Apple Stock: The impact of Twitter sentiments daily closing price of shares

Group 5
(Jiajun Du, Nate Lim, Romauli Butarbutar, Victoriya Murga)

Executive Summary / Abstract

Due to the highly volatile nature of stocks which depends on diverse political and economic factors, political factors, investor sentiment, and many other factors, predicting stock market prices is hard and encourages interest among both analysts and researchers for a long time. Recent studies in sentiment analysis have shown that there is a strong correlation between the movement of stock prices and the publication of news articles. Investors' opinions towards financial markets have been affected by the vast amount of online information in the public domain, such as Wikipedia usage patterns, news stories from the mainstream media, and social media. As such, there is a need to analyze the effects of news sentiments on the stock price. Our project proposal would use non-quantifiable data, from social media news, namely tweets about the Apple stock, to predict its future stock trend with sentiment analysis. We assume that Tweets impact the stock market, we would study the relationship between tweets and stock trends. We will retrieve, extract, and analyze the effects of news sentiments on the Apple stock price in the market, developing a dictionary-based sentiment analysis model, and evaluating the model for gauging the effects of news sentiments on the Apple stock.

Statement of the Problem

Apple Inc. (AAPL) is the world's largest multinational technology company specializing in electronics and software. In August 2018, Apple became the first publicly traded U.S. company valued at over $1 trillion and two years later, the first $2 trillion valued company. The company's stock price is evident in its success and conveys an immense pool of fascinating data. AAPL's ticker symbol represents the most-watched stock globally and is based on market capitalization - one of the largest companies in the world.

Predicting the AAPL stock price and its highly volatile trends is dependent on many variables. Hence, researchers constantly study investor behavior and the economic environment to predict stock's next moves and plan strategic activities accordingly.

Market sentiments witness different types of patterns with time, namely bullish, bearish and symmetric patterns. While a bullish or a bearish run may be an imperative consequence of an economic or geo-political vital event and therefore may be of short duration, a symmetric pattern is perceived to be the rule of vibrant economic activity, investor sentiment.

Technical and fundamental market analyses are both based on investor sentiment. Our research pursues the fundamental analysis technique to discover the future trend of AAPL stock by considering social media news such as tweets as the primary information and classifying it into positive and negative sentiments. We predict that if the overall tweet sentiment is positive, it is more likely to increase the stock price, and if the news sentiment is negative, we are likely to witness a significant decrease. We aim to build a model that predicts tweet polarity, affecting stock trends through supervised machine learning as classification and other text mining techniques to examine news polarity. We have taken the past two years of tweets and stock prices from Apple as data for our analysis.

Research Question

From $42.31 at the end of 2017 to $124.28 currently, a dramatic increase of 193.8%. Apple's stock price is sensitive to investor sentiment indicator changes. Contingent on the fact mentioned above, the question we are targeting in this research is:

Does a positive tweet sentiment increase the daily closing price of an Apple stock?

Hypotheses

In 2020, the Apple (NASDAQ: AAPL) stock increment was a little more than 60% (Forbes, 2020). Of course, Apple's Services business is a significant driver of its value. As such, it would be interesting to examine whether market sentiments, particularly Twitter sentiment, will significantly affect its stock price.

Null Hypothesis

Tweet sentiment indicators do not have any incremental effect on the average price of Apple stock.

Alternative Hypothesis

Tweet sentiment indicators have an incremental effect on the average price of Apple stock.

Literature Review

Multiple studies deal with stock return volatility from an investor's perspective. Within the scope of behavioral finance, studies have analyzed factors relating to investors' mentality, chief among others being - trends. Being an intelligent investor means - patience, discipline, and eagerness to learn; you must also harness your emotions and think for yourself. This kind of intelligence, Graham (2005), explains "is a trait more of the character than of the brain," suggesting for investors to be highly subjective to outside influence.

Apple stock is weighted at 2.408669 within the Dow Jones index (Gondo, 2021). By referring to statista.com, we found that as of December 2020, 20.4 percent of Apple's stock is held by Millennials in the United States. Millennials, also known as Generation Y, account for over 20 percent of the U.S. population. Social media and knowledge of financial markets for a substantial living - are more popular than ever. Unless Apple Inc understands the customers' requirements and adapts to emerging technologies in the market, and introduces new products and services, its business could be affected.

