OPIM 5671 - Data Mining and Business Intelligence

Stock Prediction with News

Project White Paper

May 7, 2020

Ba, Yang

Contents

1. INTRODUCTION

The stock market is extremely volatile and difficult to predict. Changes of the stock price are affected by various elements. Because stock price plays an important role in the business decision making, research of forecasting stock price attracts people from multiple areas. News is a dominant way for people to obtain daily information. Based on Efficient-Market Hypothesis and Behavioral finance concepts, news would have an impact on the stock market. This project is working to use text mining and time series to detect the connection between the news and the movement of the stock market. Two tasks are included in this project: one is to find out the way of news affecting the stock price rise and fall by machine learning and deep learning; the other is to create a time series forecasting model to predict the stock price in the next 100 days.

2. LITERATURE REVIEW

The efficient-market hypothesis (EMH) posed by University of Chicago professor William Sharp is a popular investment theory in the stock market. It is a hypothesis in financial economics that states asset prices reflect all available information, therefore current prices are the best approximation of a company's intrinsic value. Marcelo's paper mentions that the EMH claims the price changes in the stock market should be considered all the related factors rather than considering price history only.

Behavioral Finance is a study of investor market behavior that derives from psychological principles of decision making, to explain why people buy or sell the stocks. It is related to behavioral cognitive psychology, which studies human decision making, and financial market economics. (Ekanshi Gupta etc., *Efficient Market Hypothesis V/S Behavioural Finance*) Marcelo's paper mentions this concept as well, the behavioral Finance has different perceptions, which is an alternative model that accepts investors are all irrational.

Both Behavioral Finance and Efficient Market Hypothesis are trying to find the potential solution for the market trending problems. This project is an attempt to use text mining techniques and to build models to explore latent impacts of news on the movement of the stock market. My approach is to combine the traditional classification and regression models with sentiment scores generated from sentiment analysis

using Long Term Memory Networks and other features derived from texts by Natural Language Processing (NLP), which will help in improving the stock price prediction accuracy.

3. DATA

Data description

Two datasets are used in this project: one contains top 25 historical news headlines from Reddit World News Channel each day, the other is the Dow Jones Industrial Average from Yahoo Finance. The durations are both from August 8th, 2008 to July 1st, 2016.

Explore

Exploratory Data Analysis (EDA) is the first step in the data analysis process. Statistical graphics can assist to capture a big picture of the characteristics of data.

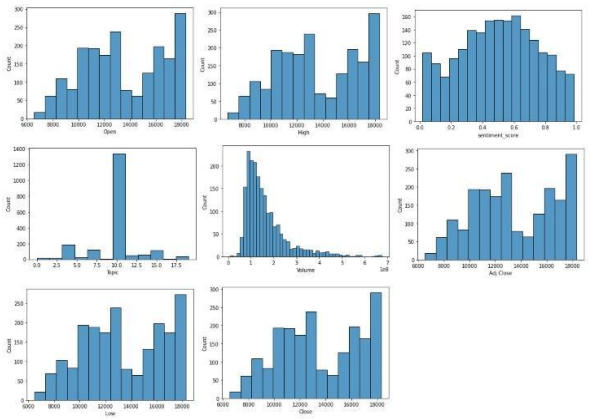


Figure 1. Data distribution

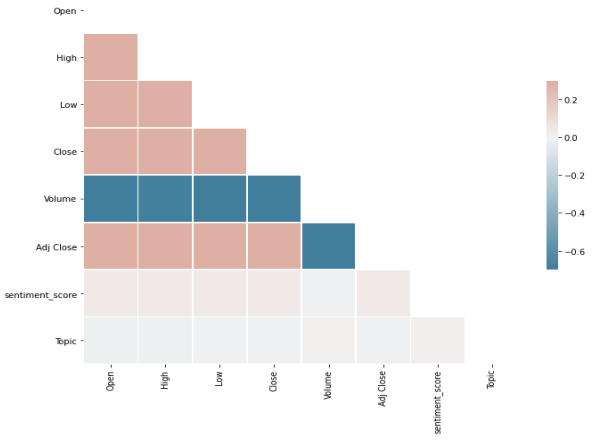


Figure 2. Correlation plot

Plots are created to understand the distribution and correlation of variables. Graphs show most of numeric features have skewness issues, especially for Volume and Topic. Normalization will be applied to them before modeling. High, low, close and Adj Close are highly related, so Adj Close is used as the label in the regression model and others are ignored.

Time Series Decomposition

R was used to decompose the time series to detect trend and seasonality. SAS miners are used to understand the seasonality and make data stationary by differencing. Then conduct different time series models to predict stock price.

