

Predicting S&P500 Index Trend Using Sentiment Analysis

Andreas Lukita
A0221743M
e0559275@u.nus.edu

Ang Yun Ting Mabel
A0211001N
e0486556@u.nus.edu

Chua Yong Lun Joshua
A0201481U
e0412906@u.nus.edu

Han Shang Yu Nathaniel
A0217700R
e0543736@u.nus.edu

Luo Meng
A0185859E
e0319150@u.nus.edu

1. Description

1.1 Introduction and Motivation

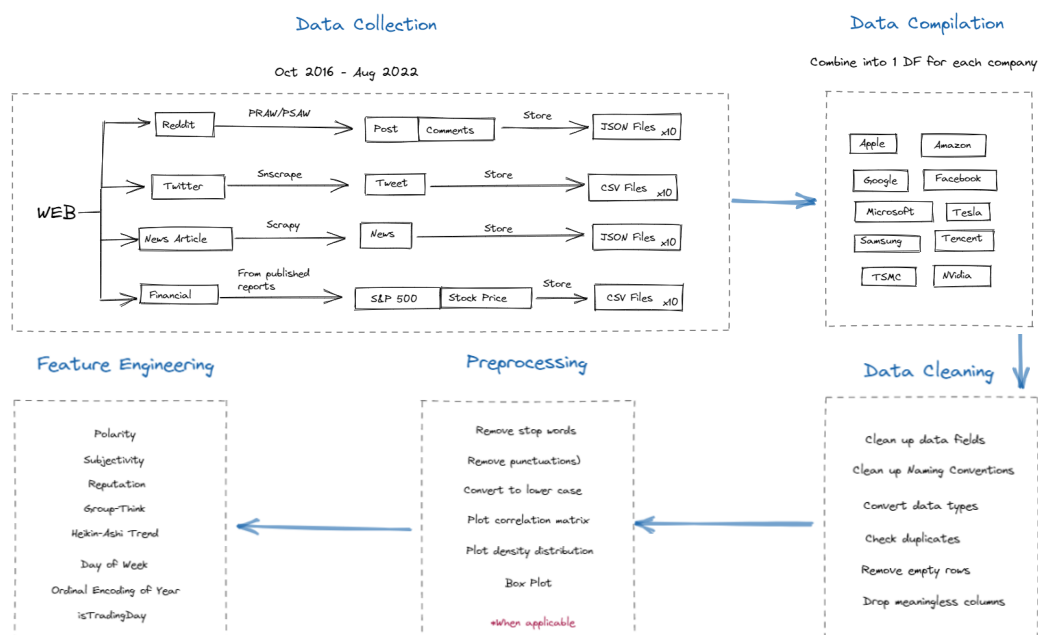
Predicting the prices of stocks remains popular amongst traders, institutional investors, and retail investors alike, as they race to find ways to beat the market to profit. Sentiment analysis is one of the methods used to predict stock market prices. More than half of the world now uses social media¹. Sentiment in the news seems to move markets as seen in examples such as a jump in Bitcoin's value by 20% after Elon Musk included "#bitcoin" to his Twitter bio², the 2020 Covid-19 stock market crash³ when fear, uncertainty and turmoil flooded the news, the 2021 WallStreetBets saga with GameStop (AMC)⁴ where Redditors teamed up to take on institutional investors by shorting the stock, and more recently in the U.S. CPI (Consumer Price Index) Reports.

1.2 Business Problem

This project aims to analyse the correlation between sentiments from various internet sources (Twitter, Reddit and Financial news) and the S&P 500 Index. Our group's hypothesis is that good sentiments observed in social media platforms as well as financial news sites would positively correlate with an upward trend in the S&P 500 Index. The Information Technology sector forms the top 28% of the S&P 500⁵, and according to a report by PricewaterhouseCoopers (PwC), technology companies represented 34% of the Global Top 100 in 2022 and 21% of the Global Top 100 5 years ago⁶. As such, our group will use a representative portfolio of the top 10 technology companies by market capitalization to approximate the S&P 500. The 10 companies will be *Apple, Microsoft, Google, Amazon, Tesla, Meta, Taiwan Semiconductor Manufacturing(TSM), Tencent, NVIDIA, and Samsung*.

