Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica

TWITTER DATA ANALYSIS FOR FINANCIAL MARKETS

Relatore:
Chiar.mo Prof.
DANILO MONTESI

Presentata da:
ALESSANDRO COCILOVA

Sessione III
Anno Accademico 2013-2014

A chi è sempre al mio fianco
negli alti e nei bassi.

Contents

Chapter 1

Sommario

Negli ultimi anni la crescita di Twitter è stata esponenziale, portando il noto social network ad avere quasi 300 milioni di utenti e circa 500 milioni di post[1], conosciuti come "tweet", per giorno. Ormai non è più, come inteso inizialmente, una semplice piattaforma di microblogging, ma è divenuto una fondamentale e popolare fonte di informazione ufficiale (la sua struttura di following unidirezionale lo rende molto popolare tra personaggi pubblici e portali web di quotidiani ed agenzie stampa, facendone di fatto l'ufficio stampa di internet) e virale.

Molti hanno iniziato a chiedersi se questi tweet, su i più disparati argomenti, avessero potere predittivo, come una sorta di rappresentazione della coscienza popolare, e sono stati fatti molti studi in vari campi: dallo studio della diffusione delle malattie[2] alla previsione dei risultati elettorali[3], fino ai risultati delle partite di calcio[4].

Sono state svolte molte ricerche anche in ambito finanziario, in cui i ricercatori hanno tentato di capire se esista un qualche tipo di correlazione tra i tweet e l'andamento dei mercati, ovvero se i tweet influenzino o predicano i mercati finanziari. In questo lavoro cerchiamo di rispondere alle seguenti domande:

1. C'è una qualche relazione tra il numero di tweet che vengono generati in un dato giorno ed il volume delle azioni scambiate in quello stesso giorno?

2. C'è una qualche relazione tra l'umore espresso dai tweet ed il prezzo delle azioni?

3. Qual è la struttura del grafo che descrive le relazioni tra gli utenti?

Per le analisi di tipo quantitativo (volume tweet-volume azioni) e qualitativo (sentiment analysis), abbiamo selezionato dal nostro dataset i tweet

riguardanti sei aziende, operanti in diversi settori, essendo queste le più twittate. Abbiamo scelto quindi di osservare l'andamento delle curve in due mesi, settembre 2012, scelto perché contiene tweet riguardanti un importante evento Apple, che abbiamo ritenuto interessante analizzare dal nostro punto di vista, e maggio 2013, scelto casualmente.

In particolare l'analisi di tipo qualitativo, che possiamo chiamare "Vocabulary Analysis" o "Term Weight Analysis", è stata fatta basandosi sul vocabolario di termini pesati manualmente descritto in [26]. Ogni tweet è stato analizzato, diviso per parole, per ogni parola che abbiamo ritrovato anche nel vocabolario, abbiamo sommato il suo peso alla somma parziale di quel tweet. I tweet la cui somma è risultata positiva sono definiti dall'umore positivo, quelli in cui il risultato è negativo sono stati definiti dall'umore negativo, in caso la somma risultasse zero, li chiameremo tweet dall'umore neutro.

I grafici che mostrano i risultati dell'analisi quantitativa e quelli che comparano, per ogni azienda, l'analisi qualitativa con l'andamento sui mercati del prezzo delle relative azioni, su base giornaliera, possono essere trovati in 5.2.

Le nostre scoperte mostrano che è presente una correlazione tra la quantità di tweet generati ed il volume delle azioni scambiate, in particolare che questi due valori sono correlati in media nel 78% dei giorni. Abbiamo anche mostrato che la distanza media fra queste due curve, nei giorni correlati, è sempre minore del 16%. La sentiment analysis mostra risultati contrastanti, evidenziando in alcuni casi un apparente potere predittivo, mentre apparendo non sufficiente in altri.

Abbiamo inoltre analizzato (e rappresentato nel grafo 5.15), in questo caso per tutti i tweet contenuti nel nostro ampio dataset, le relazioni presenti tra gli utenti, scoprendo un pattern scalabile tra chi segue e chi è seguito. Ci sono un piccolo numero di fulcri, corrispondenti ad account di prestigiosi siti di informazione economica e bloggers, ed un vasto numero di account che si limitano a seguire; abbiamo trovato inoltre un basso numero di account disconnessi, cioè che non parlano strettamente di argomenti finanziari, nella nostra rete.

Intendiamo in futuro estendere il lavoro operando su dataset più ricchi, considerando un maggior numero di azioni per un periodo di tempo più lungo, proponendo un modello su cui basare strategie di trading, migliorare la sentiment analysis, usando tecniche di natural language processing o di machine learning.

Chapter 2

Introduction

Social networking services have become more than just a tool to connect with friends. There are millions of users on these websites and these users generate vast amounts of data. Twitter is an example of such a social network platform, where the users are connected with each other in a unidirectional manner. That is a link does not mean two users are friends with each other. Users which follow a particular user are called followers of that user. Thus, it is not necessary that a user A being followed by B should also follow back B. In the last few years, the Twitter has transformed from its original intended purpose of a simple, personal, a microblogging site to mega content generator. Twitter has approximately closed to 300 million active users and about 500 million of posts, which are also called tweets, are generated by users, per day [1].

Over the time, Twitter has become a fundamental source of information for news. A post generated by a user and its subsequent viral propagation in the network, even has attracted well established news sources[5]. As a one step forward, researchers have tried to analyse if the tweets contain predictive power. This has resulted in various works in studies in different fields, such as the study of the spread of epidemics [2], the prediction of electoral results [3] or of football matches results [4]. In particular in the financial domain, a lot of research has been done to find if there exists a correlation between the tweets and the trend of the stock market. In other words, the researchers tried to investigate if tweets can affect (or predict) the financial market.

In this thesis, we present our study about understanding and findings of relation between tweets and stocks. In twitter with respect to financial tweets, users can embed cashtags, which are stock symbols embedded immediately after the $ sign. Using cashtags, users can search about the tweets related to a particular stock.

2.1 Problem Statement

Let there be a U number of users who are posting on Twitter about a particular cashtag $c_k \in C$, a set representing all the cashtags. The cashtag c_s is about the stock s, and $s \in S$, where S represents the set of all the stocks. Let T represents the set of all the tweets and T_s as set of all the tweets related to cashtag c_s. Let c_{s_p} be the stock price of a particular stock s.

In the past a lot of research[6][7][8] has been done to propose a function $f(T_s)$, which takes as input all the tweets for a particular stock or index s, analyse them and predict the stock or index price of s. In this work, we take an alternative approach: using the stock price and tweet information, we investigate following questions.

1. Is there any relation between the amount of tweets being generated and the stocks being exchanged?

2. Is there any relation between the sentiment of the tweets and stock prices?

3. What is the structure of the graph that describes the relationships between users?

2.2 Contribution

In this thesis, we analyze the vast volume of data and present our quantitative and qualitative correlation results about the relation between stocks and tweets. Our findings show that stocks and tweets volume are correlated with more than the 78% of the days (out of all the days being analyzed), in some cases, our sentiment analysis shows the predictive power on the stock price daily trend. The structure of the graph representing the relationship between users reflects the three intuitively defined categories (news sites and bloggers, investors, false positives).

Alongside, we also analysed the relationship among the various users present in the dataset. With respect to network analysis of the users we discover a long tail pattern among the follower and the followee. That is, there are a few hubs which are being followed by a large number of followers. We also found a small number of disconnected nodes in the network.

The rest of the report is organized as follows. Section 3 describes some terminologies with respect to Twitter. In section 4, we present related literature with respect to our work. We then discuss about the methodology and approach used in 5. Section 5, present our results related to the dataset we

analysed. We conclude with several future directions in section 6. In A we give a tutorial to build the dataset, in B we provide javacode.

Chapter 3

Preface

In this chapter, we give a basic background with respect to Twitter and in particular tweets related with finance.

3.1 Twitter

As with many stories of success, even more in the case of social networking services, Twitter's history is very complex and controversial[9]. Since its launch in 2006[10], initially called "twttr"[11], allows its users to write a short message of 140 characters. Doing this operation of status updating, is widely called "to tweet".

The relations in Twitter are unidirectional that is an edge from user A to B, not necessarily means vice versa. Thus, there is no concept of bidirectional friendship as present in social networks like Facebook. An user A can subscribe to another user B's feed: the operation is termed as "following". Thus, A is called a follower and B is called followee, in Twitter terminologies.

A user can i) repost someone else's tweet on its personal feed, which is termed as "retweeting" and ii) interact with other users, replying to a tweet or tagging another user in a tweet with the operator "@". Another key operation of Twitter, present since the beginning, is the tagging a tweet using keywords preceded by the symbol "#". Keywords of such kind are called hashtags and they're the base of the research on Twitter by topics. One of the key features of Twitter is the trending topics. This feature is geo specific and exploits whom a user is following.

These features, conjuncted to its typical, as said before, directed nature, unlike, for instance Facebook, which is bidirectional, contributed to ascend Twitter to the role of most popular social networking service for news, politicians, celebrities and influent bloggers.

An official online communication channel, in conclusion.

3.2 Twitter for Finance

In July 2012, Twitter introduced the possibility for the users of making a search using the ticker symbol of a stock, preceded by the symbol "$". Such a symbol, or operator, for instance $AAPL, is also called cashtag. This feature was introduced for the first time in 2008 in Stocktwits, which is equivalent of Twitter however, only related to financial domain.

The Twitter users with respect to financial domain can be categorised in the following three categories:

1. **Financial news channels:** These users drive information linking in their tweets articles from their site.
2. **Trading bloggers:** These sets of users express their opinion on the stocks.
3. **Investors:** They are users that simply follow others users for advice.

Intuitively, users of third categories are generally followers of first and second types. One of the aims of this thesis is to analyse the network structure and to cross validate our intuition.

Chapter 4

Related Works

In this section, we present various related works with respect to our work in section 4.3. As mentioned, our work is mainly related to analysis of financial tweets and correlating the tweets with the fluctuations in the stock market. However, this work is very much related to prediction of stock market using sentiment analysis. Thus, we present works done in the past, where authors have proposed various approaches for prediction of various entities in general in section 4.1 and then in particular with respect to financial domain in section 4.2.

4.1 Prediction

Researchers have always been fascinated by large amount of data on the web and exploiting and analysing it to predict various entities such as for football games [4], consumer behavior [12], music sales [13], elections [3], epidemics [2], [14], movies sales [15],[16].

Here we discuss the works with respect to Twitter data. In [4], authors proposed a predictive model tested using cross validation, to predict football outcomes. They used both Twitter and historical data for analysis and reported a 75% accuracy of the outcomes. To predict the movie box office results, a model based on the tweets creation rate is proposed in [16]. The approach uses sentiment analysis and claim to outperform market-based predictions. In [3], researchers try to predict elections using Twitter datasets by performing a sentiment analysis of the tweets. The method exploits the mentioning of party/politician and political coalitions. Apart from predictions, Twitter data is used for tracking an epidemic. In [2], authors described a tracking system for tracking the prevalence of Influenza-like Illness (ILI) in several regions of the United Kingdom.

4.2 Prediction in Financial data

In this section, we present literatures, which are related to prediction of stock market using Twitter data analysis.

Prediction of the stock market using web data such as blogs [17] and other web social platforms specially Twitter [18], [8], [19], [20], [7] has attracted a lot of researchers in the past. In [17], using SVM based technique, authors predicts the stock market with 87% accuracy. However, the data on blogs is small compared to Twitter. This motivated later studies to use Twitter data extensively for stock market prediction.

In the initial set of studies in Twitter dataset, which is done over a six month period, in [7], authors identify the tweets into six sets of moods and then use it for prediction of change in values in the Dow Jones Industrial Average (DJIA). Later in a different work, in [19], researchers argued that it is important to find experts in the dataset to predict the values. However, in this approach a lot of users who are not followers and make their opinions lost their voice. By using sentiment analysis, researchers in [21], proposed a model to predict the stock market. Their aim was to investigate if there is any correlation between Twitter and the stock market by studying sentiment, message volume, price movement and stock volume as well as the effect that a Twitter user's reputation may have on sentiment and the stock market. In the later part of the researches, researchers exploit sentiment analysis clubbing with machine learning techniques in [8], [20] and [6] on Twitter corpus. They all find out a strong correlation among tweets and stock markets.

4.3 Analysis of financial data

In this section, we present works which have also analysed the correlation between financial markets and the Twitter activity. In [22], the authors study the correlation between tweets and stock market events such as changes in price and traded volume of stocks. There are many differences between their work and this thesis.

1. First, in the Twitter API, they match in a less filtering way, getting tweets, for instance for Yahoo, that matches the regular expression "#YHOO | $YHOO | #Yahoo", while we get only tweets with the cash-tag, restricting results to financial field and preventing false positive.

2. Furthermore, their research is focused on the interaction graph that they build based on the tweets, while we investigate also on the relationship between users and on the relationship between stocks.

3. Another difference is that we also performed the sentiment analysis of the tweets and its effect on the stock market, which is missing in [22].

