Information Systems 69C (2017) 81–92

Contents lists available at ScienceDirect

Information Systems

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

Discovering public sentiment in social media for predicting stock movement of publicly listed companies

Bing Li[a,*], Keith C.C. Chan[a], Carol Ou[b], Sun Ruifeng[a]

[a] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[b] Department of Management, Tilburg School of Economics and Management, Tilburg University, Tilburg, Netherlands

ARTICLE INFO

Article history:
Received 6 September 2014
Revised 26 June 2016
Accepted 13 October 2016
Available online 02 February 2017

Keywords:
Social media analysis
Twitter
Stock prediction
Data mining
Sentiment analysis
Big data
SMeDA-SA
Parallel architecture

ABSTRACT

The popularity of many social media sites has prompted both academic and practical research on the possibility of mining social media data for the analysis of public sentiment. Studies have suggested that public emotions shown through Twitter could be well correlated with the Dow Jones Industrial Average. However, it remains unclear how public sentiment, as reflected on social media, can be used to predict stock price movement of a particular publicly-listed company. In this study, we attempt to fill this research void by proposing a technique, called SMeDA-SA, to mine Twitter data for sentiment analysis and then predict the stock movement of specific listed companies. For the purpose of experimentation, we collected 200 million tweets that mentioned one or more of 30 companies that were listed in NASDAQ or the New York Stock Exchange. SMeDA-SA performs its task by first extracting ambiguous textual messages from these tweets to create a list of words that reflects public sentiment. SMeDA-SA then made use of a data mining algorithm to expand the word list by adding emotional phrases so as to better classify sentiments in the tweets. With SMeDA-SA, we discover that the stock movement of many companies can be predicted rather accurately with an average accuracy over 70%. This paper describes how SMeDA-SA can be used to mine social media date for sentiments. It also presents the key implications of our study.

1. Introduction

Traditionally, public opinion in open societies can be studied through face-to-face, telephone or on-line surveys. Ever since social media sites, such as Twitter and the likes, have become popular, collecting and analyzing public opinions have never been any easier. Millions of Twitter users, for instance, post over 340 million messages, which are referred to as tweets, to the Twitter site everyday [1]. In most cases, these tweets represent different opinions expressed on different social, economic and political issues. Many people have started to consider social media sites like Twitter to be containing repositories for answers to all kinds of opinion poll questions. As a result, researchers have started to analyze the massive amount of social media data for public opinions on different issues ranging from product marketing to political preferences [2,3].

Ever since Milgram's work in 1967 reporting on a "small world experiment" performed to have identified a "six degrees of separation" [4] between people in a social network, researchers have started to investigate into social connections and the effect that they may have on public opinions and behavior. For example, some recent attempts have been made to analyze social data to see if movie revenues [5,6] or the trend of the Dow Jones Industrial Average (DJIA) [7,8] can be forecasted. Based on the results obtained, it is believed that the patterns embedded in social media data may provide the information needed to better understand and predict social events.

The traditional way of seeking out public opinions by the use of such tool as questionnaire survey has been effective but relatively time consuming and expensive. It is especially the case when public opinions have to be monitored continually. The popularity of social media platforms on which people exchange ideas and express opinions has provided valuable sources for sentiment to be understood relatively easily if there is an effective way that social media data can be analyzed. The enormous volume and diversity of the data that can be collected from social media sites present an excellent opportunity for the data to be mined to identify nuggets of knowledge that can be leveraged to understand public opinions and sentiments for predictions about specific events to be published. This approach of discovering opinions from various social communities facilitates the building of models that can reveal useful insights into the behavior of various stakeholders, for predicting future trends and can facilitate design of marketing and advertising campaigns [2,3].

* Corresponding author.
E-mail address: LIBing_backup@outlook.com (B. Li).

http://dx.doi.org/10.1016/j.is.2016.10.001
0306-4379/

In order to mine social media data for information that could lead to the understanding of opinions and sentiments, and to predict social events, several challenges need to be addressed. Firstly, it should be noted that social media data are collected according to time lines and are therefore temporal data. Many data mining methods handle time series data that are numerical in nature and cannot be used directly with temporal data that contain a lot of texts [9]. Mining temporal data thus requires different techniques and algorithms from those that are used to mine traditional time series data. Secondly, data from social media are usually textual data that are ambiguous. It is sometimes hard for sentiment to be understood and distinguished easily into good or bad, or positive or negative. Therefore, sentiment analysis of such ambiguous data requires an effective text mining method. Last, but not the least, social media data commonly contains billions of messages, requiring a proper database to store as well as a creative architecture to process and for them to be analyzed, data and text mining methods have to be implemented efficiently.

In this study, we attempt to address the above challenges in mining social media data. Specifically, we created a corpus of Twitter data to predict the movement of share prices of certain stocks in the stock markets in the U.S. We attempt to demonstrate how ambiguous temporal data collected from social media sites could be mined effectively and efficiently.

