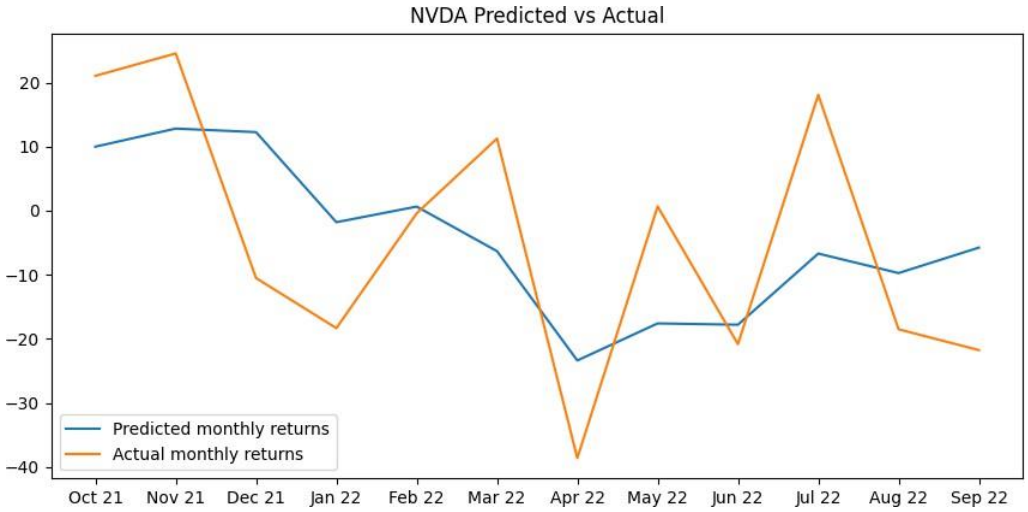
Meta



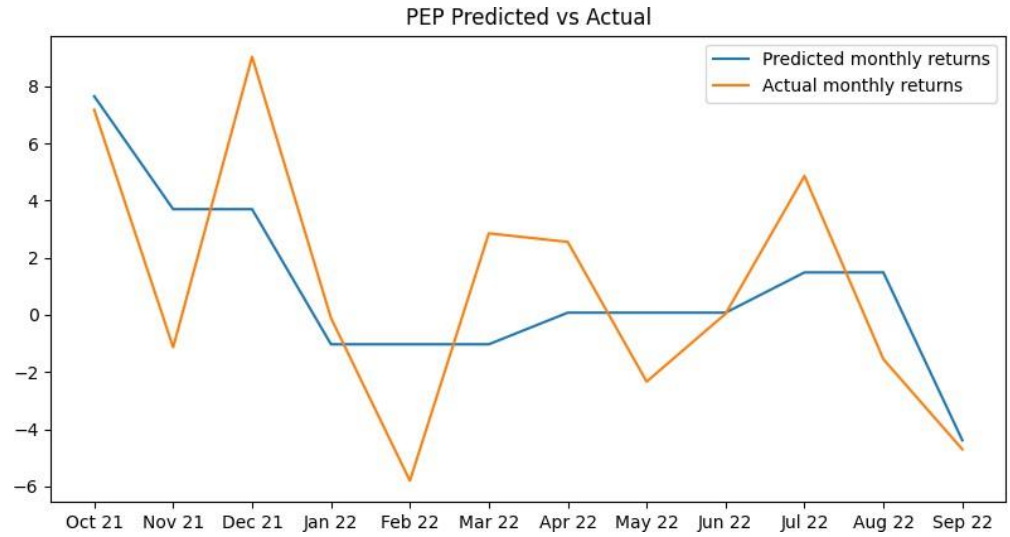
Microsoft



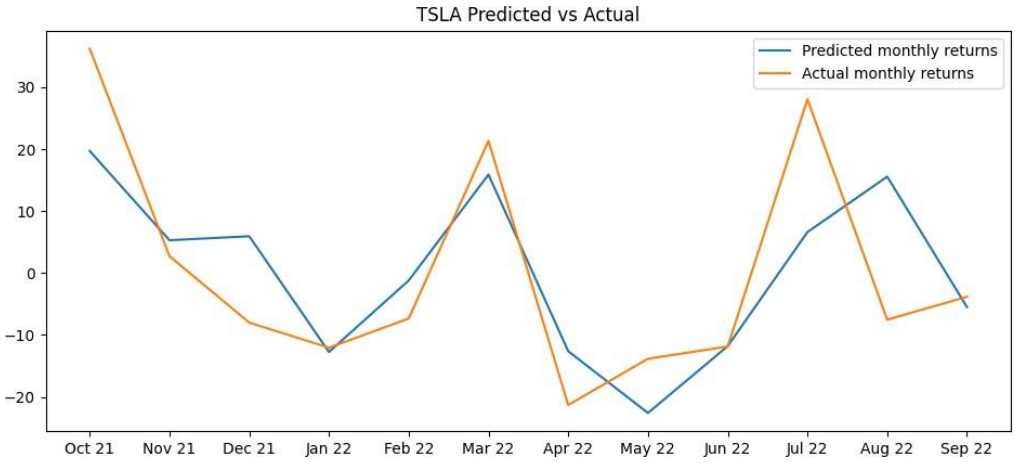
Nvidia



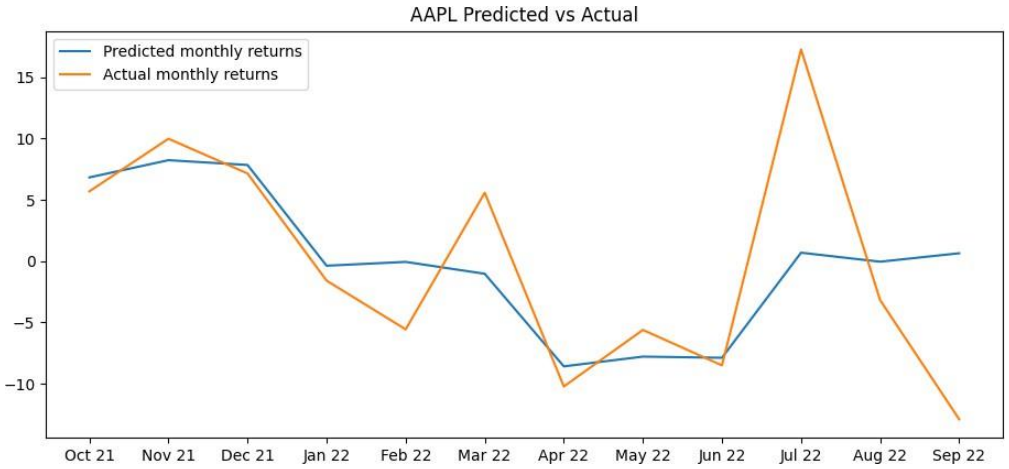
Pepsico



