Portfolio Optimisation using Machine Learning

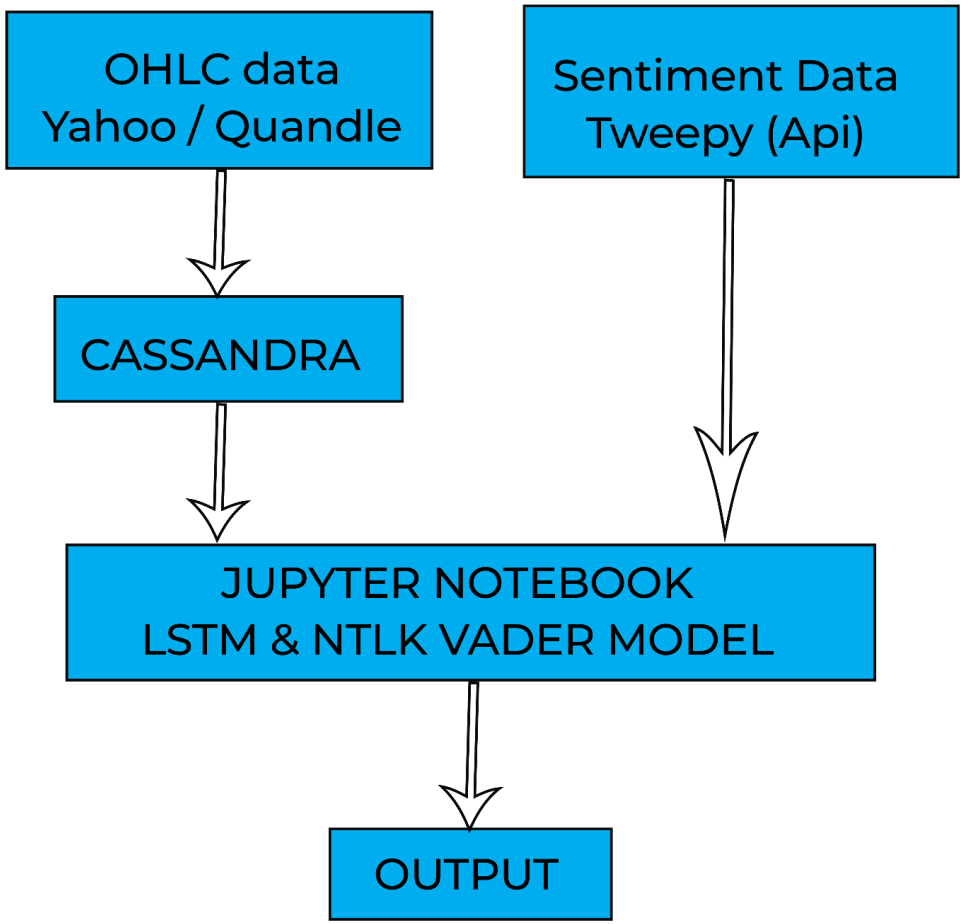
By: Tanay Dubey

Mentors: Bharath Vaddi & Satish Padmanabhan

Problem Statement

- Using historic stock price data to predict future returns.
- How does including the sentiments of financial news report change the prediction of stock prices.
- Building a stable portfolio based on these predicted values.
- Rebalancing the portfolios and optimising the rebalancing time window for optimal profits.

