

Title	株価動向予測のためのソーシャルメディアの感情分析
Author(s)	Nguyen, Hai Thien
Citation	
Issue Date	2015-09
Type	Thesis or Dissertation
Text version	ETD
URL	<a href="http://hdl.handle.net/10119/12962">http://hdl.handle.net/10119/12962</a>
Rights	
Description	Supervisor: 白井 清昭, 情報科学研究科, 博士

# Abstract

Sentiment analysis is important in academic as well as commercial point of views. It is a task to extract people’s opinions, attitudes and emotions toward entities. There are many applications of sentiment analysis such as investigation of product reviews, opinion summarization and stock market prediction. Specifically, opinion mining in the financial domain could help to make an accurate algorithm for stock market prediction. The goals of this research are: (1) studying the aspect-based sentiment analysis to identify the polarity of an aspect term or aspect category in a sentence, (2) studying the sentiment analysis on the financial domain to predict the stock market.

We propose a new topic model, Topic Sentiment Latent Dirichlet Allocation (TSLDA), to infer the topics and their sentiments simultaneously. With the observation that the topics are usually represented by nouns, whereas the opinion words are the adjectives or adverbs, words in sentences are drawn from distributions depending on its categories: topic category, opinion category and others. In addition, different topics, which are represented by word distributions, will have different opinion word distributions. Finally, to capture the sentiment meanings such as positive, negative or neutral of the opinion words for each topic, we distinguish opinion word distributions for different sentiment meanings. TSLDA is used for aspect-based sentiment analysis as well as sentiment analysis for stock market prediction.

To identify the sentiment categories for aspect terms in the first goal, an unsupervised method “ASA w/o RE” is considered. This model calculates the sentiment value of the aspect by summing over the scores of all opinion words divided by their distances to that aspect. However, the opinion words related to the aspects will have higher affection to the sentiments of them. We firstly extract the aspect-opinion relations by using a proposed tree kernel based on the constituent and dependency trees. Then, “ASA w/o RE” is extended as “ASA with RE” by integrating these aspect-opinion relations. The experiment results show that the integration of aspect-opinion relation extraction is useful for aspect-based sentiment analysis. Next, three supervised methods RNN, AdaRNN and our proposed PhraseRNN are investigated. RNN and AdaRNN convert a dependency tree of a sentence to a binary tree. PhraseRNN combines the dependency tree and list of phrases from the constituent tree to create a phrase dependency tree. This is further converted to a target dependent binary tree. In these three models, the representation model of the aspect is constructed by recursively combining the two child nodes into a parent node in bottom-up manner. The top node is used as the representation for the aspect and fed into a logistic regression to predict the sentiment category of the aspect. The results indicate that our PhraseRNN achieved better performance than unsupervised methods “ASA w/o RE” and “ASA with RE”. In addition, our PhraseRNN is better

5.35% accuracy and 7.89% F-measure than the ordinary RNN, 5.78% accuracy and 16.37% F-measure than AdaRNN. Therefore, our PhraseRNN is much effective than RNN and AdaRNN for the aspect-based sentiment analysis.

Two topic models JST and our TSLDA, are used to extract the sentiment for aspect categories in the first goal. By mapping the topics/sentiments inferred by JST or TSLDA to human topics/sentiments, the latent topics and sentiments are used to identify the aspect categories and their sentiments in each document. Our TSLDA outperforms JST model in almost all metrics in three datasets. As a result, our TSLDA is better than JST model for aspect-based sentiment analysis.

The second goal is to predict the stock price or movement using the sentiment analysis on social media. Stock price prediction is a very challenging task because the stock prices are affected by many factors. The Efficient Market Hypothesis and random walk theory said that it could not be predictable with more than about 50% accuracy. On the other hand, some researches specified that the stock market prices could be predicted at some degree. Around 56% accuracy are often reported as satisfying results.

With the assumption that integration of the sentiments from the social media can help to improve the predictive ability of models, we evaluate and compare three feature sets for stock price prediction and seven feature sets for stock movement prediction. Three employed methods to predict future stock prices are Price Only, Human Sentiment and Sentiment Classification. The first method uses only historical prices. The second combines the past prices of the stock with the sentiments annotated by posters, whereas third method uses both human and automatically classified sentiments. The results of regression models indicate that these sentiments are not useful to predict the future stock price. For the stock movement prediction, in addition to three previous models, additional four features are used: LDA-based Method, JST-based Method, TSLDA-based Method and Aspect-based Sentiment Method. Latent topics are extracted in LDA-based Method, whereas both latent topics and sentiments are exploited in the JST-based Method and TSLDA-based Method. In Aspect-based Sentiment Method, not latent but explicit topics and sentiments that appear in the sentence are incorporated into the prediction model. In addition, to address the question how automatic sentiment analysis contributes the prediction, we evaluate the automatically identified sentiment against the human annotated sentiment. The results show that our TSLDA-based and Aspect-based Sentiment Method outperform others in terms of the accuracy. The average accuracy on prediction of 5 and 18 stocks of TSLDA-based and Aspect-based Sentiment method are 56.43% and 54.41%, respectively. Besides, our method is comparable to the method using manually annotated sentiments. Therefore, the automatic sentiment analysis can be the alternative of the manual annotation. In addition, the important contribution of our experiment is that we evaluate our method for many stocks (18 stocks) and for a long time period of the test set (four months).

In future work, a nonparametric topic model for TSLDA which can guess the number of topics and sentiments by itself will be explored. In aspect term polarity identification, we will investigate the way to learn the weight parameters in “ASA with RE” method

from the training dataset to capture more accurately how the aspect-opinion relations contribute to the sentiment of the aspect. In addition, we will try to develop more sophisticated stock prediction model to also predict the degree of the change by setting more fine grained classes such as ‘great up’, ‘little up’, ‘little down’, ‘great down’ and so on.

**Keywords:** Text Mining, Sentiment Analysis, Opinion Mining, Stock Prediction, Social Media, Message Board, Tree Kernel, Topic Model, Recursive Neural Network, Support Vector Machine.