IMT Institute for Advanced Studies, Lucca

Lucca, Italy

Predictive power of web Big Data in Financial Economics

PhD Program in Economics, Markets and Institutions

XXVI Cycle

By

Gabriele Ranco

2015

The dissertation of Gabriele Ranco is approved.

Program Coordinator: Prof. Davide Ticchi,
IMT Institute for advanced studies Lucca

Supervisor: Prof. Guido Caldarelli,
IMT Institute for advanced studies Lucca

Tutor: Prof. Guido Caldarelli,

The dissertation of Gabriele Ranco has been reviewed by:

Igor Mozetič,
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

Tomaso Aste,
Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK.

IMT Institute for Advanced Studies, Lucca

2015

To Sarah

Contents

List of Figures

List of Tables

Acknowledgements

Chapter 2 is the reproduction, with some extensions, of the article ”Coupling news sentiment with web browsing data predicts intra-day stock prices” co-authored with **Prof. Guido Caldarelli** (IMT Institute for Advanced Studies, Lucca); **Prof. Fabrizio Lillo** (Scuola Normale Superiore, Pisa); **Dr. Giacomo Bormetti** (Scuola Normale Superiore, Pisa); **Dr. Ilaria Bordino** (Yahoo! Labs, Barcelona); and **Dr. Michele Treccani** (Mediobanca s.p.a.).

I want to express my deepest gratitude to the reviewer of the manuscript. For giving me the time and the attention to review my thesis.

Collaboration & Support

I would like to express my sincere gratitude to **Prof. Guido Caldarelli** for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge.

A special thanks goes to **Dr. Michele Treccani**. I will never start this study without his support and enthusiasm.

I would like to express the deepest appreciation to **Prof.Fabrizio Lillo** and **Dr. Giacomo Bormetti** to support and stimulate my research . Especially for the support given on the side of financial economics of my thesis.

I am really grateful to **Dr. Ilaria Bordino** who gave me the possibility to spend nine months of my life in the beautiful city of Barcelona, and being in contact with the group of Yahoo! Labs, Barcelona. I would like to thank her for the opportunity to work together with the researchers of one of the greatest companies of web services.

I would like, also, to thank **Dr. Darko Aleksovski** for the help and support that he gave in my Period in Ljubljana. I would like to express the deepest gratitude to **Dr. Igor Mozetič** and all the others researchers and workers of the **Jožef Stefan Institute** of Ljubljana: I would like to thank her for the opportunity to work in a clever and beautiful research team.

I am indebted to my many of my IMTLucca colleagues to have supported me. Specially I would like to thank **Stefano Sebastio, Olga Chiappinelli**,**Daniele Aurigi**,**Alessandro Belmonte**,**Riccardo Di Clemente**,**Francesca Pampaloni**,**Paolo Sgrignoli** e **Maria Gesualdo**.

I owe my deepest gratitude to my family and my girlfriend **Sarah** who have shown me unconditional love and support for this achievement.

Funding

1. FET Open Project SIMPOL "Financial Systems Simulation and Policy Modelling" nr. 610704, IP Project MULTIPLEX nr 317532 .
2. Yahoo! Inc. to have made available its web data.
3. Quantlab group to have made available its financial data.

Vita

Apr. 17, 1983 Born, Jesi (Ancona), Italy

Current Appointments

03/2011 Ph.D. Candidate in Economics, Market and Institution at **IMTLucca** Institute for Advanced Studies XXVI ciclo.

06/2013 Collaboration with Yahoo! Inc.

01/2012 Collaboration with Quantlab group

Research visit

2015 Internship at Joef Stefan Institute

2013-14 Internship at Yahoo! Labs, Barcelona

Higher Education

2010 **M.Sc.** in **Theoretical Statistical Physics**, University of Studies of Trieste .
Thesis: *"Inference f interactions on Financial Markets as inverse problem of Sttistical physics"*.
Supervisor: Prof. Matteo Marsili

2006 **B.Sc**. in **Physics**, University of Studies of Trieste .
Thesis: *"Classification of vibrational motions of small molecular clusters"*.
Supervisor: Prof. Giorgio Pastore

Ph.D. Schools and Conference

01/2015 **Lecturer at XVI Workshop on Quantitative Finance**
Universit degli Studi di Parma,Italy

10/2012 **FOC-CRISIS School on Complex Financial Networks**
IMTLucca Institute for Advanced Studies (Lucca, Italy).

06/2011 CIdE XXII Course of Econometrics for PhD students Summer School in Econometrics 2011.
CIdE - Centro Interuniversitario di Econometria (Bertinoro, Italy).

Publications

1. Ranco G., Bordino I., Bormetti G., Caldarelli G.,Lillo F. & Treccani M. Coupling news sentiment with web browsing data predicts intraday stock prices. (Submitted 01/2015 on *PLOS-ONE*)

Abstract

Due to the availability of big datasets, the digital revolution is profoundly changing our capability of understanding society and forecasting the outcome of many social and economic systems. Increasingly sophisticated semantic techniques are adopted to automatically interpret information published in articles, blogs, newspapers etc. Unfortunately, irrelevant or already commonly known information can increase the noise of these signals and make their predictive power severely affected or vanished. In this thesis we present a novel methodology which combines the information coming from the sentiment conveyed by public news with the browsing activity of the users of a finance specialized portal to forecast price returns at daily and intra-day time scale. To this aim we leverage a unique dataset consisting of a fragment of the log of Yahoo! Finance, containing the news articles displayed on the web site and the respective number of "clicks", i.e. the visualizations made by the users. Our analysis considers 100 highly capitalized US stocks in a one-year period between 2012 and 2013. Noticeably the sentiment signal and the browsing activity individually taken have very small or no predictive power. Conversely, constructing a signal which in a given time interval gives the average sentiment of the clicked news, weighted by the number of clicks, we show that for more than 50% of the investigated companies it Granger causes price returns. Our result indicates a wisdom of the crowd effect which allows to exploit users' activity to identify and weight properly the relevant and surprising news, enhancing considerably the forecasting power of the news sentiment. In addition we study the presence of predictive power between Twitter messages and price return both n terms of volumes and aggregate sentiment and we present an ″event study″ methodology to measure the impact of days of high attention on Twitter on the stock price.

Chapter 1

Introduction

The recent technological revolution with widespread presence of computers, memories and Internet has created an unprecedented situation of data deluge, changing dramatically the way in which we look at social and economic sciences. As people increasingly use the Internet for information such as business or political news, online activity has become a mirror of the collective consciousness, reflecting the interests, concerns, and intentions of the global population with respect to various economic, political, and cultural phenomena. Humans' interactions with technological systems are generating massive datasets that document collective behaviour in a previously unimaginable fashion[1, 2]. By properly dealing with such data collections, for instance representing them by means of network structures[3, 4], it is possible to extract relevant information about the evolution of the systems considered (i.e. trading[5], disease spreading[6, 7], political elections[8]).

Amongst the many fields of applications of data filtering, analysis and modeling, we present here a case of study based on financial systems. Indeed, financial turnovers, financial contagion and, ultimately, crises, are often originated by collective phenomena such as herding among investors (or, in extreme cases, panic) which signal the intrinsic complexity of the financial system [9]. Therefore, the possibility to anticipate anomalous collective behavior of investors is of great interest to policy makers [10, 11, 12] because it may allow for a more prompt intervention, when this is appropriate. It has been shown that from World-Wide-Web (WWW) and social networks data it is possible to obtain some information for the financial market[13, 14, 15].

So far, academics have focused their attention on identifying the data sources that provide a better signal to be correlated the market. Three major classes of data have been examined: web news, search engine queries and social networks. For what concerns the first class, various news datasets have been used

- to connect exogenous news with price movements[16], and to study
- stock price reaction to news[17, 18];
- the correlation between high or low pessimism of media and high market trading volume[19];
- the relation between the sentiment of news, earnings and return predictability[20],
- the role of news in the trading action[21], expecially of short sellers[22];
- the role of macroeconomic news in the performance of stock returns[23],
- and finally the high-frequency market reaction to news[24].

Regarding the analysis of search-engine queries, some works[14, 25] , [26] have studied the relation over time between the daily number of queries related to a particular stock and the amount of daily exchanges over the same stock. In another paper a similar analysis is done for a sample of Russell 3000 stocks, where an increase in queries predicts higher stock prices in the next two weeks[27]. As for social networks and microblogging platforms, Twitter data is becoming an increasingly popular choice for financial forecasting. For example some investigated whether the daily number of tweets predict SP 500 stock indicators[28]. A textual analysis approach to twitter data could be found in other works[13, 29, 30] where the authors find clear relations between mood indicators and Dow Jones Industrial Average. Some other authors used news and

wikipedia data to predict market movements[15].
However, despite the high quality of the Data set used, the level of empirical correlation between stock price derived financial time series and web derived time series remains low , especially when a textual analysis of web messages is used. This could be generated by irrelevant or already commonly known information that can increase the noise of signals and make their predictive power severely affected or vanished. In this thesis we try to overcome this problem by properly weighting the relevant news and focusing on specific moments of the evolution of the markets. The first topic is discussed in chapter 2 and the second in chapters 3 and 4.

In chapter 2 we present a novel methodology which combines the information coming from the sentiment conveyed by public news with the browsing activity of the users of a finance specialized portal to forecast price returns at daily and intra-day time scale. To this aim we leverage a unique dataset consisting of a fragment of the log of Yahoo! Finance, containing the news articles displayed on the web site and the respective number of "clicks", i.e. the visualizations made by the users. Our analysis considers 100 highly capitalized US stocks in a one-year period between 2012 and 2013. Noticeably the sentiment signal and the browsing activity individually taken have very small or no predictive power. Conversely, constructing a signal which in a given time interval gives the average sentiment of the clicked news, weighted by the number of clicks, we show that for more than 50% of the investigated companies it Granger causes price returns. Our result indicates a wisdom of the crowd effect which allows to exploit users' activity to identify and weight properly the relevant and surprising news, enhancing considerably the forecasting power of the news sentiment.

In the chapter 3 we study whether and to what extent the abnormal absolute price returns of stocks can be predicted by means of the number of financial tweets posted on twitter and vice-versa. To this aim we leverage a unique dataset consisting of a log of some messages posted on Twitter. The log we analyze spans a period of one year and a half between 2012 and may 2013, and stores the tweets containing the cash-tag (e.g. $MSFT for Microsoft) of a company of the Dow Jones Industrial Average index (DJIA). The tag is putted by the author of the message so we are sure that the company mentioned is really the topic of the message. For each of these companies we build a time series of the volume of the tweets, with this aggregate umber we mimic the wisdom-of-crowd effect observed in previous works. Then, we study if, the created time series is really the most correlated with the stock's evolution of his ticker and secondly we identify days of abnormal returns or attention.Finally we calculate the mean reaction for all the companies and days in a window of time starting five days before the event and ending five days later. Results shows that in the day of abnormal return in mean there is an abnormal attention even one days earlier and after and also that in days of abnormal attention there is an increase in the absolute price return.

In chapter 4 We test the presence of the relation between aggregate sentiment of Twitter messages and price return through the well known technique of "event study" [31],[32]. This technique has been generally used to see if the content of earning announcement conveys information for the valuation of companies. Here we use a similar approach, but instead of using the value of earnings, we use the aggregate sentiment expressed by Twitter. Results are largely consistent with the existing literature on the information content of earnings [31],[32]. The evidence strongly supports the hypothesis that Twitter messages do indeed convey information useful for the valuation of companies. Additionally there is evidence of relevant information even for days of huge attention of Twitter users that are not related to earning announcements. This fact proof the correctness of the sentiment analysis and together with the result of of Chapter 2 suggests the idea of weighting properly the news that are posted on the social network in order to increse the predictive power.

Chapter 2

Coupling news sentiment with web browsing data predicts intra-day stock prices

The recent technological revolution with widespread presence of computers, users and media connected by Internet has created an unprecedented situation of data deluge, changing dramatically the way in which we look at social and economic sciences. As people increasingly use the Internet for information such as business or political news, online activity has become a mirror of the collective consciousness, reflecting the interests, concerns, and intentions of the global population with respect to various economic, political, and cultural phenomena. Humans' interactions with technological systems are generating massive datasets documenting collective behaviour in a previously unimaginable fashion [1, 2]. By properly dealing with such data collections, for instance representing them by means of network structures [3, 4], it is possible to extract relevant information about the evolution of the systems considered (i.e.

trading [5], disease spreading [6, 7], political elections [8]).

A particularly interesting case of study is that of the financial markets. Markets can be seen as collective decision making systems, where exogenous (news) as well as endogenous (price movements) signals convey valuable information on the value of a company. Investors continuously monitor these signals in the attempt of forecasting future price movements. Because of their trading based on these signals, the information is incorporated into prices, as postulated by the Efficient Market Hypothesis [33]. Therefore the flow of news and data on the activity of investors can be used to forecast price movements. The literature on the relation between news and price movement is quite old and vast. In order to correlate news and price returns one needs to assess whether the former is conveying positive or negative information about a company, a particular sector or on the whole market. This is typically done with the sentiment analysis, often performed with dedicated semantic algorithms as described and reviewed in the Methods Section.