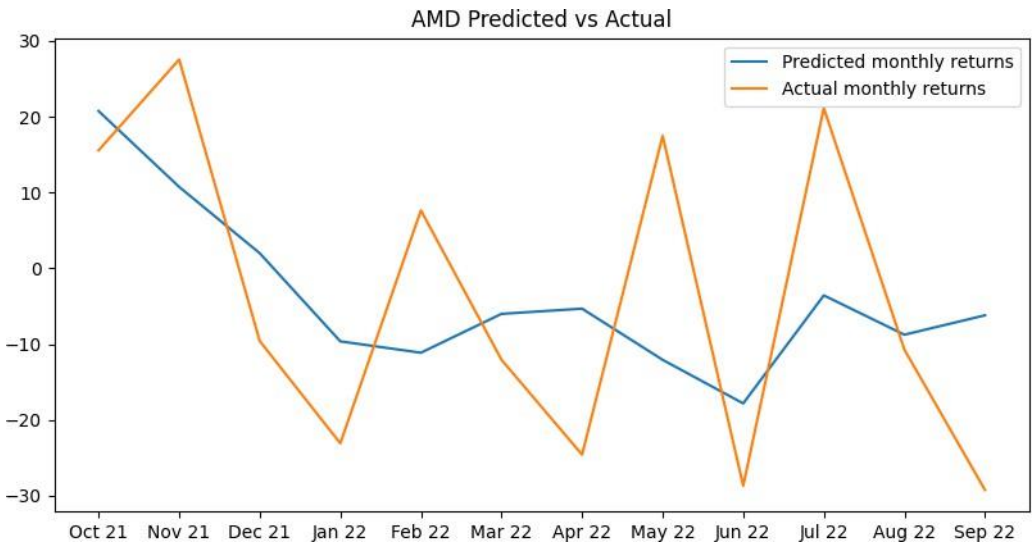
Tesla


TSLA Predicted vs Actual
Predicted monthly returns
Actual monthly returns
30
20
10
0
−10
−20
Oct 21
Nov 21
Dec 21
Jan 22
Feb 22
Mar 22
Apr 22
May 22
Jun 22
Jul 22
Aug 22
Sep 22

