Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

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

Francesco Corea *

Department of Economics and Finance, LUISS Guido Carli University, Viale Romania 32, 00197, Rome, Italy

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form 29 December 2015
Accepted 30 May 2016
Available online 3 June 2016

Keywords:
Sentiment analysis
Twitter
Microblogging
High-frequency
Stepwise regression

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.

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 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 advan-

* Tel.: +39 07916 965644.
E-mail address: fcorea@luiss.it.

http://dx.doi.org/10.1016/j.bdr.2016.05.001
2214-5796/© 2016 Elsevier Inc. All rights reserved.

tage 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 Figs. 1–3 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.

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):

$$\boldsymbol{y}_t = \boldsymbol{x}_t\beta + \epsilon_t \tag{3}$$

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

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 signifi-