Market sentiment refers to the overall attitude of investors toward marketable security. It is the feeling of a market revealed through the price movement. Sentiment can be an indicator an investor can use to gain bullish or bearish sentiment insight into the stock market's mood. Extreme readings given by these indicators can indicate impending reversals. a Analysts may use indicators that include:

CBOE Volatility Index (VIX), New York Stock Exchange (NYSE) High/Low Indicator, NYSE 200-day Moving Average, Odd-Lot Trading Statistics, the Commitment of Traders Report and social media! Companies such as Facebook and Twitter are now very influential in the financial market. They are occasionally more beneficial for traders than the traditional analysis channels such as Wall Street Journal and Financial Times. As a sentiment trader, you may want to look for what is happening in social media constantly.

Grossman and Stiglitz (1980) developed a model under asymmetric information, featuring two rational investors. Informed and uninformed investors demonstrate that it is impossible to achieve efficient markets informationally.

Natural Language Processing (NLP) techniques refer to research presented by Nagar and Hahsler in their automated text mining based approach to aggregate news stories from various sources and create a News Corpus. The Corpus is filtered down to relevant sentences and analyzed using Natural Language Processing (NLP) techniques.

A sentiment metric, called NewsSentiment, utilizing the count of positive and negative polarity words, is proposed to measure the sentiment of the overall news corpus. They have used various open-source packages and tools to develop the news collection and aggregation engine and the sentiment evaluation engine. They also state that the time variation of NewsSentiment shows a strong correlation with the actual stock price movement.

Text mining based framework Yu et al present a text mining based framework to determine the sentiment of news articles and illustrate its impact on energy demand. News sentiment is quantified, presented as a time series, and compared with energy demand and price fluctuations.

Data Collection

We will be exploring two sets of data for our analysis. The first dataset will consist of Twitter tweets consolidated with emotions such as happiness, anger, sadness lexicons, cross-referencing with another dataset that pulls the daily closing price of Apple stock.

The lexicon-based approach for sentiment analysis will represent each review as a bag of words. We will then assign sentiment values from the dictionary to all positive and negative words or phrases within the message. A combining function, such as sum or average, will be applied to the final prediction, enabling us to assess the overall sentiment of the review.

Preparing the Data

The preparation process may include imputing missing values, recording variables, creating new variables, extracting new features, and restructuring the data. We will analyze the data to address the research questions.

Data sources

Data could be in different formats or reside in various locations.
We collect the data from:

- Kaggle
- quantmod package (soucre: 'Yahoo Finance', date: 2018-01-01 to 2018-12-31)
- Tweets about the Top Companies from 2015 to 2020

Variables to include in our analysis

In our analysis, we will only use variables post_date, cleaned_data from the data.csv dataset containing tweets from the company. Variables such as the number of comments, likes, and retweets do not play an essential role in our analysis. However, a more sophisticated analysis might depend on these variables as they reflect the quality of the tweets.

Determine whether to use derived variables

We will not use derived variables since we are conducting natural language processing and sentiment analysis on tweets.

The quality of our data

The data is not complete because the tweets about the companies do not capture the entire market sentiment. However, it’s sufficient to provide a general feeling of the market about particular stocks. The original dataset contains tweets about the company we are focusing on for the past five years, while we will only take data for 2018. The dataset does not contain missing values but has the raw form of texts that need further cleaning. We will use natural language processing libraries to remove stopwords, URL, punctuation, and other unnecessary symbols.

The accuracy of the model’s predictions is directly related to the variables we select and our data quality to complete the data, along with our analysis progress. We also will conduct a statistical test to find out the outliers, do the data cleansing and fill the missing value.

The steps of data analysis

We propose to perform 3 phase system designs in this project to classify tweets articles from twitter related to APPL stock for generating stock trend.

1. Analyze and Scoring the tweets This process including news collection, text preprocessing, news articles with its polarity score
2. Classify the tweets
3. Checking for relationship between news articles and APPL stock price data.

Tweets Collection

We collected data about the Top Companies from 2015 to 2020. This dataset contains over 3 million unique tweets with supporting information such as tweet id, author of the tweet, post date, the text body of the tweet, and the number of comments, likes, and retweets matched with the related companies Apple, Google, Microsoft, and Tesla. We also collected Apple price data for one year, from 1 January 2018 to 31 December 2018. Daily stock prices contain seven columns: Open, High, Low,Close, Adjusted Close prices, Volume, and Date. For integrity throughout the project, we considered the Adjusted Close price as the everyday stock price. This data was collected using quantmod package.