In this chapter, we combine the information coming from the sentiment conveyed by public news with the browsing activity of the users of a finance specialized portal to forecast price returns at daily and intraday time scale. To this aim we leverage a unique dataset consisting of a fragment of the log of *Yahoo! Finance*, containing the news articles displayed on the web site and the respective number of "clicks", i.e. the visualizations made by the users. Our analysis considers 100 highly capitalized US stocks in a one-year period between 2012 and 2013.

For each of these companies we build a signed time series of the sentiment expressed in the related news. The sentiment expressed in each article mentioning a company is weighted by the number of visualizations of the article. In our dataset each click action is associated with a timestamp recording the exact point in time when such action took place. Thus we are able to construct time series at the time resolution of the minute. To the best of our knowledge, this is the first time that an analysis like the one described in this paper is conducted at such intra-day granularity. The main idea behind this approach is that the sentiment analysis gives information on the news, while the browsing volume enable us to properly weigh news according to the attention received from the users.

We find that news on the same company are extremely heterogeneous in the number of clicks they receive, an indication of the huge difference in their importance and the interest these news generate on users. For 70% of the companies examined, there is a significant correlation between the browsing volumes of financial news related to the company, and its traded volumes or absolute price returns. More important, we show that for more than 50% of the companies (at hourly time scale), and for almost 40% (at daily time scale), the click weighted average sentiment time series Granger-cause price returns, indicating a rather large degree of predictability.

2.1 Data

2.1.1 Stocks considered

Our analysis is conducted on 100 highly capitalized stocks traded in the US equity markets, which we monitor during a period of one year between 2012 and 2013. The ticker list of the investigated stocks with a distinctive numerical company identifier follows: 1 KBH, 2 LEN, 3 COST, 4 DTV, 5 AMGN, 6 YUM, 7 UPS, 8 V, 9 AET, 10 GRPN, 11 ZNGA, 12 ABT, 13 LUV, 14 RTN, 15 HAL, 16 ATVI, 17 MRK, 18 GPS, 19 GILD, 20 LCC, 21 NKE, 22 MCD, 23 UNH, 24 DOW, 25 M, 26 CBS, 27 COP, 28 CHK, 29 CAT, 30 HON, 31 TWX, 32 AIG, 33 UAL, 34 TXN, 35 BIIB, 36 WAG, 37 PEP, 38 VMW, 39 KO, 40 QCOM, 41 ACN, 42 NOC, 43 DISH, 44 BBY, 45 HD, 46 PG, 47 JNJ, 48 AXP, 49 MAR, 50 TWC, 51 UTX, 52 MA, 53 BLK, 54 EBAY, 55 DAL, 56 NWSA, 57 MSCI, 58 LNKD, 59 TSLA, 60 CVX, 61 AA, 62 NYX, 63 JCP, 64 CMCSA, 65 NDAQ, 66 IT, 67 YHOO, 68 DIS, 69 SBUX, 70 PFE, 71 ORCL, 72 HPQ, 73 S, 74 LMT, 75 XOM, 76 IBM, 77 NFLX, 78 INTC, 79 CSCO, 80 GE, 81 WFC, 82 WMT, 83 AMZN, 84 VOD, 85 DELL, 86 F, 87 TRI, 88 GM, 89 FRT, 90 VZ, 91 FB, 92 BAC, 93 MS, 94 JPM, 95 C, 96 BA, 97 GS, 98 MSFT, 99 GOOG, 100 AAPL. The numerical identifiers are assigned according to the increasing order of the total

number of published news in *Yahoo! Finance*.
We considered three main sources of data for these stocks:

2.1.2 Market data

The first source contains information on price returns and trading volume of the stock at the resolution of the minute. We consider different time scales of investigation, corresponding to 1, 10, 30, 65, and 130 minutes. The above values are chosen because they are sub-multiple of the trading day in the US markets (from 9:30 AM to 4:00 PM, corresponding to 390 minutes). For each time scale and each stock we extract the following time series:

- V, the traded volume in that interval of time,
- R, the logarithmic price return in the time scale,
- σ, the return absolute value, a simple proxy for the stock volatility.

The precise definition of these variables is given in the *Support Information*. Since trade volumes and absolute price returns are known to display a strong intra-day pattern, we de-seasonalize the corresponding time series (in the same SI Section we provide the details about this procedure). This procedure is necessary in order to avoid the detection of spurious correlation and Granger causality due to the presence of a predictable intra-day pattern.

2.1.3 News data

The second source of data consists of the news published on *Yahoo! Finance* together with the time series of the aggregated clicks made by the users browsing each page. *Yahoo! Finance* is the most popular web portal for news and data related to financial companies, offering news and information around stock quotes, stock exchange rates, corporate press releases, financial reports, and message boards for discussion. We analyze a portion of the web-site log, containing news articles displayed on the portal. The articles are manually tagged by a team of trained domain experts, who associate each article with the specific companies (e.g., Google, Yahoo!, Apple, Microsoft) or financial entities (e.g., market indexes, commodities, derivatives) that are mentioned in its text.

In order to automatically detect whether the article is conveying positive or negative news on the company, we perform a sentiment analysis. To obtain a sentiment score, we classify each article with *SentiStrength* [34], a state-of-the-art tool for extracting positive and negative sentiment from informal texts. The tool is based on a dictionary of "sentiment" words, which are manually picked by expert editors and annotated with a number indicating the amount of positivity or negativity expressed by them. The original dictionary of *SentiStrength* is not tailored to any specific knowledge or application domain, thus it is not the most proper choice to compute a *financial* sentiment. To solve this issue, we adapt the original dictionary by incorporating a list of sentiment keywords of special interest and significance for the financial domain [35]. Supported by previous research that studied stock price reaction to news headlines [13, 17, 19, 21], we simplify our data processing pipeline by performing the sentiment analysis computation on the title of each article, instead of using its whole content. The sentiment score is a simple sign $(-1, 0, +1)$ for each news depending on whether there are more positive or negative words in the title.

2.1.4 Browsing Data

Finally, in our analysis we use the information on the browsing volume, that is, the time series of "clicks" that the web users made on each article displayed on *Yahoo! Finance* to visualize its content. Given that the users' activity on this domain-specific portal proved to provide a clean signal of interest in financial stocks [26], we exploit it in this work to weight the sentiment of each article on a given financial company. Specifically, we use the number of clicks of an article as a proxy for the level of attention that users gave to that news. By aggregating over a time window the clicks on all the articles, even published earlier, that mention a particular company, it is possible to derive an estimation of the attention around that company.

In summary, for each time scale and for each stock, the variables we extract from the database are (see Support Information):

- C, the time series of the total number of clicks in a time window,
- S, the sum of the sentiment of all news related to each company,
- WS, the sum of the sentiment of all news weighted by the number of clicks.

The first quantity C is non negative and measures the level of attention in a given time interval for news about a specific company. The S variable is the usual sentiment indicator employed in numerous studies and provides the aggregated sentiment of the company specific news published in a given time interval. The most important and novel quantity is WS, which combines the two previous ones by assigning a sign to each click depending on the sentiment of the clicked news. As for the market variables, we remove the intra-day pattern from the click time series. In fact, both the publication of news [21] and the clicking activity of users [26] show a strong intra-day seasonality. These patterns are probably related to the way humans carry out their activities during the day (e.g. small activity during the lunch time, more hectic activity at the beginning or at the end of the business day).

2.2 Methods

2.2.1 Sentiment Analysis

Regarding the analysis of search-engine queries, some recent works [14, 25, 36, 37, 38] have studied the relation between the daily number of queries related to a particular stock and the trading volume over the same stock.

An incomplete list of contributions on the role of news includes studies investigating (i) the relation between exogenous news and price movements [15, 16, 17, 18], (ii) the correlation between high or low pessimism of media and high market trading volume [19]; (iii) the relation between the sentiment of news, earnings and return predictability [20, 39], (iv) the role of news in the trading action [21, 22, 40, 41]; (v) the role of macroeconomic news in the performance of stock returns [23], and (vi) the high-frequency market reaction to news [24].

For example, in a recent paper [14], related to ours, authors show that daily trading volumes of stocks traded at NASDAQ can be forecasted with the daily volumes of queries related to the same stocks.

In another paper a similar analysis shows that an increase in queries predicts higher stock prices in the next two weeks [27]. As for social networks and micro-blogging platforms [42], Twitter data is becoming an increasingly popular choice for financial forecasting. For example some have investigated whether the daily number of tweets predicts SP 500 stock prices [28, 43]. A textual analysis approach to Twitter data can be found in other works [13, 29, 30, 44] where the authors find clear relations between mood indicators and Dow Jones Industrial Average. Some other authors have used news, Wikipedia data or search trends to predict market movements [15, 45, 46, 47].

There are two main critical aspects in the kind of analyses described above. First, the universe of all the search engine or social network users is probably too large and the fraction of users truly interested in finance is likely quite low, with the consequence that signals based on generic Internet users are very noisy. Second, as we will empirically show below, the universe of news considered is very heterogeneous in terms of their relevance as a signal of future price movement. For example, in a day there might be several positive but almost irrelevant news and only one negative but very important news on a company. Without weighting the relevance of the news, one can easily obtain the wrong signal. The intuition behind the current work is that the number of times a news is viewed by users is a measure of its importance as well as of the surprise it conveys. Moreover the users we consider are not generic, but are those who use one of the most important news and search portals for financial information, namely *Yahoo! Finance*.

2.2.2 Pearson Correlation

To overcome these limitations we collected for each stock and for each time scale a total of six time series, namely V, R, σ, C, S, and WS, and we study their dependence by making use of two tools. First, given two time series X_t and Y_t, we consider the Pearson's correlation coefficient

$$\rho(X,Y) = \frac{\langle X_t Y_t \rangle - \langle X_t \rangle \langle Y_t \rangle}{\sqrt{(\langle X_t^2 \rangle - \langle X_t \rangle^2)(\langle Y_t^2 \rangle - \langle Y_t \rangle^2)}} \tag{2.1}$$

where $\langle \cdot \rangle$ is the time average value. The correlation $\rho(X,Y)$ quantifies the linear contemporaneous dependence. In order to assess the statistical significance of the measured value we perform a statistical test of the null hypothesis that the correlation is zero by randomizing the time series.

2.2.3 Granger Causality

Our main goal is testing for the presence of statistical causality between the variables. To this end our second tool is the Granger causality test [48]. Granger's test is a common test used in time series analysis to determine if a time series X_t is useful in forecasting another time series Y_t. X_t is said to *Granger-cause* Y_t if Y_t can be better predicted using both the histories of X_t and Y_t rather than using only the history of Y_t. The Granger causality can be assessed by regressing Y_t on its own time-lagged values and on those of X_t. An F-test is then used to examine whether the null hypothesis that Y_t is not Granger-caused by X_t can be rejected with a given confidence level (in this paper we use a p-value of 5%).

2.3 Results

The most important aspect of our analysis is to test whether one can forecast financial variables, and more specifically price returns, by using the information on the browsing activity of the users. Namely if, by weighting the sentiment of the clicked news by the number of clicks each news receives, one can improve significantly the predictability of returns.

2.3.1 Heterogenous attention

The first observation is the extreme heterogeneity of the attention that users of *Yahoo! Finance* show towards the financial news of a given company. Figure 2.1 shows the complementary of the cumulative distribution function of the number of clicks per news concerning a given stock. There we show the curves for each of the top 10 stocks and for the aggregate of 100 stocks. In all cases the tail of the distribution is very well fit by a power law behavior [49] with a tail exponent very close to 1 (see Support Information) In fact, the mean exponent across all stocks is 1.15 ± 0.30 and restricting on the top 10 it is 0.99 ± 0.08. This indicates that there is a huge heterogeneity in the number of clicks a news receives and therefore in the importance users give to it. It is also a warning that not weighting properly the importance of the news can lead to overstate the importance of the many irrelevant news and to understate the importance of the few really important ones.

2.3.2 Synchronous correlation

In order to understand how the relation between financial and news variables depends on the time scale, we perform a synchronous correlation analysis. We remark that some variables (C, σ, V) are non negative, while others (R, S, WS) are signed. Not surprisingly thus, we find that the linear correlation between any one of the former and anyone of the latter is always very close to zero. For each of the 100 companies, we compute the Pearson's correlation coefficient ρ between the three sensible pairs made by one "news" time series and one "financial" time series. Figure 2.2 summarizes the results for the 65 min time series. The x axis lists the companies, uniquely identified by a number that provides the rank of the company in the order from the least to the most cited one (as measured by the absolute number of associated news). Thus, 1 corresponds to the company KBH with the least number of news, while 100 to the most cited AAPL. We label the y axis with the pairs (C, V), (C, σ), and (WS, R), while the color scale indicates the level of correlation. We compute the correlation sampling the original time series every 65 minutes, equalizing to zero those values whose significance does not reject the null hypothesis of zero correlation with 5% confidence. Figure 2.2 shows in general a positive but small level of correlation, similarly to the result obtained by Mao [13]. In the fourth row of Table 2.1 we report the percentage of the 100 companies for which we reject the null hypothesis of zero correlation at 5% confidence level.

2.3.3 Time scale

In order to investigate how the correlation changes with the time scale, in Table 2.1 we also show the percentage of rejection for the 1, 10, 30, and 130 minutes time scales. As a general comment, we observe that the number of companies with a significant correlation becomes higher at finer time resolution. This is a known fact for market variables (e.g. volume and volatility), while we document it for the first time at intraday scale also for sentiment and browsing variables. As expected the correlations are stronger for unsigned rather than for signed variables. Nevertheless, we can state the presence of a significant relation between the attention given to news articles (signed on the basis of the sentiment expressed in them) related to a given stock and the evolution of the stock price. Remarkably, in the case of the sentiment time series a progressive increase of the number of companies can be observed, until we reach a hourly granularity, suggesting that this could be a characteristic time scale of the process.

