Stock Price Prediction- Deep learning and NLP

Abstract:

Anyone that follows the financial markets knows that the stock market is more volatile than ever. The news sentiment and even tweets from the president can dramatically move the market up or down. Recent studies have shown the vast amount of public information such as news stories can have an observable effect on investor's opinion towards financial markets. In this assignment, I evaluate the impact of sentiment analysis and how it can predict the stock market price and trend. Focusing on LSTM model, I was able to achieve directional accuracy with MSE rate at 0.00948 for the testing data.

Introduction:

For my analysis, I used VADER, a sentiment lexicon designed for social media to assign sentiment scores (Hutto and Gilbert 2014). Then focused on LSTM model with linear activation for my output layer due to the data is time series based. I separated my analysis into three parts. My initial analysis is using headline news and leverages NLP to conduct binary classification in predicting the market is going up or down. Secondarily, I combined sentiment polarity scores with daily open and close price to predict Dow Jones market price. Finally, I explored how the model works with individual stocks for further validation and observation.

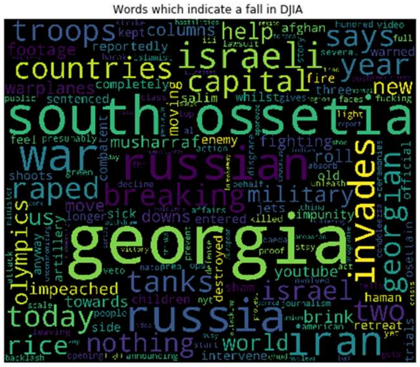
Lecture Review:

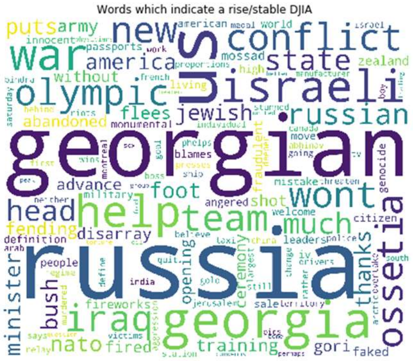
One of the main goals in the financial world is to predict stock prices to ensure profitability and control risk. Many research studies have been performed in the field of predicting stocks. Using Twitter messages to predict the stock price is a very popular approach (Bollena et al. 2011). Some of the research done in this field focuses on the idea of combining stock prices with news headlines, as outlined in the second part of this analysis (Kirange and Deshmukh, 2016.

Methods:

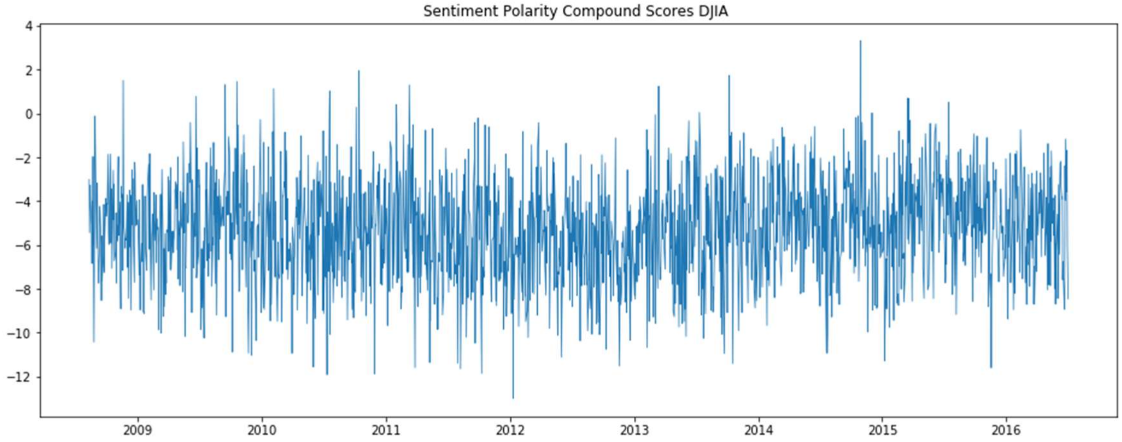
The data for this project was sourced from Kaggle, which spans from 2008-08-08 to 2016-07-01. In addition, news data consist of 25 daily headlines were extracted from reddit news and organized for analysis purpose. Within the data, each day also correspond to up "1", or down "0" for the day. Given the data scraping and cleaning can add significant amount of time to the analysis, I chose to go with this sample data.

