Exploring Sentiment on Financial Market Through Social Media Stream Analysis

Francesco Bellini and Nicola Fiore

Abstract The aim of this chapter is to present the prototype developed in the TrendMiner project in the financial domain. TrendMiner is a Research & Development project co-funded by the European Commission under the 7th Framework Programme contract nr. FP7-ICT-287863. We developed a web-based prototype summarising the media stream in terms of its likely impact on a selected financial asset from economic and political-economic perspectives. The platform is able to gather the events occurring along the social media timeline and to build a tailored visualisation/summarisation of these data with price movements of a given stock or index. The results of the prototype have been evaluated and summarised in this chapter, and three examples are used as proof of concepts for validating the prototype outcomes against the known market behaviours and the existing literature. The TrendMiner financial use case prototype shows its ability to play as another decision support tool besides the consolidated market forecast techniques such as technical and fundamental analysis.

Keywords Finance • News • Tweets • Social media • Stock • Markets • Sentiment • Corporate

1 Introduction

One of the most important research streams in finance is to understand the determinants of stock market dynamics. According to the theory of efficient financial markets [1], stock prices accurately reflect the whole public information available at all times. Hence, the stock prices adjust according to the new information almost instantaneously, so no extra returns can be achieved by trading on that. Fama and other later authors then asserted that the stock price moves at the time of information release and neither before nor after is possible to have extra returns.

F. Bellini (✉) • N. Fiore
Eurokleis, Via Romeo Romei, 27, Rome, RM 00136, Italy
e-mail: francesco.bellini@eurokleis.com; nicola.fiore@eurokleis.com

K. Corsi et al. (eds.), *Reshaping Accounting and Management Control Systems*, Lecture Notes in Information Systems and Organisation 20,
DOI 10.1007/978-3-319-49538-5_8

Subsequent studies suggest that the content of the news could influence the way people behave. Some of these works [2–4] have shown that positive and negative emotions affect in a different way individual cognition and behaviour, since “bad” news have more impact than “good” ones and bad information is processed more thoroughly than good. According to Tetlock et al. [5, 6], the fraction of negative words in firm-specific news stories forecasts low firm earnings, and firms’ stock prices underreact to the information embedded in negative words. The findings suggested that linguistic media content captures otherwise hard-to-quantify aspects of firms’ fundamentals, which investors quickly incorporate in stock prices. Schumaker and Chen [7] show that adding textual features of news can improve the forecasting accuracy of a stock prediction system.

Compared to the traditional press, Internet is a communication channel that broadcasts news much faster and enables the exchange of information at approximately zero cost. Bagnoli, Beneish and Watts [8] find that whisper forecasts (unofficial forecasts of earnings per share that circulate among traders and investors) are, on average, more accurate than First Call forecasts and constitute better proxies for market expectations of earnings than the First Call forecasts. Recent studies [9, 10] have been further concentrated in web search data showing that search volumes of stock names reveal investor attention and interest, and high search volumes thus predict higher stock prices in the short term and price reversals in the long term.

Social media feeds are becoming an important source of data to extract streams of information that could influence the investor behaviour. In the past 5 years, new contributions have shown that the information extracted from social network such as LiveJournal [11], Facebook [12] and Twitter [13] may be correlated with stock indices like the Dow Jones Industrial Average and further used to predict stock market fluctuations. This can be achieved through mood indicators resulting from the analysis of text supplied by social media and expressed in a time series format.

This chapter summarises the achievements of the TrendMiner project in the context of its financial use case (the other one was on politics) that actually started from the recognition of these early attempts. In order to contribute to this specific research field, the TrendMiner financial use case aimed at investigating any link or relationship between financial market and investors’ sentiment derived from text mining. As a proof of concept, three different examples are analysed which are discussed in the following paragraphs.

2 TrendMiner Architecture and Components

TrendMiner provides a platform for cross-lingual text mining and summarisation of large-scale stream media. The platform was developed through an interdisciplinary approach, combining deep linguistic methods from text processing, knowledge-based reasoning from web science, machine learning, economics/finance and political science.

