Social Media Activity in Relation to Share Prices

Kunaal Chaudhari
kuchaud@seas.upenn.edu

Jonathan Choi
jonchoi@seas.upenn.edu

Tiffany Moi
tmoi@seas.upenn.edu

May 7, 2019

1 Introduction

This project aims to analyze, both statistically and visually, the bearing which a company's social media presence has on its share prices. We have selected Wendy's, Dunkin Donuts, Viacom, Disney, Adidas, and Nike as our companies. These companies represent a porfolio of different industries. Specifically, Wendy's and Dunkin Donuts belong to the fast food industry, Viacom and Disney belong to the entertainment industry, and Adidas and Nike belong to the clothing/fashion industry.

As an overview, the code we have compiled allows for a user to choose a company as well as a time period that the data should be analyzed within. Using the tweety API which provides data regarding a company's Twitter activity, as well as the iex API which provides a company's stock information, the relevant data is collected. This data is visualized and in addition, Granger's Causality Test is run to assess causality of social media activity on stock prices.

2 Methodology

The codebase is made up of three files, twitterAPI.py, iexCloud.py, and dataViz.py which are describe in the following sections. The primary test performed is a Granger Causality Test explained below. Find the GitHub repository here: https://github.com/JLChoi/Social-Media-Marketing

The following equations are the basis for the Granger causality test. Granger causality takes two time series and finds if one affects the other. Let x, y be these series and we want to see if x affects y. First, using Equation 1, we apply lag values to the y time series and perform an autoregression on this data. Once this is done, we incorporate the x data as shown in Equation 2 and place lag on the x values. When the regression is performed on this data, the two trends are compared via a variety of tests. The one that will be focused on is the likelihood ratio test which outputs a ratio where larger values represent larger degrees of causality between x and y.

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_m y_{t-m} + error_t \tag{1}$$

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_m y_{t-m} + b_p x_{t-p} + \cdots + b_q x_{t-q} + error_t \tag{2}$$

2.1 twitterAPI.py

To access the number of Tweets in between two dates, we used the Twitter API. First, we used oauth2 to authenticate ourselves before querying a particular company. Then, we searched for as many posts on the company's timeline that we could. Twitter's API limits us to 200 results per call, so we had to use the max_id parameter to try to get more posts before the last result of the previous call. Unfortunately, no matter how we changed the parameters and the max_id numbers, the API would not give us any results before last month (April). We tried to use a wrapper package for Python called tweepy to see if we could get more results, but it also could only retrieve the tweets up to the beginning of April.

As we were loading the tweets, we also took into account when these tweets were posted and grouped them into tweets made on the same day. The output of the function is a list of numbers, where the number at index i corresponds to the number of posts made on the ith day in relation to the range given to the function. For example, the number at index 0 is the number of posts made on the first day of the specified range.

2.2 iexCloud.py

To gather real data from the stock market about the companies we observed, we used the IEX Cloud API. This API gave us to a wide range of historical data about any stock we queried, from earnings per share (EPS) to recommendation trends. Unfortunately, the API only offered data as recent as April 20, 2019. For this experiment, we decided to use the closing price of the relevant stocks for each day in the observed range of dates. This was to obtain a day-to-day look at the shares' value over the given period, and allowed us to directly compare the shares' price to the companies' social media activity for each day. The API took dates inputted in the form (YYYYMMDD), so the script we wrote parsed the input string to get the year, month, and day range, and iterated over all the days in the range, storing the closing price of the given stock on that day in a list. We also wrote a script for obtaining the EPS values for shares, but these values analyzed the shares over entire quarters. This did not allow us to make correlations between the values and our social media data, so we decided not to use EPS for our analysis.

2.3 dataViz.py

This file first fits the user inputs to the methods which query for data from the APIs. This involves turning dates into array indices as well as mapping company names to their stock index codes for querying. Once the data has been acquired, the raw data are plotted on the same x-axis. This is done by creating two y-axes which share one x-axis on an individual plot. In addition, arrays are initialized which store percent changes in both share prices and Twitter activity. These are graphed on the same plot in the same way. Finally, the percent change data are tested for Granger causality to determine if certain events in social media activity trigger stock price changes at all with a maximum lag of 5 days.