2.3.4 Dynamics of attention

The time scale might in principle depend on the relevance of the news. As we have seen, not all news are equal in terms of the attention they receive from the users. To investigate this dependence, we study the dynamics of the number of clicks an article received after its publication. We compute the cumulative number of clicks received by a given news until a given minute after the publication. We perform this for all minutes in a week after the publication. We then normalize this cumulative time series by dividing it by the total amount of clicks received by the news. We construct ten groups of news based on the deciles of the total number of news they eventually receive, and we compute for each group the average cumulative sum of clicks. The result is shown in Figure 2.3. The inset reports for each decile the typical time scale of attention obtained by an exponential fit of the curves. Remarkably, the time scale of attention is an increasing function of the importance of the news (as measured by the total number of clicks). Irrelevant news are immediately recognized as such, while important news continue to receive attention well after their publication. In general, the time scale of the users' attention ranges between one and two hours after the publication, suggesting that this intraday time scale is probably the most appropriate to detect dependencies among financial variables and browsing activity.

2.3.5 Causality

The synchronous correlation is an important measure of dependence, but not necessarily a sign of causality. Thus we perform a causality analysis by applying Granger's test. We present the results of this analysis, for the 65-minute time horizon, in Figure 2.4. The x axis lists the companies as done in Figure 2.2, while the y axis labels the eight tests that we perform. Black cells correspond to rejection of the null hypothesis of no Granger causality, and the opposite for the white cells. When considering the non-negative variables (V, C, and σ) we observe strong causal relations. Specifically, in 65% of the cases the clicking activity causes the trading volume and in 69% of the cases it causes price volatility. The causality is very strong also in the opposite direction, i.e. volume and volatility cause click volume. This is probably due in part to a reaction of users to anomalously high activity in the market (in terms of volume and/or volatility), while in part it might be a statistical effect due to the fact that all the three variables are very autocorrelated in time, creating strong Granger causal relations in both directions.

We obtain the most interesting and unexpected results when considering the signed variables (R, S, and WS). All these variables are weakly serially autocorrelated. When we consider the sentiment of the news (S, without the clicks), we find that only in 4% of the cases S causes returns, and in 13% of the cases price return causes S. Especially the first value is expected under the null, since at 5% confidence level we expect 5% of false positive. This means that the simple sentiment of the news does not allow to forecast price returns at intraday (hourly) time scale. On the contrary, when we consider the clicks weighted by the sentiment of the news (WS), we find that in 53% of the cases it allows predicting returns and only in 19% of the cases the opposite occurs. In general, companies with more news have higher causality.

2.3.6 Weighting news by users' browsing behavior

These results show that, on a hourly time scale, the simple news sentiment time series alone (i.e. the one without browsing activity) is not able to predict the price returns; instead, if we add the information provided by the browsing activity, we are then able to properly weigh the news (and its sentiment) by the importance the users give to it by clicking the page. Thus, we find the interesting result that the browsing activity combined with the sentiment analysis of the news increases significantly the level of prediction of price returns for the stocks.

2.3.7 Comparison with existing literature

Most of the existing studies on sentiment and predictability of returns focus on daily or longer time scale. In order to compare properly our result with the existing literature, we present in Table 2.2 the results of the above Granger tests on a daily time scale. Table 2.2 shows that, without the browsing activity, S causes R for 18% of the companies and R Granger-causes S in 9% of the cases. Thus there is now some predictability of sentiment, even if the number of companies is quite limited. This is consistent with the existing literature, which reports a weak daily predictability of returns by using sentiment. It is important to note that by adding the browsing activity we can double the number of companies for which there is predictability. In fact, WS Granger-causes R for 37% of the companies and 11% in the opposite direction.

2.4 Discussion

The semantic analysis of the news on a specific company is known to have a small predictive power on the future price movements. At the light of our findings, we argue that this effect could be related to the distribution in the attention that the news receive, as clearly emerge in Figure 1: its scale-free behaviour reflects the extreme heterogeneity in the information they convey and the surprise they generate in the readers.

We show that by adding the clicking activity of the web users, we can increase dramatically the predictive power of the news for the price returns. This occurs because the time series built with only the sentiment of the news gives the same weight to all the news. In this way even irrelevant news are considered, adding noise to the sentiment time series and reducing the predictive power of the signal. Adding the browsing activity means giving a meaningful weight to each news according to its importance, as measured by the attention it receives by the users, and this largely improves the predictability of returns.

The approach to collective evaluation that we proposed in this paper can be useful in many other non financial contexts, since the overflow of information is a common aspect in our lives. In the financial domain, a natural extension of the present work concerns market instabilities and crashes. The analysis presented here is in fact unconditional, i.e. it does not target large price movements or, more generally, abnormal returns. From our societal perspective it would be extremely valuable to have a collective evaluation system, like the one presented here, capable of sifting the relevant information from the pool of data, news, blogs, etc, and to provide early warning indicators of large price movements. Since we have shown that predictability exists also at intraday time scale, this approach could be also used for real-time indicators, as well as for high-frequency instabilities and systemic price cojumps [50], which are

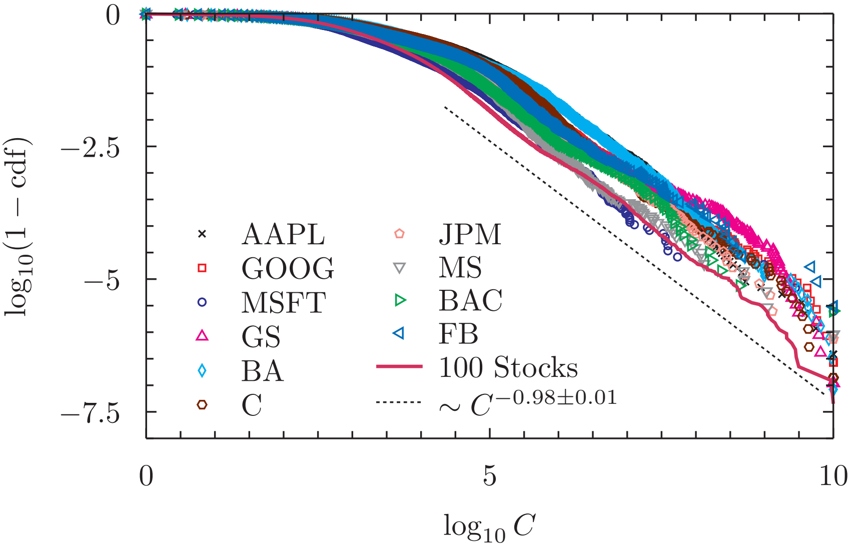

Figure 2.1: Complementary of the cumulative distribution function of the number of clicks a news receives for the ten assets with the largest number of news and the aggregate portfolio of 100 stocks. Both coordinates have been rescaled by a common factor preserving the power law scaling of the right tail and normalizing the maximum number of clicks to the value 10^{10}. The dotted line corresponds to a power law with tail exponent fitted from the portfolio time series. We provide details about the standard error and the complete list of tail exponents for all the companies in Support Information.

becoming increasingly more frequent in our highly automated financial world.

Table 2.1: Percentage of companies for which we reject the null hypothesis of zero correlation at 5% confidence level.

Time interval (minutes)	$\rho(WS, R)$	$\rho(C, \sigma)$	$\rho(C, V)$
1	6	77	78
10	8	48	69
30	10	32	62
65	13	30	57
130	8	29	52

Table 2.2: Number of companies for which we reject the null hypothesis of no Granger causality at 5% confidence level.

Causality relation	Hourly scale	Daily scale
$S \to R$	4	18
$R \to S$	13	9
$R \to WS$	19	11
$WS \to R$	53	37
$V \to C$	100	97
$C \to V$	65	52
$C \to \sigma$	69	52
$\sigma \to C$	96	16

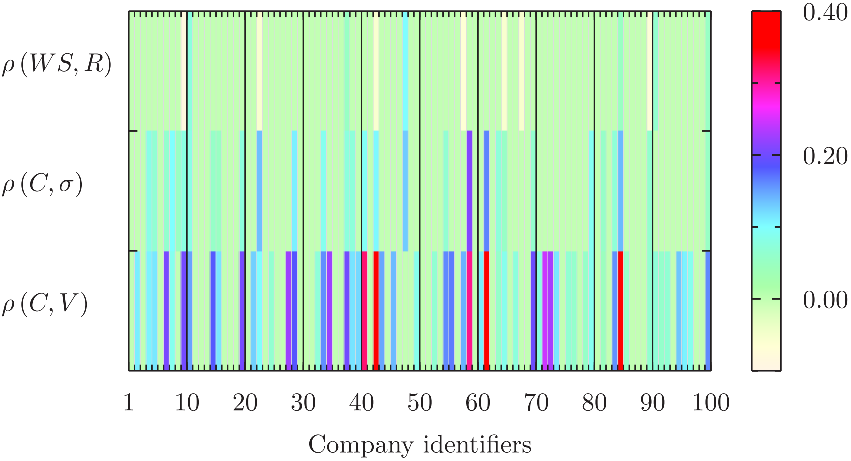

Figure 2.2: Pearson's correlation coefficients for the de-seasonalized time series of all the 100 companies at hourly scale. The x axis reports the list of companies identified by a unique number, as detailed in the main text. Among the several possibilities, we consider only three couples and the color scale corresponds to the level of correlation. We plot those values for which we reject the null of zero correlation at 5% significance level and equalize non significant values to zero (light green color).

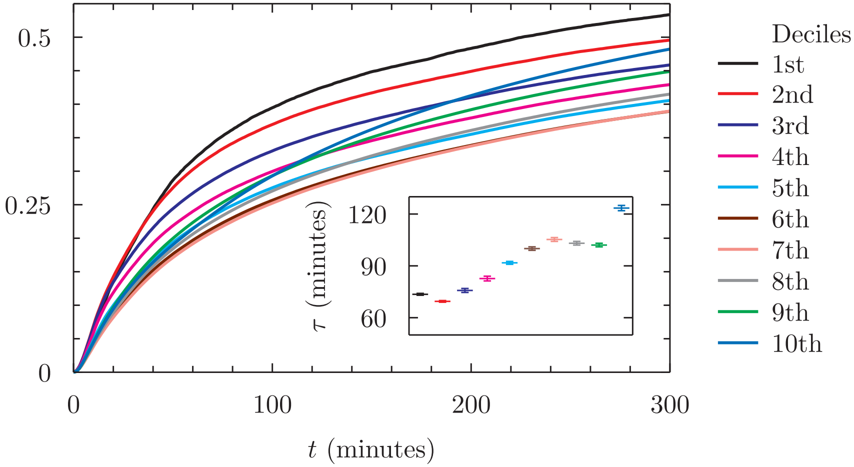

Figure 2.3: Time evolution of the cumulative number of clicks per news in a time interval of five hours after the publication. We normalize the cumulated amount by a constant which corresponds to the total number of clicks received by a single news during the first week after publication. The news are grouped in deciles according to the total number of clicks they have received until October 2013 and the curves represent average values. Inset: estimated values and standard errors of the attention time scale obtained by an exponential fit of the decile curves. We represent points and curves corresponding to the same decile with the same color.

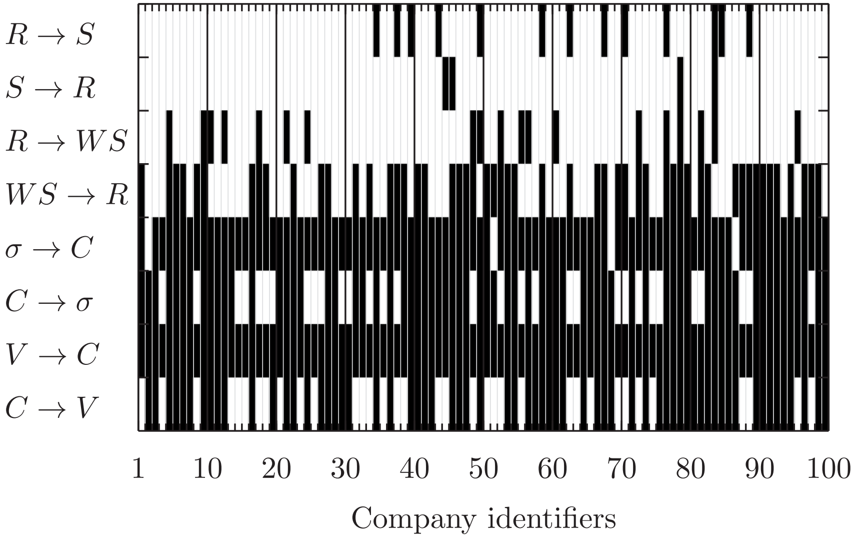

Figure 2.4: Granger Causality tests at hourly scale between de-seasonalized time series (x axis as in figure 2.2). The white cells correspond to tests for which we do not reject the null hypothesis of no Granger causality at 5% significance level. A black cell corresponds to a statistically significant Granger causality.

