Stock Market Prediction using Daily News Articles

Group Members:

Name	Roll number	Email-id
Yashwant Singh Patel	1721CS05	yashwant.pcs17@iitp.ac.in
Supriyo Mandal	1721CS03	supriyo.pcs17@iitp.ac.in

1. Abstract

In the finance field, stock trends are exceptionally crucial and volatile in nature. In recent years, it has drawn the attention of researchers to capture its volatility and predicting its trends. So that it can help the investors and market analysts to analyze the behaviour of market and plan their investment strategies accordingly. There are many factors by which the stock trends are affected, one of which is daily news articles. Now a day, these daily news articles serve the purpose of circulating organization, social or budget related information to the public and reflect their trading strategies on the stock market. Therefore, It is become necessary to deeply analyze the information to support the investors to make smart trading decision before making real investments. The aim of this project is to analyze these daily news feeds by using deep learning approaches and investigate the correlation between the large-scale daily news feeds and the stock market values over time.

Keywords: Stock trend, DJIA data set, Feature processing, News classification and Deep learning etc.

2. Introduction

Daily news articles can play a vital role for investors while judging the stock prices. These daily news articles circulate the organization, social or budget related information to the public and reflect their trading strategies on the stock market. There are numerous time series based stock prediction techniques used by stock investors. Most of these techniques use historical data to predict the market. This work uses the fundamental technique to determine the future trend of a stock by analyzing daily news articles about a company and tries to classify news as good (positive) or bad (negative). If the sentiment is found positive, there is more probability that the stock price will grow further and if it is negative, then stock price may go down. In this work, basic feature selection, vectorization and deep learning techniques are used to analyze and predict the news sentiment. These news analyses are used to correlate the future stock trends. We have taken past eight years dataset Dow Jones Industrial Average (DJIA) as stock price and corresponding news articles. The data set is divided into training as well as test data set. Then the basic feature processing and vectorization technique is applied to convert the news into a word vector form. This frequency word vectors are used in deep learning to improve the accuracy of model.

2.1. Literature survey

Stock trend prediction is a vital and active research area and requires accurate predictions. Therefore, in recent years, noteworthy efforts are put to develop prediction models for overall stock market. Few of the researchers haves shown a strong relationship between the daily news articles and stock prices belonging to a particular company. In this section previous research works are elaborated to understand their techniques and feature processing methods. Kalyani joshi et al. proposed a classification based model to predict the news polarity and relate the impact daily news articles on stock prices. Author has used the three years dataset of Apply company [1]. Stefan et al. used a simple moving average model to analyze and classify financial news' sentiment and used it to predict the corresponding stock trend [2]. Yu et al [3] present a text mining based approach for identification of news articles sentiments and demonstrate its impact on energy demand. News sentiments are quantified and represented as a time series and evaluated based on fluctuations between energy demand and prices. Existing literature methods on text mining are mostly based on simple textual representations such as bag-of-words for distinct single words. Dictionary based approach to represent the text intro word combinations or retrieved from the message corpus based on actual occurrences [4]. Other approaches like market feedback or TF-IDF [5] i.e. frequency dependent statistics of the message corpus and Noun Phrases are used to capture the semantics of text etc. Different works based on data sets, feature processing and selection methods are represented in Table I.

Table I: Related work based on data set, feature processing and machine learning techniques

Year	Author(s)	Data set	Feature Processing Methods	Machine Learning Techniques	Accuracy
2016	Kalyani Joshi et al. 2016 [1]	Apple Company	Dictionary based approach	Random Forest, SVM, Naïve Bayes	NA
2014	Stefan Lauren et al. 2014 [16]	JKSE Historical prices	Dictionary based approach	ANN	NA
2013	Michael et al. 2013[17]	German Adhoc messages	Word combinations	SVM	65.1 %
2012	Schumaker et al. 2012 [6]	US financial news	Noun phrases	SVR	59.0 %
2011	Groth et al. 2011 [14]	German adhoc announcements	Bag-of-words	SVM	NA
2010	Li 2010 [10]	Worldwide general news	Bag-of-words	K-nn, ANNs, naïve Bayes	NA
2009	Schumaker et al. 2009 [7]	US financial news	Noun phrases	SVM	58.2 %
2009	Butler et al. 2009 [15]	US annual reports	N-Gram	Proprietary distance measure	NA

