Predictive Sentiment Analysis of Tweets: A Stock Market Application

Jasmina Smailović[1,2], Miha Grčar[1], Nada Lavrač[1], and Martin Žnidaršič[1]

[1] Jožef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia
{jasmina.smailovic,miha.grcar,nada.lavrac,
martin.znidarsic}@ijs.si
[2] Jožef Stefan International Postgraduate School, Jamova cesta 39,
1000 Ljubljana, Slovenia

Abstract. The application addressed in this paper studies whether Twitter feeds, expressing public opinion concerning companies and their products, are a suitable data source for forecasting the movements in stock closing prices. We use the term predictive sentiment analysis to denote the approach in which sentiment analysis is used to predict the changes in the phenomenon of interest. In this paper, positive sentiment probability is proposed as a new indicator to be used in predictive sentiment analysis in finance. By using the Granger causality test we show that sentiment polarity (positive and negative sentiment) can indicate stock price movements a few days in advance. Finally, we adapted the Support Vector Machine classification mechanism to categorize tweets into three sentiment categories (positive, negative and neutral), resulting in improved predictive power of the classifier in the stock market application.

Keywords: stock market, Twitter, predictive sentiment analysis, sentiment classification, positive sentiment probability, Granger causality.

1 Introduction

Trying to determine future revenues or stock prices has attracted a lot of attention in numerous research areas. Early research on this topic claimed that stock price movements do not follow any patterns or trends and past price movements cannot be used to predict future ones [1]. Later studies, however, show the opposite [2]. It has also been shown that emotions have an effect on rational thinking and social behavior [3] and that the stock market itself can be considered as a measure of social mood [4].

As more and more personal opinions are made available online, recent research indicates that analysis of online texts such as blogs, web pages and social networks can be useful for predicting different economic trends. The frequency of blog posts can be used to predict spikes in the actual consumer purchase quantity at online retailers [5]. Moreover, it was shown by Tong [6] that references to movies in newsgroups were correlated with their sales. Sentiment analysis of weblog data was used to predict movies' financial success [7]. Twitter[1] posts were also shown to be useful for predicting

[1] www.twitter.com

A. Holzinger and G. Pasi (Eds.): HCI-KDD 2013, LNCS 7947, pp. 77–88, 2013.

box-office revenues of movies before their release [8]. Thelwall et al. [9] analyzed events in Twitter and showed that popular events are associated with increases in average negative sentiment strength. Ruiz et al. [10] used time-constrained graphs to study the problem of correlating the Twitter micro-blogging activity with changes in stock prices and trading volumes. Bordino et al. [11] have shown that trading volumes of stocks traded in NASDAQ-100 are correlated with their query volumes (i.e., the number of users' requests submitted to search engines on the Internet). Gilbert and Karahalios [12] have found out that increases in expressions of anxiety, worry and fear in weblogs predict downward pressure on the S&P 500 index. Moreover, it was shown by Bollen et al. [13] that changes in a specific public mood dimension (i.e., calmness) can predict daily up and down changes in the closing values of the Dow Jones Industrial Average Index. In our preliminary work [14] we used the volume and sentiment polarity of Apple financial tweets to identify important events, as a step towards the prediction of future movements of Apple stock prices.

The paper follows a specific approach to analyzing stock price movements, contributing to the research area of sentiment analysis [15,6,16,17], which is aimed at detecting the authors' opinion about a given topic expressed in text. We use the term predictive sentiment analysis to denote the approach in which sentiment analysis is used to predict the changes in the phenomenon of interest. Our research goal is to investigate whether large-scale collections of daily posts from social networking and micro-blogging service Twitter are a suitable data source for predictive sentiment analysis. In our work we use the machine learning approach to learn a sentiment classifier for classification of financial Twitter posts (tweets) and causality analysis to show the correlation between sentiment in tweets and stock price movements. In addition, visual presentation of the sentiment time series for detection of important events is proposed. We analyzed financial tweets of eight companies (Apple, Amazon, Baidu, Cisco, Google, Microsoft, Netflix and Research In Motion Limited (RIM)) but due to space limitations, detailed analysis of only two companies (Google and Netflix) is presented in this paper.

The paper is structured as follows. Section 2 discusses Twitter specific text preprocessing options, and presents the developed Support Vector Machine (SVM) tweet sentiment classifier. The core of the paper is presented in Section 3 which presents the dataset collected for the purpose of this study, and the methodology developed for enabling financial market prediction from Twitter data. The developed approach proposes *positive sentiment probability* as an indicator for predictive sentiment analysis in finance. Moreover, by using the *Granger causality test* we show that sentiment polarity (positive and negative sentiment) can indicate stock price movements a few days in advance. Furthermore, since financial tweets do not necessarily express the sentiment, we have introduced sentiment classification using the *neutral zone*, which allows classification of a tweet into the neutral category, thus improving the predictive power of the sentiment classifier in certain situations. We conclude with a summary of results and plans for further work in Section 4.