Data Analytics Final Report
Project 2C: When Can Tweets Lead Financial Markets?

Vinit Jadhav
MSc Business Analytics
vinit.jadhav.15@ucl.ac.uk
15091240

August 26, 2016

Abstract

Predicting stock markets based on social media analytics has been a topic of avid interest for researchers over the past decades. Micro-blogging sites like Twitter and Stocktwits provision for users to react to news about a particular stock using cashtags in their tweets. In our research we investigate if such tweets about a particular stock contain lead time information about the volume of trades of that particular stock. We use concepts from Information Theory and unsupervised learning to demonstrate that the volume of tweets (messages containing the cashtag) contain statistically significant lead time information about the stock trading volume within an intraday interval. Furthermore, we demonstrate that the volume of tweets having an associated sentiment value contain more lead time information about stock trading volumes; volume of bearish sentiments containing the highest lead time information about stock trading volumes. The methodology adopted for this research can be further used to test the predictive power of tweets on other stock performance parameters such as stock returns and volatility.

Contents

1 Introduction

Twitter is a micro blogging site with over 500 million users broadcasting more than 340 million messages (Tweets) daily. StockTwits is a similar platform designed for sharing ideas between investors, traders, and entrepreneurs. Users of these platforms often use cashtags in tweets (messages) to refer to a particular stock to highlight new information / news related to that particular stock. Does this information contain lead time information about the financial markets? In this research we investigate if these tweets contain any lead time information about the volumes of trade about a particular stock. Volume is one of the key indicators used by active traders to gauge money flow. Indicators derived from volume such as on-balance volume and volume by price are extensively used to formulate lucrative trading strategies and are dominant indicators for intraday trading; hence the focus on stock trading volumes. In our analysis we use the concept of entropy from Information Theory and base our research on the mutual information existing between a timeseries indicating daily changes in a twitter parameter and another timeseries indicating a change in stock trading volumes. The twitter data provided by our industry sponsor, PsychSignal, classifies the available tweets about a stock into tweets having bearish sentiments, tweets having bullish sentiments and tweets having no sentiment. We independently test four twitter parameters - volume of all tweets, volume of tweets having sentiment, volume of bullish tweets and volume of bearish tweets; to investigate if they contain lead time information about stock trading volumes by inducing lags of up to nine days in each of the four mentioned twitter parameter timeseries for 100 stocks listed on NSE and NASDAQ. We classify the lag for which mutual information is highest as the optimal lag. Along with optimal lag and maximum mutual information we take into consideration various other features such as the average daily volumes of tweets, percentage of tweets with sentiment, percentage of bullish tweets, percentage of bearish tweets, average daily volumes of bullish tweets, average daily volumes of bearish tweets etc. for 100 stocks independently for the four mentioned twitter parameters. We then perform k-means clustering on this feature dataset to identify stocks with similar features and investigate the mutual information values for stocks in each cluster having high average mutual information values and the critical feature values which resulted in the high mutual information values. We use TSNE dimensionality reduction to visualize these clusters in three reduced dimensions. We derive our final conclusions based on the critical values of these features eventually providing answers to when tweets can lead financial markets. The statistical significance of our results lies in the choice of 100 stocks we make. We only consider those stocks and lags where the mutual information is more leading than trailing and within 5% value of the average randomised mutual information over 10000 block permutations.

2 Literature Review

Prediction of stock markets has been an area of keen interest for those in the academia as well as the industry. Early research on this topic (Fama E.F.1969, 1991) was largely derived from the random walk theory and Efficient Market Hypothesis (EMH) (Fama E.F. 1965) which accounts the movement of the stock markets to new information rather than the present and past prices. However, later works on the same subject (Qian, Bo, Rasheed, & Khaled, 2007), (Gallagher, L. A & Taylor, M. P., 2002), (Kavussanos, M & Dockery, E., 2001) show that the stock market prices do not follow a random walk patter and can be predicted to an extent.

The rise of social media in the early 21st century has given this research a new dimension. Twitter, a micro blogging site with over 500 million users broadcasting more than 340 million messages (Tweets) daily, has made it possible to perceive the moods, thoughts and opinions of a large population with minimal costs to collect this data. Owing to these factors the research on predicting stock market movements based upon twitter data has gathered momentum in the past decade. A few works in this area have focused on the quantitative aspects of the data for predicting the movements in the stock markets. The works of (Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A. & Jaimes, A. 2012) , (Mao,Y., Wei,W., Wang,B.& Liu,B 2012) have focused on the volumes of twitter messages. These works however ignore the qualitative aspects of the data gathered which is a valuable source of information. Contemporary researchers have focused on the qualitative aspects of this data devising mechanisms to extract meaning from this collected data to evaluate their impact on the movement of stock market prices and volumes. Analyzing a bulk of these messages about a particular topic can offer great insights about the topic itself. This methodology of identifying the moods of a group of individuals based upon the unstructured text messages is called sentiment analysis.

The main focus of this methodology is to classify the unstructured text messages on emotional scales. Over the last decade there has been significant progress in sentiment tracking techniques. The work of Pak, A & Paroubek, P. (2010) is a significant piece of work in tracking the sentiments from twitter messages.

Following from this, a study by (Bollen et. All. 2011) investigates if public sentiment expressed in large scale collections of daily twitter posts can be used to study the stock market. Bollens study classifies the tweets into six mood dimensions and generates a time series based upon these dimensions which are then correlated to the Dow Jones Industrial Average (DJIA) to assess their ability to predict changes in the DJIA over time. The study uses granger causality to investigate the correlation and a self-organizing fuzzy neural network to test if the prediction on DJIA can be improved. Based further upon the improvised sentiment analysis techniques a study by (Ilya Zheludev, Robert Smith & Tomaso Aste 2014) demonstrates the presence of statistically significant ex-ante information in social media message sentiments on the future prices of the S&P500 index. The study analyses the performance of intraday sentiment data on intraday financial data over a period of three months. It uses the tweet volumes as well as the tweet sentiments to conduct the study and concludes that tweet sentiment carries greater powers to lead the financial data than the tweet volumes. This research serves as the foundation of our research and we compliment this study further using an unsupervised learning approach to answer the question as to when can tweets lead financial markets.

3 Research

3.1 Research Motivation

The exposure to data analysis concepts as a part of the Data Analytics module for the master's programme and application of these concepts to gain meaningful and purposeful insights supporting the decision making process has been a great motivator behind this research. Previous work experience in the financial sector, a brief understanding of financial markets and the plethora of opportunities data analysis methodologies open towards building models capable of predicting stock movements based upon psycho-social data has generated great curiosity to undertake this research. Lastly the significant contribution of the quality of this research to the module grade cannot be ignored and is a crucial motivating factor behind this research.