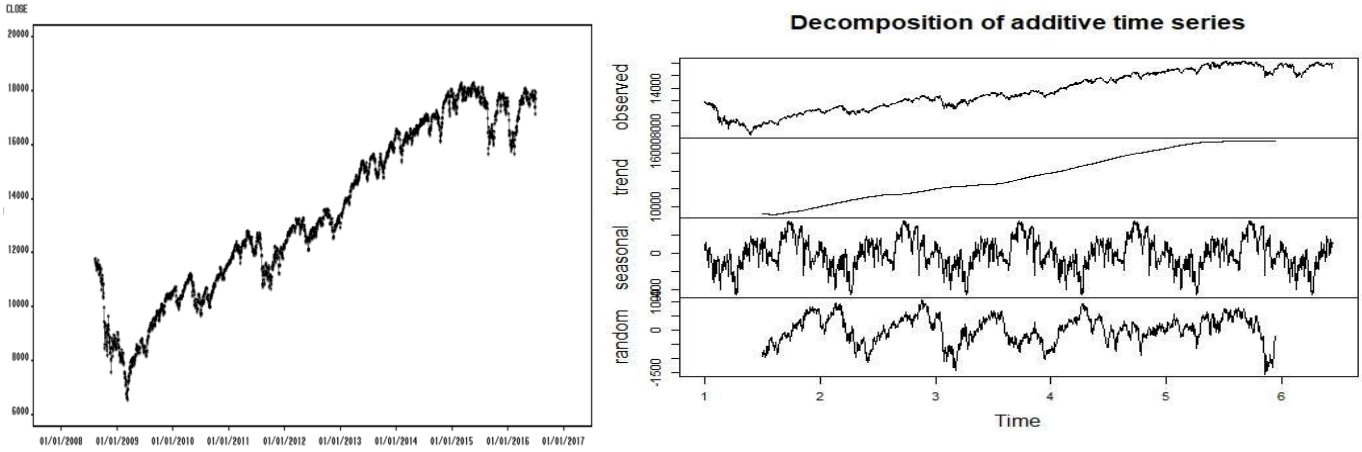


Figure 3. Stock price plot and time series decomposition

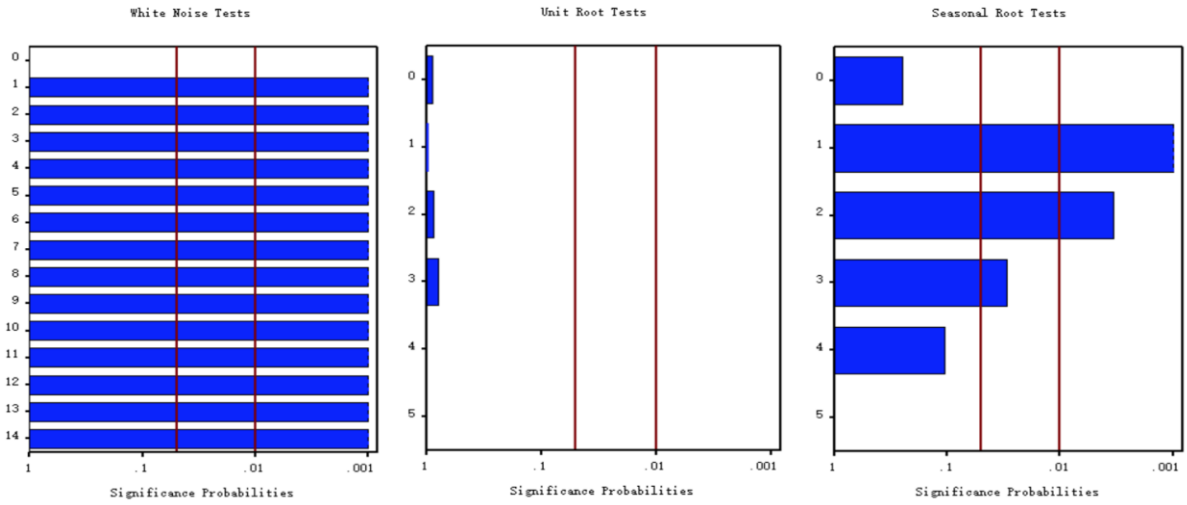


Figure 4. Test for stationarity and white noise

As graphics shown above, the time series data has a clear upward trend and a little bit of seasonality.

Data partitioning

Full datasets are split into training and validation parts and the ratio is 8:2. The training set is from 08/08/2008 to 12/31/2014 (1591 rows); the validation set is from 01/02/2015 to 07/01/2016(398 rows).

4. NATURAL LANGUAGE PROCESSING(NLP)

4.1 Sentiment Analysis with word embeddings and LSTM

Raw texts data cannot provide much valuable information. In order to detect the underlying facts hidden in these news headlines, I'm going to conduct a sentiment analysis.

Get data

I use the IMDB Movie Reviews dataset, a labelled data that consists of 50, 000 lines of sentences. The data is split into 2 parts: 80% of the overall data used to train the model, 20% of the overall data used to evaluate the accuracy of the model. The score ranges from 0 to 1, 0 representing negative and 1 representing positive.

Data preprocessing

Stop words, punctuation and other symbols affect the performance of a text classification model. The preprocessing process performs a cleaning step to remove these 'noisy' elements.

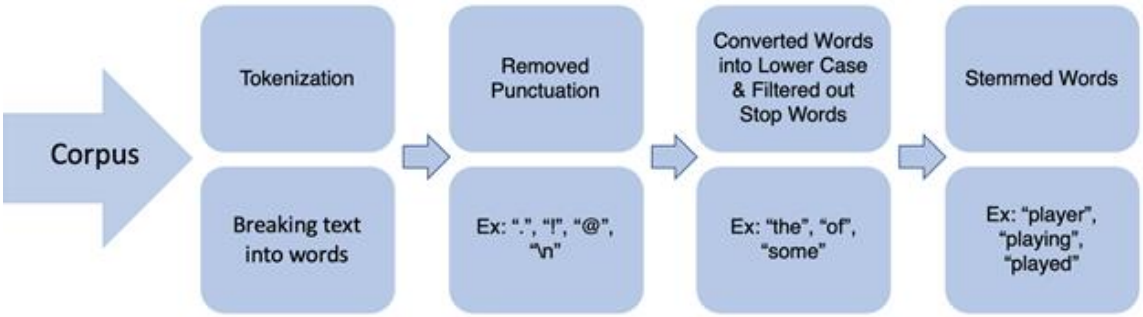


Figure 5. Text preprocessing workflow

Embeddings

Word embedding is an efficient technique of capturing underlying connections among words in a sentence, displaying more information about semantic and syntactic similarity. It converts words into corresponding dense vectors. I'm going to use a pre-trained word embedding model known as GloVe(glove.6B.100d.text). In this model, each word is represented as a 100-dimension embedding.

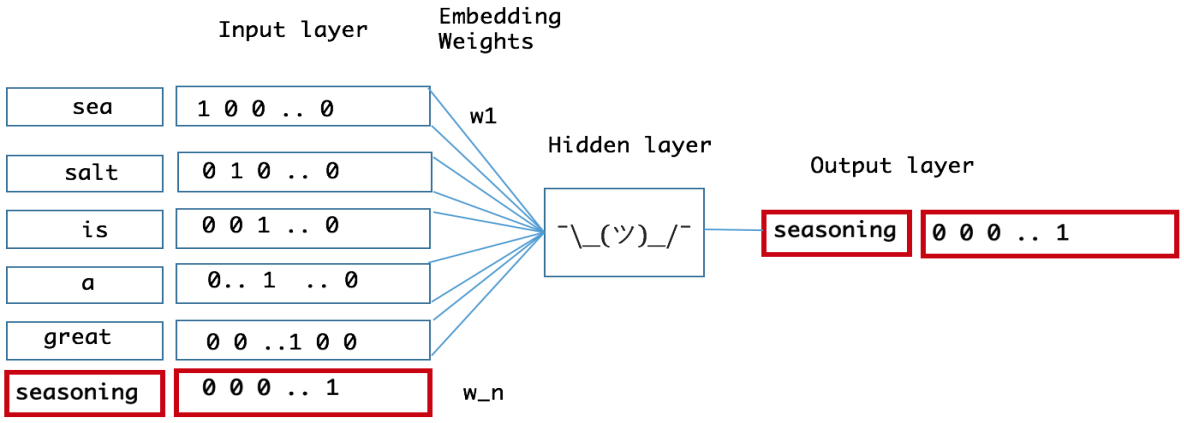


Figure 6. Embedding process

Model architecture

Convolutional neural networks are found to work well in text data, except widely used in images classification. 1D convolutional neural networks and 1 pooling layer are built in the model.