2008	Tetlock et al. 2008 [13]	US financial news	Bag-of-words	Ratio of Negative words	NA
2007	Das & Chen 2007 [12]	US message postings	Bag-of-words	Combination of different classifiers	NA
2004	Mittermayr 2004 [8]	German adhoc announcements	Bag-of-words	SVM	56.5 %
2004	Antweiler et al. 2004 [11]	US corporate filings	Bag-of-words	Combination: Bayes, SVM	NA
1998	Wüthrich et al. 1998 [9]	US financial news	Bag-of-words	SVM	NA

3. Resources:

Data Set

http://finance.yahoo.com (for stock market) and
https://www.kaggle.com/aaron7sun/stocknews (for financial news)

Tools

-PyCharm 2017.1.1
Source: https://www.jetbrains.com/pycharm/download/#section=windows
-Keras– A theano based deep learning library
Source: https://keras.io/

4. Work done:

4.1 Model Design: Following system model is used for this project:

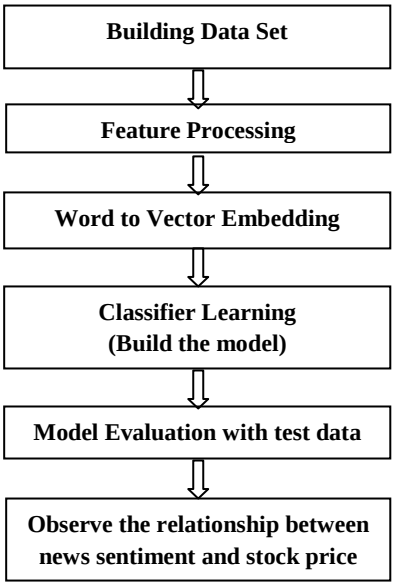

Figure 1: Model Design

4.2 Model Description: The steps of model are described as follows:

(i) Building Data set: We have collected Dow Jones Industrial Average (DJIA) data set of past eight years from 08/08/2008 to 01/07/2016 (8 years) as shown in figure 2. DJIA news data set contains three attributes date, label and news of company. DJIA stock data set contains attributes like date, Open, High, Low, Close, Adjusted Close, and Volume etc. For evaluation we have used only adjusted close price of everyday stock price. This data set is divided into 70% training and 30% test data set.

Data Set Source:

http://finance.yahoo.com (For stock market) and
https://www.kaggle.com/aaron7sun/stocknews (For financial news)

DJIA Stock Data Set

Date	Open	High	Low	Close	Volume	Adj Close
01-07-2016	17924.24	18002.38	17916.91	17949.37	82160000	17949.36914
30-06-2016	17712.76	17930.61	17711.8	17929.99	133030000	17929.99023
29-06-2016	17456.02	17704.51	17456.02	17694.68	106380000	17694.67969
28-06-2016	17190.51	17409.72	17190.51	17409.72	112190000	17409.7207
27-06-2016	17355.21	17355.21	17063.08	17140.24	138740000	17140.24023
24-06-2016	17946.63	17946.63	17356.34	17400.75	239000000	17400.75
23-06-2016	17844.11	18011.07	17844.11	18011.07	98070000	18011.07031
22-06-2016	17832.67	17920.16	17770.36	17780.83	89440000	17780.83008
21-06-2016	17827.33	17877.84	17799.8	17829.73	85130000	17829.73047
20-06-2016	17736.87	17946.36	17736.87	17804.87	99380000	17804.86914
17-06-2016	17733.44	17733.44	17602.78	17675.16	248680000	17675.16016
16-06-2016	17602.23	17754.91	17471.29	17733.1	91950000	17733.09961
15-06-2016	17703.65	17762.96	17629.01	17640.17	94130000	17640.16992
14-06-2016	17710.77	17733.92	17595.79	17674.82	93740000	17674.82031
13-06-2016	17830.5	17893.28	17731.35	17732.48	101690000	17732.48047
10-06-2016	17938.82	17938.82	17812.34	17865.34	90540000	17865.33984
09-06-2016	17969.98	18005.22	17915.88	17985.19	69690000	17985.18945
08-06-2016	17931.91	18016	17931.91	18005.05	71260000	18005.05078
07-06-2016	17936.22	18003.23	17936.22	17938.28	78750000	17938.2793
06-06-2016	17825.69	17949.68	17822.81	17920.33	71870000	17920.33008

