JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12102

WALKING DOWN WALL STREET WITH A TABLET: A SURVEY OF STOCK MARKET PREDICTIONS USING THE WEB

Michela Nardo*, Marco Petracco-Giudici and Minás Naltsidis

European Commission – Joint Research Centre

Abstract. 'A blindfolded chimpanzee throwing darts at The Wall Street Journal could select a portfolio that would do as well as the (stock market) experts' [Malkiel (2003) The efficient market hypothesis and its critics. *Journal of Economic Perspectives* 17(1): 59–82)]. However, what if this chimpanzee could browse the Internet before throwing any darts? In this paper, we ask whether online news has any influence on the financial market, and we also investigate how much influence it has. We explore the burgeoning literature on the predictability of financial movements using online information and report its mixed findings. In addition, we collate the efforts of various disciplines, including economics, text mining, sentiment analysis and machine learning, and we offer suggestions for future research.

Keywords. Big data; Financial predictions; Machine learning; Sentiment analysis; Text mining; Trading strategies; Web mining

1. Introduction

Financial prices and transactions are not predictable according to the efficient market hypothesis (EMH, Fama, 1965), which models stock markets as random walks where shocks are temporary and largely driven by new and unexpected information.[1] Instead, the critical reading of EMH offered by Behavioural Finance (Della Vigna, 2009) suggests a certain degree of predictability. Investors could be subject to waves of optimism and pessimism, which would cause prices to deviate systematically from their fundamental values (DeBond and Thaler, 1985; Schleifer and Summers, 1990), or investors may be systematically overconfident of their ability to forecast future stocks based on prices or earnings (Kahneman and Tversky, 1979). Recently, the geometric increase in digital information (online journals, dedicated blogs, social networks, etc.) has made it possible to address the predictability of financial markets from a different perspective, namely, Big Data. Terabits of data on financial transactions can be matched to a comparable amount of online news to delve into the mechanisms of decision making. Thus, the Internet has changed the way in which information is delivered to investors and how investors can act on it (Barber and Odean, 2001; Moat *et al*., 2014).

The literature relating web mining to financial predictions is relatively recent and burgeoning (to the best of our knowledge, the first study is due to Wysocki in 1998). The contributions coming from computer science and text mining have shifted the focus of attention from sophisticated prediction models

*Corresponding author contact email: michela.nardo@jrc.ec.europa.eu; Tel.: +39-0332-785968.

The views expressed in this paper are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

to the information itself, and the most useful online information is sought. Specifically, researchers have investigated how a piece of text can be an indicator of change, how complex feelings can be expressed as numbers and whether the communication structure has any effect on decision making.

This paper offers an overview of the main issues raised by the literature. We start from the simplest approach (Section 1): relating the financial movements with the frequency of web searches for certain keywords (or the presence of these keywords on the web). We examine the literature on machine learning and instant trading (Section 2) when information comes from the analysis of a whole set of texts. In addition, we discuss tonality and sentiment when text mining techniques are used to transform human feelings into numbers (Section 3). Section 4 concludes and provides suggestions for future research. This review brings up the vast array of topics in machine learning and sentiment analysis. We apologise to the experts in these fields for the lack of rigor (and also the lack of exhaustive references), but we have treated this literature as instrumental to our purpose, which is relating web buzz to financial predictions.

In addition to financial markets, online information has been used for assessing a wide variety of economic or social phenomena. A Facebook's Gross National Happiness index is calculated by Mishne and Rijke (2006). Preis *et al*. (2012) analyse the link between online behaviour and real-world economic indicators in 2010 for 45 countries, and they find that the Future Orientation Index has a 0.78 correlation with the local GDP. Doshi *et al*. (2009), Mishne and Glance (2006), Asur and Huberman (2010) and Goel *et al*. (2010) use information extracted from the web to forecast the box-office revenues of movies, videogame sales and Oscar winners. Tweets were considered an alternative to polls when forecasting the results of the 2009 German Federal Election (Tumasjan *et al*., 2010). Choi and Varian (2012) and McLaren and Shanbhogue (2011) show that Google queries improve the nowcasts of automobiles sales, unemployment claims, travel destination planning and house prices. Blog posts and blog sentiment are found to anticipate product sales (Gruhl *et al*., 2005).

The entire literature relating web buzz to economic changes has in common the difficulty in accessing high-quality web data. The relevant online information is heterogeneous, seldom freely accessible (e.g. raw data within Google Trend) and in some cases prohibitively expensive (e.g. Twitter's access to Firehose). When available, the characteristics of the universe (from which the sample is drawn) are usually not known, challenging any statistical inferences (see Morstatter *et al*., 2013 for an example with Twitter). For free online information, strong IT skills are necessary to set up machines that are able to explore the web and retrieve texts,[2] and text mining techniques are needed to classify texts according to their relevance to the question being analysed and extract the significant information. Classification algorithms and text mining are on the research frontier at the moment, and there are two main questions (Section 3): how to eliminate irrelevant information and how to decode the mood or sentiment implied in the text. Solving these problems is not an easy task, especially in finance, where textual expressions are seldom of the type 'I like' or 'I do not like' (Loughran and McDonald, 2011).

2. Blogs, Twitter and Financial News

The common starting point of the literature on financial predictions using online information is a simple assumption: '*movements in financial markets and movements in (financial) news are intrinsically interlinked*' (Alanyali *et al*., 2013). Therefore, the issue is to uncover the strength of this link and the points in time when the association is more pronounced. It is crucial to determine how precise web buzz is at anticipating financial movements. If there is any such precision, it is critical to know whether web buzz only anticipates movements or is able to say something about the direction/magnitude of change.

The simplest measures of web buzz are the number (or the share) of web texts that contain an exogenously given keyword (typically, a company name/ticker symbol) or a set of related keywords (identifying concepts such as an economic recession or speculative bubbles) or the number of searches for a given keyword.