Another work, where researchers analysed the set of tweets to understand the mood of tweets and to relate it to the stock price is done in [23]. However, the main emphasis of their work is about sentiment analysis and its effect on the stock market in general. Specifically, they evaluated the tweets from August 1, 2008 to December 20, 2008 and considering 18 events. In our case the time line is much longer. However, the main difference is that in their investigation they investigated if socio-economic phenomena (such as peaks in stock market or oil price, Presidential Election, Thanksgiving day) affect the public mood, obtained from Twitter, while we're looking in the opposite direction: if Twitter mood is useful to predict economic phenomena.

In another similar work [24], researchers analyzed tweets to find correlations among tweets and stock market in various sectors. However, their dataset is less than of three months, whereas our dataset is more than that of one year. Also, as mentioned before, we are targeting the tweets containing cashtags. Also, compared to their work, where they proposed a model, we have also performed sentiment analysis to find relation between mood of the tweets with the stock fluctuation.

Chapter 5

Evaluation

In this chapter, we first describe the dataset we used for our analysis. Later, we describe various quantitative and qualitative results of our investigation on the dataset.

5.1 Dataset

5.1.1 Twitter

The dataset contains tweets with respect to 1882 companies starting from March 2011 to June 2013 and if of size 1.31 GB. It contains approximately six million tweets which were tweeted by approximately 0.5 million unique users. Please refer to Table 5.1 for summarization of the Twitter dataset being analysed.

5.1.2 Yahoo Finance

For particular months, we analyzed tweets with respect to six companies, we also downloaded corresponding stock values from the historical archive of Yahoo Finance[25]. Yahoo Finance stores historical financial data for all the companies quoted at the New York Stock Exchange (NYSE) and National Association of Securities Dealers Automated Quotation (NASDAQ). The site provides an API and a web interface to download the data from its database in a .csv format. The fields that we got, for every day, are Date, Open, High, Low, Close, Volume and Adjusted Close, that means the close value adjusted considering the eventual dividends and splits.

Stocks Number	1.882
Size	946 MB
Tweets Number	5.927.164
Users Number	516.371
Date	2011-03-11 to 2013-06-11

Table 5.1: summary table

5.2 Results

We provide results related to top six most tweeted stocks, which ranges from technology companies (Apple, Microsoft) to banking (Bank of America, Goldman Sachs) to e-commerce (Amazon) to entertainment portals (Netflix). We found tweets related to Google and Walgreen in the top six highest tweeted however, we discarded these tweets. In case of Google, we were not able to download the financial data for the time slice we are interested. In case of Walgreen, there are high number of false positives, due to its ticker $WAG, that coincides to a popular internet slang.

5.2.1 Quantitative analysis

One of the ideas of this thesis is to compare the daily volume of stocks traded with the volume of the tweets that has been tweeted in the same days, to see if there is some kind of correlation between the two values. We obtain information about the stock price from Yahoo Finance and the volume of the total stocks traded per day was compared to per day tweet volume for each particular stock.

For all the six companies, we perform our analysis for the months of September 2012 and May 2013. We picked September 2012 as there was an event related to Apple, that is the launch of iPhone 5. We then pick the same month for the rest of the stocks. Picking May 2013, was just a random choice. For Apple stocks we conducted the experiment also for one year and and half, from the first January 2012 to May 2013.

Macro Analysis

We first present some statistics related to the correlation between the trend of the stocks and the trend of the tweets. So, as a first step, we define the **daily average correlation**, that is, for every month, in percentage, the sum of the number of days in which the two values either increases or decreases

together, divided by the total number of days of the month. We then define the **distance** between the two trends as the absolute value of the difference between the normalized values (meaning divided by the maximum value in the month we are considering) of stocks and tweets volume.

We present our findings with respect to these two measures in Table 5.2.1. We only report the maximum and minimum distance on the correlated days and the average distance in those days. It's important to note that the daily average correlation never reaches a value below 65%, while there are some months (september 2012 and may 2013 for Netflix) where it's over 90%. Looking at the distance in correlated days, we observe that the minimum values are very low: in the sample considered of twelve months is always under the 4%. Instead the maximum values of the distance in some months reaches considerably high values, in fact it's always bigger than the 26% and in some cases reaches the 60%.

Stocks	Date	DailyAvgCor	MinDstCorDay	MaxDstCorDay	AvgDstCorDays
AAPL	Sep 2012	72.41%	1.32%	47.84%	18.60%
AAPL	May 2013	80.65%	0.38%	31%	12.32%
AMZN	Sep 2012	72.41%	1.88%	59.85%	17.40%
AMZN	May 2013	70%	1%	38.67%	15.13%
BAC	Sep 2012	79.31%	0.15%	64.01%	14.39%
BAC	May 2013	83.33%	1.73%	52.89%	15.47%
GS	Sep 2012	75.86%	2.62%	47.92%	16.75%
GS	May 2013	76.67%	3.61%	33.97%	12.97%
MSFT	Sep 2012	65.52%	1.37%	26.62%	11.25%
MSFT	May 2013	80%	3.33%	61.06%	29.60%
NFLX	Sep 2012	93.10%	2.02%	26.07	10.22%
NFLX	May 2013	90%	3.81%	54.12%	15.76%

Table 5.2: Statistical results table

We also present the graph of the Apple stocks and the tweets for the extended period of one year and an half (Figure 5.1.) The graph shows that tweets and stocks are mostly correlated. To understand the relation between tweets and stock, we performed micro analysis. Anyway, for that long period of time, the daily average correlation resulted of 82.79%, while the average distance on correlated days is 7.32%.

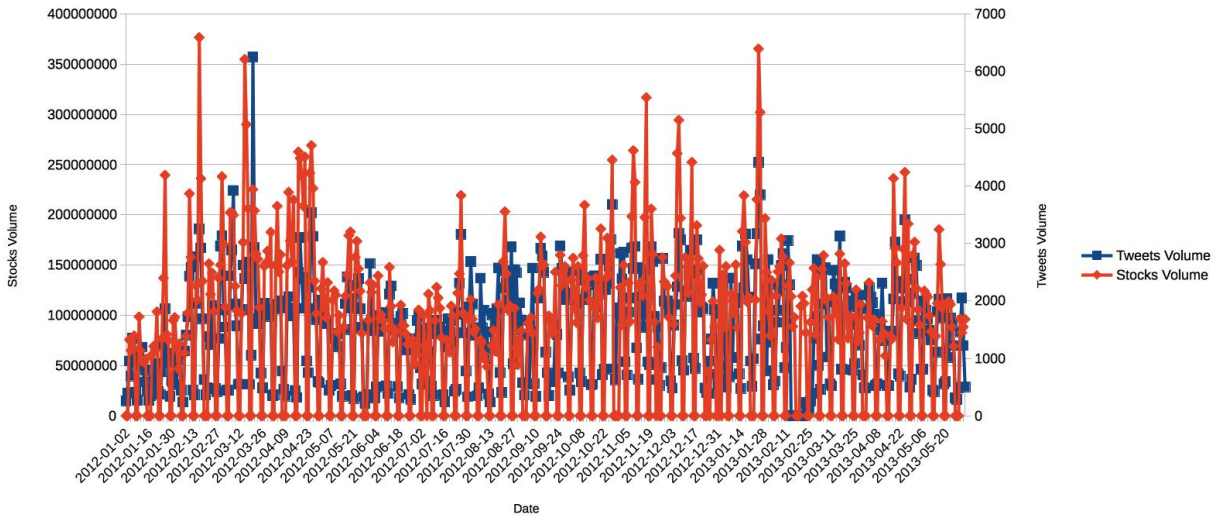


Figure 5.1: Apple extended quantitative analysis

Micro Analysis

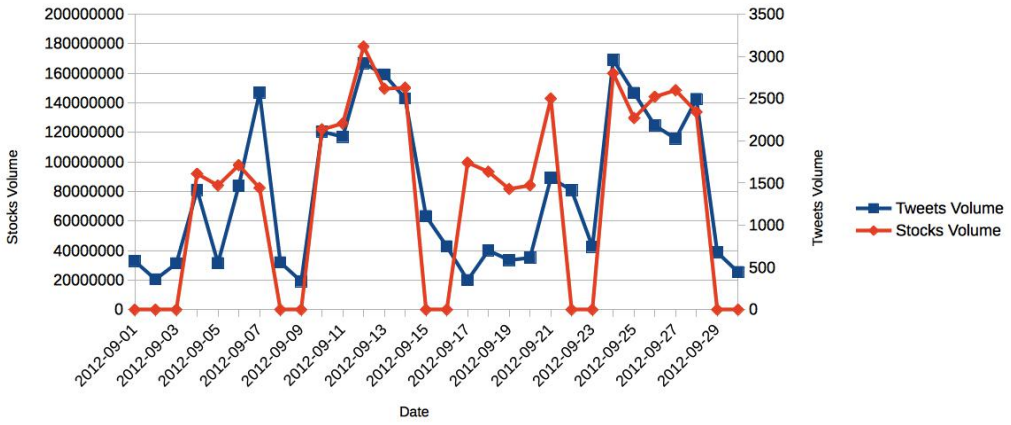
As the macro analysis performed in previous section, is not very clear. Thus, we now present the trend between the stock volume traded and tweet volume for September 2012 and May 2013.

1. **Apple**

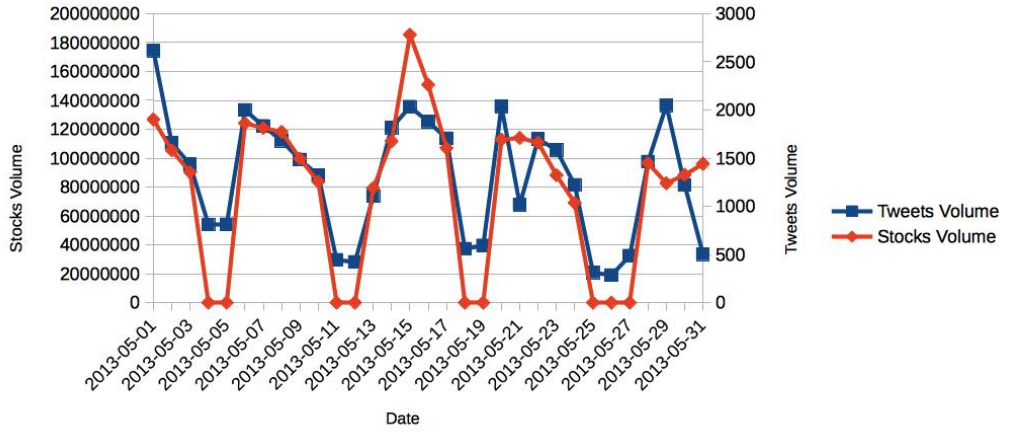
   We selected September 2012 month as it contains an important Apple event, that is the launch of iPhone 5. The two highest values of both the trends are on 12th and on 24th, of September 2012 corresponding to the two important events happened in these two days that is on 12th, Apple launched iPhone 5 and on 24th severe riot at the Foxconn plant[1], in China. In general the two months show a high correlation among the two entities, around 72%, but lower by some points to the average.

   In May 2013, the trend of the curves is very similar, this is evident in the statistical results: in this month there's an high level of correlation and a very low level of average distance on correlated days. The maximum of the two curves are not in the same day, but the one of the stocks volume corresponds to one of the highest values for the tweets volume curve and vice versa. The minimum is not not be considered, since

[1]The Foxconn company makes iPhone for Apple.



(a) Sept. 2012



(b) May 2013

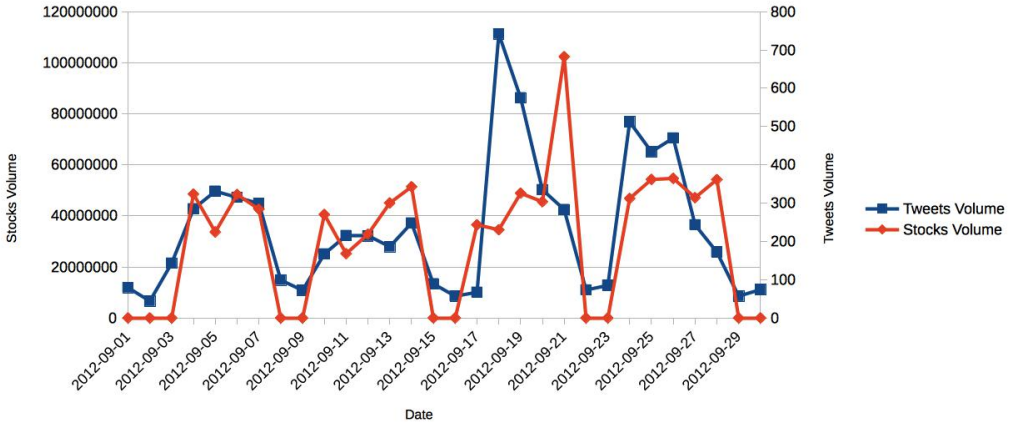
Figure 5.2: Apple quantitative analysis

in the weekends and holidays there aren't financial values (volume set to zero). However is evident that in weekend and holidays, the tweets volume reaches the lowest values.