DJIA News Data Set

Date	Label	News
09-11-2012	1	Cipla, the Indian drug company that cut prices of cancer drugs earlier this year, slashes prices of three m
12-11-2012	0	Arm raised in a Nazi-style salute, the leader of Greece's fastest-rising political party surveyed hundreds
13-11-2012	0	With more than 50,000 supporters, The uprising of women in the Arab world" facebook group is being bl
14-11-2012	0	UN says access to contraception a human right
16-11-2012	1	Mexico lawmaker introduces bill to legalize marijuana. A leftist Mexican lawmaker on Thursday present
19-11-2012	1	Turkish Prime Minister: "I say that Israel is a terrorist state, and its acts are terrorist acts"
20-11-2012	0	Israel air force drops leaflets across Gaza City warning residents to evacuate homes "immediately"
21-11-2012	1	Hamas executes six suspected collaborators with Israel
23-11-2012	1	Egypt protesters set fire to Muslim Brotherhood offices
26-11-2012	0	Having survived two assassination attempts, Mexican mayor beaten to death
27-11-2012	0	China's party paper falls for Onion joke about Kim Jong Un
28-11-2012	1	Girl, 8, to get vaccination shots after court overrules mum (xpost from /r/NewsOfTheWeird)
29-11-2012	1	Canada creates $5,000 cap on liability for file sharing lawsuits
30-11-2012	1	Less than 24 hours after General Assembly recognizes Palestine as non-member state, Israel responds b
03-12-2012	0	Internet Hangs in Balance as World Governments Meet in Secret
04-12-2012	0	Medicinal cannabis to be legalised in Ireland next year
05-12-2012	1	North Korean prisoner born in labor camp escaped after 23 brutal years
06-12-2012	1	Police and child advocates broke padlocks and busted down doors in a surprise raid of a sweatshop in In
07-12-2012	1	U.N. summit votes to support Internet eavesdropping. Uses: censorship, identifying BitTorrent and MP3
10-12-2012	1	Gunmen kill senior womens activist in Afghanistan
11-12-2012	1	Calls for a ban on helium balloons as world shortage worsens.

Figure 2: DJIA Data set

(ii) Features Processing: In this phase data cleaning is done such as removal of HTML tags(i.e. < >), common stop words (i.e. a, an, the), dealing with punctuation numbers (i.e. numeric values, spaces, comma, semicolon) and then the remaining words in the data sets are tokenized as shown below:

Sample Output:

'precision', 'guided', 'weapons', 'or', 'smart', 'bombs', 'sale', 'comes', 'as', 'human', 'rights', 'watch', 'charges', 'that', 'saudi', 'airstrikes', 'in', 'yemen', 'have', 'indiscriminately', 'killed', getting', 'in', 'way', 'deal', 'and', 'making', 'implausible', 'objections', 'say', 'delegates', 'and', 'campaigners', 'us', 'state', 'department', 'has', 'approved'

(iii) Word to Vector Embedding: In this phase we have used Bag of words and TF-IDF vectorizer. Following steps are used to convert the document into word vector form:

1. Count number of words in the dataset:

Sample Output:

[('us', 126), ('world', 102), ('government', 95), ('says', 92), ('people', 86), ('new', 83), ('police', 72), ('years', 71), ('one', 60), ('law', 56), ('war', 56), ('first', 54), ('drug', 52), ('internet', 50), ('million', 50), ('found', 46), ('said', 46), ('state', 46), ('uk', 43), ('court', 43)]

2. Convert it into word frequency

Sample Output:

```
supriyo@supriyo-Lenovo-G500:~/stockmarket/DL$ python 18march.py
Using Theano backend.
  (0, 116)  0.268855370445
  (0, 5)    0.407824086005
  (0, 432)  0.509617080092
  (0, 827)  0.434656518777
  (0, 4)    0.155669440835
  (0, 265)  0.392501077307
  (0, 968)  0.366702026887
  (1, 4)    0.0887270331534
  (1, 86)   0.273292272122
  (1, 471)  0.751982120815
  (1, 45)   0.372439483869
  (1, 12)   0.09614968043
  (1, 981)  0.257090496472
  (1, 304)  0.264513143749
  (1, 525)  0.260658935731
  (2, 4)    0.173262752488
  (2, 45)   0.363642780655
  (2, 12)   0.187757413835
  (2, 702)  0.516531137426
  (2, 964)  0.346756496897
  (2, 377)  0.619832745267
  (2, 15)   0.176787963538
  (3, 116)  0.224264768027
  (3, 4)    0.259702240502
  (3, 946)  0.366839422462
  :   :
  (1069, 346)     0.32587704589
  (1069, 858)     0.32587704589
  (1069, 179)     0.308660459623
  (1070, 265)     0.208606461045
  (1070, 12)      0.358626307281
  (1070, 15)      0.0844185234026
  (1070, 531)     0.380802042539
  (1070, 523)     0.219434670955
  (1070, 43)      0.122160633437
  (1070, 560)     0.228444813105
```

