TIM245 Project Report- Team 3
How Tweets influence a Company's Stock Price

Haoyue Gao, Umang Sardesai , Pallavi ,Runjie Chu ,Arghyadeep Giri Srikanth

June 16, 2017

1 Summary

The objective of the project is to study the correlation between a company's Twitter activities and its' stock price. We studied this correlation on Microsoft Stock Price.

2 Prediction

2.1 Problem Definition

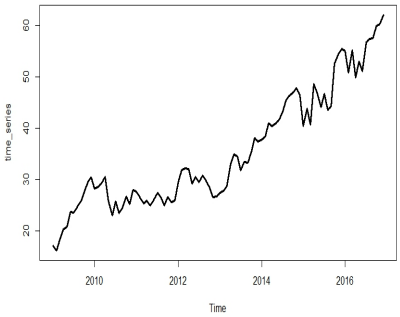
Given the monthly closing price of Microsoft Stock from January 1996 to April 2017, what is the closing price of the "Microsoft" stock in May 2017?

2.2 Dataset Collection and EDA

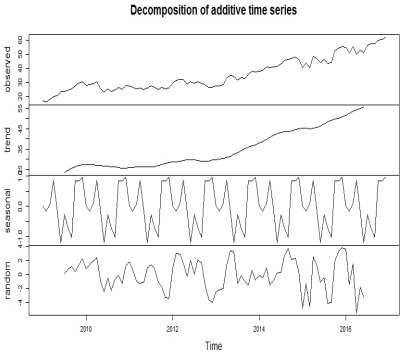
Dataset is collected from Yahoo Finance. A .csv file containing the Microsoft monthly closing price from January 2009 to December 2016 is downloaded from the website. No missing values or duplicates. So no need of any EDA.

2.3 Different Time Series Models employed

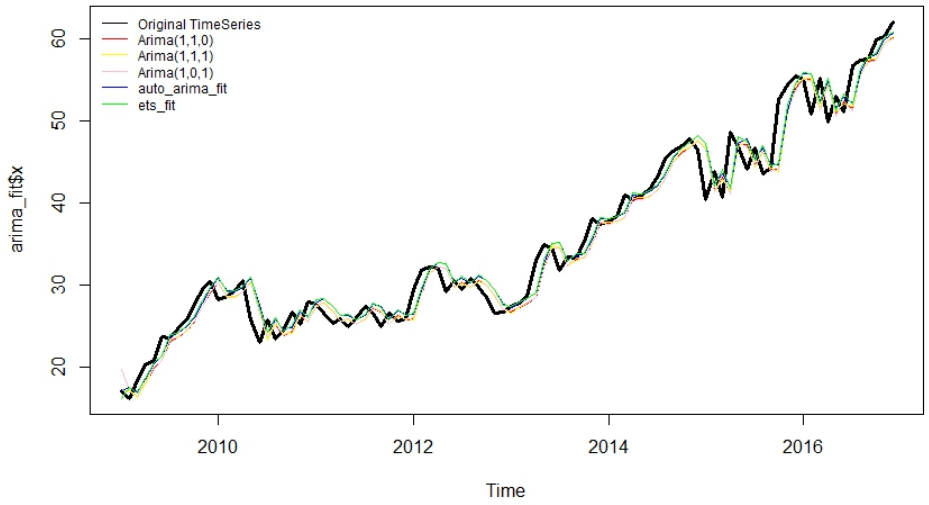
We have tried ARIMA(1,0,0), ARIMA (1,0,1) and ARIMA(1,1,1) models and compared with auto-arima fit and ETS fit. They are The noticeable differences appear during the peaks and lows. This can be attributed to the fact that the lag takes certain amount of time to respond to the sudden changes. We would prefer ARIMA(1,0,0) model because of its lowest BIC.



(a) Original TimeSeries Plot



(b) Decomposition



(c) Various TimeSeries Models

Figure 1: Time Series Modelling to predict Microsoft Monthly Closing Price

2.4 Evaluation and Next Iteration Plan

These results show huge variations when there is a high/low peaks. The results from classifications can be brought in to improve the prediction. Prediction about the high/low peak can be made in advance and some correction can be appropriated to the prediction.

