The Effects of Twitter Sentiment on Stock Price Returns

Gabriele Ranco[1], Darko Aleksovski[2]*, Guido Caldarelli[1,3,4], Miha Grčar[2], Igor Mozetič[2]

1 IMT Institute for Advanced Studies, Piazza San Francesco 19, 55100 Lucca, Italy, **2** Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia, **3** Istituto dei Sistemi Complessi (ISC), Via dei Taurini 19, 00185 Rome, Italy, **4** London Institute for Mathematical Sciences, 35a South St. Mayfair, London W1K 2XF, United Kingdom

* darko.aleksovski@ijs.si

Abstract

Social media are increasingly reflecting and influencing behavior of other complex systems. In this paper we investigate the relations between a well-known micro-blogging platform Twitter and financial markets. In particular, we consider, in a period of 15 months, the Twitter volume and sentiment about the 30 stock companies that form the Dow Jones Industrial Average (DJIA) index. We find a relatively low Pearson correlation and Granger causality between the corresponding time series over the entire time period. However, we find a significant dependence between the Twitter sentiment and abnormal returns during the peaks of Twitter volume. This is valid not only for the expected Twitter volume peaks (e.g., quarterly announcements), but also for peaks corresponding to less obvious events. We formalize the procedure by adapting the well-known "event study" from economics and finance to the analysis of Twitter data. The procedure allows to automatically identify events as Twitter volume peaks, to compute the prevailing sentiment (positive or negative) expressed in tweets at these peaks, and finally to apply the "event study" methodology to relate them to stock returns. We show that sentiment polarity of Twitter peaks implies the direction of cumulative abnormal returns. The amount of cumulative abnormal returns is relatively low (about 1–2%), but the dependence is statistically significant for several days after the events.

Introduction

The recent technological revolution with widespread presence of computers and Internet has created an unprecedented situation of data deluge, changing dramatically the way in which we look at social and economic sciences. The constantly increasing use of the Internet as a source of information, such as business or political news, triggered an analogous increasing online activity. The interaction with technological systems is generating massive datasets that document collective behavior in a previously unimaginable fashion [1, 2]. Ultimately, in this vast repository of Internet activity we can find the interests, concerns, and intentions of the global population with respect to various economic, political, and cultural phenomena.

Among the many fields of applications of data collection, analysis and modeling, we present here a case study on financial systems. We believe that social aspects as measured by social networks are particularly useful to understand financial turnovers. Indeed, 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 [3]. Therefore, the possibility to anticipate anomalous collective behavior of investors is of great interest to policy makers [4–6] because it may allow for a more prompt intervention, when appropriate.

State-of-the-art. We briefly review the state-of-the-art research which investigates the correlation between the web data and financial markets. Three major classes of data are considered: web news, search engine queries, and social media. Regarding news, various approaches have been attempted. They study: (i) the connection of exogenous news with price movements [7], (ii) the stock price reaction to news [8, 9]; (iii) the relations between mentions of a company in financial news [10], or the pessimism of the media [11], and trading volume; (iv) the relation between the sentiment of news, earnings and return predictability [12], (v) the role of news in trading actions [13], especially of short sellers [14]; (vi) the role of macroeconomic news in stock returns [15]; and finally (vii) the high-frequency market reactions to news [16].

There are several analyses of search engine queries. A relation between the daily number of queries for a particular stock, and daily trading volume of the same stock has been studied by [17–19]. A similar analysis was done for a sample of Russell 3000 stocks, where an increase in queries predicts higher stock prices in the next two weeks [20]. Search engine query data from Google Trends have been used to evaluate stock riskiness [21]. Some other authors used Google trends to predict market movements [22]. Also, search engine query data have been used as a proxy for analyzing investor attention related to initial public offerings (IPOs) [23].

Regarding social media, Twitter is becoming an increasingly popular micro-blogging platform used for financial forecasting [24–26]. One line of research investigates the relation between the volume of tweets and financial markets. For example, [27] studied whether the daily number of tweets predicts the S&P 500 stock indicators. Another line of research explores the contents of tweets. In a textual analysis approach to Twitter data, the authors find clear relations between the mood indicators and Dow Jones Industrial Average (DJIA) [28–30]. In [31], the authors show that the Twitter sentiment for five retail companies has statistically significant relation with stock returns and volatility. A recent study [32] compares the information content of the Twitter sentiment and volume in terms of their influence on future stock prices. The authors relate the intra-day Twitter and price data, at hourly resolution, and show that the Twitter sentiment contains significantly more lead-time information about the prices than the Twitter volume alone. They apply stringent statistics which require relatively high volume of tweets over the entire period of three months, and, as a consequence, only 12 financial instruments pass the test.

Motivation. Despite the high quality of the datasets used, the level of empirical correlation between stock price derived financial time series and web derived time series remains limited, especially when a textual analysis of web messages is applied. This observation suggests that the relation between these two systems is more complex and that a simple measure of correlation is not enough to capture the dynamics of the interaction between the two systems. It is possible that the two systems are dependent only at some moments of their evolution, and not over the entire time period.

In this paper, we study the relation between stock price returns and the sentiment expressed in financial tweets posted on Twitter. We analyze a carefully collected and annotated set of tweets about the previously-mentioned 30 DJIA companies. For each of these companies we build a time series of the sentiment expressed in the tweets, with daily resolution, designed to

mimic the wisdom-of-crowd effect, as observed in previous works. As first analysis we compute the Pearson correlation between price return time series and the sentiment time series generated from the tweets. We also run a Granger causality test [33] to study the forecasting power of the Twitter time series. When considering the entire period of 15 months, the values of Pearson correlation are low and only a few companies pass the Granger causality test.

In order to detect the presence of a stronger correlation, at least in some portions of the time series, we consider the relation between the stock price returns and Twitter sentiment through the technique of "event study" [34, 35], known in economics and finance. This technique has been generally used to verify if the sentiment content of earnings announcements conveys useful information for the valuation of companies. Here we apply a similar approach, but instead of using the sentiment of earnings announcement, we use the aggregate sentiment expressed in financial tweets.

