Stock Market Prediction Using LSTM and Sentiment Analysis

Sreyash Urlam[1], Bijit Ghosh[2], Dr. A. Suresh*[3]

[1]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur,Chengalpattu,Tamil Nadu, India.
[2]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur,Chengalpattu,Tamil Nadu, India.
[3]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur,Chengalpattu,Tamil Nadu, India.
[1]sr7661@srmist.edu.in, [2]bs9989@srmist.edu.in, [3]prisu6esh@yahoo.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Prediction and analysis of stock market data is paramount in today's day and age. Since the economic interactions are too complex for shallow neural networks this paper implements Long Short Term Memory (LSTM) neural networks. LSTM is chosen as it helps to vectorise the data and thus give better predictions. We've also utilised other algorithms to show the effectivity of each of these algorithms pitted against LSTM. A very important factor to consider while predicting the stock market is the mood of the people. A person's emotions have the power to influence the stock market. Sentiment analysis on twitter is used to identify a correlation amongst the future of the stock and the general public's mood. Our paper works on comparing the sentiment analysis and the predicted stock value and showing that the two are rather similar and that people's emotions affects the future of the stock prices and to do a comparative study between prediction with and without using the results of the sentiment analysis.

Keywords: - Long Short Term Memory, Recurrent Neural Network, Sentiment Analysis

1. Introduction

1.1 SENTIMENT ANALYSIS

Sentiment Analysis is a mechanism of analysing, recognizing, and categorizing by computing a set of text such as a sentence, paragraph, or even pages in order to evaluate if the essence of the text is positive, negative, or neutral. It's being exploited in several societies relating to commercial advertisements and research for better understanding the moods of the people which not only helps them in targeting the right audience but also in understanding what changes need to be made based on the result. The result, as mentioned, will be represented arithmetically in decimals between -1 (most negative) and +1 (most positive). Sentiment Analysis is often confused with Opinion Mining wherein the unit is used to identify a certain theme, or occurrence, or in other words, to measure the frequency of a similarity or dissimilarity.

Another manner to utilize this idea is to use the know-how we benefit from analysing humans' reviews of an agency on social media and discover a manner to attach it to the agency's inventory marketplace tendencies to make higher predictions of the agency's inventory expenses [1]. A top social media platform to accumulate statistics from for this reason is Twitter due to a couple of reasons such as its huge consumer base and a hundred and forty man or woman length restriction on every tweet. Many researchers have laboured in this subject with the pioneer. One of the most distinguished and progressive being [2][18]Bollen and Mao. The technique consists of numerous steps of statistics series i.e. gathering tweets primarily based totally on numerous parameters like hashtags, mentions and more. The textual content withinside the Tweets are generally polluted with undesirable characters like extra punctuation, emoticons etc. in order that they need to be pre-processed i.e. wiped clean to extract simplest beneficial textual content. Next step is to categorize the general sentiment of the Tweet into a category primarily based totally on a few scales selected through the researcher like positive, bad or impartial. The very last step is the use of the results of the above step to formulate a set of rules that predicts the stock prices as correctly as possible [3] [4] [5].

2. Literature survey

2.1 ARTIFICIAL NEURAL NETWORK

Moghaddama et al. studied the capacity of an Artificial Neural Network (ANN) in predicting the daily NASDAQ inventory trade rate. The neural community became a feed ahead neural community that became skilled in the usage of the back propagation algorithm[7][17]. They considered the short-time period historic inventory

expenses together with the day of the week as inputs for their methodology. Two different types of datasets had been taken and the distinction of their overall performance became measured. One dataset took into attention four previous days and the opposite took nine previous days. According to their findings, the synthetic neural community with 3 hidden layers and 20-40-20 neurons in every concealed layer gave the exceptional prediction and the 2 datasets gave comparable predictions that means that there's no awesome distinction among the forecasting capacity of the 4 and 9 previous operating days as enter parameters.

2.2 MODIFIED INPUT FOR ARTIFICIAL NEURAL NETWORK

To the authors Pang et al conventional neural community algorithms can carry out insufficiently or incorrectly whilst used to take a look at the impact of marketplace traits at the expense of the stocks [10][15]. This is because of the truth that the initial weight of the random choice is liable to erroneous predictions. The authors have evolved an idea of "inventory vector" wherein enter isn't always an unmarried index however is a multi-inventory and multi size historic information.The concept of an inventory vector is just like that of the phrase vector often utilized in Natural Language Processing (NLP) algorithms in order that the phrase is converted right into a shape that is comprehensible through the computer [11][16]. One-warm encoding is used to gain this. The identical idea is applied on inventory expenses and the enter parameter is multi-dimensional and is the historic information of a couple of stocks.For the prediction version, they have got used a changed Long Short Term Memory (LSTM) community that has an embedded layer and every other that has an automated encoder and feature as compared to the effects [12] [13]. They have observed that an LSTM community with the embedded layer offers a better prediction accuracy. However, this paper has studies accomplished particularly for the Chinese inventory marketplace and its overall performance may also range for European or American inventory markets.