The correlation between a measure of web buzz and stock returns or transaction volumes is the most basic approach to accounting for co-movements. The findings agree on a rather low (but significant) average correlation with trading volumes, ranging from 0.1 (Alanyali *et al*., 2013[3]) to 0.3 (Bordino *et al*., 2012[4]) with peaks (up to 0.8) for individual stocks (Bordino *et al*., 2012; see also Das and Chen, 2001; Gloor *et al*., 2009; Zhang *et al*., 2010). The average correlation with stock returns is usually lower. Some authors uncover early warning signals of financial turmoil using cross-correlations (i.e. the correlation for non-overlapping periods of time, see Preis *et al*., 2010;[5] Bordino *et al*., 2012; Ruiz *et al*., 2012[6]), Granger causality test or other non-parametric tests to cope with non-normal errors (Gilbert and Karahalios, 2010;[7] Bollen *et al*., 2011[8]). Most of these forecasts are done in-sample, that is, by using the full set of available observations.[9]

When the intent is to quantify the (average) impact of web buzz, then a correlation is not sufficient. The literature uses two types of approaches: trading strategies and estimation.

Estimation usually takes the form of a panel (Antweiler and Frank, 2004[10]) or time-series (GARCH or VAR) regressions. Whereas Wysocki (1998)[11] finds that a 10-fold increase in overnight message postings leads to a 15.6% increase in the one-day-ahead trading volume but only to 0.7% average stock returns, Gilbert and Karahalios (2010) use a VAR model to show that one standard deviation rise in the web-based Anxiety Index corresponds to an S&P500 return, that is, 0.4% lower than otherwise expected (see also Karabulut, 2013; Mao *et al*., 2014). The result is larger than the −0.08% of Tetlock (2007[12]). No relation between web buzz and stock returns one day ahead is found by Tumarkin and Whitelaw (2001[13]) or Antweiler and Frank (2004[14]).

A suggestion for why the results are so diverse is given provided by De Choudhury *et al*. (2008[15]). With a simple linear model, they show that web buzz is only helpful in leading 'big' events. Small fluctuations and trends can be captured either by sophisticated models[16] (they use support vector regressions; see Gunn, 1998, for a tutorial) or by modelling trading strategies whereby automated agents buy and sell stocks according to some web-based priors. Trading strategies are particularly useful when analysing the mechanisms of decision making and especially herding behaviours (Saavedra *et al*., 2011): when humans face a large set of choices and a large uncertainty on their pay-offs, the 'natural' tendency is to utilise the 'wisdom of the crowds'. This result is very much in line with Bentley *et al*. (2014), who propose an interesting theoretical framework for mapping collective behaviour in the big data era.

Trading strategies are usually implemented using machine learning. The great majority of contributions in this field (Section 2) are interested in the direction of price movements, not in their magnitude (with the notable exception of De Choudhury *et al*., 2008). Predicting the magnitude of price changes takes as fact the proportionality between the extension of web buzz (or the intensity of its mood) and the actual change in prices. To the best of our knowledge, no theories in behavioural finance support or investigate this hypothesis.

Examples of trading strategies that are not related to machine learning but are related to the use of communication metadata are given in Preis *et al*. (2010, 2013). In 2010, they show that the weekly transactions volumes of S&P500 companies are correlated (with at most a correlation of 0.3) with the weekly search volume of company names. In 2013, for the term *debt* the same authors find a return of the web-based trading strategy up to 2.31 standard deviations higher than the random strategy.[17] In addition, they prove that more finance-related keywords tend to provide successful strategies for market moves in the coming week (see also Curme *et al*., 2014). The starting point seems promising, and the access to text metadata is helpful.

3. Trading Strategies and Machine Learning

In the machine learning literature, the use of web information in conjunction with stock prices is essentially a question of activity monitoring (Fawcett and Provost, 1999): monitoring a stream of data and issuing an

alarm when there is a sign of positive/negative activity in it. Price trends are considered as the monitoring variable, and web *stories* (usually a set of keywords) are considered as alarms. The learning algorithms *supply a set of 'best stories': 'if the user is interested in stocks that are likely to go up in price, the learning algorithms supply the stories that are most likely to have come from a model of upward trend'* (Lavrenko *et al*., 2000b). When a new online text is found and has the same textual features of the 'best story', then the algorithm says an upward trend in prices has to be expected. The issue is not calculating the frequency of a certain set of exogenously supplied keywords, but rather finding the 'good' set of keywords from a 'good' set of texts.

Very generally, the mechanism functions in this manner: the machine starts performing a linear regression to find the price trend of a security, and then, an algorithm finds keywords that can lead to predictable outcomes (e.g. words like *earning* or *loss* are likely to predict upward/downward price movements). Stocks are then classified/weighted according to the keywords that can 'influence' them. After a training period, the machine learns the most relevant keywords and the likely price changes associated to them.

The training set is usually double-checked by humans to verify the accuracy of the machine learning. The following step is the forecast of future stock price movements: new texts are retrieved from the web and analysed for the presence of relevant keywords. Then, expected priors on the price changes are provided to guide trade strategies. The gain is computed by comparing this web-guided strategy with a baseline (usually a random choice within a set of stocks guaranteeing zero expected returns). A plethora of algorithms are available: from Bayesian rules (e.g. Gidófalvi, 2001; Gidófalvi and Elkan, 2003) to Genetic Algorithms or Support Vector Machines (see Schumaker and Chen, 2006, 2009; De Choudhury *et al*., 2008, who supply an intuitive overview of these algorithms).