Contributions. By restricting our analysis to shorter time periods around the "events" we find a statistically significant relation between the Twitter sentiment and stock returns. These results are consistent with the existing literature on the information content of earnings [34, 35]. A recent related study [36, 37], also applies the "event study" methodology to Twitter data. The authors come to similar conclusions as we do: financial Twitter data, when considering both the volume and sentiment of tweets, do have a statistically significant impact on stock returns. It is interesting that two independent studies, to the best of our knowledge the first adaptations of "event study" to Twitter data, corroborate the conclusions.

This paper presents a complementary study to [36], and uses a slightly different experimental setup. The studies use disjoint sets of stocks (S&P 500 vs. DJIA 30), non-overlapping time windows (January–June 2010 vs. June 2013–September 2014), different sentiment classification techniques (Naive Bayes vs. Support Vector Machine), different event detection algorithms, and different statistics for significance testing. We point out the differences between the two studies in the appropriate sections of the paper. Despite the methodological differences, we can confirm the main results reported in [36] to a large extent. From this perspective, one of the contributions of this work is in providing even more evidence, over a longer time period, for the conclusions drawn in both studies.

The second contribution is that the Twitter sentiment time series are made publicly available. They can be used not only to validate our results but also to carry out additional studies that do not necessarily follow the same methodology. The dataset allows one to study different sentiment aggregations, different events (points in time), and different post-event effects (such as drifts, reversals, and changes in volatility rather than abnormal returns).

The third contribution, as compared to [36], is the use of a high quality sentiment classifier, and the realistic evaluation of its performance. Our sentiment classifier was trained on a much larger training set (2,500 vs. over 100,000 annotated tweets in our case), and exhaustively evaluated. This resulted in the performance that matches the agreement between financial experts. The human annotation of such large number of tweets is relatively expensive. However, there are several advantages. First, a considerable amount of tweets can be annotated twice, by two different annotators, in order to compute the inter-annotator agreement and thus establish an upper bound on the performance. Second, there is no need to collect domain-specific vocabularies, since the annotation process itself is domain and language specific. Third, once a large enough set of tweets is assigned a sentiment label, the classifier construction is automated and the domain-specific sentiment models are available for real-time processing.

We have already applied the same sentiment classification methodology in various domains, such as: (i) to study the emotional dynamics of Facebook comments on conspiracy theories (in Italian) [38], (ii) to compare the sentiment leaning of different network communities towards

Table 1. The Twitter data for the 15 months period. For each company, there is the DJIA ticker symbol and the number of collected tweets.

Ticker	Company	Tweets
TRV	Travelers Companies Corp	12,184
UNH	UnitedHealth Group Inc	15,020
UTX	United Technologies Corp	16,123
MMM	3M Co	17,001
DD	E I du Pont de Nemours and Co	17,340
AXP	American Express Co	21,941
PG	Procter & Gamble Co	25,751
NKE	Nike Inc	29,220
CVX	Chevron Corp	29,477
HD	Home Depot Inc	30,923
CAT	Caterpillar Inc	38,739
JNJ	Johnson & Johnson	40,503
V	Visa Inc	43,375
VZ	Verizon Communications Inc	45,177
KO	Coca-Cola Co	45,339
MCD	McDonald's Corp	45,971
XOM	Exxon Mobil Corp	46,286
DIS	Walt Disney Co	46,439
BA	Boeing Co	51,799
MRK	Merck & Co Inc	54,986
CSCO	Cisco Systems Inc	57,427
GE	General Electric Co	61,836
WMT	Wal-Mart Stores Inc	63,405
INTC	Intel Corp	68,079
PFE	Pfizer Inc	71,415
T	AT&T Inc	75,886
GS	Goldman Sachs Group Inc	91,057
IBM	International Business Machines Co	101,077
JPM	JPMorgan Chase and Co	108,810
MSFT	Microsoft Corp	183,184
Total		1,555,770

doi:10.1371/journal.pone.0138441.t001

various environmental topics [39], and (iii) to monitor the sentiment about political parties before and after the elections (in Bulgarian) [40].

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 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:

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

where p_d is the closing price of the stock at day d. We use raw-returns, and not the more standard log-returns, to be consistent with the original "event study" [34, 41]. This data are 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/historical for the "Nike" stock).

Twitter data

The second source of data is from Twitter and consists of relevant tweets, along with their sentiment. The data were collected by Twitter Search API, where a search query consists of the stock cash-tag (e.g., "$NKE" for Nike). To the best of our knowledge, all the available tweets with cash-tags are acquired. The Twitter restriction of 1% (or 10%) of tweets applies to the Twitter Streaming API, and only in the case when the specified filter (query) is general enough to account for more than 1% (or 10%) of all public tweets. The data cover a period of 15 months (from June 1, 2013 to September 18, 2014), for which there are over 1.5 million tweets. The tweets for the analysis were provided to us by the Sowa Labs company (http://www.sowalabs.com/).

The Twitter sentiment is calculated by a supervised learning method. First, over 100,000 of tweets were labeled by 10 financial experts with three sentiment labels: negative, neutral or positive. Then, this labeled set was used to build a Support Vector Machine (SVM [42]) classification model which discriminates between negative, neutral and positive tweets. Finally, the SVM model was applied to the complete set of over 1.5 million tweets. The resulting dataset 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 [43], $P_d = \frac{tw_d^+ - tw_d^-}{tw_d^+ + tw_d^-}$.

The Twitter sentiment and financial time series data for the DJIA 30 stocks are available at http://dx.doi.org/10.6084/m9.figshare.1533283.

Methods

This section first describes the machine learning methodology used for sentiment classification. Then, it presents the methods used for the correlation analysis and Granger causality. Finally, it describes the event study methodology, by presenting the detection of events, the categorization of events based on Twitter sentiment, and the statistical validation of the cumulative abnormal returns.

