Capstone Project 2: Twitter Sentiment Analysis and Stock Prediction

Problem Statement: Analyze stock market movements using Twitter sentiment analysis to find the correlation between ”public sentiment” and ”market sentiment”

I. Background

The value of individual stocks often do not seem to reflect the fair value due to human error. Instead, the price stocks trade at seem to be determined more by the human perception of the stock. Behavioral economics states that the emotions and moods of individuals affect their decision making process. People and news outlets are constantly voicing their opinions about a variety of subjects, stocks included, on Twitter. Though a single tweet may not be significant, a large collection of them can provide data with valuable insight about the common opinion on a particular subject. Twitter, therefore, can be used to gauge the public sentiment and possibly predict stock price movements. The famous research paper by Bollen et al had performed Twitter sentiment analysis to predict price movement of the Dow Jones Industrial Average (DJIA). However, this study will focus on 4 individual stocks: Netflix ($NFLX), Disney ($DIS), Amazon ($AMZN), and Google ($GOOGL).

II. Impact

Applying sentiment analysis to stock movement forecasting has been prominent field due to the potential financial gains from it. Without a doubt, investment banking firms have done extensive research and built model based of the idea. Twitter API makes its data readily available to the public, so why shouldn't individual investors benefit from the wealth as well? By analyzing these trends and monitoring public opinion of companies, we can possibly build a predictive model to exploit market inefficiencies and anticipate changes in the market before they happen.

III. Data

A. **Twitter data**: The Twitter API will be used to pull tweets mentioning the stocks of interest. It is important to note that Twitter prevents users from pulling tweets past 7 days old with the standard API key. Sentiment Analysis "VADER" will be used to analyze the sentiment of the tweet. The data will include the following information:
    1. Date tweet was posted
    2. Content of actual tweet
    3. Follower count
    4. Percentage of Negative sentiment
    5. Percentage of Positive sentiment
    6. Percentage of Neutral sentiment

B. **Historical Stock Data**: Alpha Vantage API will be used to retrieve historical data for the stocks of interest. It will only be pulled starting from the earliest tweet pulled. The data will include the following information:

1. Date
2. Open
3. Close
4. High
5. Low
6. Traded Volume

Data Collection

To store all the data that I would need for the project, I created a SQLite database. SQLite is a relational database management system contained in the C library. A separate table was created for each company's twitter data and stock data. In total, 8 tables were created in the database.

Twitter Data

I collected tweets for an entire month *05-15-2019 until 06-15-2019.* I used the Standard search API to collect the tweets I need for the project. It returns a collection of relevant tweets that match the parameters given. It has parameters to specify the date range to collect tweets from as well as query to filter by the content. When a user wants to explicitly mention someone or a product, they use "@" sign to gain attention from them. Therefore, if a user wrote a tweet specifically for a product, they would "@" the product. Some of the companies I am focused on have more products than their main product. To account for those as well, I have used the following queries for the respective companies:

- Netflix: '@Netflix OR $NFLX OR Netflix'
- Disney: '@Disney OR @ESPN OR @ABCnetwork OR @Pixar OR @Marvel OR $DIS'
- Amazon: '@Amazon OR @PrimeVideo OR @awscloud OR @TwitchPrime OR @Alexa OR @WholeFoods OR $AMZN'
- Google: '@Google OR @Android OR @Waymo OR $GOOGL'

For each tweet, I collected the text of tweet, date created, and number of followers. Simultaneously, I used VADER (Valence Aware Dictionary and sEntiment Reasoner) to analyze the tweet. It is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is able to properly handle typical negations, conventional use of punctuation to signal increased sentiment intensity, slang words, and emojis. VADER then provides a positive, negative, and neutral score which are ratios for proportions of text that fall in each category. These score should add up to 1. It then computes a compound score which sums the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). The compound score is the normalized, weighted composite score.

Before using VADER to analyze the tweet, each tweet was cleaned. User mention was helpful in filtering out the tweets for the project, but it does not provide any sentiment value. Hyperlinks included in the tweet also do not provide any sentiment value. Hashtags may contain a lot of sentiment, so the text of the hashtag should be kept while the sign itself should be removed. Twitter has a "retweet" feature which allows users to quickly share a tweet by reposting it. It is prefaced with

“RT,” but it does not contain any information and may confuse the tool. All “RT” should then be removed. The Regular Expression library was used to find strings of the sort and remove them.

