Knowledge-Based Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

News impact on stock price return via sentiment analysis

Xiaodong Li [a], Haoran Xie [a,*], Li Chen [b], Jianping Wang [a], Xiaotie Deng [c]

[a] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong Special Adminstrative Region
[b] Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong Special Adminstrative Region
[c] Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 31 October 2013
Received in revised form 15 April 2014
Accepted 15 April 2014
Available online xxxx

Keywords:
News impact
Stock price return
Sentiment
Prediction
Experiment

ABSTRACT

Financial news articles are believed to have impacts on stock price return. Previous works model news pieces in *bag-of-words* space, which analyzes the latent relationship between word statistical patterns and stock price movements. However, news sentiment, which is an important ring on the chain of mapping from the word patterns to the price movements, is rarely touched. In this paper, we first implement a generic stock price prediction framework, and plug in six different models with different analyzing approaches. To take one step further, we use Harvard psychological dictionary and Loughran–McDonald financial sentiment dictionary to construct a sentiment space. Textual news articles are then quantitatively measured and projected onto the sentiment space. Instance labeling method is rigorously discussed and tested. We evaluate the models' prediction accuracy and empirically compare their performance at different market classification levels. Experiments are conducted on five years historical Hong Kong Stock Exchange prices and news articles. Results show that (1) at individual stock, sector and index levels, the models with sentiment analysis outperform the *bag-of-words* model in both validation set and independent testing set; (2) the models which use sentiment polarity cannot provide useful predictions; (3) there is a minor difference between the models using two different sentiment dictionaries.

1. Introduction

Stock market is an important and active part of nowadays financial market. Both investors and speculators in the market would like to make better profits by analyzing market information. Financial news articles, known as one major source of market information, are widely used and analyzed by investors. In *Big Data* era, the amount of news articles has been increasing tremendously. In front of such a big volume of news pieces, more and more institutions rely on the high processing power of modern computers for analysis. Predictions given by support systems could assist investors to filter noises and make wiser decisions. Therefore, how to model and analyze news articles so as to make more accurate predictions becomes an interesting problem.

Bag-of-words based approaches model news articles by vector space model which translates each news piece into a vector of word statistical measurements, such as the number of occurrences, etc. Machine learning models are then employed to capture the relationship between the word statistical patterns and the stock price movements. Although many *bag-of-words* based approaches have been reported to have prediction power in some previous works [14,41,42], they miss one important ring on the chain of the mapping from textual news articles to the final directional predictions, which is the sentiment of news. As shown in Fig. 1, one important step in the flow is that the news articles are first interpreted by investors and translated into market sentiments; the investors then make their decisions based on the sentiment interpretations; and market prices aggregate the actions of each investor and reflect them in the final price movements. Therefore, integrating the sentiment analysis into the prediction framework would become more critical.

Sentiment analysis models documents from sentiment dimensions. Instead of just measuring word frequency, each word (especially the *colorful* ones that have sentiment polarity) is decomposed and represented by a vector of sentiment features. For example, word "accelerate" can be represented by *strong* and *active* by using Harvard IV-4 psychological dictionary; using the same dictionary, word "accept" can be represented by *positive*, *submit* and *social relation* features. The number of the sentiment dimension is fixed in the dictionary. Each document can be represented by a vector of sentiment values by summing up the sentiment vectors of each word in the document. Making predictions based on the sentiment representation should have many advantages over *bag-of-words*:

* Corresponding author. Tel.: +852 68502340.
E-mail address: hrxie2@gmail.com (H. Xie).

http://dx.doi.org/10.1016/j.knosys.2014.04.022
0950-7051/

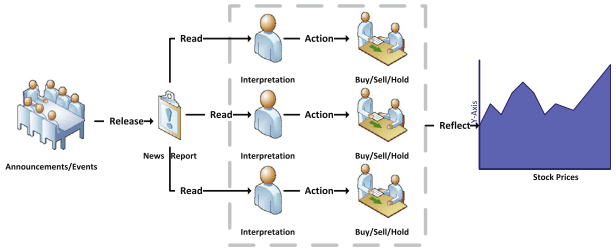

Fig. 1. The general scenario that news impact takes effect on the market prices. (1) Events happen; (2) events are reported; (3) reports are read by investors; (4) investors interpret the information according to their own knowledge; (5) investors take actions based on their interpretations, positions and budgets; and (6) various actions are translated into orders and reflected in stock price movements.

(1) **Reductive dimension**. Comparing with tens of thousands of words that are commonly observed while using words as the features, the sentiment representation largely reduces the dimensions to the order of hundreds, e.g., Harvard IV-4 psychological dictionary has 182 dimensions and Loughran–McDonald financial sentiment dictionary has 6 dimensions; (2) **Affective interpretation**. It is usually hard to interpret the mapping generated in *bag-of-words* space. In contrast, sentiment scores in different dimensions can sometimes give people more straightforward *feelings* about the document.