2. Exploratory Data Analysis

It is important to conduct exploratory data analysis before starting any modelling. This section aims to gain deep insights into the various aspects of the data collected.



2.1 Data Collection

Three popular sources were identified to collect market sentiments - Reddit, Twitter and Financial News. For financial data, we turned to Yahoo Finance, Macrotrends, and the respective companies' annual/quarterly reports. In total, we have collected 6 years of data ranging from Oct 2016 to Aug 2022. The time taken to mine 2.33gb of raw data of these 10 companies is approximately 250 hours.

2.1.1 Reddit

Data was collected from Reddit using official PRAW (Python Reddit API Wrapper) and PSAW (Python Pushshift.io API Wrapper). PRAW was used to collect reddit posts, while PSAW was used to collect reddit comments with an API rate limit of 60 requests per minute. The data of each company was then saved into JSON file format to prevent script breakage resulting in entire loss of data, then converted to CSV for data manipulation.

2.1.2 Twitter

Data was collected from Twitter using Snscape. The two biggest advantages of using Snscape over Twitter API are that there would be no limit to scraping the number of tweets and no limit to accessing historical data.

For each day, the 1000 most recent english tweets were extracted using company-related keywords such as '#Microsoft' and '#MSFT'. To prevent script breakage, we saved each year of data of a specific company into a CSV file format.

2.1.3 News Articles

Data was collected from online news articles using Scrapy, a tool for downloading, cleaning and saving data from the web. We created an automated "spider" bot to crawl and scrape from www.investing.com. For each of the 10 companies, we specified an appropriate starting point taken from either the news or the analysis section of the website and saved it as a JSON file format. To prevent script breakage, we had many processes in place such as creating a virtual environment to isolate programs, setting delays to prevent overloading website servers, and limiting the number of requests sent in parallel to each remote server.

2.1.4 Financial Data

To ensure utmost accuracy, we collected financial data from published quarterly financial reports as well as market aggregators. This includes both the stock prices of the S&P 500 index, the individual share prices of the companies, as well as their financial metrics.

2.1.5 Legality

We collected data with the user's privacy and security in mind. All data was obtained via legal and ethical means and it was ensured that our processes were not in breach of any of the website's terms

of use, and that we were only collecting publicly available data. We also limited our crawler to a reasonable speed so that it would not cause the server to noticeably slow down.

2.2 Data Cleaning

2.2.1 Understanding Data Types

We collected both structured and unstructured data from web scraping. Misspellings, format errors, mixed data values and missing values were likely included. To handle inconsistent data, it is imperative that we detect them and apply appropriate techniques of data cleaning to either replace, modify or drop them.

2.2.2 Data Cleaning Techniques

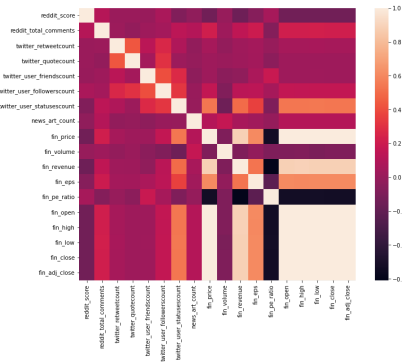
With this, we apply essential data cleaning techniques such as removing duplicates, dropping meaningless columns with only 1 unique value, converting data types such as date to *datetime* format and renaming column names to maintain good naming conventions.

Dealing with missing data is a delicate issue. For example, it is known that the stock market is closed on weekends and public holidays. If we set stock price to 0, we are assuming a share price of \$0, which would heavily skew the learning process of the machine. If we drop it, we are disregarding the fact that people are generally more active on social media on weekends. Hence, we believe that the best way to deal with this issue is to do a forward fill from the previous day when data is available.

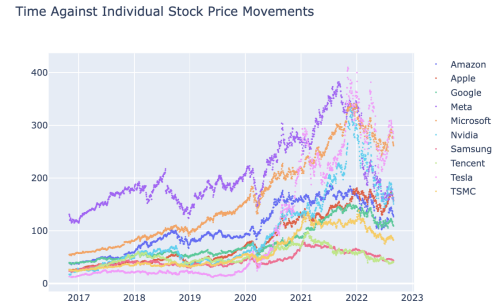
When dealing with text data, we perform NLP text preprocessing. We remove punctuations/stop words/emojis/URLs, convert to lowercase, as well as using fastText to identify whether the text is in English.