The ability to forecast stock returns or trade volumes largely depends on the learning algorithm used (especially given that transaction costs are not taken into account). No prediction power for prices or volatility is found by Antweiler and Frank (2004) with Naïve Bayesian machine learning. Schumaker and Chen (2006[18]) find that more sophisticated algorithms yield up to 57% prediction accuracy and 2.06% returns. Their results are in line with Lavrenko *et al*. (2000a), who find a return of 2% from tracking four stocks (however, Lavrenko *et al*., 2000b,[19] find a return of 0.23% when enlarging the number of stocks to 127) while Mittermayer (2004[20]) obtains a modest 0.11% using the stocks from the NASDAQ.

The widespread explanation for these results (which is similar to the explanation derived from regression analysis) is that new information is rapidly incorporated into agents' information set, so excessive returns rapidly vanish: only very short (ideally, intraday) stock price movements can be capitalised on (Schumaker and Chen, 2006, 2009). Gidófalvi (2001) and Gidófalvi and Elkan (2003) show with two different data sets that the highest association between stock prices and news occurs 20 min prior to the news publications (i.e. only access to information before it is made public can affect trading prices).[21] Saavedra *et al*. (2011[22]) show with a logistic regression that herding behaviour is economically rewarding and positively related to volatility.

Interestingly, Schumaker and Chen (2009) find that partitioning the stocks of S&P500 firms according to their industrial sectors and analysing them separately increases both the accuracy in detecting the direction of price changes (76.02% for the financial sector) and the return from automated trading strategies (up to 8.5%).

Another possible explanation for the low returns is that trading strategies are suboptimal: while the automatic trader can determine the right time to buy (or sell), it is not able to calculate the optimal time to re-sell (or re-buy), which is typically exogenously defined. The limited exception is Lavrenko *et al*. (2000a,b), who design 1-hr trading strategies with the additional possibility of selling (or buying) the stock beforehand if a threshold on profits is reached (see also Mittermayer, 2004).

There are two additional weak points in this literature, and the first is the need to supply a large set of texts that have been checked in advance by humans to allow the machine learning process to unfold. The second weakness is the circularity assumption common to this literature: price trends are used to

classify messages on the supposition that messages follow prices. However, then messages are used to predict prices, which reverses the initial hypothesis. This paradox implies that only the 'old type' of stories can be clearly detected. Brand new stories cannot be recognised by the machine unless additional (human-guided) training is done. Both issues are tackled by the literature on sentiment analysis.

4. Good or Bad News? Tonality and Sentiment

Sentiment analysis generates stories independently from price movements. The idea is that a text is intrinsically positive or negative, not positive or negative because it has been 'optimally' associated to an upward/downward trend in stock prices.[23] In this context, text classifier algorithms select fit-to-the-purpose texts and spot positive/negative tonality or more complex feelings (sentiments) according to some externally provided dictionaries and semantic rules[24] (Box 1 reports examples of text classifiers).

Several issues are crucial to determining the quality of results. The first is having a 'good' set of texts from which tonality/sentiment is derived. The typical problem is noise reduction, that is, the detection of irrelevant messages. The literature has responded to this challenge in three ways (which are not mutually exclusive): (i) by limiting the web-coverage to specialised sources, such as The Wall Street Journal (Tetlock, 2007), Yahoo! Finance (Schumaker and Chen, 2006, 2009; Antweiler and Frank, 2004), Raging Bull (Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004) and The Financial Times (Alanyali *et al*., 2013); (ii) by utilising sources with a structured or annotated corpus of messages: Tweets (with 140 characters limit and explicit hash tags and feeds, see Zhang *et al*., 2010; Bollen *et al*., 2011; Ruiz *et al*., 2012) or New York Times Annotated Corpus (Zhai *et al*., 2011) and (iii) by using a linguistic classifier based on dictionaries.

Within the linguistic classification (and mood detection), the most common textual representation is the bag-of-words (Schumaker and Chen, 2006): texts are analysed according to the presence/frequency of a given set of words or keywords. More sophisticated algorithms for text classifications can be based on categories of assigned words (Tetlock, 2007; Gilbert and Karahalios, 2010) or on lexical semantic/syntactic tagging (Schumaker and Chen, 2006, 2009). The keywords can be drawn from very generic dictionaries (e.g. the Harvard psychosocial dictionary used in Tetlock, 2007), or less frequently, from finance-specific glossaries (Mittermayer, 2004; Ruiz-Martínez *et al*., 2012). A good dictionary is crucial for the selection of a meaningful set of articles and/or the detection of their mood, especially in finance: Loughran and McDonald (2011) proved that 73.8% of the words identified as negative by the widely used Harvard Dictionary are words typically not considered negative in the financial context (e.g. tax, cost, capital and liability).

Still, if a text is too short (typically, Twitter texts with maximum 140 characters), its correct classification and the detection of its mood is extremely difficult; hence, we have a *type I error* problem. The challenge remains to reduce the number of relevant messages that cannot be classified because they are too ambiguous or because the classification algorithm is unable to classify them (Godbole *et al*., 2007). A good text classifier for a financial corpus is a good avenue for future research (see Radford *et al*., 2009; Tobback *et al*., 2014).

Researchers in this field are mainly concerned with two questions: whether the mood is correctly identified by the machine, that is, accuracy, and to what extent the tonality is related to financial movements. Overall algorithms extracting binary tonality have proven to be rather accurate in correctly detecting positive and negative feelings (70.3% accuracy in-sample and 65.9% accuracy out-of-sample are found by Koppel and Shtrimberg, 2006;[25] see also Tulankar *et al*., 2013;[26] Zhai *et al*., 2011[27]) but incapable of tracking financial variables. First in the literature, Das and Chen (2001[28]) find that message posting is reactive (with 50 min of delay) but not predictive of stock price movements. The same (negative) result is found by Tumarkin and Whitelaw (2001[29]) and Koppel and Shtrimberg (2006), while a modest positive correlation (highest equals 0.45) with stock returns is found by Gloor *et al*. (2009[30]). With a more

sophisticated measure of sentiment, Zhang *et al*. (2010[31]) find a low negative correlation between Twitter sentiments and indices such as the NASDAQ, Dow Jones and S&P500 (highest equals −0.323 with the NASDAQ).