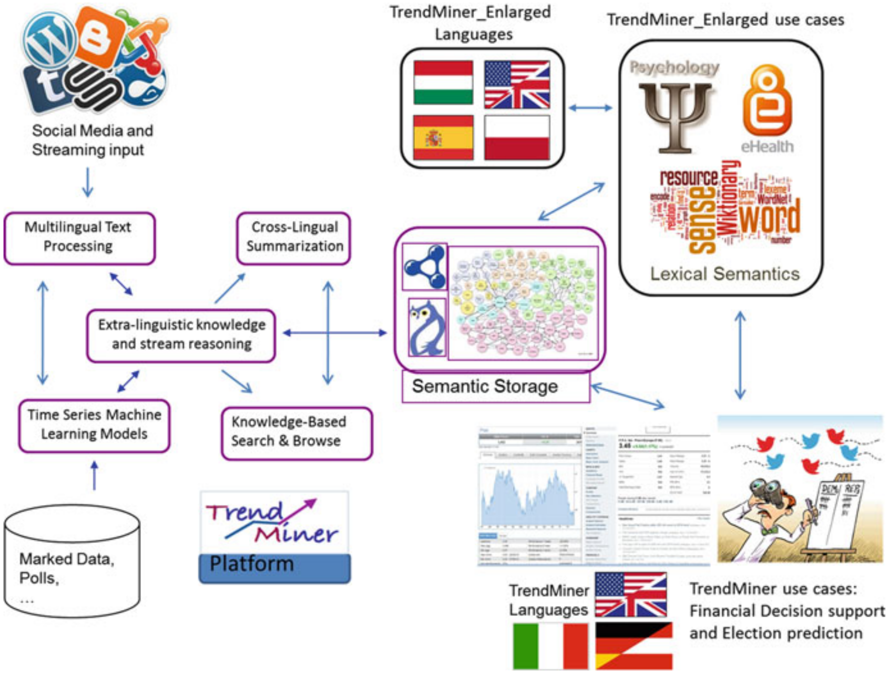

Fig. 1 TrendMiner platform

The platform covers all the phases of the social media stream processing life cycle: large-scale data collection, multilingual information extraction and entity linking, sentiment extraction, trend detection and summarization and visualisation (Fig. 1).

The high-level representation of the architecture identifies the major functional blocks of the system covering all the phases of the social media data processing chain: data collection, deployment of different analysis and transformations and summarisation and visualisation.

The data collection is supported by two complementary components that are responsible for monitoring different social media sources. The processing services are classified into two groups depending on their usage in the processing life cycle. The first group performs the resources pre-processing (metadata extraction) regardless of any context of usage, namely entities detection, language identification and geo-location detection. The results produced by these components are stored in the data repository and serve as a base for subsequent data searching and browsing. The second group of services computes collective analytical information based on user-defined context and resources selection. Finally, a presentation layer service provides an abstraction over the actual data prior to its presentation to the financial (and also political) analyst (end user).

3 Processing of Information in the Financial Domain

The process for analysing information within the financial domain considers the following phases:

- Data Collection
- Computation of polarity, annotation and sentiment
- User interface and examples of market investigations

3.1 Data Collection

Data collection was implemented with reference to the information coming from Twitter accounts, news and financial markets. The activities described in the following paragraphs have been carried out in order to isolate a consistent number of meaningful resources needed for the use case implementation.

Tweets Selection and Collection In the first stage of TrendMiner, we drove the tweets collection through specific keywords. The keywords were manually extracted from news that generated "rumours" during an observed period. Keywords allow to build a detailed sample of tweets that isolates the comments on observed companies, persons and products.

In the final version, we reversed the approach: instead of collecting the tweets that match certain keywords, we collect all the tweets originated from selected financial sources, and in the second stage we carry out the search for specific keywords. In this way, we do not restrict the collection only to the topics related to some rumours, but we collect whatever is generating interest in the financial domain and we build a dataset to be further investigated afterwards.

The identification of sources has been a time-consuming task and was carried out by following the approach below:

- Worldwide search on Twitter accounts accessed by people who share ideas on investments and global economy;
- Identification, within each account, of a list of users focused on specific issues of economics and finance. A preference was given for the Twitters acquisition. Our search was aimed to expand as much as possible the number of users who share investment ideas, and for this objective we visited each account individually;
- In the financial domain, English is the language primarily accessed worldwide, so the majority of the selected lists use this language. However, some lists in German and in Italian language were selected too.