3.2 Research Objective

This research aims to answer the question "When can tweets lead financial markets?". Our approach to answer this question is to analyse the data from Twitter and StockTwits for its potential to determine stock trading volumes of organizations belonging to various business sectors as well as various stock exchanges. We aim to investigate the determinative affinity of twitter data parameters to stock trade volumes. The ultimate aim is to identify crucial twitter data parameters and the features of these parameters which answer the above question in consideration.

3.3 Research Hypothesis

We examine the given tweet data set to test if tweets (from Twitter and StockTwits) about a particular stock contains lead time information about its volume of trades. The hypothesis is,

"The volume of tweets having an associated sentiment value for a particular asset contains lead time information about the volume of trades of that particular asset".

We use the term asset to refer to a particular stock and the terms have been used interchangeably. In order to work towards the above hypothesis, we independently test the effect of four different sentiment parameters (timeseries) on the stock-volume timeseries. The four sentiment parameters are listed as below.

- Volume of All Tweets
- Volume of Tweets with sentiment

- Volume of Bullish Tweets
- Volume of Bearish Tweets

3.4 Research Methodology

A flowchart summarizing all the activities in our methodology can be seen in the below figure. Each of the steps in the flowchart have been repeated independently for each of the above four sentiment timeseries

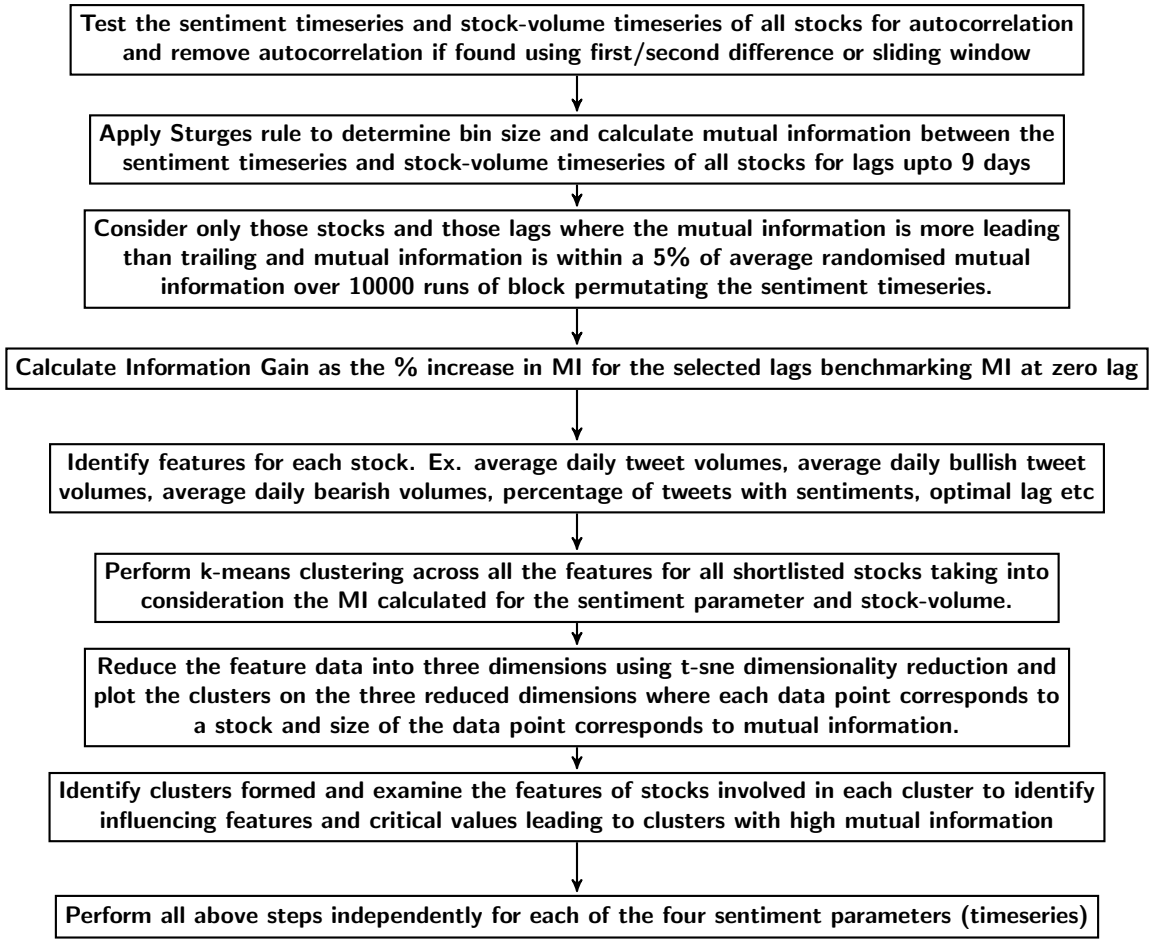


Figure 1: *Flowchart of Research Methodology*

The various methodologies adopted at various steps in the flowchart shown in Figure 1 are listed and described below.

3.4.1 Identification and removal of Autocorrelation

Autocorrelation is a measure of the internal correlation within a time series. It is a way of measuring and explaining internal association between observations in a time series. It can be termed as the similarity between observations as a function of the time lag between them. Given measurements $y_1, y_2, ..., y_n$ at time $t_1, t_2, ..., t_n$ the lag k autocorrelation function A_k is defined as

$$A_k = \frac{\sum_{i=1}^{n-k}(y_i - \hat{y})(y_{i+k} - \hat{y})}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{1}$$

Autocorrelation assumes prime importance in our context because as we undertake the calculation of mutual information, we possibly want each timeseries past to contain the least information about itself signifying minimal autocorrelation in each participant timeseries in the calculation of mutual information. Figure 2 (left exhibit) below shows the autocorrelation of the stock trading volumes timeseries for American Express Company with a lag of upto 35 days.

