Stock Movement Prediction Using Twitter Sentiment Analysis

5205 Group Project Final Report
Professor. Lala

April 15th, 2021
Junyu Zhang
Chenning Zhang
Eric Yang

Abstract

The use of machine learning in the finance realm is becoming increasingly prevalent. Data collected on social media sights are viable search objects, providing valuable resources for text mining subjects. The goal of this project is to integrate machine learning methods to interpret twitter content, hoping to provide predictive power in the stock movement of companies, Apple and Tesla, inc. Specifically, we find word lists and examine the correlation between stock price and twitter sentiments.

Introduction

Stock market prediction is an active research area, aggregating much attention. The efficient market hypothesis assumes that the stock market is driven by new information, and stock prices follow a random walk pattern. Despite the widely accepted notion, traders and researchers have attempted to extract and identify patterns in the stock market that react against external stimuli.

This research study examines the hypothesis through a behavioral economics angle, questioning whether emotions and individual moods would likely affect decision-making when picking stocks.

Research Questions

Statement of the research problem or question(s) being addressed

1. **Are there correlations between tweet sentiments and stock prices within a week? How are they related?**
2. **Could tweets sentiments be instructive in predicting stock movement**？

Our project aims to determine the relationship between tweet sentiments and stock price change within a week and use this result as a tool to help design day trading and short-term trading strategy.

Dataset & Suitability

The data we used are tweets that mention AAPL and TSLA stock symbols from 04-07-2020 to 04-13-2020 and stock price data from 04-07-2020 to 04-013-2020. The datasets' length is only a week-long, and we used two AAPL and TSLA for specific reasons. The reason for such a short period is that tweets and stock prices are time-sensitive data. The relative insights offered by Twitter sentiment in the distant past have little to no effect on the present value. According to analysis 'Time-series-Honeywell-Stock-price-prediction' found on Github, distant historical sentiment data (over one week old) provided no insight into today's stock price.

We picked Apple and Tesla stocks because they are companies of popularity, garnering large attention. By combining their market capitalization and trading volume, Apple and Tesla's stock prices are unlikely to be skewed from outlying events.

Twitter Data

We extracted Tweets information using Twitter API to web scrape tweets and retweets relevant to the stocks. Specific company names might suggest multiple meanings and create noise during text analysis, and we limited the domain of queries to only the ticker symbol of the stock, hence AAPL for the Apple company.

We only gleaned Tweets in the English language as a limitation, while sentiments in foreign languages are not compatible with our sentiment lexicon QDAP.

Stock Price Data

Stock price data is extracted using the finance portal of Yahoo finance. The stock price data consists of the ticker (stock symbol), trading date, opening price, highest trading price during the day (high), lowest trading price (low), closing price (close), and volume (the number of shares traded throughout the day) and adjusted stock price.

Rationale Behind Techniques

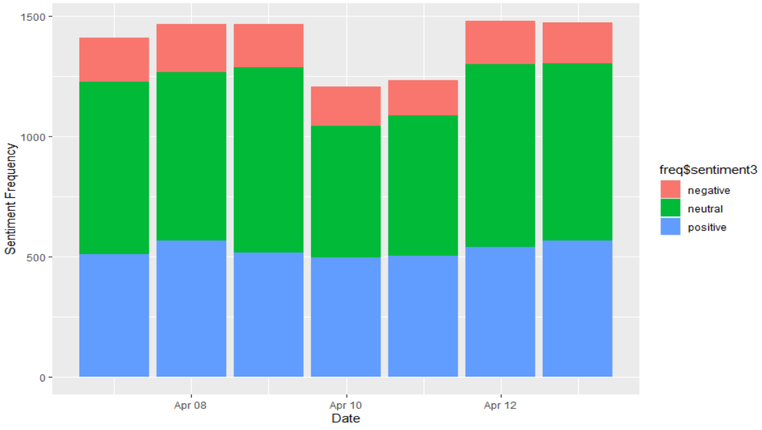
We have adopted several techniques in our analysis, primarily from text mining analysis with tools such as sentiment analysis, correlation, regression models, etc.
Sentiment analysis is key as the research topic revolves around the correlation between Twitter sentiments and stock price. Sentiment analysis is the process of determining whether a piece of word string is positive, negative, or neutral. It can derive insights or opinions from the speaker/writer and generally be served in predictive use cases indicating a person's feeling about a particular topic. The polarity of sentiment on the stock in social media is likely linked with a person's trading sentiment against the stock.

