Stock Market Prediction Using Daily News Articles

Presented by,
Yashwant Singh Patel, Supriyo Mandal
Roll No. 1721CS05, 1721CS03
Dept. of Computer Science & Engineering
Indian Institute of Technology Patna

Outline

- Introduction
- Problem Statement
- Literature Survey
- Model Design
- Performance Evaluation
- Conclusion & Future Work
- Bibliography

Introduction

- In 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.

Problem Statement

- This work is an attempt to investigate the relationship between daily news articles and stock trend by predicting the future stock trend with news sentiment feature extraction and deep learning.

- **Data Sources:**

  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	17949.3691
30-06-2016	17712.76	17930.61	17711.8	17929.99	133030000	17929.9902
29-06-2016	17456.02	17704.51	17456.02	17694.68	106380000	17694.6796
28-06-2016	17190.51	17409.72	17190.51	17409.72	112190000	17409.720
27-06-2016	17355.21	17355.21	17063.08	17140.24	138740000	17140.2402
24-06-2016	17946.63	17946.63	17356.34	17400.75	239000000	17400.7
23-06-2016	17844.11	18011.07	17844.11	18011.07	98070000	18011.0703
22-06-2016	17832.67	17920.16	17770.36	17780.83	89440000	17780.8300
21-06-2016	17827.33	17877.84	17799.8	17829.73	85130000	17829.7304
20-06-2016	17736.87	17946.36	17736.87	17804.87	99380000	17804.8691
17-06-2016	17733.44	17733.44	17602.78	17675.16	248680000	17675.1601
16-06-2016	17602.23	17754.91	17471.29	17733.1	91950000	17733.0996
15-06-2016	17703.65	17762.96	17629.01	17640.17	94130000	17640.1699
14-06-2016	17710.77	17733.92	17595.79	17674.82	93740000	17674.8203
13-06-2016	17830.5	17893.28	17731.35	17732.48	101690000	17732.4804
10-06-2016	17938.82	17938.82	17812.34	17865.34	90540000	17865.3398
09-06-2016	17969.98	18005.22	17915.88	17985.19	69690000	17985.1894
08-06-2016	17931.91	18016	17931.91	18005.05	71260000	18005.0507
07-06-2016	17936.22	18003.23	17936.22	17938.28	78750000	17938.279
06-06-2016	17825.69	17949.68	17822.81	17920.33	71870000	17920.3300

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.

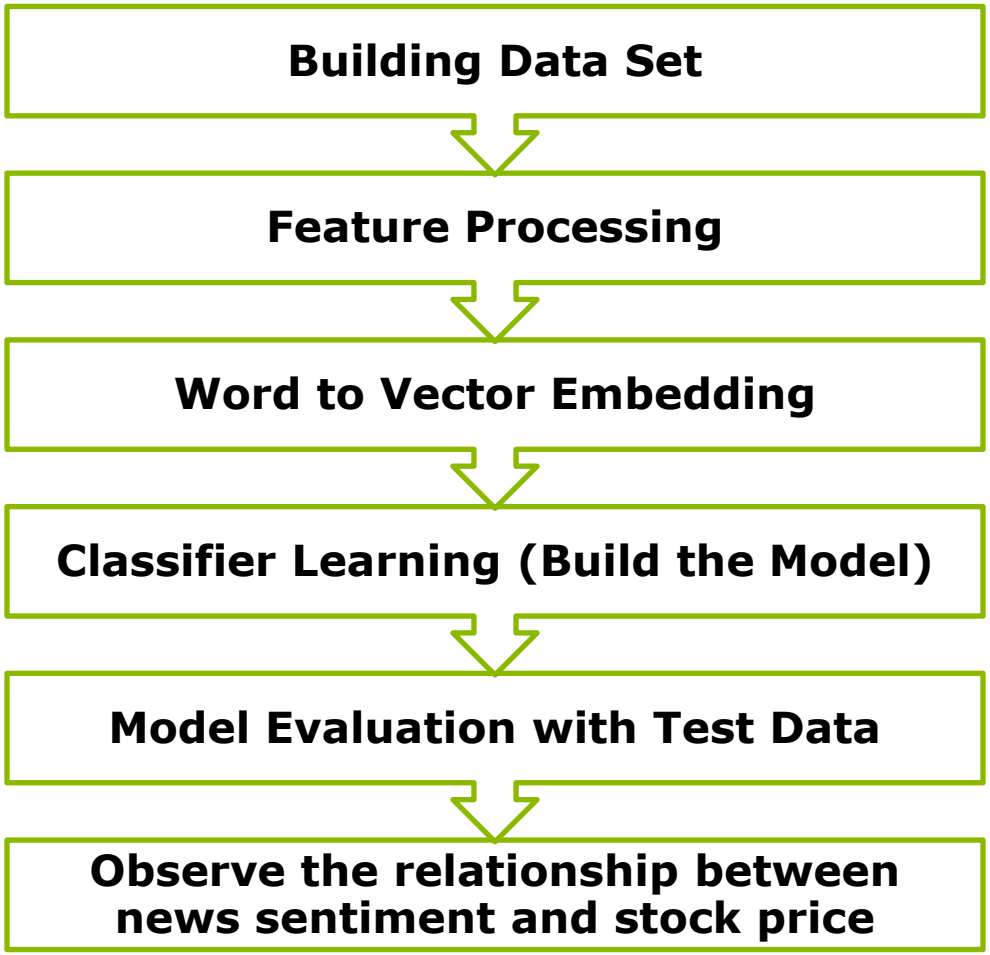
Literature Survey (1)

