David Lattimer, Harold Anderson, Veronica Warren

## Executive Summary

We used the sentiment of tweets to examine if the general feeling toward a stock on Twitter had a correlation with stock prices of that company, and more ideally, if we could use the sentiment to help us predict the upcoming trends in stock prices. In order to do this, we needed to collect tweets containing the keywords we were interested in across a 2-3 week period, and then also collect the prices of the stock/crypto during that same time period. Because the stock market is open from 9:30 am - 4:00 pm and only on weekdays, we took sentiment and stock prices in 30 minute intervals across each day and then graphed both the sentiment and stock price together. We did the same for cryptocurrencies but on hour long intervals because you can buy and sell all hours of the day including weekends. Although it would have been most ideal to be able to predict the changes in stock prices and make buying and selling decisions ahead of the actual changes in the stock prices, this doesn't seem to be a feasible strategy. Because of the volatility of sentiment changes, along with the speed at which sentiment scores and stock prices are changing, it would be very difficult to buy and sell stocks at a rate that would be comfortable investing a large budget into. Despite that, there does seem to be some correlation between sentiment of the tweets and the stock prices. Many of the trends that can be seen in a simple chart seem to mirror each other on similar time frames, which would be considered a success in that there is a relation between how people feel about a stock and the changing of prices within the stock. Changes in sentiment also mirrored changes in prices more often than just taking a random guess, which shows that there could be a chance to capitalize on this. With better tools to process tweets at a higher rate and make decisions on even short time frames, there is a chance that it is possible to make money on these changes, even if they are somewhat small for each purchase.

# Technical Report

## Background of the Problem:

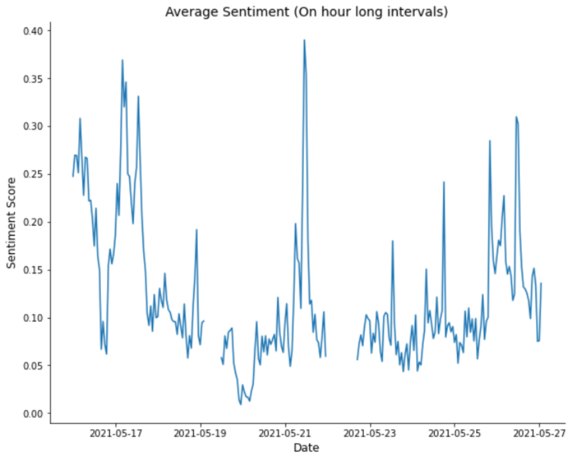
Through the years, people have tried to find any advantage they can in the stock market and make money off the success of these publicly traded companies. One of the more recent data science applications that businesses have been using for a myriad of purposes is the use of sentiment analysis. Whether it is gathering tweets or looking into reviews about a product, it can be very valuable to know what people are saying about a company and be able to gauge what people are saying to stay ahead of any problem that might arise. And although there are many forms of sentiment analysis to predict things, and using this sentiment to predict stock price isn't a new concept, it seemed like a useful project to be able to work with both text data and datetimes to see if it was even possible to predict stock prices and make money by just knowing the general sentiment around the stocks.

This project had two main objectives in order to figure out if sentiment analysis could be useful in the movement of stocks. The first of these objectives was to figure out if there was a correlation between sentiment of tweets about a stock and the stock price. After finding out if there is a correlation at all, we now have to find out if the changes in sentiment will predict the changes in prices so we can act on it. If the sentiment is instead being influenced by the changes in stock prices, we would not be able to use this to make any decisions. Although it would still be interesting to look into, the benefit of getting to buy and sell as a result of the sentiment findings would make it tough to profit off, which would be the main goal of a project and model like this. So first of all we wanted to find if there was a correlation between sentiment and stock price, and if there was, could it then be used to predict trends in the stock price ahead of time.