2.5 Conclusion

Overall, we can predict the monthly closing price with a reasonable error which we can atleast expect the baseline monthly price of a stock.

3 Classification

3.1 Problem Definition

Given the number of tweets(#tweets[1]) from the Microsoft Twitter account, number of their retweets(#tweets), favourites(#favourites) and Google trend index of Microsoft(#MSFT_GI) and CEO(#CEO_GI) in a month, can the trend(increasing or decreasing) of the monthly closing price(#Trend) be predicted from the data?

3.2 Data Collection

Our data consists of three parts:

- We have many scripts but finally landed on this script. The greatest benefit of this script is to help us scrape more than the general twitter cap of 3200 tweets. We downloaded about 9000 tweets using this script at once. It also doesnot require any developer keys.
- Google Trends: The monthly trends is downloaded directly as a csv from Google Trends.
- Yahoo Finance Dataset: We have downloaded Microsoft Stock's Closing Prices from 2009 to till 2017.

3.3 Data Cleaning

Data Cleaning is quite intense in our project. Except the Google trends data, both the twitter and yahoo finance data has to be cleaned to get the required attributes.

- **Monthly Tweet Count** The features #tweets, #retweets and # favourites are extracted in this step.

  We used a python script to access twitter API to generate semicolumn seperated data. The raw data consists of 10 attributes: username; date; retweets; favorites; text; geo;mentions; hastags; id; permalink. The goal of the preprocessing is to find out the total number of retweets and favorites for each month.
  Firstly we loaded the raw csv file into Open-Refine with the Parse text into numbers, dates box checked. It will automatically parsed the semicolumn seperated instances into 10 attributes. Then we performed some basic

[1] # is used to represent the name of the feature

data cleaning steps to eliminate duplicates based on username as well as blank value by attribute facet. Next, we applied column split rule and combination rule to transform the date from the original format "2017-05-10 20:32" into "2017-05". Then we exported the cleaned data as a csv file.
After we got the correct format for date, we need to calculate our monthly reweets and favorites using excel. We created two pivot table based on the value of date attribute, retweet attribute and favorite attribute. In the pivot table, we grouped the same month as row and sume of retweet and favorite as the values. As a result, we completed our data preprocessing for the our intetested twttier account and gained the desired monthly retweet and favorite number.

- **Closing Price Trends** The feature #Trend is obtained from this step.

  We had the daily closing price and from that we had to retrieve the monthly trend for the stock. The trend is binary either increasing (1) or decreasing (0). For every month, we plotted the closing price against the date (weekdays) and ran a linear regression model on it. This was done using Python's SciPy library. The linear regression model gives us the gradient which is the slope of the line. A positive slope indicates increasing trend and negative a decreasing trend.

3.4 Exploratory Data Analysis

- Central Tendency and Spread

Attribute	Mean	Median	Mode	Variance	Std-dev	MAD
Num of Tweet	88.33	71.5	54	5328.50	72.99	53.33
Num of Retweets	5623.23	3705	37	45445471.17	6741.33	4475.15
Num of Favorites	6107.30	3312	6	69328211.53	8326.36	5811.33
CEO_GI	24.95	20.5	18	307.79	17.54	9.54
MSFT_GI	23.76	23	24	53.15	7.29	4.77
MCP_Trend	0.66	1	1	0.23	0.48	0.45

- Symmetric/Long Tail
  The distribution of Num of Tweet is long-tailed.
  The distribution of Num of Retweets is long-tailed.
  The distribution of Num of Favorites is long-tailed.
  The distribution of CEO_GI is long-tailed.
  The distribution of MSFT_GI is long-tailed.

- Outliers Analysis There are outliers in Num of Tweet,Num of Retweets, Num of Favorites,CEO_GI and MSFT_GI. The number of tweet, retweets and favorites of each month is uncertain. So sometimes it will be very large.