Sentiment classification

Determining sentiment polarity of tweets is not an easy task. Financial experts often disagree whether a given tweet represents a buy or a sell signal, and even individuals are not always consistent with themselves. We argue that the upper bound that any automated sentiment classification procedure can achieve is determined by the level of agreement between the human experts. In order to achieve the performance of human experts, a large enough set of tweets has to be manually annotated—in our case, over 100,000. In order to measure the agreement between the experts, a substantial fraction of tweets has to be annotated by two different experts—in our case, over 6,000 tweets were annotated twice.

Our approach to automatic sentiment classification of tweets is based on supervised machine learning. The procedure consists of the following steps: (i) a sample of tweets is manually annotated with sentiment, (ii) the labeled set is used to train and tune a classifier, (iii) the classifier is evaluated by cross-validation and compared to the inter-annotator agreement, and (iv) the classifier is applied to the whole set of collected tweets. There is a lot of related work on the automatic construction of Twitter sentiment classifiers. The main difficulty is obtaining sentiment labels for a large enough set of tweets which are then used for training. One of the first approaches was by Go et al. [44] who used smileys as a proxy for sentiment labels of 1.6 million tweets. Collecting high quality human-labeled tweets is considerably more expensive. Saif et al. [45] give a survey of eight manually annotated datasets having from 500 to 14,000 labeled tweets, considerably less than our dataset. A more exhaustive review of different approaches to Twitter sentiment analysis, both lexicon-based and machine learning, is in [46]. A methodology which combines the lexicon-based and machine learning approaches is given in [47].

In this paper, as is common in the sentiment analysis literature [48], we have approximated the sentiment of tweets with an ordinal scale of three values: *negative* (−), *neutral* (0), and *positive* (+). Sentiment classification is an ordinal classification task, a special case of multi-class classification where there is a natural ordering between the classes, but no meaningful numeric difference between them [49]. Our classifier is based on Support Vector Machine (SVM), a widely used, state-of-the-art supervised learning algorithm, well suited for large scale text categorization tasks, and robust on large feature spaces. We implemented the wrapper approach, described in [50], which constructs two linear-kernel SVM [42] classifiers. Since the classes are ordered, two classifiers suffice to partition the space of tweets into the three sentiment areas. The two SVM classifiers were trained to distinguish between *positive* and *negative-or-neutral*, and between *negative* and *positive-or-neutral*, respectively. During prediction, if the target class cannot be determined as the two classifiers disagree (which happens rarely), the tweet is labeled as *neutral*.

When preprocessing tweets, we removed URLs because they normally do not represent relevant content but rather point to it. We also removed cash-tags (e.g., "$NKE") and user mentions (e.g., "@johndoe") to make a tweet independent of a specific stock (company) and/or users involved in the discussion, and thus make the first step towards generalizing our model. Last but not least, we collapsed letter repetitions (e.g., "coooool" becomes "cool"). This step is relatively easy to implement and has proven useful for sentiment classification tasks [46]. After these steps, we followed a typical bag-of-words computation procedure by applying tokenization (based on relatively simple regular expressions), lemmatization (we used LemmaGen [51] for this purpose), *n*-gram construction (we included unigrams and bigrams into the feature set), and the TF-IDF weighting scheme [52]. Note that we did not remove stop words, such as "not", as this would in some cases change the sentiment polarity of a tweet.

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 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:

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

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 [33] to check if the Twitter variables help in the prediction of the price returns. The steps of the procedure applied are summarized as follows [53]:

- 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 [41]. 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 Fig 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
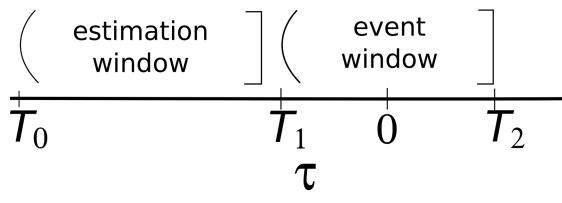
Fig 1. Time line for an event study.

and event date d, we have:

$$AR_{i,d} = R_{i,d} - E[R_{i,d}] \qquad (3)$$

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. Fig 1). For example, in an event study using daily data and the market model, the market model parameters 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, $AR_{i,\tau} \sim \mathcal{N}(0, \sigma^2(AR_{i,\tau}))$ [34]. This forms the basis for a procedure which tests whether an abnormal return is statistically significant.

Event detection from Twitter data. The following subsections first define the algorithm used to detect Twitter activity peaks which are then treated as events. Next, a method to assign a polarity to the events, using the Twitter sentiment, is described. Finally, we discuss a specific type of events for the companies studied, called earnings announcement events (abbreviated EA), 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 [54]. 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
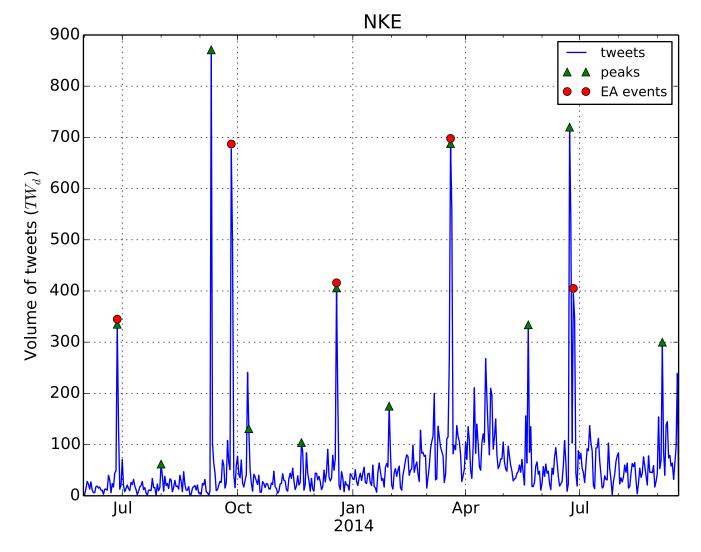
Fig 2. Daily time series of Twitter volume for the Nike company. Detected Twitter peaks and actual EA events are indicated.