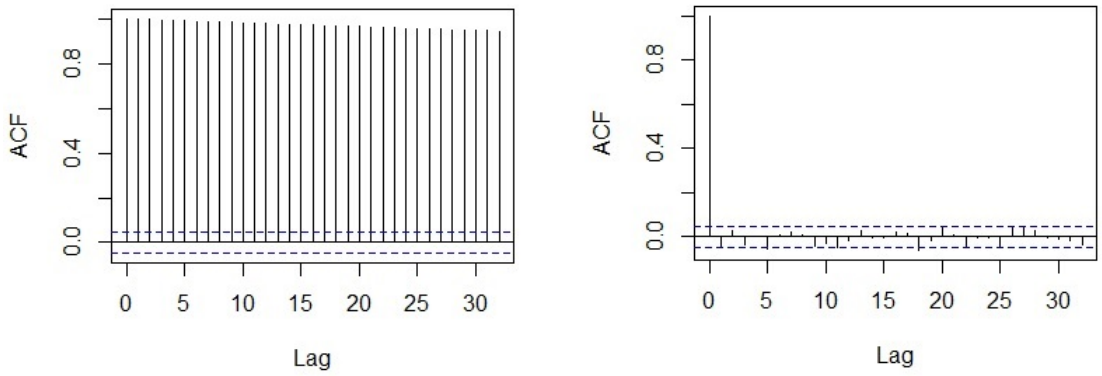


Figure 2: *Autocorrelation observed for stock trading volumes of AXP stock (left exhibit) upto 35 days. The first difference of this timeseries reduces the autocorrelation observed to insignificant levels (right exhibit)*

The removal of autocorrelation, an important task in our research due to the reason indicated above, can be done taking the first difference or subsequent difference in the timeseries values to a point where significant autocorrelation is absent. Figure 2 (right exhibit) shows the reduced autocorrelation of the stock trading volume timeseries for the American Express Company. Note that the autocorrelation at lag 0 is 1 as the time series being compared are the same without any lag. The autocorrelation index assigns a value of +1 to strong positive association, -1 to strong negative association and 0 to no association. A second approach we at times have followed is using a rolling window of width$n$ where a mean of k to n values in the timeseries corresponds to the k value in the new timeseries. The window taken is overlapping.

3.4.2 Shannon Entropy and Mutual Information

Shannon Entropy or simply entropy is a measure of uncertainty the higher the entropy, the more uncertain one is about a random variable. This statement was made quantitative by Shannon who postulated that a measure of uncertainty of a random variable X should be a continuous function of its probability distribution $P_X(x)$ and should satisfy the following conditions:

- It should be maximal when $P_X(x)$ is uniform, and in this case it should increase with the number of possible values X can take.
- It should remain the same if we reorder the probabilities assigned to different values of X.
- The uncertainty about two independent random variables should be the sum of the uncertainties about each of them.

He then showed that the only measure of uncertainty that satisfies all these conditions is the entropy, defined as

$$H(X) = -\sum_x P_X(x) log P_X(x) \tag{2}$$

Analogously, the conditional entropy is the average uncertainty about X after observing a second random variable Y, and is given by

$$H(X|Y) = \sum_y P_Y(y) \Big[-\sum_x log P_{X|Y}(x|y) log\big(P_{X|Y}(x|y)\big) \Big] \tag{3}$$

where $P_{X|Y}(x|y)$ is the conditional probability of x given y.

The Mutual Information *I(X,Y)* is the reduction in uncertainty about variable X after observing the variable Y. It measures the amount of information in variable X given the knowledge of variable Y calculated as,

$$I(X,Y) = H(X) - H(Y) \tag{4}$$

In the specific context of our research we calculate the mutual information between the timeseries of stock trading volumes and each of the four timeseries corresponding to the four sentiment parameters. In order to ascertain if the sentiment timeseries contains lead time information about the financial time series into consideration, we induce lags in the sentiment timeseries by multiples of a day and look into the mutual information calculated at these lags. Table 5 in Appendix - Exhibit 1 shows the mutual information for four stocks upto a lag of 5 days.

3.4.3 Sturges Rule to identify bin size for calculating Mutual Information

It is clear from the preceding section that the concepts of entropy and mutual information in the context of our research stem out of the measure of uncertainty in the distributions under test. The selection of the bin size to represent these distributions is hence an important consideration to achieve significant results. The Sturges Rule helps us in the selection of bin size required to calculate the entropy of our individual timeseries and there after the mutual information between them. According to the Sturges Rule, the optimal number of bins k for a distribution having n values is given by,

$$k = 1 + 3.322(log_{10}n) \tag{5}$$

We use Sturges Rule to arrive at the optimal number of bins used while calculating the mutual information in our research. The bins are of equal size.

3.4.4 Information Gain

Information Gain is the percentage change in mutual information between two time series at a certain lag from that of the mutual information between the two timeseries at no lag. The information gain at lag 0 is 0% as lag 0 is the baseline and indicates no change compared to itself. A positive information gain indicates that the sentiment parameter contains lead time information about the stock performance parameter at that particular lag while a negative information gain indicates no lead time information. Figure 3 (left exhibit) below shows a line plot of the mutual information and the corresponding information gain for a lag upto 5 days between the bearish tweet sentiment volume and the stock trading volumes for Cisco Stock (CSCO). The exhibit to the right shows a similar line plot of the mutual information and the corresponding information gain for a lag upto 5 days between the bullish tweet sentiment volume and the stock trading volumes for Chevron (CVX).

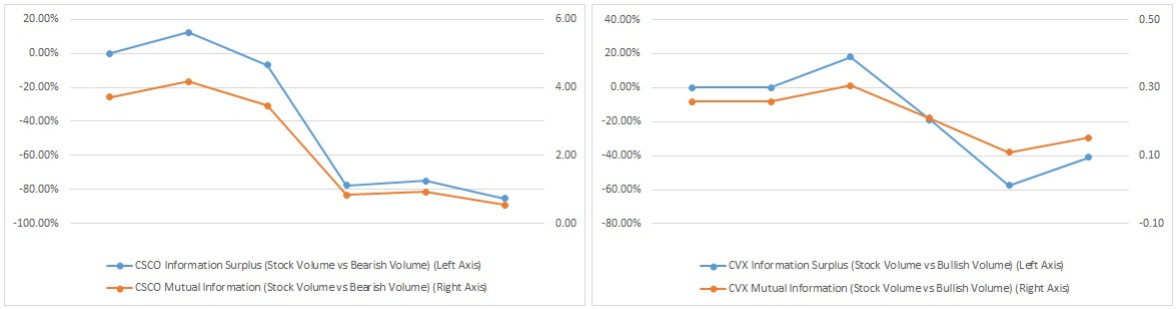


Figure 3: *Line plots showing the mutual information and information gain upto a lag of 5 days between stock volume and volume of bearish sentiment for Cisco stock (CSCO) and between stock volume and volume of bullish sentiment for Chevron stock (CVX). Note the optimal lag is 1 day for CSCO and 2 days for CVX.*

3.4.5 Feature Engineering

The ultimate question we aim to answer from our research is "When can tweets lead financial markets?". The answer to this question lies in analysis of a set of all four sentiment parameters and their characteristic features which result in them containing lead time information about the stock trading volumes of various stocks. In order to achieve this aim we intend to identify the statistical characteristics of each of the four sentiment parameter timeseries of all 100 stocks which result in the maximum mutual information for each stock. Few of the features which we look into are listed below.