By using the above-described approach, more than 3000 users were identified; more than 2000 of these are individual users and about 1000 are users operating within news providers. We started to collect the Tweets selected as above since

October 2013 and we synchronised our collection with the data warehouse of the TrendMiner platform. The dataset is now around 3 GB with more than 2.5 million of tweets representing a good starting point for our investigations. A key point for future development will be its maintenance and improvement.

In order to perform the collection of Tweets, we used the APIs provided by Twitter that allows to search within the timeline according to the keywords provided. A python routine was implemented by using these APIs. The routine creates the connection, performs the authentication and runs the search on the Twitter timeline. The routine was scheduled twice a day for the same period of collection described. During this period, the keyword list has been daily refined according to the trending topics in order to refine the selection of Tweets that will concur to the sentiment definition.

News Collection and Analysis The news documents are collected from selected sources that are supposed to provide "price-sensitive" news. The information has been crawled according to the sources identified by the financial research team and to a specific crawling strategy.

Financial Instruments Data Collection In order to combine sentiment and market prices, we have designed and implemented a structure to collect market data with the following features:

- Sentiment can be correlated not only with equity prices but also with currencies, bonds, common funds, exchange trade funds (ETF), etc. This means that the data structure shall allow the detection of each kind of prices that we generalised through the Financial Instrument (FI) concept. In this way, we can access prices and volumes (if any) of any type of market data.
- The structure is flexible in terms of time period and allows to investigate the links between sentiment and market data on monthly, weekly, daily and infra-daily basis.

The prototype focuses only on stock prices and stock indices gathered daily; however, more refined investigations could be conducted in the future.

3.2 Computation of Polarity, Annotation and Sentiment

The following step is the evaluation of entity polarity [14, 15] using the sentiment data for a specific single day. The polarity is the sentiment associated with the entity referring to the keywords list and can be positive, very positive, negative and very negative. We give a score to each word depending on the category of polarity it belongs to and increase/decrease the polarity strength when a word is preceded by a modifier, i.e.: not good = -1; good = +1; very good = +2. For this purpose we developed a tool that reads shorts notices associated with actual values of companies listed at the MIB (stock exchange in Milano) and, using a heuristic combining repeated word in the context of an increase or a decrease of the values, we generated an opinionated lexicon for the Italian language (in the financial domain). This lexicon has been merged with a "classical" computational lexicon of Italian, and it was being manually checked against longer texts, also in different domains.

On the polarity data, we also consulted the list of words deemed to be positive or negative, as they are defined in the subjectivity lexicon of Loughran and McDonald Financial Sentiment Dictionaries [16]. This is for the English text, but we also provided for Italian translation. We are currently implementing a sentence level polarity computation for Italian newspapers and porting the strategy to Tweets.

The output of the collection is a list of *json* arrays containing the Tweets. This format is processed in a pipeline approach by a series of language processing tools (Gorrell et al. [17], Preotiuc et al. [18]) to provide tokenisation, language detection, annotation and sentiment analysis. The processed data are then aggregated over a time period (Samangooei et al. [19]) to produce features suitable for describing movements in a time series, e.g. word frequencies or aggregate sentiment relating to a given person, party or company.

4 User Interface and Examples of Market Investigations

The TrendMiner user interface allows to show time-based sentiment and activity on a particular topic of interest and compares them visually with the time series of a financial instrument (price, index, etc.). Since the user interface has been designed at a prototype level and permits only to show few changes in sentiment over time, we can display daily values for max 1 month. If we go over a longer period, then we have to deal with weekly readings. In addition, the interface in the research programme has been designed and implemented to be unique for all the TrendMiner use cases that cope with finance, politics and health. For this reason, it does not include sophisticated features and quantitative tools specific for the financial domain. It allows to investigate financial markets and investor sentiment at the "explorative" level, but this permits in any case to get very interesting results. The following figures show how the user interface works according to the available components (Fig. 2).

4.1 Social-Economy

As a first example, we choose to investigate a social-economy situation creating a specific track in the system (with the parameters shown in Fig. 3) to which corresponds the entity distribution shown in Fig. 4.

