Michael Barrera

Mikeb5@bu.edu

Project packet includes

Written report of findings  
Power point presentation slide deck  
Complete R code used for analysis  
Rdata file containing global variables (original tweet data)  
2 .txt documents containing data.corpus1 and data.corpus2

A stock market sentiment analysis using twitter

Boston University

CS 688 Final Project

Michael Barrera

Spring 2, 2019

Metropolitan College

A car is lined up in a parking lot

Description automatically generated

# A traditional industry's response to rapid change...

The auto industry has seen its fair share of disruptions lately. Many companies in this space have seen significant swings in their stock performance... some good, some bad. As in other industries, rapid innovations in technology have shifted consumer habits when it comes to buying behavior. Many companies in this traditionally conservative and change-averse industry have struggled to keep pace. The following presentation examines six automotive stocks and demonstrates the correlation between stock performance and Twitter sentiment.

## Methodology

Closing prices of various automotive stocks were examined on Monday, April 29 2019. Stocks with price changes of 1% of more over the previous close were examined further. Because of the volatility of individual stocks from day to day, I also examined the long-term performance of some of these stocks.

**Biggest Gainers and Losers:**

As mentioned before, I examined the closing prices of automotive stocks on Monday, April 29 2019 with the goal of identifying 3 stocks that increased more than 1% in stock price vs. the previous day’s close and 3 stocks that decreased more than 1% over the previous day’s close. However, due to the volatility of the stock market in general, I also considered the long-term (i.e.: previous quarter) performance of each stock. For example, Honda Motor Company’s stock price was actually a bigger loser that day than Sonic Automotive. But in further examining the long-term trends, Honda’s stock was quite stable over time, whereas Sonic’s had big swings and has been trending down for the past several months.

Additionally, I tried to pick companies that were similar to each other with regard their business model. CarMax and Sonic Automotive are both established retailers who operate customer-facing, brick and mortar retail locations nationwide. Ford and GM are both manufacturers who have been in business for over 100 years. Tesla and Carvana are newer companies and operate more like technology organizations rather than traditional retailers. I felt that analyzing companies with some commonalities would result in a more reliable and insightful dataset.

So after examining these different factors, I landed on the following 6 stocks:

Gainers:

* Carmax (KMX): + 2.62
* Ford Motor Company (F): +2.47%
* Carvana (CVNA): + 2.33%

Losers:

* Tesla (TSLA): -1.96%
* General Motor (GM): - 2.04%
* Sonic Automotive (SAH): -1.23%

**Search method:**

I chose to search tweets using the cashtag option because I wanted to ensure that the data I got was related to the sentiment around each company’s financial performance. Most of these companies actually communicate with their customers via Twitter and many company employees post things about their jobs and/or co-workers via Twitter. This would have created a lot of unnecessary noise in the dataset and may have skewed the results. Since the aim of my analysis was to explore the correlation between stock performance sentiment and actual stock price, I felt that other sentiment (unrelated to financial performance)

**Storing the data:**

After collecting the raw tweets for each individual stock, I consolidated the data into two tibbles; one for the gainer stocks and another for the losing stocks. Only the tweets themselves were kept, all other information like username, dates, etc were scrubbed. I added a custom column with ticker names in order to keep track of what tweets were associated with which companies. The data was then cleaned up via a simple, customer pre-processing function and stored in 2 separate corpora. Each corpus will be provided as part of this presentation packet. They can be saved on a local directory and uploaded to R via the system.file() and DirSource() functions. They can also be accessed by simply uploading the Rdata file I included which contains the entire global environment I used for the project.

## Analysis

**Most frequent terms:**

The corpus were processed into a term-document matrix and all terms were counted for frequency. A snippet of the results of that analysis is included below:

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

While these tables don’t reveal a ton of useful information, it does start to give an idea of certain words that will help determine the overall sentiment. Words like “boom”, Top 10”, and “high” come up often in the gainers data set. In addition, there are several other stock ticker symbols that some up often in each set, like “aapl” & “nflx”. While this may not be pertinent to our analysis, it does provide some insight into other stocks that investors in this space are interested in, so further analysis could be explored on this topic if one were so inclined.

**Word Clouds:**

The word clouds in this scenario are similar to the frequency tables illustrated above. They don’t paint a completely clear picture of the sentiment around gainers and losers, but they do reveal a little more information than the tables do:

A close up of text on a white background

Description automatically generated

A close up of text on a white background

Description automatically generated

Once again, certain words associated with a positive or negative sentiment are beginning to emerge here. In the first plot (gainers) we see words like “investing”, “boom”, “top”, “higher”, “raised”, “profit”, etc. In the second plot (losers) we see “short”, “misses”, etc. While the second word cloud is not as revealing as the first, you can still see that there don’t seem to be as many positive words that jump out at you immediately.

**Sentiment Analysis:**

A sentiment analysis on the complete set of tweets wraps things up nicely. Using the sentiment\_bing() function from Module 5, I was able to get a more conclusive numerical sentiment score for each corpus as well as extract sentiment-related words from each set of data. Here are those results:

*Gainers sentiment score:* ***18***

A screenshot of a cell phone

Description automatically generated

*Losers sentiment score:* ***-2***

A screenshot of a cell phone

Description automatically generated

Because the lists are quite long, I did not include them in their entirety. However, they are easily reproducible using the code included with the project packet. What we see from the sentiment analysis is that although there are some common positive (e.g.: “upgraded”, “top”, etc.) and negative (e.g.: “death”) in both data sets, the loser data set seems to have more instances of different negative words albeit at a lower volume which is why they didn’t stick out in the word cloud. So, the moral of the story is that math is always more reliable than the naked eye!

**Conclusion:**

The sentiment scores of 18 for the gaining stocks and -2 for the losing stocks suggest that there may indeed be a correlation between social media sentiment and company performance. However, more robust analysis would be needed to determine the strength of the correlation and whether this information could be used to predict future stock performance. In other words, is the sentiment a reflection of the performance or is the performance a reflection of the sentiment.