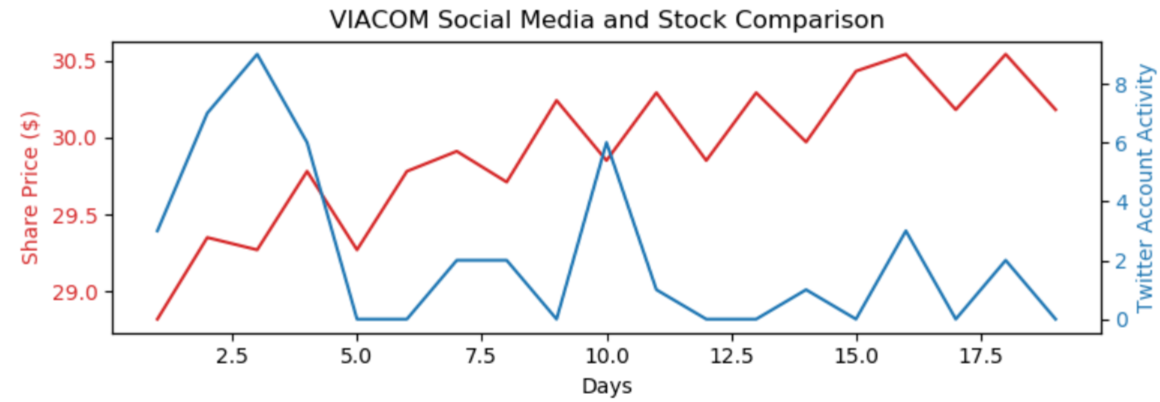
3 Results

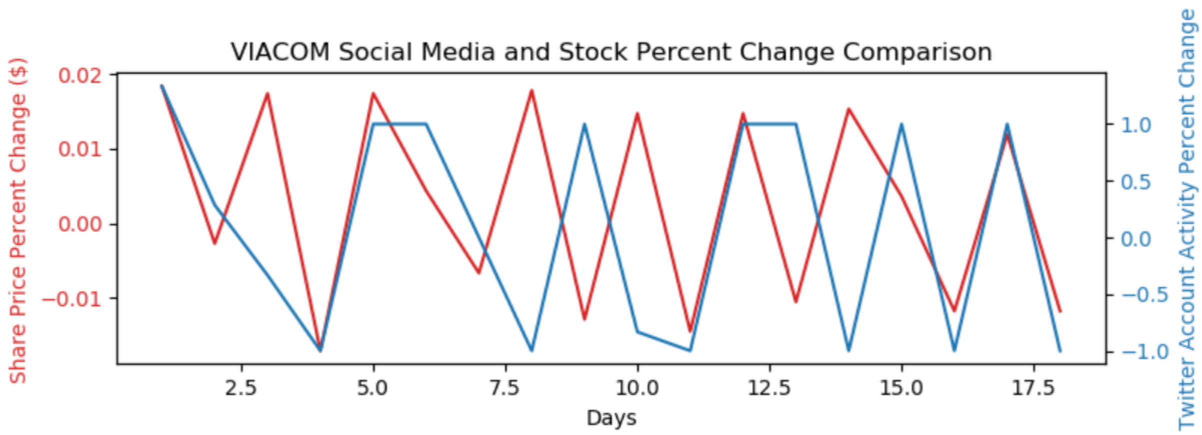
The program was run on all the companies for the entire range from 04/02/2019 to 04/19/2019. The following are the plots as well as the Granger Causality outputs with maximum lag of 5 days.

Viacom Results

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.0421  , p=0.8406  , df_denom=13, df_num=1
ssr based chi2 test:   chi2=0.0518  , p=0.8200  , df=1
likelihood ratio test: chi2=0.0517  , p=0.8201  , df=1
parameter F test:         F=0.0421  , p=0.8406  , df_denom=13, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.2725  , p=0.7669  , df_denom=10, df_num=2
ssr based chi2 test:   chi2=0.8176  , p=0.6645  , df=2
likelihood ratio test: chi2=0.7961  , p=0.6716  , df=2
parameter F test:         F=0.2725  , p=0.7669  , df_denom=10, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.3454  , p=0.7938  , df_denom=7, df_num=3
ssr based chi2 test:   chi2=2.0721  , p=0.5576  , df=3
likelihood ratio test: chi2=1.9324  , p=0.5866  , df=3
parameter F test:         F=0.3454  , p=0.7938  , df_denom=7, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=0.1951  , p=0.9287  , df_denom=4, df_num=4
ssr based chi2 test:   chi2=2.5369  , p=0.6380  , df=4
likelihood ratio test: chi2=2.3175  , p=0.6776  , df=4
parameter F test:         F=0.1951  , p=0.9287  , df_denom=4, df_num=4