2.2.3 Understanding the Data

We plot a correlation matrix to understand how each feature moves with the stock price. The lighter the colour, the greater the positive correlation. With this matrix, we aim to find out exactly which feature is showing strong relationship signals with our target “stock price”. We also plot a distribution graph to generally understand the fluctuations in stock prices over the years. These trends will be noted when we do feature engineering in the next section.



Correlation Matrix



Distribution Graph

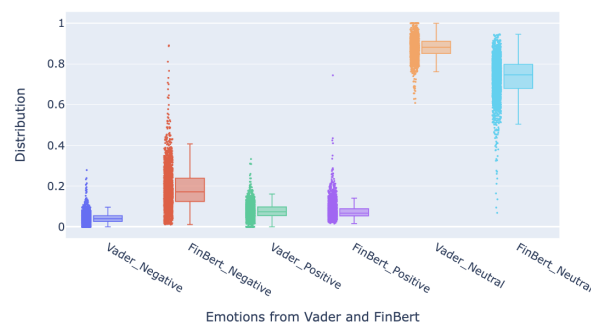
2.4 Feature Engineering and Scaling

2.4.1 Feature Engineering

Feature Engineering is one of the most important steps in creating a good machine learning model since it can directly influence the model performance. The dataset can be enriched by adding more features such as the following.

- Polarity:** This feature gives an overall sentiment score conveyed by a particular text/phrase. We use NLTK built-in pretrained social media sentiment analyzer VADER and pretrained NLP model FinBERT. VADER is best suited for language used in social media as it is able to map colloquialisms, emoticons, and slangs to an intensity value. When using VADER, we made a point to not clean the data extensively as common text pre-processing steps would cause VADER to lose some of its analytical strength. FinBERT is designed to handle NLP tasks in the financial domain. Similarly, we were careful not to clean the data extensively since context also plays an important role in the FinBERT's performance. Since they capture sentiment differently, our group has decided to use both. Since there are multiple posts from a social media platform for each day, we aggregated the sentiments by using the mean value to obtain the general sentiment.

Difference in Engineered Emotions using Different Packages



Polarity Analysis using Vader and Finbert

- Group-think:** This feature represents the groupthink phenomenon where a group of individuals reaches a consensus without critical reasoning. Here, posts with high interaction will have a

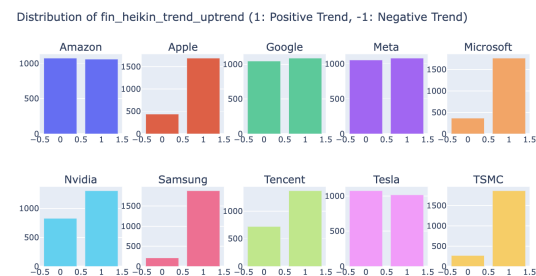
higher scoring. For Reddit, we take a weighted average of 40% polarity, 30% associated comments, 10% subjectivity, 10% score, 10% total comment. For Twitter, we take a weighted average of 50% likes count, 40% retweet count, 5% user followers count, 5% quote count.

- **Reputation:** This feature indicates how reputable our sources are. Sources with a high following or high comment ratio will have a higher scoring. In Reddit, we take the multiplication of score and total comments. In Twitter, we base it on the “user followers count”.
- **Subjectivity:** This feature indicates whether the content is factual or subjective. For all our sources, we used TextBlob because it has semantic labels to help with fine-grained analysis. Furthermore, Textblob calculates subjectivity by looking into one more parameter – intensity, which determines if a word modifies the next word.
- **Heikin-Ashi:** On top of the basic metrics such as company’s price-to-equity ratio and earnings-per-share, we want to better spot market trends by capturing trading patterns. This feature can be a good indicator of the share’s overall trend which approximates the generally positive or negative market sentiment of the company.
 - Heikin-Ashi Opening price = $(\text{open of previous bar} + \text{close of previous bar}) / 2$
 - Heikin-Ashi Closing price = $(\text{open} + \text{close} + \text{high} + \text{low of current bar}) / 4$

As seen below, the Heikin-Ashi distribution shows that it has a better ability to pick up trends compared to basic change in price.