Chapter 3

Twitter and abnormal returns

Microblogging is an increasingly popular form of communication on the web. It allows users to broadcast brief text updates to the public or to a selected group of contacts. Twitter posts, commonly known as tweets, are extremely short in comparison to regular blog posts, being at most 140 characters in length. The launch of Twitter in October 2006 is responsible for the popularization of this simple, yet vastly popular form of communication on the web. Users of these on-line communities use microblogging to broadcast different types of information:

- daily chatter, e.g., posting what one is currently doing,
- conversations, i.e., directing tweets to specific users in their community of followers,
- information sharing, e.g., posting links to web pages,
- news reporting, e.g., commentary on news and current affairs.

Starting from these data-sets, in principle, it is possible to check the opinion or the interest of Twitter users over several arguments. Until now, academics have conducted three types of analysis in order to forecast some financial index:

- Volume of tweets: stream of users messages about stocks (e.g [28]) .
- Text-based Analysis:Sentiment, mood, topic, tags (see [29], [30], [13]).
- Behavioural and Social Information: Reputation, popularity, users profile, users past performance(e.g. [43]).

In this chapter we focus on the first type of analysis. Literature shows predictive power for volumes of tweets to the traded volumes. However it lacks a relationship between them and possible anomalies in the price dynamics. We study whether and to what extent the absolute price returns of stocks can be predicted by means of the number of financial tweets posted on Twitter and vice-versa. To this aim we leverage a unique dataset consisting of a log of some messages posted on Twitter. The log we analyze spans a period of one year and a half between 2012 and may 2013, and stores the tweets containing the cash-tag (e.g. $MSFT for Microsoft) of a company of the Dow Jones Industrial Average index (DJIA). The tag is entered by the author of the message so we are sure that the company mentioned is really the topic of the message. For each of these companies we build a time series of the volume of the tweets, with this aggregate number we mimic the wisdom-of-crowd effect observed in previous works. Then, we test if, the created time series is really the most correlated with the stock's evolution of his ticker and , also, we measure the level of this correlation. From our analysis it is evident that the tagging procedure is correct, but it is also evident that the level of the correlation is not high. To reduce the noise produced by the not constant human activity we restrict our analysis to days of abnormal events in price dynamics and Twitter volumes. This procedure seems to us more promising because we believe that the relation between web variables and financial variables are more evident during peaks of attention of Twitter users. For these reasons we identify days of abnormal returns or attention. We calculate the mean reaction for all the companies and days in a window of time starting five days before the event and ending five days later. Results shows that in the day of abnormal return in average there is an abnormal attention even one days earlier and after and also that in days of abnormal attention there is an increase in the absolute price return.

3.1 Data

3.1.1 Stocks considered

Our analysis is conducted on 24 highly capitalized stocks traded in the US equity markets, which we monitor during a period of one year and half between 2012 and 2013. The ticker list of the investigated stocks is in Table 3.1

3.1.2 Market data

The first source contains information on price returns and trading volume of the stock at the resolution of the day. For each time scale and each stock we extract the following time series:

- V, the traded volume in that interval of time,
- R, the logarithmic price return in the time scale,
- σ, the return absolute value, a simple proxy for the stock volatility.

The precise definition of these variables is given in the *Methods*.

3.1.3 Twitter data

We analyze a fragment of the Twitter log, spanning 16 months between January 2012 and April 2013, and consisting of the tweets published on the portal produced by the users who post the message during the same period. For each company we create the following time series:

Table 3.1: List of considered stocks.

tag	Name	Index
AXP	American Express Co	DJIA
BA	Boeing Co	DJIA
CAT	Caterpillar Inc	DJIA
CSCO	Cisco Systems Inc	DJIA
CVX	Chevron Corp	DJIA
DIS	Walt Disney Co	DJIA
GE	General Electric Co	DJIA
HD	Home Depot Inc	DJIA
IBM	International Business Machines Co	DJIA
INTC	Intel Corp	DJIA
JNJ	Johnson & Johnson	DJIA
JPM	JPMorgan Chase and Co	DJIA
KO	Coca-Cola Co	DJIA
MCD	McDonald's Corp	DJIA
MRK	Merck & Co Inc	DJIA
MSFT	Microsoft Corp	DJIA
NKE	Nike Inc	DJIA
PFE	Pfizer Inc	DJIA
PG	Procter & Gamble Co	DJIA
UNH	UnitedHealth Group Inc	DJIA
UTX	United Technologies Corp	DJIA
VZ	Verizon Communications Inc	DJIA
WMT	Wal-Mart Stores Inc	DJIA
XOM	Exxon Mobil Corp	DJIA

1. ″tweets (TW)″ time series: a time series of the total number of tweets in a day. in formula for each tweet tw:

   $$TW_d = \sum_i tw_{d_i}$$

   where i indicates a tweet of a given day d.

let us consider the figures 3.4,3.5 to have a measure of the representativeness of our sample. Figure 3.4 shows the number of tweets for each company. Figure 3.5 shows the $\log_{10}$ of the total amount of tweets in a month for a given company. It is evident that the order of magnitude remains almost constant in the period of analysis.

3.2 Methods

3.2.1 Pearson Correlation

To measure the correctness of the tagging procedure we collected for each stock a total of two time series, namely V, TW and we study their dependence. Given two time series X_t and Y_t, we consider the Pearson's correlation coefficient

$$\rho(X,Y) = \frac{\langle X_t Y_t \rangle - \langle X_t \rangle \langle Y_t \rangle}{\sqrt{(\langle X_t^2 \rangle - \langle X_t \rangle^2)(\langle Y_t^2 \rangle - \langle Y_t \rangle^2)}} \tag{3.1}$$

where $\langle \cdot \rangle$ is the time average value. The correlation $\rho(X,Y)$ quantifies the linear contemporaneous dependence. For each of the companies, we computed Pearson's correlation between all possible pairs of TW time series and V time series to measure if the Twitter time series is really related with the traded stock. We choose the traded volume as quantity to this preliminary test because in literature it has been shown that this quantity is higly correlated with the volume of tweets [13][28].

3.2.2 High attention and abnormal returns

In order to identify peeks of abnormal tweets volume or abnormal attention we use the following procedures:

- We compute the moving average for both the time series (TW, σ) for a given number of days (in our case [5,40] days).
- we divide the original time series for the moving average time series that we have created. So if d is a given day we compute the mean over the N days before d :

  $$\hat{TW}_d = \frac{TW_d}{(\frac{1}{N}\sum_{j=d-1-N}^{d-1} TW_j)}$$

  with similar notation for σ:

  $$\hat{\sigma}_d = \frac{\sigma_d}{(\frac{1}{N}\sum_{j=d-1-N}^{d-1} \sigma_j)}$$

  where $\hat{\sigma}, \hat{TW}$ are the rescaled time series.
- we consider only days having $\hat{\sigma}, \hat{TW}$ over a given limit that we can call θ
- after having identified those days we consider an interval of 5 days before and after the events and compute the mean values over every event: independently from each company
- In order to assess the statistical significance of the measured value we perform a statistical test of the null hypothesis that the events are independent by randomizing the time series.

3.3 Results

The most important aspect of our analysis is to test whether one can find evidence of relation between writing activity of Twitter users and financial variables, and more specifically abnormal price returns.

3.3.1 First test: correctness of tagging procedure

For each of the companies, we computed Pearson's correlation between all possible pairs of TW time series and V time series to measure if the

Twitter time series is really related with the traded stock.. We choose the traded volume as quantity to this preliminary test because in the literature it has been shown that this quantity is higly correlated with the volume of tweets. Figure 3.1 summarizes the results. The x and y axes lists the companies: the first is the volume of tweets , the second is the volume of trades. The color bar indicates the level of correlation. It is possible to see that the level of cross correlation is higher both in values and in number of companies for the couples having the same tags. This first result proves that our (TW) time series is really related to the trading evolution of the stock having the same tag.

3.3.2 Relations between abnormal volume of tweets and price returns

In order to understand how the relation between financial and Twitter variables depends on the magnitude of Twitter volumes, we perform the analysis shown in figure 3.2 following the procedure explained in Methods. In order to verify the dependence for our parameters we consider two values for the N days of the mean estimation (5,40) and for the values of θ we chose 4 values (5,6,8,10). in figure 3.2 there are numbers on the upper right part of the plots that shows those values.. It is possible to see that in every plot the day in witch there is an abnormal number of tweets is related with an increase of the absolute return of the price. Moreover we can see that for all our choice an increase of price return could be found in a day before the events and when we consider $N = 40$ even in the following day. If we consider that the absolute price return could be considered as a proxy for the volatility we can say that in periods of increase of the volatility an increase of the number of tweets could be found. If we compare the measured result (the blue lines in the figure) with the hypothesis of independence (the green line) we see clearly a signal of relation between the Twitter users activity and financial events.

3.3.3 Relations between price returns and volumes of tweets

For the figure 3.3 we follow the same procedure as before, but we consider events the days of abnormal price returns.. It is possible to see that an increase of the number of tweets could be found before the abnormal price events. In particular the number of tweets increase even a couple of days before the event day for $\theta > 10$. This result could be considered a signal of the wisdom of the crowd effect of the Twitter users

3.4 Discussion

The analysis of the internet users micro-bloging activity for a specific company is known to have predictive power on the future traded volumes movements. At the light of our findings, we argue that this effect could be seen also for abnormal days of price returns, as clearly emerge in Figure 3.3 the empirical lines behaviour reflects the extreme events occurrence. We show that by only looking to the Twitter volumes, we can see, independently from the company chosen the predictive power of the Twitter data for the abnormal price returns. Our result show evidently a relation, only by consider variation in respect the local mean of tweets volume time series without too many manipulation of the time series. The approach to collective evaluation that we proposed in this paper can be useful in many other non financial contexts, since the overflow of information is a common aspect in our lives. In the financial domain, a natural extension of the present work concerns market instabilities and crashes. The analysis presented here does not consider the mood of the tweets and we remains in a daily time scale. A natural extension of this work is the use of the volumes of positive and negative tweets. A study also at the intra -day level could be useful to forecast the micro-structural instabilities that nowadays have received a huge interest by academics and policy makers.

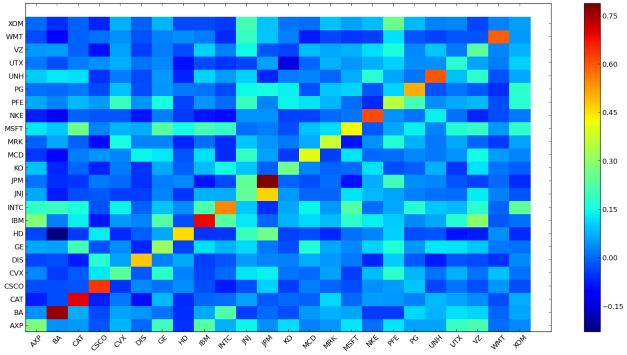

Figure 3.1: The x and y axes lists the companies: the first is the volume of tweets , the second is the volume of trades. The color bar indicates the level of correlation. It is possible to see that the level of cross correlation is higher both in values and in number of companies for the couples having the same tags.

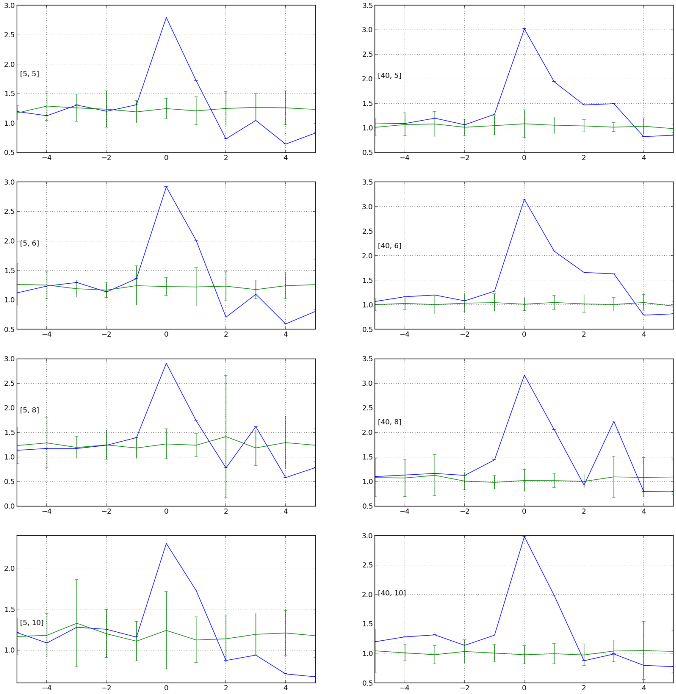

Figure 3.2: Plots of the mean evolution of price returns 5 days before and after a huge variation of tweets volume. the blue lines is empirical measure and the green line is the randomization test for the null hypothesis of independence.In the uppper right corner of each plots it is shown the number of days considered for the estimation of the moving average and the different values of θ