Granger Causality
```

```
number of lags (no zero) 5
ssr based F test:         F=2.0412  , p=0.4848  , df_denom=1, df_num=5
ssr based chi2 test:   chi2=122.4744, p=0.0000  , df=5
likelihood ratio test: chi2=28.9976 , p=0.0000  , df=5
parameter F test:         F=2.0412  , p=0.4848  , df_denom=1, df_num=5
```

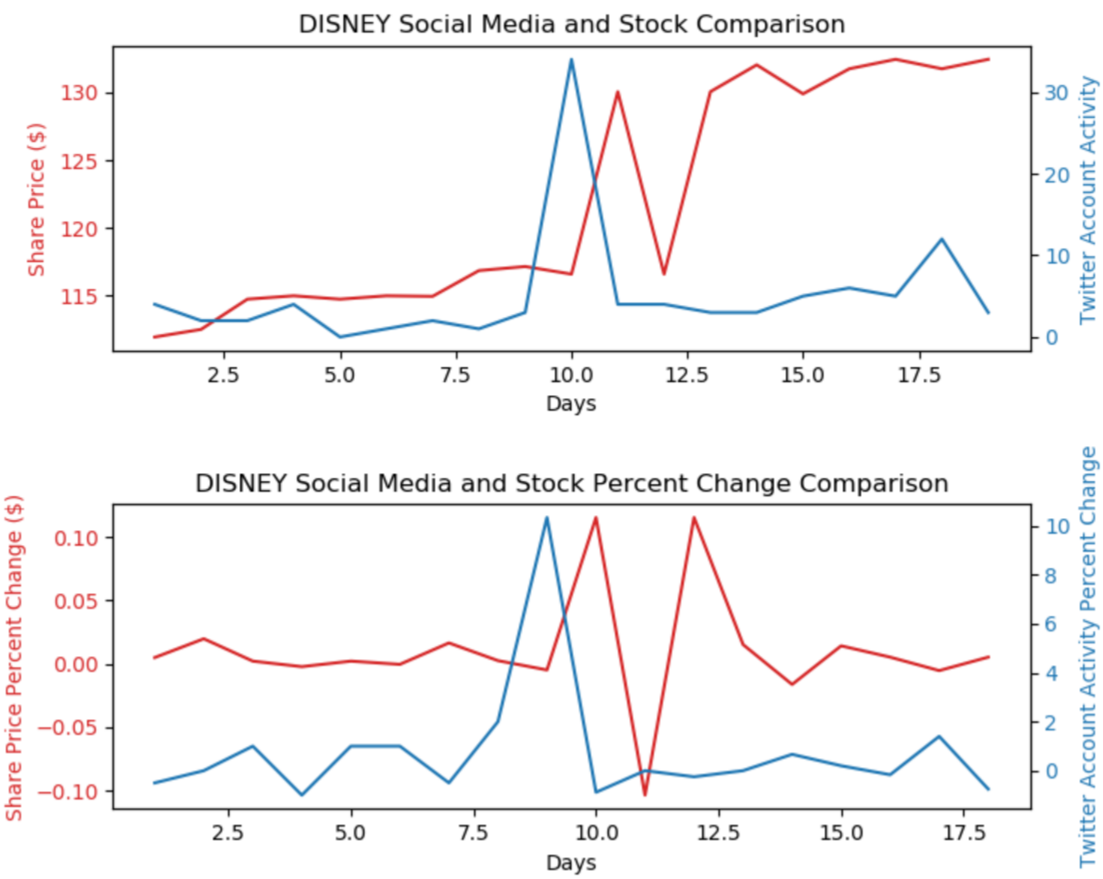




Disney Results

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=8.9633  , p=0.0104  , df_denom=13, df_num=1
ssr based chi2 test:   chi2=11.0318 , p=0.0009  , df=1
likelihood ratio test: chi2=8.3908  , p=0.0038  , df=1
parameter F test:         F=8.9633  , p=0.0104  , df_denom=13, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=3.7680  , p=0.0603  , df_denom=10, df_num=2
ssr based chi2 test:   chi2=11.3040 , p=0.0035  , df=2
likelihood ratio test: chi2=8.4251  , p=0.0148  , df=2
parameter F test:         F=3.7680  , p=0.0603  , df_denom=10, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=5.8843  , p=0.0251  , df_denom=7, df_num=3
ssr based chi2 test:   chi2=35.3059 , p=0.0000  , df=3
likelihood ratio test: chi2=17.6258 , p=0.0005  , df=3
parameter F test:         F=5.8843  , p=0.0251  , df_denom=7, df_num=3
```