As a first result, we can see that entities like Draghi, Yellen, growth, Europe and USA are more mentioned than topics like jobless, unemployment, deflation and recovery. We see also that Draghi and Yellen are mentioned when they approach

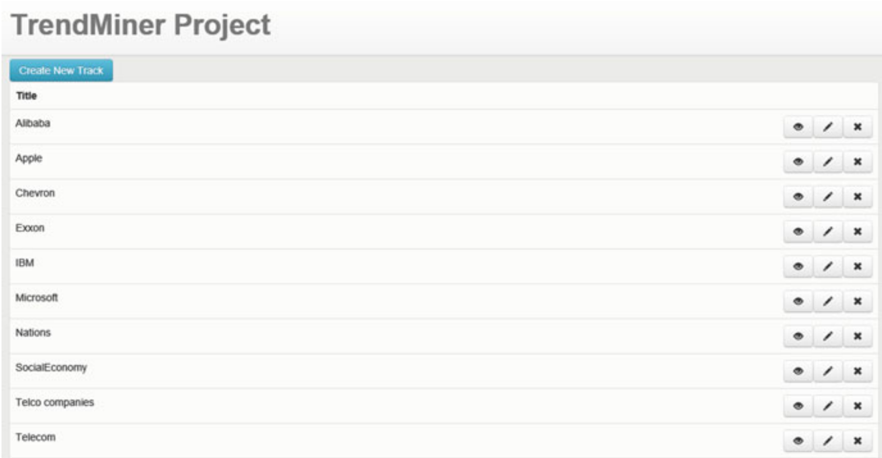

Fig. 2 Cockpit for track management

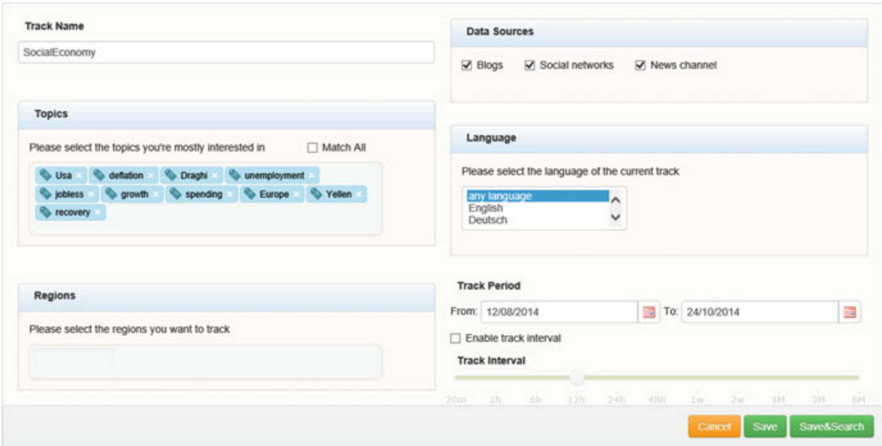

Fig. 3 Track set-up

their periodic speech, and, apart from these events, the interest of the financial community is less evident. Instead, the topic growth is evident over all the observed period, and this applies to the topic Europe that is more evident than USA.

We also tried more topics like austerity, claims, households and others, but their influence is low if compared with the above terms, and we experienced that lower activities have low impact over the sentiment readings. This first screening is important in order to establish which terms should be monitored over the time (they could be the topics with more weight on the sentiment changes) and which terms should be discarded instead.