Here’s an example of a pre-cleaned tweet:

```
'RT @Google: Toy Story is back. See the latest Toy Story 4 trailer #WithALittleHelp from Google → https://t.co/np6XbygVvi ht
tps://t.co/Hnpmy…'
```

Here is the cleaned tweet:

```
'  Toy Story is back See the latest Toy Story 4 trailer WithALittleHelp from Google   '
```

Stock Data

Alpha Vantage provides APIs for real time and historical data on stocks. With the help of the alpha_vantage package, collecting stock data was a breeze. I was able to easily search by specifying the abbreviation of the stock of interest. It had information for the date, open price, closing price, highest price, lowest price, and volume traded for each day. It is important to note that it skips weekends and holidays since the market is not open on those days.

Data Wrangling and Exploratory Data Analysis

Twitter Data

There wasn’t any missing values for any of the fields. Here is the summary statistics for each of the tables.

Netflix

	follower_count	neg_sent	neu_sent	pos_sent	compound_sent
count	6.405500e+04	64055.000000	64055.000000	64055.000000	64055.000000
mean	1.065048e+04	0.056622	0.806854	0.136505	0.179516
std	4.770410e+05	0.105180	0.178109	0.160311	0.438237
min	0.000000e+00	0.000000	0.000000	0.000000	-0.985500
25%	7.900000e+01	0.000000	0.682000	0.000000	0.000000
50%	2.650000e+02	0.000000	0.823000	0.096000	0.000000
75%	7.750000e+02	0.094000	1.000000	0.235000	0.557400
max	7.781631e+07	1.000000	1.000000	1.000000	0.990800

Disney

	follower_count	neg_sent	neu_sent	pos_sent	compound_sent
count	7.840400e+04	78404.000000	78404.000000	78404.000000	78404.000000
mean	8.773094e+03	0.053819	0.799222	0.146743	0.212293
std	3.438629e+05	0.107916	0.181663	0.158116	0.396077
min	0.000000e+00	0.000000	0.000000	0.000000	-0.985000
25%	6.800000e+01	0.000000	0.688000	0.000000	0.000000
50%	2.360000e+02	0.000000	0.812000	0.130000	0.177900
75%	6.250000e+02	0.080000	1.000000	0.245000	0.541100
max	3.548309e+07	1.000000	1.000000	1.000000	0.985300

Amazon

	follower_count	neg_sent	neu_sent	pos_sent	compound_sent
count	7.970700e+04	79707.000000	79707.000000	79707.00000	79707.000000
mean	8.407060e+03	0.050272	0.794103	0.15541	0.239563
std	1.417709e+05	0.100730	0.174021	0.15772	0.429720
min	0.000000e+00	0.000000	0.000000	0.00000	-0.986100
25%	8.900000e+01	0.000000	0.689500	0.00000	0.000000
50%	3.890000e+02	0.000000	0.814000	0.13400	0.250000
75%	1.769000e+03	0.071000	0.934000	0.24400	0.585900
max	2.048416e+07	1.000000	1.000000	1.00000	0.989900