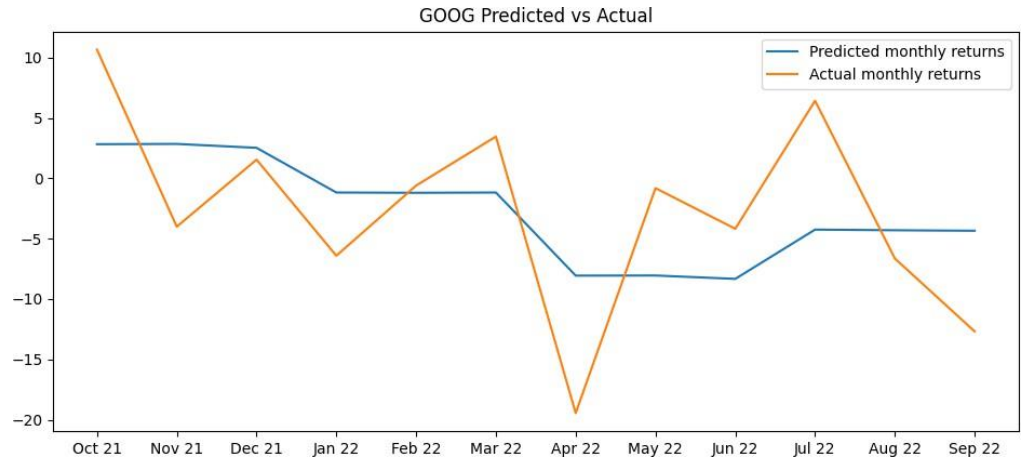
Apple


AAPL Predicted vs Actual
Predicted monthly returns
Actual monthly returns
15
10
5
0
−5
−10
Oct 21
Nov 21
Dec 21
Jan 22
Feb 22
Mar 22
Apr 22
May 22
Jun 22
Jul 22
Aug 22
Sep 22