- Average volume of daily tweets.
- Average volume of daily tweets with sentiments.
- Average volume of daily bullish tweets.
- Average volume of daily bearish tweets.
- Optimal Lag.
- Average percentage of daily tweets with sentiment.
- Average percentage of daily bullish tweets.
- Average percentage of daily bearish tweets.
- Business Sector of Stock.
- Stock Market on which stock is listed.

3.4.6 k-Means Clustering

"Birds of the same feather flock together". Clustering analysis is the basic of data mining, and K-means algorithm is the simplest clustering algorithm. The K-means clustering algorithm is an unsupervised learning algorithm which classifies items into k clusters. The items within the same cluster are similar while between the clusters are more different. The brief steps of the algorithm can be summarized as below.

- Specify a value for K, and select K items randomly as the clustering centre.
- For the rest (n-K) items, calculate their similarity (distance) to each selected K items.
- Then cluster all items into K groups and calculate the centre for each group and set them as the clustering centre.
- Repeat the process recursively until it meets the clustering criteria.

In the context of our research, we intend to visualize the stocks in the three reduced dimensions where each data point is a stock and the size of each data point is equivalent to the mutual information.

3.4.7 t-sne Dimensionality Reduction

t-distributed stochastic neighbour embedding (t-SNE) is a machine learning algorithm for nonlinear dimensionality reduction that is particularly well suited for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in a scatter plot. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modelled by nearby points and dissimilar objects are modelled by distant points.

4 Results & Testing

Our results our derived after examining the impact of the four sentiment parameters highlighted in the previous section on the stock trading volumes for several stocks between the period from January 1, 2009 to December 31, 2015. To shortlist stocks for our study we sort all the stocks for which twitter data is available in descending order of the volume of total tweets. We start with the first stock and calculate the mutual information between each of the four sentiment parameter timeseries and stock-volume timeseries independently. We only consider those stocks and lags where the mutual information is more leading than trailing and within 5% value of the average randomised mutual information over 10000 block permutations. Once we have 100 stocks which pass these tests, we proceed with k-means (k identified by elbow method - refer appendix) clustering over the identified feature data for the selected stocks and visualize these clusters in TSNE reduced dimensions. Analysis of the clustering results when each of the four sentiment parameters are tested against the stock-volume for selected 100 stocks which pass our filtering criteria, is provided below.

4.1 Stock Trading Volumes vs Volume of All Tweets

Table 1 summarizes the attributes of clusters formed after performing k-means on the feature data of the shortlisted 100 stocks. The mutual information for every stock in this case is calculated between the all tweet volume timeseries and stock-trading volume timeseries.

Cluster No.	No. of Stocks	Max MI	Optimal Lag	(Daily Total Volumes) Tweets	Bullish	Bearish	(Percentage of Tweets) Bullish	Bearish	Sentiment
1	26	7.68	0.15	167.72	28.39	14.93	67.56%	32.44%	24.99%
2	7	4.56	0.43	182.08	28.66	14.45	65.42%	34.58%	21.89%
3	1	1.19	4.00	1324.77	27.58	11.95	69.77%	30.23%	2.98%
4	2	13.80	0.00	2085.71	403.99	230.77	64.09%	35.91%	30.15%
5	7	15.06	0.43	570.95	118.37	67.34	64.58%	35.42%	32.21%
6	2	11.31	0.00	1329.85	265.09	137.00	65.74%	34.26%	30.22%
7	55	4.48	0.58	102.90	16.36	7.29	71.02%	28.98%	22.46%

Table 1: Cluster Analysis - Stock trading volumes vs volume of all tweets. *The highlighted rows show the clusters with maximum average MI and the average feature values corresponding to the cluster.*

We look at the top 3 clusters having the highest average mutual information which are clusters 5,4 and 6 with corresponding average mutual information (difference in shannon entropies in nats) of 15.06, 13.80 and 11.31 respectively. The mutual information for each stock in the cluster corresponds to the maximum mutual information from all the mutual information values calculated between the timeseries for all tweet volumes and stock-volumes for an induced lag of upto 9 days in the all tweet volume timeseries. The top three clusters account for only 11 stocks from the 100 shortlisted stocks. The average optimal lag (lag for which mutual information is highest) corresponding to each cluster is less than 1 indicating that leading mutual information is highest when looking into intra-day data. For a holistic review

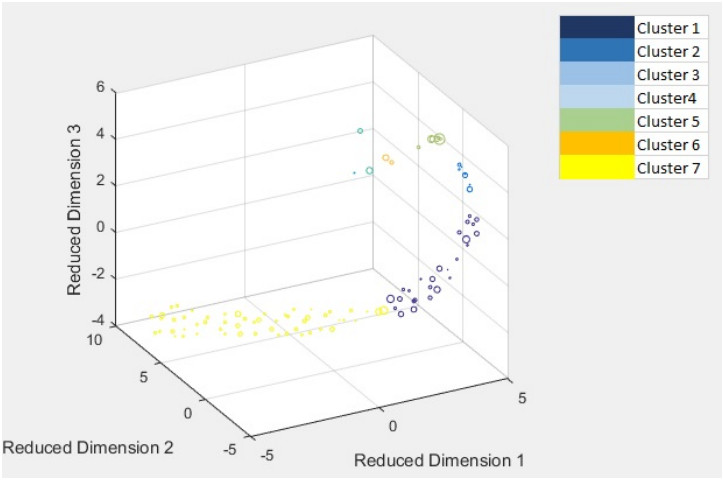


Figure 4: Cluster Visualization. *Each data point in the cluster corresponds to a stock and size of the data point corresponds to the maximum mutual information. Dimensions reduced using TSNE.*

of the feature values that result in high mutual information of stocks in the top three clusters, we normalize the feature values between 0 and 1 and study them using a radar plot (Figure 3). The top three clusters exhibit a higher percentage of daily tweets which have a sentiment value associated with them. On an average, more than 30% of daily tweets for stocks under these clusters have a sentiment value associated with them. The percentage of daily bearish tweets for stocks in the top three clusters influences the mutual information significantly. The stocks under these clusters had more than 34% daily bearish tweets on an average. The volume of all tweets including volumes of bullish and bearish tweets was significantly higher for stocks in these clusters compared to the same volumes for stocks in the less significant clusters. The average values of each of these three features for all stocks under the top three clusters was more than thrice of the corresponding values in other clusters with an exception of cluster 3. Cluster 3 which includes only one stock which is AP (Ampco Pittsburgh Corporation) is a clear outlier.