3.5 Building Different Classification Models

With five attributes : Tweets,Retweets,Favourites,CEO_GI,MSFT_GI

	ZeroR	NaiveBayes	LR	50-KNN	SMO	DT	RF
Accuracy	65.88	52.95	64.70	65.88	65.88	61.17	69.41
F-score	52.3	52.8	59.2	52.3	52.3	50	68

Removing CEO_GI

	ZeroR	NaiveBayes	LR	50-KNN	SMO	DT	RF
Accuracy	65.88	57.65	65.88	65.88	65.88	62.35	62.35
F-score	52.3	57.65	58.9	52.3	52.3	50.6	61.6

Removing CEO_GI and MSFT_GI

	ZeroR	NaiveBayes	LR	50-KNN	SMO	DT	RF
Accuracy	65.88	60	64.7	65.88	65.88	65.88	63.52
F-score	52.3	52.5	56.8	52.3	52.3	52.3	62.6

Removing Number of Tweets, CEO_GI and MSFT_GI

	ZeroR	NaiveBayes	LR	50-KNN	SMO	DT	RF
Accuracy	65.88	60	65.88	65.88	65.88	65.88	49.41
F-score	52.3	52.5	56.1	52.3	52.3	52.3	48.7

Removing Twiiter

	ZeroR	NaiveBayes	LR	50-KNN	SMO	DT	RF
Accuracy	65.88	49.41	63.52	65.88	65.88	61.17	64.70
F-score	52.3	48.7	53	52.3	52.33	50	65

3.6 Evaluation

- The first thing we want to improve is to perform the similar classification task on another task and check the pattern.
- We want to add more relevant features and check whether the accuracy improves.

3.7 Conclusion

We ran 7 different models on different sets of input attributes, but the best accuracy was always with ZeroR (barring one instance). This indicates that our choice of attributes for analyzing the market sentiment is wrong. We even tried to run the models on a specific set of attributes, but the results were more or less similar.

4 Clustering

4.1 Problem Statement

Given the details of 5000 stocks of NASDAQ market(of any given month) where monthly closing price trend and the respective twitter activity is given, are there any interesting clusters in this data?

4.2 Dataset

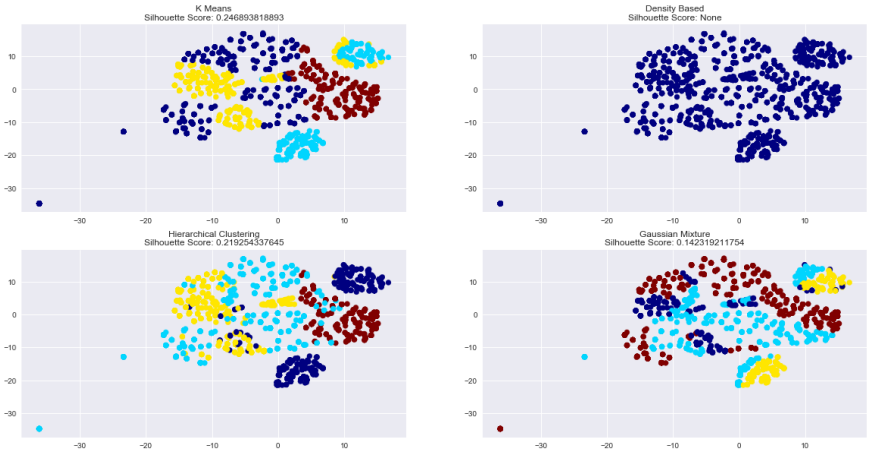
The dataset comprise the following fields: 1) Stock Name , 2) Stock Code , 3) Monthly Closing Price Trend(Increasing/Decreasing), 4) A Company's Twitter Activity (Highly Active/Medium/Low) 5) A CEO's Twitter Activity (Highly Active/Medium/Low)