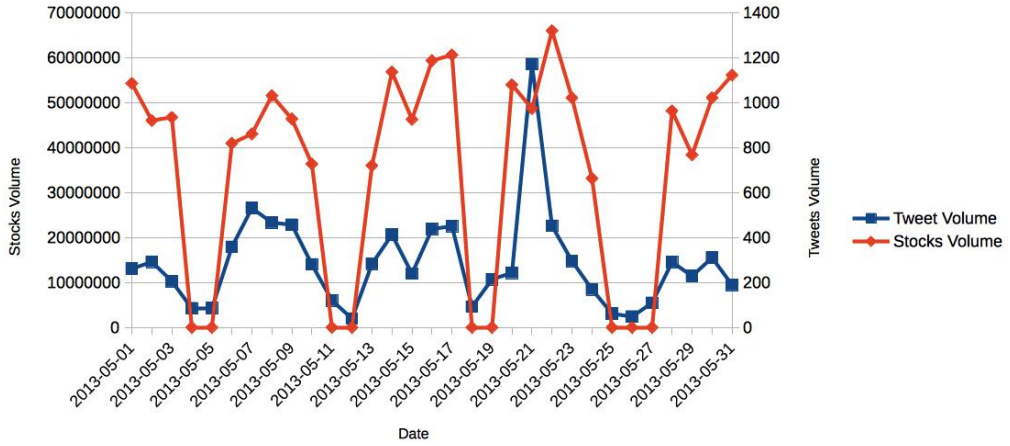
2. **Microsoft**

   In September the the two curves are quite similar, even if there are many days in which they go in different directions (one increases while the other decreases and vice versa), the distance between them is low. There's no correspondence between the highest peaks of the two curves.

   In May there is good correlation between the curve of the stocks and



(a) Sept. 2012



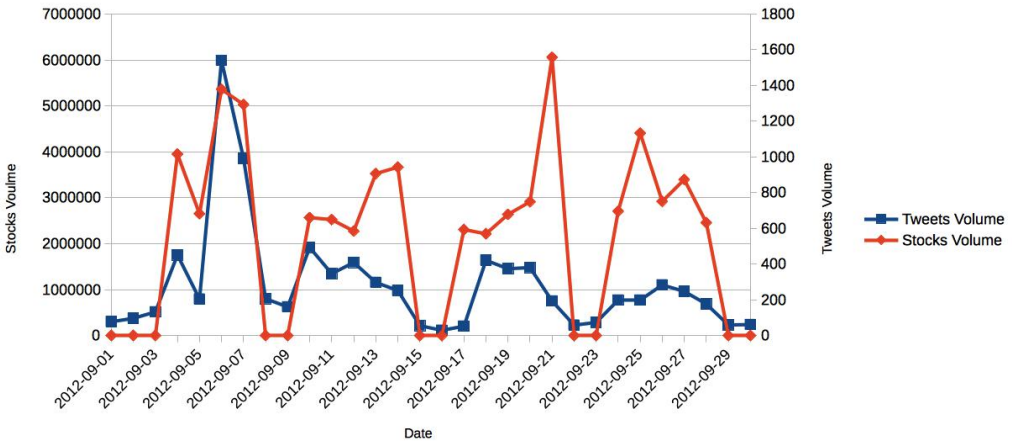
(b) May 2013

Figure 5.3: Microsoft quantitative analysis

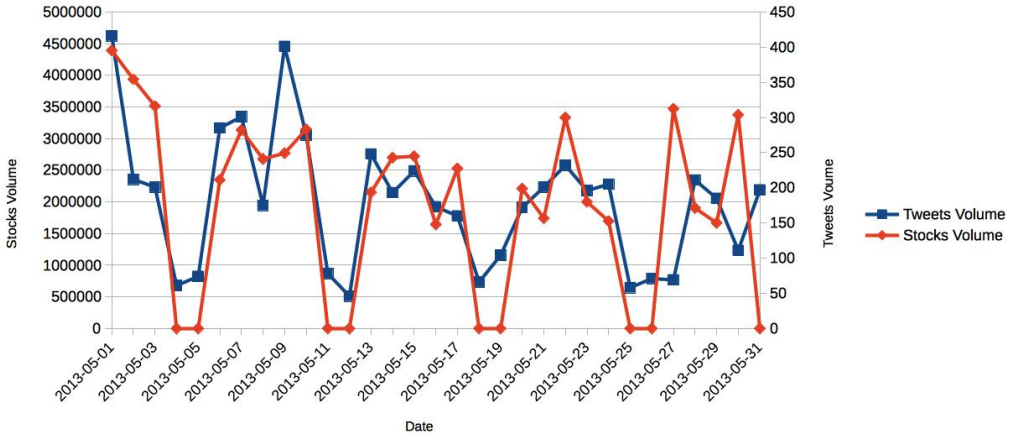
the one about the tweets, but the distance between them is high. The day with the highest volume of traded stocks is next to the one with the more elevated number of tweets.

3. **Amazon**

   For the month of September considered there is a low distance between the curves in the first half of the month, that becomes high in the second half, with an aggregated result over the mean. The same it is true also for the correlation, in particular we can see that the two curves are perfectly overlapped in the first week, with the second highest value



(a) Sept. 2012



(b) May 2013

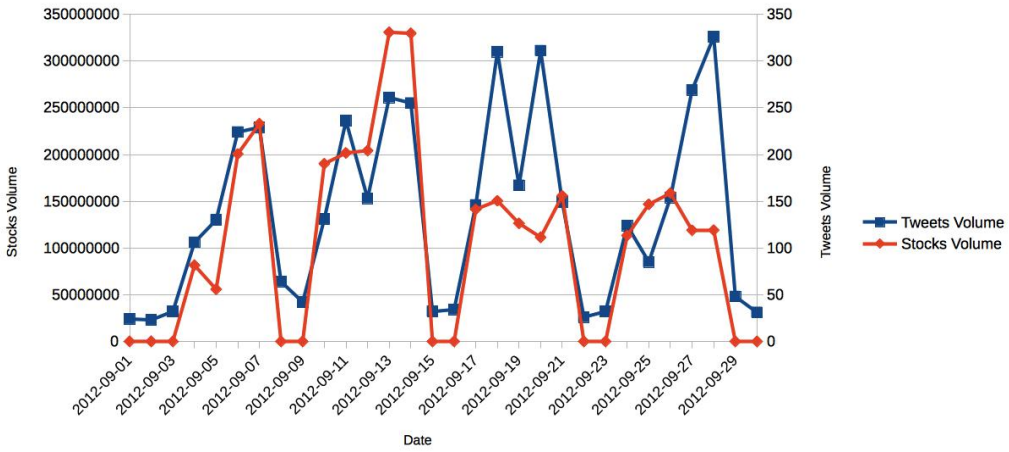
Figure 5.4: Amazon quantitative analysis

for the stocks volume corresponding to the highest value for the tweets volume.

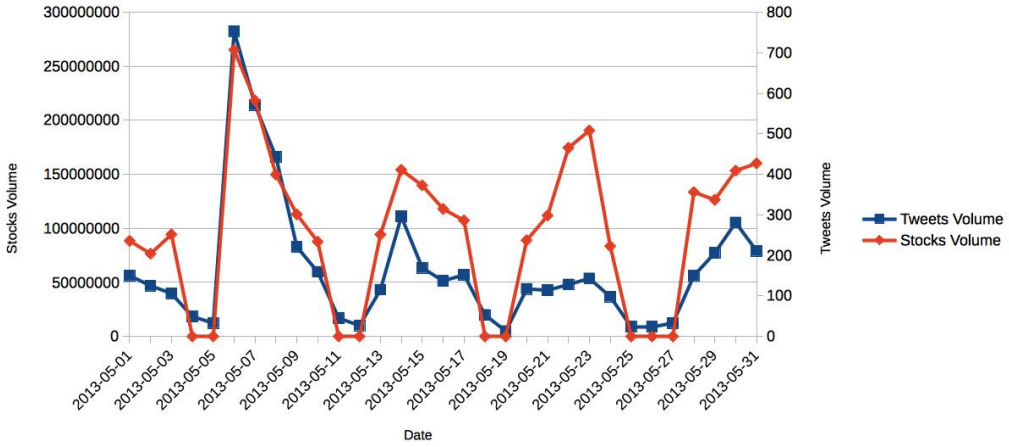
In May, even if the correlation between the two curves is not particulary high, the distance between them is very low, resulting to be lower than the mean. In the first day of the month there is correspondence between the highest value of the two trends.

4. **Bank of America**

Like for September for AMZN, in September we have got the first half



(a) Sept. 2012



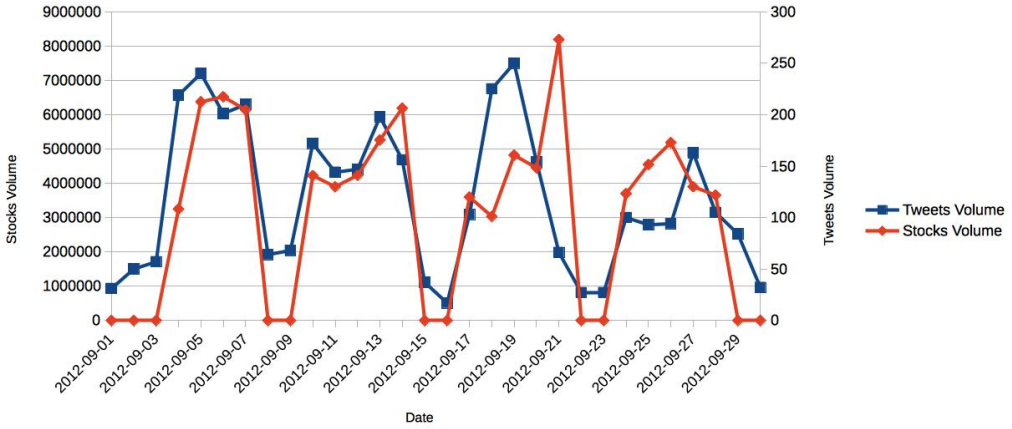
(b) May 2013

Figure 5.5: Bank of America quantitative analysis

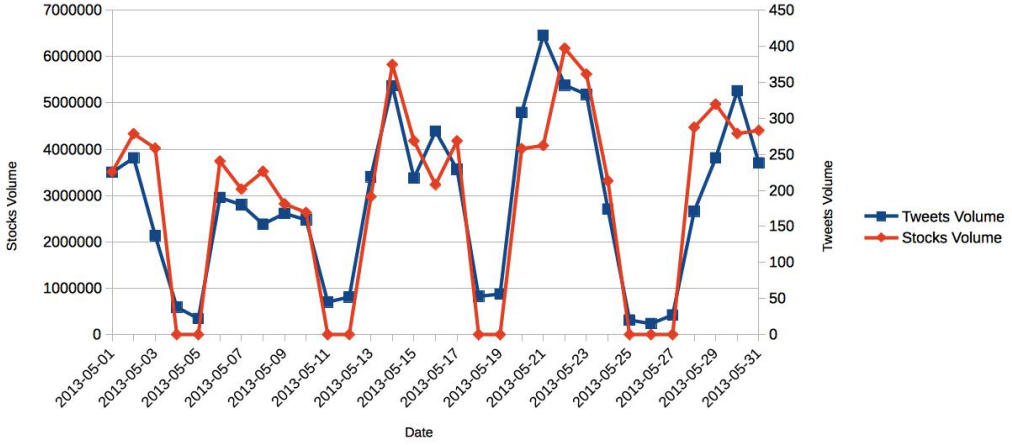
of the month with the values of the curves very close and then the second half of the month with the distance increasing. In particular in this case we can see how, in spite of the usual, the tweets volume curves overcomes the one of the stocks volume.

For the month of May the graph highlights a very good correlation among the two trends, where the first ten days are perfectly overlapped. This graphical propriety is confirmed by the statistical results: for this period we have got one of the highest values for the daily average correlation.

5. **Goldman Sachs**



(a) Sept. 2012



(b) May 2013

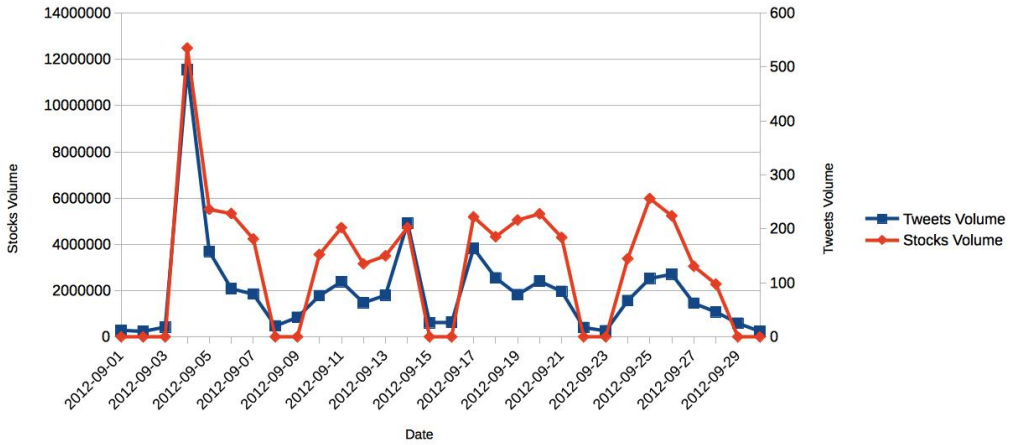
Figure 5.6: Goldman Sachs quantitative analysis

For the month of September we can see that the two trends have similar trajectories in the first half of the period of time considered and then they differ in the last part. There is one day in the middle of the two highest values, separating them.