A number of factors contribute to the unsatisfactory results. In binary tonality, slightly negative/positive messages cannot be distinguished from very negative/positive ones. If a mild negative tonality is not translated into selling/buying actions, a binary tonality would overestimate price movements by diluting the link with price returns. The extent to which the degree of tonality is related to price movements is an open question. Devitt and Ahmad (2007), Ahmad *et al*. (2006) and O'Hare *et al*. (2009) offer some hints on how to construct an indicator that fully reflects the amount of sentiment in a text.

Investment decisions could be based on more sophisticated moods than binary tonality. Indeed, Bollen *et al*. (2011[32]) and Schumaker and Chen (2006) show that more complex sentiment analysis outperforms simple binary tonality. According to Bollen *et al*. (2011) mood prediction improves model accuracy up to 86.7% and reduces the Mean Average Percentage Error by 6% when forecasting the Dow Jones with respect to simple tonality.

Finally, the modest results could be due to the symmetry with which negative and positive feelings are treated. Koppel and Shtrimberg (2006) find negative words (e.g. 'shortfall', 'negative' and 'investigation') that clearly relate sentiments to price drops, but could not find corresponding positive words.[33] This result is consistent with Preis *et al*. (2013), who in a context of trading strategies find asymmetric responses: whereas a fall in the Dow Jones tends to be preceded by a rise in the search volumes, the reverse pattern does not always happen. The manner in which positive or negative feelings relate to investment decisions has been thoroughly studied (see Lucey and Dowling (2005) for a review of this literature). Engle and Ng (1993) introduced the news impact curve, which shows that negative news has long-lasting effects on financial instruments than positive news. The results are far from being stylised, and the access to large amounts of web texts offers new research opportunities.

The last general remark is related to language. Most of the analysis on machine learning and sentiment applied to finance is based on English texts and dictionaries. The extension to other languages is growing: The Europe Media Monitor (European Commission) offers tonality detection in 14 languages, and Bautin *et al*. (2008) describe sentiment analysis on English translations of texts from eight languages. Other examples are Agić *et al*. (2010), Remus *et al*. (2009), Denecke (2008) and Ahmad *et al*. (2006). Chalothorn and Ellman (2013) offer additional literature.

Box 1. ***Examples of Language Classifiers (LC) used in the literature of web mining and stock market prediction***

Opinion Finder (OF, Bollen *et al*., 2011) is dichotomous and suitable for such moods as decreasing/increasing (based upon the share of certain words in a text).
GPOMS: Google Profile of Mood States (Bollen *et al*., 2011, a similar but free-access LC is used by Chen and Lazer, 2011). A model that is more sophisticated than the OF is based on the link between words (each word is weighted according to its relation to the desired mood).
Boosted Decision Trees (Gilbert and Karahalios, 2010, cite more LCs).
Naïve Bayesian classifiers (O'Hare *et al*., 2009; Gilbert and Karahalios, 2010).
Natural Language classifier engines (like Google Trends, and Google I4S – Insights for search in Choi and Varian, 2012).

ANEW (Affective Norms for English Words, Bradley and Lang, 1999), LC for dichotomous feelings (good/bad; active/passive; strong/week), which are used in Dodds and Danforth (2009). Useful for large texts but not for small ones, as it does not take into account the meaning of words in combination (other LCs taking into account combinations are listed in Dodds and Danforth, 2009).

WorldTracker, which it provides a metasearch: the most common keywords used in searching the web. Input: keywords of searches (to define the area); output: the most used terms, usage frequency of the words and number of searches done (Ettredge *et al*., 2005).

SentiWN is based on the English lexical database WordNet. This algorithm offers measures of polarity direction (positive/negative) and polarity intensity (Devitt and Ahmad, 2007).

5. Conclusions and Suggestions for Future Research

The sequence *economic event – investment decisions – stock price jumps* is far from being a stylised fact. Anecdotic comparisons of economic news and ex post movement in aggregated stock prices (Culter *et al*., 1989, updated by Cornell, 2013) claim that the majority of the largest movements in the S&P500 and the CRSP Total Market Index[34] cannot be tied to '*fundamental economic news sufficient to rationalize the size of the observed* [price] *move'* (Cornell, 2013). As stated by Black (1986): '*... people sometimes trade on noise as if it were information'*. We address whether online information is an ingredient of the missing link for this sequence.

This paper revises the literature linking changes in stock returns and trade volumes to measures of web buzz. We highlighted the major novelties and the limitations, and we suggested topics for future research. We observed that although the web can to some extent anticipate financial movements, the gain seldom exceeds 5%. In addition, we touched upon machine learning and sentiment detection to verify that more sophisticated models do not necessarily produce better results. Still, given the amount of money moved daily by the stock markets, even small gains could be worth exploiting, as we believe that the use of the Internet to predict financial movements is a permanent development.

A promising avenue for future research is linked to the analysis of volatility in price changes as a function of the structure and the volatility of web interactions. This research would require a modelling effort coming from the theories of bandwagons, herding behaviour and contagions, and a thorough analysis of web interactions using advanced statistical techniques would also be vital. A good starting point could be Gloor *et al*. (2009) and Ruiz *et al*. (2012). The volatility of asset pricing as a function of agents' mood states or their confidence has been already suggested by Black (1986) and Odean (1998) and Chang *et al*. (2000). In behavioural finance, herding behaviour is analysed by Scharfstein and Stein (1990), who show that for managers (who are worried about their reputations) it could be rational to mimic other managers' investment decisions regardless of any other information.