Although document sentiment analysis has been proposed and employed in many applications, e.g., recommender system [9], few work has been reported in the cross domain of computer science and algorithmic trading. In this paper, we first setup a generic framework that can take market information sets and plug in different customized prediction models. We then implement the prediction models using either sentiment analysis approach, *bag-of-words* approach or sentiment polarity approach, and compare their daily stock price return prediction accuracy on five years of Hong Kong Stock Exchange market data. In order to avoid any biases introduced by the framework, such as sentiment dictionary, the labeling method and the comparison at different market classification levels, etc., we (1) employ two sentiment dictionaries for comparison; (2) discuss and rigorously test different instance labeling methods and (3) compare models' performance at individual stock, sector and index levels in the experiment. The empirical results show that (1) at individual stock, sector and index levels, the models with sentiment analysis outperform the *bag-of-words* model in both validation and independent testing sets; (2) the models which use sentiment polarity cannot provide useful predictions; (3) there is a minor difference between the models using two different sentiment dictionaries.

The rest of the paper is organized as follows. In Section 2, we review the related work in both stock price prediction and sentiment analysis. In Section 3, we illustrate the work flow of the framework and do exploratory investigation on the data set and sentiment dictionary we use. In Section 4, we explain the setup of the experiment and show the parameter tuning related to the framework. And also, we give the experimental results and discussions. In Section 5, we provide our conclusion and future work directions.

2. Related work

The concept of *sentiment* has been known by people for a long time. It refers to *a specific view or notion* [12]. Sentiment analysis refers to *the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials* [16]. In this section, we review the related work about news impact and sentiment analysis in both finance and computer science domain.

2.1. News sentiment analysis in finance domain

The sentiment of news articles and their impacts on stock price returns have been studied in finance domain. Niederhoffer [27] analyzes *New York Times* and classifies 20 years of headlines into 19 predefined semantic categories from extreme-bad to extreme-good. He also analyzes how the markets react to the news of different categories and finds that markets have a tendency to overreact to bad news. Davis et al. [10] analyzes the effects of optimistic or pessimistic language used in news on firms' future performance. Their conclusion has two folds: (1) there is a bias between the readers' expectation and the writers' intension and (2) readers react strongly to both the content and the affective side of the reports which violate their expectations. Tetlock [36] extracts and quantifies the optimism and pessimism of *Wall Street Journal* reports, and observes that trading volume tends to increase after pessimism reports and high pessimism scored reports tend to be followed by a down trend and a reversion of market prices. Tetlock et al. also use Harvard IV-4 psychological dictionary in their work [37], where only positive and negative dimensions are exploited. They analyze the fraction of negative words in *Dow Jones News Service* and *Wall Street Journal* stories about S&P 500 firms from 1980 through 2004.

2.2. Bag-of-words approach in computer science domain

The *bag-of-words* approach has been applied to news impact analysis in financial market for many years. Seo et al. [35] build a TextMiner system (a multi-agent system for intelligent portfolio management), which could assess the risk associated with companies by analyzing news articles. Schumaker and Chen [34] build AZFinText system which is able to give directional forecast of prices based on financial news. *Bag-of-words* approach represents the textual news articles by term vectors and evaluates the "importance" of each term as their weights. After learning the mapping from the word statistical patterns to the outcome labels, the approach makes predictions for future unseen data.

2.3. Sentiment analysis in computer science domain

2.3.1. Sentiment dictionary construction

Applications evaluate word's sentiment mainly by constructing a sentiment dictionary. The construction approach could be briefly categorized as **semi-automatic** and **manual** which are described below:

> **Semi-automatic**. The dictionary is first constructed by some *seed* words that are manually selected. The dictionary is then expanded from the seeds by the rules defined by application.
> **Manual**. The dictionary is purely constructed by linguistic experts. This kind of dictionary is usually smaller in size than the one constructed semi-automatically, but more accurate.

Hatzivassiloglou and McKeown [17] make two hypotheses: (1) adjectives that are separated by "and" have the same polarity and (2) adjectives that are separated by "but" have opposite polarity. They use seed words and classify adjectives into positive and negative groups. Wiebe [40] also evaluates adjectives for polarity classification. He groups adjectives by word's tone and orientation clusters. Kim and Hovy [21] generate the polarity dictionary by using the WordNet to expand the selected seed words. They make two assumptions: (1) synonyms have the same polarity and (2) antonyms have the opposite polarity. The strength of a word

Please cite this article in press as: X. Li et al., News impact on stock price return via sentiment analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.04.022