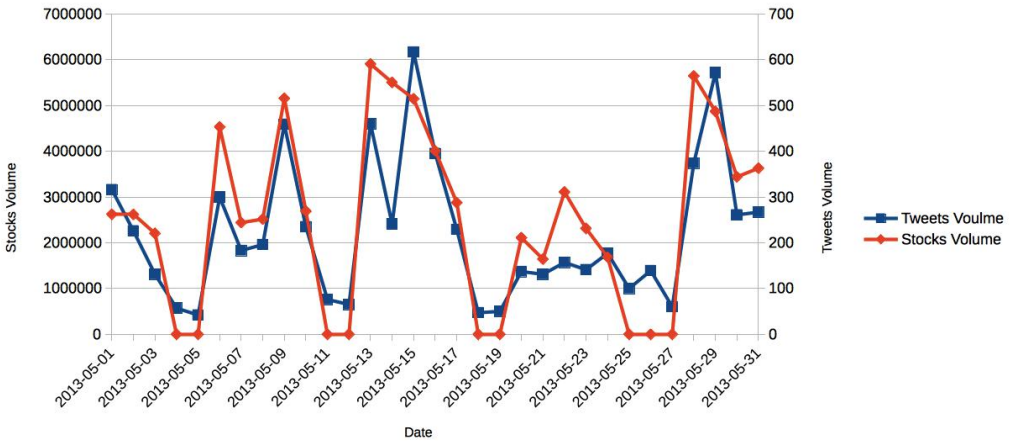
In May the two lines are particulary similar, their distance is very very low (in fact it corresponds to the second lowest value of average distance on correlated days), even if the daily correlation isn't particulary high: that's because, even if the curves doesn't take the same directions, they

stay close each other. The global highest peaks are not in the same day, but the second highest do.

6. **Netflix**



(a) Sept. 2012



(b) May 2013

Figure 5.7: Netflix quantitative analysis

In the case of September the highest values are corresponding for the two trends and the value of the correlation is particulary high, in fact it is the highest of all the month we have considered. From this follow that also the distance is the lowest value of all the months.

Also in May these curves shows intuitively good correlation, with two

cases (on the 13th and on the 28th) in which there's correspondence between some of the highest values for the tweets volume and the two highest values in the stocks volume.

5.2.2 Qualitative analysis

Naive attempt of sentiment analysis, to understand not just if someone express an opinion about something, but what opinion that user has expressed and if this opinion has some kind of effects on the price of the stocks.

That is made using the vocabulary AFINN-111 [26] of weighted words, containing 2477 words and phrases. Each word has been manually labeled with a value from minus five to plus five.

So, for each tweet, we parse all the words and, if a word is contained in the vocabulary, we sum its value to the partial sum of the previous words of that tweet.

We call the result mood of a tweet; if the mood is positive, we say that the tweet expresses a positive mood, if it is negative, we say that the tweets expresses a negative mood, otherwise the tweet has a neutral mood. In the following graphs we plot one above the other, a graph that shows for each day the percentage of tweets with moods of each kind for the top 6 stocks and a graph containing the trend of the price growth of the same stock, on a daily basis.

The relative java code can be found in B.2.

1. **Apple**

   In this case (September) correlation is not very evident. No perfect corresponding between highest and lowest value of growth and events happening in that days. Day with the highest percentage of positive mood tweets is next to the launch of the iPhone 5.

   In May there is clearly a connection between the two graphs: the days with the highest growth (5th and 19th) are next to days with highest percentage of positive tweets. In this case the reason the positive tweets were the day before, is probably because they were in the weekends and we noted that, with a lower number of tweets, the proportions between positive, negative and neutral moods often changes. Neutral moods are prevalent because many words are not contained in the dictionary. Highest decrease (15th) correspond to highest percentage of negative moods tweets.

2. **Microsoft**

In the case of September there are five days in which positive mood overtakes the neutral one, but they're not telling us very much about the future prices of the stocks: on the 3th, it corresponds to a decrease of the price, on the 8th was useless because the next day markets were closed, on the 9th it corresponded to the third highest growth in price, on 24th corresponded to one of the highest decrease and on the 26th was before a day with closed market. Peaks in negative mood weren't very significative.

Even in the case of May we cannot claim a particulary high predictive power of our sentiment analysis. The first day in which positive mood is prevalent is previous a degrowth, instead the second one anticipates a growth of the price. Days in which negative mood overtake the positive are not significant since they are in the weekends.

3. **Amazon**

   In September, the positive mood overtakes the neutral, or goes close to its percentage, only in the weekends, except for the 13th, that anticipates the highest peak of growth. Talking about the negative moods, they predict a decrease, overtaking in percentage the positive one, on the 9th and on the 24th.

   For the month of May the predictions are not good at all. There is not any day in which the mood predicts the right direction of the growth of the stock.

4. **Bank of America**

   In the month of September the second highest percentage of positive mood predicts a peak in the growth of the stock, on the 7th, and the 4th highest percentage of positive mood predicts another peak on the 11th. On the 18th the negative mood overlaps the positive and the neutral ones, but the following day the stocks prices grows up.

   In May there are not particulary good predictions: the only two days in which the positive mood overlaps the neutral one, have as next days two days of closed markets (26th and 27th). The peak of growth on the 6th is not anticipated by particular high values of positive mood in the previous day.

5. **Goldman Sachs**

   The month of September shows a typical pattern that we have encountered many times until now: the only days in which the percentage of

neutral mood are lower than the ones of positive or negative mood, are the weekends, not interesting for us since there's no financial data.

For the month of May we can see that the values are quite stable: neutral mood around 60%, positive around 30% and negative below 20%. The biggest peak in the growth of price (on the 14th) is anticipated, in the day before, by the highest value in the positive moods.

6. **Netflix**

   In September seems to be some prediction: on the 9th the negative mood distances the positive and in the next day the price decreases, same thing on the 23th (with higher percentage of decrease), on the 13th the highest peak of positive mood (excluding the weekends) is before a peak of growth of the prices.

   In May, like for the same month for the GS stocks, the values of the moods stays stable at the percentage of 60-30-10, respectively for neutral, positive and negative moods, except for the end of the month. But in that days it doesn't seem to be any correlation with the growth of the price.

5.3 Relation between users

In our dataset, there is no explicit information about the relationship among users. In Twitter's terms no information can be extracted from dataset about who follows whom. Thus, to understand relation among users of the Twitter dataset, followed a simple strategy. A user A is in a relationship with an user B, if A mentioned, replied to or retweeted B. We parsed all the tweets to search for @username, where username is the id of a Twitter user. Thus, by this definition, the relationship between two users is not symmetric.

5.3.1 Users distribution

We calculate an estimate of the number of followers of a user by the number of users that are in a relationship with it. The graph 5.14 shows the scatterplot, in logaritmic scale, among the number of followers of a user and the occurrences of users with that number of followers. The resultant graph has a power-law behavior. That is, there are a few users, who have a big following and a large number of users, so, in terms of our graph, many occurrences, who have very few followers.

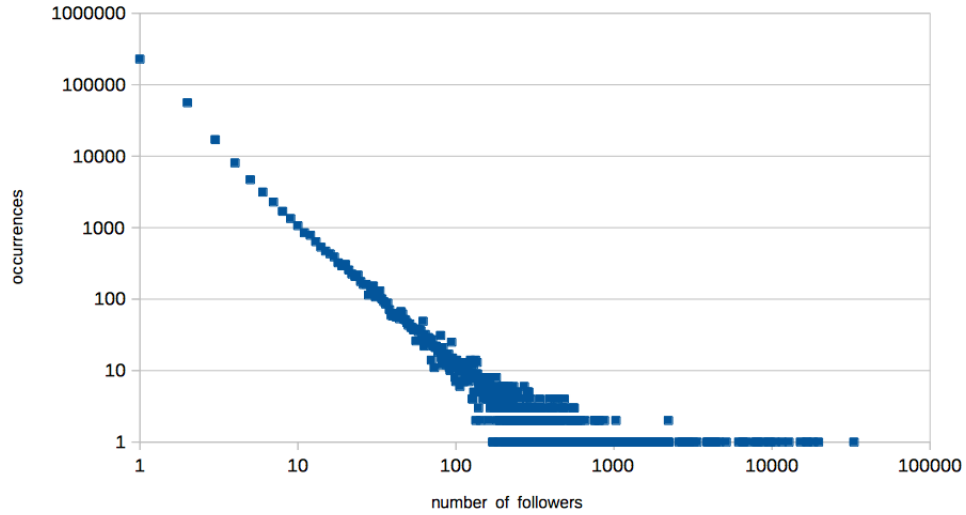


Figure 5.14: Power law distribution among Number of Followers and Occurrences

5.3.2 Network Structure

We have investigated the structure of the underlying network of the users, using an open source software platform called Cytoscape [27], initially designed to visualize molecular interaction network, but that can be used to visualize any kind of network graphs made by edges and nodes, including social networks. The input file has to be a .csv, contaning as header "Source, Target" followed in the next lines by couples of usernames, meaning that the first user is in a relationship (defined as in the previous section) with the second user.

For computational and graphical reasons the following figures are made using a subset of all the relationships among users, containing the 1% of them. The sampling has been made in R using the "sample" function, who, as the name suggests, provide us an uniformly random sample of our relationships.

Figure 5.15: Network Structure Overview

In the overview figure 5.15 we can see a long tail and a dot at the top. The dot contains a concentrated of edges and nodes: respectively, half of the nodes and two third of the edges. This means that tweets in the tail are false positives or users not particulary binded with other users in financial field, instead of the ones in the dot, that are strictly connected each other.

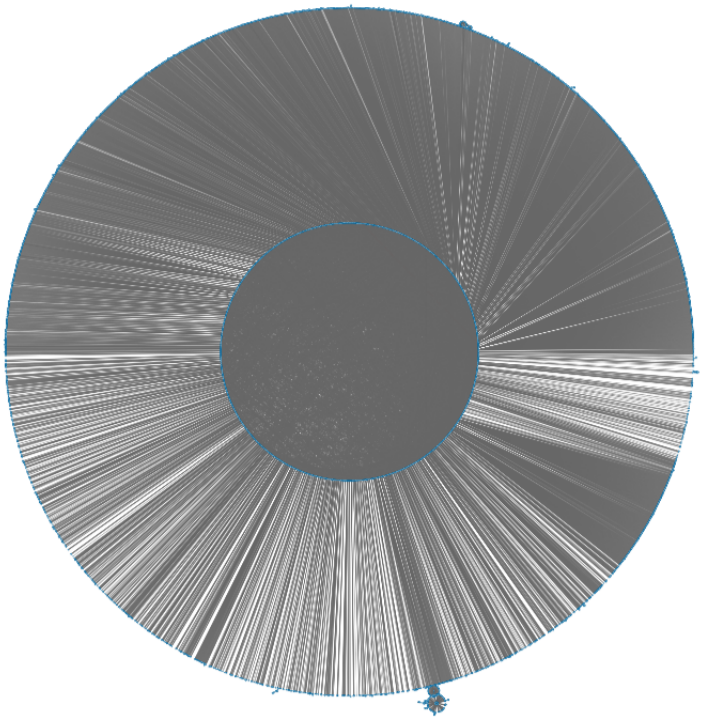


Figure 5.16: Network Structure Zoom in the Dot

Zooming (figure 5.16) we can see that the dot is in effect an annulus with two kind of users: some in the inner circle and others in the outer circle.

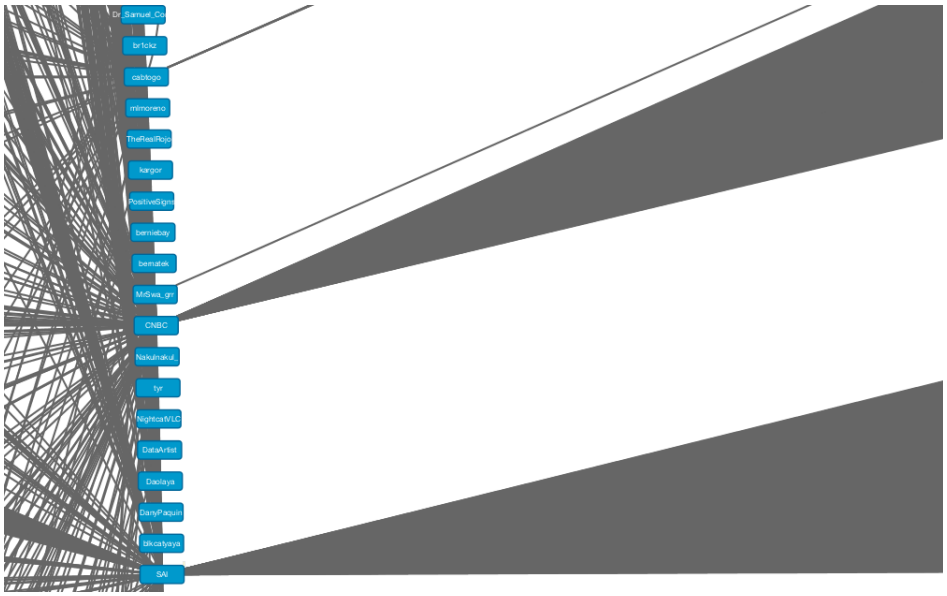


Figure 5.17: Network Structure Zoom Inner Circle

Continuing to zoom (figure 5.17) we can see that users in the inner circle are big players, like important financial news sites or influent bloggers, that follows only each other (in this case we can see SAI, twitter accont of Business Insider:tech and CNBC) The ones in the outer circle are just followers, not followed by (almost) anyone.