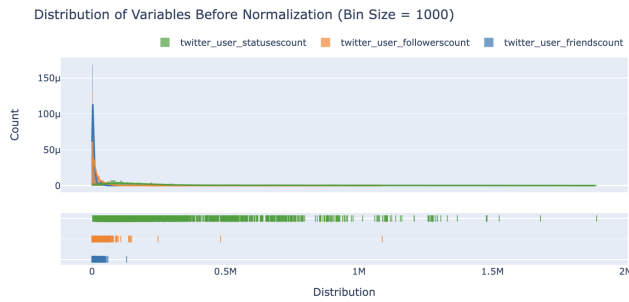
Change in Price Distribution



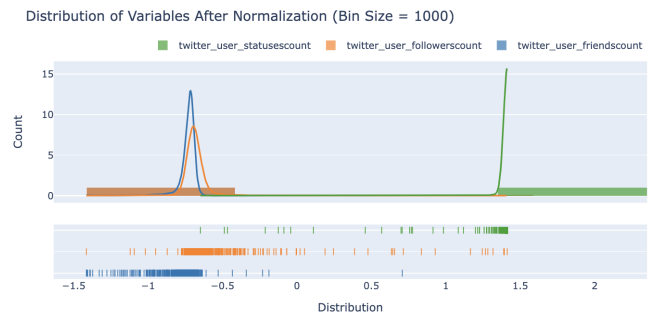
Heikin Ashi Distribution

- **Day of week:** This feature represents the day of week ranging from 1 to 7.
- **Ordinal encoding for year:** This feature helps to prevent skewing in the year of date.
- **isTradingDay:** This feature indicates a “1” if the day is trading day, and 0 otherwise, based on S&P 500 trading day records.

2.4.2 Feature Scaling



Variable Distribution Before Normalisation



Variable Distribution After Normalisation

Feature Scaling is a pre-processing technique that handles highly varying feature values by standardizing them into a fixed range. Depending on the relevancy, we applied several techniques such as log transformation and min-max scaling.

However, we made sure to keep the raw data because certain modelling packages such as PyCaret has an in-built normalization step when it is set to true.

3. Model

A simple algorithm may not make the perfect prediction when dealing with complex financial datasets. In our project, we split the modelling process into two phases. In phase 1, we want to boost the overall accuracy by building individual models for each of these 10 companies. In phase 2, we use an ensemble modelling technique to aggregate the outputs for predicting the final S&P 500 trend.

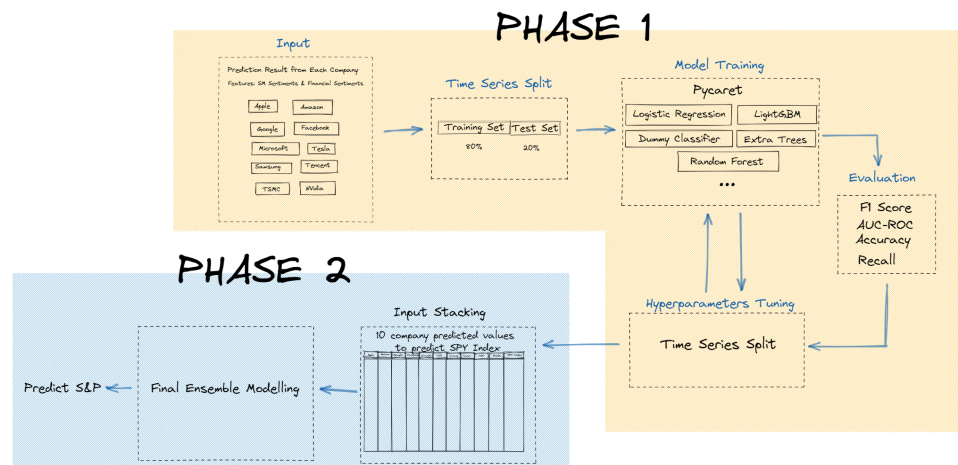
3.1 Past Experimentations

Our team attempted to build regression models to predict the share price for each company using sentiments from all the sources as input features, training the first 80% of the dataset and testing it on the last 20% of the dataset. Unfortunately, none of our modifications in feature engineering and selection could yield a satisfactory result, as the model evaluation metric (R-squared) for most cases are negative, which could indicate that there is presence of anomaly or outlier outside the range of our training dataset.

