A STOCK MARKET PREDICTION AND REVIEW USING DEEP LEARNING,SENTIMENT ANALYSIS AND MACHINE LEARNING

Karthik Ram Srinivas
Department of Computer Engineering
Mukesh Patel School of Technology Management and Engineering, NMIMS University
Mumbai , India
karthikramsrinivas@gmail.com

***Abstract*-One of the most complex and sophisticated ways of doing business is the stock or share market. In order to generate revenue and divide risks, small ownerships, brokerage firms, the banking sector all rely on this very body. Stock prices often rely significantly on new knowledge. One of the many sources of knowledge is the opinions of people on social media. Opinions of people on products from some companies will decide the reputation of the company and therefore influence the decision of people to purchase the company's stock. So, one should use opinion as primary data cautiously. In the paper various machine learning models such Applying various ML algorithms including Linear Regression, RNN, KNN and, most successfully, a LSTM deep neural network have been used for stock market prediction. Further other Machine learning models such as Sentiment Analysis , Naïve Bayes , SVM , Random Forest have been reviewed . The main concept for effective stock market prediction is not only to produce the best returns and reduce risk, but also to minimize the wrong stock price forecast.**

Keywords—Machine Learning , KNN , LSTM , Linear Regression.

I. INTRODUCTION

The financial sector is fully utilizing structured data, such as time series trading data. Consider an example of forecasting stock market patterns by fundamental analysis and technical analysis. The vigorous nature of the market makes it tough to use time series or regression algorithms. Many financial firms and traders have developed a number of closed-source software models to capture the market, but only very few have consistently achieved high returns on investment. However, the reward of stock forecasting is so enticing because an enhancement of only a few margins will boost the gains of these institutions by millions of dollars. It has, because of its financial benefit, attracted a lot of attention from both commercial and academic sides.

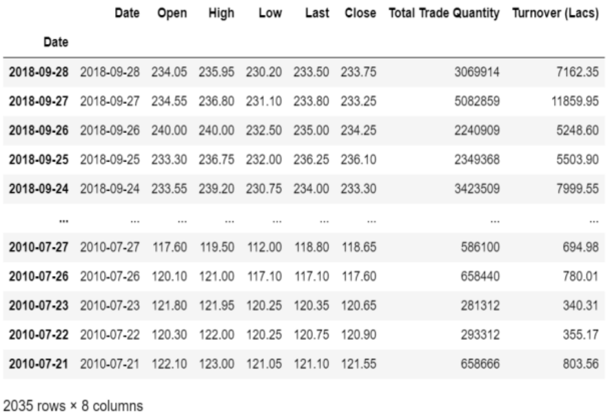
People are making their opinions on social media that can be shared by Others too. Emotion categorization can be done at word, phrase, sentence and document level. Sentiment analysis has now become a popular approach used to derive feelings and input from social media outlets like twitter.

Certain informational and emotional events, such as negative comments on Twitter/social media and news, may cause fear in the market and push investors to overwhelmingly sell a specific share or company. The opposite can also be true when positive news is released, which may translate into optimism and perhaps boost the price of a given stock. That initial rush of fear or excitement, creating outsized moves in the market can quickly create overbought or oversold conditions.

For example, after January 20th 2016, Lockheed Martin’s stock went down by 5% after Donald Trump tweeted that the F-35 program and cost was out of control and billions of dollars could be saved on military purchases. Not only Presidents but social media influencers have an effect on the stock market, in February 2018 social media influencer Kylie Jenner with over 39 million followers tweeted the following:
“sooo does anyone else not open snapchat anymore? Or is it just me... ugh this is so sad.” This tweet had a large impact on the share price of SNAP, the parent company of Snapchat. Within a day, the share price decreased by 7% and SNAP lost approximately $1.3 billion in market value.