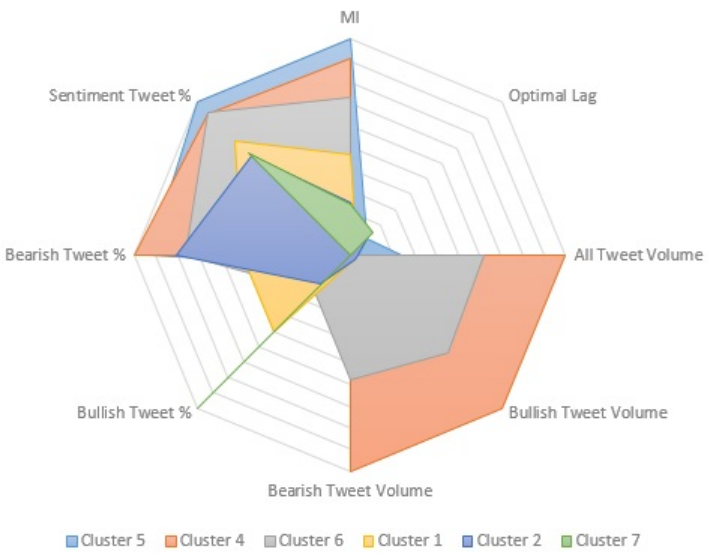


Figure 5: Influence of features on cluster. *Radar plot showing that % of sentiment tweets, % of bearish tweets and the volumes of tweets (all, bull & bear) influence high MI (all tweet volume vs stock volume.)*

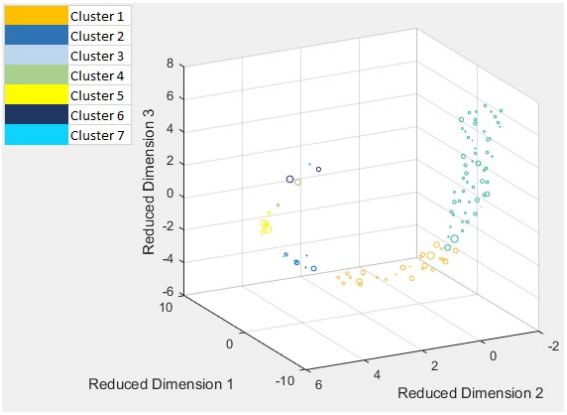
4.2 Stock Trading Volumes vs Volume of Tweets having Sentiment

Table 2 summarizes the attributes of clusters formed after performing k-means on the feature data of the shortlisted 100 stocks. The mutual information for every stock in this case is calculated between the volume-of-tweets-with-sentiments timeseries and stock-trading volume timeseries.

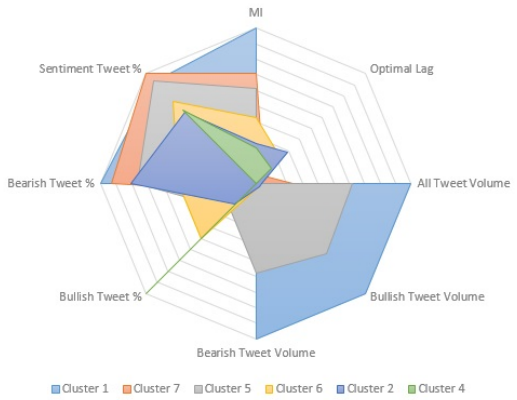
Cluster No.	No. of Stocks	Max MI	Optimal Lag	(Daily Total Volumes) Tweets	Bullish	Bearish	(Percentage of Tweets) Bullish	Bearish	Sentiment
1	26	19.48	0.00	2085.71	403.99	230.77	64.09%	35.91%	30.15%
2	7	5.74	1.14	182.08	28.66	14.45	65.42%	34.58%	21.89%
3	1	1.05	4.00	1324.77	27.58	11.95	69.77%	30.23%	2.98%
4	2	5.33	0.56	102.90	16.36	7.29	71.02%	28.98%	22.46%
5	7	12.31	0.00	1329.85	265.09	137.00	65.74%	34.26%	30.22%
6	2	8.86	0.92	167.72	28.39	14.93	67.56%	32.44%	24.99%
7	55	14.12	0.29	570.95	118.37	67.34	64.58%	35.42%	32.21%

Table 2: Cluster Analysis - Stock trading volumes vs volume of tweets with sentiment. *The highlighted rows show the clusters with maximum average MI and the average feature values corresponding to the cluster. The average MI for top three clusters is more than that observed for top three clusters when calculating MI between all tweet volume and stock-volumes.*

We look at the top 3 clusters having the highest average mutual information which are clusters 1,7 and 5 with a corresponding average mutual information (difference in shannon entropies in nats) of 19.48, 14.12 and 12.31 respectively. The average mutual information, calculated between volume of tweets having



(a) Cluster Visualization.



(b) Influence of features on cluster.

Figure 6: (a) *Each data point in the cluster corresponds to a stock and size of the data point corresponds to the maximum mutual information. Dimensions reduced using TSNE.* (b) *Radar plot showing that % of sentiment tweets and % of bearish tweets influence high MI (volume of tweet with sentiment vs stock volume.)*

sentiment (bullish+bearish) and stock-volume in this scenario, for the most significant cluster is 30% higher than the mutual information in the previous scenario where volume of all tweets was concerned. Also the number of stocks in the top three clusters has increased from 11 in the previous scenario to 88. The average optimal lag corresponding to the most significant cluster is 0 and less than 1 for the remaining two clusters indicating that leading mutual information is highest when looking at intra-day data. For a holistic review of the feature values that result in high mutual information of stocks in the top three clusters, we normalize the feature values between 0 and 1 and study them using a radar plot (Figure 4-b). Again, the top three clusters exhibit a higher percentage of daily tweets which have a sentiment value associated with them. More than 30% of daily tweets for stocks under these clusters have a sentiment value associated with them on an average compared to less than 25% observed for other clusters. The percentage of daily bearish tweets for stocks in the top three clusters influences the mutual information significantly. The stocks under these clusters had more than 34% daily bearish tweets on an average. Percentage of bullish tweets did not influence the mutual information values. The volume of all tweets including volumes of bullish and bearish tweets was significantly higher for stocks in these clusters compared to the same volumes for stocks in the less significant clusters. The average values of each of these three features for all stocks under the top three clusters was more than thrice of the corresponding values in other clusters with an exception of cluster 3. Again, AP (Ampco Pittsburgh Corporation), the single stock in cluster 3 is a clear outlier.