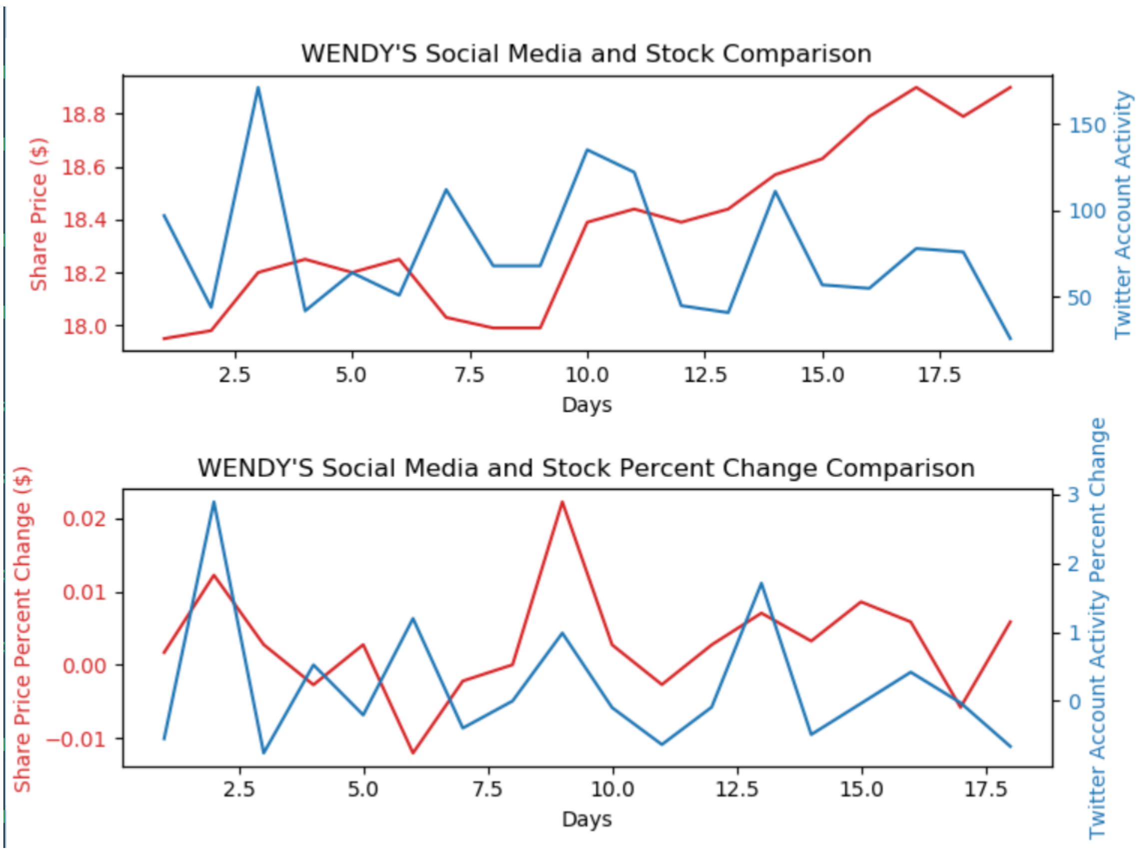
```
Granger Causality
number of lags (no zero) 4
ssr based F test:         F=5.5550  , p=0.0627  , df_denom=4, df_num=4
ssr based chi2 test:   chi2=72.2146 , p=0.0000  , df=4
likelihood ratio test: chi2=24.4429 , p=0.0001  , df=4
parameter F test:         F=5.5550  , p=0.0627  , df_denom=4, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=8.2368  , p=0.2583  , df_denom=1, df_num=5
ssr based chi2 test:   chi2=494.2052, p=0.0000  , df=5
likelihood ratio test: chi2=44.9044 , p=0.0000  , df=5
parameter F test:         F=8.2368  , p=0.2583  , df_denom=1, df_num=5
```