We believe that the anomaly could be related to the black swan events such as the Russia-Ukraine war which injected massive amounts of fear into the economy, and consecutive bearish news of a possible collapse of Evergrande Group, Credit Suisse, and a high-inflationary macro environment. These factors combined have led the stock market to face unprecedented downward pressures since November/December of 2021.

Therefore, we decided to pivot our effort to instead predict the movement of the stock market (upward or downward trend). This classification task would examine whether sentiments would value-add in the prediction of the stock market's movement as compared to a case where no sentiments were included.

3.2 Modelling Workflow



3.3 Phase 1 Predicting Stock Price Movement for each Company

3.3.1 Target Variable

Data that could confer even a marginal amount of leverage to an investor in the stock market is extremely time sensitive. The market price fluctuates due to speculation and fresh news swiftly after publication. As such, we place a window of effectiveness on our sentiments to study this phenomena.

3.3.2 Train Test Split

We split our dataset into training data and test data to simulate how well our model would perform with new data. No shuffling is done here because the ordering matters as we are mapping a sequence of features to the look ahead period of the closing price.

3.3.3 PyCaret Modelling



We use an open-source, low-code machine learning library called PyCaret to set up our base models. It allows us to experiment with many different models. Every set up in PyCaret is initialised differently depending on the company dataset. Some of the common parameters we have tuned include setting normalization and feature selection to true, setting feature selection threshold ranging from 0.3-0.8 and setting fold strategy to time-series.

For this classification task, we compare different metrics such as accuracy, precision, F1, and ROC-AUC score. As a baseline, we chose to optimise on accuracy because we want to determine the best performing model at identifying relationships and patterns between the features. The better our

model generalizes on unseen data, the better our predictions which in turn brings us closer to our objective.

3.3.4 Evaluation & Benchmarking

After finalizing the baseline model which gives us the highest accuracy, we run it on the test dataset.

	google_df_predictions = predict_model(final_model, data=google_df_test)							
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
	0 Logistic Regression	0.6842	0.7278	0.9575	0.6910	0.8027	0.1053	0.1548

Evaluation metrics when predicting on test data

On average, our accuracy score ranges from 58%-69% which we deemed as reasonable because of the uncertainty COVID-19 pandemic has brought onto the stock market. Because of the relatively small data set after feature engineering and the streamline nature of PyCaret, computation time is not an issue with most models developed within minutes. Computation time is only an issue when developing testing complicated models such as light gradient boost with run times of up to 15 minutes when testing multiple variations of it.

3.3.5 Selection of Data for S&P Ensembling

Based on the results of the testset performance, we selected 1 optimized model from the dataset without sentiment and 1 optimized model from the dataset with sentiment. As shown from the table below, for each company, we selected the final model, highlighted in yellow, based on 3 main criterias, accuracy, AUC score and prediction variance. In most cases, the final model was chosen based on accuracy and AUC score. In cases where the scores are close, we would select the model with a greater variance in prediction to boost predictive power for the later ensemble model.