4.3 Stock Trading Volumes vs Volume of Bullish Tweets

Cluster No.	No. of Stocks	Max MI	Optimal Lag	(Daily Total Volumes) Tweets	Bullish	Bearish	(Percentage of Tweets) Bullish	Bearish	Sentiment
1	26	5.22	1.71	182.08	28.66	14.45	65.42%	34.58%	21.89%
2	7	8.06	0.88	167.72	28.39	14.93	67.56%	32.44%	24.99%
3	1	4.64	0.42	102.90	16.36	7.29	71.02%	28.98%	22.46%
4	2	1.08	4.00	1324.77	27.58	11.95	69.77%	30.23%	2.98%
5	7	15.82	0.00	2085.71	403.99	230.77	64.09%	35.91%	30.15%
6	2	12.98	1.29	570.95	118.37	67.34	64.58%	35.42%	32.21%
7	55	12.08	1.00	1329.85	265.09	137.00	65.74%	34.26%	30.22%

Table 3: Cluster Analysis - Stock trading volumes vs volume of bullish tweets. *The highlighted rows show the clusters with maximum average MI and the average feature values corresponding to the cluster.*

We further investigate to what extent bullish tweet volumes contain lead time information about stock volumes. Table 3 on the previous page summarizes the attributes of clusters formed after performing k-means on the feature data of the shortlisted 100 stocks. The mutual information for every stock in this case is calculated between the bullish volume timeseries and stock-trading volume timeseries. We look at the top 3 clusters having the highest average mutual information.

Clusters 5, 6 and 7 are the most significant clusters with a corresponding average mutual information of 15.82, 12.98 and 12.08 respectively. The mutual information for each stock in the cluster corresponds to the maximum mutual information from all the mutual information values calculated between the timeseries for bullish tweet volumes and stock-volumes for an induced lag of upto 9 days in the bullish tweet volume timeseries. The top three clusters account for 64 stocks from the 100 shortlisted stocks. The average optimal lag corresponding to the most significant cluster is 0 while 1.29 and 1 for the remaining two clusters. It indicates that bullish tweet volumes carry lead time information about stock trading volumes for the same trading day as well as the next trading day. For a holistic review of the feature values that result in high mutual information of stocks in the top three clusters, we normalize the feature values between 0 and 1 and study them using a radar plot (Figure 6). The top three clusters exhibit a higher percentage of daily tweets which have a sentiment value associated with them. More than 30% of daily tweets for stocks under these clusters have a sentiment value associated with them on an average. The percentage of daily bearish tweets for stocks in the top three clusters influences the mutual information significantly. The stocks under these clusters had more than 34% daily bearish tweets on an average. The volume of all tweets including volumes of bullish and bearish tweets was significantly higher for stocks in these clusters compared to the same volumes for stocks in the less significant clusters. The average values of each of these three features for all stocks under the top three clusters was more than thrice of the corresponding values in other clusters with an exception of cluster 4 which has two stocks with high volume of tweets but significantly low MI. The optimal lag observed for this outlier cluster is 4 days.

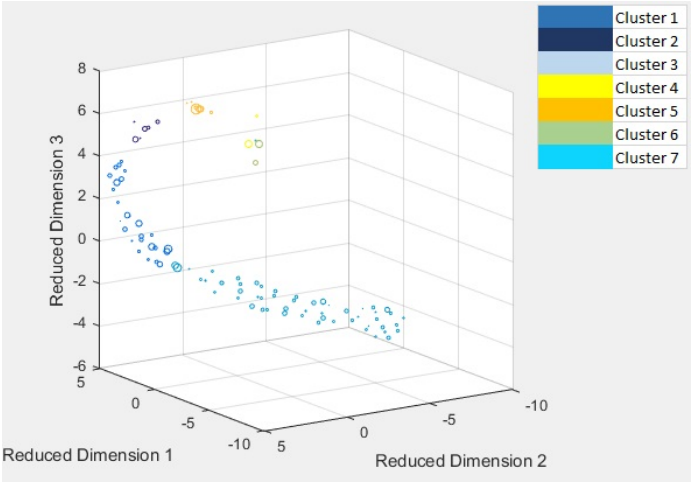


Figure 7: Cluster Visualization. *Each data point in the cluster corresponds to a stock and size of the data point corresponds to the maximum mutual information. Dimensions reduced using TSNE.*

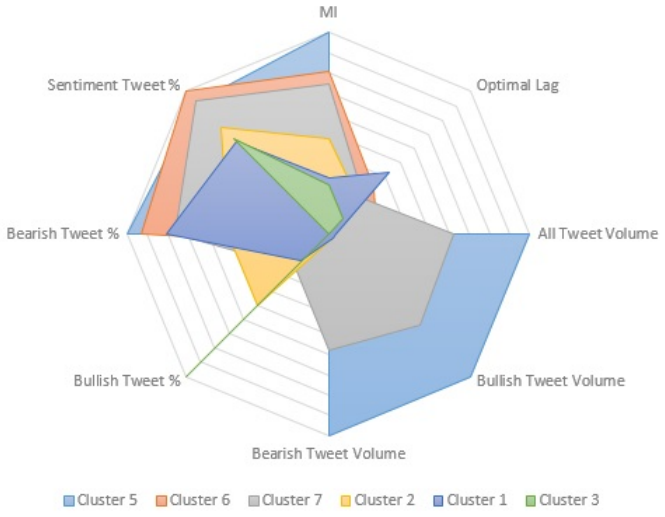


Figure 8: Influence of features on cluster.*Radar plot showing that % of sentiment tweets, % of bearish tweets and the volumes of tweets (all, bull & bear) influence high MI (bullish tweet volume vs stock volume.)*

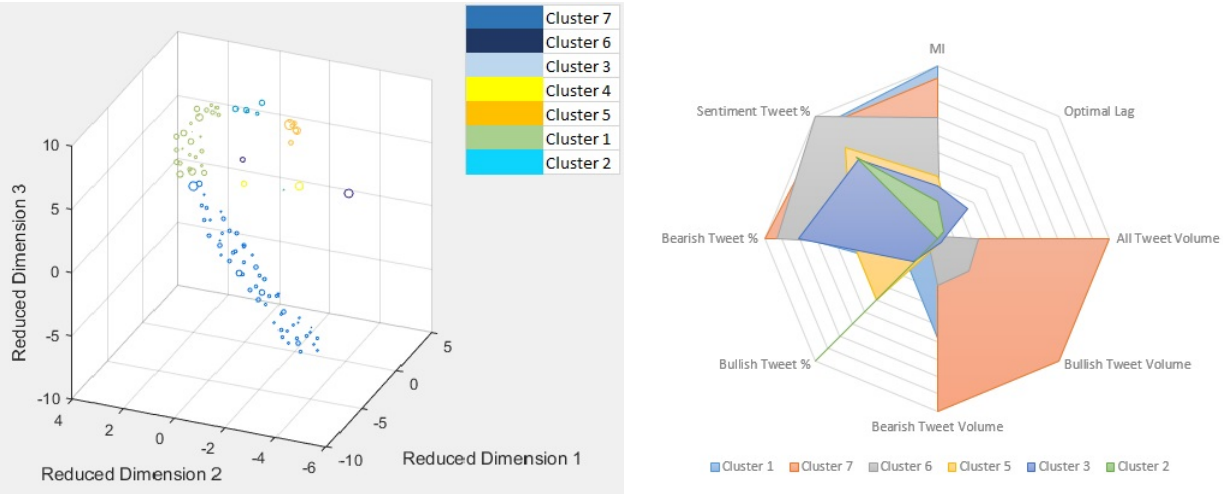
4.4 Stock Trading Volumes vs Volume of Bearish Tweets

Table 4 summarizes the attributes of clusters formed after performing k-means on the feature data of the shortlisted 100. The mutual information for every stock in this case is calculated between the bearish tweet volume timeseries and stock-trading volume timeseries.