Technology Stack



Data Collection

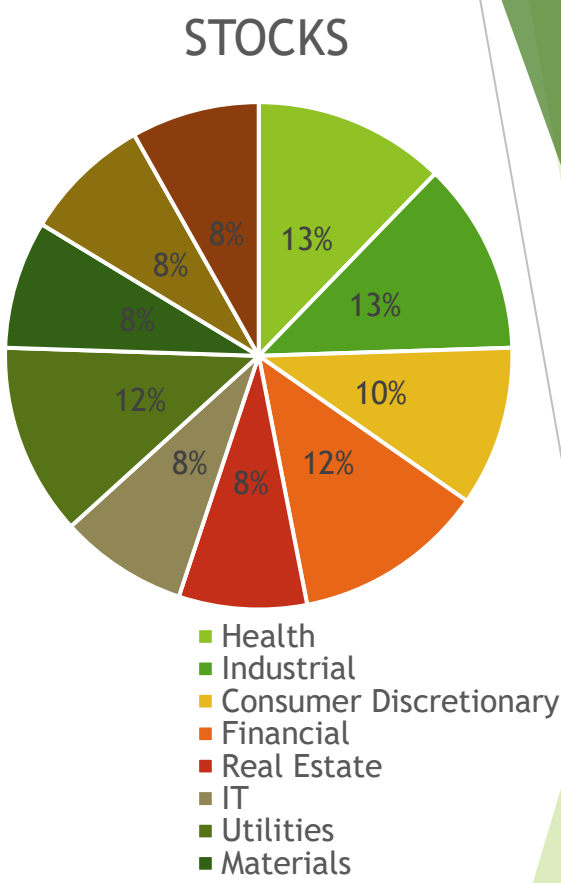
- OHLC data from Yahoo Finance

OHLC (open-high-low-close) is a type of dataset that provides open, high, low, and close prices for a stock for each and every business day from 2005. This was collected for a pool of 50 stocks.

The data was imported in **Cassandra** and was exported directly out of it for further use.

- Sentiment Analysis data

News headlines were collected for last two years using selected accounts from twitter using tweepy Api.



Correlation between various sectors

	Auto	Consumer Durables	Capital Goods	FMCG	Health Care	IT	Metal	Oil & Gas	Bankex	Power
Auto	1.00	0.77	0.65	0.62	0.83	0.86	0.79	0.48	0.74	0.56
Consumer Durables	0.77	1.00	0.77	0.60	0.78	0.69	0.73	0.62	0.85	0.69
Capital Goods	0.65	0.77	1.00	0.52	0.54	0.57	0.78	0.91	0.79	0.96
FMCG	0.62	0.60	0.52	1.00	0.65	0.45	0.35	0.36	0.39	0.42
Health Care	0.83	0.78	0.54	0.65	1.00	0.84	0.68	0.42	0.72	0.44
IT	0.86	0.69	0.57	0.45	0.84	1.00	0.69	0.43	0.68	0.44
Metal	0.79	0.73	0.78	0.35	0.68	0.69	1.00	0.77	0.84	0.80
Oil & Gas	0.48	0.62	0.91	0.36	0.42	0.43	0.77	1.00	0.73	0.95
Bankex	0.74	0.85	0.79	0.39	0.72	0.68	0.84	0.73	1.00	0.78
Power	0.56	0.69	0.96	0.42	0.44	0.44	0.80	0.95	0.78	1.00

Data Preparation

- **OHLC data**

Python Stocker module makes it relatively easy to get OHLC data for each ticker as a Pandas data frame. Used a 90 day window (a quarter) to construct each row as X containing all OHLC data for that period and put Adj. Close as Y.

- **Sentimental Data**

The scraped data was highly unstructured as it contains hashtags, symbols and numerical tables. Cleaned these files by removing unrelated data and then used them to parse HTML into clean text data which was then used to analyse the sentiments. For cleaning and parsing, regex, NumPy and BeautifulSoup were used.

The data was filtered to remove the duplicate headlines and was paired with the tickers.

Sentiment Analysis Using NLTK VADER

Used NLTK VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analyser. In this approach, each of the words is rated as to whether it is positive, negative or neutral and scores are calculated based on how positive, negative or neutral a sentiment is. VADER was updated with financial terms using existing Financial Sentiment Word Lists from Loughran-McDonald.

Obtained positive, negative and neutral sentiment scores for the entire sections of each of the headlines using VADER model.