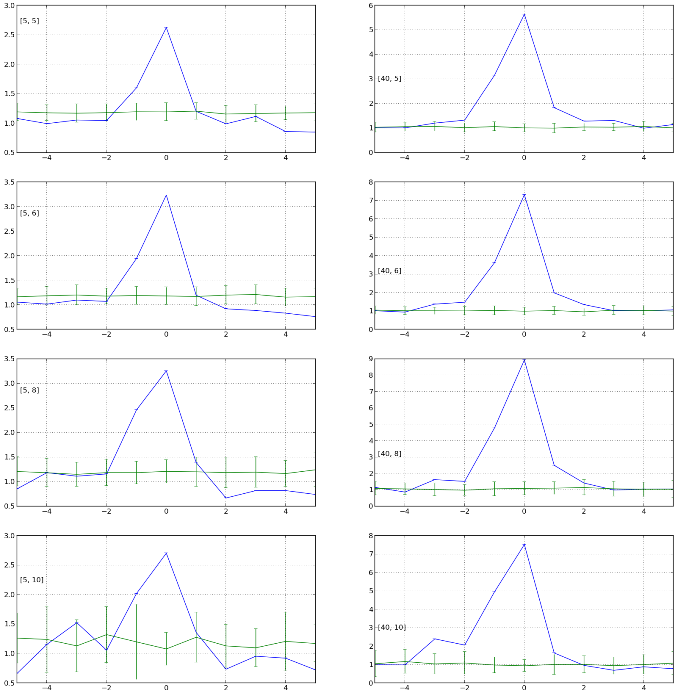

Figure 3.3: Plots of the mean evolution of tweets volume 5 days before and after a huge variation of price return. the blue lines is empirical measure and the green line is the randomization test for the null hypothesis of independence.In the uppper right corner of each plots it is shown the number of days considered for the estimation of the moving average and the different values of θ

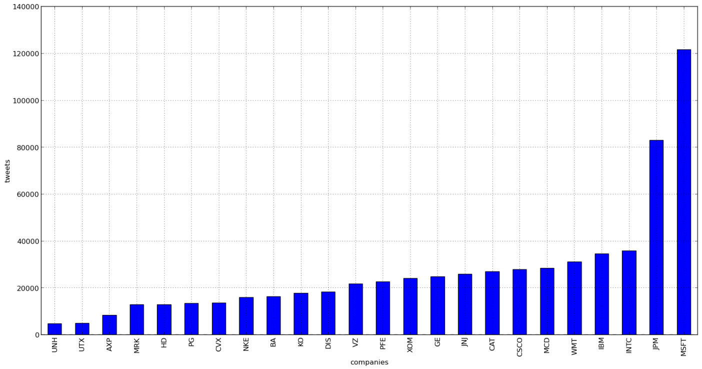

Figure 3.4: Distribution of tweets for each company

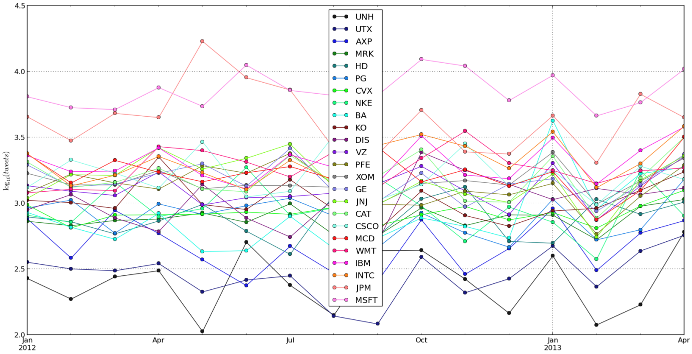

Figure 3.5: Distribution of tweets for each month

Chapter 4

Event Study from Tweets volume and sentiment

In the previous Chapter 3 we have seen that exists a relation between Tweets volumes and price return. This relation is present at least at daily level and for abnormal variation of these two variables. Also it has been shown that aggregate tweets sentiment forecasts at daily level aggregate price index like Dow Jons [13],[29] and that at intra day level social media message sentiment can contain statistically-significant ex-ante information on the future prices [51]. In this Chapter we would like to test the presence of predictive power for sentient time series at daily resolution for stock price index. The data sample of Twitter message for each company is created following the rule of the Chapter 3(e.g. $MSFT for MIcrosoft inc.). in the following sections we have tried to follow the approach of the Chapter 2, but the Result that we have both for Person Correlation and Granger Causality are very poor. In order to be sure that there is in any case a relation between the computed sentient and the price return we try to test the null hypothesis of independence between our time series. We test the presence of this relation through the well known technique of ″event study″ [31],[32]. This technique has been generally used to see if the content of earning announcement conveys information for the valuation of companies. Here we use a similar approach, but instead of using the value of earnings, we use the aggregate sentiment expressed by Twitter. Results are largely consistent with the existing literature on the information content of earnings [31],[32]. The evidence strongly supports the hypothesis that Twitter messages do indeed convey information useful for the valuation of companies. Additionally there is evidence of relevant information even for days of huge attention of Twitter users that are not related to earning announcements. This fact proof the correctness of the sentiment analysis and together with the result of of Chapter 2 suggests the idea of weighting properly the news that are posted on the social network in order to increse the predictive power.

Data

Our analysis is conducted on 30 stocks of the DJIA index. The stock data are collected for a period of 15 months between 2013 and 2014. The ticker list of the investigated stocks is shown in Table 4.1. In the analysis we investigate the relation between price/market data, and Twitter data. The details of both are given in the remainder of this section.

Market data

The first source of data contains information on price returns of the stock, with daily resolution. For each stock we extract the time series of daily returns, R_d:

$$R_d = \frac{p_d - p_{d-1}}{p_{d-1}} \tag{4.1}$$

where p_d is the closing price of the stock at day d. This data is publicly available and can be downloaded from various sources on the Internet, as for example the Nasdaq web site (i.e., http://www.nasdaq.com/symbol/nke/hist for the ”Nike” stock).

Ticker	Company	Tweets
TRV	Travelers Companies Corp	12184
UNH	UnitedHealth Group Inc	15020
UTX	United Technologies Corp	16123
MMM	3M Co	17001
DD	E I du Pont de Nemours and Co	17340
AXP	American Express Co	21941
PG	Procter & Gamble Co	25751
NKE	Nike Inc	29220
CVX	Chevron Corp	29477
HD	Home Depot Inc	30923
CAT	Caterpillar Inc	38739
JNJ	Johnson & Johnson	40503
V	Visa Inc	43375
VZ	Verizon Communications Inc	45177
KO	Coca-Cola Co	45339
MCD	McDonald's Corp	45971
XOM	Exxon Mobil Corp	46286
DIS	Walt Disney Co	46439
BA	Boeing Co	51799
MRK	Merck & Co Inc	54986
CSCO	Cisco Systems Inc	57427
GE	General Electric Co	61836
WMT	Wal-Mart Stores Inc	63405
INTC	Intel Corp	68079
PFE	Pfizer Inc	71415
T	AT&T Inc	75886
GS	Goldman Sachs Group Inc	91057
IBM	International Business Machines Co	101077
JPM	JPMorgan Chase and Co	108810
MSFT	Microsoft Corp	183184

Table 4.1: The collected Twitter data for the 15 months period: the company names and the number of tweets.

Twitter data

The second source of data is from Twitter and consists of relevant tweets, along with their sentiment. The data was collected by Twitter search API, where search query consists of the stock cashtag (e.g., $NKE for Nike). The data covers a period of 15 months (from June 1, 2013 to September 18, 2014), for which there is a total of 1.7 million tweets. The tweets and their sentiment were provided by the Sowa Labs[1] company.

The Twitter sentiment is calculated by a supervised learning method. First, a large fraction of tweets were labeled by financial experts with three sentiment labels: negative, neutral or positive. Then, this labeled set was used to build a support vector machine (SVM [52]) classification model which discriminates between the negative, neutral and positive tweets. Finally, the SVM model was applied to the complete set of tweets. The final data set is in the form of a time series of negative, neutral and positive tweets for each day d. In particular, we create the following time series for each company:

- Volume of tweets, TW_d: the total number of tweets in a day.
- Negative tweets, tw_d^-: the number of negative tweets in a day.
- Neutral tweets, tw_d^0: the number of neutral tweets in a day.
- Positive tweets, tw_d^+: the number of positive tweets in a day.
- Sentiment polarity, P_d: the difference between the number of positive and negative tweets as a fraction of non-neutral tweets[53], $P_d = \frac{tw_d^+ - tw_d^-}{tw_d^+ + tw_d^-}$.

Methods

Correlation and Granger causality

For an initial investigation of the relation between the Twitter sentiment and stock prices, we apply the Pearson correlation and Granger causality

[1] http://www.sowalabs.com/

tests. We use the Pearson correlation to measure the linear dependence between P_d and R_d. Given two time series, X_t and Y_t, the Pearson's correlation coefficient is calculated as:

$$\rho(X,Y) = \frac{\langle X_t Y_t \rangle - \langle X_t \rangle \langle Y_t \rangle}{\sqrt{(\langle X_t^2 \rangle - \langle X_t \rangle^2)(\langle Y_t^2 \rangle - \langle Y_t \rangle^2)}} \tag{4.2}$$

where $\langle \cdot \rangle$ is the time average value. The correlation $\rho(X,Y)$ quantifies the linear contemporaneous dependence.

We also perform the Granger causality test [54] to check if the Twitter variables help in the prediction of the price returns. The steps of the procedure applied are summarized as follows [55]:

- Determine if the two time series are non-stationary, by the Augmented Dickey-Fuller (ADF) test.
- Build a Vector Autoregressive (VAR) model and determine its optimal order by considering four measures: AIC, BIC, FPE, HQIC.
- Fit the VAR model with the selected order from the previous step.
- Perform the Ljung-box test for no autocorrelation in the residuals of the fit.
- Perform the F-test to detect statistically significant differences in the fit of the baseline and the extended models (Granger causality test).

Event study

The method used in this paper is based on an event study, as defined in financial econometrics [56]. This type of study analyzes the abnormal price returns observed during external events. It requires that a set of abnormal events for each stock is first identified (using prior knowledge or automatic detection), and then the events are grouped according to some measure of "polarity" (whether the event should have positive, negative or no effect on the valuation of the stock). Then, the price returns for events of each group are analyzed. In order to focus only on isolated events affecting a particular stock, the method removes the fluctuations (influences) of the market to which the stock belongs. This is achieved by using the market model, i.e., the price returns of a selected index.

Event window. The initial task of conducting an event study is to define the events of interest and identify the period over which the stock prices of the companies involved in this event will be examined: the event window, as shown in Figure 4.1. For example, if one is looking at the information content of an earnings announcement on day d, the event will be the earnings announcement and the event window $(T_1, T_2]$ might be $(d-1, d+1]$. The reason for considering one day before and after the event is that the market may acquire information about the earnings prior to the actual announcement and one can investigate this possibility by examining pre-event returns.

Normal and abnormal returns. To appraise the event's impact one needs a measure of the abnormal return. The abnormal return is the actual ex-post return of the stock over the event window minus the normal return of the stock over the event window. The normal return is defined as the return that would be expected if the event did not take place. For each company i and event date d, we have:

$$AR_{i,d} = R_{i,d} - E[R_{i,d}] \tag{4.3}$$

where $AR_{i,d}$, $R_{i,d}$, $E[R_{i,d}]$ are the abnormal, actual, and expected normal returns, respectively. There are two common choices for modeling the expected normal return: the constant-mean-return model, and the market model. The constant-mean-return model, as the name implies, assumes that the mean return of a given stock is constant through time. The market model, used in this paper, assumes a stable linear relation between the overall market return and the stock return.

Estimation of the normal return model. Once a normal return model has been selected, the parameters of the model must be estimated using a subset of the data known as the estimation window. The most common choice, when feasible, is to use the period prior to the event window for the estimation window (cf. Figure 4.1). For example, in an event study using daily data and the market model, the market model parameters

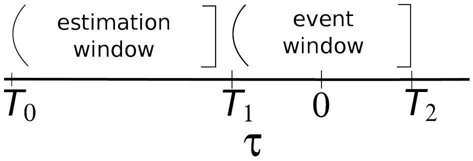

Figure 4.1: Time line for an event study.

could be estimated over the 120 days prior to the event. Generally, the event period itself is not included in the estimation period to prevent the event from influencing the normal return model parameter estimates.

Statistical validation. With the estimated parameters of the normal return model, the abnormal returns can be calculated. The null hypothesis, H_0, is that external events have no impact on the returns. It has been shown that under H_0, abnormal returns are normally distributed, $AR_{i,\tau} \sim \mathcal{N}(0, \sigma^2(AR_{i,\tau}))$ [31]. This forms the basis for a procedure which tests whether an abnormal return is statistically significant.

Event detection using Twitter activity peaks. This part first discusses the algorithm used to detect Twitter activity peaks, which are then treated as events. Next, it describes the method used to assign a polarity to the events, using the Twitter sentiment. Finally, it discusses a specific type of events for the companies studied, called earnings announcement events, which are already known to produce abnormal price jumps.