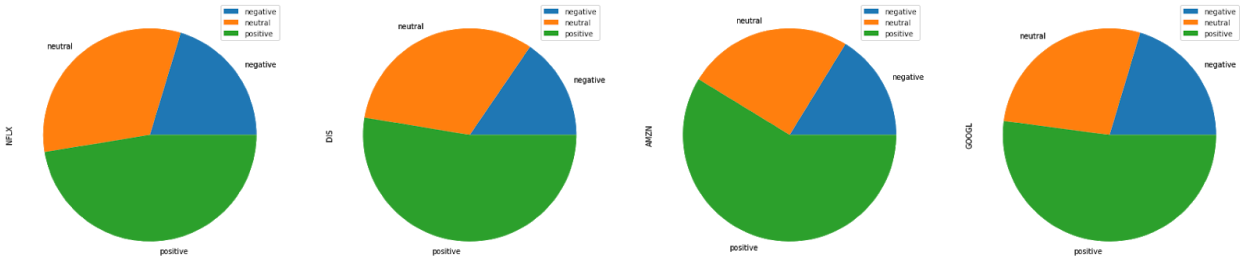
Google

	follower_count	neg_sent	neu_sent	pos_sent	compound_sent
count	8.972800e+04	89728.000000	89728.000000	89728.000000	89728.000000
mean	1.162305e+04	0.057093	0.805894	0.136841	0.184412
std	2.718562e+05	0.109389	0.172226	0.156312	0.453263
min	0.000000e+00	0.000000	0.000000	0.000000	-0.988300
25%	7.400000e+01	0.000000	0.702000	0.000000	0.000000
50%	3.110000e+02	0.000000	0.821000	0.103000	0.102700
75%	1.281000e+03	0.087000	1.000000	0.240000	0.585900
max	2.804650e+07	1.000000	1.000000	1.000000	0.988000

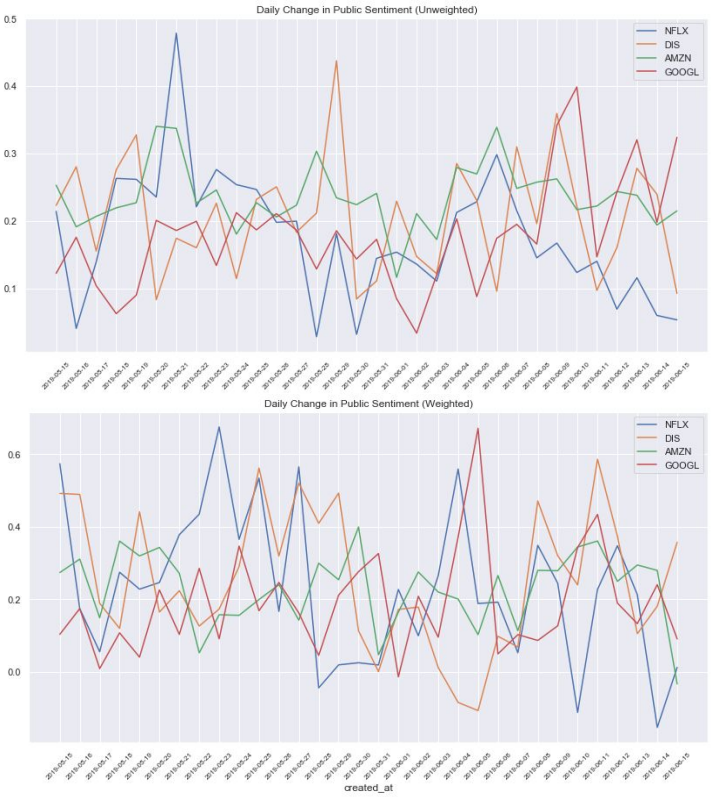
Google, overall, had the highest number of tweets collected (89,728 tweets) Netflix had the lowest number (64,055 tweets). The following is the average number of daily tweets per company.

Company	average_daily_tweets
AMZN	2490.84375
DIS	2450.12500
GOOGL	2804.00000
NFLX	2001.71875

For easier interpretation of the compound scores, I have classified the scores into 3 groups in the “sentiment” column. The documentation for VADER advises that compound scores greater than 0.05 should be categorized as positive, less than -0.05 as negative, and those in between to be considered neutral. I also added a column for the name of the company so that all the data can still be distinguished once I merge all the tables together.



As shown by the pie plot, most of the tweets the companies received were positive. Amazon had the highest ratio for positive tweets while Netflix had the highest ratio for neutral tweets.



This table on the left illustrates the daily change in average public sentiment per stock. All the tweets are given equal weight when computing the average. However, twitter accounts that have more followers have more influence on the community since their tweet reaches more people. To incorporate this, we have given weight to each tweet according to proportion of followers they have compared to the total amount of followers for each day. The resulting graph is the graph below.

<u>Stock Data</u>

There wasn’t any missing values for any of the fields for these table either other than the weekend and holiday gaps. However, to see the trend in the time series, I opted to have continuous data. The information for the weekends and holidays were filled in using forward-filling linear interpolation. It estimates a new value by connecting two adjacent known values with a straight line.

