News-Driven Stock Price Prediction Based on a Pretrained BERT Models

Mingxi Liu
MIDS UCB
mingxi.liu@berkeley.edu

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style file for ACL 2023. The document itself conforms to its own specifications, and is, therefore, an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

News-based stock price prediction usually consists of two tasks: sentiment analysis of the news and predicting the price movement based on the sentiment. With the help of all kinds of pre-trained BERT models, sentiment analysis has become less challenging. However, accurate sentiment analysis does not necessarily bring accurate stock price prediction. This paper compares several predicting methods based on either the sentiment analysis results or the pooler output of a BERT model pretrained on financial news and articles. The best result comes from the model which has a simplified transformer structure and combines historical news and stock price returns.

2 Background

It's no news that news is used to predict stock price movements(e.g. **?**). Since it's an interdisciplinary task involving NLP and financial engineering, the solution evolves with the progress on either side. On the NLP side, the essential problem is how to extract numeric information from the text of news. Early work used the bag-of-words as text representation to perform sentiment analysis for the news. For example, **?** examines a predictive machine learning approach for financial news article analysis using several different textual representations: bag of words, noun phrases, and named entities. Later work adopted word embeddings such as Word2Vec.**?** propose a Hybrid Attention Networks to predict the stock trend based on the sequence of recent related news and employ self-paced learning for effective and efficient learning. They use a pre-trained Word2Vec embedding layer to calculate the embedded vector for each word and then average all the words' vectors to construct a news vector. Other techniques include character-based neural language model (**?**) and topic modeling (**?**). Since BERT became the state-of-art NLP model, researchers have been trying to apply it to the task above(see **?** and **?**).

As with other applications of BERT, fine-tuning may further improve the performance. A lot of recent works have been focused on further training BERT on financial domain corpora(see **?**) or even training BERT from scratch on financial domain corpora(see **?** and **?**). **?** points out that continual training from the original BERT model is more effective than domain-specific pretraining from scratch. This paper proposes a third option: fine-tuning with the stock price data. This can be done by retraining certain layers of BERT during the training for the prediction task. The process is similar to sentiment analysis, only the sentiment is labeled by the change in stock price.

Focusing on the NLP part instead of the financial engineering part, this research aims to compare the performance of the following four model settings in predicting the direction of stock price movement:

1. the original BERT;

2. BERT further trained with financial domain corpora;

3. BERT trained from scratch with financial domain corpora;

4. BERT to be fine-tuned with stock price data.

3 Methods

3.1 Data

This paper experiments on predicting the daily change of SP 500 Index based on financial headline reports of Reuters. In order to get as many samples as possible, two sources of news data are combined together. One is from **?**, which covers the period from October 2006 to November 2013, totally 105,343 pieces. This time span witnesses a severe economic downturn in 2007-2010, followed by a modest recovery in 2011-2013. The other is provided by Kaggle, which includes 32,770 Reuters news from March 2018 to July 2020, which also covers a volatile period during Cov-19. Since both sources are from Reuters and are comparable in frequency and length, we think it's reasonable to combine them together. The gap between 2013 to 2018 is not a big issue if we use the first source as training set and the second validation/test set.

The lead paragraph (normally the first paragraph) of each news is selected as the input. It typically contains 30 50 words, providing more complete information than the title which contains less than 10 words and requring less computing resources than the full content.

The SP 500 Index price data comes from Yahoo Finance. It contains daily series cover both periods above. The log returns of the index is used as the predicting traget. The size of the labels is 2,393 after aligning the dates with the news data.

Note that the number of news is more than 50 times of the number of labels. This is because there are usually 40 to 60 news per day. Therefore, we need to aggregate the information of news of the same day. Different methods will be dicussed in the following sections.

3.2 Experiments

In this section, we compare five methods of extracting and processing the information from news text. One of them applies lexicon-based sentiment analysis, and the rest are based on the sentiment output or pooler output of FinBert, a BERT model pre-trained on financial contents.

3.2.1 Lexicon-based

1) a lexicon-based sentiment analysis; 2) sentiment analysis output by FinBert a BERT model pre-trained on financial contents; 3) predicting based on pooler output of FinBert model; 4) predicting based on pooler output of FinBert model with attention mechanism; 5) predict based on pooler output

Each tweet is embedded with one of the BERT variants above. The output vector for the [CLS] token goes to the next step (subject to change).

3.3 Daily aggregation

Since there may be more than one tweet for a stock in a day, the information from these tweets must be aggregated to generate one signal. RNN, attention-based weighted average, or simple average may be used for the aggregation, depending on time and computing resources.

3.4 Prediction

The final layer of the model is a Softmax layer predicting the next day's price movement in three direction categories: up, down, or flat(within a certain range). Metrics such as loss and accuracy can be compared at this step.

3.5 Trading simulation

According to the prediction from the model, we can decide to buy, sell, or do nothing with the targeted stock. Then we can calculate this simulated investment return and evaluate its investing performance with financial metrics such as the Sharpe Ratio, which are more important than loss or accuracy for investment.

3.6 Customized loss function (Optional)

The last two steps can be combined together using a customized loss function which is directly related to the investment performance. Models trained under this customized loss function may generate better investment results.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1415–1425, Doha, Qatar. Association for Computational Linguistics.

Leonardo dos Santos Pinheiro and Mark Dras. 2017. Stock market prediction with deep learning: A character-based neural language model for event-based trading. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 6–15, Brisbane, Australia.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 261–269, New York, NY, USA. Association for Computing Machinery.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization. Special Track on AI in FinTech.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364, Beijing, China. Association for Computational Linguistics.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is domain adaptation worth your investment? comparing BERT and FinBERT on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. 2021. Quantitative day trading from natural language using reinforcement learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4018–4030, Online. Association for Computational Linguistics.

Robert P. Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2).

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Gakuto Tsutsumi and Takehito Utsuro. 2022. Detecting causes of stock price rise and decline by machine reading comprehension with BERT. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 27–35, Marseille, France. European Language Resources Association.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.