ScienceDirect

Download

Big Data Research
Volume 4, June 2016, Pages 70-74

# Can Twitter Proxy the Investors' Sentiment? The Case for the Technology Sector

Francesco Corea

Show more

Outline | Share Cite

https://doi.org/10.1016/j.bdr.2016.05.001

Get rights and content

## Abstract

The stock market is influenced by several factors, such as macroeconomics, regulatory, purely speculative ones, and many others. However, one of the most relevant and meaningful is the general opinion and the overall investors' sentiment, i.e., what investors think about a certain firm and, as a consequence, about the relative stock. This investors' sentiment is here proxied by the Twitter content, and the study sums up to the recent outbreak of works that exploit sentiment analysis and Twitter data for stock market predictions. The sample analyzed concerns three major technology companies over a two-months period, on a minute basis. Using microblogging activities and a scoring algorithm for each tweet, it was possible to formulate interesting forecasting models identifying a new set of variables and indicators of the stock market future movements. A selection model has been used to implement the study, and the evidences found were encouraging, since it has been possible to draw the conclusion that this new source of data may increase the explanatory power of financial forecasting models. More in detail, it looks like that the average sentiment associated to any tweet is not so relevant as expected in prediction terms, while the posting volume has a greater forecasting power and it could be used to augment the models. Although this kind of analysis are becoming mainstream and quite common, this work represents an interesting case study for the technological sector rather than advancing fundamental new techniques in the field.

Previous Next

## Keywords

Sentiment analysis; Twitter; Microblogging; High-frequency; Stepwise regression

## 1. Introduction

Big Data is becoming nowadays a buzzword used in several contexts and many different ways. Even if it presents controversial and still ambiguous definitions, it is generally identified as a common feature to all these definitions of the

FEEDBACK

presence of a huge variety of high-speed unstructured data. Hence, probably one of the best examples of big data applications is the use of social media and web contents data, that is generally known as sentiment analysis [1]. A manifold spectrum of utilizations of this new source of data has been studied over the last few years: in medical and epidemics contexts for instance [14], or to try to predict the presidential elections [34].

Although medical or political applications have been deeply explored, one of the most prolific fields of research concerned the use of social media for business and financial purposes. So, no matter whether it dealt with movie revenues [25], commercial sales [11], or music albums forecasts [18] from one hand, or with different social networks sources, such as blogging activities [17], stock messages board [2], [21], or web search queries [9] from the other hand, the importance of this new available dataset has grown and it is currently used for trying to predict the future [3].

Nevertheless business and finance in general were under a "social" attack in the last five years and a lot of different works have been implemented (e.g., [29], [24]), a subset of them – the ones that regard the stock markets – have been particularly analyzed. The main instrument was the data coming from Twitter, and it has been extensively preferred to other sources, such as for instance analysts' recommendations [6], or financial news [22], [30], because of the tweets standard length, common language and symbolism, and high availability and variety.

Thus, Bollen et al. in a first place [8], and then others in following works [7], [23], [26], used financial tweets and their associated investors' mood in order to predict the Dow Jones Industrial Average Index. Corea [12] and Corea and Cervellati [13] instead used Twitter data about major technologies companies to predict the Nasdaq-100 movements, while Brown [10] investigated how Twitter user's reputation could affect the stock market, and Oliveira et al. [28] found a positive correlation between the tweets posting volume and the stock market variations.

Although the use of social media data in order to anticipate the stock markets' oscillations is quite new, the idea of exploiting the investor sentiment and financial news to gain a competitive advantage is well established in literature [15], [16]. It has been showed that financial news with negative words [32], [33], or investors' sentiment [4], [5] have a certain degree of prediction power for the stock markets, as well as it is for tactical allocation [19]. Finally, the gap between traditional finance view on the topic and sentiment analysis has been filled by Oh and Sheng [27], Sprenger and Welpe [31], as well as many others mentioned above.