Since text is a sequence of words, I also use an LSTM (Long Short-Term Memory network), a Recurrent Neural Network, to train the model. As we can see in the picture below, the LSTM network can pass the preview information to the next LSTM cell. It enables to encapsulate the notion of forgetting part of its former stored memory, as well as to add the new information. This advantage makes it a better choice to solve sentiment classification problems.

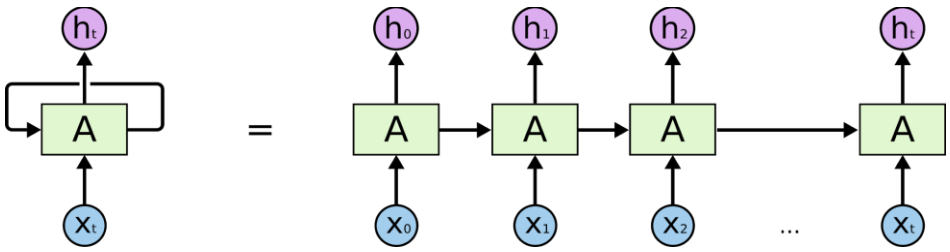


Figure 7. Long Short-term Memory Cell

Compared with CNN, LSTM has a much smaller difference between training and test sets and the loss values are also negligible.

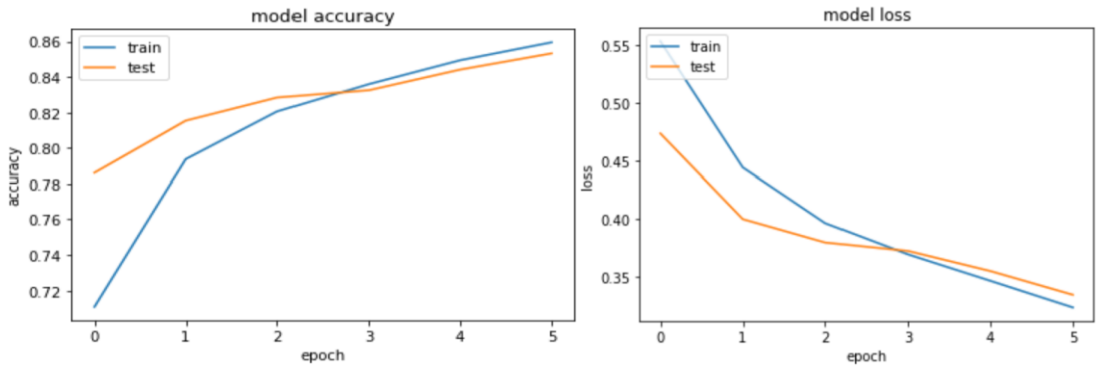


Figure 8. LSTM model performance

Predict the news data set

After the same cleaning process, the news text data is put into the selected LSTM model to generate sentiment scores. The LSTM model can accurately capture subtle information rather than the NLTK classifier, which generates scores from -1 to 1. For example, the sentence "Great! It is raining today!", which contains negative words, will be predicted a score 0.34 by LSTM model while a score 0.74 by the NLTK toolkit.

4.2 Feature Engineering

Latent Dirichlet Allocation (LDA)

In spite of the dazzling news headlines, Latent Dirichlet allocation (LDA) can extract topics from them. LDA is a generative probabilistic model of a corpus. LDA assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents share a common Dirichlet prior. Each latent topic in the LDA model is also represented as a probabilistic distribution over words and the word distributions of topics share a common Dirichlet prior as well.

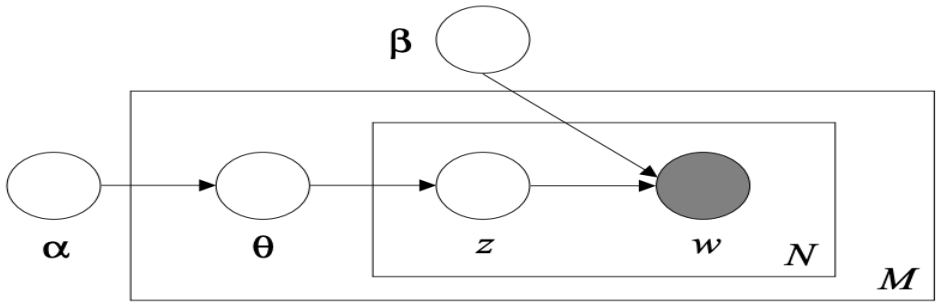


Figure 9. Plate notation of LDA model parameters

Using the Gensim toolkit, 20 topics and 200 iterations are set in the LDA model. These 20 topics as a feature are used in the machine learning algorithms.

SVD, Term-Document Matrix and Word Cloud

The Top 25 news headlines contain 250-350 words on average every single day, which means so many features in them. Singular Value Decomposition method can reduce the dimensionality of these texts. 150 columns of SVDs are produced to be the output of SAS Enterprise Miner as below.