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

Literature Survey (2)

Year	Author(s)	Data set	Feature Processing Methods	Machine Learning Techniques	Accuracy
2009	Schumaker et al. 2009 [7]	US financial news	Noun phrases	SVM	58.2 %
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

Model Design



Step-1 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). DJIA news data set contains three attributes date, label and news of company.

- This data set is divided into 60% training and 40% test data set.

Step- 2 Feature 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 training 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'

Step-3 Word to vector embedding

- In this phase we have used bag of words and TF-IDF vectorizer.

Document 1		Document 2	
this	1	this	1
is	1	is	1
a	2	a	2
brown	1	brown	2
box	2	bag	4

- tf("this", d1) = 1/7 =0.14, tf("this", d2) = 1/10 = 0.1
- idf("this", D) = log(2/2) =0
- tfidf("this",d1) =0.14*0 =0
- tfidf("this",d2) =0.1*0 =0

Sample Output

```
  :   :
  (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
  (1070,  990)      0.174386266654
  (1070,  713)      0.250557075816
  (1070,  134)      0.236631076376
  (1070,  775)      0.223714111291
  (1070,  678)      0.228444813105
  (1070,  239)      0.223714111291
  (1070,  700)      0.211933919307
  (1070,  957)      0.259567217966
  (1070,  745)      0.285958921013
  (1070,  737)      0.254836516152
  (1071,  12)       0.179081603071
  (1071,  702)      0.492663497205
  (1071,  219)      0.47265102011
  (1071,  302)      0.509014881584
  (1071,  943)      0.492663497205
[[ 0.  0.  0. ...,  0.  0.  0.]
 [ 0.  0.  0. ...,  0.  0.  0.]
 [ 0.  0.  0. ...,  0.  0.  0.]
 ...,
 [ 0.  0.  0. ...,  0.  0.  0.]
 [ 0.  0.  0. ...,  0.  0.  0.]
 [ 0.  0.  0. ...,  0.  0.  0.]]
('X_train shape:', (1072, 1000))
('X_test shape:', (917, 1000))
```

Step-4 Classifier Learning

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

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

Step-5 Performance Evaluation (1)