To best quantify the qualitative data into insights, we used the analyzeSentiment function provided by dplyr package, dividing words into sentiment directions (positive, negative, and neutral). By combining tweet sentiment, a mean sentiment score is derived per stock per day.

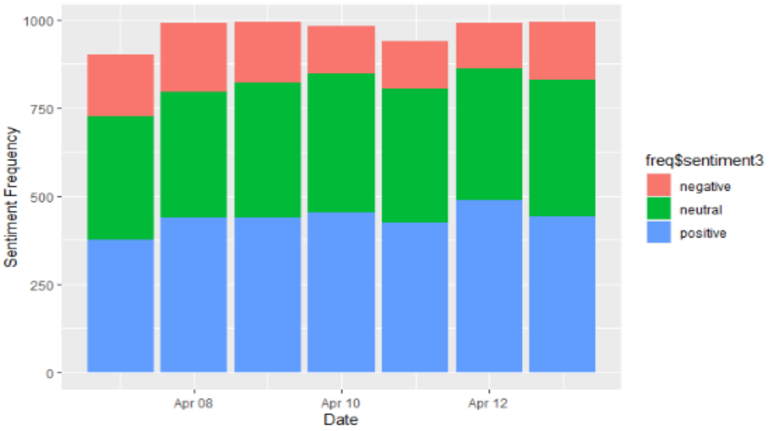
We also calculated the z-score of the stock price for the duration of observation. Z-score is an instructional metric trader often refers to as it captures the statistical measure of an observation's variability, helping traders determine market volatility.

To examine the relationship between tweet sentiments and stock price, we also included drawing correlations between tesla stock closing price and tweet sentiments.

Finally, we built regression models between the percentage of tweets (positive or negative) against stock closing price. We used generalized linear regression, given it is the optimal choice through visualization.



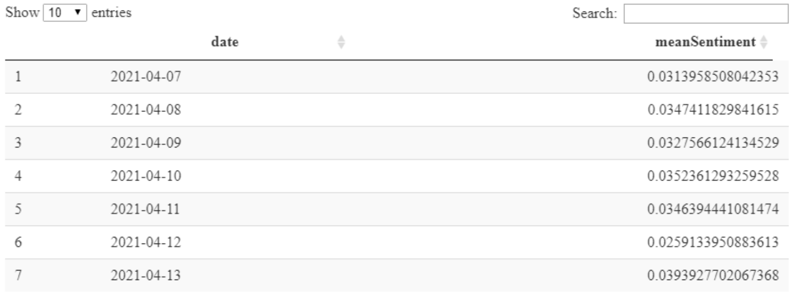
Apple sentiments frequency distribution



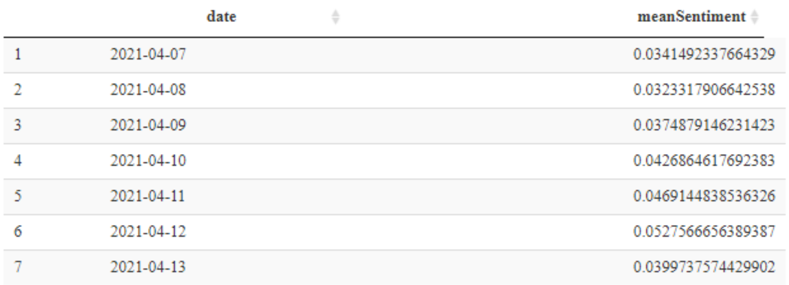
Tesla sentiments frequency distribution

Tesla and Apple are the industry leaders and share a history of impressive performance, so the distribution of sentiment frequency reflects their reputation. Positive(blue) and neutral(green) comments are above 70%. On the other hand, when comparing Apple and Tesla, we can see that Tesla has more negative sentiments than Apple does, implying its potential risks concerned by the public.