It would be worth investigating when web buzz has the largest prediction power, and particularly if this power is observed when special events occur. The results of Choi and Varian (2012) and of De Choudhury *et al*. (2008) seem to point in this direction: the explanatory power of web mining increases only after the occurrence of 'big' events (not by chance were the good results of Gilbert and Karahalios obtained with 2008 data). When prices are flat, the contribution of web variables seems to be less compelling (Ettredge *et al*., 2005). The findings so far are consistent with researches in behavioural finance: whereas stocks become overpriced in periods of high sentiment, they become underpriced in periods of low sentiment (Bennet *et al*., 2011, Bennet and Selvam, 2013, but also Baker and Wurgler, 2006, 2007; Lemmon and Portniaguina, 2006). Thus, discriminating events and moods is another promising avenue for research. In spite of the disappointing results, mood extraction seems to also be good business according to the number of new firms selling it.[35]

Most of the literature (both in machine learning and in sentiment detection) is concerned with indicators of the general mood and associates these indicators to general market indicators such as the NASDAQ, S&P500 or Dow Jones. As noticed by Tetlock (2007), this pattern resembles the use of a consumer confidence index as a forward-looking barometer of economic worry. The results of Schumaker and Chen (2009) and Preis *et al*. (2013) call for more sector-specific and industry-specific analyses and ad hoc dictionaries.

Finally, we believe that the most promising and challenging avenue for research is related to the metadata of the communication flow (De Choudhury *et al*., 2008; Ruiz *et al*., 2012). The properties of Internet communications, such as the number of posts and comments, the length and response time, the 1 strength of comments, the frequency of discriminating terms and the extension and dimension of nodes and links, could be useful in analysing bubbles. More generally, these properties could be used to analyse the mechanisms of decision making in financial markets.

Acknowledgements

We thank S. Lechner and anonymous referee for comments and suggestions.

Notes

1. Cutler *et al*. (1989), and more recently Cornell (2013), find that important news items do not seem to explain large movements in the S&P500, as macroeconomic events instead explain these movements.
2. See, for example, the Europe Media Monitor created by the European Commission: http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html.
3. Alanyali *et al*. (2013) calculate the average correlation between the daily mentions of a company's name in the Financial Times and the transaction volumes for the company's stock. Their daily data comprise 31 companies in the Dow Jones in the period from January 2008 to December 2012.
4. Bordino *et al*. (2012) compute the correlation between trading volumes and Yahoo! ticker searches for the set of firms included in the NASDAQ. The daily data are from the period between May 2010 and April 2011.
5. The authors associate returns and transaction volumes to the Google Trend data for each of the companies listed in the S&P500. Their weekly data are from 2004 to 2010.
6. Ruiz *et al*. (2012) correlate the traffic on Twitter to returns and trading volumes for 150 companies in S&P500 with daily data from the first half of 2010.
7. Gilbert and Karahalios (2010) use over than 20 million posts in LiveJournal to construct an Anxiety Index for the US mood and find a relationship between the peaks of this Index and the drops of the S&P500.
8. Bollen *et al*. (2011) use a measure of general mood obtained from Twitter to match the behaviour of daily price movements of the Dow Jones.
9. To the best of our knowledge, only De Choudhury *et al*. (2008) deal with out-of-sample observations, confirming that web information is useful in predicting stock returns (the same results are obtained by Karabulut, 2013).
10. Antweiler and Frank (2004) correlate 1.5 million messages on Yahoo! Finance and Raging Bull with stock returns corresponding to 45 companies on the Dow Jones. They use intraday data from 2000.
11. Wysocki (1998) relates the amount of messages (from the Yahoo! Message Board) regarding over 3000 firms and their corresponding stock revenues and traded volumes in the NYSE. The daily data are from January 1998 to August 1998.
12. Tetlock (2007) relates Wall Street Journal news to Dow Jones daily returns with data from January 1984 to September 1999.

13. Tumarkin and Whitelaw (2001) analyse investor opinions (from Raging Bull) and the returns and trading volumes of 73 stocks quoted on the NYSE. Time-series analysis is limited to a subset of 50 firms. The daily data are from April 1999 to February 2000.
14. Antweiler and Frank (2004) correlate 1.5 million messages on Yahoo! Finance and Raging Bull with stock returns corresponding to 45 companies on the Dow Jones. They use intraday data from 2000, and they show an association between web information and market volatility.
15. De Choudhury *et al*. (2008) compare several features of posts and comments in financial blogs (their daily data are from January to November 2007) with returns for four individual stocks and for the NASDAQ.
16. The gain goes from 11% for the prediction of the magnitudes of fluctuations to 15.6% for the trend. Still, the prediction error remains at 22% for the magnitude and at 13.4% for the trend.
17. With the same trading strategy, Moat *et al*. (2013) analyse the prediction power of Wikipedia views for the companies listed on the Dow Jones. They find an average log return equals a return of 0.5 (0.0002 is the return of the random strategy).
18. Schumaker and Chen (2006) investigate financial news articles and stock quotes of the S&P500. They use intraday data from October 26 to November 28, 2005. The same set of data is used in Schumaker and Chen (2009).
19. Lavrenko *et al*. (2000a,b) relate 38,469 news articles for 127 stocks to stock prices using data from October 1999 to February 2000.
20. Mittermayer (2004) investigates over 6000 press release about companies in the USA National Market System with intraday data from January to December 2002.
21. Gidófalvi (2001) obtain their data from Lavrenko *et al*. (2000a,b). Gidófalvi and Elkan (2003) use minute-by-minute price information and news about the top 30 companies in the Dow Jones. Their data are from July 26, 2001 to March 16, 2002.
22. The authors relate all of the second-to-second stock trades on the NYSE with the instant messages of 66 traders using data from the period September 2007–February 2009.
23. See Chalothorn and Ellman (2013), Liu (2012) and Pang and Lee (2008) for a general overview on opinion mining and sentiment analysis.
24. The difference between machine learning and sentiment analysis is artificial. If sentiment problems can be presented as classification problems, they can be analysed with machine learning (Ren *et al*., 2013).
25. Koppel and Shtrimberg (2006) construct an algorithm to extract binary tonality from a Multex Significant Developments corpus and relate tonality to the company's prices on the S&P500. They use data from 2000 to 2002.
26. Tulankar *et al*. (2013) derive the binary mood from different web sources (blogs, news, research reports, etc.) and find that brokers' predictions are similar to those obtained via sentiment analysis (with an accuracy of 87.5%). Then, positive feelings are associated to upward trends in closing EOD prices for the Indian Market. The data are from November 2012.
27. Zhai *et al*. (2011) use articles from The New York Times Annotated Corpus and stock returns of S&P500 companies. Daily data from January 1987 to June 2007 are used.
28. Das and Chen (2001) compare several text classifiers and examine their ability to extract messages with positive/negative feelings.
29. Tumarkin and Whitelaw (2001) show that the tonality of web messages from Raging Bull does not predict industry-adjusted returns or an abnormal trading volume of 73 stocks. They use daily data from April 1999 to February 2000.
30. Gloor *et al*. (2009) use a semantic social network analyser (Condor) to classify messages according to a binary tonality and investigate the dynamics of web interactions in online news sites, company websites, blogs and forums. They use daily stock data (21 stocks) from April to October 2008.