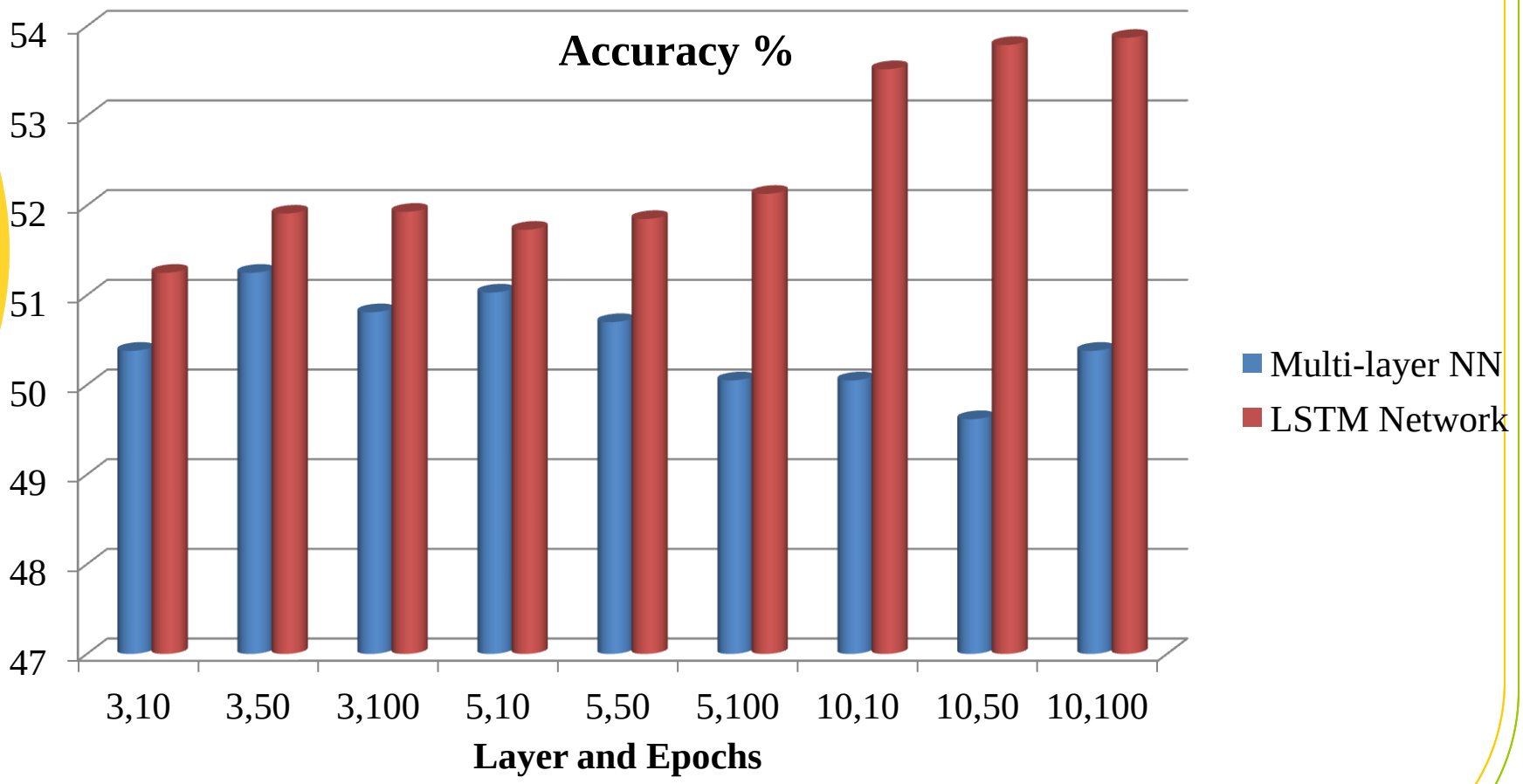


Figure 5: Comparison based on Layers and Epochs

Step-5 Performance Evaluation (2)