Wendy's

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.0037  , p=0.9523  , df_denom=13, df_num=1
ssr based chi2 test:   chi2=0.0046  , p=0.9460  , df=1
likelihood ratio test: chi2=0.0046  , p=0.9460  , df=1
parameter F test:         F=0.0037  , p=0.9523  , df_denom=13, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.5099  , p=0.6154  , df_denom=10, df_num=2
ssr based chi2 test:   chi2=1.5297  , p=0.4654  , df=2
```

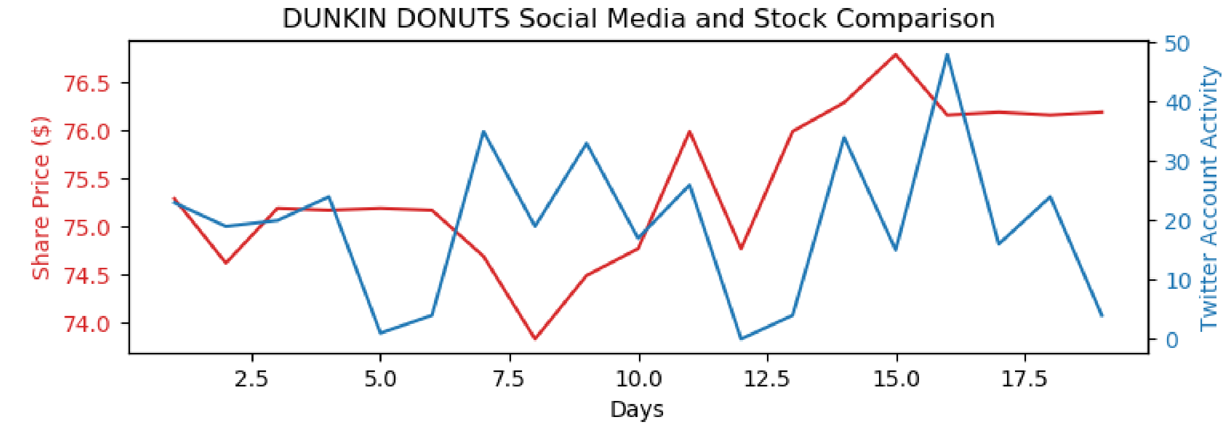
```
likelihood ratio test: chi2=1.4566   , p=0.4827   , df=2
parameter F test:         F=0.5099   , p=0.6154   , df_denom=10, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=1.3620   , p=0.3304   , df_denom=7, df_num=3
ssr based chi2 test:   chi2=8.1722   , p=0.0426   , df=3
likelihood ratio test: chi2=6.4370   , p=0.0922   , df=3
parameter F test:         F=1.3620   , p=0.3304   , df_denom=7, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=7.5669   , p=0.0377   , df_denom=4, df_num=4
ssr based chi2 test:   chi2=98.3696  , p=0.0000   , df=4
likelihood ratio test: chi2=27.9228  , p=0.0000   , df=4
parameter F test:         F=7.5669   , p=0.0377   , df_denom=4, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=23.8629  , p=0.1541   , df_denom=1, df_num=5
ssr based chi2 test:   chi2=1431.7751, p=0.0000   , df=5
likelihood ratio test: chi2=57.4813  , p=0.0000   , df=5
parameter F test:         F=23.8629  , p=0.1541   , df_denom=1, df_num=5