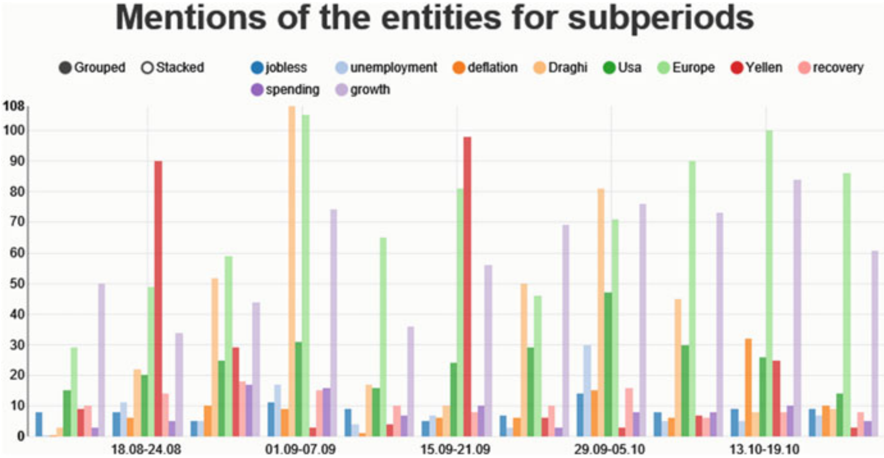

Fig. 4 Number of entities for each topic

The sentiment is then reported in Fig. 5. Here the average value is 0.24; an increasing trend is observed in the first period with a peak in the middle of September. After that the sentiment drops. A possible interpretation of this pattern could be an increasing expectation from Draghi and Yellen speeches of mid-September and a decreasing interest after this event. The FTSE100 time series (Fig. 6) seems to be in line with the sentiment.

The following figures show how the user interface works according to the available components. The cockpits in Figs. 7 and 8 show how a query is built according to topics, data source, locations, language and time period.

4.2 Initial Public Offering (IPO)

As a second example, we investigated the IPO of Alibaba Group that took place on 19th September 2014. The entities chart of Fig. 9 shows a high interest around the IPO date. We compared Alibaba vs some peers, in order to see if some discontinuity happens in the period. As can be seen, when the listing is over all the peers seem to have the same mood (as it is logic), and this is an important result in terms of reliability of our dataset.

The sentiment chart (Fig. 10) shows an average value of 0.47 that is higher than the average of the social-economy case. On the other hand, the trend seems the opposite to the one of the stock price, meaning that sentiment roughly increases when stock price decreases.

The stock price (Fig. 11) decreases immediately after the listing date and increases at the end of the observed period. During the days close to the listing date, activities are driven by Alibaba's topic while the sentiment strongly decreases

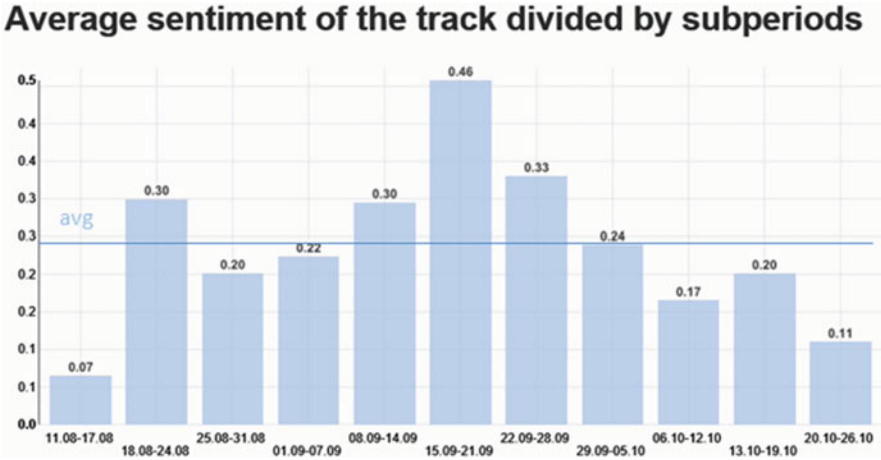

Fig. 5 Sentiment pattern

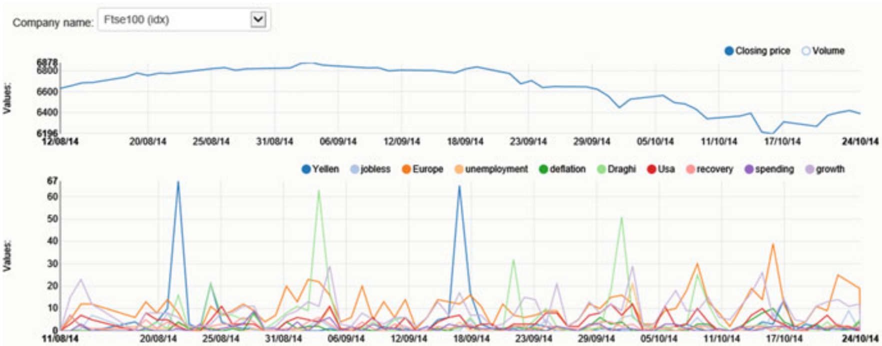

Fig. 6 Topics activity and FTSE100 time series

the day after the listing. This behaviour may be expected as the attention of investors typically drops right after an important event (Simon [20]).

