Information Processing and Management 50 (2014) 426–441

Contents lists available at ScienceDirect

Information Processing and Management

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

Evaluating and understanding text-based stock price prediction models

Enric Junqué de Fortuny [a,*], Tom De Smedt [b], David Martens [a], Walter Daelemans [b]

[a] Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium
[b] Faculty of Arts, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium

ARTICLE INFO

Article history:
Received 18 September 2012
Received in revised form 20 August 2013
Accepted 23 December 2013
Available online 31 January 2014

Keywords:
Stock
Prediction
Support Vector Machine
Sentiment mining
Evaluation

ABSTRACT

Despite the fact that both the Efficient Market Hypothesis and Random Walk Theory postulate that it is impossible to predict future stock prices based on currently available information, recent advances in empirical research have been proving the opposite by achieving what seems to be better than random prediction performance. We discuss some of the (dis)advantages of the most widely used performance metrics and conclude that is difficult to assess the external validity of performance using some of these measures. Moreover, there remain many questions as to the real-world applicability of these empirical models. In the first part of this study we design novel stock price prediction models, based on state-of-the-art text-mining techniques to assert whether we can predict the movement of stock prices more accurately by including indicators of irrationality. Along with this, we discuss which metrics are most appropriate for which scenarios in order to evaluate the models. Finally, we discuss how to gain insight into text-mining-based stock price prediction models in order to evaluate, validate and refine the models.

1. Introduction

Is there a way to outperform other investors on the markets? It is a question that has attracted the attention of many a trader since the advent of stock markets in Europe during the late Middle Ages. With the emergence of companies, financial institutions, financial products and government-imposed regulations on these products, the nature of stock markets has changed substantially since those days. Nevertheless, stock price prediction remains an attractive topic for both researchers and investors.[1] During the past decades different theories have been developed to motivate why stock price prediction is not feasible under the assumption of rational actors in a market.

The *efficiënt market hypothesis* was first introduced by Fama (1965) and posits that the financial markets are informationally efficient. This implies that one cannot design a system to predict the change in stock price based on any information because all information is already reflected in the current stock price. Similarly, Malkiel (1985) argues that the stock market prices of assets evolve in a pattern comparable to that of a Random Walk (hence its name *Random Walk Theory*). The implication is that stock prices cannot be predicted better than a "blindfolded chimpanzee throwing darts" at a numerical scale board.

The contributions of this publication are twofold. First, we build empirical models to try to counter-act the validity of the previously mentioned theories, based on the fact that humans do not always act rationally. In order to do so, we combine text-mining techniques in a novel hybrid modelling technique. Second, we discuss the difficulties in evaluating such a model

* Corresponding author. Tel.: +32 32654393.
E-mail address: enric.junquedefortuny@uantwerp.be (E. Junqué de Fortuny).

[1] Querying Google for "stock prediction" reveals over 1.9 million results, including 1360 scientific publications.

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.12.002

as a *decision making tool*. We discuss how using many different evaluation metrics can remedy this situation and we show how the model can be used as a *decision support tool* without the aforementioned drawbacks.

2. Empirical research on stock prediction

2.1. Design of empirical models using text mining

A selection of recent empirical research is shown in Table 1, together with the main design choices in these studies. As can be seen from the table, most of the models predicted classes of movement (e.g. up/down) instead of the actual values. Although traditionally most research has centred on various short- and long-term technical performance indicators of a stock (e.g., Lavrenko, Schmill, Lawrie, & Ogilvie (2000)), more recent research has focused on building models based on textual information to perform directional predictions of stock movement (e.g., Schumaker & Chen (2009b); Mittermayer (2006)). The behavior explained by both theories mentioned in the Introduction is based on the assumption that investors act rationally. One way to counter them is to counter the assumption of rationality of the trader. Irrational behavior could for example occur as a reaction to news in the popular written media, i.e. information that comes in the format of text. Since there is a lag between the appearance of an article text and the trading action of the reader, automatic trading systems could outperfom human reaction in a high-frequency trading environment.

2.1.1. Text mining

Text mining concerns the process of automatically extracting novel, non-trivial information from unstructured text documents (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), by combining techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR) and knowledge management (Mihalcea et al., 2007). Common text mining tasks involve document classification, summarization, clustering of similar documents, concept extraction and sentiment analysis. Text mining has had a wide range of applications to date. Prevalent applications include: forecasting petitions (Suh, Park, & Jeon, 2010), guiding financial investments (Rada, 2008) and sentiment detection (Tang, Tan, & Cheng, 2009; Junqué de Fortuny, De Smedt, Martens, & Daelemans, 2012).

In our setup we look for patterns in the occurrence of words (the so called *bag-of-words*-approach) or the sentiment of the message that have an influence on the stock market price of a commodity. In related work, typically only the direction (rise/fall) of the stock movement is predicted and the patterns come in the form of a linear model, in which each word of a certain vocabulary receives a weight towards the stock price either going up or down. The weighted sum of the word scores of all words in the article is then used in the prediction of a new article. Reported results on independent test sets in terms of accuracy have been in the order of a 10% increase when compared to random predictions (Mittermayer, 2006). We will explore the specifics of text mining for stock price prediction in Section 4 when we discuss the construction of the empirical models.

2.1.2. Combining information

Amongst others, Li, Wang, Dong, and Wang (2011) and Schumaker and Chen (2009b) remark that using only textual information can be too limiting because the approach disregards other (complementary) information. Imagine that a negative news article concerning a certain asset is published during an upward trend of the asset's price. This article might influence the positive trend in a negative way by reducing the slope of that trend, yet the overall trend for the asset can stay upward (see Fig. 1). In this case a negative directional prediction would be wrong, although the impact of the message itself was

Table 1
Literature overview of studies containing text analysis for stock prediction with key design choices. Popular techniques include Naive Bayes (NB), genetic algorithms (GA) and Support Vector Machines (SVM).

Reference	Prediction window	Exchange/index	Technique	Metrics	Target
Wuthrich and Cho (1998)	closing price	Mixed	NB	acc., return	+/±/–
Lavrenko et al. (2000)	1 h	Mixed	NB	profit	++/+/±/–/–
Thomas (2000)	closing price	NASDAQ, NYSE	hybrid GA	excess returns	+/±/–
Gidofalvi (2001)	1 h	NASDAQ	NB	precision/recall	+/±/–
Peramunetilleke (2002)	1-3 h	Currency rates DEM/USD JPY/USD	decision rules	acc.	+/+–/–
Pui Cheong Fung and Xu Yu (2003)	1 h	Hongkong Stock Exchange	SVM	return	+/±/–
Mittermayer (2006)	15 m	S& P 500	SVM	acc., profit, return	+/±/
Zhai et al. (2007)	20 m	BHP Billion Ltd.	SVM	acc., profit	+/–
Schumaker and Chen (2009a)	20 m	S& P 500	SVR	acc., return	+/– and value
Li et al. (2011)	5–30 m	Hang Seng Index	SVM	acc.	+/–
This publications	1 m-64 m/1 day	Euronext Brussels	SVM	acc., AUC, return, Sharpe	+/–