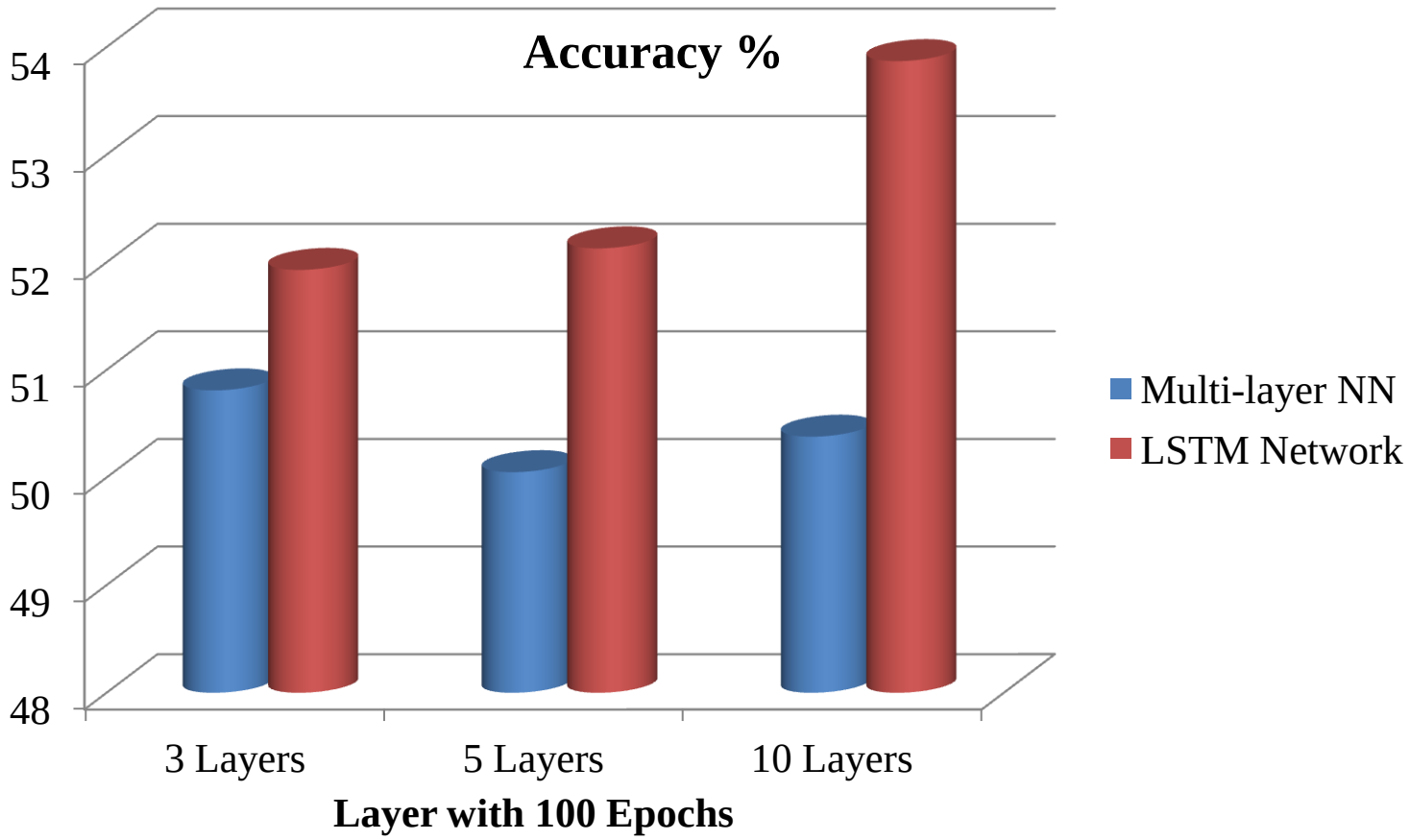


Figure 6: Overall comparison of MLP and LSTM

Step-5 Performance Evaluation (3)