I conducted preliminary EDA to understand the dataset presented from the headline news. Key words associated with market up or down were then displayed via word cloud as below.


Words which indicate a fall in DJIA

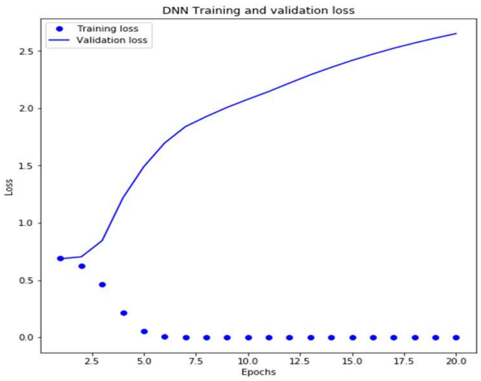

Words which indicate a rise/stable DJIA

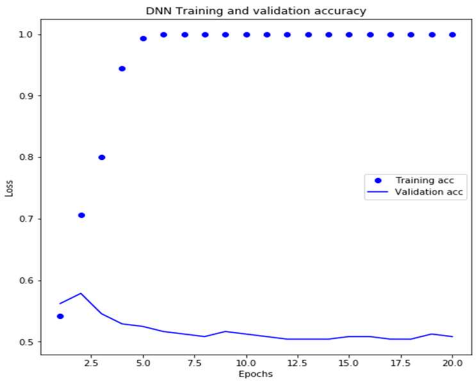
Visualization suggests same key words can be associated with up as well as down, for example, “Georgia”, “Russia” “Ossetia” and “War”. Correlation between market direction and key words may not be significant enough. Next, I passed all headlines through the Vader analyzer to capture sentiment polarity score. As the graph indicated below, news in this dataset is more on the negative side. This may not be surprising, since the news can be biased based on the media source.

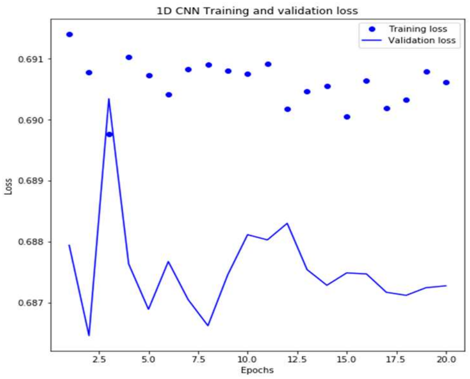


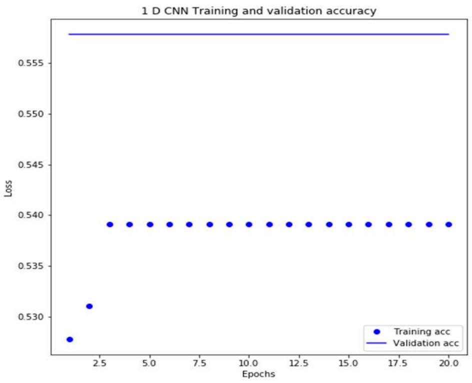
Part 1 Analysis-Binary Classification

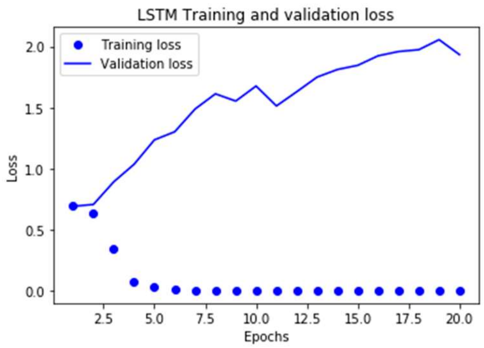
The initial focus is to predict up or down of the market based on just the headline news. Since this is a time series-based data, we can’t randomly break the dataset for train and test. For train = data[data['Date'] < '2015-01-01'], and the test = data[data['Date'] > '2014-12-31']. To set a simple benchmark, I first ran the logistic regression model, and accuracy output was 0.4781. Not quite impressive considering it’s less than 50/50 chance, so proceed to our neural network models, and evaluate the model accuracy against the benchmark. Dense Neural Network with 1 layer, took 43 seconds over epochs of 20, and produced 0.5343 accuracy. 1D CNN took 294 seconds and produced 0.5079 accuracy score. LSTM model took 795 seconds and produced 0.5 accuracy. All models were conducted with same activation, loss functions as well as optimizer to keep the baseline constant.