II. DATASET

NSE Tata Global Dataset was imported from Kaggle , it consists of 2035 row x 8 columns

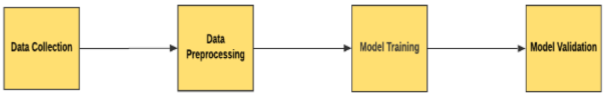


	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
Date								
2018-09-28	2018-09-28	234.05	235.95	230.20	233.50	233.75	3069914	7162.35
2018-09-27	2018-09-27	234.55	236.80	231.10	233.80	233.25	5082859	11859.95
2018-09-26	2018-09-26	240.00	240.00	232.50	235.00	234.25	2240909	5248.60
2018-09-25	2018-09-25	233.30	236.75	232.00	236.25	236.10	2349368	5503.90
2018-09-24	2018-09-24	233.55	239.20	230.75	234.00	233.30	3423509	7999.55
...	...	...	...	...	...	...	...	...
2010-07-27	2010-07-27	117.60	119.50	112.00	118.80	118.65	586100	694.98
2010-07-26	2010-07-26	120.10	121.00	117.10	117.10	117.60	658440	780.01
2010-07-23	2010-07-23	121.80	121.95	120.25	120.35	120.65	281312	340.31
2010-07-22	2010-07-22	120.30	122.00	120.25	120.75	120.90	293312	355.17
2010-07-21	2010-07-21	122.10	123.00	121.05	121.10	121.55	658666	803.56

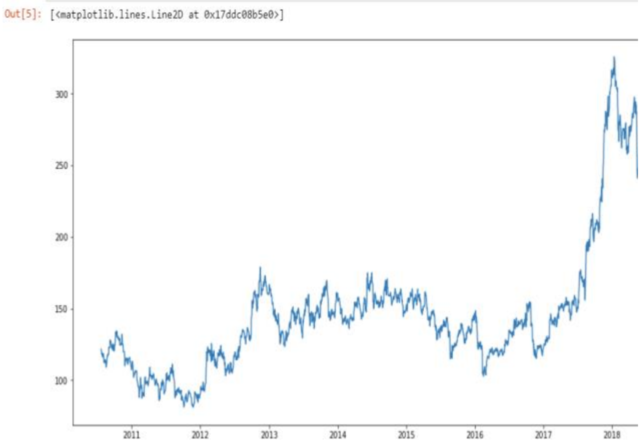
2035 rows × 8 columns

III. RESULTS

I have followed the methodology of collecting data then pre processed it , then trained the model and finally validated the model .



I analyzed the closing prices from dataframe:

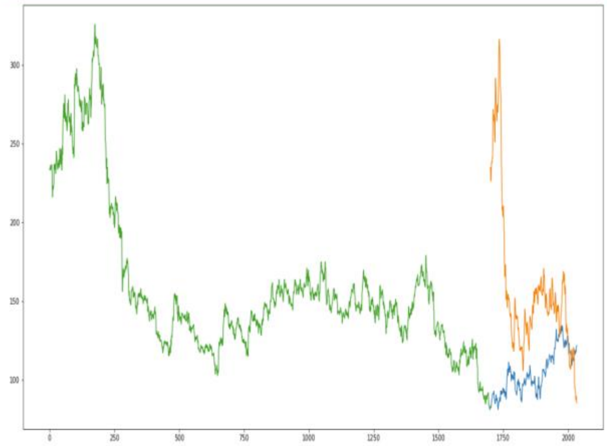


```
Out[5]: [<matplotlib.lines.Line2D at 0x17ddc08b5e0>]
```

KNN:

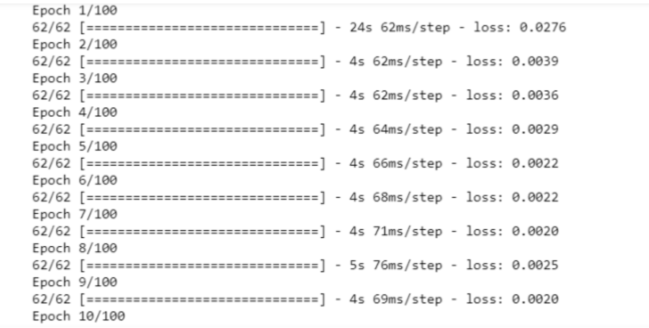
The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness).First I converted the data to only date and closing price, then train test split , further used KNN , scaled the data , found the best parameters using cross validation , then fit , predicted and plotted the data and viewed an accuracy of 78.00387709000377 and the graph predicted stock price similar to actual stocks

KNN's main disadvantage of becoming significantly slower as the volume of data increases makes it an impractical choice in environments where predictions need to be made rapidly.