Model	Accuracy	AUC	Recall	Prec.	F1	Feature Selection Threshold	Company	Sentiment	Predicted 1	Predicted 0
Logistic Regression	0.6711	0.7321	0.9527	0.6746	0.7899	0.4	Amazon	No	418	38
Logistic Regression	0.6711	0.7386	0.9155	0.6843	0.7832	0.6	Amazon	Yes	396	60
Logistic Regression	0.6842	0.7283	0.8599	0.7233	0.7857	0.5	Apple	No	365	91
Logistic Regression	0.6732	0.7172	0.7524	0.7599	0.7561	0.4	Apple	Yes	304	152
K Neighbors Classifier	0.682	0.7094	0.9869	0.6817	0.8064	0.15	Google	No	443	13
Logistic Regression	0.6842	0.7278	0.9575	0.691	0.8027	0.9	Google	Yes	424	32
Logistic Regression	0.6689	0.7478	0.9729	0.6674	0.7917	0.5	Meta	No	430	26
Logistic Regression	0.6623	0.7433	0.8915	0.6831	0.7735	0.8	Meta	Yes	385	71
Logistic Regression	0.6557	0.6959	0.9966	0.6556	0.7909	0.3	Microsoft	No	453	3
Logistic Regression	0.6425	0.6891	0.9161	0.6642	0.7701	0.3	Microsoft	Yes	411	45
Logistic Regression	0.6535	0.7253	0.9236	0.6731	0.7787	0.3	Nvidia	No	413	43
Logistic Regression	0.6667	0.7506	1	0.6645	0.7984	0.4	Nvidia	Yes	445	11
Logistic Regression	0.5833	0.6264	0.6696	0.5969	0.6311	0.3	Samsung	No	258	174
Logistic Regression	0.588	0.6327	0.7217	0.5929	0.651	0.5	Samsung	Yes	280	152
Random Forest Classifier	0.6713	0.7017	0.6359	0.6866	0.6603	0.5	Tencent	No	201	231
Random Forest Classifier	0.6597	0.6762	0.6037	0.6823	0.6406	0.4	Tencent	Yes	192	240
Light Gradient Boosting Machine	0.6644	0.6142	0.9682	0.6683	0.7908	0.4	Tesla	No	410	22
Logistic Regression	0.6667	0.5805	0.9399	0.6768	0.787	0.7	Tesla	Yes	393	39
Light Gradient Boosting Machine	0.6338	0.7156	0.8136	0.6818	0.7419	0.35	TSMC	No	365	91
Logistic Regression	0.6425	0.6915	0.9119	0.6626	0.7675	0.1	TSMC	Yes	406	50

3.4 Phase 2 Predicting S&P 500 Price Movement

Using the result derived from Phase 1 on the predicted stock movement for each company, we are able to build our final model, which is optimised for Precision, to predict the trend of S&P 500 Index. The following shows the model evaluation with and without sentiments, the best being logistic regression as highlighted in yellow.

WITH Sentiments							WITHOUT Sentiments						
Model	THRESHOLD	ACCURACY	PRECISION	RECALL	F1SCORE	ROCAUC	Model	THRESHOLD	ACCURACY	PRECISION	RECALL	F1SCORE	ROCAUC
lr	0.500	0.681	0.685	0.893	0.775	0.618	lr	0.200	0.736	0.808	0.750	0.778	0.732
lr	0.600	0.681	0.690	0.875	0.772	0.623	dt	0.500	0.615	0.615	1.000	0.762	0.500
gbc	0.500	0.648	0.643	0.964	0.771	0.554	dt	0.800	0.615	0.615	1.000	0.762	0.500
lightgbm	0.500	0.637	0.639	0.946	0.763	0.545	dt	0.900	0.615	0.615	1.000	0.762	0.500
lr	0.700	0.670	0.686	0.857	0.762	0.614	lr	0.700	0.626	0.628	0.964	0.761	0.525

3.5 Logic to choose the best F1 Score

In the scenario where the buy signal is incorrect, it would represent an unrealised loss to the investor. And in the scenario where the sell signal is incorrect, it would represent an opportunity cost to the investor.

As it is harder for investors to sell, we can assume that even with an incorrect sell signal, less investors would act which would diminish its effectiveness. In addition, an opportunity cost is arguably less painful than an unrealised loss since the latter requires capital to already be deployed and exposed to market fluctuations and risk.

Therefore, it makes sense for us to predict the true positive buy signal since it carries more weight. Since F1 score takes into consideration both recall and precision, this will be our main criteria to determine the best model, which is Logistic Regression as seen in the table above.