Stock Price Prediction

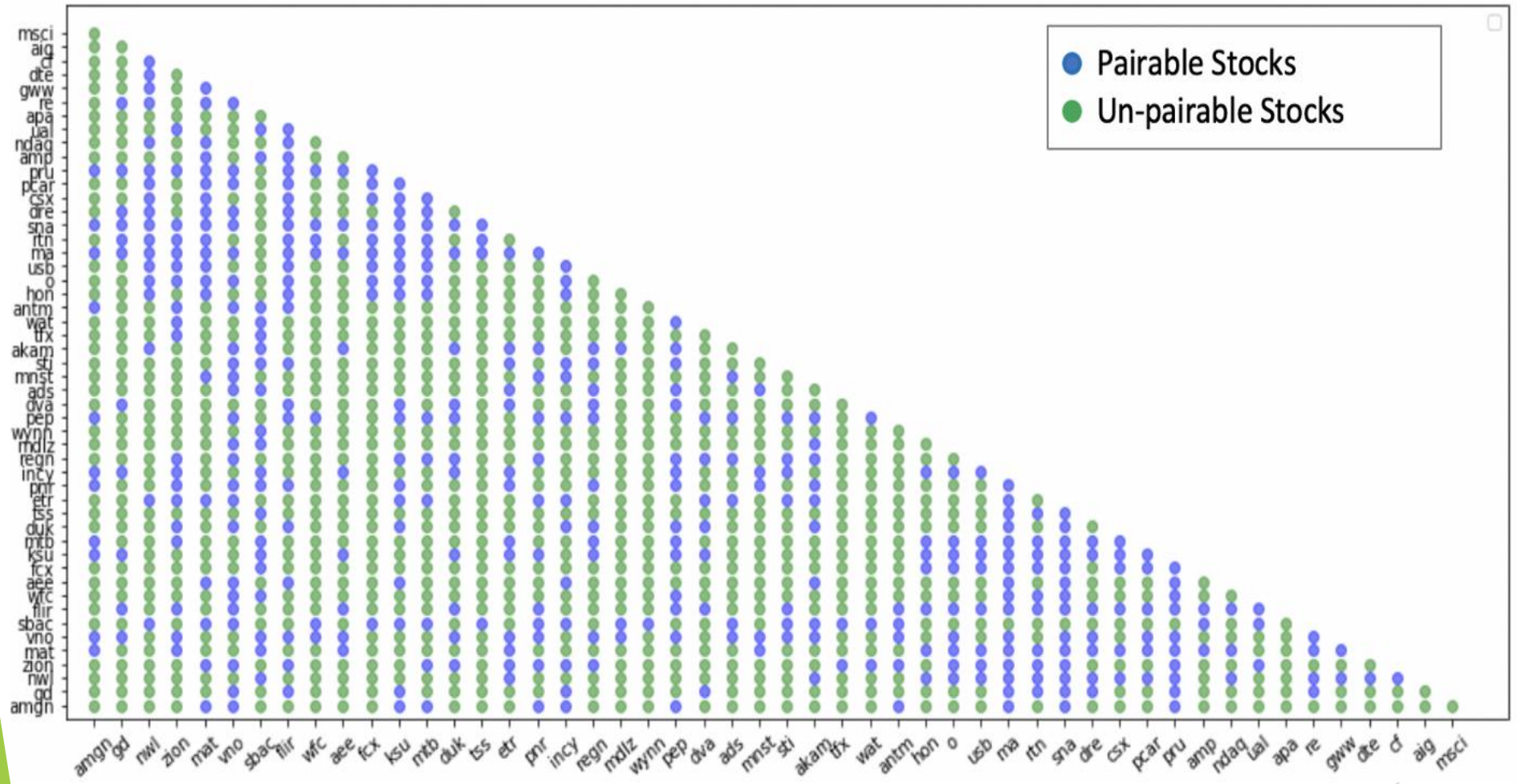
- Constructed pandas data frame for each stock using a 90 day window.
- Pre-processing and analysing was done on the data.
- Transformed the data into respective train, validation and test dataset for each of these stocks.
- Stock price prediction was done using LSTM model(**Long short-term memory**) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections)
- The predicted price and sentiment scores were then combined to produce the final predicted prices for the stocks.

