Expert Systems With Applications 42 (2015) 9603–9611

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Sentiment analysis on social media for stock movement prediction

Thien Hai Nguyen [a,*], Kiyoaki Shirai [a], Julien Velcin [b]

[a] School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[b] University of Lyon (ERIC, Lyon 2), 5 Avenue Pierre Mendes-France, 69676 Bron Cedex, France

ARTICLE INFO

Keywords:
Sentiment analysis
Opinion mining
Classification
Prediction
Stock
Social media
Message board

ABSTRACT

The goal of this research is to build a model to predict stock price movement using the sentiment from social media. Unlike previous approaches where the overall moods or sentiments are considered, the sentiments of the specific topics of the company are incorporated into the stock prediction model. Topics and related sentiments are automatically extracted from the texts in a message board by using our proposed method as well as existing topic models. In addition, this paper shows an evaluation of the effectiveness of the sentiment analysis in the stock prediction task via a large scale experiment. Comparing the accuracy average over 18 stocks in one year transaction, our method achieved 2.07% better performance than the model using historical prices only. Furthermore, when comparing the methods only for the stocks that are difficult to predict, our method achieved 9.83% better accuracy than historical price method, and 3.03% better than human sentiment method.

1. Introduction

Stock price forecasting is very important in the planning of business activity. However, building an accurate stock prediction model is still a challenging problem. In addition to historical prices, the current stock market is affected by the mood of society. The overall social mood with respect to a given company might be one of the important variables which affect the stock price of that company. Nowadays, the emergence of online social networks makes available large amounts of mood data. Therefore, incorporating information from social media with the historical prices can improve the predictive ability of models.

The goal of our research is to develop a model to predict the stock price movement (whether the price will be up or down) using information from social media (Message Board). In our proposed method, a model that predicts the stock value at t using features derived from information at $t-1$ and $t-2$, where t stands for a transaction date, will be trained by supervised machine learning. Apart from the mood information, the stock prices are affected by many factors such as microeconomic and macroeconomic factors. However, this research only focuses on how the mood information from social media can be used to predict the stock price. We will mainly aim at extracting the mood information by sentiment analysis on social media data. Then, these sentiments will be integrated into a model to predict stocks. To achieve this goal, discovering the topics and sentiments in a large amount of social media is very important to get the opinions of investors. However, sentiment analysis on social media is difficult. The text is usually short, contains many misspellings, uncommon grammar constructions and so on. In addition, the literature shows conflicting results in sentiment analysis for stock market prediction. Some researchers report that sentiments from social media have no predictive capabilities (Antweiler & Frank, 2004; Tumarkin & Whitelaw, 2001), while other researchers have reported either weak or strong predictive capabilities (Bollen, Mao, & Zeng, 2011). Therefore, how to use opinions in social media for stock price predictions is still an open problem.

One contribution of this paper is that we propose a novel feature 'topic-sentiment' to improve the performance of stock market prediction. It is important to recognize what topics are discussed in social media and how people feel about these topics. The 'topic-sentiment' feature, which represents the sentiments of the specific topics of the company (product, service, dividend and so on), are used for prediction of stock price movement. This feature is obtained in two ways: by using the existing topic model called the joint sentiment/topic model (JST) and by our own proposed method. The extracted topics and sentiments in the former method are hidden (latent), whereas not hidden in the latter. To the best of our knowledge, this is the first research trying to extract topics and sentiments simultaneously and utilize them for stock market prediction. Another contribution is a large scale evaluation. The effectiveness of the sentiments in social media in stock market prediction is still uncertain because a

* Corresponding author. Tel.: +81 80 2956 5927.
E-mail addresses: nhthien8x@gmail.com, nhthien@jaist.ac.jp (T.H. Nguyen), kshirai@jaist.ac.jp (K. Shirai), julien.velcin@univ-lyon2.fr (J. Velcin).

http://dx.doi.org/10.1016/j.eswa.2015.07.052
0957-4174/

relatively small data was used for evaluation in the previous work. This paper investigates whether the sentiments in the social media are really useful on the test data containing many stocks and transaction dates.

The rest of the paper is organized as follows. Section 2 introduces some previous approaches on sentiment analysis for stock prediction. Section 3 describes our dataset. Section 4 describes our proposed method. We also propose a novel feature for stock prediction based on the topics and the sentiments associated with them. Section 5 assesses the results of the experiments. Finally, Section 6 concludes our contribution.

2. Related work

Stock market prediction is one of the most attracted topics in academic as well as real life business. Many researches have tried to address the question whether the stock market can be predicted. Some of the researches were based on the random walk theory and the Efficient Market Hypothesis (EMH). According to the EMH (Fama, 1991; Fama, Fisher, Jensen, & Roll, 1969), the current stock market fully reflects all available information. Hence, price changes are merely due to new information or news. Because news in nature happens randomly and is unknowable in the present, stock prices should follow a random walk pattern and the best bet for the next price is current price. Therefore, they are not predictable with more than about 50% accuracy (Walczak, 2001). On the other hand, various researches specify that the stock market prices do not follow a random walk, and can be predicted at some degree (Bollen et al., 2011; Qian & Rasheed, 2007; Vu, Chang, Ha, & Collier, 2012). Degrees of directional accuracy at 56% hit rate in the predictions are often reported as satisfying results for stock predictions (Schumaker & Chen, 2009b; Si, Mukherjee, Liu, Li, Li, & Deng, 2013; Tsibouris & Zeidenberg, 1995).

