Expert Systems with Applications 41 (2014) 7653–7670

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Text mining for market prediction: A systematic review

Arman Khadjeh Nassirtoussi [a,*], Saeed Aghabozorgi [a], Teh Ying Wah [a], David Chek Ling Ngo [b]

[a] Department of Information Science, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[b] Research & Higher Degrees, Sunway University, No. 5, Jalan University, Bandar Sunway, 46150 Petaling Jaya, Selangor DE, Malaysia

ARTICLE INFO

Article history:
Available online 9 June 2014

Keywords:
Online sentiment analysis
Social media text mining
News sentiment analysis
FOREX market prediction
Stock prediction based on news

ABSTRACT

The quality of the interpretation of the sentiment in the online buzz in the social media and the online news can determine the predictability of financial markets and cause huge gains or losses. That is why a number of researchers have turned their full attention to the different aspects of this problem lately. However, there is no well-rounded theoretical and technical framework for approaching the problem to the best of our knowledge. We believe the existing lack of such clarity on the topic is due to its interdisciplinary nature that involves at its core both behavioral-economic topics as well as artificial intelligence. We dive deeper into the interdisciplinary nature and contribute to the formation of a clear frame of discussion. We review the related works that are about market prediction based on online-text-mining and produce a picture of the generic components that they all have. We, furthermore, compare each system with the rest and identify their main differentiating factors. Our comparative analysis of the systems expands onto the theoretical and technical foundations behind each. This work should help the research community to structure this emerging field and identify the exact aspects which require further research and are of special significance.

1. Introduction

The welfare of modern societies today depends on their market economies. At the heart of any market economy, lies the financial markets with their supply and demand equilibriums. Therefore, it is crucial to study markets and learn about their movements. Understanding market movements primarily facilitates one with the ability to predict future movements. Ability to predict in a market economy is equal to being able to generate wealth by avoiding financial losses and making financial gains. However, the nature of markets is as such that they are extremely difficult to predict if at all.

In general the predictive measures are divided into technical or fundamental analyses. They are differentiated based on their input data, with historic market data to be used for the former and any other kind of information or news about the country, society, company, etc. for the latter. Most of the research in the past has been done on technical analysis approaches, mainly due to the availability of quantitative historic market data and the general desire among traders for technical quantitative methods. Fundamental data is more challenging to use as input specially when it is unstructured. Fundamental data may come from structured and numeric sources like macro-economic data or regular financial reports from banks and governments. Even this aspect of fundamental data has been rarely researched; but occasionally it has been demonstrated to be of predictive value as in the works of Chatrath, Miao, Ramchander, and Villupuram (2014), Khadjeh Nassirtoussi, Ying Wah, and Ngo Chek Ling (2011) and Fasanghari and Montazer (2010). Nevertheless, fundamental data available in unstructured text is the most challenging research aspect and therefore is the focus of this work. Some examples would be the fundamental data available online in the textual information in social media, news, blogs, forums, etc. (see Fig. 1).

In this work a systematic review of the past works of research with significant contribution to the topic of online-text-mining for market-prediction has been conducted, leading to the clarification of the today's cutting-edge research and its possible future directions. The main contributions of this work in summary are:

1. The relevant fundamental economic and computer/information science concepts are reviewed and it is clarified how they are tied into the currently proposed solutions for this research problem.
2. The most significant past literature has been reviewed with an emphasis on the cutting-edge pieces of work.
3. The main differentiating factors among the current works are identified, and used to compare and contrast the available solutions.

* Corresponding author. Tel.: +49 1626230377.
E-mail address: armankhnt@gmail.com (A. Khadjeh Nassirtoussi).

http://dx.doi.org/10.1016/j.eswa.2014.06.009
0957-4174/

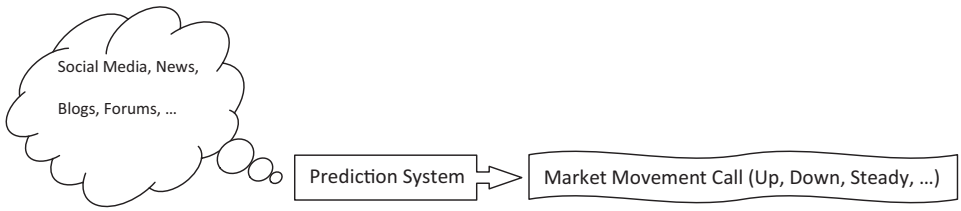

Fig. 1. Online text-mining for market-prediction.