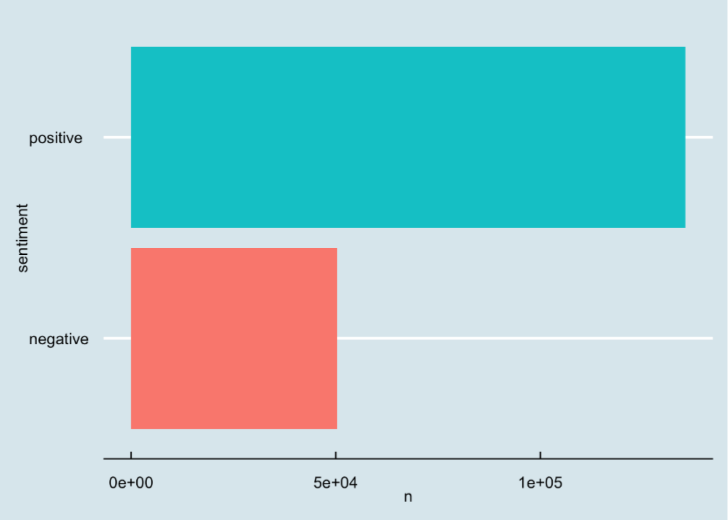
Pre Processing

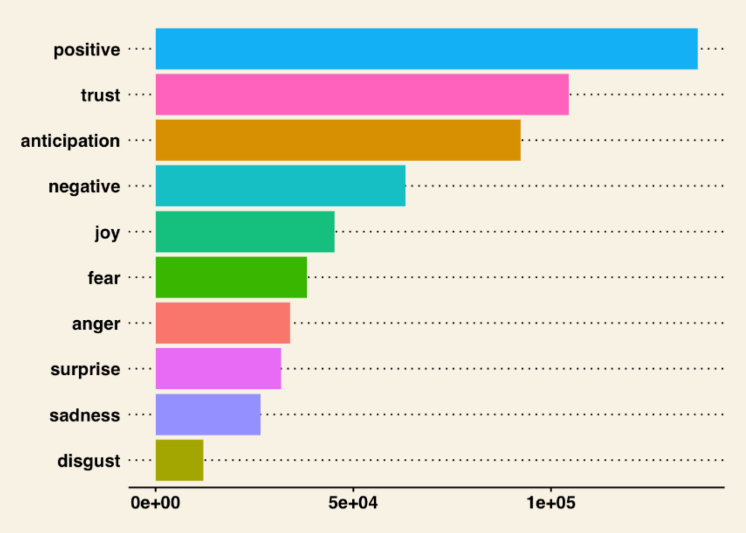
For raw data, we first merged tweets.csv with company_tweets.csv by "tweet_id". Next, we subsetted the data to only include the ticker symbol 'AAPL' and the year 2018. We also converted the type of date column to date data type. As the text is unstructured data, we cannot provide raw test data to the classifier as an input, so we need to tokenize the sentences into words to operate on word level. We need to drop the comments creating noise instead of meaning. In addition, text data may contain numbers, white spaces, tabs, punctuation characters, stop words, etc. We also need to clean data by removing all those words. We removed HTML tags from textual data.

For stock prices, we extracted the data using quantmod package from yahoo finance. We saved the volume and closing price of AAPL stocks as two new dataframes. We then used these two dataframes to join with our original data for future analysis.

Binary sentiment lexicons

We examined lexicons that classify tokens into two categories based on valence, positive or negative. As displayed, there are more positive sentiments than negative sentiments in the Apple tweet data.