Detection of Twitter peaks. To identify Twitter activity peaks, for every company we use the time series of its daily Twitter volume, TW_d. We use a sliding window of $2L + 1$ days ($L = 5$) centered at day d_0, and let d_0 slide along the time line. Within this window we evaluate the baseline volume activity TW_b as the median of the window [57]. Then, we define the outlier fraction $\phi(d_0)$ of the central time point d_0 as a relative difference of the activity TW_{d_0} with respect to the median baseline TW_b: $\phi(d_0) = [TW_d - TW_b]/max(TW_b, n_{min})$. Here, $n_{min} = 10$ is a minimum activity level used to regularize the definition of $\phi(d_0)$ for low activity values. We say that there is an activity peak at d_0 if $\phi(d_0) > \phi_t$, where ϕ_t

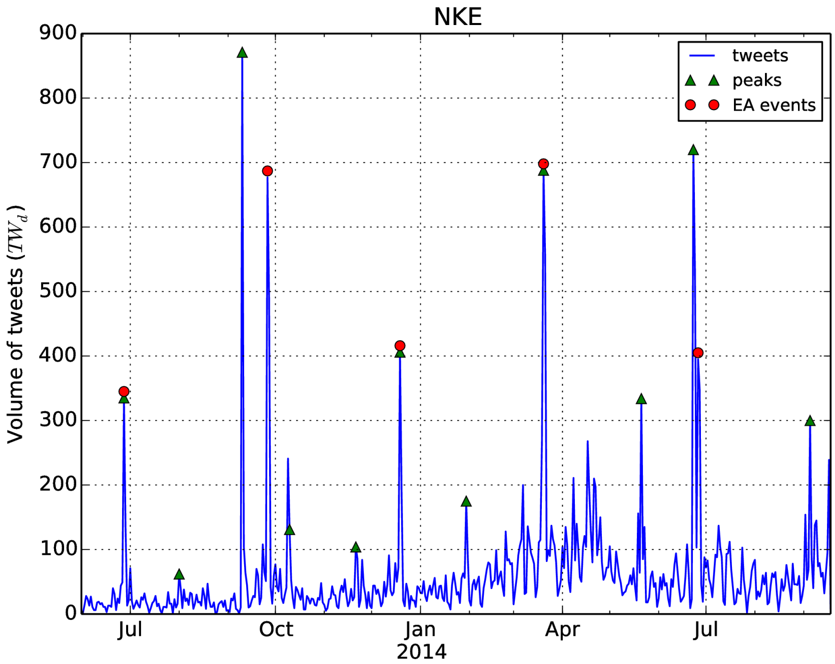

Figure 4.2: Daily time series of Twitter volume with indicated peaks for the Nike company.

is a threshold value, set to $\phi_t = 2$. As an illustration, the resulting activity peaks for the Nike company are shown in Figure 4.2. After the peak detection procedure, we treat all the peaks detected as events. These events are then assigned polarity (from Twitter sentiment) and type (earnings announcement or not).

Polarity of events. Each event is assigned one of the three polarities: negative, neutral or positive. The polarity of an event is derived from the sentiment polarity P_d of tweets for the peak day. From our data we detected 260 events. The distribution of the P_d values for the 260 events is not uniform, but prevailingly positive, as shown in Figure 4.3. To obtain three sets of events with approximately the same size, we use threshold values extracted from the distribution. We determined the following thresholds, and defined the event polarity as follows:

- If $P_d \in [-1, 0.1)$ the event is a *negative event*,
- If $P_d \in [0.1, 0.65]$ the event is a *neutral event*,
- If $P_d \in (0.65, 1]$ the event is a *positive event*.

Putting thresholds on signal is always somewhat arbitrary, therefore we make a simple assumption. Since the sentiment of the tweets is biased towards positive, we consider negative all the tweets with a value of polarity around 0 or below. We then put the threshold when the trend is inverted.

Event types. For a specific type of events in finance, in particular quarterly *earnings announcements* (EA), it is known that the price return of a stock abnormally jumps in the direction of the earnings [31, 32]. In our case, the Twitter data shows high posting activity during the EA events, as expected. However, there are also other peaks in the Twitter activity, which do not correspond to EA, abbreviated as non-EA events. See Figure 4.2 for an example of Nike.

The total number of peaks that our procedure detects in the period of the study is 260. Manual examination reveals that in the same period, there are 151 EA events[2]. Our event detection procedure detects 118 of them, the rest are non-EA events. This indicates that there is a large number of interesting events on Twitter which cannot be explained by earnings announcement. The impact of the EA events on price returns is already known in the literature, and our goal is to reconfirm these results. On the other hand, the impact of the non-EA events is not known, and it is interesting to verify if they have similar impact on prices as the EA events. Therefore, we perform the event study in two scenarios, with explicit detection of the two types of events, all the events (including EA) and non-EA events only:

1. Detecting **all events** from the complete time interval of the data, including the EA days. In total, 260 events are detected, 118 out of these are the EA events.

[2] As obtained from `http://www.zacks.com/`

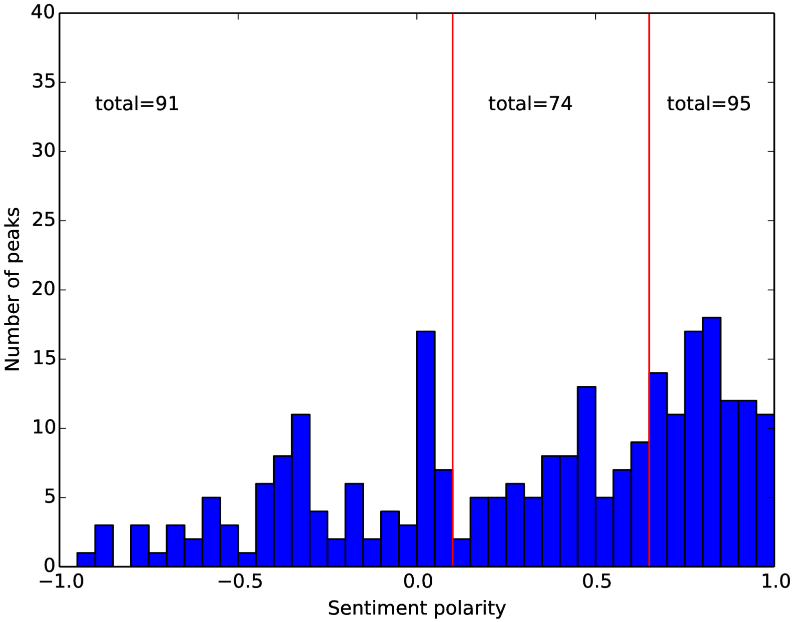

Figure 4.3: Distribution of sentiment polarity for the 260 detected Twitter peaks. The two red bars indicate the chosen thresholds of the polarity values.

2. Detecting **non-EA events** from a subset of the data. For each of the 151 EA events, where d is the event day, we first remove the interval $[d-1, d+1]$, and then perform the event detection again. This results in 182 non-EA events detected.

The first scenario allows to compare the results of the sentiment analysis with the existing literature [31]. It is worth noting, however, that the variable used to infer ″polarity″ of the events there is the difference between the expected and announced earnings. The analysis of the non-EA events in the second scenario tests if the Twitter sentiment data contains useful information about the behavior of investors for other types of events, in addition to the already well-known EA events.

Estimation of normal returns. Here we briefly explain the market model procedure for estimation of normal returns. Our methodology follows the one presented in [31] and [33]. The market model is a statistical model which relates the return of a given stock to the return of the market portfolio. The model's linear specification follows from the assumed joint normality of stock returns. We use the DJIA index as a normal market model. This choice helps us avoid adding too many variables to our model and simplifies the computation of the result. The aggregated DJIA index is computed from the mean weighted prices of all the stocks in the index. For any stock i, and date d, the market model is:

$$\begin{aligned} R_{i,d} &= \alpha_i + \beta_i R_{DJIA,d} + \epsilon_{i,d} & (4.4) \\ E(\epsilon_{i,d}) &= 0, \;\; var(\epsilon_{i,d}) = \sigma^2_{\epsilon_{i,d}} & (4.5) \\ E[R_{i,d}] &= \hat{\alpha_i} + \hat{\beta_i} R_{DJIA,d} & (4.6) \end{aligned}$$

where $R_{i,d}$ and $R_{DJIA,d}$ are the returns of stock i and the market portfolio, respectively, and $\epsilon_{i,d}$ is the zero mean disturbance term. $\alpha_i, \beta_i, \sigma^2_{\epsilon_{i,d}}$ are the parameters of the market model. To estimate these parameters for a given event and stock, we use an estimation window of $L = 120$ days, according to the hint provided in [31]. Using the notation presented in Figure 4.1 for the time line, the estimated value of $\sigma^2_{\epsilon_{i,d}}$ is:

$$\hat{\sigma}^2_{\epsilon_{i,d}} = \frac{1}{L-2} \sum_{d=T_0+1}^{T_1} (R_{i,d} - \hat{\alpha_i} - \hat{\beta_i} R_{DJIA,d})^2 \tag{4.7}$$

where $\hat{\alpha_i}, \hat{\beta_i}$ are the estimated parameters following the OLS procedure[31]. The abnormal return for company i at day d is the residual :

$$AR_{i,d} = R_{i,d} - \hat{\alpha_i} - \hat{\beta_i} R_{DJIA,d}. \tag{4.8}$$

Statistical validation. Our null hypothesis, H_0, is that external events have no impact on the behavior of returns (mean or variance). The distributional properties of the abnormal returns can be used to draw inferences over any period within the event window. Under H_0, the distribution of the sample abnormal return of a given observation in the event window is normal:

$$AR_{i,\tau} \sim \mathcal{N}(0, \sigma^2_{AR}) \tag{4.9}$$

Equation 4.9 takes into account the aggregation of the abnormal returns.

The abnormal return observations must be aggregated in order to draw overall conclusions for the events of interest. The aggregation is along two dimensions: through time and across stocks. By aggregating across all the stocks [33], we get:

$$\overline{AR}_\tau = (1/N) \sum_{i=1}^{N} AR_{i,\tau} \tag{4.10}$$

The cumulative abnormal return (CAR) from time τ_1 to τ_2 is the sum of the abnormal returns:

$$CAR(\tau_1, \tau_2) = \sum_{\tau=\tau_1}^{\tau_2} \overline{AR}_\tau \tag{4.11}$$

To calculate the variance of the CAR, we assume $\sigma^2_{AR} = \sigma^2_{\epsilon_{i,t}}$ (shown in e.g., [31, 33]):

$$var(CAR(\tau_1, \tau_2)) = (1/N^2) \sum_{i=1}^{N} (\tau_2 - \tau_1 + 1)\sigma^2_{\epsilon_i} \tag{4.12}$$

where N is the total number of events. Finally, we introduce the test statistic θ. With this quantity we can test if the measured return is abnormal:

$$\frac{CAR(\tau_1, \tau_2)}{\sqrt[2]{var(CAR(\tau_1, \tau_2))}} = \theta \sim \mathcal{N}(0, 1) \tag{4.13}$$

where τ is the time index inside the event window, and $|\tau_2 - \tau_1|$ is the total length of the event window.

Results

Correlation and Granger causality

Correlation. Table 4.2 shows the computed Pearson correlations, as defined in the Methods section. The computed coefficients are small, but

Ticker	Pearson correlation	Granger causality			
	$\rho(P_d, R_d)$	P_d & R_d		TW_d & $\lvert R_d \rvert$	
TRV	0.1178	←	→		→
UNH	0.2565	←			
UTX	0.1370	←			
MMM	0.1426	←	→	←	
DD	0.2680	←			
AXP	0.1566	←	→		→
PG	0.2145				
NKE	0.2460				
CVX	0.2053				
HD	0.2968	←			→
CAT	0.3648				
JNJ	0.2220				
V	0.2995	←			
VZ	0.1775				
KO	0.1203				
MCD	0.2047				→
XOM	0.2738				
DIS	0.2305	←			→
BA	0.2408				→
MRK	0.1758				
CSCO	0.2393		→		→
GE	0.1450				
WMT	0.2710		→		
INTC	0.2703				→
PFE	0.1252				
T	0.1409				→
GS	0.3405				
IBM	0.3462		→		→
JPM	0.1656	←			
MSFT	0.2700				→
		10	3	2	10

Table 4.2: Results of the Pearson correlation and Granger causality tests. Companies are ordered as in Table 4.1. The arrows indicate a statistically significant Granger causality relation for a company, at the 5% level. A right arrow indicates that the Twitter variable (sentiment polarity P_d or volume TW_d) Granger-causes the market variable (return R_d), while a left arrow indicates that the market variable Granger-causes the Twitter variable. The counts at the bottom show the total number of companies passing the Granger test.

are in line with the result of [13]. In our opinion, these findings and the one published in [13] underline that when considering the entire time period of the analysis, days with a low number of tweets affect the measure.

Granger causality. The results of the Granger causality tests are also in Table 4.2. They show the results of the causality test in both directions: from the Twitter variables to the market variables and vice-versa. The table gives the Granger causality links per company between a) sentiment polarity and price return, and b) the volume of tweets and absolute price return. The conclusions that can be drawn are:

- The polarity variable is not useful for predicting the price return, as only three companies pass the Granger test.
- The number of tweets for a company Granger-causes the absolute price return for one third of the companies. This indicates that the amount of attention on Twitter is useful for predicting the price volatility. Previously, this was known only for an aggregated index, but not for individual stocks [13, 29].

Cumulative abnormal returns

The results of the event study are shown in Figures 4.4 and 4.5, where the cumulative abnormal returns (CAR) are plotted for the events defined earlier. The results are largely consistent with the existing literature on the information content of earnings [31, 32]. The evidence strongly supports the hypothesis that tweets do indeed convey information relevant for stock returns.