5.4 Relation between stocks

In the graph 5.18 two stocks are connected if exists an user that tweeted about both. Thicker is the line between a couple of stocks, more are the users who tweeted about both. False positives has been removed using the same sample simple filter that can be found in B.1.

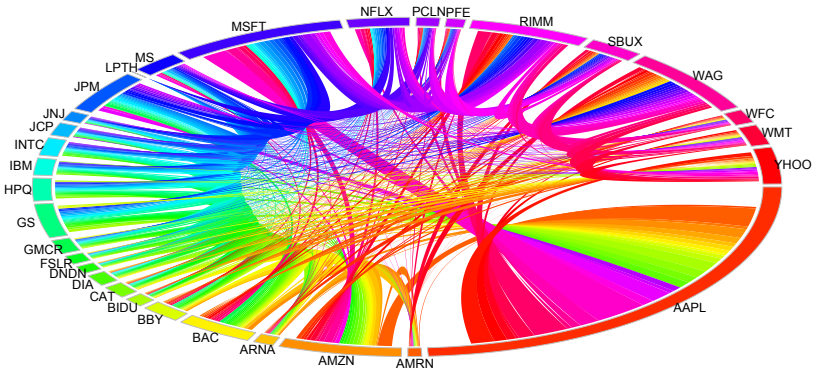
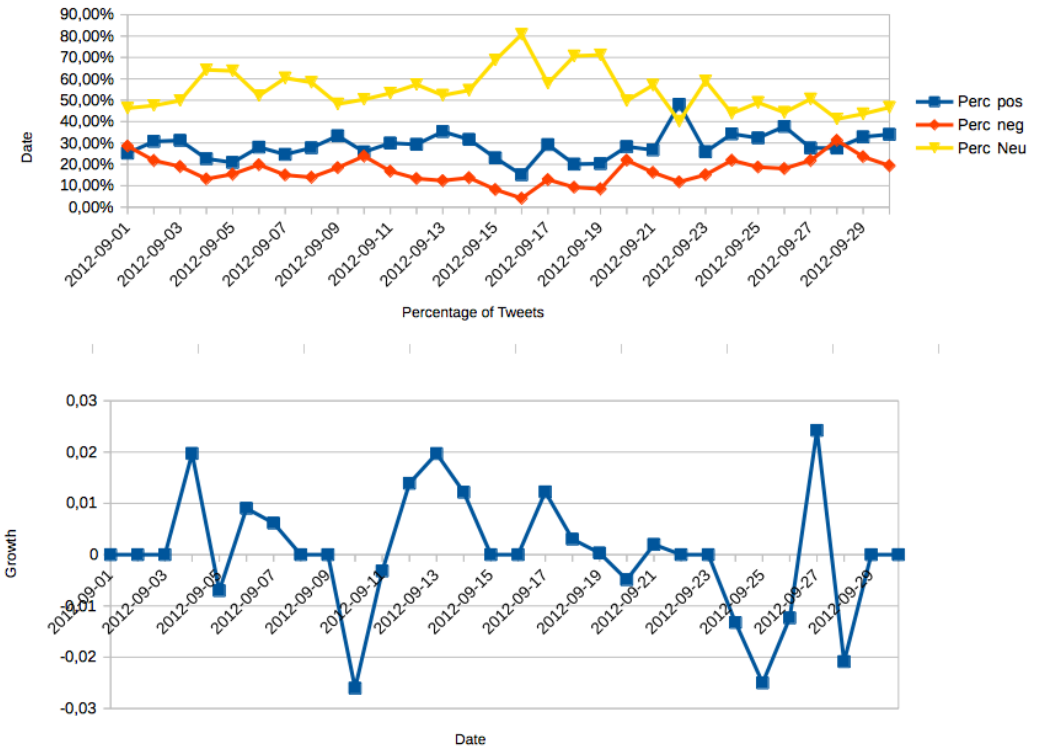
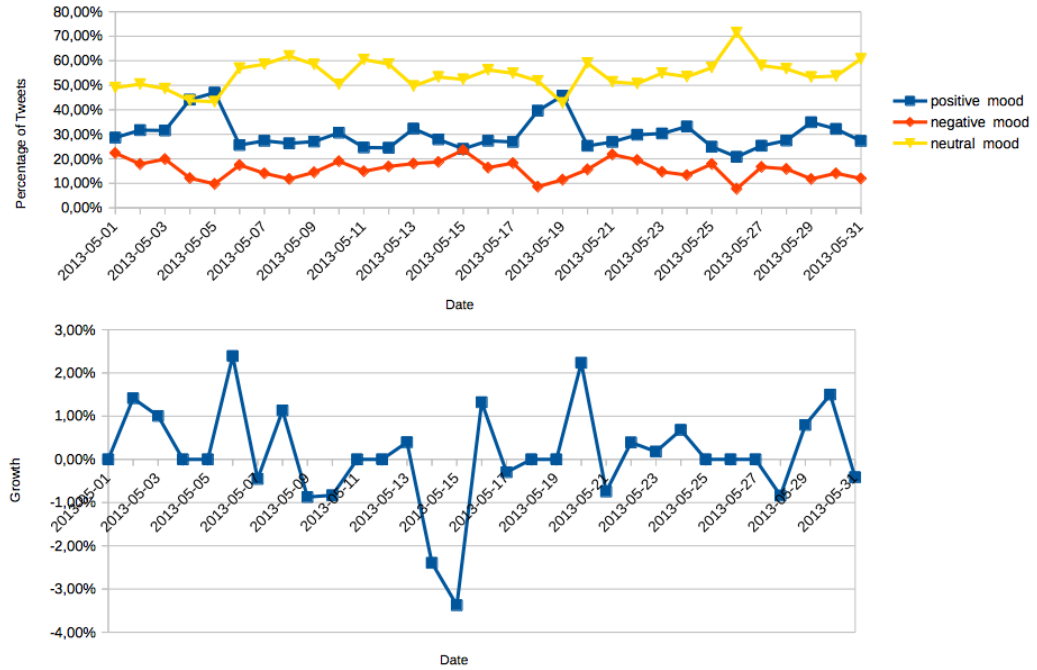