2.3 DEEP LEARNING MODELS

The authors M. et al. on this paper have compared the numerous deep learning prediction models and additionally linear models like Auto Regressive Integrated Moving Average (ARIMA) and their man or woman capacity in predicting inventory expenses from one of a kind inventory exchanges particularly the National Stock Exchange (NSE) and the New York Stock Exchange (NYSE) [6]. The models had been skilled at the NSE dataset and had been made to expect the NYSE datasets. On doing this it becomes observed that the deep getting to know fashions done nicely whilst predicting values in each the exchanges considering the fact that they ought to efficaciously perceive the common underlying sample in each the exchanges. However, the linear fashions couldn't achieve this. Additionally,out of all of the deep getting to know fashions used, the CNN outperformed the others i.e., effortlessly and continually observed the fashion of the values accordingly reinforcing that CNN can manage abrupt modifications withinside the fashion withinside the education set effortlessly [8] [9].

2.4 HEURISTIC ALGORITHMS

Small start-up investments do now no longer have the coins float important to make huge and long-time investments withinside the inventory marketplace. They are greatly involved and become making medium sized investments for intervals starting from per week to 3 months.The authors of the paper Zhang et al. have tried to cope with the trends of stock value projection during such periods [14]. In this paper, it is proposed that a singular information-pushed structure known as Xuanwu. The system completes all of the numerous machines getting to know processes involved in creating a suitable stock prediction version that consist of producing education samples from the unique transaction information to constructing the prediction fashions. The goal of the gadget is to complete all of the aforementioned procedures with none human intervention. The fundamental gadget structure of the Xuanwu gadget as designed through Zhang et al. The first step the gadget follows is that it makes use of a sliding window technique to reduce the historic transactional information of every inventory into a couple of Clips. Each clip has a duration that's identical to a predefined prediction period deemed appropriate for the funding in question. Then, consistent with the shapes that the near expenses of those Clips appear to be whilst visualized graphically, the gadget then makes use of an unmanaged heuristic set of rules to categorise them into 4 primary training: Up, Down, Flat, and Unknown. For the Clips that belong to the training Up and Down, they get similarly labelled as one-of-a-kind degrees which assist to resonate the extents in their increase and receding rates. This is accomplished with admire to each absolute near rate and relative go back rate. The education units are derived from those Clips through sampling one of a kind training of samples for unequal elegance dissemination. Finally, the getting to know fashions are skilled from those training sets irrespective of feature selection using the random forest and decision tree algorithms. The characteristic choice technique used is Forward Sequential Search technique.

3. Proposed work

3.1 DATASET

Two different data sets have been used here: one of which contains the tweets and the other contains the prices of the stock. The data in the 1st data set is the date of that particular day and the contents of the tweet at random on Twitter which was filtered using "Amazon" as the keyword from 2010 to 2020. The Twitter Scraper API of the PYPI library was used to get the dataset. The time period 2010 to 2020 was chosen because people used Twitter as a major social media platform starting in 2010. The second dataset is composed of Time, Open stock value, High stock value, Low stock value, Close stock value, and Volume of the stock traded from 2000 to 2020.

3.2 PRE-PROCESSING AND CLEANING OF DATASET

The original dataset consists of tweets posted by people as they are without any alteration or filtering of any form. Before applying sentiment analysis on them, they need to be "cleaned" and filtered of the unwanted words and characters in them keeping only what's essential for the sentiment analysis process. The first step is to remove all the pictographs, symbols, emoticons, transport and map symbols, and flags as they are unrecognizable by a sentiment analyser.

3.3 LSTM

Long Short Term Memory, abbreviated to LSTM, is a segment of Recursive Neural Network (RNN) engineering (an artificial neural network) put forth by Hochreiter and Schmidhuber in 1997. RNN is a generalization of a feed forward neural network that has a memory of its own. A RNN is named Recurrent as it performs a fixed function for all the inputs of data where the outcome of the current input is dependent on the previous calculation. Unlike the usual feed forward neural networks, RNNs are able to utilise their internal state or memory in order to compute sequences of inputs. In other neural networks, all the inputs are independent of each other. But in RNN, all the inputs are related to one another. The demerit of using an RNN is they will be afflicted by the vanishing and exploding gradients hassle wherein gradients are both reduced to 0 or incremented certain throughout back propagation via a big quantity of time steps.

LSTM is delivered mostly to triumph over these problems, especially the vanishing gradients. The LSTM community is appropriate to analyse based on experience to categorize, system and predict time collection whilst there are time lags of unknown length and bound among crucial incidents. LSTM is capable of managing long time dependencies with the aid of using the usage of a memory unit. It has a sequence like structure, having 3 neural community layers or gates that are applied to the usage of diverse mathematical functions. Along with the gates, there may be additionally a cell state which acts as a transport channel that transfers relative records all the way down the series chain. As the cell state is going on its journey, records get added, updated or eliminated to the cell state via gates.