	TextCluster_SVD1	TextCluster_SVD2	TextCluster_SVD3	TextCluster_SVD4	TextCluster_SVD5	TextCluster_SVD6	TextCluster_SVD7	TextCluster_SVD8	TextCluster_SVD9	TextCluster_SVD10	TextCluster_SVD11	TextCluster_SVD12
2	0.070340114	-0.053418193	0.007627758	-0.039590066	0.013758202	-0.006961034	0.003005235	-0.009582263	0.004257301	-0.015628509	0.058124621	-0.042869334
3	0.051980009	-0.058077784	-0.015399651	-0.007447694	0.046361831	0.002541248	-0.087866241	0.085166965	-0.003734233	-0.008320733	-0.028623084	0.040409213
4	0.189980843	-0.047806397	0.068441172	-0.014392762	-0.006441791	-0.066779379	-0.037231105	-0.009753008	0.01598861	-0.205987968	0.049159784	0.169125813
5	0.024443524	-0.021403577	-0.011992042	-0.000997722	-0.00653632	-0.001815523	0.003278639	-0.003551917	0.007168791	-0.028825008	0.007590198	-0.007292146
6	0.164037696	-0.043497936	0.001303352	-0.023022701	0.040366017	-0.009300889	-0.043737962	-0.097056443	0.051675435	0.006076514	0.020189106	0.042213933
7	0.022845085	-0.026412823	0.00237397	-0.011166965	0.022534413	0.003231199	-0.028910541	0.065960516	-0.001504602	-0.024255527	0.001175459	0.041740192
8	0.019050642	0.009963616	0.003062893	-0.004661739	0.005712028	-0.004547871	-0.002583416	-0.000678885	-0.002706236	-0.002488063	0.004489107	-0.001347041
9	0.038653896	-0.030654349	-0.018975181	-0.032178735	-0.013936403	-0.031999196	-0.011542517	-0.002357799	0.01899002	-0.030517009	0.034466831	0.016088629
10	0.093979263	-0.0584685	-0.079588067	0.034351136	-0.042836619	0.015973401	0.039963583	-0.029485775	0.113888741	-0.153767558	-0.129295436	-0.200174317
11	0.203856461	-0.059061531	0.060715736	-0.015158041	-0.001850374	-0.067786184	-0.03461337	-0.009673705	0.038880213	-0.205166494	0.060636768	0.168384779
12	0.063664911	-0.03024549	-0.028315712	0.006820442	-0.010995704	-0.027024326	0.02278198	-0.009156626	-0.007835013	-0.030042661	0.045423971	0.044509165
13	0.091443856	-0.06439135	0.030097187	-0.080111643	-0.00886799	-0.066892487	0.035260809	0.040433321	0.033938872	-0.05361177	0.075213073	0.041921938
14	0	0	0	0	0	0	0	0	0	0	0	0
15	0.083658382	-0.078153086	-0.004264374	-0.108983984	0.015181546	-0.105942493	0.049888482	-0.025588709	-0.028271171	-0.007732764	0.031790746	-0.01187221
16	0.072300735	-0.060578235	-0.01869404	-0.005804895	0.003909334	-0.029597402	0.009547726	-0.024326343	0.044202402	-0.085949886	0.192680405	0.059683953
17	0.112875878	-0.068114056	-0.028129729	0.000212553	-0.004061076	-0.039581599	0.013030523	0.020570232	0.024581967	-0.054890046	0.091547083	0.080291923
18	0.078680444	-0.06849239	-0.011581867	-0.004920417	-0.019587377	-0.056774455	0.0190757	0.031387567	0.008813421	-0.035873173	0.070470697	-0.085248112
19	0.02499197	-0.018234487	-0.008459492	-0.07696668	0.023426381	-0.142225403	0.118980028	0.052682578	0.005489451	0.070533531	0.017244322	0.018590447
20	0.156409081	-0.005735907	0.071947412	-0.036877187	-0.015909345	-0.04404395	-0.069916662	-0.056318259	-0.035817042	-0.018838818	0.005398463	-0.044112537
21	0.17281272	-0.005425332	-0.057471556	-0.006723045	0.014304292	0.06060406	0.013813422	-0.048937709	0.080279088	0.020066993	0.041928018	0.00036135
22	0.011647794	-0.009990526	-0.009900178	0.002837714	0.002838105	-0.0031502	0.002919955	0.0017087	0.003124701	-0.003148973	0.01572092	-0.006879299
23	0.033196722	-0.039544226	0.002963567	-0.016496975	0.036120036	-0.012975143	-0.011183228	-0.010316478	-0.01355966	0.004191077	0.011363453	-0.018875335
24	0.127315644	0.021287191	0.028968825	-0.04018426	-0.017442655	-0.054303784	-0.080457182	-0.029032631	-0.017524223	-0.060145342	0.016111107	-0.013128053
25	0.039950629	-0.043527116	0.025045523	-0.001028218	0.025849156	-0.012272677	0.0004864	-0.039377073	-0.003696623	-0.02641277	0.05429854	0.028055031

Figure 10. SVD matrix

In order to calculate the term frequency of the news texts, I built a Term-Document Matrix (TDM). The higher the tf value a term gets, the more important it is. A high value is reached when the term frequency in the news is high. The dimension of TDM is 55127 x 1989.

These words that are more than 500 times are picked, based on two categories-- stock rise or fall, and create two WordClouds. The larger the word is, the more frequent it is.

Figure 11. WorldCloud for stock rise

Figure12. WordCloud for stock fall

By comparing the compositions of word clouds, some insights can be obtained, even though there are somewhat overlaps for both categories.

- The news headlines concentrate on politics.
- The stock rise or fall are differential when news involving different countries. Specifically, China and Israel both affect stock rise and fall. But Iran will have a significance of stock fall, whereas Russia and Korea would play an important role in stock rise.
- News related to the bank links closely to the stock rise.

5. Modeling

5.1 Classification model

Model architecture with texts data

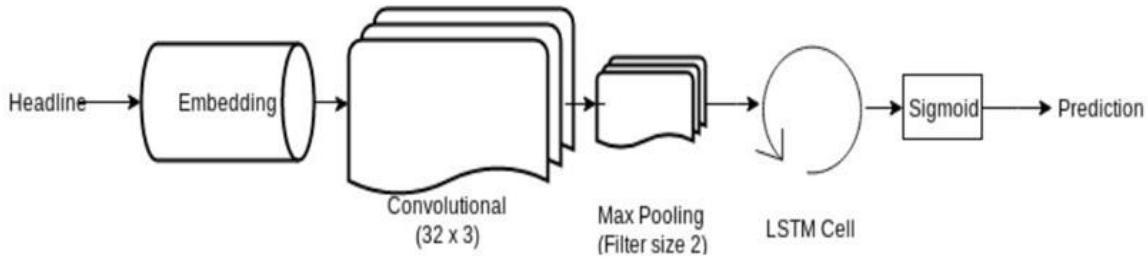


Figure 13. Classification model architecture with texts data

Similar to the sentiment analysis process mentioned above, the model is constructed following by the step: preprocessing and embedding. Rather than generating sentiment scores, Convolution layers, a set of filters whose weights are updated when the algorithm learns, are followed by the Embedded matrix. Max pooling is a part of Convolution where layers are introduced in successive layers to help reduce parameters to control overfitting. Then a LSTM cell is passed from the previous layer. Lastly, sigmoid activation function is used to predict probability ranging from 0 to 1.

Model architecture with extra features

Instead of converting texts into vectors, I derived features from text data, including sentiment scores, the number of LDA topics and SVDs. Then build classification models by applying traditional machine learning algorithms, Deep Learning and hybrid CNN-LSTM.

