Stock Price Prediction with Social Media Comments*

Niyang Bai

Background

With the development of the technology, social media nowadays affects almost all sectors of the society, including financial market. In January 2021, GameStop (NYSE: GME) experienced a sustained short squeeze, which had a significant financial impact on some hedge funds in the market. The short squeeze was primarily triggered by users of a discussion board named `r/wallstreetbets` on Reddit, a famous social networking site, mostly through free trading applications. During this period, as a result of the incidence, `r/wallstreetbets` saw a spike in visits, peaking at 73 million visits in 24 hours.

It is reasonable to doubt that the change in GameStop's stock price might be related to the intense discussion on social media. So, in this project, I want to build a natural language processing (NLP) based convolutional neural network (CNN) to predict stock prices.

For NLP itself, Gan et al.[1] built a sentence-based convolutional neural networks. Kiros et al.[2], on the other hand, built a BERT method to study web community data. For the application of NLP in financial forecasting, Frank et al. [4] reviewed articles from 1980 onwards. Many of them used Twitter as a source of information because of its small size and ease of analysis. Oncharoen et al.[3] also built a NLP based deep learning structure to predict the S&P 500 index.

Objectives

In this project, I plan to use natural language data from `r/wallstreetbets` during the GameStop's short squeeze event to build a convolutional neural network that can predict the GameStop stock prices, and then compare it with classic time series predictions to ensure the validity of the model.

In terms of data, stock prices can be obtained directly from internet sites like the Google finance, and similarly, with the help of Reddit API, all discussions and comments during the incident can be captured. For deep learning, I plan to divide the work into several specific parts. The first is data pre-processing, which requires uniform coding of all textual information. Then comes sentence-

*for more information: `https://github.com/niyangbai/master_thesis`

and document-based embedding, followed by the construction of deep learning structures. For the time series analysis (baseline), I plan to estimate a classic time series model on a longer period of time without introducing the impact of social media.

For NLP part, packages like `spaCy` and `NLTK` are necessary, and for deep learning tools, packages like `scikitlearn`, `PyTouch` and `Scipy` can be very helpful.

Regarding potential difficulties, in terms of data processing, unlike Twitter, Reddit does not impose a limit on the number of words users can comment, which can cause very huge data load. Also, commonly used slang and network terms can also lead to biased prediction as they can be very hard for computers to recognize. For the deep learning part, different deep learning structures need to be explored and also further compared. Regarding the classic time series analysis, more information might be necessary besides only stock price.

References

[1] Zhe Gan et al. "Unsupervised learning of sentence representations using convolutional neural networks". In: *arXiv preprint arXiv:1611.07897* (2016).

[2] Ryan Kiros et al. "Skip-thought vectors". In: *Advances in neural information processing systems* 28 (2015).

[3] Pisut Oncharoen and Peerapon Vateekul. "Deep learning for stock market prediction using event embedding and technical indicators". In: *2018 5th international conference on advanced informatics: concept theory and applications (ICAICTA)*. IEEE. 2018, pp. 19–24.

[4] Frank Z Xing, Erik Cambria, and Roy E Welsch. "Natural language based financial forecasting: a survey". In: *Artificial Intelligence Review* 50.1 (2018), pp. 49–73.