Methods:

In order to solve this problem, we needed to collect two different sets of data across the same time frame in order to compare the sentiment to the stock prices. The first of these was the tweets we would be using. This was by far the most finicky because of the way the tweepy library works on Python and the limitations that the Twitter API has put in place. Without paying for a premium Twitter API or enterprise version, we were stuck with several limitations that had to be worked around, such as number of tweets we could grab and the maximum date we could go back in to grab tweets. If the prices were reasonable for three students working on a project, we may have considered buying access to an easy to run API that could collect higher volumes of tweets and go back in time even further, but it was not something we could make happen. If this was a project that would be run by a company, we would assume that we could pay for a better way to collect the tweets, and then possibly run multiple sentiment analysis projects with the data to make it worth it. Because of this limitation, we were forced to run the code multiple times to gather tweets and try to get a continuous set of sentiment across a timeframe. The more a stock or cryptocurrency was tweeted about, the more often we had to run the codes, each of which took close to five hours.

The volume of tweets for cryptocurrencies like Etherium and Dogecoin made it so we needed to run the code often, and unfortunately we had to cut Etherium because we had too many gaps in the sentiment. Dogecoin also had some gaps in the sentiment, but for the most part, can see all the trends and



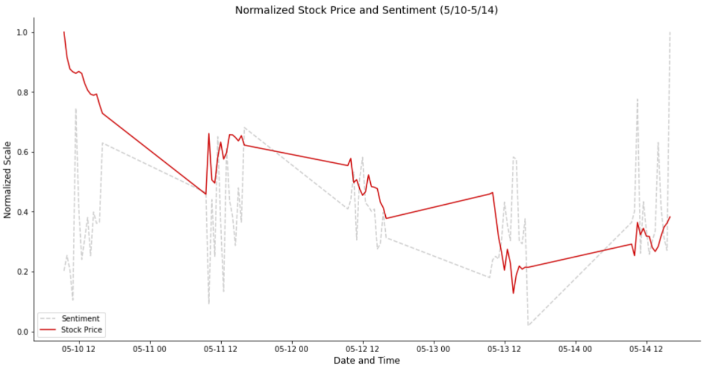
sentiment across the 12-day period. These would have been easily avoided with access to better tools, but we did what we could. After running the code for the sentiment to gather tweets, we then needed to put the tweets through some code to clean the text and get it ready to get a sentiment score tied to it. From here, we took a mean of each half an hour (or hour) long interval to get the general sentiment to match up with a stock price and compare.

Speaking of stock prices, this was another challenging aspect of the project. We wanted to collect stock prices on 30-minute intervals on days that the stock market was open. This meant between 9:30-4 on Monday through Friday, we needed to know the prices on every 30-minute interval. For whatever reason, it is difficult to receive this data for any stock more than five days ago. Anytime longer than 5 days was given as a mean price of that day, which did not help us at all. So once again, we needed to stay on top of gathering information and manually enter all these prices into excel files since no website allowed us to just download hourly data. Fine. So once we had the stock prices (or crypto prices), we could match it up with the sentiment over the time frame leading up to it and finally get some results and see if there was a correlation in the data. And if there was a way to predict the peaks and dips in the stock prices before they happened.

Results:

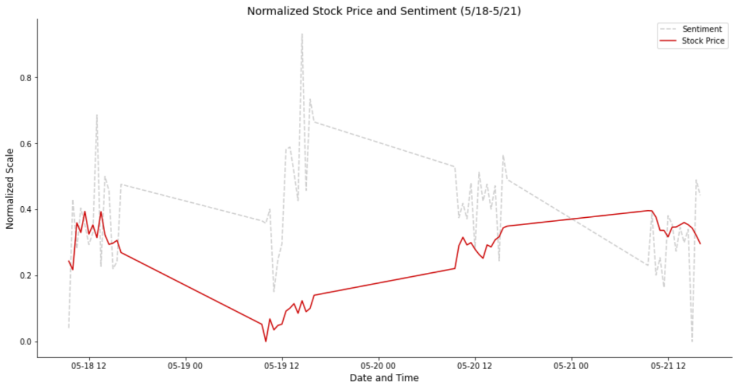
There were two objectives we had with this project to figure out if sentiment could be measured and would be correlated with the stock prices, as well as if we could actually predict the stock prices with the sentiment as it came in. Sentiment over these time periods was a lot more volatile than stock price is, with big jumps and falls between half an hour periods. Stocks also have lots of peaks and valleys throughout a day or week, but over short time frames like half an hour, this seems to smooth itself out a lot more and doesn’t have the issues that sentiment does. We will take a bit of a deeper dive into each of the stocks (and cryptocurrencies) that we chose and then look into the overall results of the project and what we would do differently next time if we had the chance.