Figure 5.18: Kaleidoscopic Stocks Relation Graph

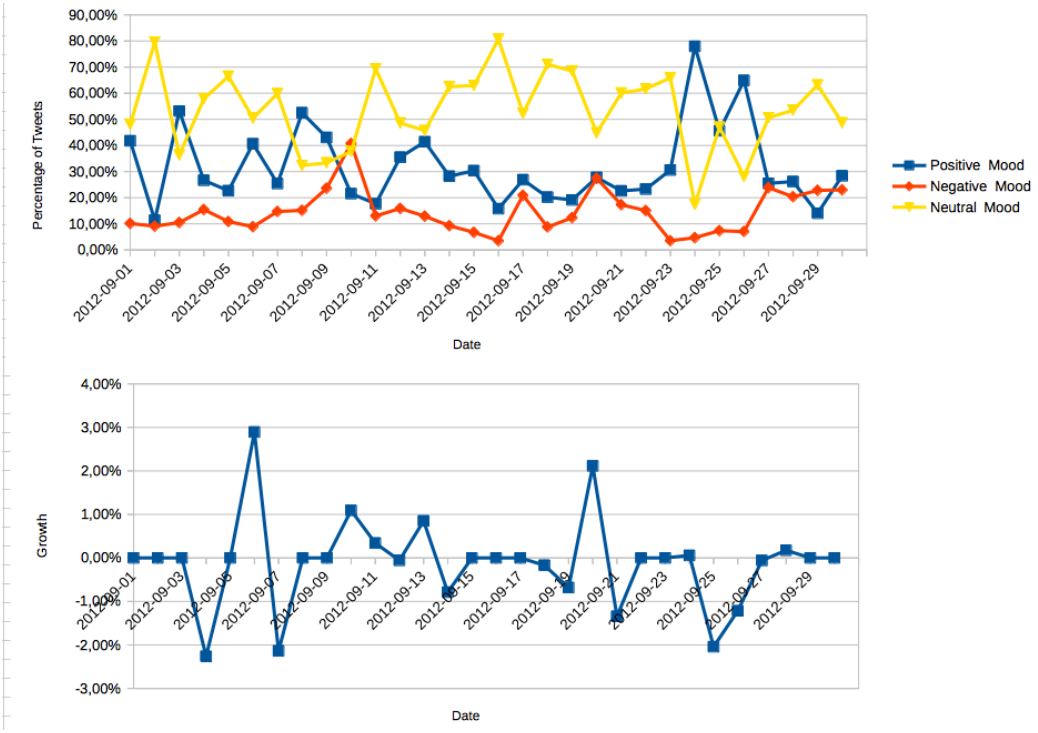


(a) Sept. 2012

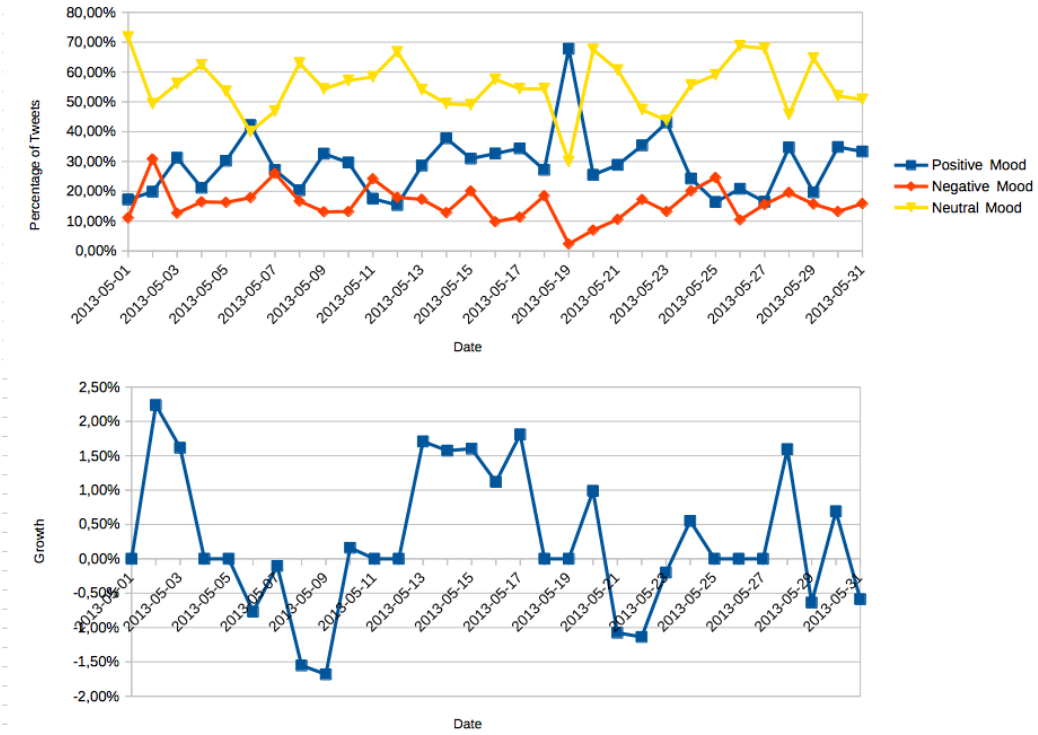


(b) May 2013

Figure 5.8: Apple qualitative analysis

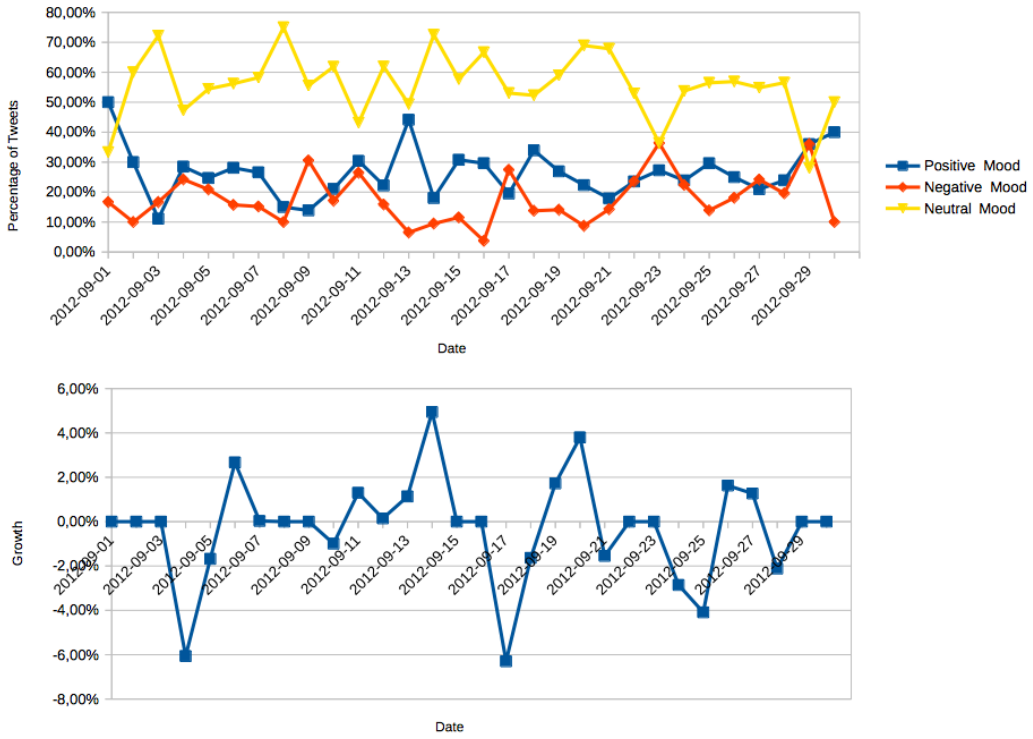


(a) Sept. 2012

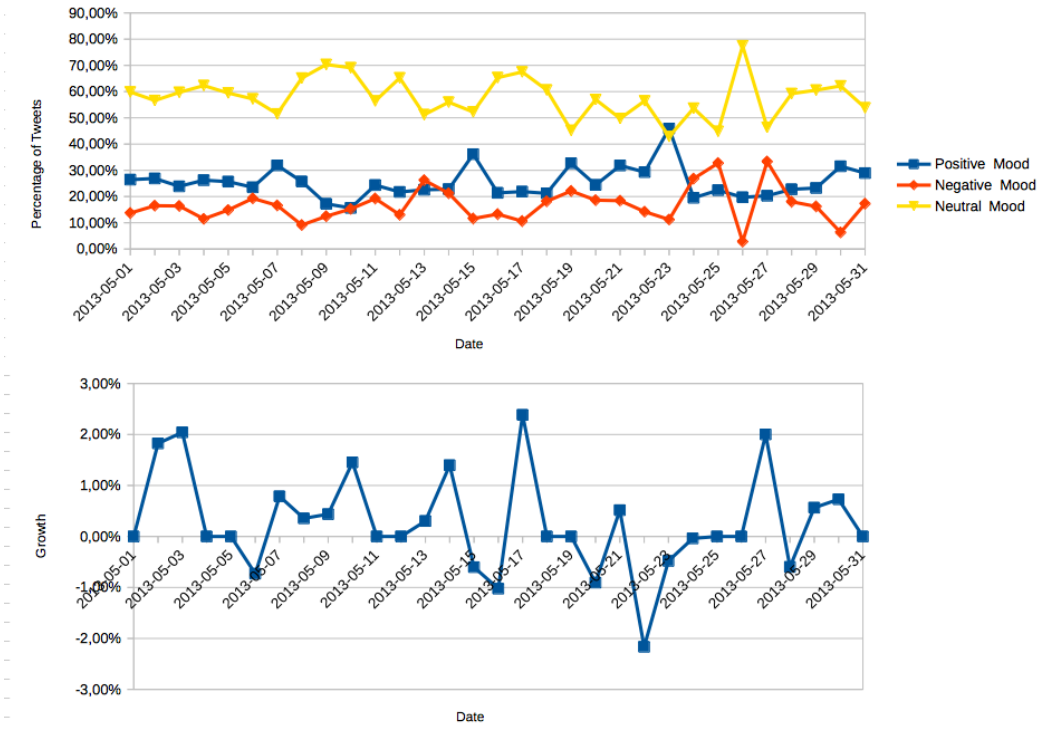


(b) May 2013

Figure 5.9: Microsoft qualitative analysis

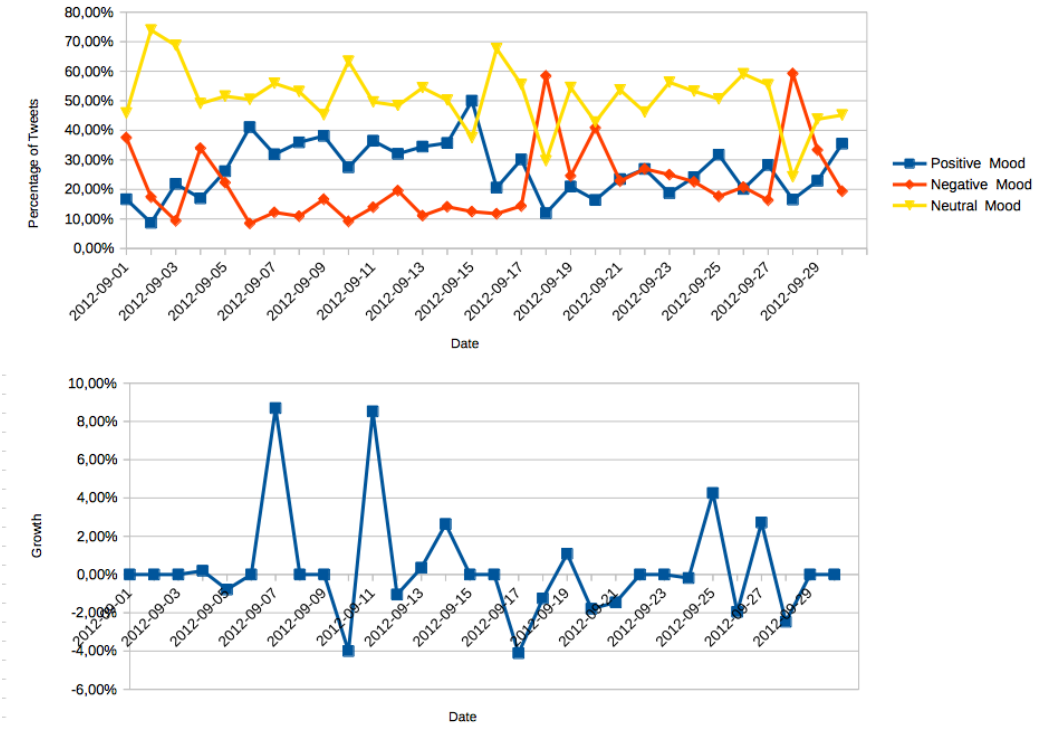


(a) Sept. 2012

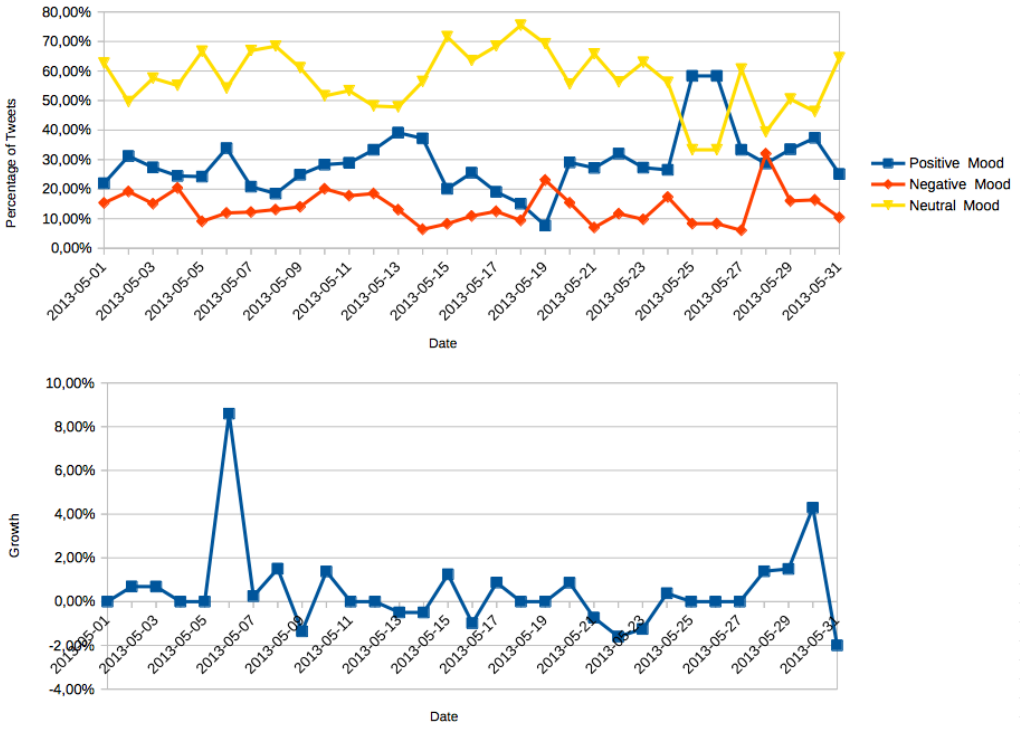


(b) May 2013

Figure 5.10: Amazon qualitative analysis

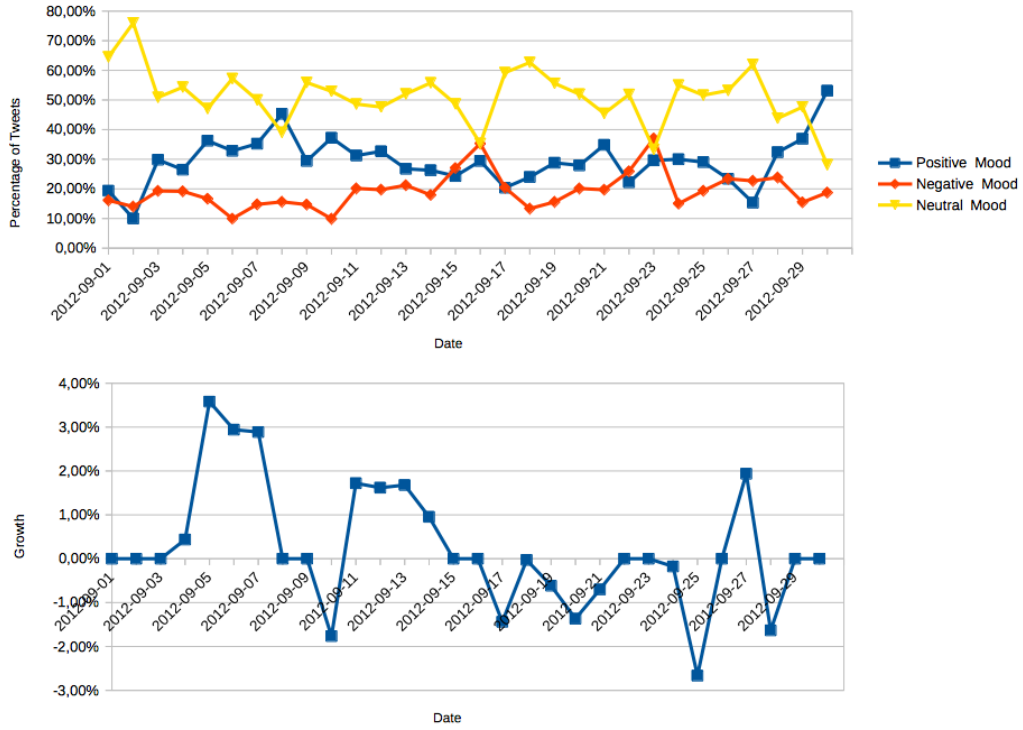


(a) Sept. 2012

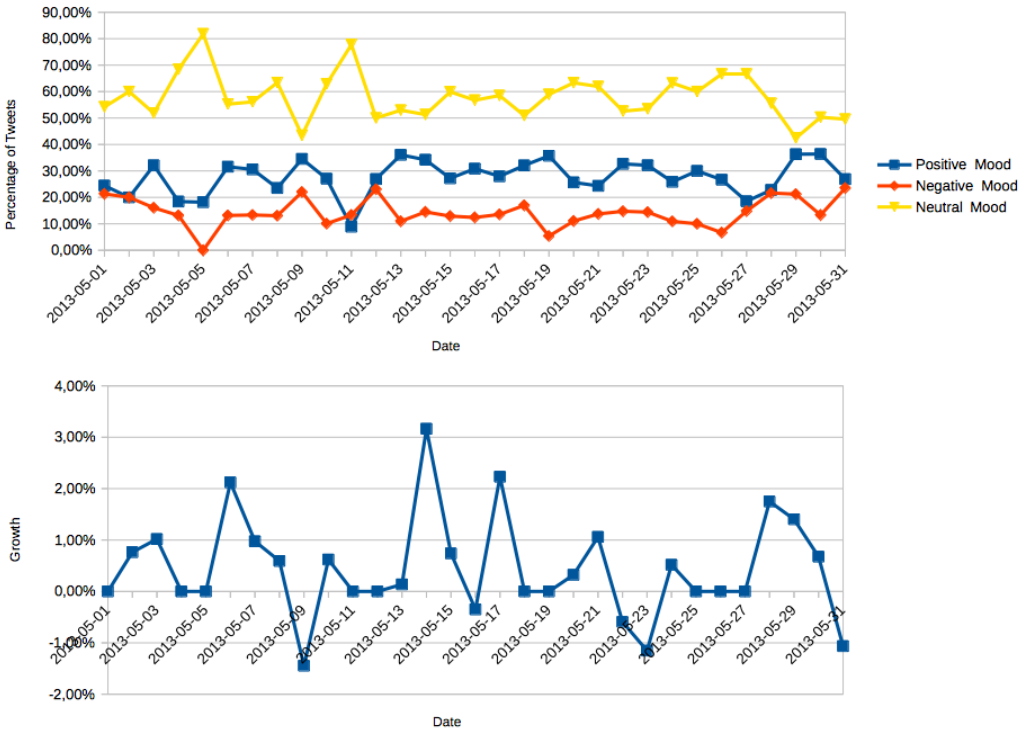


(b) May 2013

Figure 5.11: Bank of America qualitative analysis

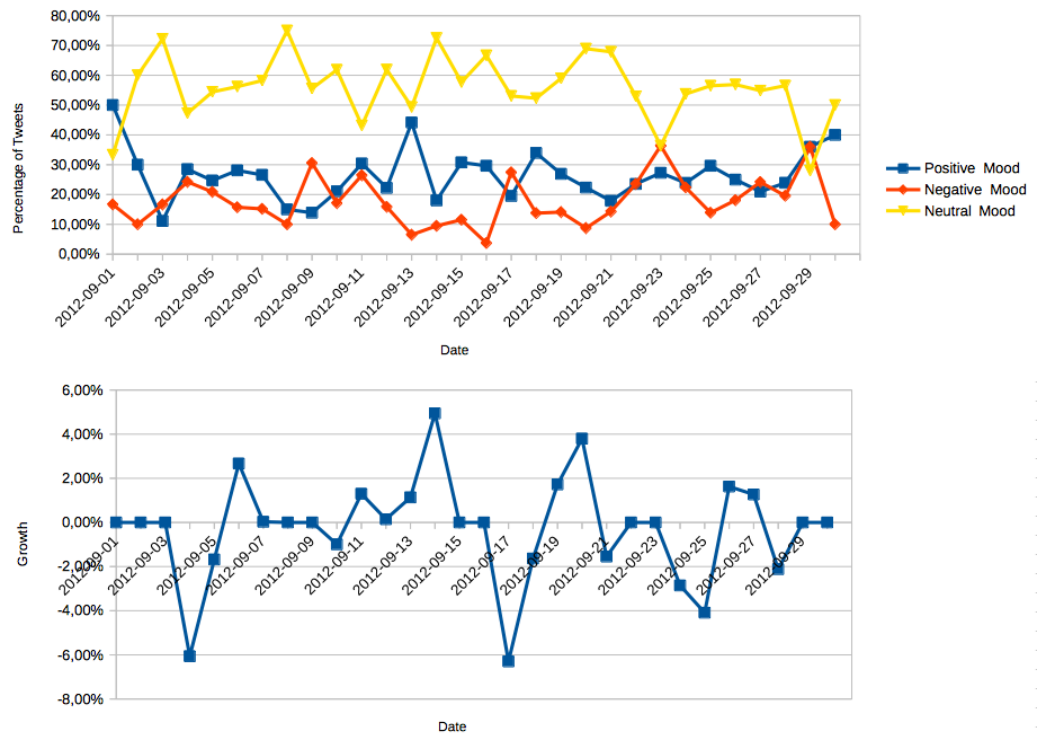


(a) Sept. 2012

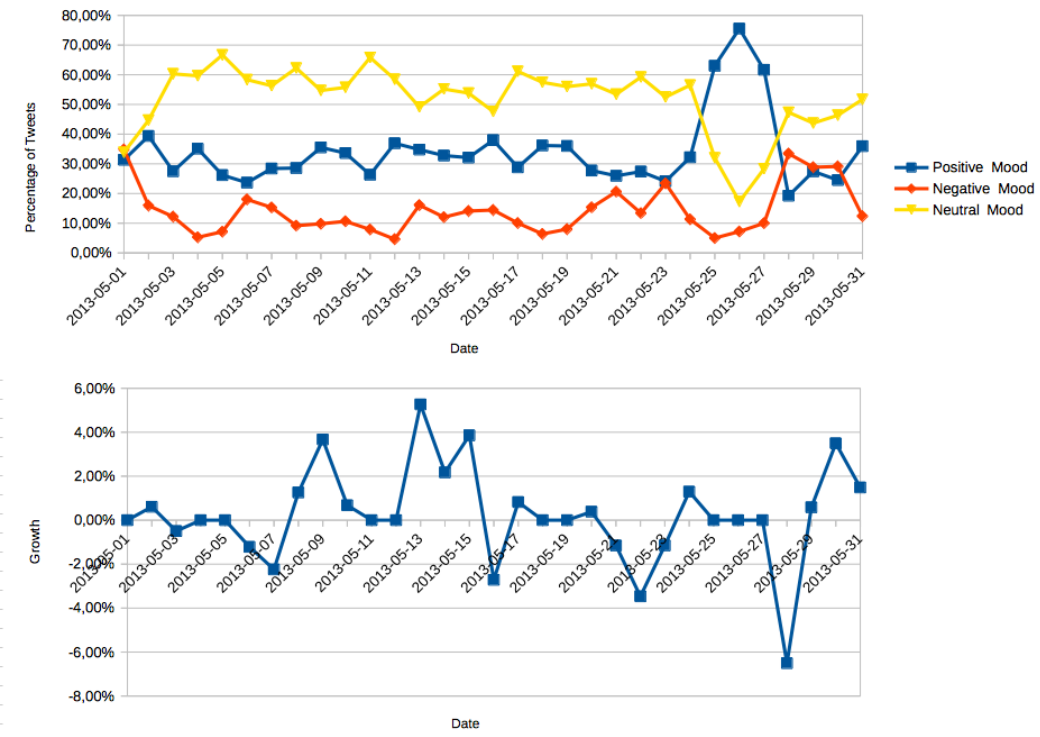


(b) May 2013

Figure 5.12: Goldman Sachs qualitative analysis



(a) Sept. 2012



(b) May 2013

Figure 5.13: Netflix qualitative analysis

Chapter 6

Conclusions

In this thesis, we analysed Twitter dataset to investigate if there is any correlation between tweets and fluctuation of the stock market. We analysed a much bigger dataset compared to the studies done in the past, selecting top six most tweeted stocks for analysis, ranging from various domains. Out of total days for which we performed our experiment, we find out that for 78.3 % of days the values are positively correlated and the average distance between stocks volume and tweets volume, on correlated days, is lower than 16%. We also performed sentiment analysis on the tweets to find correlation among the sentiment and fluctuation in the stock price. To understand the correlation between the daily sentiment about one specific stock and the performance of that stock in the markets we performed sentiment analysis on the tweets. As a supplement, we also analyzed the network structure of the dataset to understand the relation among users of the Twitter dataset.

6.1 Future Work

We are planning to extend this work in following dimensions.

1. One of the important lines of work is to analyse richer dataset. By richer, we mean following

   (a) By clubbing web data (for example, from blogs) with Twitter dataset.

   (b) Dataset which spans across over a much longer period of time.

   (c) By considering not only cashtagged tweets, however also the one which uses the name of the companies in their tweets.

2. In this thesis, we only performed experiments on six top most tweeted stocks. However, we would like to perform similar experiments on a larger number of stocks. Also, we would like to perform experiments on larger number of months.

3. We would also like to propose a model based on our rich dataset which can predict the stock market in the future, like in [19][18].

4. Our sentiment analysis method being used is just a first attempt, we can call it Term Weight Analyisis or Vocabulary Analysis. In future we would like to improve our sentiment analysis based on techniques such as the Natural Language Processing[23] or other techniques rooted in machine learning, using Support Vector Machines, like in [8][6][19].

Ringraziamenti

Vorrei ringraziare innanzitutto il Professor Montesi, per avermi introdotto in questo ambito di ricerca.

In secondo luogo un grande ringraziamento va a Rajesh, il cui preziosissimo aiuto ha contribuito fortemente a plasmare questa tesi.

Un grazie va anche al Dottor Poda ed ai suoi preziosi silenzi.

Ringrazio tutti gli amici con cui ho condiviso birre, pizze, chiacchiere e risate in questi anni.