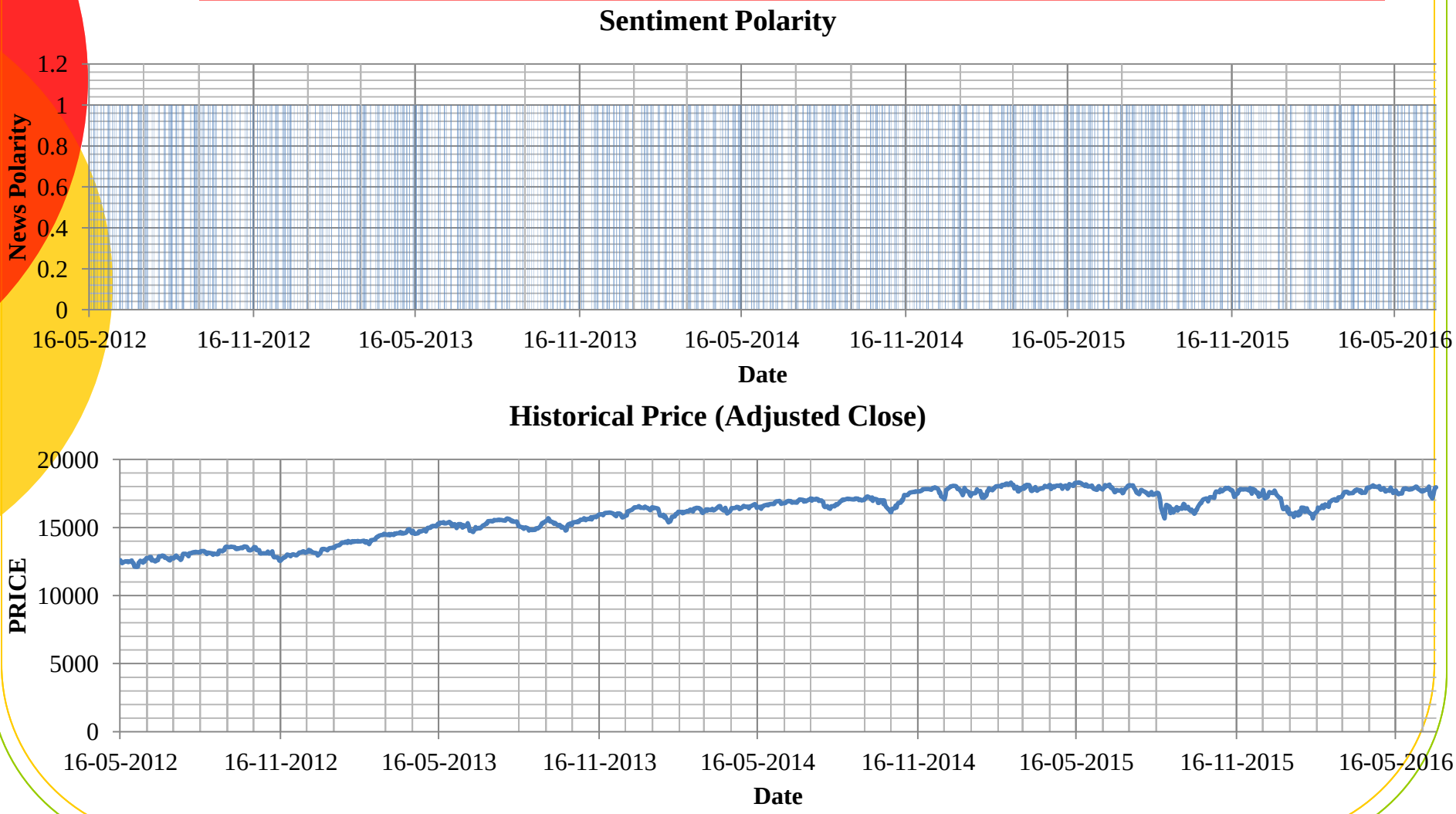


Figure 7: Time series plot of News Sentiment Score and Stock Price for Test Data set

Figure 9: Last 10 Days plot of News Sentiment Score and Stock Price

Overall View of Model

Data set		Text Mining Features Processing			Machine Learning	
		Type of features	Selection Methods	Market Feedback	Method	Accuracy
DJIA News	Stock Market Price	Bag of words	Removal of HTML tags, Stop words, Dealing with Punctuations, Select top 1000 words using Tf-Idf vectorizer	Yes	Deep Learning	53.87% with LSTM (10 layers and 100 epochs)

Github Source

- Complete source code is available at:

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

Conclusions & Future Work

- We have automated the sentiment detection from news articles and calculated the accuracy of predicted result using the techniques of deep learning models i.e. MLP and LSTM
- Then we have investigated the relationship between predicted news sentiment and future stock market. In future we will try to optimize the model with:

(i) Other Deep learning approaches i.e. RBM, RNN, CNN

(ii) Other Feature processing approaches i.e. Dictionary, N-gram, Noun-phrases and

(iii) Data set of other companies.

Bibliography (1)

[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.

Bibliography (2)

[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.

Bibliography (3)

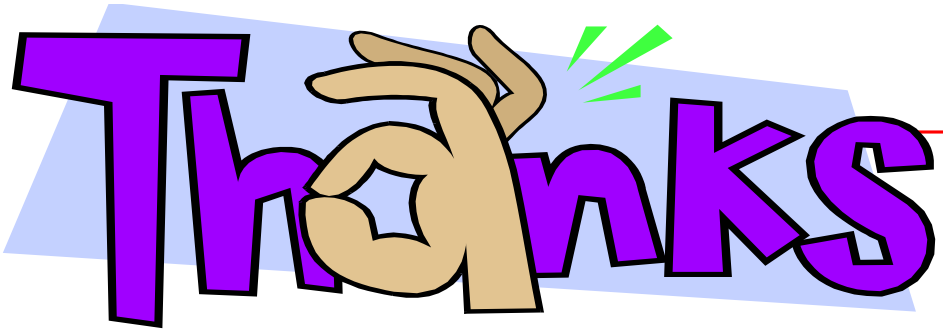
[13] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More Than Words: Quantifying Language to Measure Firms' Fundamentals, 2008, pp. 1437–1468.

[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



ANY QUERIES ???????