Plots comparing actual Adj. Closing prices for GWW with predicted prices with and without negative sentiment scores

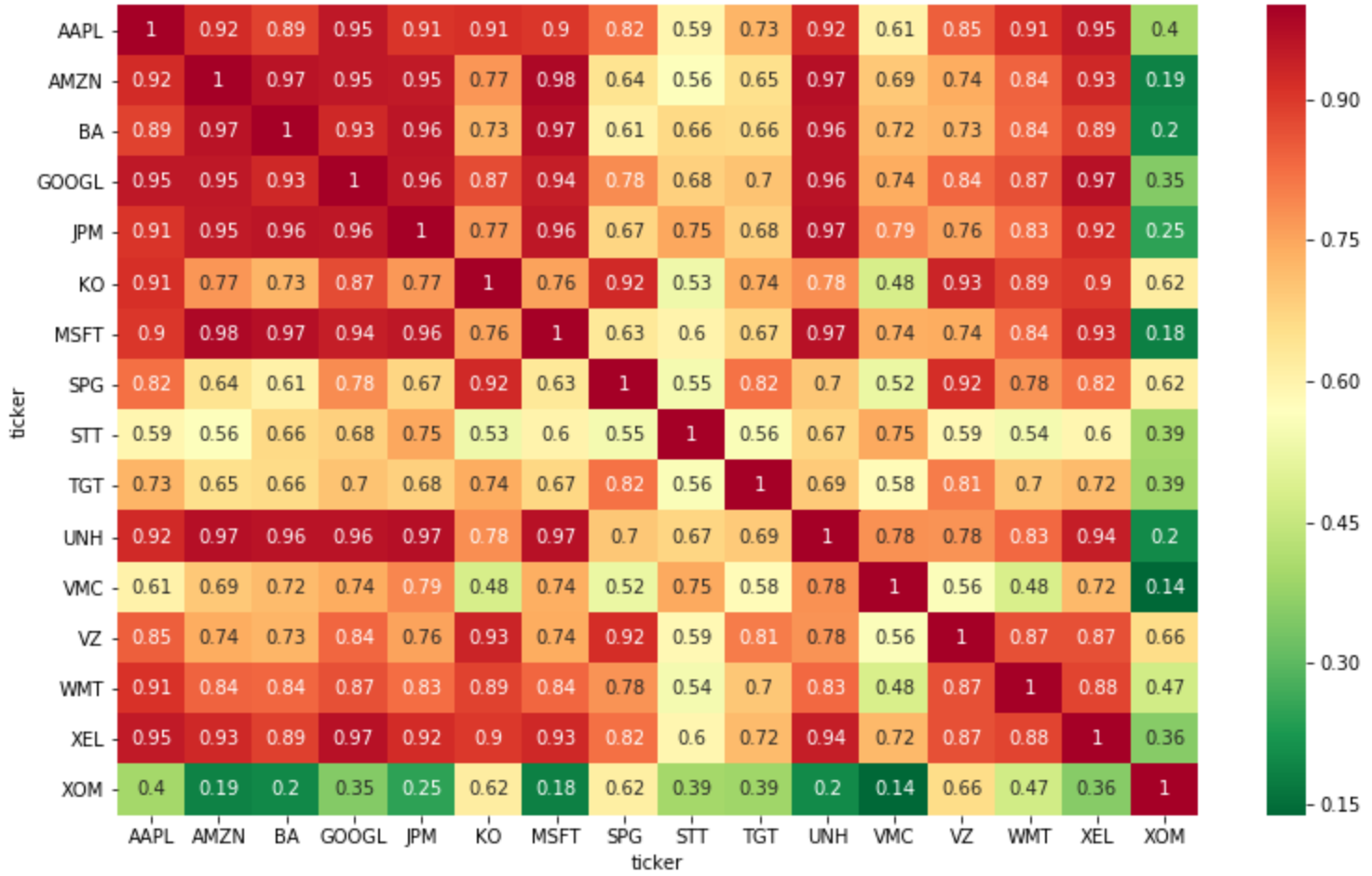
Constructing stable Portfolios

- The strategy followed in order to create stable portfolios is to keep uncorrelated stocks in same portfolio to avoid any huge loss in a portfolio, because if one stock goes down the others will balance the loss.
- Considered a correlation value less than 0.5, and a covariance value less than mean covariance. Based on analysed the “pairable” stocks and “unpairable” stocks using the plot shown on next page.