Apple sentiment scores



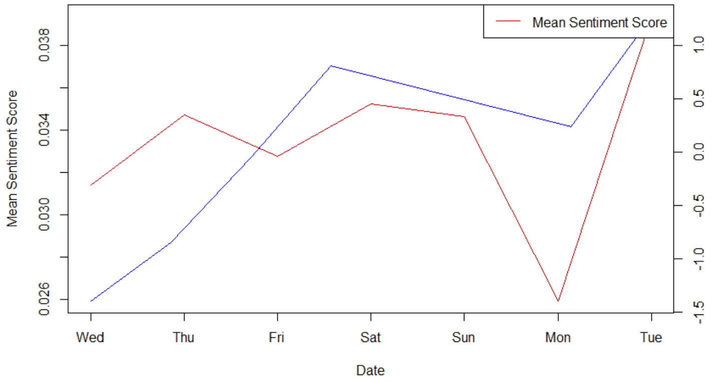
	date	meanSentiment
1	2021-04-07	0.0313958508042353
2	2021-04-08	0.0347411829841615
3	2021-04-09	0.0327566124134529
4	2021-04-10	0.0352361293259528
5	2021-04-11	0.0346394441081474
6	2021-04-12	0.0259133950883613
7	2021-04-13	0.0393927702067368



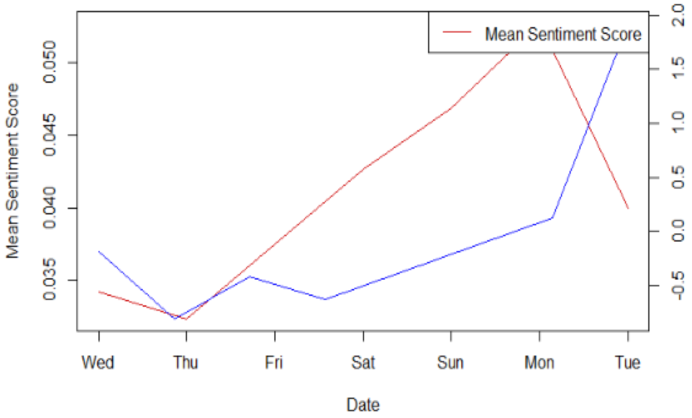
	date	meanSentiment
1	2021-04-07	0.0341492337664329
2	2021-04-08	0.0323317906642538
3	2021-04-09	0.0374879146231423
4	2021-04-10	0.0426864617692383
5	2021-04-11	0.0469144838536326
6	2021-04-12	0.0527566656389387
7	2021-04-13	0.0399737574429902

Tesla sentiment scores

This picture is an example of our final dataset in our sentiment analysis. Meansentiment column is the average score of sentiments of a particular day.



