Information Sciences 285 (2014) 181–203

Contents lists available at ScienceDirect

Information Sciences

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

Stream-based active learning for sentiment analysis in the financial domain

Jasmina Smailović [a,b,*], Miha Grčar [a,b], Nada Lavrač [a,b,c], Martin Žnidaršič [a,b]

[a] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Jamova 39, 1000 Ljubljana, Slovenia
[c] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

ARTICLE INFO

Article history:
Received 22 March 2013
Received in revised form 31 March 2014
Accepted 17 April 2014
Available online 25 April 2014

Keywords:
Predictive sentiment analysis
Stream-based active learning
Stock market
Twitter
Positive sentiment probability
Granger causality

ABSTRACT

Studying the relationship between public sentiment and stock prices has been the focus of several studies. This paper analyzes whether the sentiment expressed in Twitter feeds, which discuss selected companies and their products, can indicate their stock price changes. To address this problem, an active learning approach was developed and applied to sentiment analysis of tweet streams in the stock market domain. The paper first presents a static Twitter data analysis problem, explored in order to determine the best Twitter-specific text preprocessing setting for training the Support Vector Machine (SVM) sentiment classifier. In the static setting, the Granger causality test shows that sentiments in stock-related tweets can be used as indicators of stock price movements a few days in advance, where improved results were achieved by adapting the SVM classifier to categorize Twitter posts into three sentiment categories of positive, negative and neutral (instead of positive and negative only). These findings were adopted in the development of a new stream-based active learning approach to sentiment analysis, applicable in incremental learning from continuously changing financial tweet streams. To this end, a series of experiments was conducted to determine the best querying strategy for active learning of the SVM classifier adapted to sentiment analysis of financial tweet streams. The experiments in analyzing stock market sentiments of a particular company show that changes in positive sentiment probability can be used as indicators of the changes in stock closing prices.

1. Introduction

Predicting the value of stock market assets is a challenge investigated by numerous researchers. One of the reasons for addressing this challenge is the controversy of the efficient market hypothesis [17], which claims that stocks are always traded at their fair value. Based on this market theory, claiming that it is not possible for investors to buy undervalued stocks or sell stocks for overestimated prices, it is impossible for traders to consistently outperform the average market returns. This hypothesis is based on the assumption that financial markets are informationally efficient (i.e., that stock prices always reflect all the relevant information at investment time). The unpredictable nature of stock market prices was first investigated by Regnault [51] and later by Bachelier [4]. Fama [17], who proposed the efficient market hypothesis, also claimed that stock price movement is unpredictable and that past price movements cannot be used to forecast future stock prices.

* Corresponding author at: Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia. Tel.: +386 1 4773 143.
E-mail address: jasmina.smailovic@ijs.si (J. Smailović).

http://dx.doi.org/10.1016/j.ins.2014.04.034
0020-0255/

However, as the efficient market hypothesis is controversial, researchers from various disciplines (including economists, statisticians, finance experts, and data miners) have been investigating the means to predict future stock market prices. The findings vary: from those claiming that stock market prices are not predictable to those presenting opposite conclusions [9,33].

This paper addresses the described challenge in the context of the explosive growth of social media and user-generated content on the Internet. Through blogs, forums, and social networking media, more and more people share their opinions about individuals, companies, movements, or important events. Such opinions both express and evoke sentiments [49]. Recent research indicates that analysis of these online texts can be useful for trend prediction. For example, it was shown that the frequency of blog posts can be used to forecast spikes in online consumer purchasing [23]. Moreover, it was shown by Tong [72] that references to movies in newsgroups are correlated with their sales. Sentiment analysis of weblog data was successfully used to predict the financial success of movies [40]. Twitter[1] posts were also shown to be useful for predicting box-office revenues of movies before their release [3].

Twitter is currently the most popular microblogging platform [46] allowing its users to send and read short messages of up to 140 characters in length, known as *tweets*, via SMS, the Twitter website, or a range of applications for mobile devices. Twitter gained global popularity very quickly with over 500 million active users in 2012, writing over 340 million tweets daily [16,41]. Twitter data (and data from other social network websites) are very interesting because of their large volume, popularity, and capability of near-real-time publishing of individuals' opinions and emotions about any subject. Given that this massive amount of user-generated content became abundant and easily accessible, many researchers became interested in the predictive power of microblogging messages, especially in the domain of stock market prediction, prediction of election results, or prediction of the financial success of movies or books. Many of these studies use *sentiment analysis* [36,75] as a basis for prediction. The term *sentiment*, used in the context of automatic analysis of text and detection of predictive judgments from positively and negatively opinionated texts, first appeared in the papers by Das and Chen [14] and Tong [72], where the authors were interested in analyzing stock market sentiment. Even though there are many studies on predicting the phenomenon of interest using sentiment analysis of online texts, there is still an urge to develop methods and tools for adaptive dynamic sentiment analysis of microblogging posts, which would enable handling changes in such data streams. This field of research is still insufficiently explored and represents a challenge, which is addressed in this work through *active learning* [61].

This work contributes to sentiment analysis and to active learning research, and partly towards better understanding of phenomena in financial stock markets. While sentiment analysis is generally aimed at detecting the author's attitude, emotions or opinions expressed in the text, our study is concerned with the development of an approach to *predictive sentiment analysis*. With this term, we denote an approach in which sentiment analysis is used to predict a specific phenomenon or its changes, postulating that the proposed methodology for predictive sentiment analysis of streams of microblogging messages should be capable of predicting the financial phenomenon of interest. The indication that there may be a relationship between emotions and stock market prices relies on findings in psychological research which indicate that emotions are crucial to rational thinking and social behavior [13], and can influence the choice of actions. Given that the general mood of a society is propagated through social interactions, the collective social mood can be transferred through the investors to the stock market and consequently, the sentiment can be reflected in stock price movements. As a result, the stock market itself can be considered as a measure of social mood [44]. It is, thus, reasonable to expect that the analysis of the public mood can be used to predict price movements in the stock market. We hypothesize that this assumption may hold in situations when people actually express positive or negative opinions about some topic concerning the stock market, whereas in situations when people do not express opinions, but mostly neutral facts, we anticipate finding no correlations. In accordance with this hypothesis, we propose a mechanism for distinguishing opinionated (positive and negative) from non-opinionated (neutral) tweets in Twitter data streams.

In an effort to build an active learning approach to sentiment analysis, applicable in incremental learning from continuously changing financial tweet data streams, we first addressed a static Twitter data analysis problem, which was explored in order to determine the best Twitter-specific text preprocessing setting for training the Support Vector Machine (SVM) sentiment classifier. In the static setting, the Granger causality test showed that sentiment in stock-related tweets can be used as an indicator of stock price movements a few days in advance, where improved results were achieved by adapting the SVM classifier to categorize Twitter posts into three sentiment categories of positive, negative and neutral (instead of positive and negative only). These findings were successfully used in the development of a new stream-based active learning approach to sentiment analysis, applicable in incremental learning from continuously changing financial tweet data streams.

Using stream data for sentiment analysis makes sense when the information about the changes in the sentiment is time-critical and a proper data flow is available, for example, in the analysis of streams of financial tweets in which people express their opinions about stocks in real time. The main idea of active learning [56,61,65], adapted in this study for continuously updating the sentiment classifier from a tweet stream, is that the algorithm is allowed to select new examples to be labeled by the oracle (e.g., a human annotator) and added to the training set. It aims at maximizing the performance of the algorithm with as little human labeling effort as possible. The main challenge of active learning is the selection of the most suitable

[1] www.twitter.com.