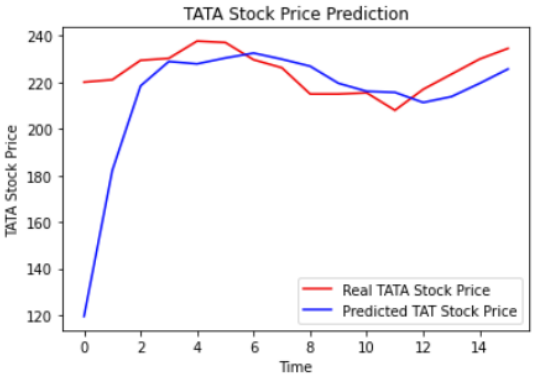
RNN:

Recurrent Neural Network (RNN) are a type of Neural Network where the output from previous step are fed as input to the current step . RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is Hidden state, which remembers some information about a sequence. First compiled the RNN model and added LSTM layer and fit the RNN to the training dataset



```
Epoch 1/100
62/62 [==============================] - 24s 62ms/step - loss: 0.0276
Epoch 2/100
62/62 [==============================] - 4s 62ms/step - loss: 0.0039
Epoch 3/100
62/62 [==============================] - 4s 62ms/step - loss: 0.0036
Epoch 4/100
62/62 [==============================] - 4s 64ms/step - loss: 0.0029
Epoch 5/100
62/62 [==============================] - 4s 66ms/step - loss: 0.0022
Epoch 6/100
62/62 [==============================] - 4s 68ms/step - loss: 0.0022
Epoch 7/100
62/62 [==============================] - 4s 71ms/step - loss: 0.0020
Epoch 8/100
62/62 [==============================] - 5s 76ms/step - loss: 0.0025
Epoch 9/100
62/62 [==============================] - 4s 69ms/step - loss: 0.0020
Epoch 10/100
```

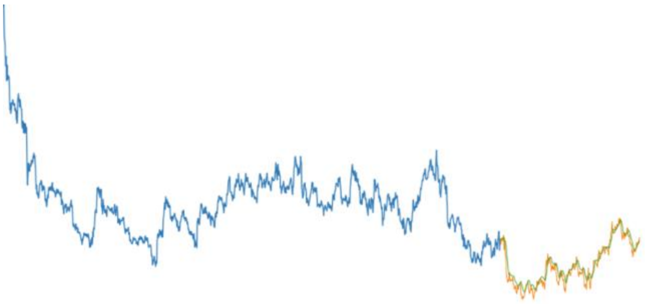
Then predicted and visualized the results



LSTM

A stacked LSTM model is trained on the data, and evaluated by measuring the Mean Squared Error (MSE) and Mean Absolute Diviation (MAD) of the model. 90% of the data is used as training while the remaining 10% is used as testing and validation. Stacked LSTM networks provide a deeper model for learning and are composed of multiple hidden layers of LSTMs. These multiple hidden layers act as a Deep Recurrent Neural Network (DRNN).

Adding 3 layers of LSTM and some Dropout regularization, LSTM provides the highest accuracy while prediction as seen in the graph.



IV. REVIEW AND RELATED WORK

SENTIMENT ANALYSIS

It is a method for classifying the polarity of opinion using natural language processing and text analysis. Sentiment Lexicon is a word in the sentence that contains the sentences emotion and plays an important role in deciding polarity. Positive and negative lexicons are the two kind of lexicons that exist. Positive lexicon is an expression reflecting positive emotion and negative lexicon is a phrase expressing negative feeling or emotion.

Examples of positive lexicon are good, fantastic and examples of negative lexicon are bad and ugly. Classifying by supervised learning (Machine learning algorithms like Random Forest, SVM etc.) is one of the approaches that can be used in sentiment analysis. The approach classifies views derived from the data.

POS Tagging: POS stands for Part of Speech. They are typically used to identify the nature of word in a sentence. Extracting information is easy if type of the word is known. For example using POA tagging we can determine whether the word was a noun, adjective and verb etc.

