Headlines and Market Trends

Executive Summary - Deep Learning Project - Erdős Institute, Summer 2024

Research Question: Can we use news article sentiment in order to better predict stock price movement?

Team: Jem Aizen Guhit | Nawaz Sultani | Samson Johnson | Saeid Hajizadej

GiTHub: https://github.com/jguhit/Erdos-2024-DL-Newsworthy

Approach:

- Explore news sentiment data to find features that correlate with stock price movement.
- Develop and validate predictive models that use data from financial news headlines, alongside features from stock data, to predict the stock price movement 15 stocks across three sectors
- Run a simulation of an investment portfolio directed by the best model, and evaluate how it performs. Compare the model's yield to the performance of a simple "Buy & Hold" strategy.

Data Gathering & Processing:

- Using Stock News API and Alpaca API, 5 years worth of news & stock data was collected, corresponding to the timespan: 2019/03/15 - 2024/03/15.
- Stock data was collected at 15 minute intervals over extended market hours. News articles were matched to the closest stock time after publishing time of the article.

Sentiment Analysis:

- Sentiment analysis scores were obtained on article headlines via FinVADER, base RoBERTa, a pretrained RoBERTA model trained on financial terms, and a custom fine-tuned RoBERTa model.
- To train the customized RoBERTa model, the OpenAI API was used to obtain sentiment classifications (positive, neutral, negative) and sentiment scores on the collected articles which served as ground truth labels.
- The fine-tuning process involved training the model initially with 100 epochs to find the epoch where the model has the lowest validation loss. It was trained again to fine-tune the learning rate and the best model was saved and used to predict on the whole dataset.
- The fine-tuned model performance was compared with the RoBERTa model (not fine-tuned) and we gained a 35% increase in accuracy.

Modeling :

- Train-Test-Split: From the data set that spans 5 years, the last year of data (March, 2023 - March, 2024) was kept aside as the testing set. Among the first 4 years, the last year (March, 2022 - March 2023) was broken further into four 3 months increments as validation sets.
- Baseline Model: Buy and Hold, Fiducial (Variance Balanced)
- Other Models: LSTMs, CNNs, and Transformers (Informer model).
- Time Series Data was appropriately transformed into images for input into the CNN

Final Results and Conclusions:

- Developed a fine-tuned RoBERTa model that returned 89% accuracy on returning ground truth labels for article sentiment, a 35% increase compared to the base RoBERTa.
- Fiducial baseline model gave -5.08% annual growth on the simulation portfolio, while LSTM gives 1.2% annual growth.
- Have working models for CNN and Informer, but still need to be finetuned for simulation use.

Future Developments:

- Finish development of the CNN and Informer models
- Understand the causality of sentiment from news affecting stock price and vice versa
- Generalizing models to capture sector-wise trends rather than those related to a single stock