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 $\phi_t = 2$. The threshold ϕ_t determines the number of detected peaks and the overlaps between the event windows—both increase with larger ϕ_t. One should maximize the number of detected peaks, and minimize the number of overlaps [41]. We have analyzed the effects of varying ϕ_t from 0.5 to 10 (as in [54]). The decrease in the number of overlaps is substantial for ϕ_t ranging from 0.5 to 2, for larger values the decrease is slower. Therefore, we settled for $\phi_t = 2$. As a final step we apply filtering which removes detected peaks that are less then 21 days (the size of the event window) apart from the other peaks.

As an illustration, the resulting activity peaks for the Nike company are shown in Fig 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 Fig 3. To obtain three sets of events with approximately the same size, we select the following thresholds, and define the event polarity as follows:

- If $P_d \in [-1, 0.15)$ the event is a *negative event*,

- If $P_d \in [0.15, 0.7]$ the event is a *neutral event*,

- If $P_d \in (0.7, 1]$ the event is a *positive event*.

Putting thresholds on a signal is always somewhat arbitrary, and there is no systematic treatment of this issue in the event study [41]. The justification for our approach is that sentiment should be regarded in relative terms, in the context of related events. Sentiment polarity has no absolute meaning, but provides just an ordering of events on the scale from −1 (negative) to +1 (positive). Then, the most straightforward choice is to distribute all the events uniformly between the three classes. Conceptually similar approaches, i.e., treating the sentiment in relative terms, were already applied to compare the sentiment leaning of network communities towards different environmental topics [39], and to compare the emotional reactions to conspiracy and science posts on Facebook [38]. Additionally, in the closely related work by Sprenger et al. [36], the authors use the percentage of positive tweets for a given day d, to determine the event polarity. Since they also report an excess of positive tweets, they use the median share of positive tweets as a threshold between the positive and negative events.
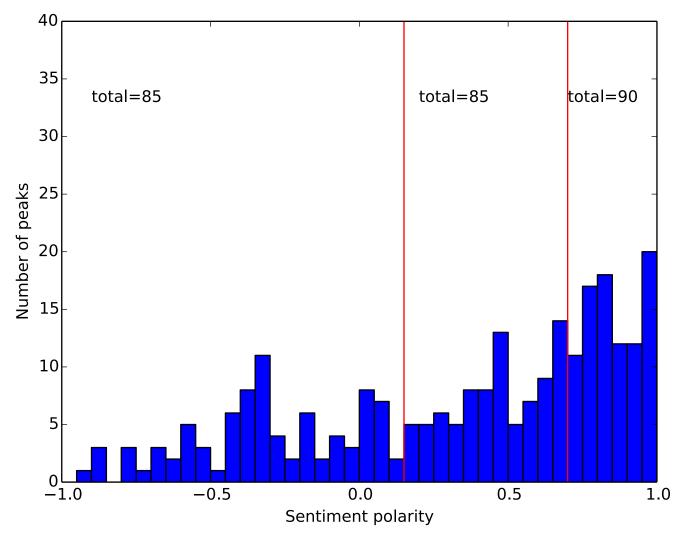
Event types. For a specific type of events in finance, in particular quarterly *earnings announcements* (EA), it is known that the price return of a stock abnormally jumps in the direction of the earnings [34, 35]. In our case, the Twitter data show high posting activity during the EA events, as expected. However, there are also other peaks in the Twitter activity, which do not correspond to EA, abbreviated as non-EA events. See Fig 2 for an example of Nike.
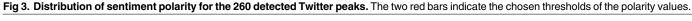
The total number of peaks that our procedure detects in the period of the study is 260. Manual examination reveals that in the same period, there are 151 EA events (obtained from http://www.zacks.com/). Our event detection procedure detects 118 of them, the rest are non-EA events. This means that the recall (the fraction of all EA events that were correctly detected as EA) of our peak detection procedure is 78%. In contrast, Sprenger et al. [36] detect 224 out of 672 EA events, yielding the recall of 33%. They apply a simpler peak detection procedure: a Twitter peak is defined as one standard deviation increase of the tweet volume over the previous five days.

The number of the detected peaks 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. 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.

Fig 3. Distribution of sentiment polarity for the 260 detected Twitter peaks. The two red bars indicate the chosen thresholds of the polarity values.

doi:10.1371/journal.pone.0138441.g003

We report all the detected peaks, with the dates and their polarity, in S1 Appendix. The EA events are in Table 1, and the non-EA events are in Table 2.

The first scenario allows to compare the results of the Twitter sentiment with the existing literature in financial econometrics [34]. 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 contain 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 [34] and [55]. 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

Table 2. A comparison of the inter-annotator agreement and the classifier performance. The inter-annotator agreement is computed from the examples labeled twice. The classifier performance is estimated from the 10-fold cross-validation.

	Annotator agreement	Sentiment classifier
No. of hand-labeled examples	6,143	103,262
Accuracy(−,0,+)	77.1%	76.0 ± 0.3%
Accuracy±1(−,+)	98.8%	99.4 ± 0.1%
$\overline{F}_1(-,+)$	49.4%	50.8 ± 0.5%
Precision / Recall(−)	48.0/48.0%	71.3/38.9%
Precision / Recall(+)	50.9/50.9%	68.6/40.9%

doi:10.1371/journal.pone.0138441.t002

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:

$$R_{i,d} = \alpha_i + \beta_i R_{DJIA,d} + \epsilon_{i,d} \tag{4}$$

$$E(\epsilon_{i,d}) = 0, \quad var(\epsilon_{i,d}) = \sigma^2_{\epsilon_{i,d}} \tag{5}$$

$$E[R_{i,d}] = \hat{\alpha}_i + \hat{\beta}_i R_{DJIA,d} \tag{6}$$

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 [34]. Using the notation presented in Fig 1 for the time line, the estimated value of $\sigma^2_{\epsilon_{i,d}}$ is:

$$\hat{\sigma}^2_{\epsilon_{i,d}} = \frac{1}{L-2} \sum_{d=T_0+1}^{T_1} (R_{i,d} - \hat{\alpha}_i - \hat{\beta}_i R_{DJIA,d})^2 \tag{7}$$

where $\hat{\alpha}_i, \hat{\beta}_i$ are the estimated parameters following the OLS procedure [34]. The abnormal return for company *i* at day *d* is the residual:

$$AR_{i,d} = R_{i,d} - \hat{\alpha}_i - \hat{\beta}_i R_{DJIA,d}. \tag{8}$$

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:

$$AR_{i,\tau} \sim \mathcal{N}(0, \sigma^2_{AR}) . \tag{9}$$

Eq (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 [55], we get:

$$\overline{AR}_{\tau} = (1/N)\sum_{i=1}^{N} AR_{i,\tau} \; . \tag{10}$$

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

$$CAR(\tau_1, \tau_2) = \sum_{\tau=\tau_1}^{\tau_2} \overline{AR}_{\tau} \; . \tag{11}$$

