Paper Trading From Sentiment Analysis on Twitter and Reddit Posts

Eden Wang (eyw@stanford.edu), Chinmaya Andukuri (andukuri@stanford.edu), Shobha Dasari (sdasari1@stanford.edu)
CS 224N, Winter 2023

Problem

Through large pre-trained word-embeddings models and data from social media platforms, researchers have been investigating the validity of the widely-accepted **Efficient Market Hypothesis**. This hypothesis states that stock market prices are driven by new information and follow a random-walk pattern. Therefore, stock prices cannot be predicted, since new information constantly enters the market. Solutions to this problem are of interest in financial engineering.

Background

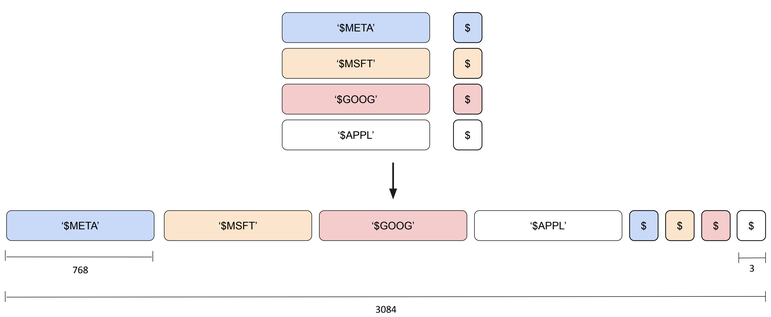
Many papers develop NLP neural network models to predict stock movement based on financial sentiment, typically using only Twitter data and predicting the output of a single pre-selected stock price. We extended these current approaches to use both Twitter and Reddit data and make predictions for multiple stocks tickers.

Our goals were to investigate potential applications of neural network-based NLP methods to:

1. Evaluate and generalize community sentiment (in Reddit and Twitter comments) towards the most popular stock tickers.
2. Understand the potential influence of public sentiment and community discussion on multiple stocks' price movement.
3. Develop an intelligent portfolio strategy based on predictions for multiple stock tickers.

Data

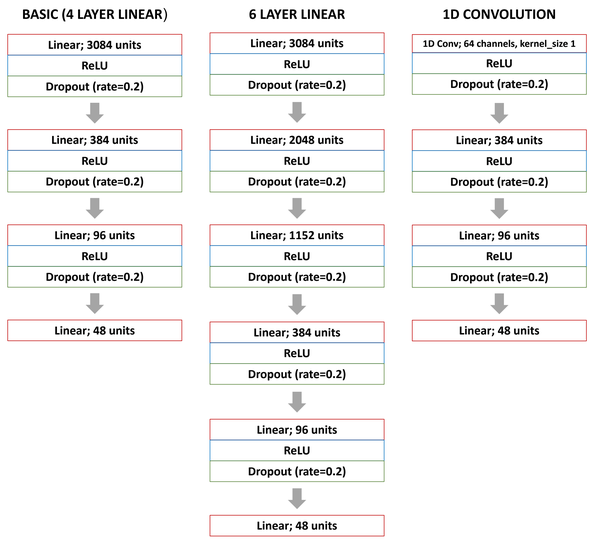
- Using text and stock price data for Meta, Google, Microsoft, and Apple
- 150,000 Tweets and Reddit posts from December 1, 2022 to February 28, 2023
  - Web-scraped comments that specifically mentioned the stock tickers names "$META," "$GOOGL," "$MSFT," and "$AAPL"
- Corresponding daily stock data scraped from Yahoo! Finance and NASDAQ data, for ground-truth mean and covariance data on the change in daily stock prices



Methods

The model's inputs are clustered sentence embeddings (generated by SBERT and performing k-means), concatenated with historical stock mean vector and covariance matrix over the past *d* days. The model's outputs are the mean vector and Cholesky factorization of the GT covariance.

We evaluated three model architectures:



Experiments

Task: Predict average stock prices and pairwise covariances for each of 4 companies, for day *i*, given:

- Embeddings for a sample of online mentions for companies from day *i*
- Stock prices for days *i - 1, i - 2, i - 3*

Results from Predicted Paper Trading Schemes

Baseline (uniform investment distribution): 0.0059% mean daily returns
Experiment 1 (4 linear layers): 0.0070% mean daily returns
Experiment 2 (6 linear layers): 0.0078% mean daily returns
Experiment 3 (1D conv + 3 linear layers): 0.0081% mean daily returns

Model Simulated Returns & Variance

Portfolio optimization was simulated with various risk-aversion parameters. Tables for our results (with varying gamma=0,1,...,5) can be seen below:

Gamma	Mean	Variance
0	0.00690	0.000328
1	0.00688	0.000325
2	0.00672	0.000315
3	0.00654	0.000313
4	0.00651	0.000313
5	0.00615	0.000303

Table 1: Exp. 1: 4 Linear Layer

Gamma	Mean	Variance
0	0.00758	0.000358
1	0.00746	0.000348
2	0.00716	0.000341
3	0.00678	0.000333
4	0.00647	0.000315
5	0.00614	0.000289

Table 2: Exp. 2: 6 Linear Layer

Gamma	Mean	Variance
0	0.00791	0.000358
1	0.00779	0.000345
2	0.00791	0.000352
3	0.00708	0.000330
4	0.00741	0.000319
5	0.00690	0.000293

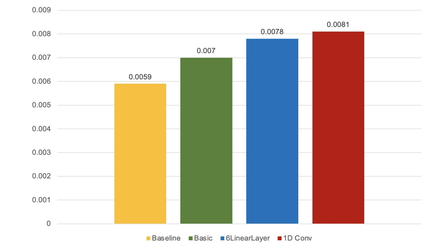
Table 3: Exp. 3: 1D Conv + Linear

Analysis

Compared to the baseline, the Basic Stock Prediction model produced higher mean daily returns. The 6LinearLayer model produced higher returns than the Basic model, and the 1DConvolutionLayer model produced the best returns of the three models.

Generally, returns decreased in all model architectures as risk tolerance decreased, with the 6LayerLinear model having the largest decrease in yield of all three model architectures.

Mean Yield on Investment from Different Model Architectures



Mean Yield on Investment for Model Architectures at Varying Risk Tolerances

Discussion / Learning Points

- The results provide evidence against the Efficient Market Hypothesis: NLP neural network-based models *can* achieve greater returns than uniform investment strategies given only recent financial history and sentiment.
- The 1D Convolution Model performs best in optimized portfolio prediction.
- Quality of text data is heavily dependent on availability of efficient and well-documented scraping tools.
- Weighted sampling necessary to combat sparsity in raw data when generating training pairs.

Future Work

- Extending stock prediction model to a wider variety of companies or stock indices through generation of additional data.
- Expanding the search keywords through query engineering during the data collection process to increase scope of inputs to the model.