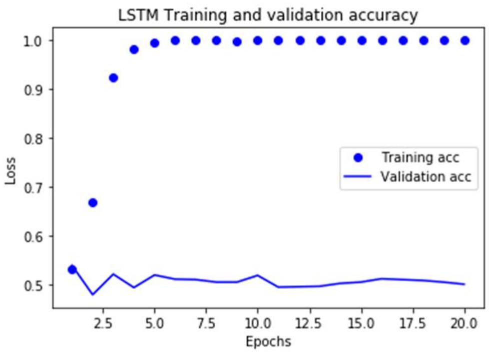








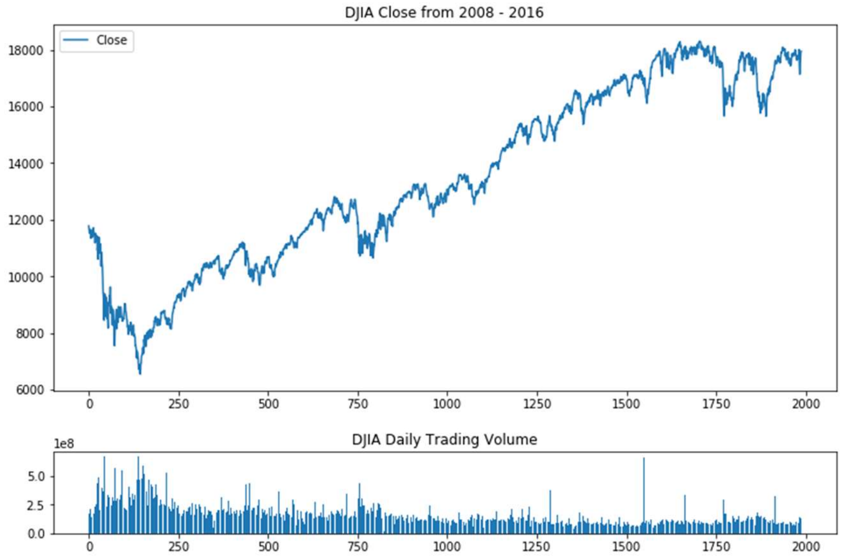




Results suggests an overfitting issue across all three models. Since using headline news to predict the market up or down is not much better than a coin toss, I decided to dive into the next part of the analysis. Combining sentiment polarity and stock price as the new dataset for further evaluation.

Part 2 Sentiment polarity time series analysis

The polarity score consists of “positive”, “negative”, “neutral” and “compound” was already created by using the Vader analyzer, I merged the scores with the daily DJIA stock price by “Date”.



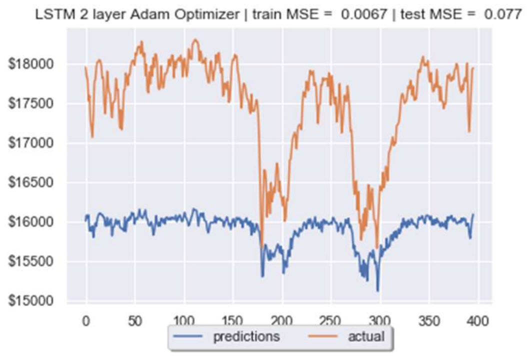
During the data preparation, the close stock price and compound score were shifted to next day. During the modeling build, I ran 5 models with adjustment to the optimizer, neural network layer, and number of nodes to minimize “mean squared error” loss function. Since the data is time series, output activation has been called using “linear” as the activation.