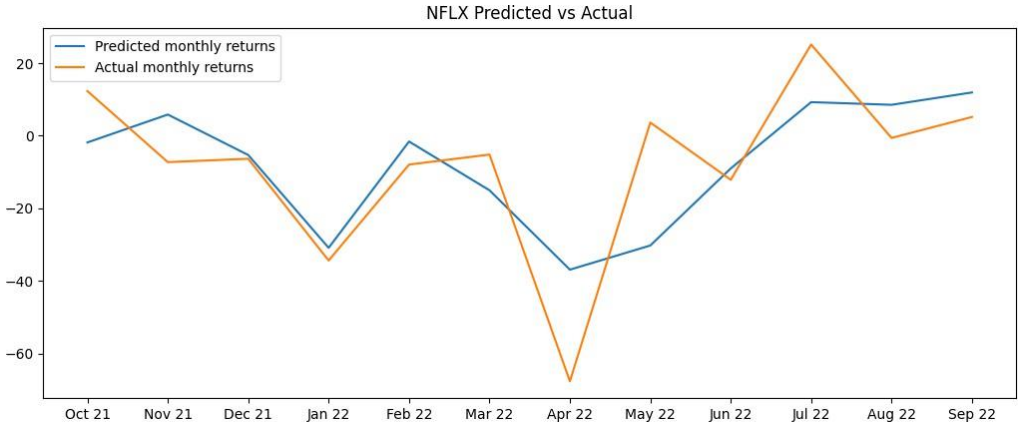
Advanced Micro Devices



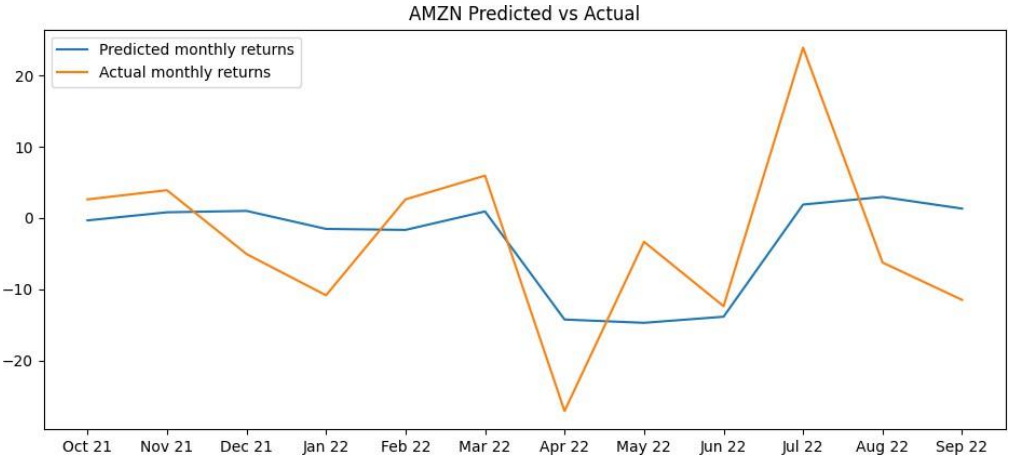
Google



Netflix



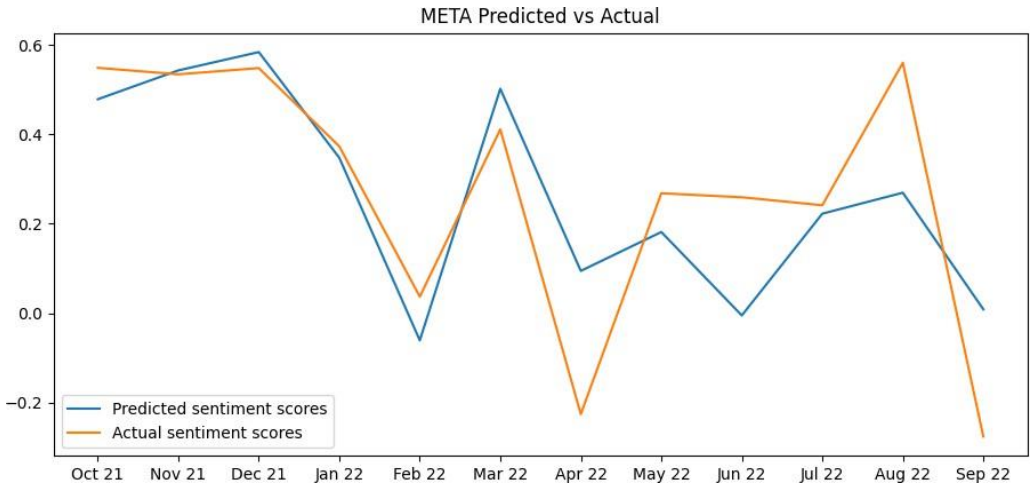