To calculate the variance of the *CAR*, we assume $\sigma^2_{AR} = \sigma^2_{\epsilon_{i,t}}$ (shown in e.g., [34, 55]):

$$var(CAR(\tau_1, \tau_2)) = (1/N^2)\sum_{i=1}^{N}(\tau_2 - \tau_1 + 1)\sigma^2_{\epsilon_i} \tag{12}$$

where *N* is the total number of events. Finally, we introduce the test statistic $\hat{\theta}$. With this quantity we can test if the measured return is abnormal:

$$\frac{CAR(\tau_1, \tau_2)}{\sqrt[2]{var(CAR(\tau_1, \tau_2))}} = \hat{\theta} \sim \mathcal{N}(0, 1) \tag{13}$$

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

Results

This section first presents an exhaustive evaluation of the Twitter sentiment classification model. Then it shows the correlation and Granger causality results over the entire time period. Finally, it shows statistically significant results of the event study methodology as applied to Twitter data.

Twitter sentiment classification

In machine learning, a standard approach to evaluate a classifier is by cross-validation. We have performed a 10-fold cross-validation on the set of 103,262 annotated tweets. The whole training set is randomly partitioned into 10 folds, one is set apart for testing, and the remaining nine are used to train the model and evaluate it on the test fold. The process is repeated 10 times until each fold is used for testing exactly once. The results are computed from the 10 tests and the means and standard deviations of different evaluation measures are reported in Table 2.

Cross-validation gives an estimate of the sentiment classifier performance on the application data, assuming that the training set is representative of the application set. However, it does not provide any hint about the highest performance achievable. We claim that the agreement between the human experts provides an upper bound that the best automated classifier can achieve. The inter-annotator agreement is computed from a fraction of tweets annotated twice. During the annotation process, 6,143 tweets were annotated twice, by two different annotators. The results were used to compute various agreement measures.

There are several measures to evaluate the performance of classifiers and compute the inter-annotator agreement. We have selected the following three measures to estimate and compare them: *Accuracy*, *Accuracy±1*, and $\overline{F_1}$. *Accuracy*(−,0,+) is the fraction of correctly classified examples for all three sentiment classes. This is the simplest and most common measure, but it doesn't take into account the ordering of the classes. On the other extreme, *Accuracy±1*(−,+) (a

shorthand for *Accuracy within 1* neighboring class) completely ignores the neutral class. It counts as errors just the negative sentiment examples predicted as positive, and vice versa. $\bar{F}_1(-,+)$ is the average of F_1 for the negative and positive class. It does not account for the misclassification of the neutral class since it is considered less important than the extremes, i.e., negative or positive sentiment. However, the misclassification of the neutral sentiment is taken into account implicitly as it affects the precision and recall of the extreme classes. F_1 is the harmonic mean of *Precision* and *Recall* for each class. *Precision* is a fraction of correctly predicted examples out of all the predictions of a particular class. *Recall* is a fraction of correctly predicted examples out of all actual members of the class. $\bar{F}_1(-,+)$ is a standard measure of performance for sentiment classifiers [56].

Table 2 gives the comparison of the inter-annotator agreement and the classifier performance. The classifier has reached the annotator agreement in all three measures. In a closely related work by Sprenger et al. [36], they use Naive Bayes for sentiment classification. Their classifier is trained on 2,500 examples, and the 10-fold cross-validation yields *Accuracy* of 64.2%.

We argue that in our case, there is no need to invest further work to improve the classifier. Most of the hypothetical improvements would likely be the result of overfitting the training data. We speculate that the high quality of the sentiment classifier is mainly the consequence of a large number of training examples. In our experience in diverse domains, one needs about 50,000—100,000 labeled examples to reach the inter-annotator agreement.

If we compare the F_1 measures, we observe a difference in the respective *Precision* and *Recall*. For both classes, − and +, the sentiment classifier has a considerably higher *Precision*, at the expense of a lower *Recall*. This means that tweets, classified into extreme sentiment classes (− or +) are likely indeed negative or positive (*Precision* about 70%), even if the classifier finds only a smaller fraction of them (*Recall* about 40%). This suits well the purpose of this study. Note that it is relatively easy to modify the SVM classifier, without retraining it, to narrow the space of the neutral class, thus increasing the recall of the negative and positive classes, and decreasing their precision. One possible criterion for such a modification is to match the distribution of classes in the application set, as predicted by the classifier, to the actual distribution in the training set.