Original Table:

	date	open	high	low	close	volume
0	2019-05-15	1122.55	1178.30	1121.40	1170.80	2965117.0
1	2019-05-16	1171.84	1194.16	1168.45	1184.50	1765388.0
2	2019-05-17	1175.83	1186.29	1166.42	1168.78	1268050.0
3	2019-05-18	NaN	NaN	NaN	NaN	NaN
4	2019-05-19	NaN	NaN	NaN	NaN	NaN

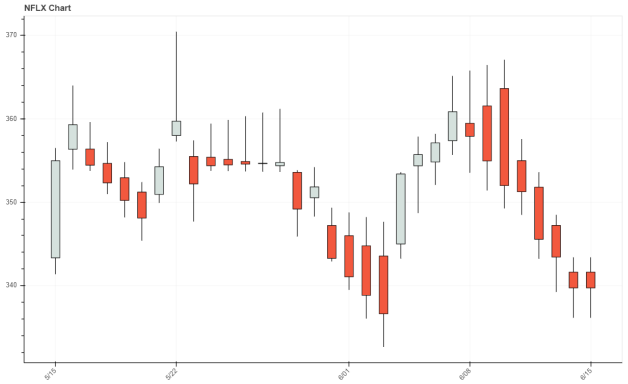
New Table After Interpolation:

	date	open	high	low	close	volume
0	2019-05-15	1122.55	1178.300000	1121.40	1170.80	2.965117e+06
1	2019-05-16	1171.84	1194.160000	1168.45	1184.50	1.765388e+06
2	2019-05-17	1175.83	1186.290000	1166.42	1168.78	1.268050e+06
3	2019-05-18	1168.22	1175.193333	1156.99	1160.74	1.355409e+06
4	2019-05-19	1160.61	1164.096667	1147.56	1152.70	1.442767e+06

Here are the summary statistics candlestick graph for each stock. The days in green means that the price of the stock increased throughout the day while red means the price decreased throughout the day.

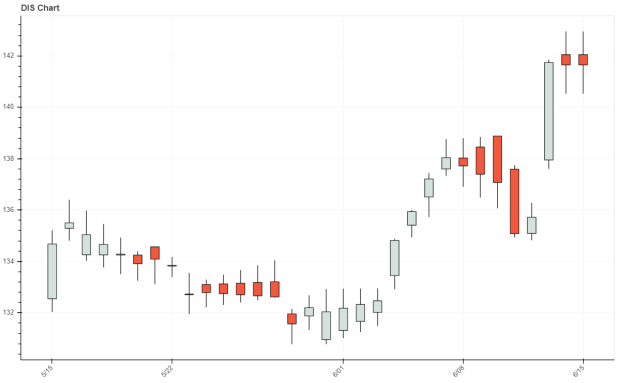
Netflix

	open	high	low	close	volume
count	32.000000	32.000000	32.000000	32.000000	3.200000e+01
mean	352.288437	356.663394	347.494766	350.973594	5.380332e+06
std	5.736398	6.787584	6.665745	6.588239	1.183147e+06
min	341.630000	343.400000	332.650000	336.630000	3.709955e+06
25%	347.227500	353.312000	343.152500	347.472500	4.612312e+06
50%	354.385000	357.319667	348.605000	352.868333	5.019546e+06
75%	355.432500	360.430000	353.649438	354.825833	6.214358e+06
max	363.650000	370.460000	357.300000	360.870000	7.891632e+06