4.3 *New Product Announcement*

As a third example we analysed the Apple stocks around the date of the iPhone 6 announcement. The topics have been chosen by selecting the terms most related to this stock (Fig. 12). It shows the weights of each entity during the selected period and that iPhone 6 is the one most relevant. Figures 13 and 14 show sentiment, activity and stock price.

Fig. 7 Co-occurrence of entities

The chart of activity combined with the stock price shows a peak in the activity around the 8th September 2014. This corresponds, on one hand to a negative change in sentiment, while on the other hand to a positive reverse in the stock price. This is an interesting configuration to be examined in order to decide for a long position on the stock. The same also happens from the third to the fourth week, even if the peaks and the increase of stock price are less evident.

This trend can be explained by the fact that the iPhone 6 announcement was seen by the market as the most significant driver of the stock price for a long time. A reduction of the sentiment at the same time of the peak in activity is considered by some traders/investors as an acceptable behaviour. This may happen (Simon [20]) because a greater amount of information is exchanged among the actors when a new product (the iPhone 6) is expected to drive the future stock price. In general, during this phase, expectation and activity increase around the new factor and decrease

Fig. 8 Topics, word cloud and source indications

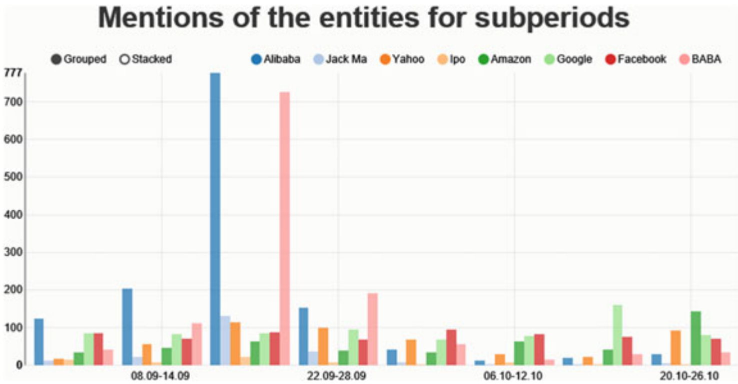

Fig. 9 Alibaba entity distribution

when more details are known and the decision investment has been taken. Another theory, which leads to the same conclusion, considers the investor sentiment as a contrarian indicator (Thorp [21]) that foresees a bullish market when the sentiment reaches low values and vice versa.

Of course, this is not a unique strategy for a long position in the stock. Strategies for short positions could be also detected. As a result, the trader/investor can base his strategy not only considering the market price but also with the information provided by the text mining.