Tutti i professori che hanno stimolato la mia curiosità.

La mia famiglia, che mi sostiene sempre.

Infine ringrazio Adele, che sa quando serve uno schiaffo o un abbraccio per risvegliarmi, scuotermi e farmi tornare a credere in me.

Appendices

Appendix A

How to build a dataset

In this section we'll show how to build a dataset from scratch using twitter's APIs and we'll give a working example.

A.1 An example

The first obvious step is to subscribe on Twitter, then, to subscribe as a developer on `http://dev.twitter.com`. Doing so is necessary to create a new app. It's enough to compile the form in the requested fields to complete this operation.

Now we have all the necessary to download tweets from Twitter. By default the app permissions are read-only, but this is okay for our purpose. All we need to do now is get the Consumer key and the Consumer secret, that we'll use in our R program, a freely available open-source statistics package[28], using the R package called StreamR, that gives us the access of the Twitter stream API via R.

```
library(streamR)
library(ROAuth)

requestURL <− "https://api.twitter.com/oauth/request_token"
accessURL = "https://api.twitter.com/oauth/access_token"
authURL = "https://api.twitter.com/oauth/authorize"
#these keys are just placeholders
consumerKey <− "xxxxxyyyyyzzzzzz"
consumerSecret <− "xxxxxxyyyyyzzzzzzz111111222222"
my_oauth <− OAuthFactory$new(consumerKey=consumerKey,
consumerSecret=consumerSecret, requestURL=requestURL,
accessURL=accessURL, authURL=authURL)
```

```
my_oauth$handshake(cainfo = system.file("CurlSSL", "cacert.pem",
 package = "RCurl"))
```

After the execution of this last line R will output a message saying that it's necessary to follow the given link and get a PIN to proceed with the handshake and therefore with the tweet download. After the insertion of the PIN, finally we can start to actually download the tweets.

```
## capture tweets containing the $AAPL cashtag tweeted in the next 120 seconds
filterStream( file.name="tweets_$AAPL.json", track="$AAPL",
timeout=120, oauth=my_oauth )
## convert tweets in JSON to data frame
x <- parseTweets("tweets_$AAPL.json")
## output the dataframe into a .csv file
write.csv(x, "downloadedTweets.csv")
```

What we finally obtain is a .csv file with many more fields than our dataset, due to an updates made during the time of the APIs themselves, such as retweeted count, favorited, truncated, friends count, country code, ecc ecc.

Appendix B

Java Code for Data Cleaning and Elaboration

This section includes all the java programs used for data cleaning and elaborations on the dataset.

Thus our dataset is essentially a .csv table, where each line corresponds to a tweet and its metadata, divided each other by commas: "id" is a 17 digits number, unique identifier of a tweet, "lang" indicates with two letters the language of the tweets, "date" indicates, using the USA date notation, the day on which dates back the tweet, same thing for the "time" field, "userid" is a 9 digits unique user identifier, "user" contains the username, and, at last, the field "text" contains the effective 140 character that compose the tweet. After building the dataset, the first step is to parse it in order to clean it, keeping only the informations we consider important to our analysis. In our case we've decided to keep only the tweetId, useful for avoid duplicates, date, time, username, the list of the stocks tickers mentioned in the tweet and, if present, the URL.