31. Zhang *et al*. (2010) summarise in a binary tonality different measures of general mood (*hope happy, fear, worry, nervous, anxious, upset*) using Twitter. They test the (in-sample) prediction ability of a Twitter Volatility Index with respect to several general market indices (Dow Jones, NASDAQ, S&P500, VIX). The data are from March to September 2009.
32. Bollen *et al*. (2011) analyse daily Twitter posts and Dow Jones movements from the period March–December 2008.
33. From an analysis of 20 years of NY Times headlines, Niederhoffer (1971) and Akhtar *et al*. (2012) find that markets react to news with a tendency to overreact to bad news.
34. The CRSP Total Market Index covers all New York, American and NASDAQ stocks http://www.crsp.com/.
35. See, for example, https://www.stockpulse.de/en/, http://www.sntmnt.com/, http://www.downsidehedge.com/twitter-indicators/ or http://growthintel.com/technical/. Another example is the company FINIF at http://www.finif.com/.

References

Agić, Z., Ljubešić, N. and Tadić, M. (2010) Towards sentiment analysis of financial texts in Croatian. LREC *Proceedings of the European Language Resource Association*. Available at: http://www.lrec-conf.org/proceedings/lrec2010/ (Last accessed 26 November 2013).

Ahmad, K., Cheng, D. and Almas, Y. (2006) Multi-lingual sentiment analysis of financial news streams. *Proceeding of the Conference: Grid Technology for Financial Modelling and Simulation*, Palermo, Italy.

Akhtar, S., Faff, R., Oliver, B. and Subrahmanyam, A. (2012) Stock salience and the asymmetric market effect of consumer sentiment news. *Journal of Banking and Finance* 36: 3289–3301.

Alanyali, M., Moat, H.S. and Preis, T. (2013) Quantifying relationship between financial news and stock market. *Scientific Reports* 3: 3578, http://www.nature.com/srep/2013/131220/srep03578/full/srep03578.html (Last accessed 16 March 2014).

Antweiler, W. and Frank, M.Z. (2004) Is all that talk just noise? The information content of internet Stock Message Boards. *Journal of Finance* 59(3): 1259–1294.

Asur, S. and Huberman, B.A. (2010) Predicting the future with social media. Available at http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf (Last accessed 16 November 2013).

Baker, M. and Wurgler, J. (2006) Investor sentiment and the cross-section of stock returns. *Journal of Finance* 61(4): 1645–1680.

Baker, M. and Wurgler, J. (2007) Investor sentiment in the stock market. NBER Working paper series, 13189. Available at: http://www.nber.org/papers/w13189 (Last accessed 16 November 2013).

Barber, B. and Odean, T. (2001) The internet and the investor. *Journal of Economic Perspectives* 15(1): 41–54.

Bautin, M., Vijayarenu, L. and Skiena, S. (2008) International sentiment analysis for news and blogs. *Proceedings of the Second International Conference on Weblogs and Social Media*, Seattle, Washington, USA.

Bennet, E. and Selvam, M. (2013) The influence of stock specific factors on the sentiment of equity investors: evidence from Indian stock market. *Proceedings of ASBBS (*American Society of Business and Behavioral Sciences*) Annual Conference*, Las Vegas, USA.

Bennet, E., Selvam, M. and Shalin, E.E. (2011) Investors' sentiment measures. *Global Conference on Innovations in Management*, London UK.

Bentley, R.A., O'Brien, M.J. and Brock, W.A. (2014) Mapping collective behavior in the Big-Data era. *Behavioral and Brain Sciences* 37: 63–119.

Black, F. (1986) Noise. *Journal of Finance* 41(3): 528–543.

Bollen, J., Mao, H. and Zeng, X. (2011) Twitter mood predicts the stock market. *Journal of Computational Science* 2: 1–8.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A. and Weber, I. (2012) Web search queries can predict stock market volumes? *PLoS One* 7(7): e40014.

Bradley, M. and Lang, P. (1999) Affective norms for English Words (anew): stimuli, instruction manual and affective ratings. *Technical Report c-1*, Gainesville, University of Florida, USA.

Chalothorn, T. and Ellman, J. (2013) Sentiment analysis: state of the art. *Proceedings of the International Conference on Advances in Computer and Electronics Technology – ACET 2013*, Hong Kong.

Chang, E.C., Cheng, J.W. and Khorana, A. (2000) An examination of herd behavior in equity markets: an international perspective. *Journal of Banking & Finance* 24: 1651–1679.

Chen, R. and Lazer, M. (2011) Sentiment analysis of Twitter feeds for the prediction of stock market movements. CS 229 Machine Learning Final Projects, Autumn 2011. Available at: http://cs229.stanford.edu/projects2011.html (Last accessed 27 November 2013).