- The Sharpe ratio is calculated by subtracting the risk-free rate from the return of the portfolio and dividing that result by the standard deviation of the portfolio’s excess return.
- A higher value of Sharpe ratio means better risk-adjusted return. A sharpe ratio greater than 1 is considered good, greater than 2 is considered very good and greater than 3 is considered excellent.
- Based on the above three parameters, stable portfolios were predicted.

Plot showing pairable and unpairable stocks.



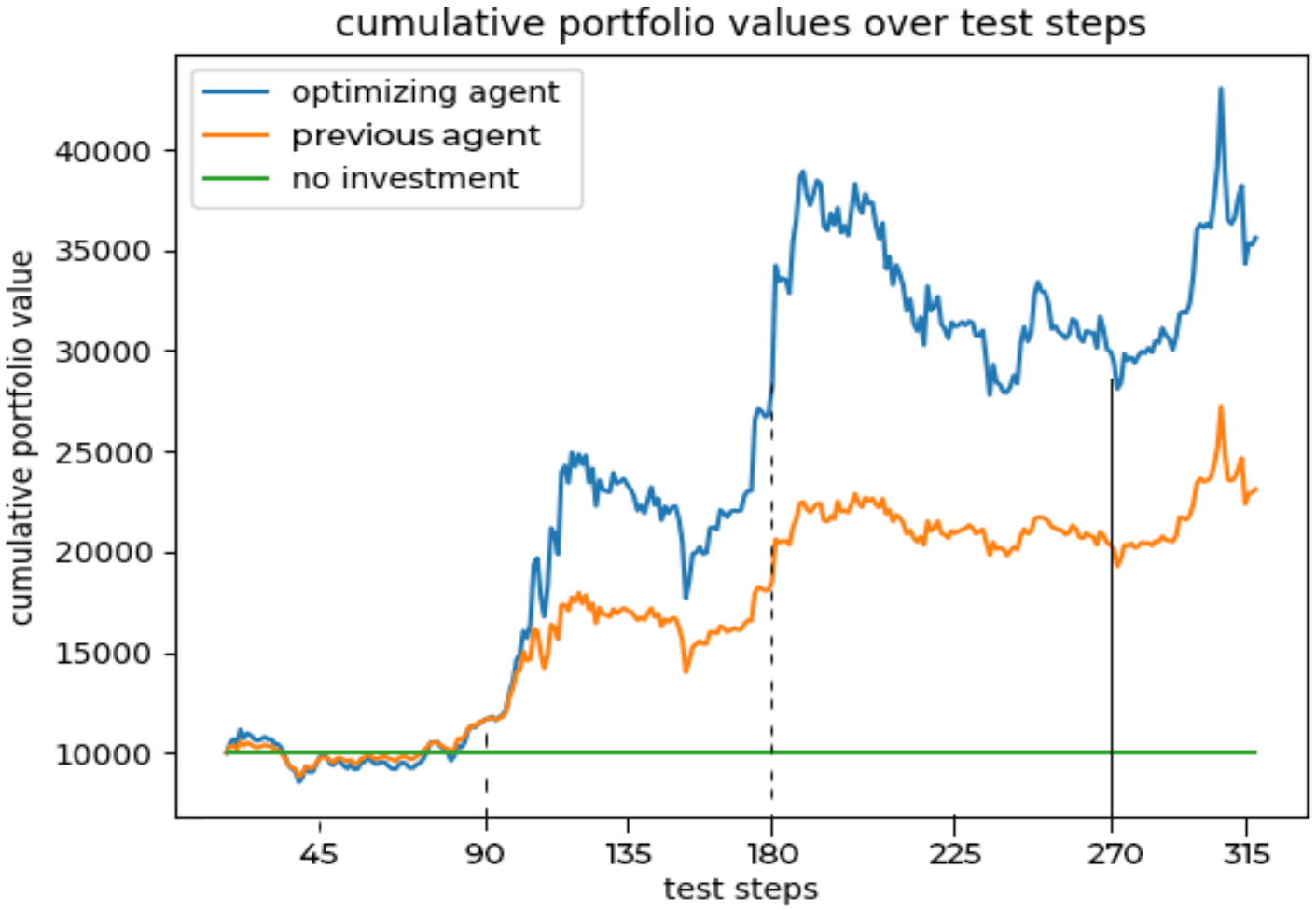
Heat map showing the correlations between various stocks