Words and its weight: Calculating the weight of the words in the text makes it easy to decide which word is relevant and which is not.

Sentiment shifters: Sentiment shifters are a set of words or expression that converses the emotion or feel of a sentence.

The algorithm used to calculate the document's sentiment score is given below.

1. Perform text tokenization as part of pre-processing.
2. Form word vector of the tokens.
3. Making a dictionary containing words with their polarity. Polarity can be positive or negative.
4. Check for each word weather it matches with the words from positive word dictionary or negative words dictionary.
5. Find total number of words that belong to positive as well as negative polarity.
6. Score of a document can be calculated by subtracting the count of positive matches by the count of negative matches:

   Score of documents = count (PM) – count (NM)

   PM=Positive matches

   NM=Negative matches
7. If the result is 0 or less, then we consider the document is negative otherwise it is considered positive.

There are various algorithms like Naïve Bayes, Random Forest, Decision Tree, and Support Vector Machine (SVM) that can be used for supervised classification.

A. *Random Forest*

The Random Forest is made up of a large number of decision-making trees where each tree gives a class prediction. The class with maximum votes becomes the model prediction. The basic concept of a random forest is simple but robust, the wisdom of the masses.

Random forest also de-correlates the independently built trees by randomly choosing a subset of the available features (predictors) to build the tree. This procedure avoids the possibility of the independently constructed trees being highly correlated on account of one or two very strong predictors.

Yahya Eru Cakra [1] used basic sentiment analysis to estimate the Indonesian stock market. Random Forest is one of the algorithms used to classify tweets in order to measure a company's sentiment. The outcomes of which is used in estimating the stock price of the company. To construct the prediction model, he used the linear regression method. The output of his experiment showed that prediction models having a hybrid function (created by a combination of stock price on a given day and the positive tweet percentage on that day) and previous stock price as a predictor can predict with a determination coefficient of 0.9989 and 0.9983. Classification of sentiment into positive and negative obtained by using Random Forest with an accuracy of 60.39% was the highest among all the other algorithms used. The result of price fluctuation prediction model was also highest for random forest with an accuracy of 67.37% for a tweet data of 4 days.

Kalyani Joshi et al [2] considered news articles about a business as prime details, their research follows the Fundamental Analysis Methodology to identify a stock's future trend and classifying news as positive and negative. Chances the stock price will increase if news sentiment is good, while for the negative news sentiment, the stock price will decrease. They build their model on Apple's stock price and news articles related to the company from last three years as data. They implemented three algorithms Random Forest, SVM and Naïve Bayes in different situations. After finding the results for the different models, Random Forest stood out with an accuracy between 88% to 92% in the different test cases.

Bhaskar Tiwari [4] used the twitter dataset for his thesis which was collected using a custom crawler, that relies upon twitter's native search functionality to extract the relevant tweets. The crawler fetches data from twitter based on relevant search keywords. They used the name of the company and the abbreviation of the company name (for example: ICICI for ICICBANK) as relevant keywords in their data collection strategy and extracted a total of 1,19,116 tweets.

In their analysis they varied the number of trees(M) in the random forest model and ran a 10-cross validation on the training set. This way found the number of trees(m=300) giving the highest accuracy.

With this value of the parameter, they trained their classifier with the entire training data.

After testing the model on the test dataset their optimized classifier with M = 300 gave an overall accuracy of 67.8%

B. *Support Vector Machine(SVM)*

SVM generates a linear algorithm, that creates a function to maximize the distance between classes. Class function is created using instance data at the edge of the class. The data points lying closest to the decision boundary are the most difficult to identify and they have a direct impact on the optimum position of the decision surface. SVM algorithm can achieve this by creating a function that maximizes the distance.

The sentiment analysis model presented by Yahya Eru Cakra [1] showed SVM had an accuracy of 38.64% which was lowest from Random Forest, Naïve Bayes and Decision Tree. In price fluctuation forecast, he presented models that can predict an increase or decrease in the future stock price. SVM showed an accuracy at 66.34 % on all 5 days which was just behind that of Random Forest whose highest accuracy of 67.37 was achieved on day 4.