Choi, H. and Varian, H. (2012) Predicting the present with Google Trends. *Economic Record* 88: 2–9.

Cornell, B. (2013) What moves stock prices? Another look. *Journal of Portfolio Management* 39(3): 32–38.

Curme, C., Preis, T., Stanley, H.E. and Moat, H.S. (2014) Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences of the United States of America* 111(32): 11600–11605.

Cutler, D., Poterba, J. and Summers, L. (1989) What moves stock prices? *Journal of Portfolio Management* 15: 4–12.

Das, S.R. and Chen, M. (2001) Yahoo! for Amazon: sentiment parsing from small talk on the web (August 5, 2001). *EFA 2001 Barcelona Meetings*. Available at SSRN: http://ssrn.com/abstract=276189 or http://dx.doi.org/10.2139/ssrn.276189 (Last accessed 22 November 2013).

DeBond, W.F.M. and Thaler, R. (1985) Does the stock market overreact? *Journal of Finance* 40: 793–805.

De Choudhury, M., Sundaram, H., John, A. and Seligmann, D.D. (2008) Can blog communication dynamics be correlated with stock market activity? *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia,* Pittsburgh, Pennsylvania, USA.

DellaVigna, S. (2009) Psychology and economics: evidence from the field. *Journal of Economic Literature* 47(2): 315–372.

Denecke, K. (2008) Using SentiWordNet for multilingual sentiment analysis. *Data Engineering Workshop*, ICDEW 2008, IEEE 24th International Conference, Cancun, Mexico.

Devitt, A. and Ahmad, K. (2007) Sentiment polarity identification in financial news: a cohesion-based approach. *Proceedings of the 45th Annual Meeting of the Association of Computer Linguistics* (pp. 984–991), Prague, Czech Republic.

Dodds, P.S. and Danforth, C.M. (2009) Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies* doi:10.1007/s10902-009-9150-9.

Doshi, L., Krauss, J., Nann, S. and Gloor, P. (2009) Predicting movie prices through dynamic social network analysis. *Proceeding Collaborative Innovations Network Conference* (COINs 2009), Savannah, USA.

Engle, R. and Ng, V.K. (1993) Measuring and testing the impact of news on volatility. *Journal of Finance* 48(5): 1749–1777.

Ettredge, M., Gerdes, J. and Karuga, G. (2005) Using web-based search data to predict macroeconomic statistics. *Communications of the ACM* 48(2): 87–92.

Fama, E. (1965) The behavior of stock market prices. *The Journal of Business* 38(1): 34–105.

Fawcett, T. and Provost, F. (1999) Activity monitoring: noticing interesting changes in behavior. *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA.

Gidófalvi, G. (2001) Using news articles to predict stock market movement. Mimeo, Department of Computer Science and Engineering, University of California San Diego, USA.

Gidófalvi, G. and Elkan, C. (2003) Using news articles to predict stock price movements. Technical Report, Department of Computer Science and Engineering, University of California San Diego, USA.

Gilbert, E. and Karahalios, K. (2010) Widespread worry and the stock market. *Fourth International AAAI Conference on Weblogs and Social Media*, (ICWSM), Washington (DC), USA.

Gloor, P., Krauss, J., Nann, S., Fischbach, K. and Schroder, D. (2009) *Web Science 2.0: identifying trends through semantic social network analysis*. 2009 International Conference on Computational Science and Engineering, Vancouver, Canada.

Godbole, N., Srinivasaiah, M. and Skiena, S. (2007) Large scale sentiment analysis for news and blogs. *Proceeding of the First International AAAI Conference on Weblogs and Social Media (ICWSM) 2007*, Boulder, Colorado, USA.

Goel, S., Hofman, J., Lahaie, S., Pennock, D. and Watts, D. (2010) Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences of the United States of America* 107(41): 17486–17490.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. (2005) The predictive power of online chatter. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 78–87), New York, USA.

Gunn, S.R. (1998) Support Vector Machines for classification and regression. Technical Report, Faculty of Engineering Science and Mathematics, School of Electronics and Computer Science, University of Southampton, UK.

Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47(2): 263–291.

Karabulut, Y. (2013) Can Facebook predict stock market activity? Mimeo, Goethe University, Frankfurt, Germany.

Koppel, M. and Shtrimberg, I. (2006) Good news or bad news? Let the market decide. In J.G. Shanahan, Y. Qu, J. Wiebe (eds.) *Computing Affect and Attitude In Text: Theory and Applications, The Information Retrieval Series*, Vol. 20, Chapter 22, Springer, Dordrecht, NL.

Lavrenko, V.M., Schmill, M., Lawrie, D. and Ogilvie, P. (2000a) Mining of concurrent text and time series. *6th ACM SIGKDD International Knowledge Discovery and Data Mining*, Boston, USA.

Lavrenko, V.M., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. and Allan, J. (2000b) Language models for financial news recommendations. *Proceedings of the 9th International Conference on Information and Knowledge Management*, Washington (DC), USA.

Lemmon, M. and Portniaguina, E. (2006) Consumer confidence and asset prices: some empirical evidence. *Review of Financial Studies* 19(4): 1499–1529.

Liu, B. (2012) Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies* 5(1): 1–167.

Loughran, T. and McDonald, B. (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66: 67–97.

Lucey, B.M. and Dowling, M. (2005) The role of feelings in investor decision-making. *Journal of Economic Surveys* 19(2): 211–237.

Malkiel, B.G. (2003) The efficient market hypothesis and its critics. *Journal of Economic Perspectives* 17(1): 59–82.

Mao, H., Counts, S. and Bollen, J. (2014) Quantifying the effects of online bullishness on international financial markets. *ECB Workshop on Using Big Data for Forecasting and Statistics*, Frankfurt, Germany.