Figure 3: Word Frequency

3. Use normalization to convert into a word vector form

4. Selection of top 1000 most frequency words.

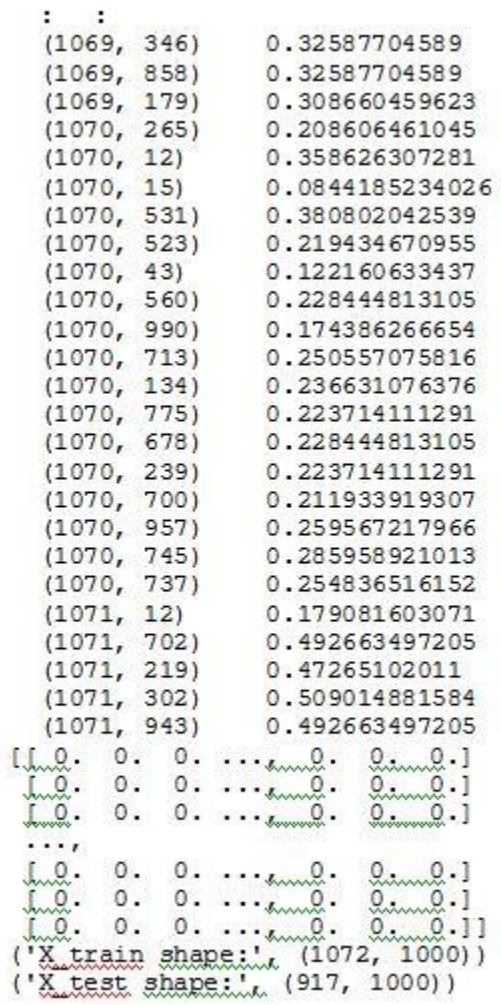

Figure 4: Word to vector embedding

(iv) Classifier Learning (Build the model): In this phase, two deep learning models are created: (i) Multi-layer Perceptron Model (MLP) and (ii) LSTM (Long Short Term Memory Model) and trained by giving the input of word vector.

(v) Model evaluation with the test data: Data set is divided into two categories train and test set. Accuracy of model is evaluated with the test data set. The model is optimized with number of layers, activation functions, epochs and batch size etc. as shown in table II. The performance comparison of both deep learning models is represented in figure 5 and figure 6.

Table II: Configuration of Deep Learning Models

Depth of Network	Epoch Size	Batch Size	Activation Functions	Loss Function	Optimizer
3,5,10	10,50,100	100	Relu, Softmax	Cross entropy	RMSprop