The problem of stock price prediction has been a popular research problem in the last two decades but not many approaches have been proposed to effectively tackle it [10]. For instance, prediction based on the assumption of Random Walk has not so far been very satisfactory. There has been some effort to focus on prediction based on detailed financial news analysis about listed companies [11] basing on such assumptions as the classical Efficient Market Hypothesis (EMH) [12]. However, predicting news trends have not been shown to be any easier. Consequently, it appears that any prediction based on unpredictable factors like financial news is likely to be arbitrary. Even though relatively higher prediction accuracy has been reported in some studies such as [6,7], these studies cannot be easily generalized since too many specific parameters and conditions are required for predictions to be made more accurately.

Motivated by the challenges and the practical significance, we have developed a novel approach, called Social Media Data Analyzer – Sentiment Analysis (SMeDA-SA), to mine ambiguous temporal social media data collected from Twitter to determine the movement of the US stock markets, namely NYSE and NASDAQ. It has been widely accepted by economists that there is a potential connection between a company's stock price and the information published about it [13]. Given that data about public opinion can be collected relatively easily from social media sites, we attempted to collect and mine such data to find out public sentiments about products and services to predict stock price of listed companies directly and indirectly. In this paper, we present details of SMeDA-SA which we develop for such a purpose.

To perform its task, SMeDA-SA takes several steps. First, we consider each tweet's structure as a combination of words and phrases. We apply neuro-linguistic programming (NLP) techniques to classify a tweet's sentiment into five categories (*Positive$^+$, Positive, Neutral, Negative, Negative$^-$).* We then use the concept of *adjusted residuals* [24] to identify interesting patterns between public sentiments and stock market prices. To evaluate the effectiveness of the proposed approach, we have performed a number of experiments. In our experiments, we selected 30 listed companies from different industries in NYSE and NASDAQ to test out how accurate prediction of stock movements can be made based on mining social media data using the proposed approach. For mining social media data for public sentiment, we had collected approximately 15 million records of Twitter data that mentioned these 30 companies either directly or indirectly by mentioning their products or services. For instance, for the company "Apple Inc", which we had selected, we looked for tweets that mention "AAPL" the stock market code for the company, as well as the keywords of its products, such as "iPad", "iTunes" and "iPhone", etc., and also product characteristics such as "CPU Speed" and "Color", etc. In order to identify correlation between the sentiment as reflected by the posting on Twitter and the movement of stock prices, we made use of the proposed algorithms to compute a degree of sentiment for each of the 30 listed companies that we have chosen to determine how much it is correlated with the price movement of selected stocks.

Following this introduction, in Section 2 we describe the background of the proposed work and review related literature. In Section 3, we present details of our proposed methods to tackle the research problem. Specifically, we describe the process of analyzing and extracting valuable information from the ambiguous temporal textual data collected from the social media platform. Section 4 presents our experiments and the results. We conclude this article and suggest directions for future work in Section 5.

2. Related works

Although social media analytics is becoming increasingly popular as a research topic, not much work has been done in discovering temporal patterns of ambiguous contents in social media and relate them to other time-dependent events that take place in the real world. The pioneers in this research domain are Jansen and his colleagues [14]. They have investigated how word-of-mouth advertisements in social media may change the information recipients' sentiments in the related brand and products. Their work provided insights on how ambiguous information from social media can be analyzed. Despite their effort, the potential of social media analytics remains very much unexplored. This is especially the case with the problem of the kind of time-varying temporal patterns in social media data that we are directly concerned with here.

There has been some previous effort to analyze blog contents to determine if they are correlated with any business performance indicators such as spikes in the sales volume of books [15]. There have also been some attempts to determine if movie ticket sales can be predicted based on social media contents. The predictions are primarily made based on meta-data information about the movies, including such information as the Motion Picture Association of America (MPAA) ratings, the genre, the number of screens on which the movie debuted, running time, release date and the presence of particular actors or actresses in the cast, etc. Based on such information, linear regression was used to predict earnings about the posted movies [5]. In [16], instead of linear regression, the prediction problem is treated as a traditional classification problem and artificial neural networks are used to classify movies into categories ranging from'blockbuster' to'flop'. Apart from the fact that they predicted the ranges instead of the actual sales volume of a movie, the use of these approaches does not seem to be able to allow very accurate models to be constructed for prediction. The accuracy of these models was tested and found to be relatively low.

Of the work related to social media analytics, it is worth pointing out that Asur and Huberman [6] mined the temporal data from social media. In their paper, they show how a model can be constructed based on the popularity rating of movies, as determined by relevant tweets, can be used to predict the actual box office revenue of a movie. The dataset that was used in the studies was collected from the Twitter site on an hourly basis and aggregated for analysis. In order to ensure that the tweets obtained all referred to a specific movie, the keywords obtained from a movie title were used for searching of relevant tweets over a period of three months. These tweets were then used for prediction and the accuracy of the predicted results were found to be higher than other methods as described in [17–19]. Their studies provided some