Machine learning

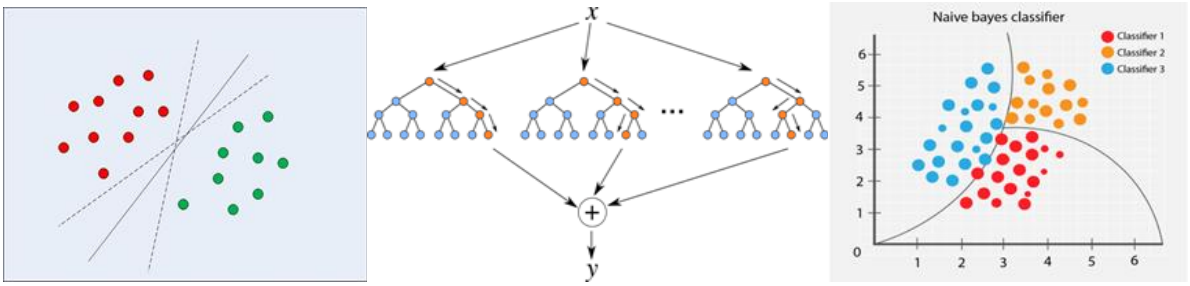


Figure 14. Machine learning algorithms

Except for well-known machine learning algorithms, an ensemble model that includes Logistic Regression, K-nearest neighbors, Decision Tree and SVM, is built.

Dip Deeper in Machine Learning Model

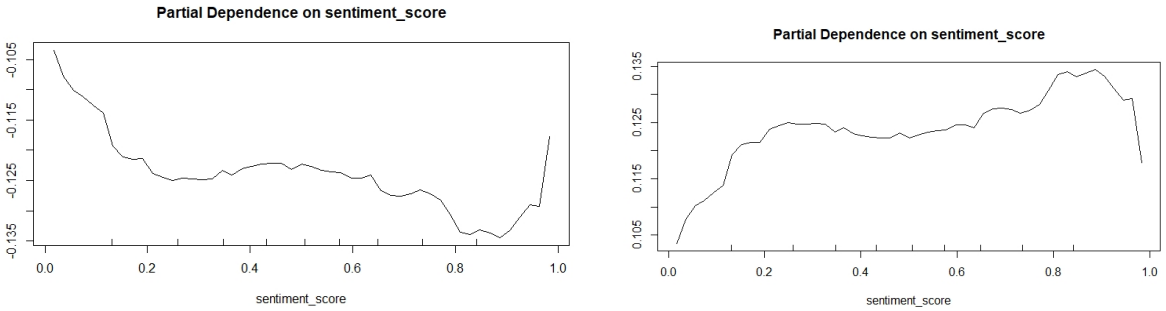


Figure 15. Partial dependence on sentiment scores (left: Label=0, right: Label=1)

The Partial Dependence Plot derived from the Random Forest model could tell more information about how sentiment score affects the label. The sentiment score has a positive effect on stock rise and the effect increases with the sentiment scores increasing. On the contrary, the sentiment score has a negative effect on stock rise and the effect decreases with the sentiment scores increasing.

Deep Learning

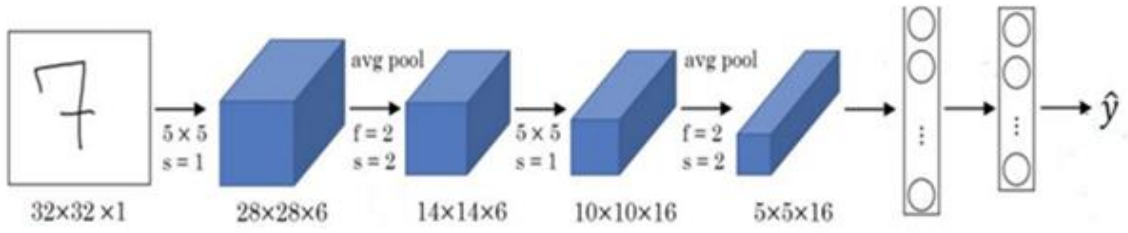


Figure 16. Convolutional Neural Network architecture

Features used in the Deep Learning Neural network are the same as machine learning models. This time, another Convolution layer is added by the first Convolution layer. Each complete Convolution layer part contains one Convolution layer, Max pooling layer and a Dropout layer with parameter 0.2.

Since text is actually a sequence of words, the LSTM model is an automatic choice. A hybrid CNN-LSTM model is applied: a LSTM cell connects with two 1D CNN layers. Each 1D CNN layer part has one Convolution layer, Max pooling layer and a Dropout layer with parameter 0.2. A LSTM layer with parameter 128 comes after two CNN layers. The model performance can be seen as below.

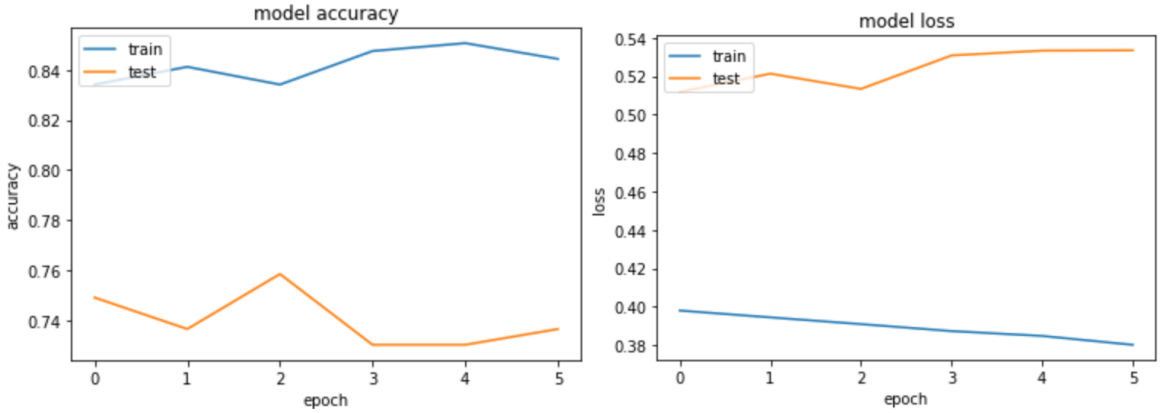


Figure 17. Hybrid model (CNN+LSTM) performance

5.2 Regression model

Autocorrelation Plot

The ACF plot and PACF plot that is not significant after order 2 prove a strong trend and a subtle seasonality again. After applying first differencing and seasonal differencing, ARIMA models, Holt winters and Exponential Smoothing are implemented. RMSE is selected as the evaluation metric. Sentiment scores and different regressors are added in the models to improve the model accuracy.

