Chapter 3

Using Big Data Opinion Mining to Predict Rises and Falls in the Stock Price Index

Yoosin Kim
University of Texas – Arlington, USA

Michelle Jeong
University of Pennsylvania, USA

Seung Ryul Jeong
Kookmin University, Korea

ABSTRACT

In light of recent research that has begun to examine the link between textual “big data” and social phenomena such as stock price increases, this chapter takes a novel approach to treating news as big data by proposing the intelligent investment decision-making support model based on opinion mining. In an initial prototype experiment, the researchers first built a stock domain-specific sentiment dictionary via natural language processing of online news articles and calculated sentiment scores for the opinions extracted from those stories. In a separate main experiment, the researchers gathered 78,216 online news articles from two different media sources to not only make predictions of actual stock price increases but also to compare the predictive accuracy of articles from different media sources. The study found that opinions that are extracted from the news and treated with proper sentiment analysis can be effective in predicting changes in the stock market.

1. INTRODUCTION

With the development of ICT, emergence of the Internet, and the expansion of the smartphone-led mobile environment, the digital revolution is accelerating. A huge amount of digital data is being created from both daily and corporate activities. In recent years, the collective form of such large volumes of digital data has taken on the label of ‘big data’ and much interest has begun to be taken in their utiliza-
Using Big Data Opinion Mining to Predict Rises and Falls in the Stock Price Index

Using Big Data Opinion Mining to Predict Rises and Falls in the Stock Price Index

ion (Madden, 2012; Manovich, 2011). Particularly compelling is the need to find and take advantage of the rapidly expanding supply of available unstructured text data, as such information can directly and indirectly reflect certain tendencies of their respective authors, including their opinions, emotions, interests, and preferences (Pang & Lee, 2008).

Unstructured text data exist in diverse content forms, ranging from news articles, blogs, and social network service (SNS) posts, to information found in Q&A forums and Voice of the Customer (VOC) feeds. Particularly of interest to this study is the data found in news articles. The news media produce and digitize a large quantity of news articles daily, distributing them worldwide. Such mass-produced news articles influence all sectors of society, but are especially believed to be closely correlated with stock price. For one, while changes in the stock market, due to their accompanying complexities, cannot be explained solely by changes in the market’s fundamentals, one possible alternative explanation involves news. Even without any distinctive changes in the market fundamentals, the price can often fluctuate depending on the appearance of a particular news story. This may be due to the fact that news contains explanations of various real-world events, as well as predictions of future changes and directions in politics, society, and the economy. This close relationship between news and stock prices leads people to expect to learn of new investment opportunities and/or earn profits via news, and allows market participants to predict, albeit partially, the stock market fluctuations. It is thus feasible to imagine that with the proper analysis of news content, and the accurate distinction between favorable and adverse issues that arise in the stock market, stock price can be predicted, thereby creating economic profits (Fu, Lee, Sze, Chung, & Ng, 2008; Gillam et al., 2002; Mitchell & Mulherin, 1994; Mittermayer & Knolmayer, 2006; Schumaker & Chen, 2009; Sehgal & Song, 2009).

Although previous studies have indicated that news may influence stock prices (Fu et al., 2008; Mittermayer & Knolmayer, 2006), they mainly targeted news on particular events or individual corporate news, making it difficult to generalize these effects to the real world in which a myriad of news on different topics is produced and distributed daily. It is a challenge to identify not only which piece of news, among a continuous flow of a variety of news stories, is especially crucial in influencing stock prices, but how each news story brings about such effects. Moreover, a variety of news content, including daily market situations, future prospects, and evaluations of corporate performances, is written and disseminated in real time, and it is not easy to clearly determine whether such contents have actual causal impact on the market, and if so, whether those effects are positive or negative. After all, hard news attempts to cover both positive and negative aspects of the stock markets in an effort to keep the tone neutral, making it difficult to identify the underlying valence that may or may not be present (Kim, Kim, & Jeong, 2012).

Existing studies have attempted to analyze relatively simple news stories that cover particular events in conjunction with the stock prices responding to such news, as well as the link between stock price fluctuations and the presence of related news that could potentially have influenced such results. Given the slew of news that is being produced by diverse media platforms and sources, there has also been a recent increase in attention to the need for converting unstructured forms of massive amounts of data into meaningful information, but so far, only preliminary efforts have been made.

Thus, in order to overcome such limitations in the field, this study proposes the intelligent investment decision-making support model based on opinion mining, in which a continuous flow of massive news is regarded as unstructured text big data and leads to the automatic clipping and parsing of news and tagging of vocabulary sentiment, and consequently, inferences regarding the opinions on stock price fluctuations. The prototype experiment for sentiment polarity tagging developed and applied a subject-oriented sentiment dictionary specifically for the stock market domain, rather than a general purpose