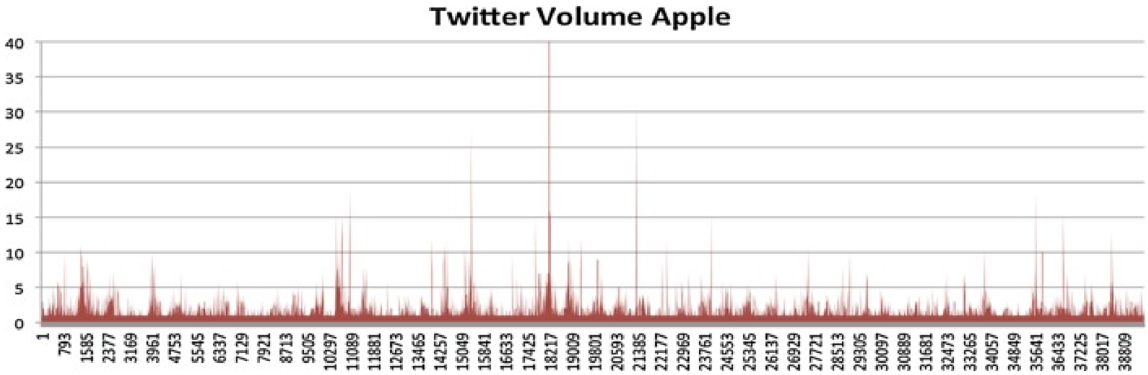
Hence, the purpose of this study is to sum up to the existing literature providing new insights and methods for sentiment analysis forecasting. Using data from three major technology companies over a two-month period, a single high-frequency price-forecasting model will be provided for each of them, as well as a trend one, i.e., whether the prices are experiencing a bullishness or bearishness second by second. The work is then structured as follows: section 2 will deal with the data collection, variables creation, and methodology used, while section 3 will show some results from the analysis implemented. Finally Section 4 will draw some conclusions, suggesting further future improvements for the field of study.

## 2. Data collection and methodology

The data used in the study have been obtained through two different sources: the Twitter one comes from a data provider named DataSift, while the prices for the three stocks have been extracted by Bloomberg. The time period considered spanned over two months from September 24th to November 21st 2014, and only the English tweets regarding Apple, Facebook, and Google have been collected. Other languages represented a minority of tweets and were out of the scope of this analysis, and so there were not considered, while concerning the choice of the companies to analyze, the decision has been driven from two factors: the high presence of tweets on the selected companies, and the existing studies who proved that sentiment analysis works in the technology sector [12], [13]. As the frequency considered, the data were analyzed on a minute basis.

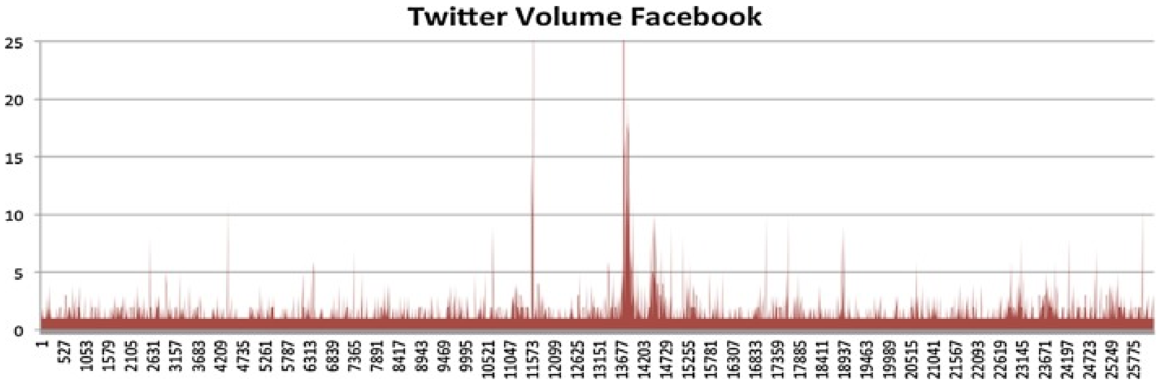
All the noise coming from meaningfulness tweets or information has been depurated taken into account only the tweets posted by individuals with some degree of financial literacy. This has been obtained considering only the tweets that showed the company's ticker, where the presence of the ticker is meant to be a good proxy of individuals' financial knowledge. Hence, overall almost 88,000 thousands of tweets has been gathered for the Apple stock, about 44,000 for Facebook, and less than 32,000 concerning Google.

The Fig. 1, Fig. 2, Fig. 3 illustrate the amount of tweets per minute relatively to each single stock. This gives an idea of the intensity of the microblogging phenomenon, and it could be used in future studies to deepen sentiment analysis with respect to specific tweets-intensive minutes (e.g., reaction to announcements). Furthermore, from the figures can be inferred that there are neither intraday patterns nor seasonality that might bias the results. The pictures also exclude any intuitive correlation in posting activities between stocks so similar. In the period considered, it seems indeed that no event affected all the stocks at the same time and with the same magnitude. In addition, the contagion effect that usually characterizes stock belonging to the same sector or geographic area seems to be missing here.



