Predicting Stable Portfolios using Machine Learning

Muhammad Rafay Aleem, Nandita Dwivedi, Kiran Rawat

Simon Fraser University

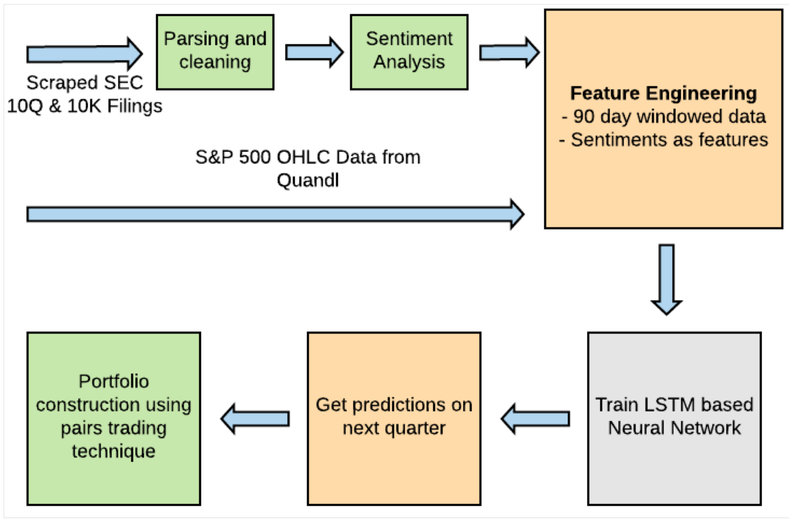
MOTIVATION AND GOALS

- Investment firms that manage equity portfolios for their clients predict stable portfolios for risk free returns.
  Can we make this better using Machine Learning / Deep Learning?
- US firms file quarterly reports (10Q) and annual reports (10K) with Security and Exchange Commission (SEC). As first part of this project we try to establish to what extent these financial reports of S&P 500 companies have implications on future returns using sentiment analysis.
- We predict quarterly stock prices by using deep learning on OHLC data and sentiments of SEC filings.
- We construct stable portfolios using these quarterly predictions on historical stock prices.

APPROACH

- **Data Extraction and Pre-Processing:**
  - Scraped, cleaned and parsed S&P500 10-Q and 10-K filings from SEC EDGAR for last 10 years. (approx. 60GB)
  - Retrieved OHLC data for the past 10 years from Quandl.
  - Used Pandas and NumPy.
- **Sentiment Extraction:** Leveraged NLP to extract sentiment scores using NLTK VADER API on selected SEC filings using financial lexicons.
- **Predictive Modeling:** Trained deep learning models using Keras for stock price prediction.
- **Portfolio Generation:** Selected stocks to construct optimized and stable portfolios using pair trading strategy for next quarter.

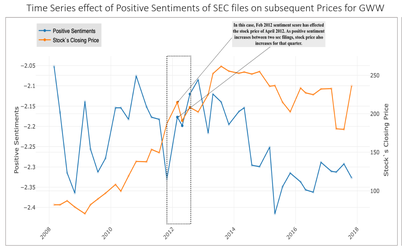
DATA PIPELINE



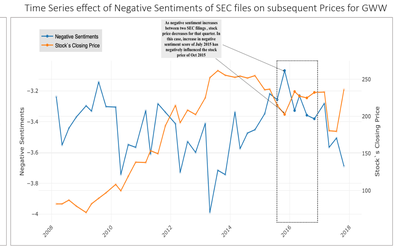
SENTIMENT ANALYSIS

- We used NLTK VADER sentiment analyzer that is based on lexicons of sentiment-related words.
- Vader lexicons were updated with financial lexicons from Loughran-McDonald Financial Sentiment Word Lists.
- We obtained Positive, Negative and Neutral scores for each SEC filing.(approx. 4600 SEC files)
- To reduce the sentiment calculation time content was split into batches of 2000 words.
- Mapped quarterly OHLC data for chosen tickers with the 10K and 10Q SEC filings' sentiment scores using a forward window.
- Analyzed the effect of SEC sentiments on the stock prices.