4.3 Data Preparation

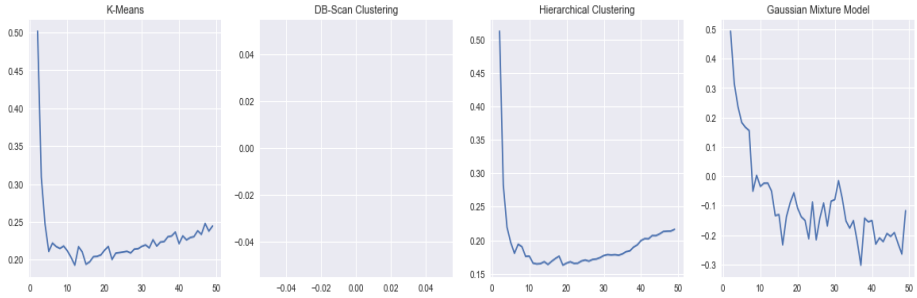
Data Preparation took a lot of time in preparing the data. We ran a script to download 5000 stocks monthly data from yahoo finance and compute the trend of Monthly Closing Prices for each stock. The next big i sto segregate each of the companies and CEOs based on there Twitter Profiles. The company or the CEO who does more than 85 tweets is considered highly active, from 10-85, its medium and less than 10, its low. This kind of binning allows us to find clusters more efficiently.

4.4 Clustering Results

Firstly, We create four clustering models with the four different methods. We can get the cluster labels for each instance. We compute the silhouette score for each clustering. Then plot the results using the T-SNE representation.



Secondly, We loop through different values for k and compute the silhouette score. At the same time, we loop through different values for epsilon and compute the silhouette score. After this step, we plot silhouette score as function of k and epsilon.Then we choose one best method to do the clustering.



From the figure, we can know the best performance is kmeans method. The silhouette coefficient of dbscan is none. So this data-set can't be used in dbscan method. The silhouette coefficient of gaussian_mm is lower than kmeans and hierarchical method. These two methods is useless for the data-set. Although the silhouette coefficient of hierarchical is positive. But the data-set is not hierarchical. So I recommend using kmeans method for the data-set.
From the figure of kmeans, I think the optimal k is 4.

4.5 Conclusion

Actually there are not many interesting clusters we could notice. One interesting trend we could capture is that we could spot the increasing trend when there are CEOs who are very inactive on Social Media. Some of the recommendations would be adding more features. One another thing is to try running this algorithm on different stock market datasets.

5 Association Analysis

5.1 Problem Statement

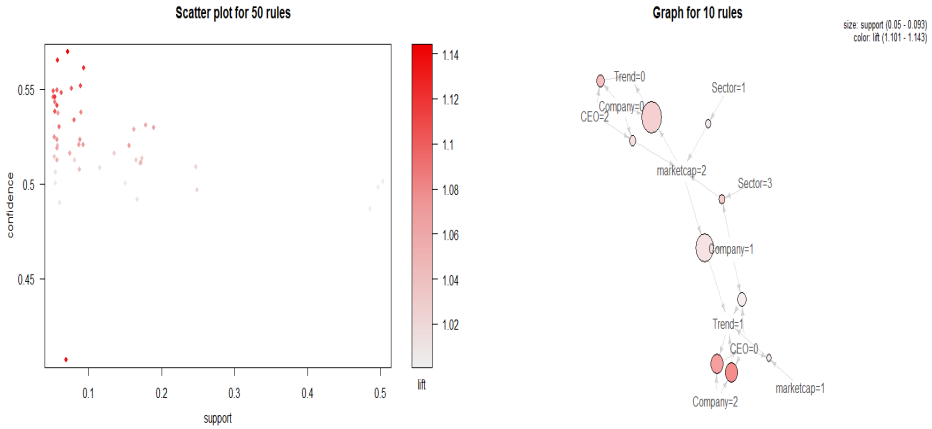
With the association model, we want to find out which factor will contribute to the up trend and down trend of stocks.

5.2 Model Parameters

In order to find the best support and confidence. We tried a lot of different support and confidence. Finally, we choose the support is 0.05. The confidence

is 0.4.