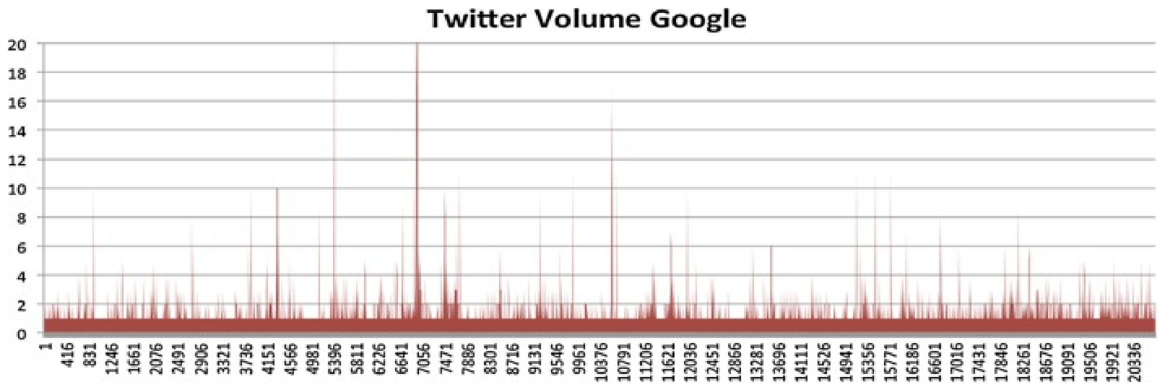
Download : Download high-res image (135KB) Download : Download full-size image

Fig. 1. Amount of tweets posted for each minute about Apple stock.



Download : Download high-res image (128KB) Download : Download full-size image

Fig. 2. Amount of tweets posted for each minute about Facebook stock.



Download : Download high-res image (149KB) Download : Download full-size image

Fig. 3. Amount of tweets posted for each minute about Google stock.

Once the tweets have been extracted, their sentiment was assessed (by the data provider) through an algorithm that scored them with a value ranging from –20 to +20, depending on the strongly negativity or positivity of each tweet's content. A second different score – the klout score – has also been included in the dataset. This is a value that indicates the degree of social influence of certain individual in the social media world, and it varies between 1 and 100 – to a higher value corresponds a higher influence power.

In order to analyze not only the relations with the prices but also those ones with the trend, a set of different variable has been constructed, similarly to what previously observed in [28]:

- *Sentiment Mean* (**SM**): the simple mean of the sentiment score per minute;
- *Sentiment Ratio* (**SR**): the ratio between the Sentiment Mean at $t$ and $t-1$;
- *Bull-Bear Sentiment* (**BBS**) positive/negative: the Sentiment Mean per minute only for positive/negative tweets;
- *Bull-Bear Sentiment Ratio* (**BBSR**): the ratio between the Bull-Bear Sentiment for positive and for negative tweets;
- *Twitter Volume* (**TV**): the volume of tweets at a particular minute $t$;
- *Bull-Bear Volume* (**BBV**) positive/negative: the Sentiment Volume per minute only for positive/negative tweets;
- *Bull-Bear Volume Ratio* (**BBVR**): the ratio between the Bull-Bear Volume for positive and for negative tweets;
- *Twitter Volume 5-minutes Moving Average* (**TVMA**):

$$TVMA_t = \frac{1}{5}\sum_{i=t-4}^{t} TV_i \tag{1}$$

- *Twitter Sentiment 5-minutes Moving Average* (**SMMA**):

$$SMMA_t = \frac{1}{5}\sum_{i=t-4}^{t} SM_i \tag{2}$$

- *Klout Score*: it has been computed the average of the score per day.