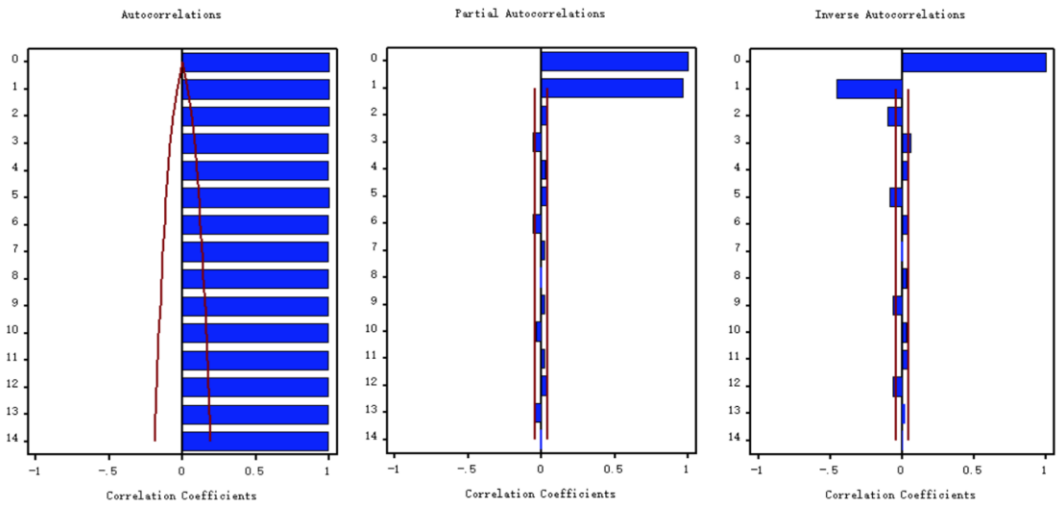


Figure 18. ACF, PACF & IACF plot

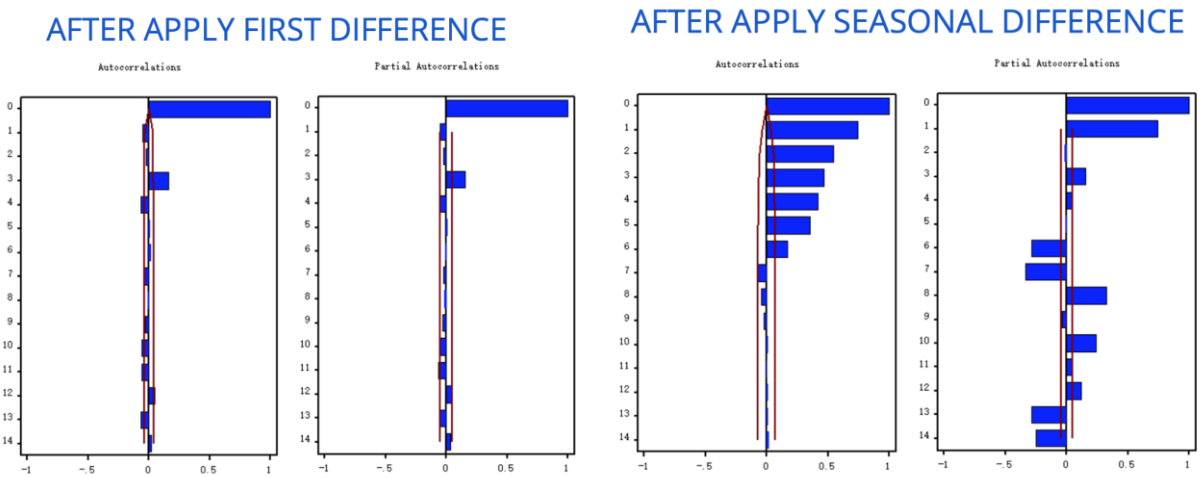


Figure 19. Comparison before & after first/seasonal differencing

From the results, sentiment scores are indeed useful to reduce RMSE, but LDA topic seems not to be a good regressor for ARMIA models. Comparing changes of models' performance, adding events is not helpful to improve models' accuracy. One speculation is that the stock market could be followed by black swan events, which is hard to predict and be a reference in the future.

Prophet Model

Facebook's Prophet model is employed using R. It uses a Bayesian based curve fitting method to forecast time series data. It's good for historical data with a strong seasonality. After checking the model

performance, the RMSE for the Prophet model is 282.4, which is not a good shape for this stock data. The plot of the model is shown below.

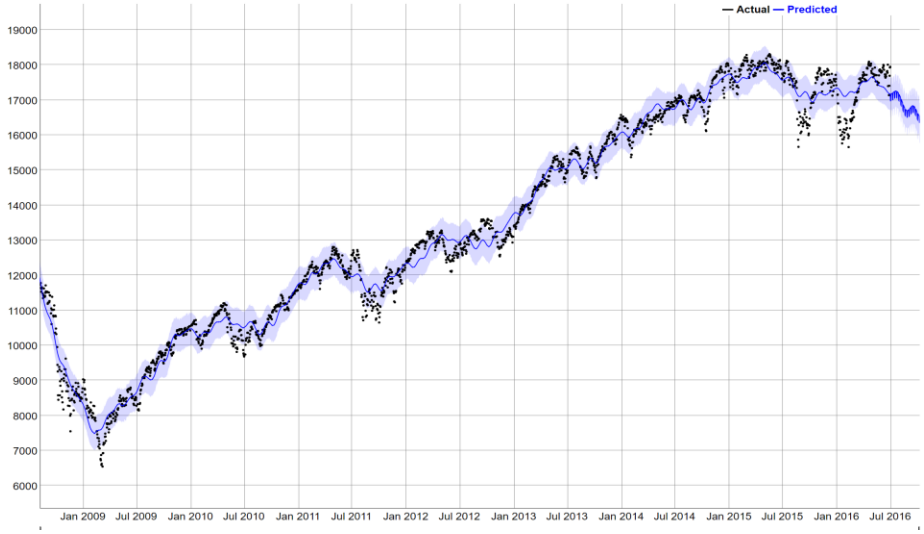


Figure 20. Actual(black) vs Predicted(blue) stock price on test data using Prophet

By comparison, the best model is the Seasonal ARIMA model with trend by adding sentiment scores with a lagging period of 5. These plots show autocorrelation, white noise test and stationarity test for the best model.