Amazon



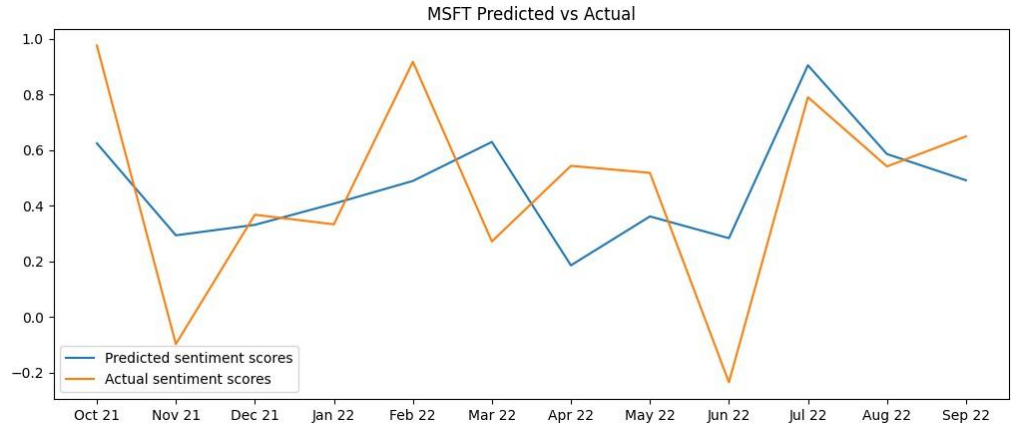
Charts (continued)

MLR 2: Predicting Monthly Average Sentiment Score Using Earning Ratios and Average Monthly Returns