B.1 Data Cleaning

```
package transform;

import java.io.BufferedReader;
import java.io.FileReader;
import java.io.PrintWriter;
import com.sun.org.apache.xalan.internal.xsltc.runtime.Hashtable;

public class DataCleaning {
```

```
public static void main(String [] args) throws Exception{
    PrintWriter out = new PrintWriter("log_ft.txt", "UTF−8");
    System.out.println(args[0]);
    try {
    formatTweets(args[0]);
    }
    catch (Exception e) {
        out.println("Problem␣with␣formatTweet␣"+e); //out.write
        System.err.print("Problem␣with␣formatTweet␣"+e);
    }
    out.close();
}
public static void formatTweets(String inputFile) throws Exception{
    BufferedReader br = new BufferedReader(new FileReader(inputFile));
    String line;
    String tweetId;
    String date;
    String time;
    String userName;
    String StockList;
    String URL;
    char a, b, c;
    int i, j, comma;
    PrintWriter writer = new PrintWriter("formattedTweetsJava.csv", "UTF−8");
    writer.println("userName␣StockList,␣URL,␣time,␣date");
    Hashtable noDuplicates = new Hashtable();
    br.readLine(); //I don't transcribe the header line, since we'll have different header
    while ((line = br.readLine()) != null) {
        i = 0; //line index
        j = 0; //misc index
        comma = 0;
        URL = "none";
        StockList = "␣";
        if (line.charAt(0) < '0' && line.charAt(0) > '9') continue;
        //while((a = line.charAt(i)) != ',' && i < (line.length() −1) ) i++;
        while((a = line.charAt(i)) != ',') i++;
        tweetId = line.substring(0, i); // i − 1?
        i++;
        if (line.charAt(i) != ',') {i+=3;}//consume the lang field, if exists
        else i++;
        date = line.substring(i, i+10);
```

```
i+=10;
i++; //consume the comma between date and time
time = line.substring(i, i+8);
i+=8;
i++; //consume the comma between time and userid
if (line.charAt(i) != '0'){ //for the lines with not null uid
    while((line.charAt(i)) != ',') {i++;}
    i++; //consume the comma
}
else {
    i+=2; //consume "0,"
}
while((a = line.charAt(i+j)) != ',') j++; //count the length of username
userName = line.substring(i, i + j);
i += j;
j = 0; //I need it later in this cycle
while (comma == 0){ //while I haven't parsed all the tweet
    i++;
    if (line.charAt(i) == ',' && line.charAt(i − 1) == ','
    && line.charAt(i − 2) == ',') {comma = 1;}
    else if (line.charAt(i) == '#' || line.charAt(i) == '$'){//stocks parsing
        b = line.charAt(i+1);
        c = line.charAt(i+2);
    if(( (b>='A' &&b<='Z') || (b>='a' &&b<='z'))
    && ( (c>='A' &&c<='Z') || (c>='a' &&c<='z'))){
            j++;
            while(((a = line.charAt(i+j)) >= 'A'
            && a <= 'Z') || a >= 'a' && a <= 'z') j++;
            if (j <= 6){ //Don't transcribe hash/cashtags that aren't stocks
            StockList+=line.substring(i, i + j);
            StockList+="␣";
            }
             i += j;
             j = 0;
        }
    }
    if (line.charAt(i) == 'h'){//URL parsing
        b = line.charAt(i+1);
        c = line.charAt(i+2);
        if ((b == c) && b == 't'){
            while ((((a = line.charAt(i+j)) >= 'a') && a <= 'z') ||
```

```
                        (a >= '0' && a <= '9')|| (a >= 'A' && a <= 'Z')||
                        a == ':' || a == '/' || a == '.') j++;
                    URL = line.substring(i, i + j);
                    i += j;
                    j = 0;
                }
            }
        }
        if (noDuplicates.put(tweetId, userName) == null){//if it's not a duplicate
            if (! (StockList.contains("␣$wag␣") || StockList.contains("␣$wAg␣")
                    || StockList.contains("␣$WAg␣") || StockList.contains("␣$wAG␣")
                    || StockList.contains("$waG") || StockList.contains("$hit")
                    || StockList.contains("$HIT"))){ //filter only for WAG stocks
            writer.println(userName+",␣"+StockList+",␣"+URL+",␣"+time+",␣"+date+"
            }
        }
    }
    br.close();
    writer.close();
  }
}
```

B.2 Sentiment Analysis

This program has to be run two times for each input file, one time to get the number of positive mood tweets per day, one for the negative ones. The number of neutral mood tweets is obtained as:

$$Neutral = TweetsVolume - PosTweets - NegTweets \tag{B.1}$$

```
package transform;

import java.io.BufferedReader;
import java.io.FileReader;
import java.io.PrintWriter;
import java.util.HashMap;
import java.util.Iterator;
import java.util.Map;
import java.util.Set;
import java.util.StringTokenizer;
```

```
public class mood {
   private static Map<String, Integer> dictionary = new HashMap<String, Integer>();
   private static Map<String, Integer> moods = new HashMap<String, Integer>();
   private static Map<String, Integer> counts = new HashMap<String, Integer>();

   public static void main(String [] args) throws Exception{
      PrintWriter out = new PrintWriter("log_ft.txt", "UTF-8");
      System.out.println(args[0]);
      try {
         fillHash(args[0]);
         }
         catch (Exception e) {
            out.println("Problem with fillHash "+e); //out.write
            System.err.print("Problem with fillHash "+e);
         }
      try {
      createMoods(args[1]);
      }
      catch (Exception e) {
         out.println("Problem with createMoods "+e); //out.write
         System.err.print("Problem with createMoods "+e);
      }
      try {
         Print();
         }
         catch (Exception e) {
            out.println("Problem with Print "+e); //out.write
            System.err.print("Problem with Print "+e);
         }
      try {
         PrintBis();
         }
         catch (Exception e) {
            out.println("Problem with PrintBis "+e); //out.write
            System.err.print("Problem with PrintBis "+e);
         }
      out.close();
   }

   public static void fillHash(String inputFile) throws Exception{
      BufferedReader br = new BufferedReader(new FileReader(inputFile));
```

```
        String line;
        String word;
        int i, val;
        while ((line = br.readLine()) != null) {
            i = 0;
            while(line.charAt(i)!='\t') i++;
            word = line.substring(0, i);
            i++; //to skip the tab
            if (line.charAt(i) == '-'){
                val = Character.getNumericValue(line.charAt(i+1));
                val=val*(-1);
            }
            else {
                val = Character.getNumericValue(line.charAt(i));
            }
            //if the word hasn't already been inserted in the hashtable.
            if (!dictionary.containsKey(word)){
                dictionary.put(word, val);
            }
        }
        br.close();
    }

    public static void createMoods(String inputFile) throws Exception{
        BufferedReader br = new BufferedReader(new FileReader(inputFile));
        PrintWriter writerNotIn = new PrintWriter("wordsNotInDictionary.csv", "UTF-8")
        PrintWriter writerIn = new PrintWriter("wordsInDictionary.csv", "UTF-8");
        String line;
        String tweetId;
        String date;
        String time;
        String userName;
        String StockList;
        String URL;
        String tweet;
        char a, b, c;
        int i, j, tmp, wordVal, tweetMood, count;
        double posMoodPerc;
        //Hashtable noDuplicates = new Hashtable(); <------Think about this
        br.readLine(); //I don't transcribe the header line, since we'll have different header
        while ((line = br.readLine()) != null) {
```

```
i = 0; //line index
j = 0; //misc index
tweetMood = 0;
count = 0;
URL = "none";
StockList = "␣";
if (line.charAt(0) < '0' && line.charAt(0) > '9') continue;
while((a = line.charAt(i)) != ',') i++;
tweetId = line.substring(0, i); // i − 1?
i++;
if (line.charAt(i) != ',') {i+=3;}//consume the lang field, if exists
else i++;
date = line.substring(i, i+10);
i+=10;
i++; //consume the comma between date and time
time = line.substring(i, i+8);
i+=8;
i++; //consume the comma between time and userid
if (line.charAt(i) != '0'){ //for the lines with not null uid
    while((line.charAt(i)) != ',') {i++;}
    i++; //consume the comma
}
else {
    i+=2; //consume "0,"
}
while((a = line.charAt(i+j)) != ',') j++; //count the length of username
userName = line.substring(i, i + j);
i += j;
j = 0; //I need it later in this cycle
i++; //to skip the comma between username and the tweet
writerNotIn.println();
writerNotIn.print(tweetId+"␣");
writerIn.println();
writerIn.print(tweetId+"␣");
tweet = line.substring(i, line.length());
StringTokenizer st =new StringTokenizer(tweet);
//System.out.println("words count: " + st.countTokens());
// iterate through st object to get more tokens from it
while (st.hasMoreElements()) {
    String token = st.nextElement().toString();
    token= token.toLowerCase();
```

```
                while ((token.charAt(0) < 'a' || token.charAt(0) > 'z') && token.length()>1
                    token = token.substring1);
                while ((token.charAt(token.length() −1) < 'a'
                || token.charAt(token.length()−1) > 'z') && token.length()>1)
                    token = token.substring(0, token.length()−1);
            if (dictionary.containsKey(token)){
                wordVal = dictionary.get(token);
                tweetMood+=wordVal;
                writerIn.print(token+"␣");
            }
            else {
                writerNotIn.print(token+"␣");
            }
        }
        //Day mood perc (pos, neg, neu)
        //all words of this tweet have been considered. TweetMood is now definitive.
        //I increase (or set to 1) the value in hash only if mood is positive.
        if (tweetMood > 0){ //change this line to obtain number of neg tweets
        //if the word is contained into the dictionary, update the mood of that day
            if (moods.containsKey(date)){
                tmp = moods.get(date);
                moods.put(date, tmp+1);
            }
            else {
                moods.put(date, 1);
            }
        }
        //Tweets per day counter is increased anyway
        if (counts.containsKey(date)){
            tmp = counts.get(date);
            counts.put(date, tmp+1);
        }
        else {
            counts.put(date, 1);
        }
    }
    br.close();
    writerNotIn.close();
    writerIn.close();
}
```

```
    public static void Print() throws Exception{
        Double x = 0.0;
        PrintWriter writer = new PrintWriter("PosMoodsBACsep12.csv", "UTF-8");
        Set<String> setOfKeys = moods.keySet();
          Iterator<String> iterator = setOfKeys.iterator();
          while(iterator.hasNext()){
           String key = (String) iterator.next();
           Integer value = moods.get(key);
           writer.println(key+": "+ value);
           x++;
          }
        writer.close();
        System.out.println("There are "+x+ " total days.");
    }

    public static void PrintBis() throws Exception{
        Double x = 0.0;
        PrintWriter writer = new PrintWriter("CountStocks.csv", "UTF-8");
        Set<String> setOfKeys = counts.keySet();
          Iterator<String> iterator = setOfKeys.iterator();
          while(iterator.hasNext()){
           String key = (String) iterator.next();
           Integer value = counts.get(key);
           writer.println(key+": "+ value);
           x++;
          }
        writer.close();
        System.out.println("There are "+x+ " total days.");
    }

}
```

Bibliography

[1] `https://about.twitter.com/company`

[2] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector: Tracking epidemics on twitter," in *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, (Berlin, Heidelberg), pp. 599–602, Springer-Verlag, 2010.

[3] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," 2010.

[4] S. Kampakis and A. Adamides, "Using Twitter to predict football outcomes," Nov. 2014.

[5] Twitter and T. M. R. or Lovers?, "http://mashable.com/2013/06/03/twitter-traditional-media/,"

[6] L. Zhang, "Sentiment analysis on twitter with stock price and significant keyword correlation," 2013.

[7] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, 2011.

[8] J. Smailović, M. Grčar, N. Lavrač, and M. Žnidaršič, "Predictive sentiment analysis of tweets: A stock market application," in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, vol. 7947, pp. 77–88, Springer Berlin Heidelberg, 2013.

[9] T. R. H. of Twitter, "http://www.businessinsider.com/how-twitter-was-founded-2011-4,"

[10] O. R. Twttr, "http://techcrunch.com/2006/07/15/is-twttr-interesting/,"

[11] J. D. first tweet, "https://twitter.com/jack/status/20,"

[12] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with web search," *Proceedings of the National Academy of Sciences*, vol. 107, no. 41, pp. 17486–17490, 2010.

[13] R. E. Vossen, "Does chatter matter? predicting music sales with social media," 2013.

[14] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1014, 2009. doi:10.1038/nature07634.

[15] G. Mishne and N. Glance, "Predicting movie sales from blogger sentiment," pp. 155–158, 2006.

[16] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, (Washington, DC, USA), pp. 492–499, IEEE Computer Society, 2010.

[17] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, "Can blog communication dynamics be correlated with stock market activity?," in *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, HT '08, (New York, NY, USA), pp. 55–60, ACM, 2008.

[18] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting stock market indicators through twitter i hope it is not as bad as i fear," *Procedia - Social and Behavioral Sciences*, vol. 26, no. 0, pp. 55 – 62, 2011. The 2nd Collaborative Innovation Networks Conference - {COINs2010}.

[19] R. Bar-Haim, E. Dinur, R. Feldman, M. Fresko, and G. Goldstein, "Identifying and following expert investors in stock microblogs," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, (Stroudsburg, PA, USA), pp. 1310–1319, Association for Computational Linguistics, 2011.

[20] P. G. S. Timm O. Sprenger, Andranik Tumasjan and I. M. Welpe, "Tweets and Trades: The Information Content of Stock Microblogs," *European Financial Management*, vol. 20, pp. 926–957, Nov. 2013.

[21] E. D. Brown, "Will twitter make you a better investor? a look at sentiment, user reputation and their effect on the stock market," 2012.

[22] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, (New York, NY, USA), pp. 513–522, ACM, 2012.

[23] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *CoRR*, vol. abs/0911.1583, 2009.

[24] Y. Mao, W. Wei, B. Wang, and B. Liu, "Correlating s&p 500 stocks with twitter data," in *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, HotSocial '12, (New York, NY, USA), pp. 69–72, ACM, 2012.

[25] `http://finance.yahoo.com/`

[26] F. Å. Nielsen, "A new ANEW: evaluation of a word list for sentiment analysis in microblogs," *CoRR*, vol. abs/1103.2903, 2011.

[27] http://www.cytoscape.org/

[28] R-project, "http:// cran.r-project.org,"