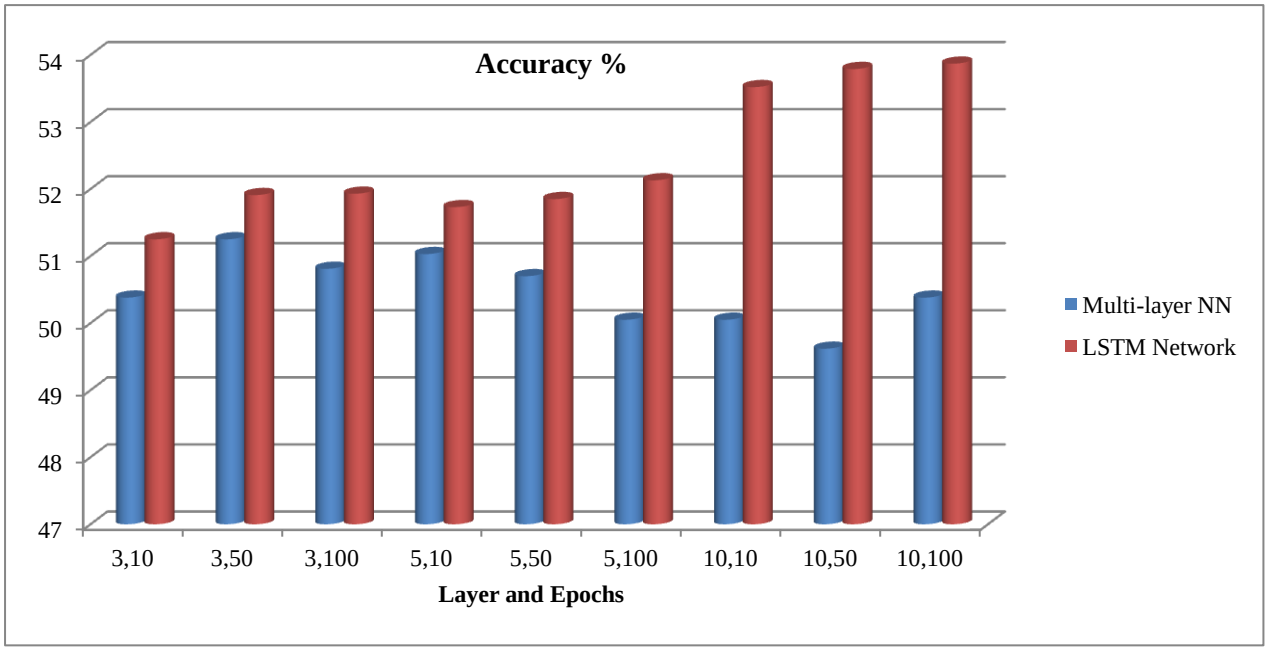

Figure 5: Comparison based on Layers and Epochs

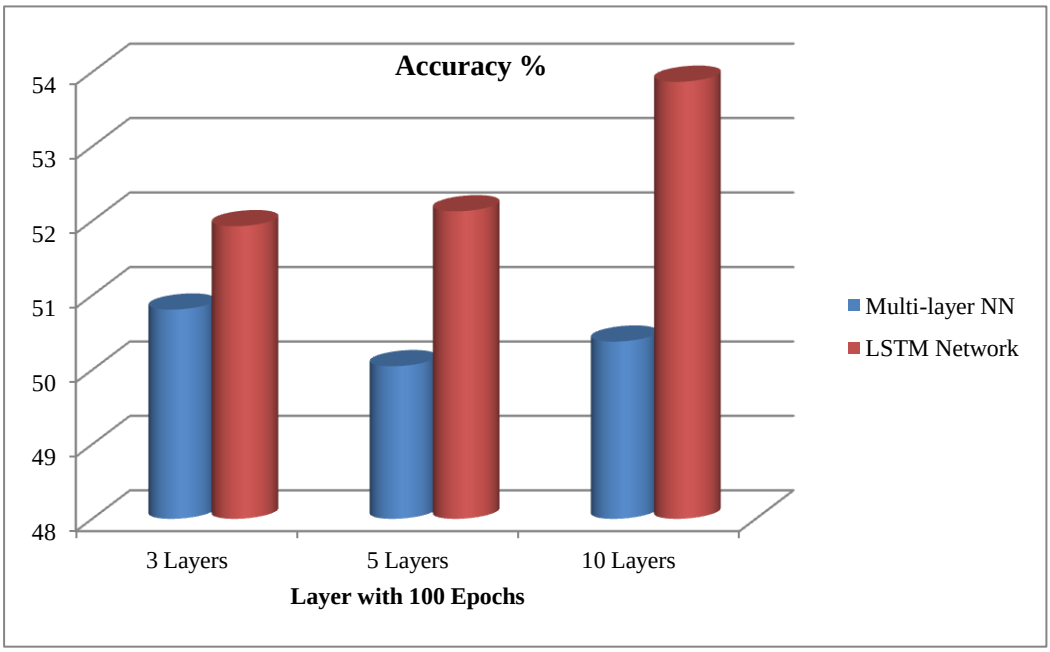

Figure 6: Overall comparison of MLP and LSTM Network

(vi) Observe the relationship between news sentiment and stock price: This phase investigate the relation between the predicted news sentiment and the DJIA stock data set. For evaluation we have used only adjusted close price of everyday stock price. If the news sentiment is found positive, then there is more probability that the stock price will grow

further and if it is negative, then stock price may go down. The relationship between the predicted news sentiment and stock price is represented in figure 7 and 8. The last 10 sample of news and stock are plotted in figure 9 and 10. In this plot, it can be observed that the negative news between 23-06-2016 to 27-06-2016 brings down the stock price.