Time Series effect of Positive Sentiments of SEC files on subsequent Prices for GWW

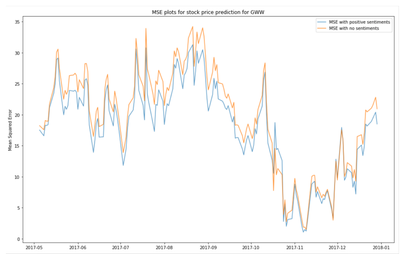


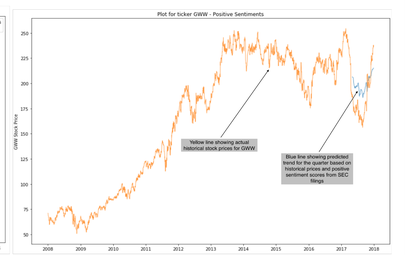
Time Series effect of Negative Sentiments of SEC files on subsequent Prices for GWW



STOCK PRICE PREDICTION

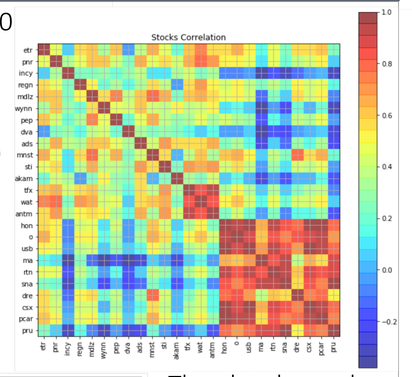
- A separate LSTM model was trained (20 epochs) for 50 companies from the S&P 500 index.
- Training was done on OHLC data along with temporal and sentiment features.
- We created 90 days window to map OHLC data to quarters and train each model using it.
- Train/Validation/Test set split was 2087/261/130 for each model. Test set was mapped to the next financial quarter.
- Models showed improvement with sentiments, proving that sentiments do influence stock prices.
- The results were not consistent across all models, concluding that SEC sentiment scores are not the only factors influencing prices for the next quarter.

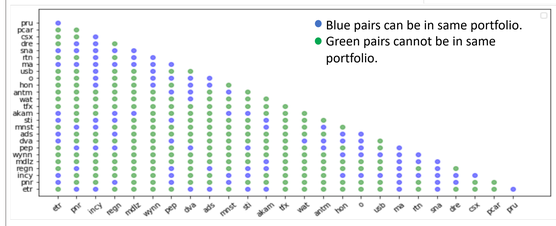




PORTFOLIO OPTIMIZATION

- Portfolios were constructed for 50 stocks (visualizations for 25 stocks is shown in the plots here)
- We performed pair trading on 50 stocks by generating all possible pair combinations. (1225 pairs)
- We included the stocks with correlation < 0.5 and covariance less than mean covariance in same portfolio.

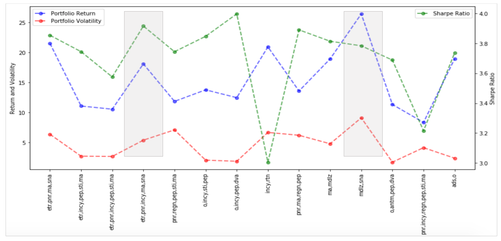




- The plot shows that tickers with blue markers have less correlation and covariance values, and can be paired.
- We generated portfolio return, portfolio volatility and Sharpe Ratio for each portfolio by taking most efficient set of weights out of 200 randomly generated weights.

	Portfolio	Weights	Sharpe Ratio	Portfolio Return	Portfolio Volatility
3	[etr, pnr, incy, ma, sna]	[0.05172413793103448, 0.11001642036124795, 0.1...	3.916348	18.10	5.382979
10	[mdlz, sna]	[0.10251450676982592, 0.8974854932301741]	3.784285	26.43	9.110109

The portfolio: [etr, pnr, incy, ma, sna] is a less risky weighted investment with a return of 18.1%



The portfolio: [mdlz, sna] is a more risky weighted investment with a return of 26.43%

CONCLUSION AND FUTURE WORK

- On training the model for only 20 epochs we could see that adding sentiments extracted from SEC filings was influencing the predictions for most of the stocks.
- Training for more epochs might produce more interesting results.
- We extracted only positive, negative and neutral sentiments from SEC filings, other sentiments like uncertainty, litigious, constraining, superfluous can be extracted and analyzed.
- We used pair trading for stable portfolio selection on predicted prices, this gave considerably good results. Other rebalancing strategies can be investigated.