Expert Systems With Applications 42 (2015) 9192–9210

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms

Alya Al Nasseri[a,*], Allan Tucker[b], Sergio de Cesare[a]

[a] Brunel Business School, Brunel University London, UK
[b] Department of Computer Science, Brunel University London, UK

ARTICLE INFO

Keywords:
Decision tree
Filter approach
Text mining
Trading Strategy

ABSTRACT

Growing evidence is suggesting that postings on online stock forums affect stock prices, and alter investment decisions in capital markets, either because the postings contain new information or they might have predictive power to manipulate stock prices. In this paper, we propose a new intelligent trading support system based on sentiment prediction by combining text-mining techniques, feature selection and decision tree algorithms in an effort to analyze and extract semantic terms expressing a particular sentiment (sell, buy or hold) from stock-related micro-blogging messages called "StockTwits". An attempt has been made to investigate whether the power of the collective sentiments of StockTwits might be predicted and how the changes in these predicted sentiments inform decisions on whether to sell, buy or hold the Dow Jones Industrial Average (DJIA) Index. In this paper, a filter approach of feature selection is first employed to identify the most relevant terms in tweet postings. The decision tree (DT) model is then built to determine the trading decisions of those terms or, more importantly, combinations of terms based on how they interact. Then a trading strategy based on a predetermined investment hypothesis is constructed to evaluate the profitability of the term trading decisions extracted from the DT model. The experiment results based on 122-tweet term trading (TTT) strategies achieve a promising performance and the (TTT) strategies dramatically outperform random investment strategies. Our findings also confirm that StockTwits postings contain valuable information and lead trading activities in capital markets.

1. Introduction

Stock market prediction is an attractive and challenging area of research different methodologies that has been developed with the aim of predicting the direction of securities' prices as accurately as possible (Guresen, Kayakutlu, & Daim, 2011). The aim has been to create accurate models that have the ability to predict stock price behavioral movements in the stock market. However, predicting these changes is very challenging (and appealing to researchers to investigate), due to the fact that stock market data are noisy and time varying in nature (Atsalakis & Valavanis, 2009). To address the topic of future stock price predictions, several theories become relevant in this regard. Several works have attempted to study stock market prediction while providing an answer to the common question: can stock prices really be predicted? There are two theories that are mostly relevant in answering such a question: (1) Efficient Market Hypothesis and (2) Random Walk Theory. According to the Efficient Market Hypothesis (EMH), market prices reflect all publicly available information (Fama, 1970). This implies that past and current information is immediately incorporated into the stock prices, thus any price changes can only be explained by new information or "news". Due to the random arrival of new information, the stock price is said to follow a random walk pattern and it is impossible to predict the stock market, since prices are randomly determined. If this hypothesis is held; therefore the attempts to predict the stock market will be ineffective. The researchers continuing efforts in accurately forecasting stock markets using various methods and techniques have proved that underlying assumptions of the EMH and random walk turn out to be unrealistic and that some degree of predictability might be possible (Darrat & Zhong, 2000). A variety of machine learning techniques have been proposed to predict the future movement and trend of stock prices in capital markets. However, most of these studies focus on predicting the movement in stock prices rather than predicting the investment decisions that derive from and cause the movement itself, such as buying, selling and holding decisions. For example, Xue-shen, Zhong-ying, Da-ren, Qing-hua, and Hui (2007) adopted classification complexity of Support Vector Machines (SVM) as a feature selection

* Corresponding author. Tel.: +447703346765.
E-mail addresses: Alya.Al-Nasseri@brunel.ac.uk, aliaaa83@hotmail.com (A.A. Nasseri), Allan.Tucker@brunel.ac.uk (A. Tucker), Sergio.deCesare@brunel.ac.uk (S. de Cesare).

http://dx.doi.org/10.1016/j.eswa.2015.08.008

criterion to predict the Shanghai Stock Exchange Composite Index (SSECI). Huang, Yang, and Chuang (2008) employed a wrapper approach to select the optimal feature subset and apply various classification algorithms to predict the trend in the stock markets of Taiwan and South Korea. Lee (2009) proposed a prediction model based on a hybrid feature selection method and SVM to predict the trend of the stock market.

Investor sentiment has been proven significant in affecting the behavior of stock prices (Baker & Wurgler, 2006). Investors' expectations and their psychological thinking, from which the sentiments are derived, are considered the main factors that affect stock price movements in capital markets (Tan, Quek, & Ng, 2007). Behavioral finance theory suggested that the existence of different types of traders and the effect of their trading behavior has substantial effect on influencing price changes in financial markets (DeLong, Shleifer, Summers, & Waldmann, 1991). There are two types of traders in financial markets, the "irrational noise trader" or so-called "day trader" is one who does not possess fundamental information (Kyle, 1985) and the "rational trader" or "arbitrageur" who holds rational beliefs (DeLong, Shleifer, Summers, & Waldmann, 1990) and, in a way, is always updating their beliefs according to the new information available to them (Baberies and Thaler, 2003). The presence of noise traders in financial markets, who make irrational decisions regarding buying, selling or holding stocks, can then cause price levels and risks to deviate from expected levels, even if all other traders are rational (De Long et al., 1990). Noise traders are always taking part in the discussion and conversations related to financial information in capital markets. In the context of online investment forums, conversations among investors including the noise traders involve making predictions, exchanging opinions, asking questions, sharing analyses, and reporting financial information (Oh & Sheng, 2011). Therefore, the ability of noise traders to affect price changes will also appear in online investment forums, where information and opinion is widely spread among investors through the investment communication platforms (Zhang & Swanson 2010). It is therefore important to highlight the critical role played by trading decisions in the stock market. Trading decisions have a great effect on the profitability position of an investor in the capital market. Therefore, the ability to predict an intelligent trading support mechanism would help investors to make profitable investment decisions concerning a particular security in the capital market. Making correct investment decisions is a substantially difficult task for investors due to the problem of high nonlinearity embodied in the behavior of financial markets.