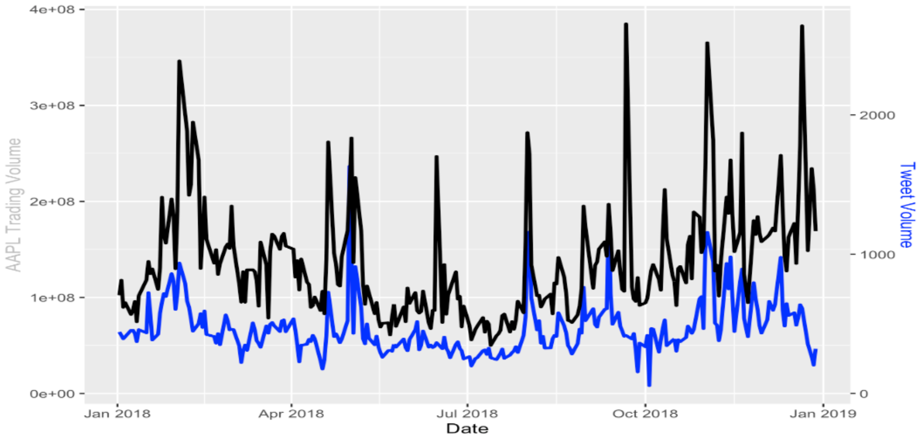
Emotion lexicons

The afinn lexicon scores each word based on the extent to which it is positive or negative. We examine the emotions expressed in the tweets by grouping based on post date and calculate the sentiment score Lexicons. We can see that the average of positive sentiment reach 88%.

Volume Analysis

We also performed analysis on the effect of tweet volume on the trading volume of the stocks.

First we collect the tweet volume per day using table function, bind the number of tweet with the stocks volume and adjust the data between the tweet volume and stocks volume by scaling the tweet volume.



Based on the above visualization, we can see that the trading volume peaks as tweet volume peaks, indicating a possible correlation. We will examine this relationship more closely later in the project.

Efforts to prepare the data

We needed to conduct data wrangling, cleaning, and tidying several times, and the challenging aspect of stock price prediction was making use of available data to make an informed decision. The dataset consisted of data from many companies and, if these data were to be processed manually, it would not be easy to decide in time.

After obtaining the data, we performed an initial data analysis using the summary or str function. Next, we determined which variables to include in the analysis. The original dataset contains five companies, where we filtered only Apple, which we were interested in analyzing. We will also subset the dataset for the year 2018. Lastly, we needed to remove English stop words, punctuation, URLs, etc., by establishing a corpus object, un-listing it, and binding it with our dataset.

Limitations

However, this conclusion is hindered by the limitations incurred by this analysis. Sentiment analysis using R's package assigns a sentiment score based on individual words. However, it is not sophisticated enough to evaluate the context of the string or sarcasm. Thus, sentiment analysis scores are limited.

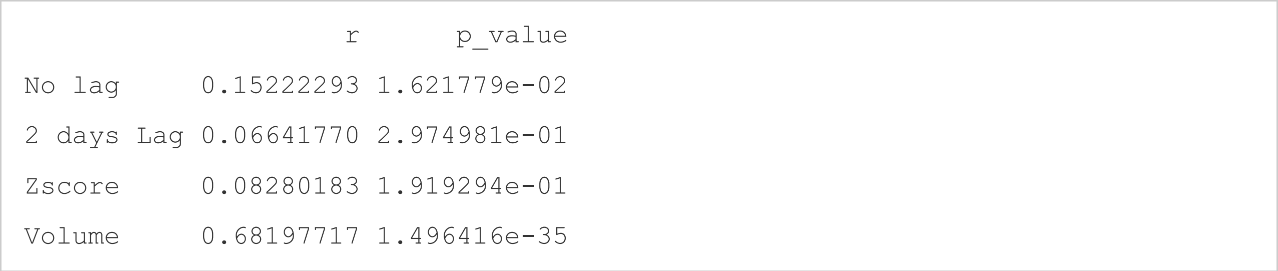
Further investigation is required to make a concrete deduction of the impact of market sentiments on the stock price. These include gathering more data over a larger time frame and conducting a more sophisticated sentiment analysis using additional sources (news articles, other social media channels, media, podcasts).

Recommendations

Based on the analysis results, if any predictions were to be made on APPL closing prices, they should be made within the same day of a tweet. Making a stock forecast 2-days after a tweet would not be meaningful, as concluded from our 2-day lag prediction analysis, with a high p-value (0.3).