Kalyani Joshi et al [2] created their own dictionary to remove stop words and added stop words related to finance. Stop words are words in the language that do not add significant meaning to a sentence. Common stop words are he, his, your, does, do, if etc. They can be easily ignored without losing the meaning of the sentence. On analysing this data, three classification models were implemented and tested under different situations. After comparing the performances, SVM had second highest accuracy of 86% behind Random Forest which showed an accuracy ranging from 88% to 92% accuracy.

Bhaskar Tiwari [4] had tested the classifier on four test sample data sets. Their optimized classifier with C = 105 and γ = 10-2 gives an overall accuracy of 70.8% (C is the regularization coefficient that controls the trade-of between the simplicity of the decision boundary). Performance of SVM with Gaussian RBF kernel (C = 105 and γ = 10-2) on the four test data sets AXISBANK showed 70%, ICICIBANK 70.44%, HDFCBANK 72.66%.

C. Naïve Bayes

Naïve Bayes is a machine learning model uses Bayes theorem to classify objects. This model is used with vast amounts of data. Whenever we are dealing with data that has millions of data records, Naïve Bayes is the preferred solution. This provides very good results when it comes to NLP activities, such as sentimental analysis. It's a fast and uncomplicated classification algorithm.

Yahya Eru Cakra [1] showed that sentiment analysis model based on Naïve Bayes classification can classify tweet data with an overall accuracy of 56.50% which was second highest after Random Forest. In price fluctuation prediction, over 5 days Naïve Bayes had its highest accuracy at 66.34% which was constant until day 3, on day 4 it went down to 61.39 and on day 5 to 57.43.

Kalyani Joshi et al [2] executed various classification models and tested in various test conditions. Their performance analysis showed that Naïve Bayes algorithm performance is around 83%, behind Random Forest and SVM.

D. Decision Tree

The decision tree is a supervised machine learning algorithms used to solve regression and classification problems without requiring much computation.

Yahya Eru Cakra [1] presented that accuracy of Created sentiment analysis model using decision tree was 46.02 %. This result showed that this approach was less effective than Random Forest, Naïve Bayes but more effective than SVM. In price fluctuation prediction Decision Tree had a constant accuracy of 66.34% over 5 days like SVM.

V. CONCLUSION

This study proposes the use of a stacked LSTM network model for predicting stock market behavior, using data from NSE Tata Global Dataset Composite . The model was trained and the results obtained show that the model was able to predict stock market behavior with some accuracy. The Indian stock market has numerous ups and downs. It is very critical for investors to forecast the stock market situation in order to invest money in the stock market to purchase shares. Sentiment analysis for the stock market has been demonstrated by fetching live data values at various time intervals to forecast status of the stock market. It will assist investors to predict what shares should be invested with money.

Social Media forums such as twitter, Facebook, Instagram, blogs, news channels offer a variety of information about people's views, opinions, feelings and sentiments about a specific problem or product. Even after using this as a source for sentiment analysis data, it is insufficient further analysis and research is required

Through our survey we found that:

1. LSTM comes closest to predicting the stock price accurately
2. RNN Model can predict price very close to the actual price
3. KNN Model is less accurate than the KNN model but it is also useful to predict the actual price.
4. R square value can be reduce by high percent of positive or optimistic tweets
5. Among Sentiment analysis models Random Forest identifies twitter data with highest precision, followed by Naïve Bayes.
6. News may cause fluctuation in the stock market. So, after analyzing the relationship between news and stock market, we arrived to the conclusion that news articles and previous price history can help in prediction of movement in stock price. Positive news has positive effect on stock price and negative news negative effect.
7. Public opinion and sentiment can be recorded from twitter , Instagram , Facebook feeds using basic NLP techniques like sentiment analysis.
8. The research on Twitter data shows a trend of predicting the outcome of the election results.
9. Molla et al [7] demonstrated that news articles with similar relation to a paticular stock improved the outcome of financial prediction. Many kernel learning techniques were used to segment the data from five separate groups collected from news articles based on industries, sectors, sub-sectors, etc.
10. Anshul Mittal [3] Used machine learning concepts to estimate a correlation between the opinion of the general public regarding the stock and the views of the stock traders/ stock market veterans. They used twitter data to forecast the public opinion towards the stock and predict its movement in the market. He proposed new cross validation approach to predict financial data which resulted into 75.56 percent accurate results from June 2009 to December 2009 using Self Organizing Fuzzy Neural Networks (SOFNN) on Twitter feeds and DJIA values. He also adopted a Naïve algorithm approach on their forecasted values. His research is based on Bollen et al's seminal paper"Twitter mood as a Stock Market predictor" that gave 87% accurate result.