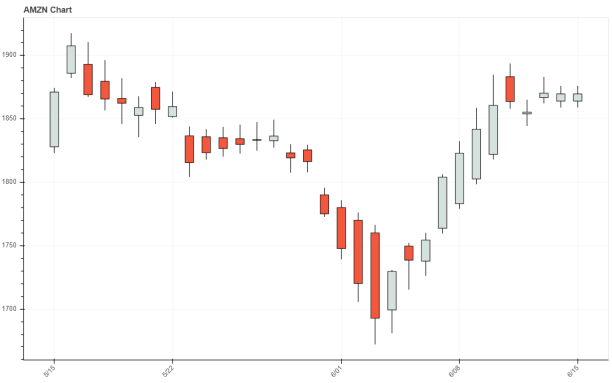
Disney

	open	high	low	close	volume
count	32.000000	32.000000	32.000000	32.000000	3.200000e+01
mean	134.831094	135.779687	133.980147	134.971719	7.934529e+06
std	2.911039	2.976506	2.597020	2.853834	2.407592e+06
min	130.960000	132.150000	130.778300	131.570000	4.570132e+06
25%	133.002500	133.531875	132.175000	132.694375	6.751520e+06
50%	134.251667	134.903333	133.320000	134.475000	7.745408e+06
75%	136.780000	137.516250	135.137500	136.222500	8.521091e+06
max	142.050000	142.950000	140.530000	141.740000	1.793954e+07



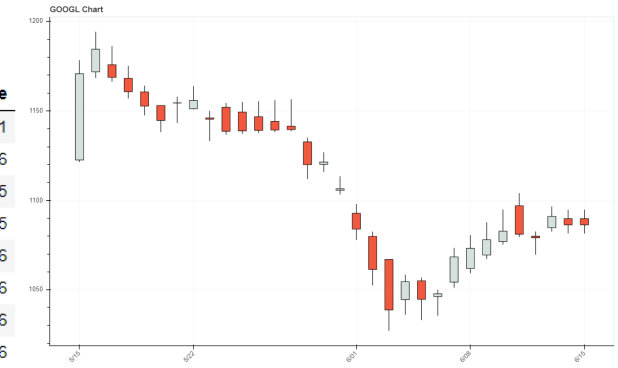
Amazon

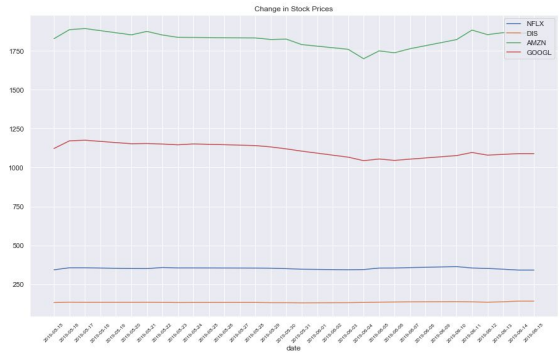
	open	high	low	close	volume
count	32.000000	32.000000	32.000000	32.000000	3.200000e+01
mean	1824.333750	1842.896719	1805.932500	1824.001406	4.322092e+06
std	48.262917	48.373435	56.777636	52.901046	1.399366e+06
min	1699.240000	1730.820000	1672.000000	1692.690000	2.678335e+06
25%	1788.290833	1823.665000	1777.420000	1812.617500	3.274212e+06
50%	1833.927500	1848.331250	1821.412500	1834.786250	4.175393e+06
75%	1864.000000	1876.750000	1847.352500	1862.660000	4.754525e+06
max	1893.050000	1917.510000	1882.290000	1907.570000	9.098708e+06