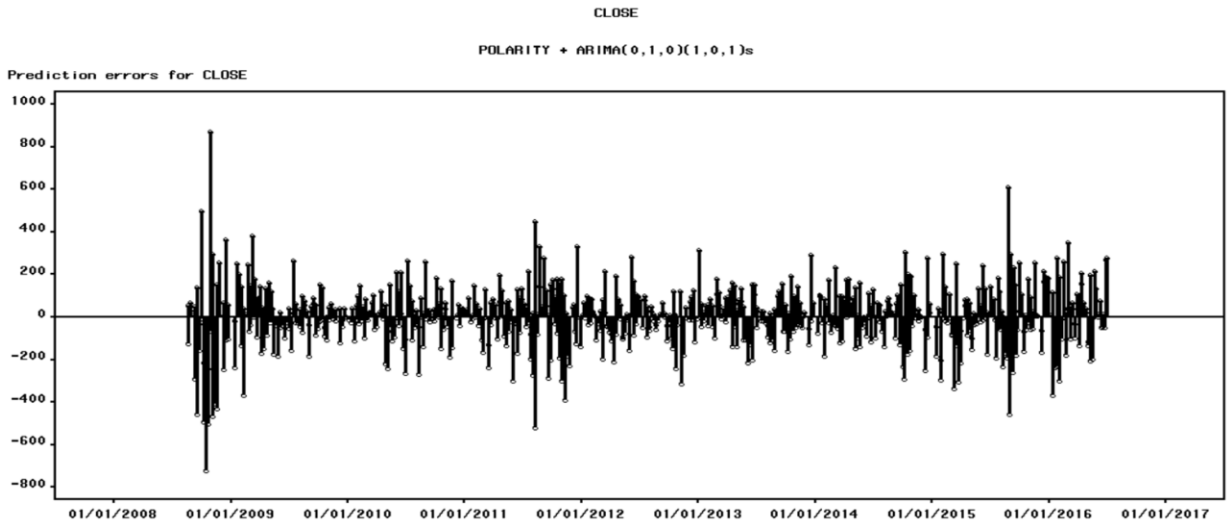


Figure 21. Prediction error plot for ARIMA model

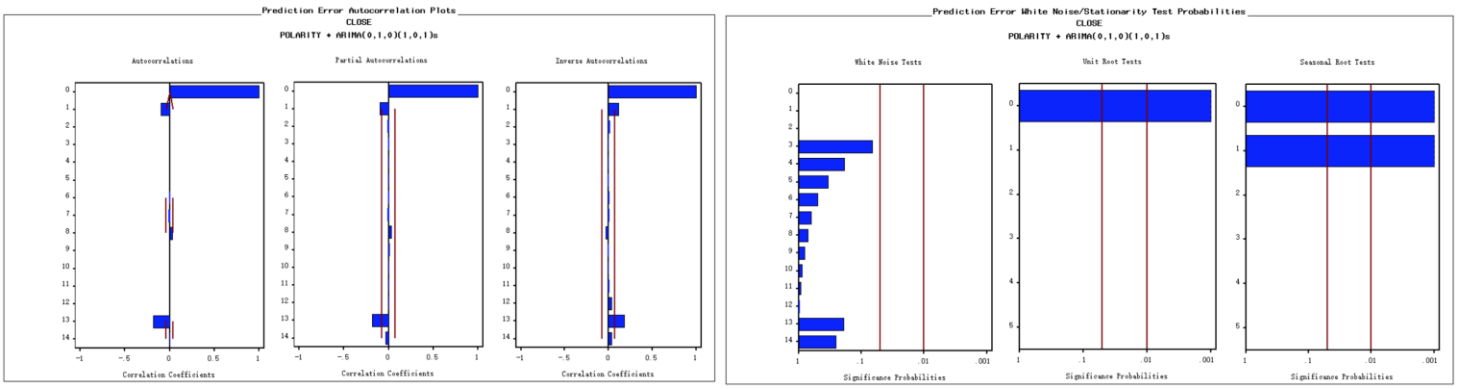
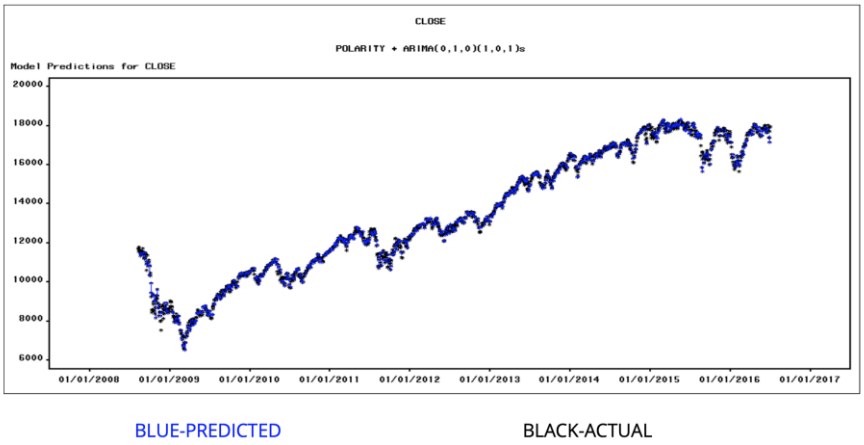


Figure 22. Plots for the best regression model



Predict stock price in the future 100 days based on the best model from ARIMA. As we can see, the fitting is pretty close.