Correlation and Granger causality

Pearson correlation. Table 3 shows the computed Pearson correlations, as defined in the Methods section. The computed coefficients are small, but are in line with the result of [30]. In our opinion, these findings and the one published in [30] 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 3. 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 [28, 30].
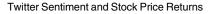
Table 3. Results of the Pearson correlation and Granger causality tests. Companies are ordered as in Table 1. The arrows indicate a statistically significant Granger causality relation for a company, at the 5% significance level. A right arrow indicates that the Twitter variable (sentiment polarity P_d 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.
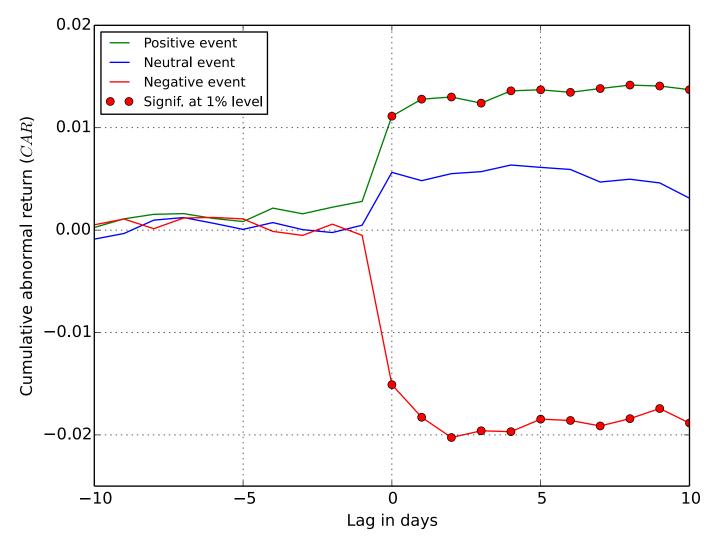
	Pearson correlation	Granger causality			
Ticker	$\rho(P_d, R_d)$	P_d & R_d		TW_d & $|R_d|$	
TRV	0.1178	←			
UNH	0.2565	←			
UTX	0.1370	←			
MMM	0.1426	←		←	
DD	0.2680	←			
AXP	0.1566	←			→
PG	0.2145				
NKE	0.2460				
CVX	0.2053				
HD	0.2968	←			→
CAT	0.3648				
JNJ	0.2220				
V	0.2995	←			
VZ	0.1775				
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				→
Total		10	3	2	10

Cumulative abnormal returns

The results of the event study are shown in Figs 4 and 5, where the cumulative abnormal returns (*CAR*) are plotted for the two types of events defined earlier. The results are largely consistent with the existing literature on the information content of earnings [34, 35]. The evidence strongly supports the hypothesis that tweets do indeed convey information relevant for stock returns.
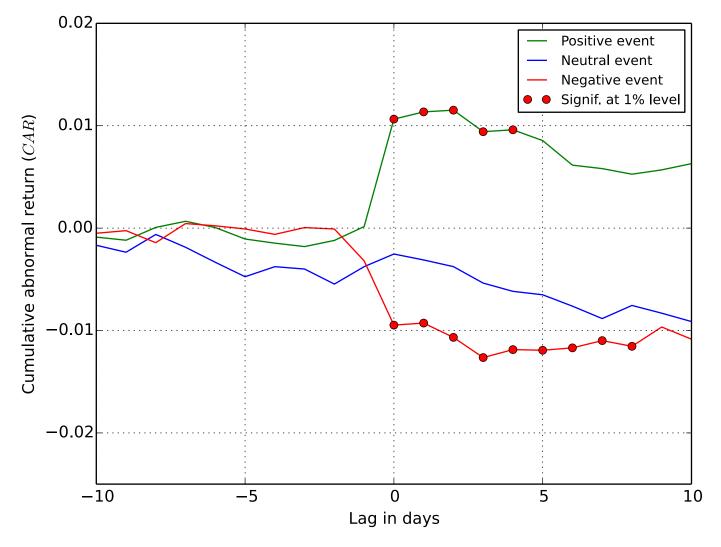
Fig 4. _CAR_ for all detected events, including EA. The _x_ axis is the lag between the event and _CAR_, and the red markers indicate days with statistically significant abnormal return.

Fig 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. The values of _CAR_ are significant at the 1% level for ten days after the positive sentiment events. Given this result, the null hypothesis that the events have no impact on price returns is rejected. The same holds for negative sentiment events, but the _CAR_ (actually loss) is twice as large in absolute terms. The _CAR_ after the neutral events is very low, and barely significant at the 5% level at two days; at other days one cannot reject the null hypothesis. We speculate that the positive _CAR_ values for the neutral events, barely significant, are the result of the uniform distribution of the Twitter peaks into three event classes (see Fig 3). An improvement over this baseline approach remains a subject of further research.

Fig 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.

doi:10.1371/journal.pone.0138441.g005

A more interesting result concerns the non-EA events in Fig 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 shows that after the event days the values of *CAR* remain significant at the 1% level for four days after the positive events, and for eight 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 values of *CAR* are lower, but in both cases the impact is statistically significant. The *CAR* for the neutral events tend to be slightly negative (in contrast to the EA events), albeit are not statistically significant. However, this again indicates that the distribution of Twitter peaks into the event classes could be improved.

Table 4. Values of the $\hat{\theta}$ statistic for each type of event. Significant results at the 1% level ($|\hat{\theta}| > 2.58$) are denoted by **, and at the 5% level ($|\hat{\theta}| > 1.96$) by *.

Lag	All events (including EA)			Non-EA events		
(days)	negative	neutral	positive	negative	neutral	positive
-10	0.6408	-1.0730	0.3208	-0.5281	-1.5168	-1.0017
-9	0.9495	-0.2828	0.9806	-0.1847	-1.5060	-0.9509
-8	0.0977	0.6852	1.1197	-0.8646	-0.3225	0.0458
-7	0.7302	0.7470	1.0126	0.2333	-0.8464	0.3790
-6	0.6865	0.3657	0.6419	0.1069	-1.3505	0.0276
-5	0.5536	0.0356	0.4295	-0.0358	-1.7525	-0.4941
-4	-0.0580	0.3377	1.0212	-0.2430	-1.2873	-0.6304
-3	-0.2255	0.0207	0.7089	0.0200	-1.2781	-0.7248
-2	0.2395	-0.0961	0.9382	-0.0302	-1.6476	-0.4560
-1	-0.1981	0.1849	1.1148	-1.0632	-1.0765	0.0535
0	-5.6350**	2.0709*	4.2197**	-3.0057**	-0.6897	3.6489**
1	-6.5332**	1.6975	4.6436**	-2.8173**	-0.8118	3.7254**
2	-6.9559**	1.8629	4.5338**	-3.1146**	-0.9436	3.6325**
3	-6.4855**	1.8582	4.1682**	-3.5557**	-1.2979	2.8611**
4	-6.2936**	1.9989*	4.4168**	-3.2240**	-1.4419	2.8187**
5	-5.7154**	1.8655	4.3086**	-3.1383**	-1.4721	2.4297*
6	-5.5829**	1.7492	4.1047**	-2.9850**	-1.6720	1.6956
7	-5.5822**	1.3478	4.0987**	-2.7250**	-1.8837	1.5573
8	-5.2308**	1.3889	4.0868**	-2.7867**	-1.5667	1.3732
9	-4.8243**	1.2552	3.9575**	-2.2729*	-1.6803	1.4462
10	-5.0916**	0.8288	3.7645**	-2.4901*	-1.8009	1.5622

doi:10.1371/journal.pone.0138441.t004

These results are similar to the ones reported by Sprenger at al. [36]. In addition, the authors show statistically significant increase in the *CAR* values even before the positive event days. They argue that this is due to the information leakage before the earnings announcements. We observe a similar phenomena, but with very low *CAR* values, and not statistically significant (cf. the positive events at day −1 in Fig 4).