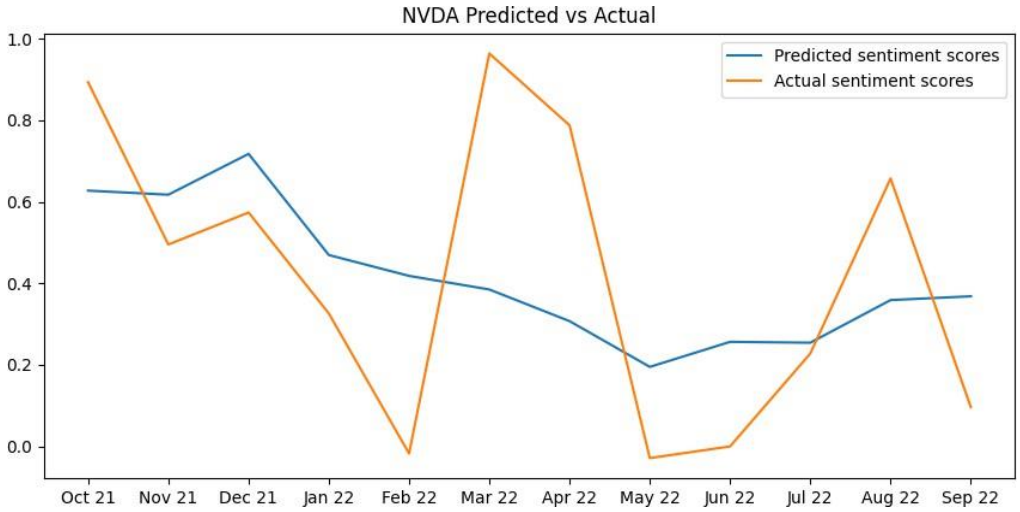
Meta



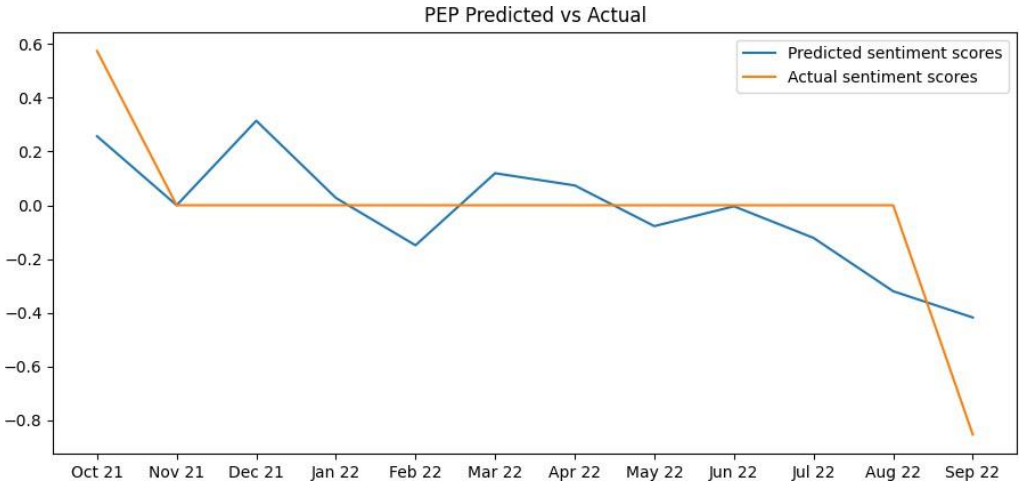
Microsoft



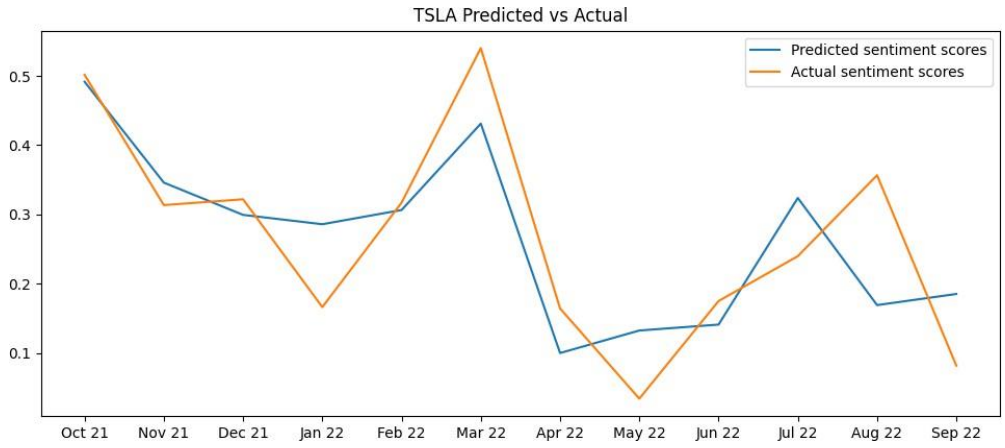
Nvidia



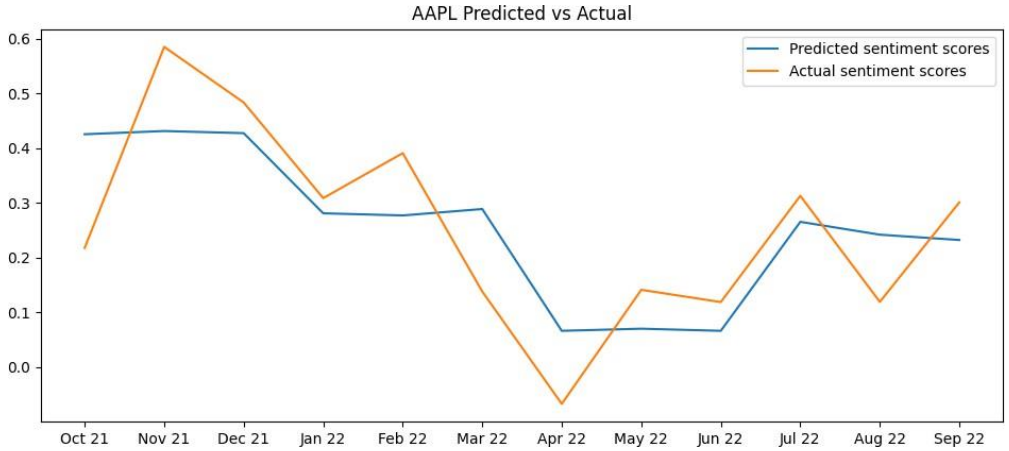
Pepsico



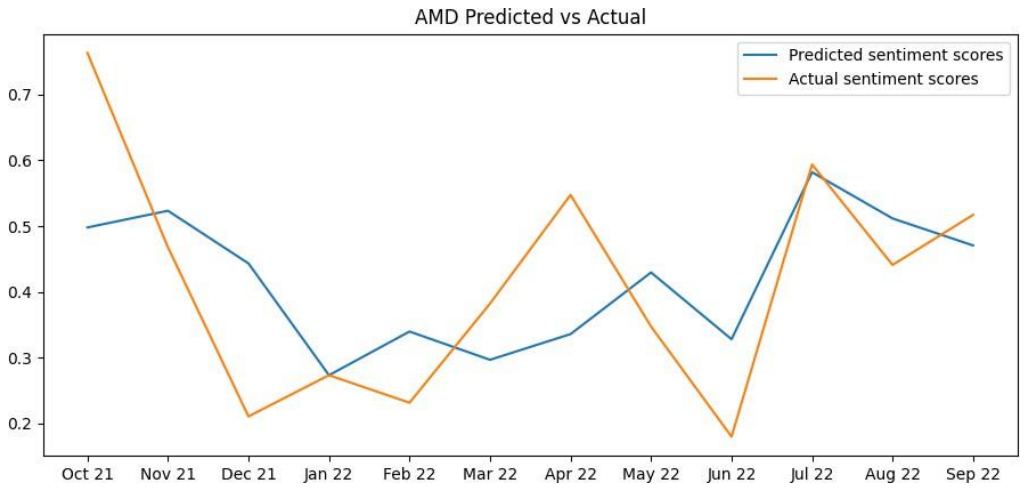
Tesla


TSLA Predicted vs Actual
Predicted sentiment scores
Actual sentiment scores
0.5
0.4
0.3
0.2
0.1