Model	ANN	Layer	Neurons	Optimizer	Train MSE	Test MSE
1	LSTM	2	25	Adam	0.0067	0.07729
2	LSTM	1	25	Adam	0.0011	0.00948
3	LSTM	1	25	RMSprop	0.0014	0.01361
4	LSTM	1	25	SGD	0.0013	0.01265
5	LSTM	1	50	SGD	0.0323	0.07444

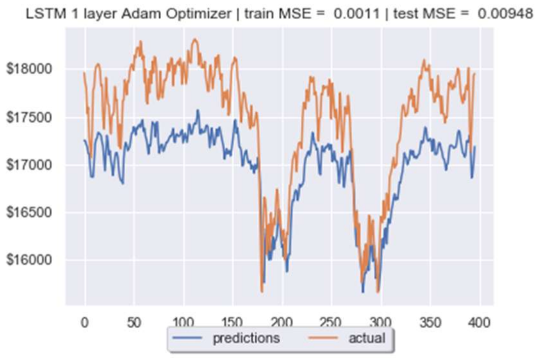
Comparing these 5 models, Model 2 produced the lowest mse rate in both training and testing. Model 2 is also 25% improvement comparing the model 3 based on mse rate. Observation based on plots suggests while all the models can generally predict the overall trend. Model 2 presented has the best fit between actual and predicted close price. 2 layered LSTM performed the worst, followed by additional nodes added to the model, from 25 to 50.

Model 5 exacerbated the volatility of the market, while Model 2, 3, and 4 did a good job in predicting the 2 main market dips in test data.