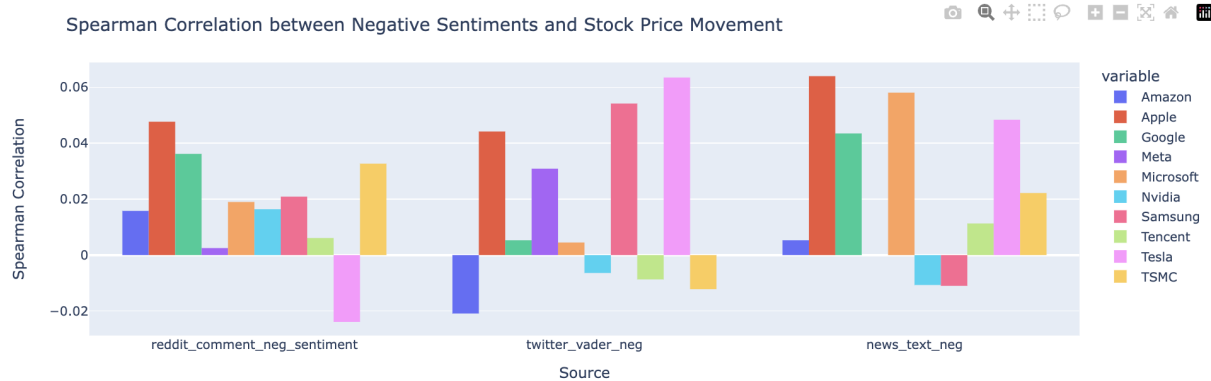
4. Findings

4.1 Features

When Twitter data are utilized and evaluated with TextBlob and VADER, the majority of tweets have a neutral sentiment. There are many highs and lows that are very close to zero, which suggests that utilizing Twitter data to anticipate stock prices may be unreliable, as emotion and stock movement could have very low correlation overall.

4.2 Usefulness of Sentiment Analysis in non-turbulent periods

As mentioned in section 3, our testing data (20%) would be heavily affected by the turbulent macroeconomic events which would skew our result since our model is trained based on a much stable period (2016-2021). That being the case, looking at the outcomes of our Phase 2 of predicting the S&P 500 Index, we see that the models with and without sentiment are not too far apart. Our group believes that if the model is tested on a more stable period with proportionately more positive sentiments, it could perform better since it has decent performance on our test dataset.



From this chart, we can see that our finding may be supported. Looking at the overall period used in our dataset, there is a slight positive correlation between the negative sentiments and negative change in stock market prices.

4.3 Model Performance

Our model suggests that the effect of sentiments on the S&P 500 Index movement diminishes by the day. We attempted to shift the prediction window from 1 day to 3 days and observe that by the time we move beyond 3 days, the AUC value tends towards 0.5, which indicates that the model is as best as making random guesses. This could be linked to the time-sensitive nature of financial data as mentioned in section 3.3.1 – as sentiments are priced in by the market and subsequently loses its influence.

5. Limitations

From our regression analysis, even with feature selection and hyperparameter tuning, our models yielded unsatisfactory results and merely have marginal improvement. Assuming our methods of data collection were sound, we believe that the issue could lie with the quality of the raw data collected from social media sources. As users from social media sites such as Twitter and Reddit may have limited expertise in analysing a company's performance and are largely filled with gibberish at times, data collected from these sites may be limited in being able to quantify a company's performance in the stock market.

Another assumption being made is that the users are direct investors or consumers of the brand. Buzzing social media discussions generated may not directly result in stock price changes. The presence of confounders may be at play.

References

The Global State of Digital in July 2022 — DataReportal – Global Digital Insights. (2022). Retrieved 2 October 2022, from <https://datareportal.com/reports/digital-2022-july-global-statshot>

Bitcoin spikes 20% after Elon Musk adds #bitcoin to his Twitter bio. Retrieved 2 October 2022, from <https://www.cnbc.com/2021/01/29/bitcoin-spikes-20percent-after-elon-musk-adds-bitcoin-to-his-twitter-bio.html>

The Coronavirus Crash Of 2020, And The Investing Lesson It Taught Us. Retrieved 2 October 2022, from <https://www.forbes.com/sites/lizfrazierpeck/2021/02/11/the-coronavirus-crash-of-2020-and-the-investing-lesson-it-taught-us/?sh=3da0421146cf>

The GameStop, Reddit WallStreetBets short selling situation explained. Retrieved 2 October 2022, from <https://lanthorn.com/80122/news/the-gamestop-reddit-wallstreetbets-short-selling-situation-explained/>

⁵Where to Find a List of the Top Stocks in the S&P 500. Retrieved 2 October 2022, from <https://www.investopedia.com/ask/answers/08/find-stocks-in-sp500.asp>

⁶Global Top 100 Companies by Market Capitalisation. Retrieved 2 October 2022, from <https://www.pwc.com/gx/en/audit-services/publications/top100/pwc-global-top-100-companies-by-market-capitalisation-2022.pdf>