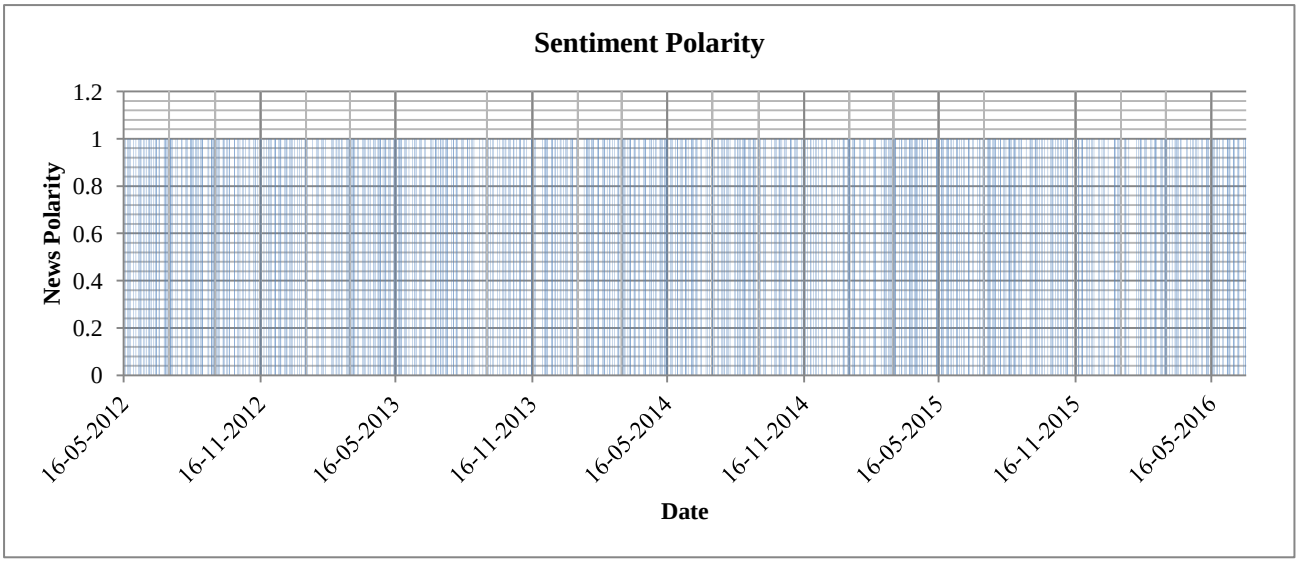

Figure 7: Time series plot of News Sentiment Score for Test Data set

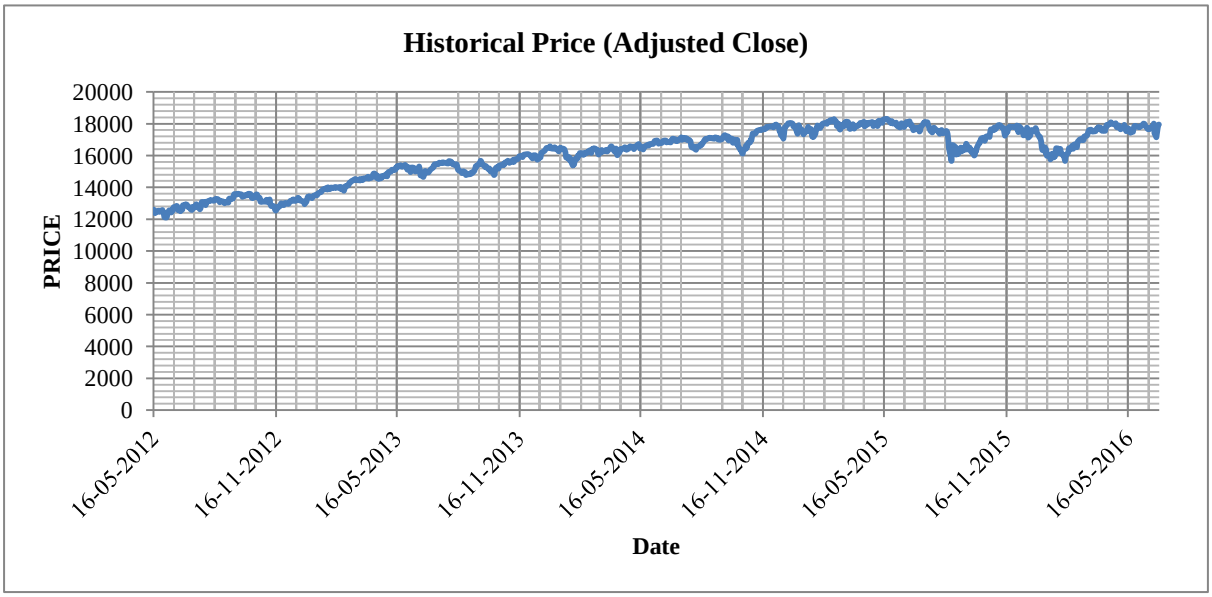

Figure 8: Adjusted Stock Price Plot for Test Data set

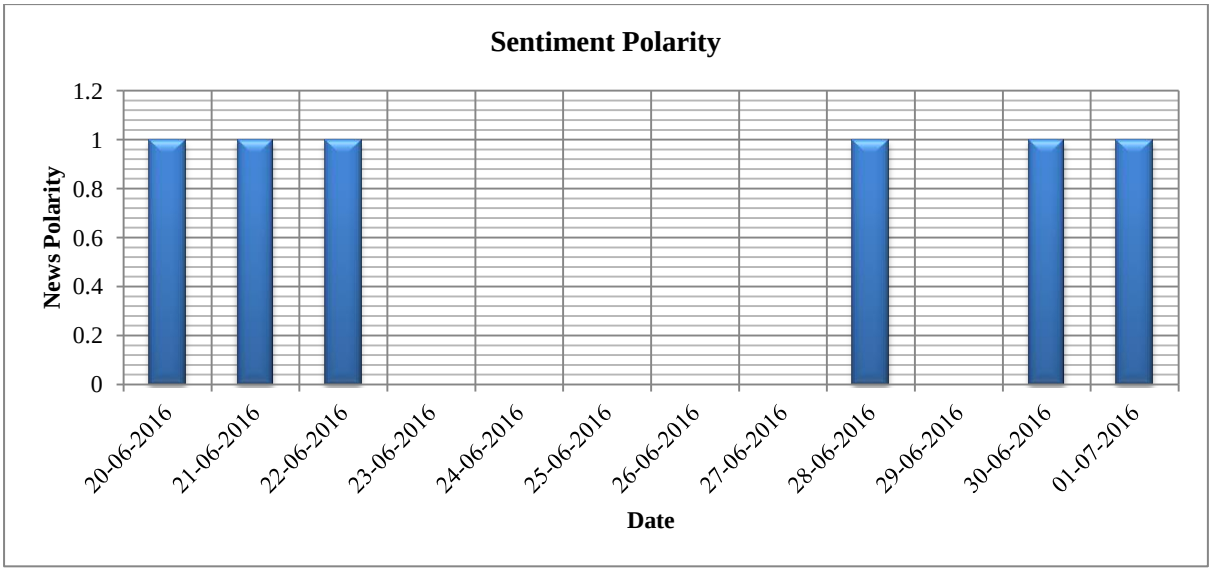