Oct 21
Nov 21
Dec 21
Jan 22
Feb 22
Mar 22
Apr 22
May 22
Jun 22
Jul 22
Aug 22
Sep 22

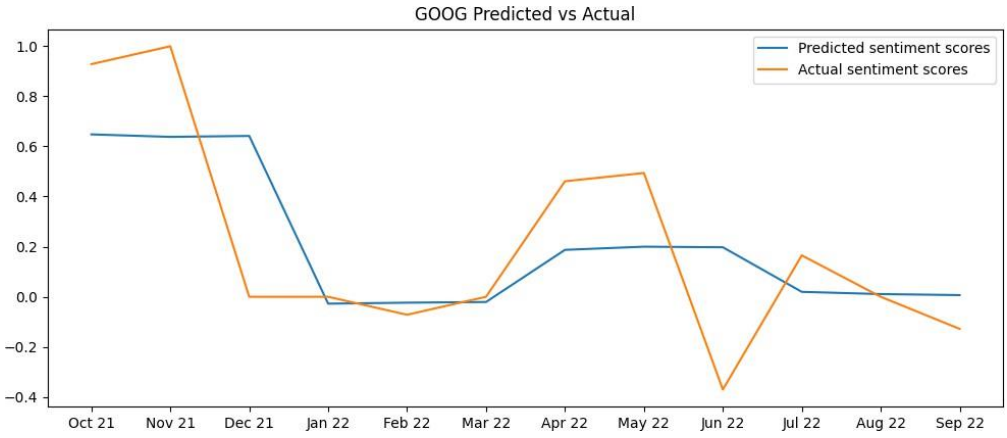
Apple


AAPL Predicted vs Actual
Predicted sentiment scores
Actual sentiment scores
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Oct 21
Nov 21
Dec 21
Jan 22
Feb 22
Mar 22
Apr 22
May 22
Jun 22
Jul 22
Aug 22
Sep 22