Figure 23. Actual(black) vs Predicted(blue) stock price on test data

5.3 Model Comparison

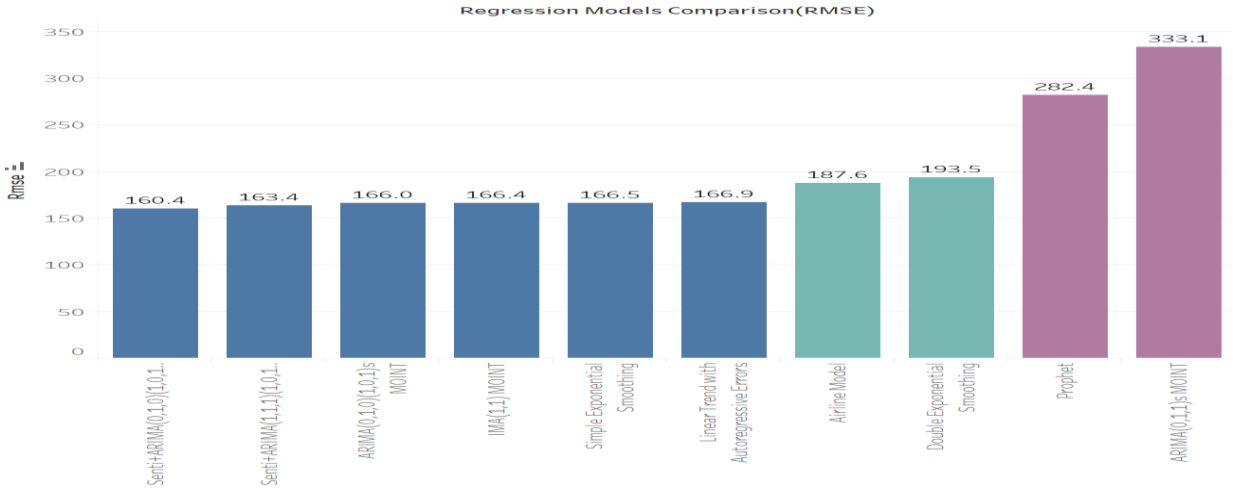


Figure 24. Regression Model comparison

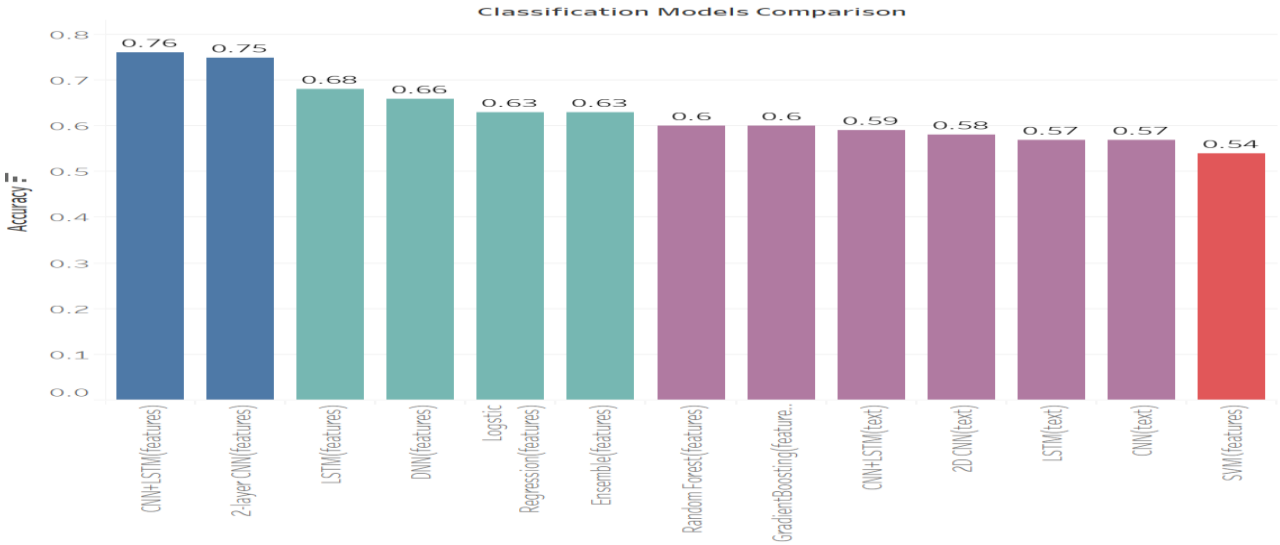


Figure 25. Classification Model Comparison

To sum up: the best classification model is a hybrid model with two CNN layers followed by one LSTM layer. The best regression model is ARIMA (0,1,0) (1,0,1) with sentiment scores.

6. INSIGHTS

Even though the unpredictable nature of stock price, the purpose for the project is to explore the underlying connections between the news and the movement of the stock market. Based on all the work I did, the conclusion is that the news and stock price are related to each other. I hope this project can provide some business insights for investors and stock companies as well, and take actions upon this project suggestions.

- International relationships are important parts of daily news headlines. The news about China and Israel could result in stock price rise and fall equally. Russia and Korea are probably related to stock rise. The appearance of Iran in the news headlines have a higher probability to cause stock fall. Being a wise investor, one should be aware of the daily news related to these countries.
- The market sentiment varies with the sentiment of news and the stock price will react correspondingly. Positive words in the news are more likely to contribute to stock price rise while negative words will lead to stock price fall.
- As one of the effective indicators of stock, the performance of bank business is closely related to the stock rise. Keep an eye on the news about banks.

- Unpredictable events will lead to the short-term stock fall but can't determine the trend in the long term. Some specific past events could not help to predict the future stock price trend.
- For world news, topic modeling can rapidly decompose the original news into countries and the specific events that happened.

7. CHALLENGES FACED

The first problem I encountered is the stock price data has missing data for holidays and weekends. Instead of imputing these missing values, I create blank observations to ensure predictive accuracy.

Besides, due to the long-time span of the dataset, unpredictable events are unavoidably abundant in the data set. Especially because of the financial Crisis of 2007-08, the stock market has a downward trend around these years, as opposed to the tendency in the next several years. If the dataset collected stock data from 2009 to 2016, the RMSE of time series models will be significantly decreased.

8. FUTURE IMPROVEMENT

1. Dow Jones Industrial Average consists of 30 well-known companies in the world, including various kinds of industries. In order to improve predictive accuracy, I would add news diversity, such as tech and business news, covering more industries and providing more insights rather than focusing on world news.
2. Continue to explore the possibility of improving the model accuracy and avoiding overfitting, try different hyperparameters combinations and increase the validity of classification models.
3. Apply the Long short-term memory (LSTM) in the regression model part to predict the stock price.

9. REFERENCES

1. Sun, J. (2016, August). *Daily News for Stock Market Prediction*, Version 1. Retrieved from https://www.kaggle.com/aaron7sun/stocknews
2. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* Journal of machine Learning research, 2003/1/3: p. 993-1022.
3. Marcelo Beckmann, "*Stock Price Change Prediction Using News Text Mining"*, 2017/01/24
4. Sharpe, William F. "*The Arithmetic of Active Management*"
5. Ekanshi Gupta; Preetibedi; Poonamlakra, *Efficient Market Hypothesis V/S Behavioural Finance*
6. Kavita Ganesan / Hands-On NLP, *Text Mining,Tutorial: Extracting Keywords with TF-IDF and Python's Scikit-Learn*
7. Anwar Ur Rehman1 & Ahmad Kamran Malik1 & Basit Raza1 & Waqar Al, *A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis*, 2019/5/17
8. Usman Malik, *Python for NLP: Movie Sentiment Analysis using Deep Learning in Keras*
9. GloVe: Global Vectors for Word Representation