```

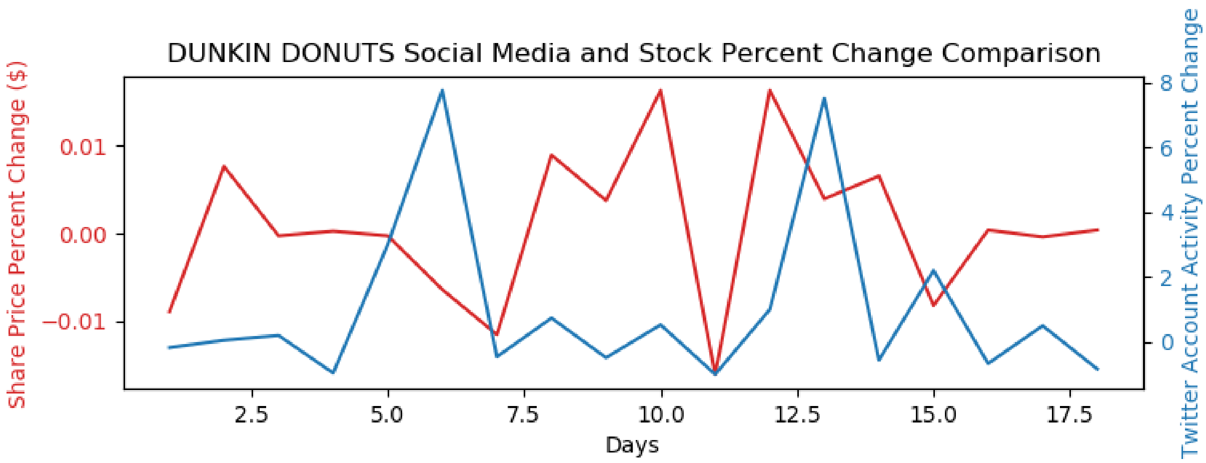


Dunkin Donuts

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=1.4399   , p=0.2516   , df_denom=13, df_num=1
ssr based chi2 test:   chi2=1.7722   , p=0.1831   , df=1
```

```
likelihood ratio test: chi2=1.6808  , p=0.1948  , df=1
parameter F test:         F=1.4399  , p=0.2516  , df_denom=13, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.8160  , p=0.4696  , df_denom=10, df_num=2
ssr based chi2 test:   chi2=2.4479  , p=0.2941  , df=2
likelihood ratio test: chi2=2.2675  , p=0.3218  , df=2
parameter F test:         F=0.8160  , p=0.4696  , df_denom=10, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=1.1353  , p=0.3985  , df_denom=7, df_num=3
ssr based chi2 test:   chi2=6.8118  , p=0.0781  , df=3
likelihood ratio test: chi2=5.5505  , p=0.1356  , df=3
parameter F test:         F=1.1353  , p=0.3985  , df_denom=7, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=1.0849  , p=0.4695  , df_denom=4, df_num=4
ssr based chi2 test:   chi2=14.1041 , p=0.0070  , df=4
likelihood ratio test: chi2=9.5516  , p=0.0487  , df=4
parameter F test:         F=1.0849  , p=0.4695  , df_denom=4, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=1.9567  , p=0.4933  , df_denom=1, df_num=5
ssr based chi2 test:   chi2=117.3995, p=0.0000  , df=5
likelihood ratio test: chi2=28.5360 , p=0.0000  , df=5
parameter F test:         F=1.9567  , p=0.4933  , df_denom=1, df_num=5
```