Discussion

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

Similar results were corroborated in a recent, independent study by Sprenger et al. [36]. The authors have made an additional step, and classified the non-EA events into a comprehensive set of 16 company-specific categories. They have used the same training set of 2,500 manually classified tweets to train a Naive Bayesian classifier which can then reasonably well discriminate between the 16 categories. In our future work, we plan to identify topics, which are not predefined, from all the tweets of non-EA events. We intend to apply standard topic detection algorithms, such as Latent Dirichlet allocation (LDA) or clustering techniques.

Studies as this one could be well used in order to establish a direct relation between social networks and market behavior. 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. One possible direction is to test the presence of forecasting power of the sentiment time series. Following an approach similar to the one presented by Moat et al. [57] one can decide to buy or sell according to the presence of a peak in the tweet volume and the level of polarity in the corresponding direction. However, detection of Twitter events should rely just on the current and past Twitter volume.

Also, during the events, we might move to a finer time scale, e.g., from daily to hourly resolution, as done by [32]. 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.

Supporting Information

S1 Appendix. Event dates and polarity. Detailed information about the detected events from the Twitter data and their polarity. We show the 118 detected EA events and 182 detected non-EA events.
(PDF)

Acknowledgments

Author Contributions

Conceived and designed the experiments: GR DA IM. Performed the experiments: GR DA MG. Analyzed the data: GR DA IM GC MG. Wrote the paper: GR DA IM GC MG.

References

1. King G. Ensuring the data-rich future of the social sciences. Science. 2011; 331(6018):719–721. doi: 10.1126/science.1197872 PMID: 21311013

2. Vespignani A. Predicting the behavior of techno-social systems. Science. 2009; 325(5939):425–428. doi: 10.1126/science.1171990 PMID: 19628859

3. Bouchaud JP. The (unfortunate) complexity of the economy. arXiv preprint. 2009; Available from: http://arxiv.org/arbs/0904.0805

4. Haldane AG, May RM. Systemic risk in banking ecosystems. Nature. 2011; 469(7330):351–355. doi: 10.1038/nature09659 PMID: 21248842

5. Schweitzer F, Fagiolo G, Sornette D, Vega-Redondo F, Vespignani A, White DR. Economic Networks: The New Challenges. Science. 2009; 325(5939):422–425. PMID: 19628858

6. Bouchaud JP. Economics needs a scientific revolution. Nature. 2008; 455(7217):1181–1181. doi: 10.1038/4551181a

7. Cutler DM, Poterba JM, Summers LH. What moves stock prices? Journal of Portfolio Management. 1989; 15:4–12. doi: 10.3905/jpm.1989.409212

8. Chan WS. Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics. 2003; 70(2):223–260. doi: 10.1016/S0304-405X(03)00146-6

9. Vega C. Stock price reaction to public and private information. Journal of Financial Economics. 2006; 82(1):103–133. doi: 10.1016/j.jfineco.2005.07.011

10. Alanyali M, Moat HS, Preis T. Quantifying the relationship between financial news and the stock market. Scientific reports. 2013; 3. doi: 10.1038/srep03578 PMID: 24356666

11. Tetlock PC. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance. 2007; 62(3):1139–1168. doi: 10.1111/j.1540-6261.2007.01232.x

12. Tetlock PC, Saar-Tsechansky M, Macskassy S. More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance. 2008; 63(3):1437–1467. doi: 10.1111/j.1540-6261.2008.01362.x

13. Lillo F, Miccichè S, Tumminello M, Piilo J, Mantegna RN. How news affect the trading behavior of different categories of investors in a financial market. Quantitative Finance. 2014; doi: 10.1080/14697688.2014.931593

14. Engelberg JE, Reed AV, Ringgenberg MC. How are shorts informed? Short sellers, news, and information processing. Journal of Financial Economics. 2012; 105(2):260–278. doi: 10.1016/j.jfineco.2012.03.001

15. Birz G, Lott JR Jr. The effect of macroeconomic news on stock returns: New evidence from newspaper coverage. Journal of Banking & Finance. 2011; 35(11):2791–2800. doi: 10.1016/j.jbankfin.2011.03.006

16. 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. 2011; 18(2):321–340. doi: 10.1016/j.jempfin.2010.11.009

17. Preis T, Reith D, Stanley HE. 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. 2010; 368(1933):5707–5719. doi: 10.1098/rsta.2010.0284

18. Bordino I, Battiston S, Caldarelli G, Cristelli M, Ukkonen A, Weber I. Web search queries can predict stock market volumes. Plos One. 2012; 7(7):e40014. doi: 10.1371/journal.pone.0040014 PMID: 22829871

19. Bordino I, Kourtellis N, Laptev N, Billawala Y. Stock trade volume prediction with Yahoo Finance user browsing behavior. In: Proc. 30th IEEE Intl. Conf. on Data Engineering (ICDE); 2014. p. 1168–1173.

20. Da Z, Engelberg J, Gao P. In search of attention. The Journal of Finance. 2011; 66(5):1461–1499. doi: 10.1111/j.1540-6261.2011.01679.x

21. Kristoufek L. Can Google Trends search queries contribute to risk diversification? Scientific reports. 2013; 3. doi: 10.1038/srep02713

22. Curme C, Preis T, Stanley HE, Moat HS. Quantifying the semantics of search behavior before stock market moves. Proc National Academy of Sciences. 2014; 111:11600–11605. doi: 10.1073/pnas.1324054111

23. Varkman T, Kristoufek L. Underpricing, underperformance and overreaction in initial public offerings: Evidence from investor attention using online searches. SpringerPlus. 2015; 4(84):1–11.