Cluster No.	No. of Stocks	Max MI	Optimal Lag	(Daily Total Volumes) Tweets	Bullish	Bearish	(Percentage of Tweets) Bullish	Bearish	Sentiment
1	26	19.65	0.00	1329.85	265.09	137.00	65.74%	34.26%	30.22%
2	7	4.90	0.38	102.90	16.36	7.29	71.02%	28.98%	22.46%
3	1	6.63	1.71	182.08	28.66	14.45	65.42%	34.58%	21.89%
4	2	0.79	7.00	1324.77	27.58	11.95	69.77%	30.23%	2.98%
5	7	7.56	0.81	167.72	28.39	14.93	67.56%	32.44%	24.99%
6	2	14.00	0.14	570.95	118.37	67.34	64.58%	35.42%	32.21%
7	55	18.32	0.00	2085.71	403.99	230.77	64.09%	35.91%	30.15%

Table 4: Cluster Analysis - Stock trading volumes vs volume of bearish tweets. *The highlighted rows show the clusters with maximum average MI and the average feature values corresponding to the cluster.*

We look at the top 3 clusters having the highest average mutual information which are clusters 1, 7 and 6 with a corresponding average mutual information (difference in shannon entropies in nats) of 19.65, 18.32 and 14.00 respectively. The mutual information for each stock in the cluster corresponds to the maximum mutual information from all the mutual information values calculated between the timeseries for bearish tweet volumes and stock-volumes for an induced lag of upto 9 days in the bearish tweet volume timeseries. The top three clusters account for 83 stocks from the 100 shortlisted stocks which is more than the number of stocks observed in the bullish scenario. The average optimal lag corresponding to the top two clusters is 0 and almost 0 for the remaining cluster indicating that leading mutual information is highest when looking at intra-day data in this scenario.



(a) Cluster Visualization.

(b) Influence of features on cluster.

Figure 9: (a) *Each data point in the cluster corresponds to a stock and size of the data point corresponds to the maximum mutual information. Dimensions reduced using TSNE.* (b) *Radar plot showing that % of sentiment tweets, % of bearish tweets and the volumes of tweets (all, bull & bear) influence high MI (bearish tweet volume vs stock volume.)*

For a holistic review of the feature values that result in high mutual information of stocks in the top three clusters, we normalize the feature values between 0 and 1 and study them using a radar plot (Figure 7-b). The top three clusters exhibit a higher percentage of daily tweets which have a sentiment value

associated with them. More than 30% of all daily tweets for stocks under these clusters have a sentiment value associated with them on an average. The percentage of daily bearish tweets for stocks in the top three clusters influences the mutual information significantly. The stocks under these clusters had more than 34% daily bearish tweets on an average. The volume of all tweets including volumes of bullish and bearish tweets was significantly higher for stocks in these clusters compared to the same volumes for stocks in the less significant clusters. The average values of each of these three features for all stocks under the top three clusters was more than thrice of the corresponding values in other clusters with an exception of cluster which has two stocks with high volume of tweets but significantly low MI. The optimal lag observed for this outlier cluster is 7 days.

4.5 Statistical Significance

It is important to ascertain that our results our valid and have statistical significance. In obtaining the results there are two major steps which ensure the robustness of our results. The first one is when we filter stocks based upon the fact that the mutual information is more leading than trailing. To achieve this we calculate the mutual information by inducing lags in the stock-volume time series upto 9 days lags and compare the mutual information with the original mutual information achieved by lagging the sentiment timeseries for those particular lags. We only consider those stocks and those lags where the mutual information with lagging sentiment timeseries for a specific lag is greater than the mutual information with lagging stock-volume time timeseries for that particular lag.

The second one is when we ensure that the mutual information is significant. To achieve this we use block permutations on each of the sentiment timeseries 10000 times and calculate the average mutual information for these 10000 runs. We compare this average mutual information values to the original mutual information values and consider only those stocks where the original mutual information values are within a 5% range of the average mutual information values.

5 Discussion and Future Work

The below points highlight the important aspects of our results which were aimed to uncover when do tweets contain lead time information about the stock trading volumes.

- **Tweets contain more lead time information about stock trading volumes when the volumes of tweets having sentiment are considered.** The top three clusters with highest mutual information in section 4.1, where we considered the volume of all tweets against the stock volume, contained only 11 among 100 stocks which went up to 88 in section 4.2 where we considered only the volume of tweets having sentiments against stock volumes. Also the average mutual information for stocks in the most significant cluster in the later scenario was 30% higher than the average mutual information of stocks in the most significant cluster in the former scenario.

- **Tweets contain lead time information about stock trading volumes when the volume of tweets is significantly high.** It was seen in the section 4 that the average volume of tweets (all, bullish and bearish) for stocks belonging to the most significant clusters in each scenario was at least four times than that of the stocks belonging to the non significant clusters. A daily tweet volume of more than 500 tweets for a stock corresponded to significantly high mutual information values. Similarly bullish tweet volumes exceeding 100 tweets a day and bearish tweet volumes exceeding 60 tweets a day corresponded to high mutual information values.

- **Tweets contain lead time information about stock trading volumes when the tweets are bearish in nature.** Section 4.3 and 4.4 helped us analyze the effect of bearish and bullish tweets respectively on stock volumes independently. The average mutual information of all stocks in the most significant cluster in the bearish scenario (4.4) was 25% higher than that of all stocks in the most significant cluster in the bullish scenario (4.3). Also the number of stocks in the top three clusters with high mutual information increased from 64 in the bullish scenario to 83 in the bearish scenario. In all subsections of section 4, percentage of bearish tweets has been more influential to high mutual

information values than the percentage of bearish tweets. This explains the panic selling prevalent in the stock market when certain critical news about an organization is available.