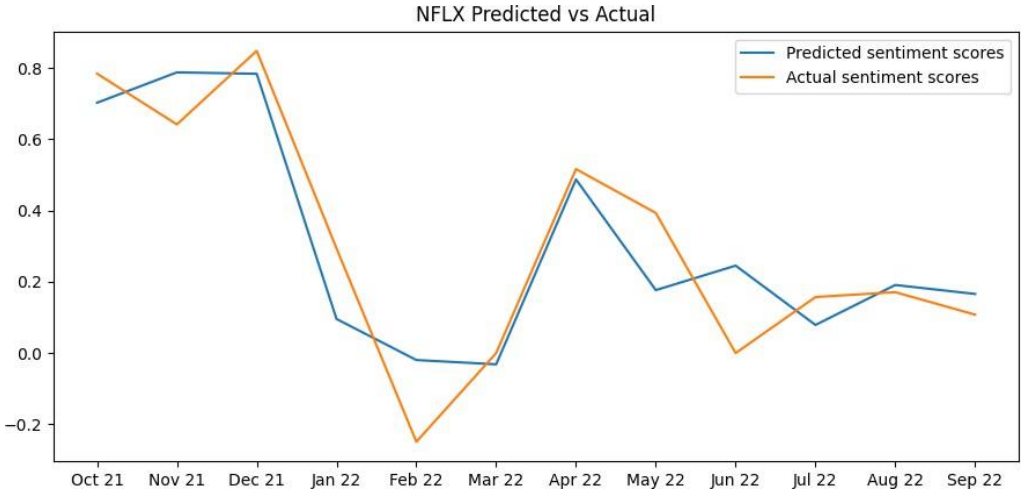
Advanced Micro Devices



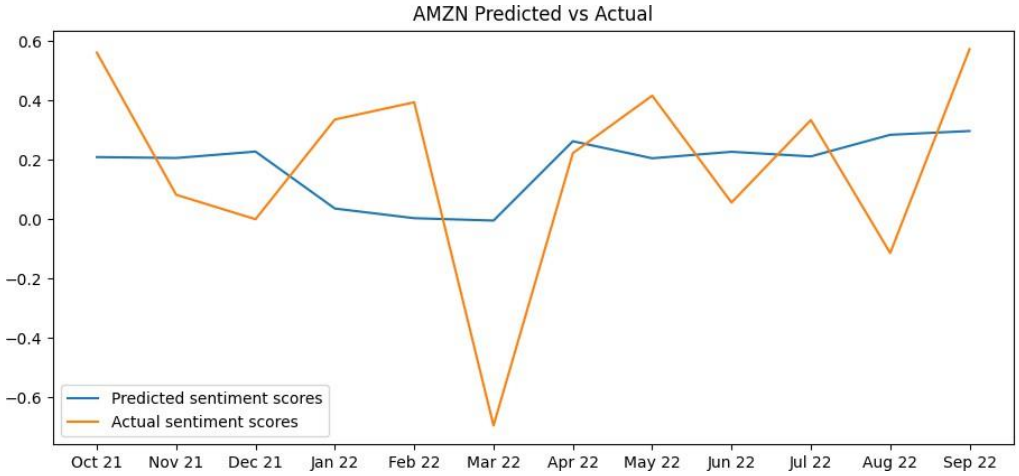
Google



Netflix



Amazon



Code:

Link to Code:
https://drive.google.com/drive/folders/1O5GT8b00-cBGt42MYCpoST69Jch9II9v?usp=share_link