3J1-5in2

Stock price movement prediction using distributed representations of financial report

*1Yosuke Yasuda *2Junichiro MORI *3Ichiro Sakata

University of Tokyo, Engineering Department

In this research, we aim at predicting stock price movements based only on text data from financial reports by applying recent natural language processing techniques. In particular, we propose neural network-based document embedding models to represent text information from the large amount of financial reports to show the possibility that natural language processing approach can be used for stock price prediction. The contribution of this research is to show how to acquire proper text representation that can be useful to predict stock price movement.

1. Introduction

1.1 Background and Purpose

Financial reports, which explain accounting and financial information of companies, are important sources for investors to know about situations and financial performance of companies. In particular, public companies have to report periodic financial situations through the EDGAR database in the U.S. Therefore, financial reports are useful sources for stock price movement prediction. However, interpreting the contents of financial reports is difficult and time consuming because they include many facts that humans should judge their importance. Several studies have tried to extract what parts, in particular numerical information, are important to lighten the burden of human judges. While such approaches potentially support decision making for investors, there are still limitations. In this research, we are aiming at predicting stock price movements based only on text data from financial reports by applying recent natural language processing techniques. The documents have rich text information about how the business is going on. They must be related to the evaluation of the company and should be included to overcome the weakness of existing system trading. In particular, we propose neural network-based document embedding models to represent text information from the large amount of financial reports to show the possibility that natural language processing approach can be used for stock price prediction. Our proposed method enables stock price prediction models to include textual information in addition to numerical factors, which improves accuracy of automatic trading system eventually. The contribution of this research is to show how to acquire proper text representation that can be useful to predict stock price movement.

1.2 Related Works

Financial reports include both quantitative and qualitative factors. Several researches have suggested that incorporating quantitative and qualitative factors improves prediction accuracy of stock price movement [Lin 2011, Ding 2014]. We employ a distributed representation approach [Mikolov 2013] to combine both quantitative and qualitative factors in texts of financial reports to predict stock price movements.

Contact: Yosuke Yasuda, School of Engineering, The University of Tokyo, City Heights.201, 3-4-5, Ryusen, Taito-ku, Tokyo, 110-0012, Japan, E-mail: double.y.919.quick@gmail.com

2. Method

2.1 Data Set

Financial reports are downloaded from edgar (https://www.sec.gov/edgar/searchedgar/webusers.htm). Stock price data is downloaded from yahoo finance. Those data is separated into rise and sink based on the rule below.

Rise of stock price is defined as follows.

$$\text{if } \frac{peak - open_s}{open_s} > 0.03 \text{ and } \frac{average - open_s}{open_s} > 0.02 \quad (1)$$

Sink is defined as follows.

$$\text{if } \frac{open_s - drop}{open_s} > 0.03 \text{ and } \frac{open_s - average}{open_s} > 0.02 \quad (2)$$

They are ratio of stock price change in 2 days relative to open price of the publishing day.

2.2 Experiment

Keywords are selected by creating logistic regression from TFIDF and coefficients of logistic regression. Top 100 words are chosen as keywords from each rise prediction and sink prediction. Documents are filtered only with paragraphs including those keywords. Paragraph2vec model that learns whole text is created from texts that have paragraphs with positive keywords or negative keywords. Another model uses 2 paragraph2vec models. One learns texts with only positive paragraphs and another deals with negative parts. They are combined and used as document representations.

3. Result and Discussion

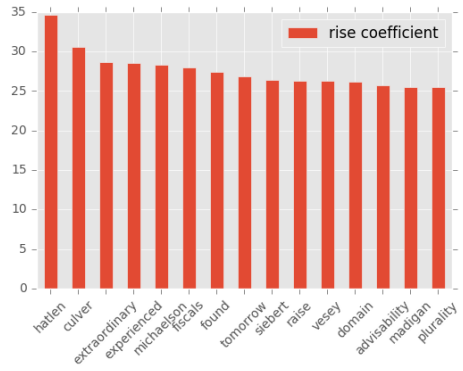

Figure 1: Rise keywords

These are top 15 positive keywords. Positive keywords like "extraordinary", "experienced" and "raise" are acquired. However, there are keywords that are clearly proper nouns.

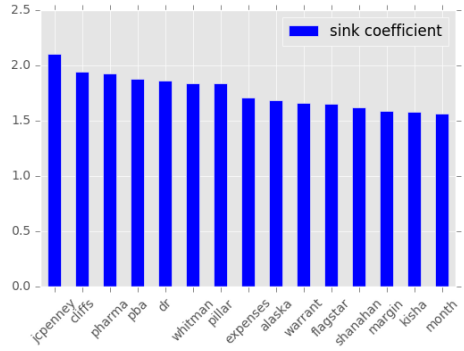

Figure 2: Sink keywords

"expenses", "warrant" and "margin" are chosen as relatively strong negative keywords. There are also proper nouns in the negative keywords.

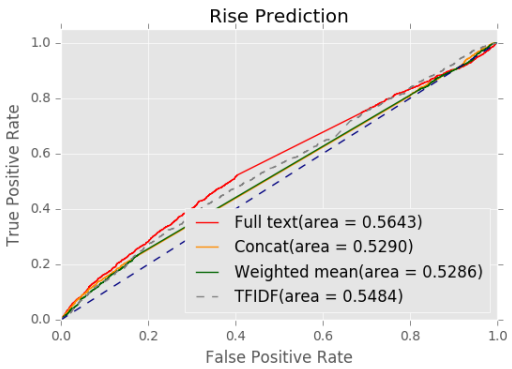

Figure 3: Rise prediction ROC curve

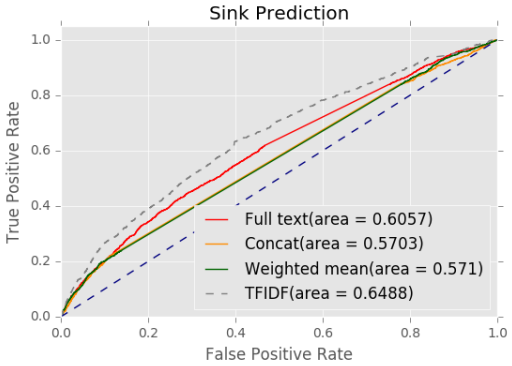

Figure 4: Sink prediction ROC curve

As for paragraph2vec models, full text model has the best performance, although TFIDF model has better performance in sink prediction. Concat and weighted mean are the models that learn positive and negative parts separately but performance was worse.

4. Conclusion

In this research, proper way of filtering financial reports that affects stock prices of companies and creating distributed representations of financial reports are examined. For filtering process, TFIDF representation and weights of logistic regression were used. As the result, it extracted some keywords that were likely to be related to the performance of companies. This can be a way to filter unnecessary part from documents for a task to be applied. To create distributed representations of the documents, 3 ways of creating paragraph2vec models were examined. One was to use whole text. Others used 2 paragraph2vec models that learned positive part and negative part separately. Then, they were combined. One used concatenation and one used weighted mean for merge. Among those three models, full text model had the best although TFIDF had better performance in sink prediction. This result indicates that neural network-based document embedding, which includes contextual information, can be useful for stock price prediction.

References

[Lin 2011] Lin, Lee and Kao, Chen: Stock price movement prediction using representative prototypes of financial reports(2011)

[Ding 2014] Ding, Xiao and Zhang, Yue and Liu, Ting and Duan, Junwen: Using Structured Events to Predict Stock Price Movement: An Empirical Investigation(2014)

[Mikolov 2013] Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey: Efficient estimation of word representations in vector space(2013)