McLaren, N. and Shanbhogue, R. (2011) Using internet search data as economic indicators. *Bank of England Quarterly Bulletin* 51(2): 134–140.

Mishne, G. and Glance, N. (2006) Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford University, California, USA.

Mishne, G. and Rijke, M. (2006) Capturing global mood levels using blog posts. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford University, California, USA.

Mittermayer, M.A. (2004) Forecasting intraday stock price trends with text mining techniques. *Proceeding of the 37th Hawaii International Conference on System Sciences*, Hawaii.

Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, H.E. and Preis, T. (2013) Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, 3: 1801, http://www.nature.com/srep/2013/130508/srep01801/full/srep01801.html (Last accessed 6 March 2014).

Moat, H.S., Preis, T., Olivola, C., Liu, C. and Chater, N. (2014) Using Big Data to predict collective behavior in the real world. *Behavioral and Brain Sciences* 37(1): 92–93.

Morstatter, F., Pfeffer, J., Liu, H. and Carley, K. (2013) Is the sample good enough? Comparing data from Twitter's Streaming API with Twitter's Firehose. *Association for the Advancement of Artificial Intelligence* (www.aaai.org), http://www.public.asu.edu/~fmorstat/paperpdfs/icwsm2013.pdf (Last accessed August 2014).

Niederhoffer, V. (1971) The analysis of world events and stock prices. *Journal of Business* 44(2): 193–219.

Odean, T. (1998) Volume, volatility, price and profit when all traders are above average. *Journal of Finance* 53: 1887–1934.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C. and Smeaton, A. (2009) Topic-dependent sentiment analysis of financial blogs. *TSA '09, 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, Hong Kong.

Pang, B. and Lee, L. (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1–2): 1–135.

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

Preis, T., Moat, H.S., Stanley, H.E. and Bishop, S.R. (2012) Quantifying the advantage of looking forward. *Scientific Reports* 2: 350, http://www.nature.com/srep/2012/120405/srep00350/full/srep00350.html (Last accessed 16 May 2014).

Preis, T., Moat, H.S. and Stanley, H.E. (2013) Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports* 3: 1684, http://www.nature.com/srep/2013/130425/srep01684/full/srep01684.html (Last accessed 16 May 2014).

Radford, W., Hachey, B., Curran, J.R. and Milosavljevic, M. (2009) Tracking information flow in financial text. *Proceedings of the Australasian Language Technology Workshop*, ALTW (2009), University of New South Wales, Sydney, Australia.

Remus, R., Heyer, G. and Ahmad, K. (2009) Sentiment in German language news and blogs and the DAX. *Text Mining Services 2009* (pp. 149–158). Leipzig: Leipziger Beitrage zur Informatik, Germany.

Ren, J., Wang, W., Wang, J. and Shaoyi Liao, S. (2013) Exploring the contribution of unlabelled data in financial sentiment analysis. *Proceedings of the XXVII AAAI Conference on Artificial Intelligence*, Bellevue, Washington, USA.

Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A. (2012) Correlating financial time series with micro-blogging activity. In E. Adar, J. Teevan, E. Agichtein and Y. Maarek (eds.). *Proceedings of the Fifth International Conference on Web Search and Web Data Mining* (pp. 513–522). Seattle (WA), USA. ACM 2012 ISBN 978–1–4503–0747–5.

Ruiz-Martínez, J.M., Valencia-García, R. and García-Sánchez, F. (2012) Semantic-based sentiment analysis in financial news. *Proceedings of the 1st International Workshop on Finance and Economics on the Semantic Web* (pp. 38–51). Available at: http://ceur-ws.org/Vol-862/FEOSWp4.pdf (Last accessed 21 October 2013).

Saavedra, S., Hagerty, K. and Uzzi, B. (2011) Synchronicity, instant messaging, and performance among financial traders. *PNAS Early Edition*, S. L. Levin (ed.), Princeton: Princeton University Press.

Scharfstein, D. and Stein, J. (1990) Herd behavior and investment. *The American Economic Review* 80(3): 465–479.

Schumaker, R. and Chen, H. (2006) Textual analysis of stock market prediction using breaking financial news: the AZFinText system. *12th Americas Conference on Information Systems* (AMCIS-2006), Acapulco, Mexico.

Schumaker, R. and Chen, H. (2009) A quantitative stock prediction system based on financial news. *Information Processing and Management* 45(5): 571–583.

Shleifer, A. and Summers, L.H. (1990) The noise trader approach to finance. *The Journal of Economic Perspectives* 4(2): 19–33.

Tetlock, P. (2007) Giving content to investor sentiment: the role of media in the stock market. *The Journal of Finance* 62(3): 1139–1168.

Tobback, E., Daelemans, W., Junqué de Fortuny, E., Naudts, H. and Martens, D. (2014) Belgian economic policy uncertainty index: improvement through text mining. *ECB Workshop on Using Big Data for Forecasting and Statistics*, Frankfurt, Germany.

Tulankar, S., Athale, R. and Bhujbal, S. (2013) Sentiment analysis of equities using data techniques and visualizing the trends. *International Journal of Computer Science Issues* 10(4): 265–269.

Tumarkin, R. and Whitelaw, R.F. (2001) News or noise? Internet message board activity and stock prices. *Financial Analysts Journal* 57: 41–51.

Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M. (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. *Fourth International AAAI Conference on Weblogs and Social Media* (ICWSM), Washington (DC), USA.

Wysocki, P.D. (1998) Cheap talk on the Wweb: the determinants of posting on stock message boards. Working Paper n. 98025, University of Michigan Business School.

Zhang, X., Fuehres, H. and Gloor, P. (2010) Predicting stock market indicators through Twitter 'I hope it is not as bad as I fear'. *Procedia – Social and Behavioral Science*, 2010.

Zhai, J., Cohen, N. and Atreya, A. (2011) Sentiment analysis of news articles for financial signal prediction. *Mimeo*, University of Stanford, USA.