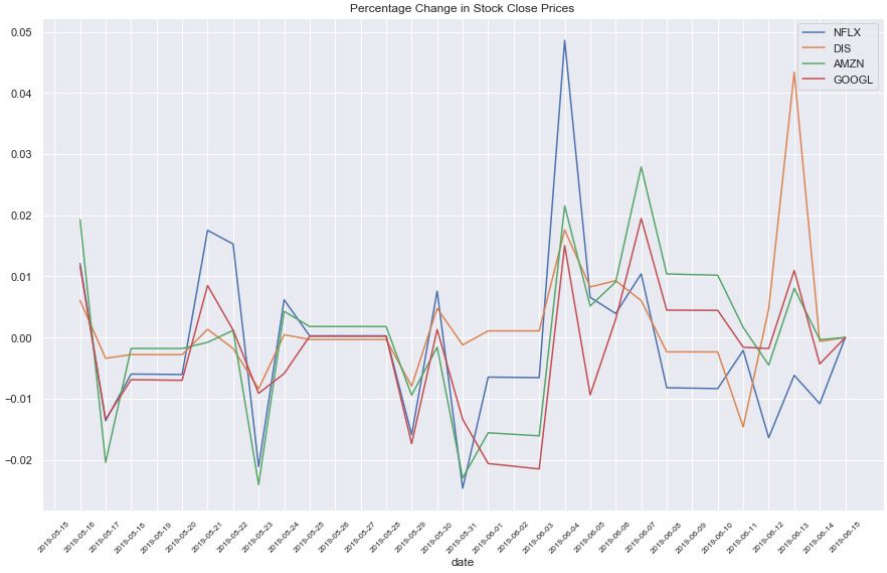
Google

	open	high	low	close	volume
count	32.000000	32.000000	32.000000	32.000000	3.200000e+01
mean	1112.029062	1121.782031	1101.736875	1110.561406	1.702416e+06
std	41.566028	42.652710	43.399708	42.342269	9.023679e+05
min	1044.490000	1050.000000	1027.030000	1038.740000	9.044190e+05
25%	1079.125000	1086.407500	1069.009167	1078.815833	1.045831e+06
50%	1112.895000	1120.100000	1107.650000	1113.220000	1.434116e+06
75%	1149.840000	1156.090625	1138.142500	1144.830000	1.842647e+06
max	1175.830000	1194.160000	1168.450000	1184.500000	4.844480e+06





As shown by the average price per stock, Netflix and Disney are sold at way lower prices than Amazon and Google. It would be hard to see the trend in all the stock prices on one graph. Therefore, the prices must be standardized so that price movements would be comparable. I have computed the percentage change in price via difference in natural log of stock close prices.



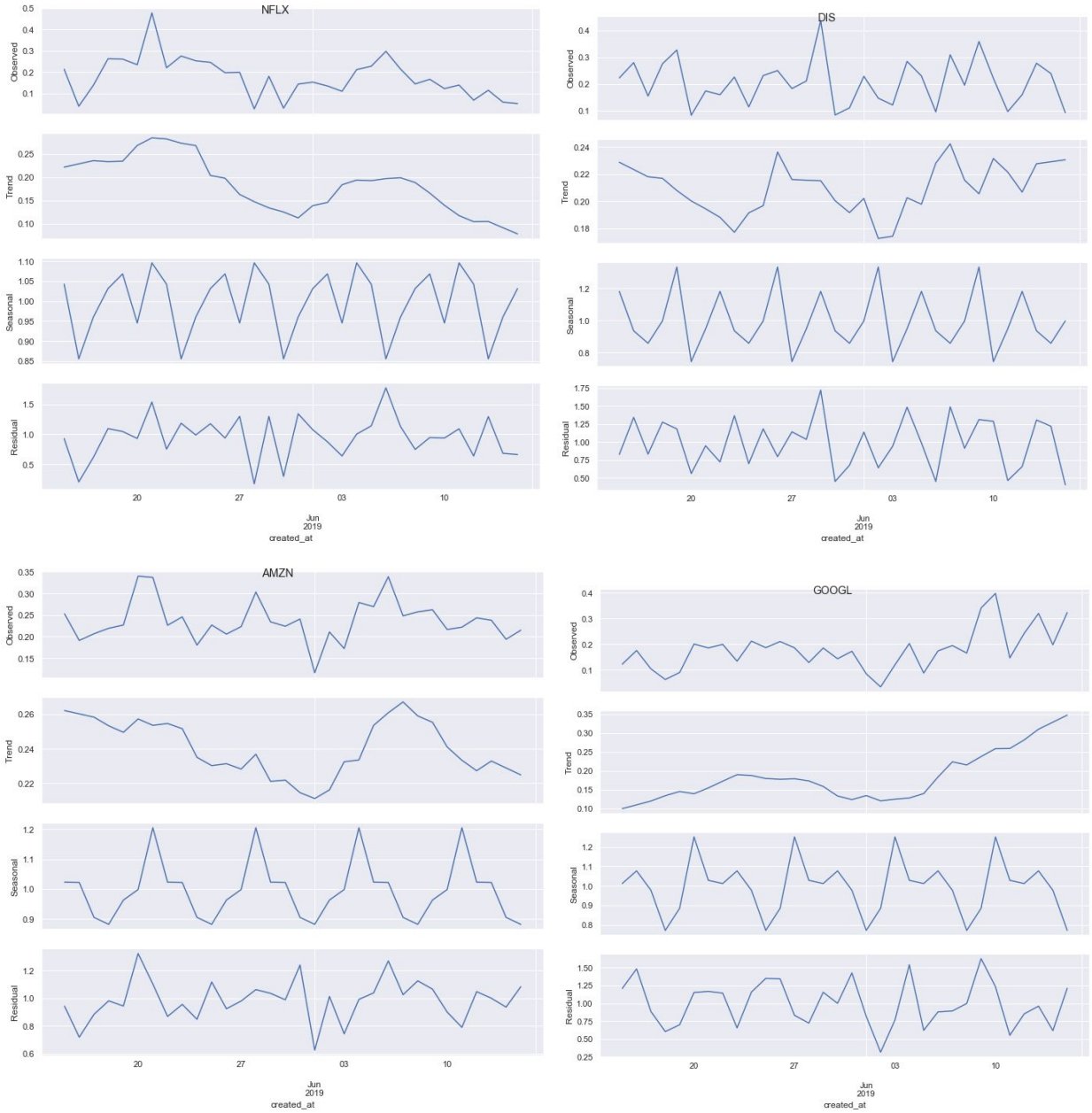
With this graph, it is more apparent when stock prices decrease or increase from day to day. Most of the stocks plummeted on 5-23-2019 then spiked up on 06-04-2019 and spiked again 6-13-2019.

