CS230

Long-Term Stock Prediction using Quantitative and Text-Based Data

Category: Natural Language Processing & Recurrent Neural Network

Dylan M. Crain
Department of Energy Resources Engineering
Stanford University
cooper96@stanford.edu

Project Description

Stock market prediction is a challenging and fairly well-researched topic within deep learning. Due to the *Efficient Market Hypothesis*[3], all available public data should already be incorporated into the current stock price for a given company. This means that it should be unfeasible to garner any predictive information from these trends. However, according to my literature review, this appears to not be the case in reality.[3][4] Using stock price data can lead to excellent next day predictions.

Furthermore, there has been work on honoring the *Efficient Market Hypothesis* by taking in news reports or even Tweets to inform a deep learning network that predicts future stock prices.[2] What is not as common, though there are cases such as in reference [1], is the idea of combining quantitative information, e.g., end of day price & volume, with text-based data, such as news articles or social media posts.

Moreover, I have currently found no source that combines this data space with longer term stock prediction, i.e., beyond just the next day. Therefore, my goal in this project is to use both bases of data (text as well as quantitative) to build either a Recurrent Neural Net or Long-Short Term Memory structure to predict stock prices at different time frames, i.e., not just the next day. These methods (RNN & LSTM) are chosen, since they appear to be well used for such a problem due to the time series nature of stock prices. If time and results permit, I would like to use this model to possibly predict economic downturns close to a week in advance.

Challenges

The challenges for this project are immense. My initial worry is that the scope may be too large. To rectify this, I will continue gathering information as I go and possibly cut functionality to make sure I have a clean and interesting final project.

Data Collection

Consequently, data collection will also be a challenge. It is possible to get end of day stock prices (as well as volumes) for the companies under NASDAQ back 10 years at the following link: https://www.nasdaq.com/market-activity/quotes/historical.

The more challenging data collection will be the news articles and/or social media posts. This can be done, however. When I took CS229, my project revolved around predicting political bias from news articles based on Tweets from U.S. Senators, so I have some experience with this data gathering. Some work will have to go into cleaning this data, for sure.

Results Evaluation

The results will be evaluated on accuracy of predicted stock prices in the NASDAQ grouping for differing time periods: 1 day, 3 days, 1 week, et cetera. It is certainly possible that these results can be conglomerated into one value, such as the mean squared error of all of these end day errors for a given time period prediction.

Furthermore, if time permits getting to the economic downturn prediction, it can be displayed how far out from the most recent crash due to Covid-19, the model can accurately predict generally falling prices.

Reading for Context & Background

Some of the references used to come up with this proposal are listed in the next section. However, I have only scratched the surface on this front. There is plenty of additional room to explore other sources while I collect and clean data for the upcoming network.

Additionally, a good source of material was from previous CS230 projects. Reading these sources, I hope to learn from my predecessors successes and failures. I found it curious that many teams submitted projects involving stock market prediction, yet none have attained the coveted "Outstanding Project" title. A goal of mine will be to try and change this trend.

References

[1] R. Akita, A. Yoshihara, T. Matsubara and K. Uehara, "Deep learning for stock prediction using numerical and textual information," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016, pp. 1-6, doi: 10.1109/ICIS.2016.7550882.

[2] 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)*. Association for Computing Machinery, New York, NY, USA, 261269. DOI:https://doi.org/10.1145/3159652.3159690

[3] F. Kamalov, L. Smail and I. Gurrib, "Stock price forecast with deep learning," *2020 International Conference on Decision Aid Sciences and Application (DASA)*, 2020, pp. 1098-1102, doi: 10.1109/DASA51403.2020.9317260.

[4] Stoean C, Paja W, Stoean R, Sandita "Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations". PLOS ONE 14(10): e0223593. https://doi.org/10.1371/journal.pone.0223593