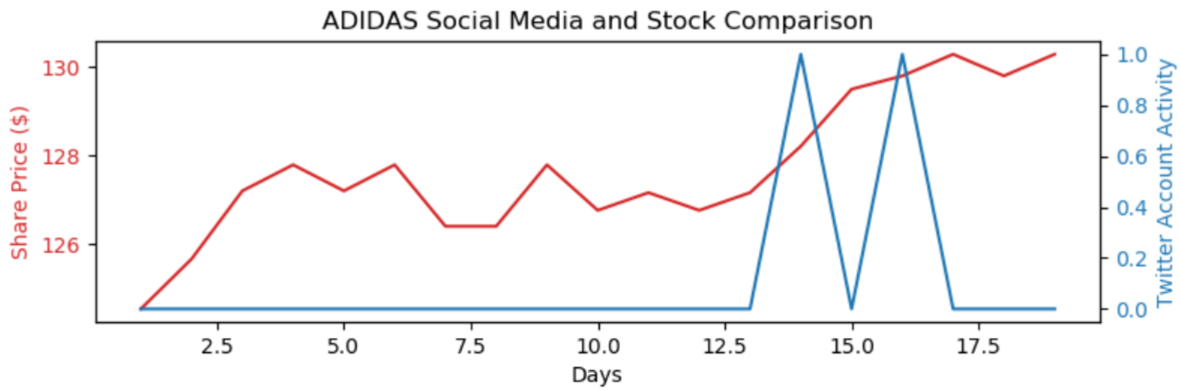


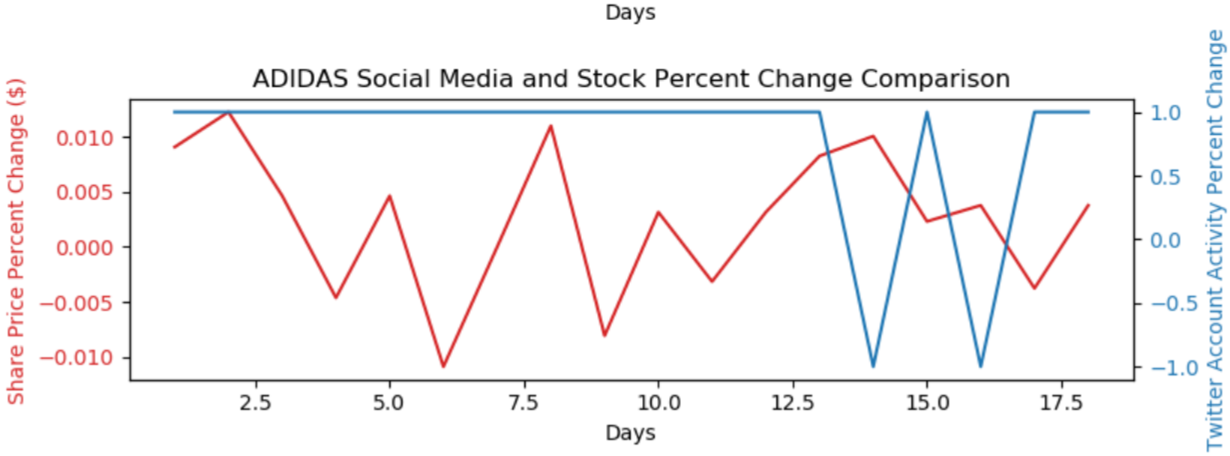


Adidas

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.1266  , p=0.7277  , df_denom=13, df_num=1
ssr based chi2 test:   chi2=0.1558  , p=0.6930  , df=1
likelihood ratio test: chi2=0.1551  , p=0.6938  , df=1
parameter F test:         F=0.1266  , p=0.7277  , df_denom=13, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.1159  , p=0.8917  , df_denom=10, df_num=2
ssr based chi2 test:   chi2=0.3477  , p=0.8404  , df=2
likelihood ratio test: chi2=0.3437  , p=0.8421  , df=2
parameter F test:         F=0.1159  , p=0.8917  , df_denom=10, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.1951  , p=0.8965  , df_denom=7, df_num=3
ssr based chi2 test:   chi2=1.1707  , p=0.7600  , df=3
likelihood ratio test: chi2=1.1244  , p=0.7712  , df=3
parameter F test:         F=0.1951  , p=0.8965  , df_denom=7, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=0.2194  , p=0.9145  , df_denom=4, df_num=4
ssr based chi2 test:   chi2=2.8519  , p=0.5829  , df=4
```

```
likelihood ratio test: chi2=2.5785  , p=0.6306  , df=4
parameter F test:         F=0.2194  , p=0.9145  , df_denom=4, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=0.0918  , p=0.9849  , df_denom=2, df_num=5
ssr based chi2 test:   chi2=2.7546  , p=0.7378  , df=5
likelihood ratio test: chi2=2.4798  , p=0.7795  , df=5
```





Nike

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.0357  , p=0.8531  , df_denom=13, df_num=1
ssr based chi2 test:   chi2=0.0439  , p=0.8340  , df=1
likelihood ratio test: chi2=0.0439  , p=0.8341  , df=1
parameter F test:         F=0.0357  , p=0.8531  , df_denom=13, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.1026  , p=0.9034  , df_denom=10, df_num=2
ssr based chi2 test:   chi2=0.3079  , p=0.8573  , df=2
likelihood ratio test: chi2=0.3048  , p=0.8587  , df=2
parameter F test:         F=0.1026  , p=0.9034  , df_denom=10, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.0728  , p=0.9727  , df_denom=7, df_num=3
ssr based chi2 test:   chi2=0.4366  , p=0.9326  , df=3
likelihood ratio test: chi2=0.4299  , p=0.9340  , df=3
```

```
parameter F test:         F=0.0728  , p=0.9727  , df_denom=7, df_num=3

Granger Causality
number of lags (no zero) 4
ssr based F test:         F=2.8883  , p=0.1644  , df_denom=4, df_num=4
ssr based chi2 test:   chi2=37.5477 , p=0.0000  , df=4
likelihood ratio test: chi2=17.6536 , p=0.0014  , df=4
parameter F test:         F=2.8883  , p=0.1644  , df_denom=4, df_num=4

Granger Causality
number of lags (no zero) 5
ssr based F test:         F=5.4353  , p=0.3142  , df_denom=1, df_num=5
ssr based chi2 test:   chi2=326.1158, p=0.0000  , df=5
likelihood ratio test: chi2=40.0618 , p=0.0000  , df=5
parameter F test:         F=5.4353  , p=0.3142  , df_denom=1, df_num=5
```