The first stock we looked into was a more traditional stock, Tesla. We got the stock prices across the same time frame that we gathered the tweets and measured the sentiment. When lining up the data to try to compare both of these together, we used a dataframe to match up the times, so when seeing a timestamp of 10:00 am, this includes the stock price at precisely 10:00 am as well as the average sentiment from 9:30-10:00. This allowed us to see if the trends lined up and although the amount of time to react is relatively low, seeing an increase in sentiment may lead to an increase in stock price. Here is a view of these two lined up together:

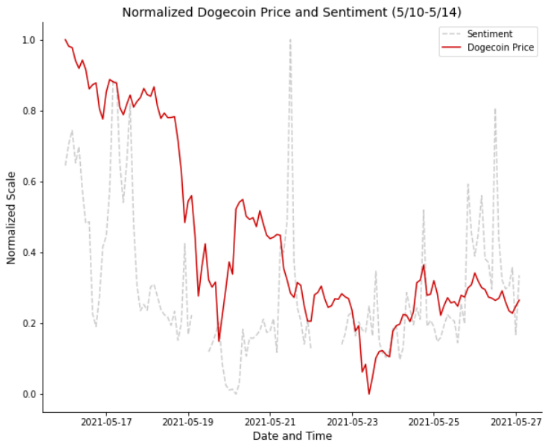


As you can see, sentiment is a lot less predictable, but we also had a lot of fluctuations in stock prices here. Clearly, the first day is all over the place and hard to match up, but this was a day of only decreasing in stock price, so the fluctuating pattern of sentiment is hard to match up. Luckily day two shows a bit of a pattern. They both are a bit erratic, but they mirror each other and show increases and decreases simultaneously. Unfortunately, day four looks to throw a wrench in our plan, where every increase in sentiment coincides with decreases in stock prices, and vice versa. The last day here shows a predictable pattern, although sentiment is a lot more drastic. But they follow a similar pattern and would be able to act on predictions and receive relatively predictable results.

Our second week follows a similar pattern, as you can see. There are some days here that clearly show a correlation and mirroring between the sentiment and stock price and everything looks great, and then there are other days where everything gets flipped on its head and suddenly all of our predictions are wrong. There does seem to be some correlation between sentiment and prices, but there is absolutely no way looking at the erratic sentiment that we are finding that we could suggest buying and selling stocks based on the results. Looking at our data points, there are 168 timestamps across the two weeks, and during this, our sentiment has an increase in sentiment from the last 30-minute time frame a total of 81 times. This is where if the sentiment prediction could be acted upon and worked well, we would get a larger number of increases on those time frames than decreases. Unfortunately, over those 81 predictions we end up with only 39 of them equating to increases in stock price, and 42 of them led to decreases. This is concerning and would not be able to be acted on. Sadly Tesla is a bit of a hard stock to crack, although there is still a chance that we don't have the right resources or method for solving this, but we will talk about suggestions for possibly expanding on that project later.