The regression models used in order to understand the relations between the prices and the tweets' sentiment are respectively an ordinary least square (OLS) regression and a linear probability model (LPM):

$$\boldsymbol{y}_t = \boldsymbol{x}_t\beta + \epsilon_t$$

$$\boldsymbol{y}_t^* = \boldsymbol{x}_t\beta + \epsilon_t \tag{4}$$

where $\boldsymbol{y}_t^*$ is a latent variable observable only in terms of his sign. In other words:

$$\boldsymbol{y}_t^* = \begin{cases} 0, & (\frac{P_t}{P_{t-1}}) \leq 1 \\ 1, & (\frac{P_t}{P_{t-1}}) > 1 \end{cases} \tag{5}$$

This is one of the differences with respect to the literature so far mentioned: the work studies both the impact of the sentiment on the simple stock price but also on the directional trend that the stock is experiencing, that is whether it is growing or decreasing over the following minute. As it has been noticed, the dummy variable indeed assumes value 1 whether the prices are *up*-moving, while 0 if they are *down*-moving.

Furthermore, instead of selecting by hand which of those variables to be included in the model or testing different models, it has been decided to use a selection model that automatically inserts or excludes a certain variable on the base of a threshold significance level. In this case, the value for a variable to be part of the model is 0.05, while 0.1 for being removed. There are different types of *stepwise regression* model, and here the backward version has been implemented. The backward stepwise regression assumes to estimate the full model with all the explanatory variables in a first place. Then, if the least-significant term is statistically insignificant, it removes that variable and reestimates the model (otherwise it stops). The process is then reiterated. At the same time, for each step, if the most-significant excluded term is statistically significant, it adds that variable back and reestimates the model (otherwise it stops). The algorithm is thus alternatively choosing the least significant variable to drop and to be reintroduced in the model. It is a particular smart and convenient way to select the statistical meaningful variables on the base of pre-fixed significance threshold values without having to deal with each one by hand.

A difference with respect to some previous works is the choice of not taking into account any corporate information at this stage [20], and only considering the information coming directly from Twitter, as well as the stock price.

## 3. Empirical results

Hence, two regressions have been run for each company's stock, one for the price – OLS – and one for trend – LPM. The results are shown in the Table 1.

Table 1. Stepwise variable selection for the high-frequency prices and trends. T-statistics in parentheses.

| | Apple | | Facebook | | Google | |
|---|---|---|---|---|---|---|
| | Price | Trend | Price | Trend | Price | Trend |
| **Price** | 1.000*** | | 1.000*** | | 1.000*** | |
| | (22483.99) | | (4341.11) | | (13475.16) | |
| **BBVR** | 0.00504* | | 0.0141* | | | 0.0463* |
| | (2.01) | | (1.69) | | | (1.72) |
| **BBVn** | 0.00319* | | | | | 0.0850** |
| | (2.18) | | | | | (2.72) |
| **TV** | −0.00218** | | | | | |
| | (−2.97) | | | | | |

| | Apple | | Facebook | | Google | |
|---|---|---|---|---|---|---|
| | Price | Trend | Price | Trend | Price | Trend |
| **Trend** | | 0.395*** | | 0.486*** | | 0.413*** |
| | | (17.36) | | (12.52) | | (5.14) |
| **Klout** | | 0.00836*** | | 0.00862*** | | |
| | | (10.82) | | (9.47) | | |
| **SR** | | 0.00860* | –0.0115* | | | |
| | | (1.99) | (–2.28) | | | |
| **BBSp** | | 0.0160*** | | | | |
| | | (3.68) | | | | |
| **BBSn** | | –0.00739* | 0.00748* | | | –0.0448** |
| | | (–1.65) | (2.47) | | | (–3.21) |
| **BBVp** | | 0.0115* | | | | |
| | | (2.03) | | | | |
| **SMMA** | | | 0.00875* | | | |
| | | | (2.26) | | | |
| **TVMA** | | | | 0.00615* | | |
| | | | | (2.27) | | |
| **BBSR** | | | | –0.0354** | | –0.0666** |
| | | | | (–2.80) | | (–3.11) |

* $p < 0.1$.

** $p < 0.01$.

*** $p < 0.001$.

As it can be observed from the Table 1, the results are quite mixed but encouraging, meaning that there is a single feature/variable that is present in every regression. Nonetheless, some consistency can be noted. The most interesting thing is that the simple sentiment mean has been excluded from each regression, and in general the sentiment, whether it is positive or negative, it is not so relevant every time and for every stock. In general, it is true that the variables where the sentiment was considered in any form have more predictive power in term of stock trend than stock point-forecasting. On the other hand though, the tweets volume seems to have a strong impact both in terms of price forecasting and directional prediction. This suggests that maybe is more valuable how much people talk about a certain stock with respect to what they actually think about it. To confirm this hypothesis, it can be observed that negative sentiments have a negative impact on the stock price – as intuitively should be – while an increase in the posting volume of negative tweets has anyway a positive impact on the stock price.