- **Tweets contain lead time information about stock trading volumes dominantly within an intraday interval.** For all subsections under section 4 we observed that the average value of optimal lag for top three clusters with highest mutual information was less than one. This means that the maximum mutual information for most stocks was without any lag in the sentiment timeseries. In the case of observing bullish volumes for lead time information the optimal lag was 1 day for two significant clusters.
- The same methodology can be used to study the effect of other sentiment parameters like sentiment intensities on other financial parameters like stock returns or volatility.
- Carrying forward from this methodology, the mutual information between tweet volumes and stock trading volumes can be studied by inducing hourly lags taking into consideration the tweet data at a granularity of hours. The granularity of tweets in the dataset provided to us was daily and hence such analysis could not be undertaken.
- Pointwise mutual information (PMI) can be used as an alternative methodology where we calculate the mutual information on a daily basis irrespective of stocks. Calculating PMI's and examining the underlying daily feature values resulting in the PMI values can thereafter be used to construct models which predict the future PMI's giving us lead time information about financial parameters like volumes, log returns etc.

6 Code

The results are derived from the code setup in R language. R facilitates various packages which have been used to achieve these results. I have extensively used the package "sqldf" to munge the data from csv files into tables which can be queried on the fly for further analysis. The "tseries" package provides ample functionality for time series operations. The "entropy" package is used for mutual information calculations. The TSNE dimensionality reduction and clustering plots have been created in MATLAB. Our group has also setup a code framework in Python entailing the same research methodology.

7 References

- Ilya Zheludev, Robert Smith & Tomaso Aste. When Can Social Media Lead Financial Markets? (2014)
- Tharsis T. P. Souza & Tomaso Aste. A nonlinear impact: evidences of causal effects of socialmedia on market prices (2016)
- Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, Miha Grar, Igor Mozeti. The Effects of Twitter Sentiment on Stock Price Returns. (2015) available at *http://journals.plos.org/plosone/article?id= 10.1371/journal.pone.0138441*
- Johan Bollen,Huina Mao1,,Xiao-Jun Zeng. Twitter mood predicts the stock market. (2011) available at *http://arxiv.org/pdf/1010.3003.pdf*
- Guy Nason. Time Series Introduction. (2016) available at *http://www.maths.bris.ac.uk/ guy/Research/LSTS/STSIntro.html*
- Dr. Peter E. Latham, Dr. Yasser Roudi, Gatsby Computational Neuroscience Unit, University College London. Mutual information. (2009) available at *http://www.scholarpedia.org/article/Mutual_informationa*
- Herve Abdi, Lynne J. Williams. Principal Component Analysis. (2010) available at *http://ead.ipleiria.pt/ucs 201415/pluginfile.php/168687/mod_resource/content/14/ABDI-WIRE%20CS-PCA%202010.pdf*

- Hong Zhang, Hong Yu, Ying Li, Baofang Hu. Improved K-means Algorithm Based on the Clustering Reliability Analysis. (2015) available at *http://www.atlantis-press.com/php/pub.php?publication=isci-15&frame=http%3A//www.atlantis-press.com/php/paper-details.php%3Fid%3D17709*
- Mao, Y.,Wei, W., Wang, B.& Liu, B. Correlating S&P500 stocks with Twitter data. (2012)
- Pak, A & Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. (2010)
- Fama, E. F. Journal of Finance (1991)
- Fama, E. F. The Journal of Business (1965)
- Fama, Eugene F, e. a. International Economic Review 1 (1969)
- Gallagher, L. A & Taylor, M. P. Southern Economic Journal (2002)
- Kavussanos, M & Dockery, E. Applied Financial Economics (2001)
- Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A. & Jaimes, A. Correlating Financial Time Series with Micro Blogging Activity (2012)
- TSNE. Laurens van der Maaten. available at *https://lvdmaaten.github.io/tsne/*
- Course Notes. COMPG011: Data Analytics.

Appendix

Exhibit 1. Sample Mutual Information at lags upto 5 days.

Stock Symbol	Financial Parameter	Sentiment Parameter (Lagged)	Mutual Information at Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5
AXP	Stock Volume	Sentiment Tweet Volume	2.1820857	2.6096223	2.6867141	1.5072481	1.4941785	1.1590174
CSCO	Stock Volume	Bearish Volume	3.7155792	4.1697848	3.4664862	0.8349384	0.9319044	0.5409299
CVX	Stock Volume	Bullish Volume	0.2588746	0.2589571	0.3056294	0.210148	0.1102846	0.1527313
DD	Stock Volume	All Tweet Volume	5.9289487	4.2274612	2.2523109	0.6340903	1.042265	1.3869931

Table 5: *Mutual information for a few stocks from taking into consideration stock volume and all four identified sentiment parameters independently. The mutual information is calculated when the sentiment timeseries is not lagged as well as lagged upto 5 days with the stock volume timeseries*

Exhibit 2. Sample Information Gain

Stock Symbol	Financial Parameter	Sentiment Parameter	Information Gain at Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5
AXP	Stock Volume	Sentiment Tweet Volume	0.00%	19.59%	23.13%	-30.93%	-31.53%	-46.88%
CSCO	Stock Volume	Bearish Volume	0.00%	12.22%	-6.70%	-77.53%	-74.92%	-85.44%
CVX	Stock Volume	Bullish Volume	0.00%	0.03%	18.06%	-18.82%	-57.40%	-41.00%
DD	Stock Volume	All Tweet Volume	0.00%	-28.70%	-62.01%	-89.31%	-82.42%	-76.61%

Table 6: *Information Gain for the stocks from Table 5. The information gain at no lag corresponds to 0%. The information gain of lags upto 5 days is shown in the table. Lag 2 is the optimal lag for Chevron stock (CVX) as it corresponds to the highest information gain.*

Exhibit 3. Selection of k

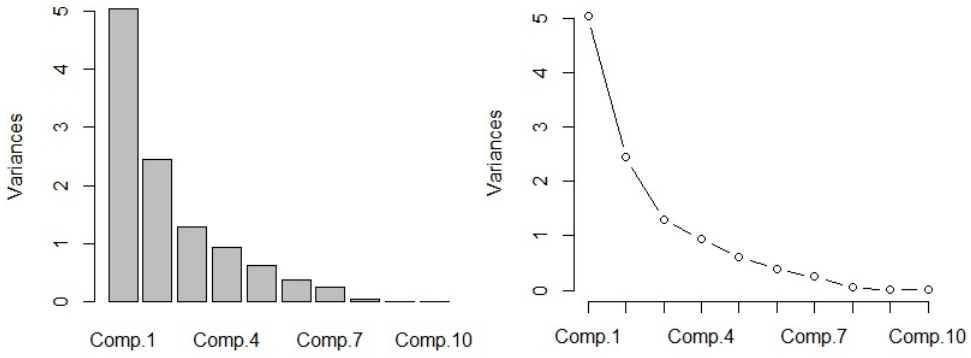


Figure 10: Selection of k. *We select the k corresponding to the number of principle components which cover 90% variance of the entire feature dataset.*