Apple sentiment score and stock prices



Tesla sentiment score and stock prices

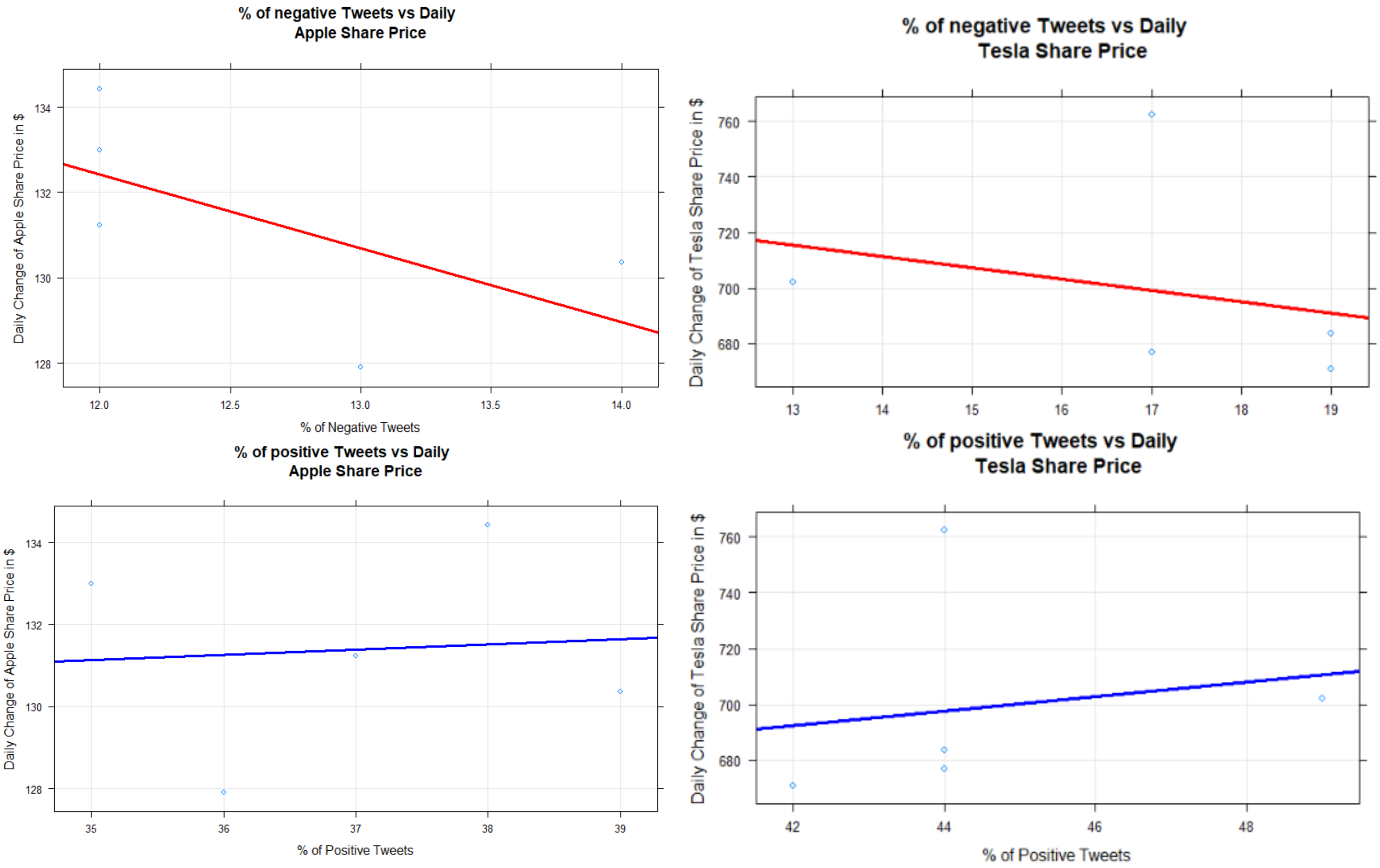
On the left, the graph contains a blue line(stock prices) and a red line(mean sentiment scores). We can see by simply eyeballing a certain level of correlation between apple stock price and its tweets sentiments. For example, the sentiment score went down on Saturday and Sunday and the apple stock price also dropped on Monday. Tesla demonstrated a similar correlation is well. Although it is not shown in the graph, we saw a drop in stock price on the following Wednesday(April 14 2020), matching the sentiment drop on Tuesday; These graphs support the possibility that tweets sentiments can be used as an indicator of stock movements.

Correlation table	Apple	Tesla
Negative Sentiments	-0.68	-0.27

Positive Sentiments	0.08	0.181

We can see a negative relationship between (percentage of negative tweets vs. stock price) from the scatterplot. This means as the percentage of negative tweets increase, the value of the stock decreases. We did not use positive sentiments and stock price correlations because the correlation was only 0.08 for Apple and 0.181 for Tesla, insufficient to prove correlational relationships.