Moreover, variables that capture the sentiment mean have on average a higher magnitude with respect to the posting volume ones.

A second interesting consideration is that the Klout score is significant in more than one case. Hence, it seems that individuals with a higher influence power within the social media worlds can effectively influence the stock direction with their posts and opinions, although the magnitude is extremely low.

Finally, contrarily to Tetlock [32] and Tetlock et al. [33], negative sentiment causes a downward pressure of a lower magnitude than the upward push of a positive tweet.

## 4. Conclusions

New sources of data are daily used for trying to capture the stock market behavior. One of the currently most used and innovative is without a doubt the social media data. It has been analyzed here how Twitter in particular could be used in financial contexts. Different variables embedding tweets' content sentiment and volume as well have been created and used for price and trend forecasting. Three major technology companies have been studied for a two months period. Consistently with previous studies, as it can be inferred from the extensive survey proposed by Kearney and Liu [20], linear regressions perform often far better than more complicated regressions, and are then used also for the sake of this study. In spite of that, no previous work use selection models such as the stepwise regression, which seems to optimize the variable selection process in the present analysis.

The results provided gave an overview on the kind of insights that can be achieved through microblogging and social media more in general. The results are also quite mixed, probably reflecting structural and specific intrinsic differences for each company, but at the same time they show some degree of consistency and comparability.

Further implementations could be studied in the next future, such as considering longer timeframes, different companies and sectors, or analyzing special situations such as IPO and company's announcements (dividends, etc.). Of particular interest would also be studying the structure of the network who is talking about a certain stock or firm and assess how this affect the company's evaluation on the stock market.

Recommended articles Citing articles (13)

## References

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau
**Sentiment analysis of Twitter data**
LSM '11 Proceedings of the Workshop on Languages in Social Media (2011), pp. 30-38
View Record in Scopus Google Scholar

[2] W. Antweiler, M.Z. Frank
**Is all that talk just noise? The information content of internet stock message boards**
J. Finance, 59 (3) (2004), pp. 1259-1294
CrossRef View Record in Scopus Google Scholar

[3] S. Asur, B. Huberman
**Predicting the future with social media**
Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference, vol. 1 (2010), pp. 492-499
CrossRef View Record in Scopus Google Scholar

[4] M. Baker, J. Wurgler
**Investor sentiment and the cross-section of stock returns**

J. Finance, 61 (4) (2006), pp. 1645-1680
CrossRef View Record in Scopus Google Scholar

[5] M. Baker, J. Wurgler
**Investor sentiment in the stock market**
J. Econ. Perspect., 21 (2) (2007), pp. 129-151
CrossRef View Record in Scopus Google Scholar

[6] B. Barber, R. Lehavy, M. McNichols, B. Trueman
**Can investors profit from the prophets? Security analyst recommendations and stock returns**
J. Finance, 56 (2) (2001), pp. 531-563
CrossRef View Record in Scopus Google Scholar

[7] J. Bollen, H. Mao
**Twitter mood as a stock market predictor**
IEEE Comput., 44 (10) (2011), pp. 91-94
CrossRef View Record in Scopus Google Scholar

[8] J. Bollen, H. Mao, X. Zeng
**Twitter mood predicts the stock market**
J. Comput. Sci., 2 (1) (2011), pp. 1-8
Article Download PDF View Record in Scopus Google Scholar

[9] I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen, I. Weber
**Web search queries can predict stock market volumes**
PLoS ONE, 7 (7) (2012), Article e40014
CrossRef View Record in Scopus Google Scholar

[10] E.D. Brown
**Will Twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market**
SAIS 2012 Proceedings (2012)
Paper 7
Google Scholar

[11] H. Choi, H. Varian
**Predicting the present with Google trends**
Special Issue: Selected Papers from the 40th Australian Conference of Economists, Econ. Record, 88 (Issue Supplement 1) (2012), pp. 2-9
CrossRef View Record in Scopus Google Scholar

[12] F. Corea
**Why social media matters: the use of Twitter in portfolio strategies**
Int. J. Comput. Appl., 128 (6) (2015), pp. 25-30
CrossRef View Record in Scopus Google Scholar