24. Graham M, Hale SA, Gaffney D. Where in the world are you? Geolocation and language identification in Twitter. The Professional Geographer. 2014; 66(4):568–578. doi: 10.1080/00330124.2014.907699

25. Nguyen VD, Varghese B, Barker A. The royal birth of 2013: Analysing and visualising public sentiment in the uk using twitter. In: 2013 IEEE International Conference on Big Data. IEEE; 2013. p. 46–54.

26. Java A, Song X, Finin T, Tseng B. Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM; 2007. p. 56–65.

27. Mao Y, Wei W, Wang B, Liu B. Correlating S&P 500 stocks with Twitter data. In: Proc. 1st ACM Intl. Workshop on Hot Topics on Interdisciplinary Social Networks Research; 2012. p. 69–72.

28. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of Computational Science. 2011; 2(1):1–8. doi: 10.1016/j.jocs.2010.12.007

29. Bollen J, Pepe A, Mao H. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: Proc. 5th Intl. AAAI Conf. on Weblogs and Social Media; 2011. p. 450–453.

30. Mao H, Counts S, Bollen J. Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint. 2011; Available from: http://arxiv.org/abs/1112.1051

31. Souza TTP, Kolchyna O, Treleaven PC, Aste T. Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry. arXiv preprint. 2015; Available from: http://arxiv.org/abs/1507.00784

32. Zheludev I, Smith R, Aste T. When Can Social Media Lead Financial Markets? Scientific Reports. 2014; 4. doi: 10.1038/srep04213

33. Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society. 1969;p. 424–438. doi: 10.2307/1912791

34. Campbell JY, Lo AWC, MacKinlay. The econometrics of financial markets. Princeton University Press; 1997.

35. Boehmer E, Masumeci J, Poulsen AB. Event-study methodology under conditions of event-induced variance. Journal of Financial Economics. 1991; 30(2):253–272. doi: 10.1016/0304-405X(91)90032-F

36. Sprenger TO, Sandner PG, Tumasjan A, Welpe IM. News or Noise? Using Twitter to Identify and Understand Company-specific News Flow. Journal of Business Finance & Accounting. 2014; 41(7–8):791–830. doi: 10.1111/jbfa.12086

37. Sprenger TO, Tumasjan A, Sandner PG, Welpe IM. Tweets and trades: The information content of stock microblogs. European Financial Management. 2014; 20(5):926–957. doi: 10.1111/j.1468-036X.2013.12007.x

38. Zollo F, Novak PK, Del Vicario M, Bessi A, Mozetič I, Scala A, et al. Emotional Dynamics in the Age of Misinformation. To appear in Plos One, arXiv preprint. 2015; Available from: http://arxiv.org/abs/1505.08001

39. Sluban B, Smailović J, Battiston S, Mozetič I. Sentiment Leaning of Influential Communities in Social Networks. Computational Social Networks. 2015; 2(9).

40. Smailović J, Kranjc J, Grčar M, Žnidaršič M, Mozetič I. Monitoring the Twitter sentiment during the Bulgarian elections. In: Proc. IEEE Intl. Conf. on Data Science and Advanced Analytics. IEEE; 2015.

41. MacKinlay AC. Event studies in economics and finance. Journal of economic literature. 1997;p. 13–39.

42. Vapnik VN. The Nature of Statistical Learning Theory. New York, USA: Springer; 1995.

43. Zhang W, Skiena S. Trading Strategies to Exploit Blog and News Sentiment. In: Proc. 4th Intl. Conf. on Weblogs and Social Media; 2010.

44. Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. Stanford University; 2009. CS224N Project Report.

45. Saif H, Fernandez M, He Y, Alani H. Evaluation datasets for twitter sentiment analysis. In: Proc. 1st Workshop on Emotion and Sentiment in Social and Expressive Media. Turin, Italy; 2013.

46. Smailović J. Sentiment analysis in streams of microblogging posts. PhD Thesis, Jozef Stefan International Postgraduate School. Ljubljana, Slovenia; 2015.

47. Kolchyna O, Souza TTP, Treleaven PC, Aste T. Methodology for Twitter Sentiment Analysis. arXiv preprint. 2015; Available from: http://arxiv.org/abs/1507.00955

48. Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008; 2(1–2):1–135. doi: 10.1561/1500000011

49. Gaudette L, Japkowicz N. Evaluation methods for ordinal classification. In: Advances in Artificial Intelligence. Springer; 2009. p. 207–210.

50. Frank E, Hall M. A simple approach to ordinal classification. In: Proceedings 12th European Conference on Machine Learning. Springer; 2001. p. 145–156.

51. Juršič M, Mozetič I, Erjavec T, Lavrač N. LemmaGen: Multilingual Lemmatisation with Induced Ripple-Down Rules. Journal of Universal Computer Science. 2010; 16:1190–1214.

52. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann; 2011.

53. Piškorec M, Antulov-Fantulin N, Novak PK, Mozetič I, Grčar M, Vodenska I, et al. Cohesiveness in Financial News and its Relation to Market Volatility. Scientific reports. 2014; 4. doi: 10.1038/srep05038 PMID: 24849598

54. Lehmann J, Gonçalves B, Ramasco JJ, Cattuto C. Dynamical classes of collective attention in Twitter. In: Proceedings of the 21st international conference on World Wide Web; 2012. p. 251–260.

55. Malkiel BG, Fama EF. Efficient capital markets: A review of theory and empirical work. The Journal of Finance. 1970; 25(2):383–417. doi: 10.1111/j.1540-6261.1970.tb00518.x

56. Kiritchenko S, Zhu X, Mohammad SM. Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research. 2014;p. 723–762.

57. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T. Quantifying Wikipedia usage patterns before stock market moves. Scientific reports. 2013; 3. doi: 10.1038/srep01801