Figure 4.4 shows CAR for all the detected Twitter peaks, including the EA events (45% of the detected events are earnings announcements). The average CAR for the events is abnormally increasing after the positive peaks and decreasing after the negative sentiment peaks. This is confirmed with details in Table 4.3. The value of θ remains above two standard deviations for ten days after the positive sentiment events. Given this result, the null hypothesis that the event has no impact on price returns is rejected. The same holds for negative sentiment events. However, as expected, the abnormal return after the neutral events is not significant and in this case one cannot reject the null hypothesis.

A more interesting result concerns the non-EA events in Figure 4.5. Even after removing the earnings announcements, with already known impact on price returns, one can reject the null hypothesis. In this case, the average CAR of the non-EA events is abnormally increasing after the detected positive peaks and decreasing after the negative peaks. Table 4.3 shows that after the event days the value of θ remains above two

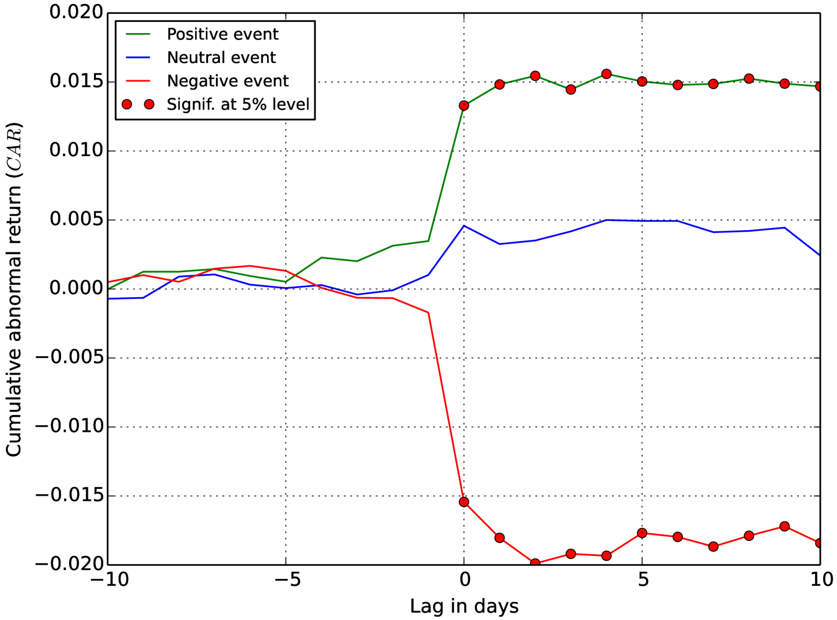

Figure 4.4: CAR for all detected events, including EA. The x axis is the lag between the event and CAR, and the red markers indicate days with statistically significant abnormal return.

standard deviations for 5 days after the positive events, and for 10 days after the negative events. The period of impact of Twitter sentiment on price returns is shorter when the EA events are removed, and the CARs are lower, but in both cases the impact is statistically significant.

Discussion

In this work we present significant evidence of dependence between stock price returns and Twitter sentiment in tweets about the same companies. As a series of other papers have already shown, there is a signal worth investigating which connects social media and market behaviour. This opens the way, if not to forecasting, then at least to ”now-casting”

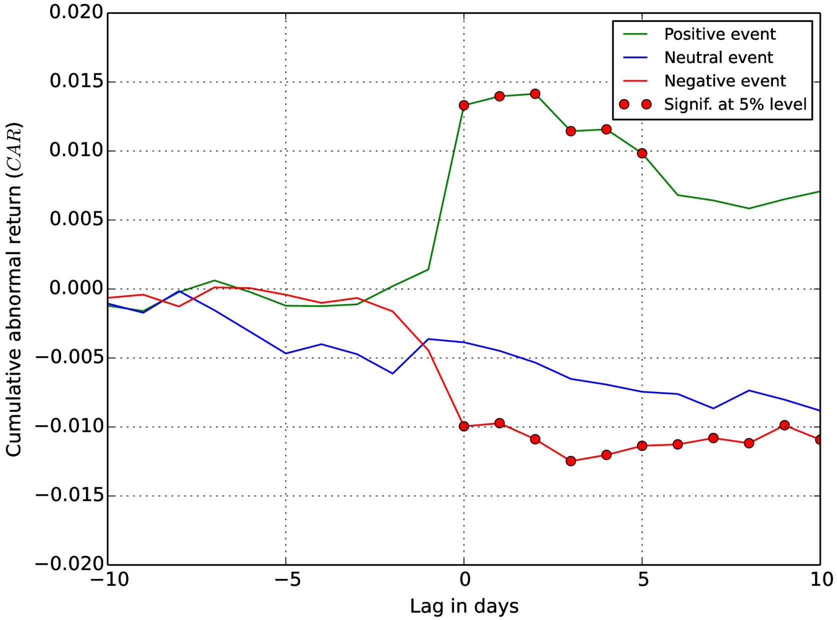

Figure 4.5: CAR for non-EA events. The x axis is the lag between the event and CAR, and the red markers indicate days with statistically significant abnormal return.

financial markets. The caveat is that this dependence becomes useful only when data are properly selected, or different sources of data are analyzed together. For this reason, in this paper, we first identify events, marked by increased activity of Twitter users, and then observe market behaviour in the days following the events. This choice is due to our hypothesis that only at some moments, identified as events, there is a strong interaction between the financial market and Twitter sentiment. Our main result is that the aggregate Twitter sentiment during the events implies the direction of market evolution. While this can be expected for peaks related to ”known” events, like earnings announcements, it is really interesting to note that a similar conclusion holds also when peaks do not correspond to any expected news about the stock traded.

Studies as this one could be well used in order to establish a direct relation between social networks and market behaviour. A specific application could, for example, detect and possibly mitigate panic diffusion in the market from social network analysis. To such purpose there is some additional research to be done in the future. For one, detection of Twitter events should rely just on the current and past Twitter volume, in order to be applicable for real-time monitoring. It might be worth characterizing the topics of Twitter events, specially the non-EA events, by applying text analysis tools, and investigating their impact on price movements. Also, during the events, we might move to a finer time scale, e.g., from daily to hourly resolution. Finally, our short term plan is to extend the analysis to a larger number of companies with high Twitter volume, and over longer period of time.

Table 4.3: Values of the θ statistic for each type of event. The bold numbers indicate significant values at the 5% level, i.e., $|\theta| > 2$.

Lag	All events (including EA)			Non-EA events		
(days)	negative	neutral	positive	negative	neutral	positive
-10	0.649887	-0.803961	-0.017561	-0.729825	-0.896790	-1.319828
-9	0.924240	-0.516087	1.130406	-0.334372	-1.033670	-1.226448
-8	0.391043	0.590919	0.923003	-0.838579	-0.071650	-0.140536
-7	0.955298	0.599959	0.921514	0.068941	-0.656765	0.336994
-6	0.971351	0.160289	0.536642	0.031506	-1.178662	-0.108518
-5	0.696531	0.029952	0.273652	-0.194874	-1.621344	-0.537235
-4	0.040918	0.123647	1.096456	-0.436805	-1.286092	-0.510411
-3	-0.291666	-0.161277	0.907073	-0.263249	-1.420006	-0.427799
-2	-0.286039	-0.034943	1.330531	-0.622001	-1.736869	0.075007
-1	-0.699695	0.366674	1.399473	-1.614968	-0.974915	0.486124
0	**-6.034968**	1.574068	**5.107868**	**-3.448207**	-0.990421	**4.364840**
1	**-6.753299**	1.071353	**5.452248**	**-3.226228**	-1.099833	**4.384798**
2	**-7.156974**	1.109907	**5.458699**	**-3.470841**	-1.257358	**4.266317**
3	**-6.653879**	1.271347	**4.922402**	**-3.830621**	-1.480477	**3.324482**
4	**-6.474863**	1.470405	**5.127602**	**-3.567463**	-1.520210	**3.249010**
5	**-5.733431**	1.404257	**4.791116**	**-3.264718**	-1.583226	**2.673873**
6	**-5.652182**	1.361331	**4.569458**	**-3.138621**	-1.568814	1.795027
7	**-5.707076**	1.105312	**4.463371**	**-2.925752**	-1.734561	1.644837
8	**-5.321658**	1.099470	**4.456279**	**-2.946521**	-1.434466	1.455898
9	**-4.988644**	1.131591	**4.239979**	**-2.537126**	-1.525630	1.582168
10	**-5.212109**	0.601714	**4.081477**	**-2.739847**	-1.636512	1.679536

Appendix A

Materials and Methods for Chapter 2

A.1 Financial time series

We consider only trading hours (i.e. from 9:30 AM to 4:00 PM) and trading days (i.e. business days at the New York Stock Exchange), and we disregard the trading events that occur out of this time window. From the financial data we create three time series: the first one is the traded volume for each minute for each company, the second one is the logarithmic returns, and the last time series is the absolute value of the logarithmic returns.

1. **Trade volume (V) time series**. It consists of the traded volume of a given company at minute time scale. We build a time series at time scale t by summing the traded volumes v_τ at the smallest time scale τ (in our case one minute). Then, the total volume $\overline{v_t}$ can be defined as follows:

   $$\overline{v_t} = \sum_{\tau > t-1}^{t} v_\tau \, . \tag{A.1}$$

   Traded volumes at intra-day time t are rescaled by a factor ζ_t^v, which is computed as the average, over all days, of the volume at time t

normalized by the total daily volume. If $\overline{v_{d,t}}$ is the raw volume of day d and intra-day time t, we define the rescaled time series as

$$V_{d,t} = \frac{\overline{v_{d,t}}}{\zeta_t^v}\,, \tag{A.2}$$

where

$$\zeta_t^v = \frac{1}{T}\sum_{d'} \frac{\overline{v_{d',t}}}{\Lambda_{d'}}\,, \tag{A.3}$$

and

$$\Lambda_d = \sum_t \overline{v_{d,t}}\,, \tag{A.4}$$

i.e. the total volume trades on day d, and T is the total number of trading days.

2. **Price return (R) time series**. We preliminary compute the logarithmic return, defined as

$$r_t = \log_{10}(\frac{p_t}{p_{t-1}})\,, \tag{A.5}$$

where p_t is the last recorded price in the interval $(t-1, t]$. Returns at intra-day time t are rescaled by a factor ζ_t^r, which is computed as the average, over all days, of absolute returns at time t rescaled by the average volatility. More precisely, if $r_{d,t}$ is the raw return of day d and intra-day time t, we define the rescaled time series as

$$R_{d,t} = \frac{r_{d,t}}{\zeta_t^r}\,, \tag{A.6}$$

where

$$\zeta_t^r = \frac{1}{T}\sum_{d'} \frac{|r_{d',t}|}{\Xi_{d'}}\,, \tag{A.7}$$

and

$$\Xi_d = \text{mean}(|r_{d,t}|)\,, \tag{A.8}$$

where we compute the mean value for all t in a day d.

3. **Volatility (σ) time series**. We employ as a simple proxy for the de-

seasonalized intra-day volatility the absolute value of the rescaled return

$$\sigma_t = |R_t| \,. \tag{A.9}$$

Notice that the volatility σ_t is already de-seasonalized, because of the definition of R

A.2 Click and sentiment time series

Consistently with the financial time series, we consider only trading time and trading days. This implies that we neglect the clicks that occur out of this time window, because we are only interested in the click behaviour during the trading hours evolution of markets. From the click-through history of each news we create two time series: the first one is the total amount of clicks for each minute for each company, and the second one is the number of clicks multiplied by the sentiment score of the associated news.

1. **Click (C) time series**. Starting from the number of clicks of each news at the smallest time scale τ (in our case one minute), we build a time series at the time scale t by aggregating the number of clicks of all the news of a given company. So if we denote by N the number of news of a company, and by c^i_τ the number of clicks for news i at scale τ, the total volume $\overline{c}_t$ can be defined as

$$\overline{c_t} = \sum_{\tau > t-1}^{t} \sum_{i=1}^{N} c^i_\tau \,. \tag{A.10}$$

The news that are not viewed in the time interval have zero clicks. We filter out the intra-day pattern from clicks by means of a simple methodology. Clicks at intra-day time t are rescaled by a factor ζ^c_t, which is computed as the average, over all days, of the click volume at time t normalized by the total number of daily clicks. More precisely, if $\overline{c_{d,t}}$ is the raw click volume of day d and intra-day time t, we define the time series of rescaled clicks as:

$$C_{d,t} = \frac{\overline{c}_{d,t}}{\zeta_t^c} \,, \tag{A.11}$$

where

$$\zeta_t^c = \frac{1}{T} \sum_{d'} \frac{\overline{c}_{d',t}}{\Gamma_{d'}} \,, \tag{A.12}$$

and

$$\Gamma_d = \sum_t \overline{c}_{d,t} \,, \tag{A.13}$$

with Γ_d the total number of clicks in a day.

2. **Sentiment (S) time series**. To construct this time series we consider the sentiment of the headlines of the news. With the same notation previously used

$$S_t = \sum_{\tau > t-1}^{t} \sum_{i=1}^{N} s_\tau^i \,, \tag{A.14}$$

where s_τ^i is the sign $(-1, 0, 1)$ of the sentiment of a news headline published at time τ.