[13] F. Corea, E.M. Cervellati
**The power of micro-blogging: how to use Twitter for predicting the stock market**
Eurasian J. Econ. Finance, 3 (4) (2015), pp. 1-7
CrossRef View Record in Scopus Google Scholar

[14] A. Culotta
**Towards detecting influenza epidemics by analysing Twitter messages**
Proceedings of the First Workshop on Social Media Analytics (2010), pp. 115-122

CrossRef View Record in Scopus Google Scholar

[15] Z. Da, J. Engelberg, P. Gao
**In search of attention**
J. Finance, 66 (5) (2012), pp. 1461-1499
Google Scholar

[16] Z. Da, J. Engelberg, P. Gao
**The sum of all FEARS investor sentiment and asset prices**
Rev. Financ. Stud., 28 (1) (2015), pp. 1-32
CrossRef View Record in Scopus Google Scholar

[17] M. De Choudhury, H. Sundaram, A. John, D.D. Seligmann
**Can blog communication dynamics be correlated with stock market activity?**
Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia (2008), pp. 55-60
CrossRef Google Scholar

[18] V. Dhar, E.A. Chang
**Does chatter matter? The impact of user-generated content on music sales**
J. Interact. Mark., 23 (4) (2009), pp. 300-307
Article Download PDF View Record in Scopus Google Scholar

[19] K.L. Fisher, M. Statman
**Investor sentiment and stock returns**
Financ. Anal. J., 56 (2) (2000), pp. 16-23
CrossRef View Record in Scopus Google Scholar

[20] C. Kearney, S. Liu
**Textual sentiment in finance: a survey of methods and models**
Int. Rev. Financ. Anal., 33 (2014), pp. 171-185
Article Download PDF View Record in Scopus Google Scholar

[21] J.L. Koski, E.M. Rice, A. Tarhouni
**Day trading and volatility: evidence from message board postings in 2002 vs. 1999**
Working paper under review by Management Science
(2008)
Google Scholar

[22] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan
**Language models for financial news recommendation**
Proceedings of the Ninth International Conference on Information and Knowledge Management (2000), pp. 389-396
View Record in Scopus Google Scholar

[23] H. Mao, J. Bollen, S. Counts
**Predicting financial markets: comparing survey, news, Twitter and search engine data**
Working Paper
(2011)
Google Scholar

[24] H. Mao, S. Counts, J. Bollen
**Quantifying the effects of online bullishness on international financial markets**
ECB Stat. Paper Ser., 9 (2015)

Google Scholar

[25] G. Mishne, N. Glance
**Predicting movie sales from blogger sentiment**
AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (2006)
Google Scholar

[26] A. Mittal, A. Goel
**Stock prediction using Twitter sentiment analysis**
Working Paper Stanford University CS 229
(2012)
Google Scholar

[27] C. Oh, O.R.L. Sheng
**Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement**
ICIS 2011 Proceedings (2011)
Google Scholar

[28] N. Oliveira, P. Cortez, N. Areal
**On the predictability of stock market behaviour using StockTwits sentiment and posting volume**
Progress in Artificial Intelligence, Lecture Notes in Computer Science, vol. 8154 (2013), pp. 355-365
CrossRef View Record in Scopus Google Scholar

[29] E.J. Ruiz, V. Hristidis, C. Castillo, A. Gionis
**Correlating financial time series with micro-blogging activity**
Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (2012), pp. 513-522
CrossRef View Record in Scopus Google Scholar

[30] R.P. Schumaker, H. Chen
**Textual analysis of stock market prediction using breaking financial news: the azfin text system**
ACM Trans. Inf. Syst. (TOIS), 27 (2) (2009), p. 12
Google Scholar

[31] T. Sprenger, I. Welpe
**Tweets and trades: the information content of stock microblogs**
Social Science Research Network Working Paper Series: 1–89
(2010)
Google Scholar

[32] P.C. Tetlock
**Giving content to investor sentiment: the role of media in the stock market**
J. Finance, 62 (3) (2007), pp. 1139-1168
CrossRef View Record in Scopus Google Scholar

[33] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy
**More than words: quantifying language to measure firms' fundamentals**
J. Finance, 63 (2008), pp. 1437-1467
CrossRef View Record in Scopus Google Scholar

[34] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe
**Predicting elections with Twitter: what 140 characters reveal about political sentiment**
Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)

Google Scholar

View Abstract