Time Series Analysis

Time Series Decomposition

Time series may be split into: Base Level, Trend, Seasonality, Error. Below, I have used the multiplicative seasonal decomposition to break down the time series for public sentiment per company.

Multiplicative Time Series: Value = Base Level x Trend x Seasonality x Error



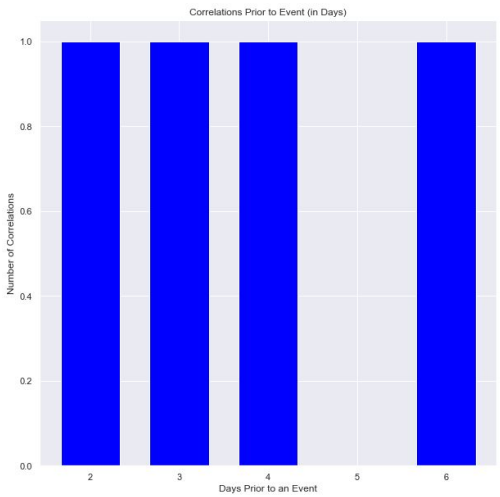
After decomposing the seasonality from the time series, it is apparent that the public sentiment for Netflix has decreased over time. Google, on the other, has an upward trend in public sentiment. The other two have a more balanced trendline.

Granger Causality

Granger causality tests whether the time series in the second column Granger causes the time series in the first column. Granger causality means that past values of x2 have a statistically significant effect on the current value of x1, taking past values of x1 into account as regressors. We reject the null hypothesis that x2 does not Granger cause x1 if the p-values are below a desired size of the test. In this case, I am testing if the public sentiment Granger causes shifts in the stock price at the 5% significance level.

H0: The public sentiment for the company does NOT Granger cause the movement in stock price for that company.
***H1*:** The public sentiment for the company Granger causes the movement in stock price for that company.

```
============================================================

No Causality

NFLX closing stock price and NFLX twitter sentiment showed NO causality, count: 31
DIS closing stock price and DIS twitter sentiment showed NO causality, count: 31
AMZN closing stock price and AMZN twitter sentiment showed NO causality, count: 31

------------------------------------------------------------

Causality

GOOGL closing stock price and GOOGL twitter sentiment showed causality, count: 31

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Causality Count: 1
No Causality Count: 3

std 1.479019945774904
avg 3.75
Percent showing causality: 0.25
```



It looks like the only relationship that was able to reject the null hypothesis was Google's public sentiment and market sentiment time series. There was a correlation shown between the public sentiment and market sentiment for 2,3,4 and 6 days prior to the stock price movements.

There's not enough evidence to prove that the public sentiment Granger caused the stock price movements for the other companies at the 95% confidence level.