VI. FUTURE WORK

Using advanced deep learning algorithms to improve accuracy An open issue remains in that the volatility of the stock market cannot be mitigated using only historic data, but factors of the present also need to be analysed including current news in the world of politics and economics that could affect the behavior of investors and ipso facto the behavior of stock markets.

VII. REFERENCES

[1] Stock Price Prediction using Linear Regression based on Sentiment Analysis Yahya Eru Cakra Faculty of Computer Science Universitas Indonesia Depok,Indonesiayahya.eru@ui.ac.id Bayu Distiawan Tri sedya Faculty of Computer Science Universitas Indonesia Depok, Indonesia b.distiawan@cs.ui.ac.id

[2] STOCK TREND PREDICTION USING NEWS SENTIMENT ANALYSIS Kalyani Joshi1 , Prof. Bharathi H. N.2 , Prof. Jyothi Rao3 1Department of Computer Engineering, KJSCE, Mumbai 2Department of Computer Engineering, KJSCE, Mumbai 3Department of Computer Engineering, KJSCE, Mumbai

[3] Stock Prediction Using Twitter Sentiment Analysis Anshul Mittal Stanford University anmittal@stanford.edu Arpit Goel Stanford University argoel@stanford.edu

[4] Intraday Stock Trend Prediction Using Sentiment Analysis by Bhaskar Tiwari under the guidance of Dr. Diganta Mukherjee Associate Professor Sampling and Official Statistics Unit

[5] Celikyilmaz, A., Hakkani-Tur, D., & Feng, J, Probabilistic model-based sentiment analysis of twitter messages, IEEE Spoken Language Technology Workshop, 2010, pp. 79-84.

[6] Wang, X., & Luo, X, Sentimental Space Based Analysis of User Personalized Sentiments, IEEE 9th International Conference on Semantics, Knowledge and Grids, October 2013, pp. 151-156.

[7] Molla, A., Biadgie, Y., & Sohn, K. A. Network based visualization of opinion mining and sentiment analysis on

twitter, IEEE International conference on It Convergence and Security, 2014, pp.1-4.

[8] Yauheniya Shynkevich, T.M. McGinnity, Sonya Coleman, Ammar Belatreche, Predicting Stock Price Movements Based on Different Categories of News Articles, 2015 IEEE Symposium Series on Computational Intelligence

[9] R. Goonatilake and S. Herath, The volatility of the stock market and news, International Research Journal of Finance and Economics, 2007, 11: 53-65.

[10] L. Breiman, Random forests. Machine Learning, 45(1):5-32, 2001

[11] Zhang, X., Fuehres, H., & Gloor, P. A, Predicting stock market indicators through twitter, ELSEVIER Procedia-Social and Behavioral Sciences, 2011, pp. 55-62

[12] Public opinion and sentiment can be recorded from twitter , Instagram , Facebook feeds using basic NLP techniques like sentiment analysis.

[13] Molla et al [7] showed how the outcomes of financial prediction can be enhanced when news

[14] Anshul Mittal [3] Used machine learning concepts

[15] A.E.Stefano Baccianella andF. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC. LREC.

[16] A. Lapedes and R. Farber. Nonlinear signal processing using neural network: Prediction and system modeling. In Los Alamos National Lab Technical Report.

[17] G. P. Gang Leng and T. M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks. Neural Networks, 17(10):1477–1493.

[18] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[19] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining - Practical Machine Learning Tools and Techniques (3rd Ed). Burlington: Morgan Kaufmann, pp. 124-125