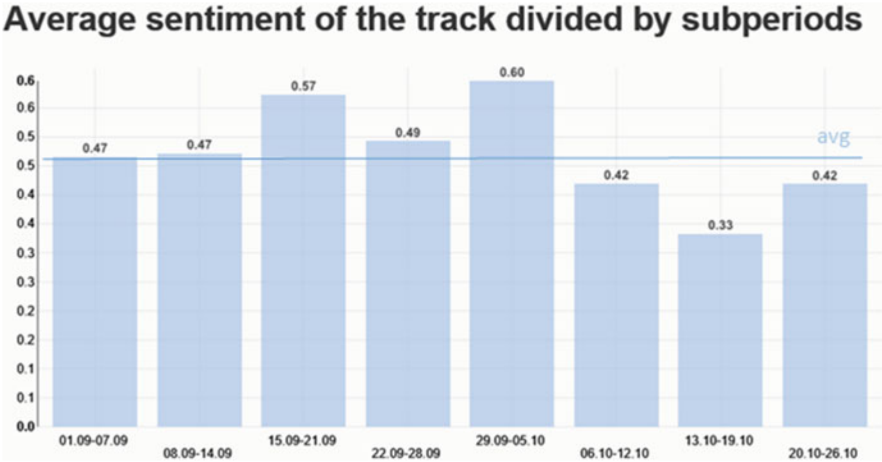

Fig. 10 Alibaba sentiment pattern

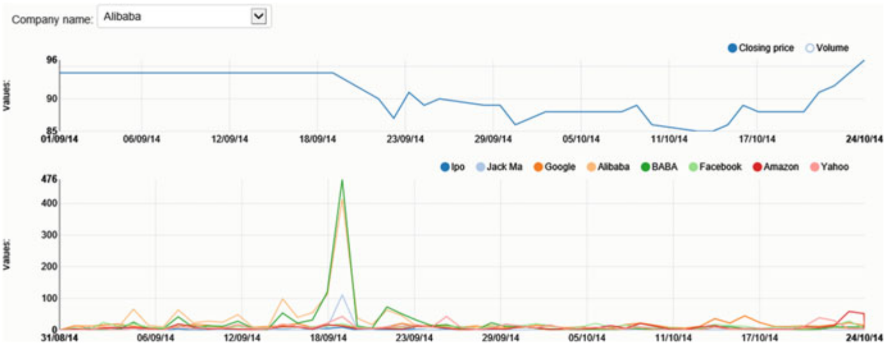

Fig. 11 Alibaba activity and stock exchange price

5 Conclusions and Future Developments

From the three examples analysed, we can draw the following considerations.

Sentiment can be considered as an additional source of information to drive investment decisions. This can be used together with the consolidated tools of technical analysis. However, at this stage quantitative approaches for the computation of the sentiment are still missing. It is not yet clear whether the absolute value of a sentiment can be associated with a bullish or to a bearish market. Contrarian strategy is one of the most used in the market and helps to discover situations in which extremely bullish or extremely bearish configurations happen in order to decide to go long or short in investment. Actually, these configurations are

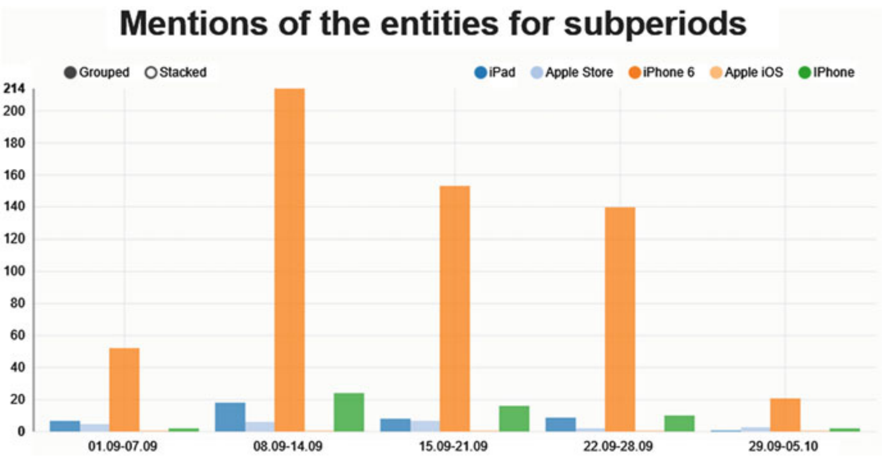

Fig. 12 Number of entity mentions in a period

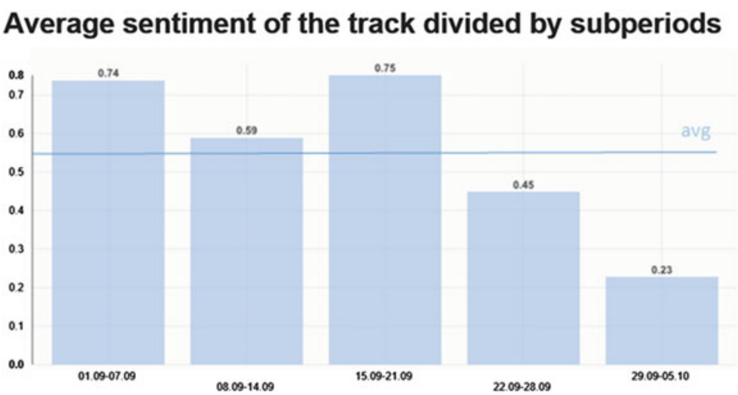

Fig. 13 Sentiment of the track

investigated through mathematical and statistical tools, but a second source value coming from moods and opinions could be of great influence.

Of course, either the visual links or the quantitative readings will lead in any case to subjective valuations for investments, because the decision of which amount the sentiment must change in order to shift from a bullish configuration to a bearish one is very subjective. This stresses the concept that a decision support system is the most valuable aid for investment choices.

Given these considerations, our examples of market investigations assess some important points:

- The validity of underlying dataset. All the examples, although not similar in the content, have found feedback and compliance in the system. This means that this dataset must be maintained and improved.