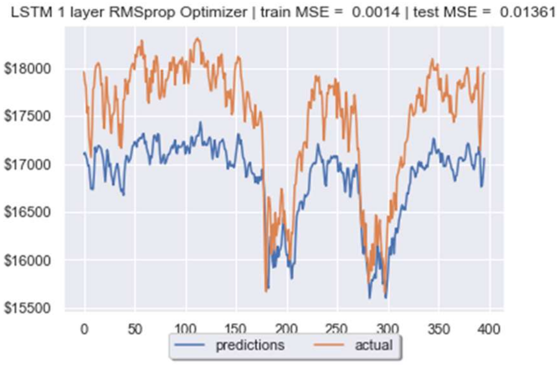
Model 1



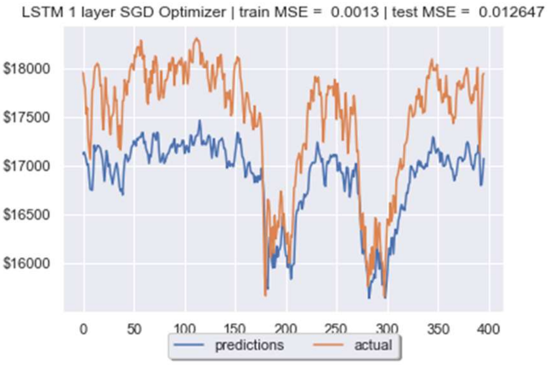
Model 2



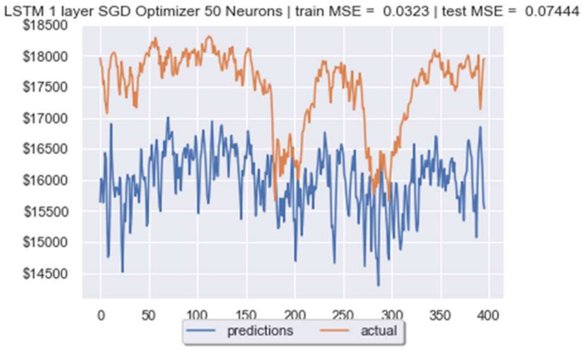
Model 3



Model 4

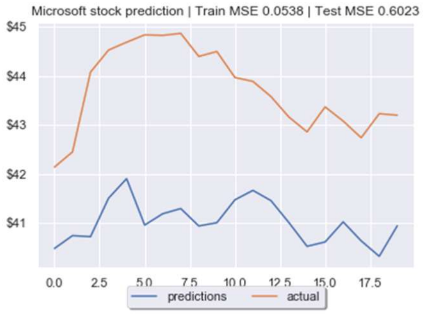


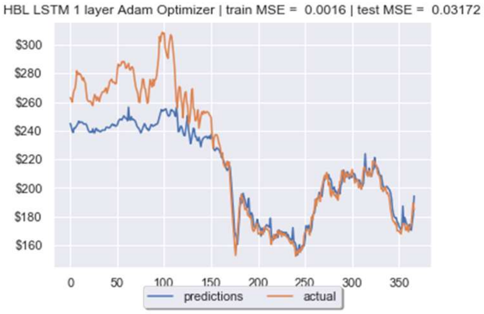
Model 5



Part 3 Single stock evaluation

I tried to see if this approach would work by using general market sentiment against an individual stock ticker. I downloaded 6 months, April to August of 2014, Microsoft stock price and paired it with the polarity score of the market in the existing data. Result was not impressive. I think the short time duration of 6 months versus 8 years of the previous data range could be one of the key contributors. Perhaps for a model to perform well at an individual level, more specific news collection is which then stacked with technology industry news to improve accuracy.





I came across a set of raw data from a stock symbol HBL, which is a small traded bank along with news attached spanning from January 2011 all the way to June 2018. Using the same approach, the prediction plot shows this approach can be successful when sourcing the relevant news along with sentiment polarity for the price prediction.

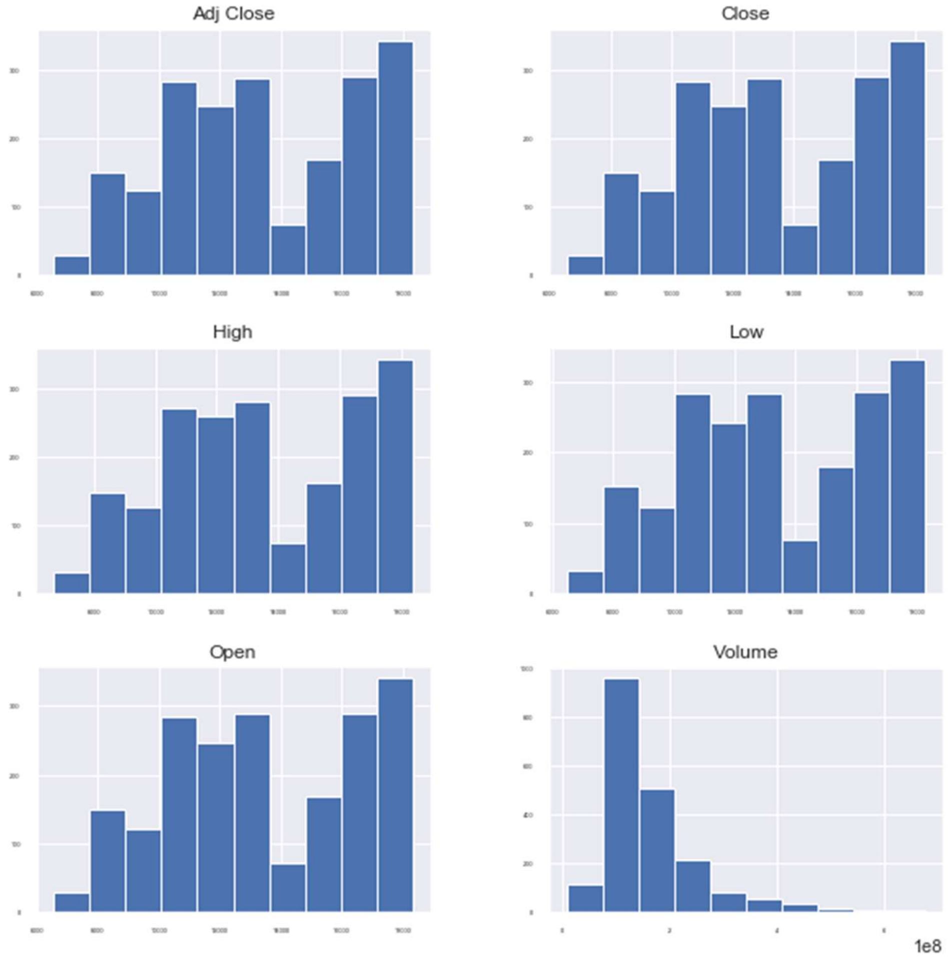
Conclusion:

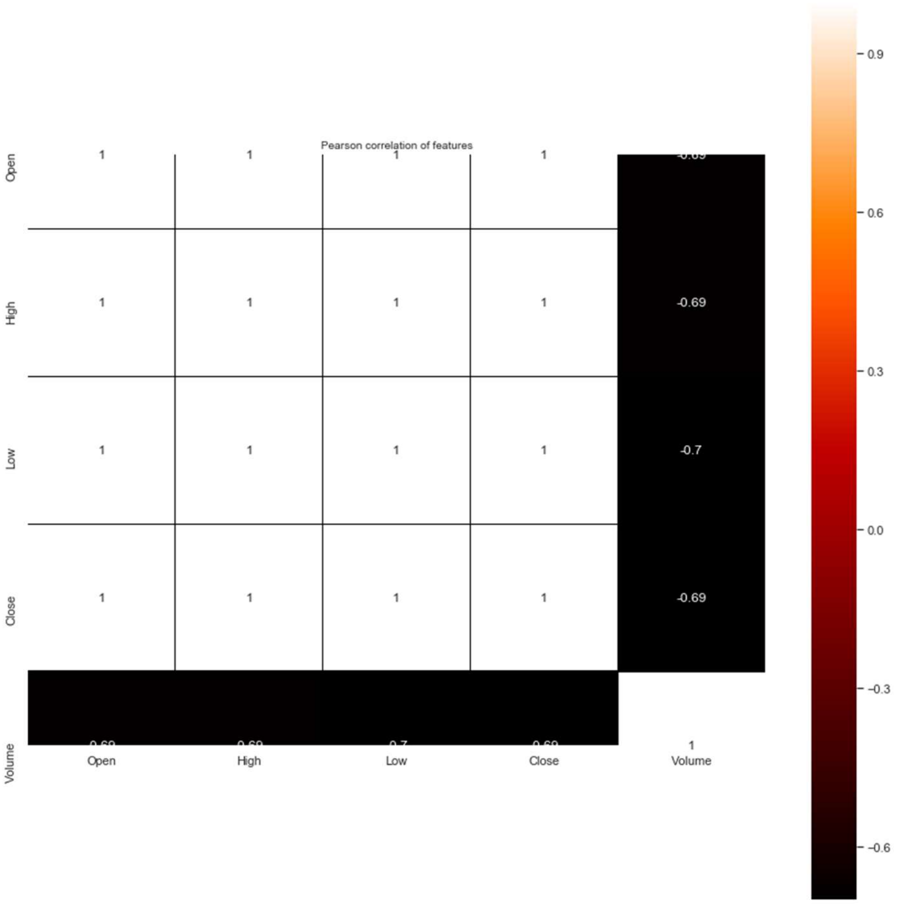
Leveraging sentiment analysis in addition to the traditional statistical modeling has becoming more popular in the past few years. An app called “Trump2cash” has been developed by Max Braun to monitor Trump’s tweet when he mentions a public traded stock and making recommendation on buying or selling via Twitter handle @Trump2Cash.

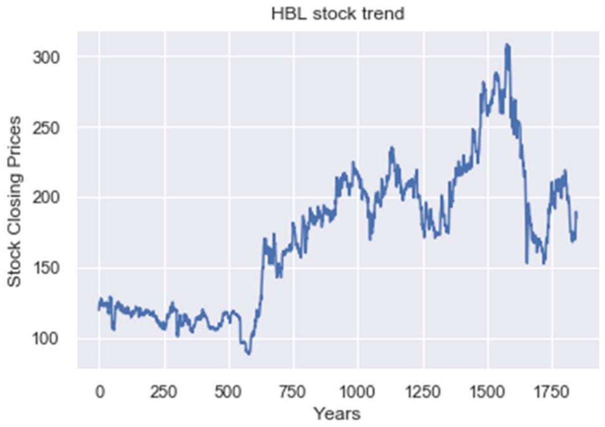
In this analysis, some of the models achieved directional accuracy by using a pre-defined data to predict Dow market trend. Our individual stock analysis shows some promising success in predicting daily stock price based on relevant news sentiment. However, further modeling can be conducted by combining conventional network, GRU, and simple RNN to the existing model to optimize and make it more robust. Daily news from various news sources can be bridged via API to automate the news scraping procedure. In this model, the sentiment was conducted over headline news, however, additional analysis can be done using news body itself instead of the headlines. Adding the price delta by day can also be explored further as an additional model input.