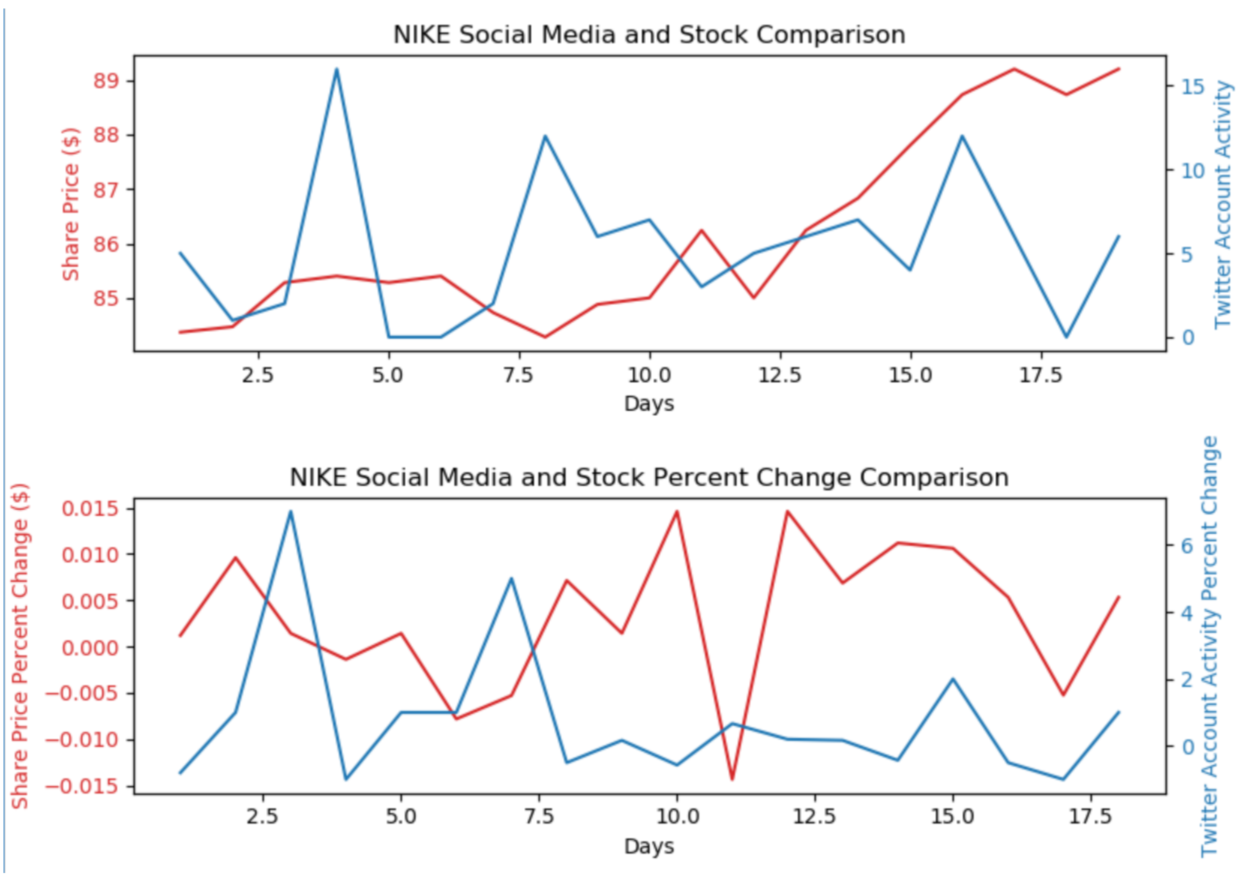


Table 1: Company and Likelihood Ratio Significant Lag

Company	Ratio	Lag(Days)
Wendy's	6.44	3
Dunkin Donuts	9.55	4
Nike	17.65	4
Adidas	2.75	>5
Viacom	29.00	5
Disney	8.39	1

4 Analysis

Within the results, the likelihood ratio test will be focused on. We use an alpha value of 0.10 due to the ambiguity involved with the data since the range is relatively small. The table reports the first instance of a statistically significant p-value (< 0.10) for the likelihood ratio test. In otherwords, the smallest amount of lag is reported for each company. The ratio represents the degree to which the stock prices seem to be affected by Twitter activity. It is clear that the ratios are largest within the entertainment industry, followed by the fast food industry and finally the sports fashion industry.

Though we cannot be certain that the results imply causality, it is clear that in many cases, social media trends are followed by similar trends in share prices. This indicates possible usefulness as a predictor of share price trends if a more comprehensive classification for the social media activity can be generated.