Reproduce Stock Price Prediction Tasks with News Analysis

NYU FRE-7871 Final Project Proposal

Pei-Lun Liao
New York University
pll273@nyu.edu

ABSTRACT

In this project, selected related works will be reproduced with different financial news dataset. The goal is to understand techniques and challenges in stock price prediction.

1 INTRODUCTION

Nowadays, traders apply machine learning technique to extract information from financial news data. The information can be used to boost the accuracy of stock price prediction [1, 2, 9]. The project aims to reproduce selected related works [9]. Hence, people can learn the challenge and technique that are crucial in stock price prediction.

2 APPROACHES

2.1 Data

Data quality is crucial to make the accurate prediction [11]. Unfortunately, it is not easy to find a publicly available financial news dataset. For example, the public financial news dataset published in [1] was currently unavailable due to license issue [7]. Another corpus available online requires membership and subscription fee [8]. Also, it is not practical to crawl dataset from business news media in this one month project. Most of the time will be spent on data collecting, cleaning, and processing instead of understanding the technique that works for stock price prediction.

Fortunately, an available financial news dataset was found on Webhose.io [10]. The data was crawled from the Internet from July to October in 2015. 47,851 news articles were collected in machine-readable format. However, there is no paper shows the dataset could help stock price prediction.

2.2 Plan

2.2.1 *Stage 1: bag-of-words.* The goal in stage 1 is reproducing the experiment result in AZFinText system [9]. AZFinText represented article in the bag-of-words with only proper nouns. The dataset were Yahoo Finance news articles and the S & P 500 index. Reproducing the experiment in AZFinText can prove that the data from Webhose.io also works for stock price prediction. Moreover, we examined whether only using proper nouns helps the performance.

The experiment setup will follow the AZFinText paper. We will represent articles in the bag-of-words and use SVR to predict stock price. Finally, we evaluate the result by the rate of return. The strategy is buying the stock as the predicted price is greater than or equal to 1% movement from the stock price at the time the article was released. Then, sell the stock after 20 minutes. In stage 1, we will have four different results to compare.

- 1. Stock price
- 2. Stock price + News in bag-of-words
- 3. Stock price + News in bag-of-words without stopwords
- 4. Stock price + News in bag-of-words with only proper nouns

The expected performance will be 4 > 3 > 2 > 1 as the paper described.

2.2.2 *Stage 2: word and sentence representation.* In this stage, We are curious how modern deep learning techniques like word2vec, seq2seq, and CNN-LSTM language models could help improve the performance.

- 5. Stock price + News feature in average word2vec [5, 6]
- 6. Stock price + News feature in Skip-Thought vector[4]
- 7. Stock price + News feature in CNN-LSTM encoding [3] (may try to find pre-trained model)

The performance in stage 2 should be better than the one in stage 1.

REFERENCES

[1] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. In *EMNLP*.

[2] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep Learning for Event-driven Stock Prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 2327–2333. http://dl.acm.org/citation.cfm?id=2832415.2832572

[3] Zhe Gan, Yunchen Pu, Ricardo Henao, Chunyuan Li, Xiaodong He, and Lawrence Carin. 2016. Unsupervised Learning of Sentence Representations using Convolutional Neural Networks. (11 2016).

[4] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 3294–3302. http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[7] Philippe RÃĬmy. 2014. Financial News Dataset from Bloomberg and Reuters. https://github.com/philipperemy/financial-news-dataset. (2014).

[8] Evan Sandhaus. 2008. The New York Times Annotated Corpus. https://catalog.ldc.upenn.edu/LDC2008T19. (2008).

[9] Robert P. Schumaker and Hsinchun Chen. 2009. A quantitative stock prediction system based on financial news. *Information Processing & Management* 45, 5 (2009), 571 – 583. https://doi.org/10.1016/j.ipm.2009.05.001

[10] Webhose.io. 2018. Webhose.io. Available at: https://webhose.io/datasets/ accessed on 2018-09-17. (2018).

[11] Jinjian Zhai, Nicholas Cohen, and Anand Atreya. 2011. CS224N Final Project: Sentiment analysis of news articles for financial signal prediction. (March 2011).