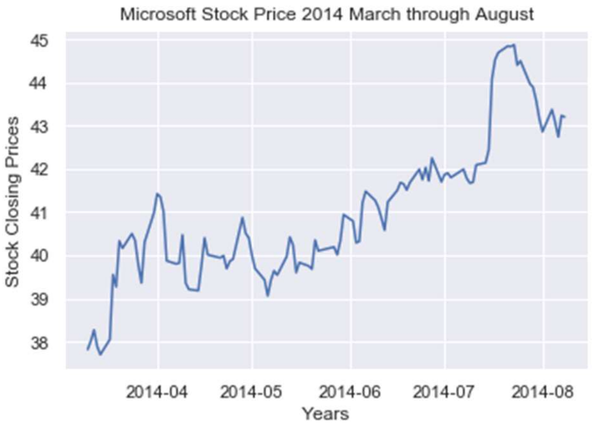
From our part 3 analysis, portfolio specific modeling can yield better results by targeting individual stocks price and news sentiment. None the less, model suggests, we may not yet able to predict the exact close price for the day, but we can understand the trend of the market and better prepare our personal portfolio when facing a downturn by closely monitoring the news sentiment.

Appendix- Visualization




Pearson correlation of features
Open
High
Low
Close
Volume
1
1
1
1
-0.69
1
1
1
1
-0.69
1
1
1
1
-0.7
1
1
1
1
-0.69
-0.69
-0.69
-0.7
-0.69
1
Open
High
Low
Close
Volume
0.9
0.6
0.3
0.0
−0.3
−0.6


HBL stock trend
Stock Closing Prices
300
250
200
150
100
0
250
500
750
1000
1250
1500
1750
Years


Microsoft Stock Price 2014 March through August
Stock Closing Prices
45
44
43
42
41
40
39
38
2014-04
2014-05
2014-06
2014-07
2014-08
Years

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 256)               102912
_________________________________________________________________
activation_1 (Activation)    (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               32896
_________________________________________________________________
activation_2 (Activation)    (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 2)                 258
_________________________________________________________________
activation_3 (Activation)    (None, 2)                 0
=================================================================
Total params: 136,066
Trainable params: 136,066
Non-trainable params: 0
```

Reference:

Sun, J. (August 2016). Daily News for Stock Market Prediction, Version 1. Retrieved from https://www.kaggle.com/aaron7sun/stocknews

Maas, A. L. (January 2011). Learning word vectors for sentiment analysis. *Annual Meeting of the Association for Computational Linguistics and ... Conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the Conference, 49,* 142-150.Retrieved from: http://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf

Hutto, C. J., and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In ICWSM.

Lee, H. & Surdeanu, M. & MacCartney, B. & Jurafsky, D.. (2014). On the importance of text analysis for stock price prediction. Proceedings of the 9th Edition of the Language Resources and Evaluation Conference (LREC). 1170-1175.

Heinz, S. (November 2017). A simple deep learning model for stock price prediction using TensorFlow. Medium. Retrieved from https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877

Bigg, J. (2017). Trump2Cash lets you invest automatically whenever the president mentions a publicly traded company. *TechCrunch.* Retrieved from https://techcrunch.com/2017/02/10/trump2cash-lets-you-invest-automatically-whenever-the-president-mentions-a-publicly-traded-company/

GitHub data. Retrieved from https://raw.githubusercontent.com/ZainUlMustafa/Stock-Prediction-using-News-Info-Sentiment/master/final_data_hbl.csv

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation* 9.8 (1997): 1735-1780.

Bollena, J. H. Maoa, & X. Zengb, (2010). Twitter mood predicts the stock market, *Journal of Computational Science*. 2.

D. Kirange and R. R. Deshmukh (2016). Sentiment analysis of news headlines for stock price prediction. Retrieved from: https://www.researchgate.net/publication/299536363_Sentiment_Analysis_of_News_Headlines_for_Stock_Price_Prediction