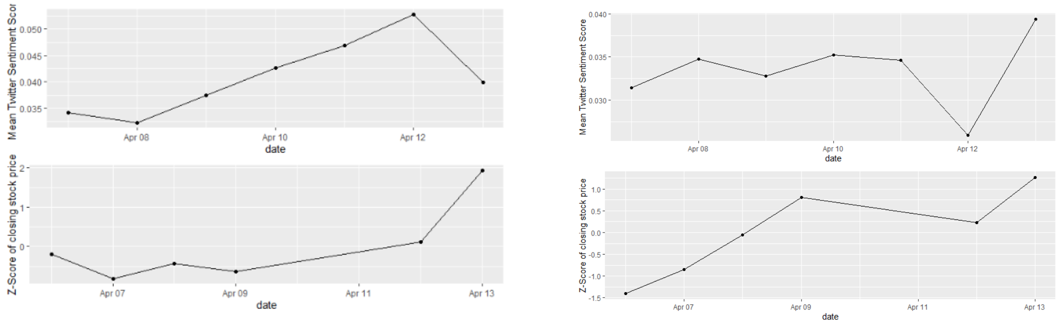
From the correlation coefficient, we can see a medium to significant negative correlation (-0.618) for Apple and a low to medium correlation for Tesla.



(AAPL) (TSLA)

The above graphs were generated from generalized linear regression models. Both Apple and Tesla stock prices demonstrate a positive correlation with the sentiments of tweets about them. Out of the four graphs, Tesla's stock price and percentage of negative sentiments showed a most obvious correlation; as the percentage of negative tweets increased from 13% to 19%, the stock price dropped around $30. Unfortunately, our data points are not sufficient for a generalized linear regression model to be counted as solid proof to our hypothesis but an indication of possibility.

(Tesla tweets mean sentiment + z-score)



Recommendations

We found a positive correlation between stock prices and tweet sentiments that mentioned the stock ticker through our analyses. The correlations between negative sentiments and drop in stock price are pronounced and were reflected in sentiment analysis, correlation analysis, and generalized linear regression model. However, the correlation between positive sentiments and a rise in stock price is little and neglectable. It means that positive sentiments cannot be used as an indicator for rising future stock prices. In saying so, we think that tweets' sentiments cannot predict stock price. Instead, traders can use tweets sentiments as an indicator of the stock drop to minimize risk in day trading or short-term trading strategies.

If we were to conduct the research another time, we would have retrieved more data points from Twitter and analyze tweet sentiments with a longer time horizon. This is likely to reduce variability and p-score examining the correlation. A longer time horizon will also neutralize the effects of stock movement caused by outlying macro-economic policies. For example, the implementation of the stimulus bill or announcement of products within a particular industry might lead to differing effects in stock price without much discussion on the individual ticker on Twitter.

We should also expand our stock data pool to include more stock tickers, ideally, around 10 to 15 stocks differing in industries, market capitalization, and varying performance patterns. The assumption states that the stock we pick would warrant enough discussion and generate sentiments that might have predictive power over next-day stock movement. So, more stock tickers included would reflect a broader range of discussion topics on Twitter.

We could have better categorized human emotions and quantified the descriptive vocabulary more intelligently. Implementing a more exhaustive dictionary to include various types of human emotions can be conducive to our study, given the complexity of emotions span beyond the narrow scope of “positive, negative or neutral.” If more wordlists are integrated, such as the Profile of Mood States questionnaire (POMS), which provides a scale of 1 to 5 on 65 words mapped onto six standard moods: tension, depression, anger, vigor, fatigue, and confusion. Aside from wordlist generation, we could also implement other tweet filtering measures to limit the types of tweets extracted to tweets expressing a feeling.

Finally, it is worth mentioning that our research has not taken into account many factors. Our dataset did not map real public sentiment as it only captures active English-speaking, twitter users’ behavior. We may hypothesize people’s emotions indeed pose effects on their investment decisions. However, There’s no direct correlation between people who invest in the stock market and twitter using behavior—although people’s investment decisions may be affected by the moods and comments of people surrounding them. All of these are explorable subjects open for future research.