3. **Weighted Sentiment (WS) time series**. We multiply the click count of each news by its sentiment score. With the same notation of the click time series, we have:

$$\overline{WS}_t = \sum_{\tau > t-1}^{t} \sum_{i=1}^{N} c_\tau^i s_\tau^i \,. \tag{A.15}$$

This way, we weight the sign of every news by the level of attention that the news has received. We remark that for each trading day we consider all the clicks of all the news clicked in that day, not only the news that have been released in that day. Then, we define the weighted sentiment as

$$WS_t = \mathrm{sign}(\overline{WS}_t) C_t \,. \tag{A.16}$$

A.3 Estimation of power law exponents

We estimate the power law tail exponent of the distribution of the number of clicks per news by employing the R package PoweRlaw developed and maintained by Colin Gillespie, and described in [58].

Table A.1: Tail exponent α and lower bound $x_{\min} > 0$ with standard errors ϵ_α and $\epsilon_{x_{\min}}$ estimated from 100 highly capitalized stocks traded in the US equity markets. Following the procedure detailed in [**?**], we estimate the probability distribution associated with integer numbers larger than $x_{\min}$ whose expression reads to $p(x) = x^{-1-\alpha}/\zeta(1+\alpha, x_{\min})$. The normalizing constant corresponds to the Hurwitz zeta function ζ. We report in bold the top 10 assets, as measured by the absolute number of associated news.

	α	ϵ_{α}	$x_{\min}$	$\epsilon_{x_{\min}}$		α	ϵ_{α}	$x_{\min}$	$\epsilon_{x_{\min}}$
AA	1.05	0.09	482	114	**JPM**	**0.99**	**0.04**	**665**	**205**
AAPL	**0.97**	**0.05**	**2881**	**1451**	KBH	1.33	0.20	402	98
ABT	1.39	0.33	530	261	KO	0.96	0.09	565	242
ACN	1.84	0.37	1074	313	LCC	0.95	0.08	722	371
AET	1.21	0.13	261	54	LEN	0.86	0.19	256	197
AIG	1.16	0.13	854	193	LMT	1.41	0.14	808	159
AMGN	1.60	0.33	1596	445	LNKD	0.90	0.13	735	297
AMZN	1.06	0.05	969	206	LUV	0.82	0.10	617	282
ATVI	1.47	0.23	863	192	MA	0.97	0.13	335	213
AXP	0.89	0.07	447	93	MAR	1.34	0.17	318	90
BAC	**0.95**	**0.05**	**807**	**200**	MCD	0.71	0.06	462	305
BA	**1.01**	**0.05**	**802**	**250**	M	0.86	0.09	384	200
BBY	0.74	0.13	957	526	MRK	1.53	0.20	966	213
BIIB	1.88	0.41	1081	251	MSCI	1.12	0.09	205	62
BLK	1.20	0.12	547	122	**MS**	**1.15**	**0.09**	**615**	**457**
CAT	1.14	0.11	834	224	**MSFT**	**1.06**	**0.05**	**2247**	**507**
CBS	0.82	0.07	169	75	NDAQ	1.24	0.08	213	42
C	**1.10**	**0.06**	**627**	**256**	NFLX	0.94	0.07	613	141
CHK	1.72	0.18	1479	215	NKE	1.09	0.14	566	143
CMCSA	1.41	0.24	1186	357	NOC	1.46	0.22	894	228
COP	1.76	0.24	1708	303	NWSA	0.89	0.06	234	117
COST	0.79	0.10	448	372	NYX	0.98	0.18	176	192
CSCO	1.32	0.11	1290	271	ORCL	0.95	0.20	523	422
CVX	1.37	0.15	782	180	PEP	0.92	0.13	565	220
DAL	1.01	0.12	959	220	PFE	1.54	0.14	1306	250
DELL	1.00	0.07	944	246	PG	1.02	0.11	809	327
DIS	0.72	0.04	283	198	QCOM	1.44	0.23	1748	363
DISH	1.05	0.17	515	391	RTN	1.83	0.47	1379	425
DOW	1.21	0.13	818	313	SBUX	0.80	0.07	453	236
DTV	1.11	0.13	438	132	S	1.09	0.14	684	424
EBAY	0.91	0.11	1612	459	TRI	0.89	0.04	122	62
FB	**0.95**	**0.08**	**1211**	**1077**	TSLA	0.98	0.11	3024	850
F	0.96	0.05	1188	254	TWC	1.06	0.09	403	176
FRT	1.19	0.08	412	87	TWX	0.84	0.08	260	119
GE	0.98	0.08	587	236	TXN	1.55	0.42	1060	475
GILD	1.87	0.43	1300	294	UAL	1.00	0.12	1153	659
GM	0.84	0.05	863	296	UNH	1.21	0.15	378	147
GOOG	**0.92**	**0.03**	**1786**	**450**	UPS	1.93	0.49	2499	721
GPS	1.28	0.17	390	158	UTX	1.49	0.20	733	209
GRPN	0.93	0.13	644	193	V	1.17	0.13	505	243
GS	**0.89**	**0.05**	**462**	**143**	VMW	1.50	0.45	1003	547
HAL	1.56	0.18	880	213	VOD	1.87	0.42	1850	711
HD	0.79	0.07	456	169	VZ	1.11	0.10	1001	359
HON	1.18	0.17	632	296	WAG	1.06	0.23	512	322
HPQ	1.13	0.11	786	274	WFC	0.95	0.05	874	271
IBM	1.11	0.11	1014	283	WMT	0.67	0.06	2073	517
INTC	1.35	0.32	1777	738	XOM	1.29	0.12	802	147
IT	1.32	0.13	410	98	YHOO	1.15	0.12	2546	605
JCP	0.75	0.07	2827	959	YUM	0.90	0.10	483	225
JNJ	1.51	0.19	1058	174	ZNGA	1.38	0.16	1578	338

References

[1] King, G. Ensuring the data-rich future of the social sciences. *Science* **331**, 719–721 (2011). 1, 5

[2] Vespignani, A. Predicting the behavior of techno-social systems. *Science* **325**, 425–428 (2009). 1, 5

[3] Bonanno, G. *et al.* Networks of equities in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems* **38**, 363–371 (2004). 1, 5

[4] Caldarelli, G. *Scale-Free Networks: complex webs in nature and technology* (Oxford University Press, Oxford, 2007). 1, 5

[5] Tumminello, M., Aste, T., Di Matteo, T. & Mantegna, R. N. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10421–10426 (2005). 1, 6

[6] Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H. & Liu, B. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, 702–707 (IEEE, 2011). 1, 6

[7] Tizzoni, M. *et al.* Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC medicine* **10**, 165 (2012). 1, 6

[8] Caldarelli, G. *et al.* A multi-level geographical study of Italian political elections from Twitter data. *PloS one* **9**, e95809 (2014). 1, 6

[9] Bouchaud, J.-P. The (unfortunate) complexity of the economy. *arXiv preprint arXiv:0904.0805* (2009). 1

[10] Haldane, A. G. & May, R. M. Systemic risk in banking ecosystems. *Nature* **469**, 351–355 (2011). 2

[11] FAGIOLO, G. *et al.* Economic networks: The new challenges. *science* **325**, 422–425 (2009). 2

[12] Bouchaud, J.-P. Economics needs a scientific revolution. *Nature* **455**, 1181–1181 (2008). 2

[13] Mao, H., Counts, S. & Bollen, J. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051* (2011). 2, 9, 11, 14, 24, 27, 35, 47, 48

[14] Bordino, I. *et al.* Web search queries can predict stock market volumes. *PloS One* **7**, e40014 (2012). 2, 10, 11

[15] Curme, C., Preis, T., Stanley, H. E. & Moat, H. S. Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences* **111**, 11600–11605 (2014). 2, 3, 11

[16] Cutler, D. M., Poterba, J. M. & Summers, L. H. What moves stock prices? *Journal of Portfolio Management* **15**, 4–12 (1989). 2, 11

[17] Chan, W. S. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics* **70**, 223–260 (2003). 2, 9, 11

[18] Vega, C. Stock price reaction to public and private information. *Journal of Financial Economics* **82**, 103–133 (2006). 2, 11

[19] Tetlock, P. C. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance* **62**, 1139–1168 (2007). 2, 9, 11

[20] Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance* **63**, 1437–1467 (2008). 2, 11

[21] Lillo, F., Miccichè, S., Tumminello, M., Piilo, J. & Mantegna, R. N. How news affect the trading behavior of different categories of investors in a financial market. *Quantitative Finance* (2014). DOI: 10.1080/14697688.2014.931593. 2, 9, 10, 11

[22] Engelberg, J. E., Reed, A. V. & Ringgenberg, M. C. How are shorts informed? Short sellers, news, and information processing. *Journal of Financial Economics* **105**, 260–278 (2012). 2, 11

[23] Birz, G. & Lott Jr, J. R. The effect of macroeconomic news on stock returns:

New evidence from newspaper coverage. *Journal of Banking & Finance* **35**, 2791–2800 (2011). 2, 11

[24] Gross-Klussmann, A. & Hautsch, N. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance* **18**, 321–340 (2011). 2, 11

[25] Preis, T., Reith, D. & Stanley, H. E. Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **368**, 5707–5719 (2010). 2, 10

[26] Bordino, I., Kourtellis, N., Laptev, N. & Billawala, Y. Stock trade volume prediction with yahoo finance user browsing behavior. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, 1168–1173 (2014). 2, 9, 10

[27] Da, Z., Engelberg, J. & Gao, P. In search of attention. *The Journal of Finance* **66**, 1461–1499 (2011). 2, 11

[28] Mao, Y., Wei, W., Wang, B. & Liu, B. Correlating S&P 500 stocks with Twitter data. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, 69–72 (ACM, 2012). 2, 11, 24, 27

[29] Bollen, J., Mao, H. & Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science* **2**, 1–8 (2011). 2, 11, 24, 35, 48

[30] Bollen, J., Pepe, A. & Mao, H. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 450–453 (2011). 2, 11, 24

[31] Campbell, J. Y., Lo, A. W., MacKinlay, A. C. & Whitelaw, R. F. The econometrics of financial markets. *Macroeconomic Dynamics* **2**, 559–562 (1998). 4, 35, 36, 41, 43, 44, 45, 46, 48

[32] Boehmer, E., Masumeci, J. & Poulsen, A. B. Event-study methodology under conditions of event-induced variance. *Journal of Financial Economics* **30**, 253–272 (1991). 4, 35, 36, 43, 48

[33] Malkiel, B. G. & Fama, E. F. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* **25**, 383–417 (1970). 6, 45, 46

[34] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. & Kappas, A. Sentiment

strength detection in short informal text. *Journal of the American Society for Information Science and Technology* **61**, 2544–2558 (2010). 9

[35] Loughran, T. & McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* **66**, 35–65 (2011). 9

[36] Bank, M., Larch, M. & Peter, G. Google search volume and its influence on liquidity and returns of german stocks. *Financial markets and portfolio management* **25**, 239–264 (2011). 10

[37] Kristoufek, L. Can google trends search queries contribute to risk diversification? *Scientific Reports* **3** (2013). 10

[38] Vlastakis, N. & Markellos, R. N. Information demand and stock market volatility. *Journal of Banking & Finance* **36**, 1808–1821 (2012). 10

[39] Schumaker, R. P. & Chen, H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)* **27**, 12 (2009). 11

[40] Alanyali, M., Moat, H. S. & Preis, T. Quantifying the relationship between financial news and the stock market. *Scientific Reports* **3** (2013). 11

[41] Zhang, Y. *et al.* Internet information arrival and volatility of sme price index. *Physica A: Statistical Mechanics and its Applications* **399**, 70–74 (2014). 11

[42] De Choudhury, M., Sundaram, H., John, A. & Seligmann, D. D. Can blog communication dynamics be correlated with stock market activity? In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, 55–60 (ACM, 2008). 11

[43] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A. & Jaimes, A. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 513–522 (ACM, 2012). 11, 24

[44] Zhang, X., Fuehres, H. & Gloor, P. A. Predicting stock market indicators through Twitter i hope it is not as bad as i fear. *Procedia-Social and Behavioral Sciences* **26**, 55–62 (2011). 11

[45] Preis, T., Moat, H. S., Stanley, H. E. & Bishop, S. R. Quantifying the advantage of looking forward. *Scientific Reports* **2** (2012). 11

[46] Preis, T., Moat, H. S. & Stanley, H. E. Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports* **3** (2013). 11

[47] Moat, H. S. *et al.* Quantifying wikipedia usage patterns before stock market moves. *Scientific Reports* **3** (2013). 11

[48] Granger, C. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969). 12

[49] Zipf, G. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA, 1949). 13

[50] Bormetti, G. *et al.* Modelling systemic price cojumps with Hawkes factor models. *arXiv preprint arXiv:1301.6141* (2013). 17

[51] Zheludev, I., Smith, R. & Aste, T. When can social media lead financial markets? *Scientific reports* **4** (2014). 35

[52] Joachims, T. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms* (Kluwer/Springer, 2002). 38

[53] Zhang, W. & Skiena, S. Trading strategies to exploit blog and news sentiment. In *ICWSM* (2010). 38

[54] Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 424–438 (1969). 39

[55] Piškorec, M. *et al.* Cohesiveness in financial news and its relation to market volatility. *Scientific reports* **4** (2014). 39

[56] MacKinlay, A. C. Event studies in economics and finance. *Journal of economic literature* 13–39 (1997). 39

[57] Lehmann, J., Gonçalves, B., Ramasco, J. J. & Cattuto, C. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, 251–260 (ACM, 2012). 41

[58] Clauset, A., Shalizi, C. R. & Newman, M. E. Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009). 57