Figure 9: Top 10 Days plot of News Sentiment Score

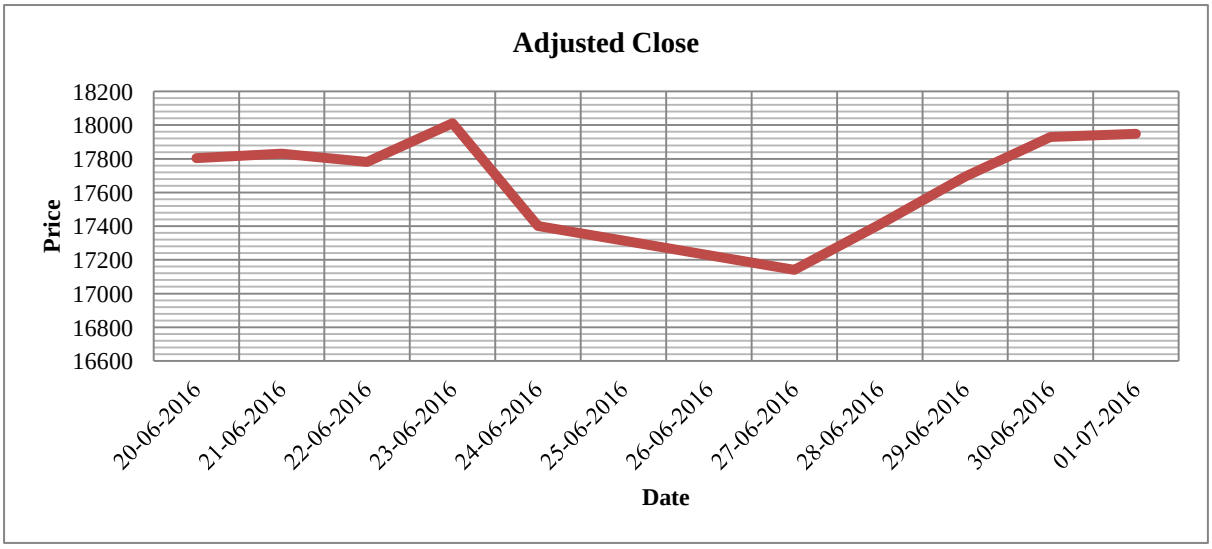

Figure 10: Top 10 Days plot of Actual Stock Price

4.3 Github location:

The complete code and description for this project is available at:

https://github.com/SUPRIYOPHD/Prediction/

Table III: Overall view of Model

Data set		Text Mining Features Processing			Machine Learning	
		Type of features	**Selection Methods**	**Market Feedback**	**Method**	**Accuracy Achieved**
DJIA News	Stock Market Price	Bag of words	Remove HTML tags, Stop words, Dealing with Punctuations, Select top 1000 words using Tf-Idr vectorizer	Yes	Deep Learning (MLP, LSTM)	53.87%

7. Conclusion and Future Work

In the finance field, stock trends are exceptionally crucial and volatile in nature. There are many factors by which the stock trends are affected, one of which is daily news articles. This work investigates the correlation between the large-scale daily news feeds and the stock market values over time. We have automated the sentiment detection from news articles based on words and formed word vector. Then classified the word vector and calculate the accuracy of predicted result using the techniques of deep learning models i.e. MLP and LSTM. Out of these two, the LSTM model gives better performance and achieves 53.87 % accuracy. Further, we have investigated the relation between predicted news sentiment and future stock market. If the sentiment is found positive, there is more probability that the stock price will grow further and if it is negative, then stock price may go down. In future we will try to optimize the model with- (i) other deep learning approaches such as Recurrent Neural Network (RNN), Restricted Boltzmann Machine (RBM) and Convolutional Neural Network (CNN) etc., (ii) other feature processing approaches i.e. Dictionary based and N-Gram etc. and (iii) data set of other companies.

References:

[1] Kalyani Joshi, Prof. Bharathi H. N., Prof. Jyothi Rao, Stock Trend Prediction Using News Sentiment Analysis, 2016

[2] S. Lauren and S. D. Harlili, "Stock trend prediction using simple moving average supported by news classification," 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA), Bandung, 2014, pp. 135-139.

[3] W.B. Yu, B.R. Lea, and B. Guruswamy, A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting, International Journal of Electronic Business Management. 2011, 5(3): 211-224
[4] Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S, 2008. "More than words: Quantifying Language to Measure Firms' Fundamentals", The Journal of Finance, Volume 63, Number 3, June 2008 , pp. 1437-1467
[5] Mittermayr, M.-A. ,"Forecasting Intraday Stock Price trends with Text Mining techniques", Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004
[6] R.P. Schumaker, Y. Zhang, C. Huang, H. Chen, Evaluating sentiment in financial news articles, Decision Support Systems, 2012, pp. 458–464.
[7] R.P. Schumaker, H. Chen, Textual analysis of stockmarket prediction using breaking financial news: the AZFin text system, ACMTransactions on Information Systems, 2009
[8] M.-A.Mittermayr, Forecasting intraday stock price trendswith textmining techniques, Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004.
[9] B. Wüthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, Daily stock market forecast from textual web data, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, 1998.
[10] F. Li, The information content of forward-looking statements in corporate filings — a naïve Bayesian machine learning approach, Journal of Accounting Research, 2010, pp. 49–102.
[11] W. Antweiler,M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, Journal of Finance, 2004, pp. 1259–1294.
[12] S.R. Das, M.Y. Chen, Yahoo! for Amazon: sentiment extraction from small talk on the web, Management Science, 2007, pp. 1375–1388.
[13] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More Than Words: Quantifying Language to Measure Firms' Fundamentals, 2008, pp. 1437–1468.
[14] S.S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Decision Support Systems 50 (2011) 680–691.
[15] M. Butler, V. Keselj, Financial forecasting using character N-Gram analysis and readability scores of annual reports, Advances in AI, 2009.
[16] S. Lauren and S. D. Harlili, "Stock trend prediction using simple moving average supported by news classification," 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA), Bandung, 2014, pp. 135-139.
[17] Michael Hagenau, Michael Liebmann, Markus Hedwig, Dirk Neumann, Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Specific Features, HICSS '12 Proceedings, 45th Hawaii International Conference on System Sciences, 2012