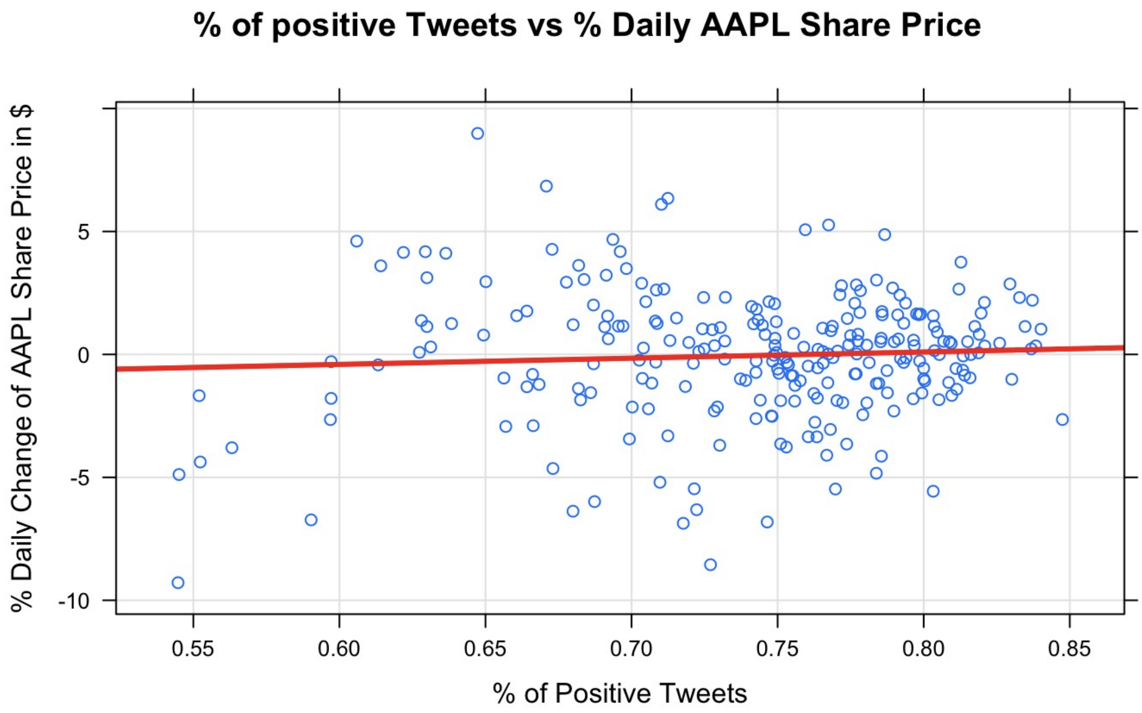
While Twitter continues to be an active source of gathering market sentiment, one should not solely rely on it as a predictor, especially on a long-term basis. We would propose to devise a future model that would analyze and combine the results of other social media resources such as Facebook posts, Instagram posts, Economic news reports, Marketing Time Series Analysis. It would be interesting to analyze if there were any confounding variables such as media coverage that would indirectly affect APPL's closing stock price.

Conclusion



```
                     r      p_value
No lag      0.15222293 1.621779e-02
2 days Lag  0.06641770 2.974981e-01
Zscore      0.08280183 1.919294e-01
Volume      0.68197717 1.496416e-35
```

Based on the data collected and analyzed, positive Twitter sentiments and the increase in AAPL stock prices have a significant correlation on the same day due to a low p-value of 0.016 <0.05. This indicates the sensitivity of stock prices to tweets. As mean Twitter sentiment score increases, so do the closing stock price. There was also no correlation of Zscore against Apple stock price change due to the high p-value of 0.19. An extremely low p-value of 1.50e-30 indicates a significant correlation between Apple trading volume and Apple Tweet volume. Finally, with the 2-day lag scenario, no significant correlation is seen as the p-value of 0.16 is greater than 0.05. This is evident and reconfirmed in the linear model as shown below.



References

[1] Anurag Nagar, Michael Hahsler, Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams, IPCSIT vol. XX (2012) IACSIT Press, Singapore

[2] W.B. Yu, B.R. Lea, and B. Guruswamy, A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting, International Journal of Electronic Business Management. 2011, 5(3): 211-224

[3] Dow Jones Industrial Average' (2003) Essential Investment, pp. 71–72. Available at: http://search.ebscohost.com.ezproxy.cul.columbia.edu/login.aspx?direct=true&AuthType=ip&db=bth&AN=26024016&site=bsi-live (Accessed: 13 June 2021).

[4] GRAHAM, B., & ZWEIG, J. (2005). The intelligent investor: a book of practical counsel. New York, Collins Business Essentials.

[5] Moseki, K. K., & KS, M. R. (2017). Analysing stock market data—Market sentiment approach and its measures. Cogent Economics & Finance, 5(1) doi:http://dx.doi.org/10.1080/23322039.2017.1367147