Portfolio Rebalancing

- Rebalancing was performed on one of the stable predicted portfolio.
- CNN model (Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.) was used with Sharpe ratio for predicting the weights of the assets in the portfolio.
- The model was trained to rebalance the portfolio every 90 days initially and evaluation of results were carried on keeping this assumption in mind.
- Rebalancing was done taking in consideration the transaction costs incurred while rebalancing.

Results Indicating Total Portfolio Value



Conclusion

- Learning about finance domain to understand what can potentially work was definitely a challenge. Moreover, working on how to manipulate time-series data, windowing methods, and training LSTM and CNN models using that proved interesting.
- Sentiments extracted from financial news are significant to predict future stock trends. Learnt to extract sentiments on very large text data (100,000 words in some case) using NTLK VADER.
- Learned how to construct portfolios by leveraging concepts such as correlation, covariance, Sharpe ratio and volatility. Relevant visualizations such as colormaps and correlation matrix were very useful in confirming the obtained results.
- Rebalancing of portfolio using CNN model for training gave satisfactory results in for a 90 day time window which is also visualized in the graphs plotted for it.

Assumptions/Limitations

- Currently, have trained separate models for each of the stocks considered. This was done because we obtained very poor results after training a single model.
- Due to constraints on hardware resources, failed to train a single model using all the 500 S&P stocks data and had to eventually limit it to 50 stocks.
- Twitter was used for sentiment analysis data due to constraints on the requests to be sent on severs of financial news sites.
- Transaction cost during rebalancing was considered to be 20%.
- Risk free rate was assumed to be 2% based on the source referred in appendix.

Future works

- For sentiment analysis, we can consider word-embedding models namely: Word2Vec, FastText and Universal Sentence Encoder.
- Extracted only positive, negative and neutral sentiments. There are a lot of other sentiments tied to the financial industry, like uncertainty, litigious, constraining, superfluous, which can be extracted and analysed further.
- There are a lot of other portfolio construction methods like ‘Global Minimum Variance Portfolio (GMV)’ and ‘Inverse Volatility Portfolio IVP’, etc.
- Also various fundamental factors such as Beta, Alpha, Volatility, etc that can be considered while prediction.

References

1. Lazy Prices http://laurenhcohen.com/wp-content/uploads/2017/09/lazyprices.pdf
2. S&P 500 Companies https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
3. Python Stocker https://github.com/WillKoehrsen/Data-Analysis/tree/master/stocker
4. Universal features of price formation in financial markets: perspectives from Deep Learning https://arxiv.org/pdf/1803.06917.pdf
5. https://medium.com/datadriveninvestor/why-financial-time-series-lstm-prediction-fails-4d1486d336e0
6. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks https://www.uts.edu.au/sites/default/files/ADG_Cons2015_Loughran%20McDonald%20JE%202011.pdf
7. Financial lexicon from Loughran-McDonald Sentiment Word Lists. https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists
8. The Most Rewarding Portfolio Construction Techniques https://seekingalpha.com/article/1710142-the-most-rewarding-portfolio-construction-techniques-an-unbiased-evaluation