Dogecoin was a unique look at a similar methodology without some of the limitations that stocks have. For one, trading Dogecoin is never closed, so we can get pricing data, as well as sentiment data across a 24 hour period without weekends. This helps both for the way we handle after hour prices and sentiment (which of course we don't have) and allows our graphs to be a lot more coherent. Instead of having six and half hours of stock prices and then seventeen and a half hours of nothing, we get to constantly roll through the sentiment and compare it to prices. Because we have longer time periods, we opted to use 2-hour long chunks of sentiment and prices. This will help with both the erraticness of the sentiment as



well as have meaningful movements in the stock prices that show trends across the entire day and week. Without these limitations, we get a much simpler chart when we lay the sentiment over the prices of Dogecoin. Although it doesn't always line up, there are some interesting trends that seem to follow in both of these. The first large drop in sentiment, even with the hitch in it, is also shown in the price drop lining up with it. There are two gaps in the sentiment across this timeline (Sorry, I didn't run the code early enough), but there are definitely some similarities and trends that line up relatively well. Our data shows that we predicted price changes with the sentiment correctly 65 times and failed 69 times. That is discouraging, especially when there are some moments that really look to move together.

Now that we have seen the results of both Tesla and Dogecoin, why do we have these problems? Well let's give a few reasons that we may have gotten results that weren't ideal but possibly could be encouraging for making this sort of analysis work. One of the reasons that could be a factor is that over both of the timeframes, both of the prices were dropping more often than they were rising. Across the 168 points of data in the Tesla set, we had 89 points lower than the last, and in the Dogecoin set we had a 145 to 121 split of decreasing vs. increasing. Dogecoin started at 54 cents per coin and dropped down to 32 cents by the end, with a low of 25 cents. Inevitably our sentiment will increase and decrease, but the trend of price steadily declines which means unless we are trading on small changes across very short windows, the most profitable way to handle Dogecoin was to not get involved at all.

Another problem with running the code the way we did was that we specified either 30-minute or 2-hour intervals, mostly because they line up pretty well with the time the stock market is open, or a 24-hour day. But sentiment and stock prices shouldn’t match up perfectly and it's possible that some of the trends we see mirrored aren’t lined up correctly and there is a lag from the sentiment to the change in stock price. Unfortunately for us, it is very difficult to find what that lag is, and more than likely it could be sentiment that is predicted by stock price. It makes sense, people get upset if a stock price drops, people get happy if the stock rises. The problem is, that's not profitable.

Although the numbers that get spit out seem to suggest that there is no correlation or way to predict the stocks and this performed worse than just a random guess of when to buy and sell, we can see that there is more to it than that. Even if the times don’t necessarily match up, there does seem to be a correlation between the sentiment on Twitter and stock prices. We cannot determine which is affecting which for certain and there is still hope that sentiment could predict the price changes.

Discussion/Conclusion:

We have determined that there is more to this than the numbers might suggest, but let’s talk about what we would do differently given the chance to work on this again (and with a little bit of a business type budget). For one, the chance to look into more companies, especially a wider variety would have been great. Both Tesla and Dogecoin are relatively unstable and rely on a ton of factors, weirdly enough both of which are closely tied to Elon Musk’s Twitter use, so you can imagine how all over the place these can be. Being able to research more of a variety of industries and some lower frequency of tweet types of companies would be a good starting point. Once that is done, reducing the limits of the Twitter API with some extra money would be extremely helpful. Being able to run the code to gather tweets just a single time, even if it takes a lot longer, would be extremely helpful. This API also is less limited by the number of tweets and time we can go back to retrieve tweets, so a lot can be done. By adding more companies along with a longer time frame, we can ensure that any predictions we made will be correct. This is especially important because ideally once the model is built and we are likely going to be putting money into these predictions. And if our predictions are mistakes or misleading, we will lose money at a high rate.

From there we would look into more chunks of time to see if looking at smaller time intervals helps make small profits on a lot of buying opportunities that aren't being held long, or if we should look at longer time intervals that are less volatile to sentiment shifts that may be by chance. Although we had problems with four of the companies we planned to get done, there do seem to be some correlations between the sentiment and prices of the companies that did come out alright. Given the opportunity and budget we may be able to use tweets to predict the market and there could be a chance there, but we cannot know for sure. There are many factors that could go into it and would be amazing to get the chance to expand and keep learning from this project.

References:

Tesla Inc. Historical Stock Prices. https://www.marketwatch.com/investing/stock/tsla.

Market Watch

Dogecoin Price Index. https://www.coindesk.com/price/dogecoin. Coin Desk.