Besides the efficient market hypothesis and the random walk theories, there are two distinct trading philosophies for stock market prediction: fundamental analysis and technical analysis. The fundamental analysis studies the company's financial conditions, operations, macroeconomic indicators to predict stock price. On the other hand, the technical analysis depends on historical and time-series prices. Price moves in trends, and history tends to repeat itself. Some researches have tried to use only historical prices to predict the stock price (Cervelló-Royo, Guijarro, & Michniuk, 2015; Patel, Shah, Thakkar, & Kotecha, 2015a, 2015b; Ticknor, 2013; Zuo & Kita, 2012a, 2012b). To discover the pattern in the data, they used Bayesian network (Zuo & Kita, 2012a, 2012b), time-series method such as Auto Regressive model, Moving Average model (Patel et al., 2015a, 2015b), Auto Regressive Moving Average model (Zuo & Kita, 2012a) and so on.

While these previous methods did not consider the sentiments on the social media, in this paper our work aims at incorporating them to improve the performance of the stock market prediction.

Most of the research tried to predict only one stock (Bollen et al., 2011; Qian & Rasheed, 2007; Si et al., 2013) and the number of instances (transaction dates) in a test set is very low such as 14 or 15 instances (Bollen et al., 2011; Vu et al., 2012). With only a few instances in the test set, the conclusion might be insufficient. To the best of our knowledge, there is no research showing a good prediction result on a data consisting of many stocks in a long time period. Our research tried to solve this issue by predicting 18 stocks over a period of one year.

2.1. Use of opinions from text for stock market prediction

Sentiment analysis has been found to play a significant role in many applications such as product reviews and restaurant reviews (Liu & Zhang, 2012; Pang & Lee, 2008). There are some researches trying to apply sentiment analysis on an information source to improve the stock prediction model (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014). There are two main sources from which authors have incorporated information aggregated from textual content into financial models. In the past, the main source was the news (Schumaker & Chen, 2009a, 2009b), and in recent years, social media sources. Then, these sentiments are integrated into prediction models. A simple approach is combining the textual content with the historical prices through the linear regression model.

Most of the previous work primarily used the bag-of-words as text representation that are incorporated into the prediction model. Schumaker and Chen (2009b) tried to use different textual representations such as bag-of-words, noun phrases and named entities for financial news. Then this information was integrated with linear regression and support vector machine regression as predictive models. They applied their models to estimate a discrete stock price 20 min. after a news article was released. The results show 0.04261 mean square error, 57.1% directional accuracy, and 2.06% return in a simulated trading engine. However, the textual representations are just the words or named entity tags, not exploiting so much about the mood information.

Antweiler and Frank (2004) used naive Bayes to classify the messages from message boards into three classes: buy, hold and sell. The number of relevant messages in these three classes was aggregated into a single measure of bullishness. They investigated three aggregation functions as a number of alternatives to bullishness. They were integrated into the regression model. However, they concluded that their model does not successfully predict stock returns.

Zhang, Fuehres, and Gloor (2011) measured collective hope and fear on each day and analyzed the correlation between these indices and the stock market indicators. They used the mood words to tag each tweet as fear, worry, hope and so on. They concluded that the emotional tweet percentage significantly negatively correlated with Down Jones, NASDAQ and S&P 500, but had significant positive correlation to VIX. However, they did not use their model to predict the stock price values.

Two mood tracking tools, OpinionFinder and Google Profile of Mood States, were used to analyze the text content of daily Twitter (Bollen et al., 2011). The former measures positive and negative mood. The latter measures mood in terms of six dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). They used the Self Organizing Fuzzy Neural Network model to predict DJIA values. The results show 86.7% direction accuracy (up or down), Mean Absolute Percentage Error 1.79%. However, their test period is very short (from December 1 to December 19, 2008). Even though, they achieved high accuracy, there are only 15 transaction dates in their test set. With such a short period, it might not be sufficient to conclude the effectiveness of their method.

Xie, Passonneau, Wu, and Creamer (2013) proposed a novel tree representation based on semantic frame parsers. They indicated that this representation performed significantly better than bag-of-words. By using stock prices from Yahoo Finance, they annotated all the news with labels in a transaction date as going up or down categories. However, the weakness of this assumption is that all the news in one day will have the same category. In addition, this becomes a document classification problem, not stock prediction.

Rechenthin, Street, and Srinivasan (2013) incorporated Yahoo Finance Message Board into the stock movement prediction. They tried to use various classification models to predict stock. They used the explicit sentiments and predicted sentiments obtained by a classification model with the bag-of-words and meta-features.

A keyword-based algorithm was proposed to identify the sentiment of tweets as positive, neutral and negative for stock prediction (Vu et al., 2012). Their model achieved around 75% accuracy. However, their test period is very short, from 8th to 26th in September, 2012 which contains only 14 transaction dates.