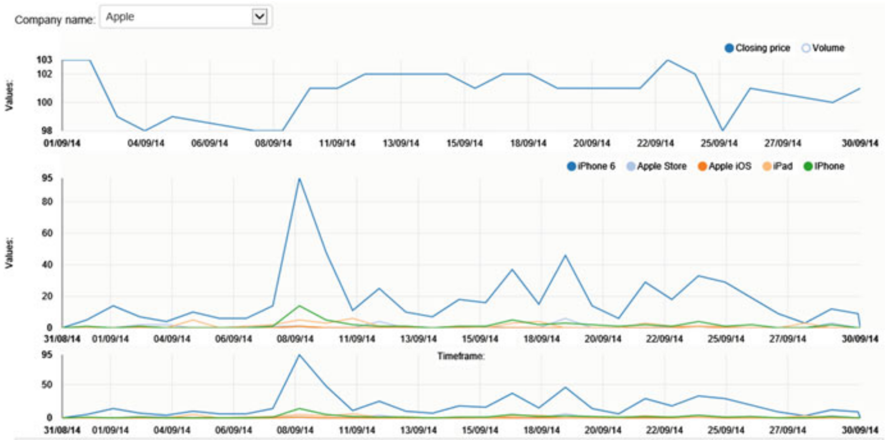

Fig. 14 Stock market and Twitter data in the time frame

- Some topics must be identified and observed during the time. This helps to find out average values and quantitative changes in sentiment. These topics must be started and investigated on financial instruments with a consistent "volume" of moods around them, which are primarily the most important indices, and secondly the stocks with high volume exchanged.
- An investigation must be done about the forecasting power of the sentiment when quantitative values will arise. In our examples, we used daily prices but we need to investigate if the sentiment indicator would be more appropriate to forecast weekly or infra-day market values or if long daily time series are necessary to analyse the next infra-day behaviour.

References

1. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance, 25*(2), 383–417.
2. Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370.
3. Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. New York: McGraw-Hill.
4. Brief, A. P., & Motowidlo, S. J. (1986). Prosocial organizational behaviors. *Academy of Management Review, 11*, 710–725.
5. Tetlcok, P. C., Saar-Tsechansky, M., & Mackassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance, 63*(3), 1437–1467.
6. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.
7. Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news. *ACM, 27*(2), 1–19.
8. Bagnoli, M., Beneish, M. D., & Watts, S. G. (1999). Whisper forecasts of quarterly earnings per share. *Journal of Accounting and Economics, 28*(1), 27–50.

9. Da, Z., Engelberand, J., & Gao, P. (2011). In search of attention. *Journal of Finance, 66*, 1461–1499.
10. Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google trends. *Nature UK Scientific Reports, 3*, 1684.
11. Gilbert, E., & Karahalios, K. (2010). *Widespread worry and the stock market*. International AAAI Conference on Weblogs and Social Media.
12. Karabulut, Y. (2011). *Can Facebook predict stock market activity?* National Bureau of Economic Research, Behavioral Finance Meeting, Standford, CA.
13. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predict the stock market. *Journal of Computational Science, 2*(1), 1–8.
14. Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). *Large-scale sentiment analysis for news and blogs*. ICWSM.
15. Zhang, W., & Skiena, S. (2010). *Trading strategies to exploit blog and news sentiment*. Fourth International Conference on Weblogs and Social Media (ICWSM).
16. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance, 66*(1), 35–65.
17. Gorrell, G., Petrak, J., Bontcheva, K. Emerson, G., & Declerck, T. TrendMiner D2.3.2 Multilingual resources and evaluation of knowledge modelling—v2.
18. Preotiuc-Pietro, D. Samangooei, S., Varga, A., Gelling, D., Cohn, T., & Niranjan, M. (2014). TrendMiner D3.3.1 Tools for mining non-stationary data—v2 Clustering models for discovery of regional and demographic variation—v2.
19. Samangooei, S., Preotiuc-Pietro, D., Jing, L., Niranjan, M., Gibbins, N., & Cohn, T. Trendminer D3.1.1 Regression models of trends in streaming data.
20. Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.
21. Thorp W. A. (2004, September/October). CFA Using investor sentiment as a contrarian indicator. CI Features.