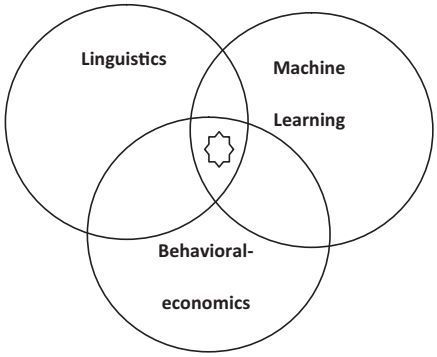

Fig. 2. Interdisciplinary between linguistics, machine-learning and behavioral-economics.

4. Observations are made on the areas with lack of research which can constitute possible opportunities for future work.

The rest of this paper is structured as follows. Section 2 provides insight into the interdisciplinary nature of the research problem at hand and defines the required foundational concepts for a robust comprehension of the literature. Section 3 presents the review of the main available work. Section 4 makes suggestions for future research. And Section 5 concludes this work.

2. A review of foundational interdisciplinary background concepts

Essentially what is being targeted by this research is utilization of computational modeling and artificial intelligence in order to identify possible relationships between the textual information and the economy. That is the research problem.

In order to address this research problem adequately at least three separate fields of study must be included namely: linguistics (to understand the nature of language), machine-learning (to enable computational modeling and pattern recognition), behavioral-economics (to establish economic sense) (see Fig. 2).

The main premise is aligned with recent findings of behavioral-economic principles whereby market conditions are products of human behaviors involved (Bikas, Jurevičienė, Dubinskas, & Novickytė, 2013; Tomer, 2007).

Our research has identified the below background topics as essential to develop a robust comprehension of this research problem:

2.1. Efficient Market Hypothesis (EMH)

The idea that markets are completely random and are not predictable is rooted in the efficient-market hypothesis (Fama, 1965) which asserts that financial markets are "informationally efficient" and that consequently, one cannot consistently achieve returns in excess of average market returns on a risk-adjusted basis, given the information available at the time the investment is made. However, this absolute hypothesis is found not to be completely accurate and Fama himself revises it to include 3 levels of efficiency as strong, semi-strong and weak (Fama, 1970). This indicates that there are many markets where predictability is plausible and viable and such markets are termed as "weakly efficient". Market efficiency is correlated with information availability and a market is only "strongly efficient" when all information is completely available, which realistically is rarely the case. Hence, he conceded that his theory is stronger in certain markets where the information is openly, widely and instantly available to all participants and it gets weaker where such assumption cannot be held concretely in a market.

2.2. Behavioral-economics

Cognitive and behavioral economists look at price as a purely perceived value rather than a derivative of the production cost. The media do not report market status only, but they actively create an impact on market dynamics based on the news they release (Robertson, Geva, & Wolff, 2006; Wisniewski & Lambe, 2013).

People's interpretation of the same information varies greatly. Market participants have cognitive biases such as overconfidence, overreaction, representative bias, information bias, and various other predictable human errors in reasoning and information processing (Friesen & Weller, 2006).

Behavioral finance and investor sentiment theory have firmly established that investors' behavior can be shaped by whether they feel optimistic (bullish) or pessimistic (bearish) about future market values (Bollen & Huina, 2011).

2.3. Adaptive market hypothesis (AMH)

The dilemma of market efficiency with regards to its degree and applicability to different markets is still a vibrant and ongoing topic of research with very contradictory results. For every paper producing empirical evidence supporting the market efficiency, a contradictory paper can perhaps be found which empirically establishes market inefficiency (Majumder, 2013). Hence, a few years ago research has produced a counter-theory by the name of Adaptive Markets Hypothesis in an effort to reconcile Efficient Markets hypothesis with Behavioral Finance (Lo, 2005). Urquhart and Hudson (2013) have conducted a comprehensive empirical investigation on the Adaptive Markets Hypothesis (AMH) in three of the most established stock markets in the world; the US, UK and Japanese markets using very long run data. Their research provides evidence for adaptive markets, with returns going through periods of independence and dependence although the magnitude of dependence varies quite considerably. Thus the linear dependence of stock returns varies over time but nonlinear dependence is strong throughout. Their overall results suggest that the AMH provides a better description of the behavior of stock returns than the Efficient Market Hypothesis.