5.3 Association Results



Support = 0.05, Confidence = 0.4

5.4 Evaluation

	lhs		rhs	support	confidence	lift
[1]	{CEO=0,Company=2}	=>	{Trend=1}	0.07103825	0.5697211	1.143418
[2]	{Trend=1,Company=2}	=>	{CEO=0}	0.07103825	0.4074074	1.137463
[3]	{CEO=2,Company=0}	=>	{Trend=0}	0.05812221	0.5652174	1.126517
[4]	{Sector=3,Company=1}	=>	{marketcap=2}	0.05315450	0.5459184	1.120218
[5]	{marketcap=2,Company=0}	=>	{Trend=0}	0.09289617	0.5615616	1.119231
[6]	{CEO=2,Company=0}	=>	{marketcap=2}	0.05563835	0.5410628	1.110254
[7]	{marketcap=2,Company=1}	=>	{Trend=1}	0.08693492	0.5520505	1.107954
[8]	{Sector=1}	=>	{marketcap=2}	0.05265772	0.5380711	1.104115
[9]	{marketcap=1,CEO=0}	=>	{Trend=1}	0.05017387	0.5489130	1.101657
[10]	{CEO=0,Company=1}	=>	{Trend=1}	0.06159960	0.5486726	1.101174

Support = 0.05, Confidence = 0.4

From the rules we can know when CEO doesn't post tweet and the company posts tweet frequently. The stock trend of this company is up. When the CEO posts tweet frequently and the company doesn't post tweet. The stock trend of this company is down.

5.5 Conclusion

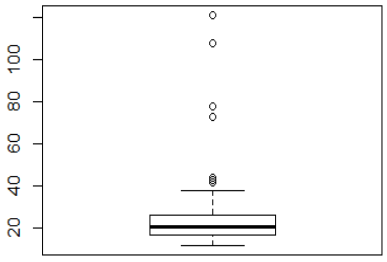
We can get some interesting rules. Although the lift is not very high. It shows our model is useful. Some of the recommendations that can be made are try running the same analysis on the stock exchange data of other countries and check the pattern. The second addition would be adding more relevant features. For example, inclusion of Google Index is one more interesting feature.

6 GitHub Repository

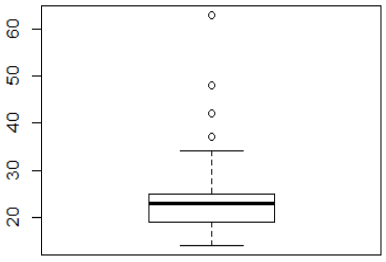
https://github.com/UmangSardesai/TIM245-StockPriceAnalysis

7 Appendix

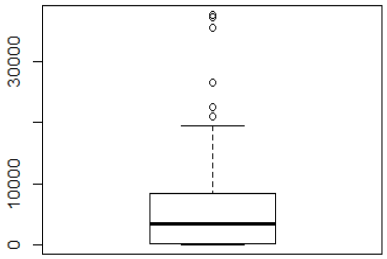
7.1 Box Plots of Classification Datasets



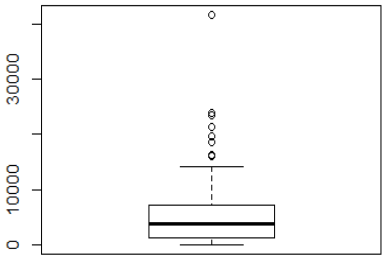
(a) boxplot of CEO_GI



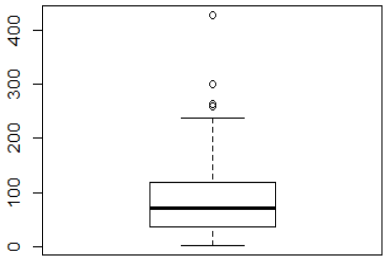
(b) boxplot of MSFT_GI



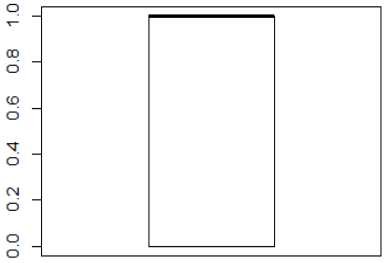
(c) boxplot of Num of Favorites



(d) boxplot of Num of Retweets



(e) boxplot of Num of Tweet



(f) boxplot of MCP_Trend