Many attempts have been made to provide investors and other financial professionals with consistently profitable autonomous trading support systems. Motivation for such trading systems comes from various fields of studies ranging from fundamental analysis and financial econometric modeling to evolutionary computation (Hu, Feng, Zhang, Ngai & Liu, 2015), machine learning (Booth Gerding, and McGroarty, 2014) and text mining (Gong, Zeng, & Zhang, 2011; Nuij, Milea, Hogenboom, Frasincar, & Kaymak, 2014). In this context, numerous financial researchers have progressively provided investors and their peers in capital markets with decision-making support systems in order to improve and enhance their ability to make a better-informed investment decision that will lead to greater return on their investments (Kodogiannis & Lolis, 2002; Li & Kuo, 2008; Skabar, 2005; Sun, Liang, Zhang, Lee, Lin et al., 2005; Chun & Park, 2005). Some of these studies are based on traditional time series predictions (Kodogiannis & Lolis, 2002; Skabar, 2005; Sun et al., 2005) and trend prediction (Cheng Wei, and Chen, 2009; Tsai & Hsiao, 2010) that mainly focused on historical past prices in predicting the future value of stocks. Most of the capital market players, however, are much more interested in time series predictions of future trends rather than exact future prices. In addition to the traditional time series approach, the application of artificial intelligence (AI), such as expert systems (Kee & Koh, 1994), fuzzy systems (Abraham Nath & Mahanti, 2001; Chang & Liu, 2008), and artificial neural networks (ANN) (Chiang, Urban, & Baldridge, 1996; Duan et al., 2009; Masoud, 2014), has received extensive attention by researchers with an attempt to make the forecast of future prices more reliable. Despite the effectiveness demonstrated by such methodologies, there are some drawbacks associated with their applications. For example, the main drawback with ANNs and other black-box techniques is that the results obtained from such methodologies are misleading and very difficult to interpret (Lai, Fan, Huang, & Chang, 2009). Another drawback is the lack of investigating the nature of interactions between technical indicators and stock market fluctuations. Methodologies that provide a greater insight into market procedures must therefore be developed (Chi, Chen, & Cheng, 1999; Zhang, 2007). However, most recent studies tend to provide accurate trading strategies by combining machine learning techniques (e.g., SVM) with all other techniques, namely robust feature selection, transactional volume incorporation, pattern models and technical analysis. The research community has had a long-standing argument on the effectiveness of technical analyses in stock trading. Some argue that stock prices are not predictable while others, such as Brock, Lakonishok and LeBaron, (1992) and Blume Easley, and O'hara. (1994), have presented positive empirical evidence on the effectiveness of technical analyses (Kaucic, 2010). Kara, Boyacioglu, and Baykan (2011) provide a comparable pattern whereby neutral networks and plain SVM were compared for the purpose of making stock price movement prediction with the extensive use of several technical indicators. Rosillo, Giner, and de la Fuente (2014) used Volatility Index and technical analysis with the aim to forecast weekly change in S&P 500. Dai, Shao, and Lu (2013) incorporated MARS splines for attribute selection, which then was used as an input for the Support Vector Regression model. Recently Żbikowski (2015) applied a modified Support Vector Machine (SVM) classifier (volume weighted SVM) with walk forward testing and the Fisher method for feature selection for the purpose of creating a stock trading strategy and forecasting short-term trends on the stock market. Hu et al. (2015) proposed a hybrid long-term and short-term evolutionary trend following algorithm (eTrend) that combines TF investment strategies with the eXtended Classifier Systems (XCS) for the purpose of providing effective trading guidance for investors in the capital market.

The provision of an accurate and timely trading support mechanism is the key success for traders to make a profitable decision in capital markets. This study presents a novel approach for developing a new decision support system based on tweet semantic terms extracted from the decision tree model (Quinlan, 1993) which then can be implemented as a trading strategy and constitute three different portfolios (sell, buy and hold). The decision tree proved successful in searching for rules hidden in large amounts of data. The visibility of the connected relationships between nodes branches and leaves in the tree makes it most suitable approach for feature selection and prediction of investment trading decisions in capital markets. It has also proved efficient for time series analysis. In addition, decision tree techniques have already been shown to be interpretable, efficient, problem independent and able to deal with large-scale applications. The decision tree model provides a visualized insight into the StockTwits data by highlighting the individual relationships with respect to the class as well as the combined associations of features with respect to the decision class. One would expect that the decision effect of individual terms (feature) appearing in a tweet posting would have a different decision effect than if it had appeared in combination with other terms. The ability of the decision tree model to explore the related interactions between the selected terms and their ability to predict trading decisions makes it a better and more suitable model for this research.

This research takes a different approach by proposing an automatic decision support system that integrates text mining techniques, feature selection and decision tree algorithm. This research