4. IMPLEMENTATION

4.1 IMPLEMENTATION OF LSTM

In our study, the LSTM model was used two times, once without twitter sentiment analysis as a feature and once with twitter sentiment analysis as a feature. For the first result the normalised data is used as an input into the LSTM. Then sentiment analysis is performed on all tweets pertaining to the company Amazon and the polarity of people's moods is recorded. This polarity is then added as a feature to the second run of the LSTM RNN to compare the accuracy of the two. The initial input without the twitter sentiment analysis takes in 4 inputs and then takes 5 as the input for the second time with twitter sentiment analysis in the LSTM. The number of outputs however remains the same (Open, close, high and low values). This particular LSTM architecture uses 200 neurons and 4 layers out of which, two layers are input and output layers and 2 are hidden layers. The learning rate is set at 0.001 so that the algorithm doesn't miss a local minima during the gradient descent and the learning rate seemsto be the most suitable for a dataset of this magnitude. The gradient descent optimizer used is "ADAM optimiser". ADAM

is a powerful learning rate optimiser that utilises the potential of adaptive learning rate methods to identify separate learning rates of every parameter.

The LSTM uses Leaky ReLU as the activation function. The main advantage of using Leaky ReLU instead of ReLU is that in this way we avoid running into the problem of vanishing gradients.

The issue is that in certain situations, the gradient will be negligibly tiny, directly prohibiting the weight from altering its value.

Meaning that the weights won't change their values aftera certain iteration but Leaky RElu solves this problem. The MSE drops for every epoch as the above designed LSTM is trained.

4.2 TWITTER SENTIMENT ANALYSIS IMPLEMENTATION

Twitter sentiment analysis is the process of analysing the mood of either a set of tweets from a set of people or even tweets from every person tweeting about Amazon's stock which passes the filtering criteria. After the first run of the LSTM model on the Amazon dataset, the moods of the people are read and classified into positive, negative, or neutral. The Tweets are retrieved using Twitter scraper. In order to process the tweets to receive the polarity of the public's mood, the tweets need to be pre-processedi.e. all the unnecessaryelements of the tweets must be eliminated.

The pre-processing includes:

a) Removal of emojis and symbols.
b) Convert all the characters to lowercase
c) Removal of hyperlinks
d) Removal of numbers
e) Filling in missing values
f) Tokenization

5. Results and discussion

LSTM prediction models were implemented at the stock prices dataset of Amazon from 2010 to 2020, its shows in figure 1. LSTM follows a multivariate approach and is suitable for time series and sequential facts. The tweets published at the various date stamps in the stock dataset had been collected and filtered the use of the key-word Amazon. 20 tweets had been taken for every day whose polarities had been averaged to locate the cumulative polarity of tweets for one given day. Two different variants of the LSTM version had been used; one which was trained best at the inventory expenses of Amazon and the opposite version changed into educated on the dataset of inventory expenses included with the tweets. The intention is to look at the distinction of their overall performance in prediction of the identical data points. It may be concluded that even the LSTM version that did now no longer consist of the emotions has confirmed higher results as compared to the version that did consist of the emotions shows figure 2, it's far noteworthy that the latter attempts to bear in mind greater elements like people's opinion whilst predicting something as risky and unpredictable as inventory expenses. This paper targets contrasting the two. Future work in the region may be done with the aid of using the use of ensemble deep getting to know techniques that have a higher danger at capturing the intricacies of a tough sample like people's opinion and locating a different manner to cope with the lack of facts whilst managing social media systems in phrases of the 30, forty years of facts normally had to educate a LSTM community well, it's shows in figure 3. Tweets scraped may be filtered more specifically to look at the individual effects of numerous parameters at the prediction.

Figure 1: Dataset

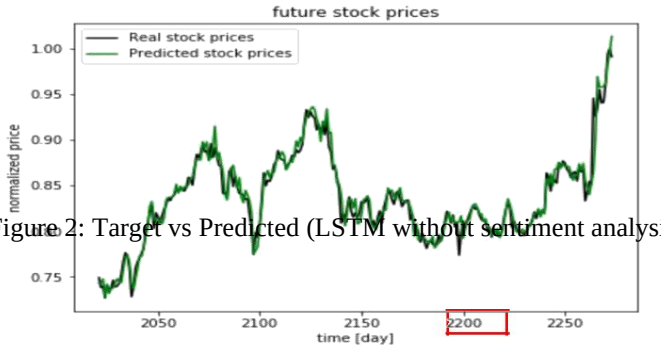

Figure 2: Target vs Predicted (LSTM without sentiment analysis)

Figure 3: Target vs Predicted (LSTM with sentiment analysis)

6. Conclusion

LSTM prediction module was applied to the stock prices dataset of Amazon from 2010 to 2020/. A multivariate approach is followed by LSTM and is suitable for time series and sequential data. The tweets posted on the various date stamps in the stock dataset were collected and filtered using the keyword Amazon. 20 tweets were taken for each day whose polarities were averaged to find the cumulative polarity of tweets for one given day. Two different variations of the LSTM model were used; one that was trained only on the stock prices of Amazon and the other model was trained on the dataset of stock prices integrated with the tweets. The aim is to see the difference in their performance in prediction of the same data points. It can be concluded that even the LSTM model that did not include the sentiments has shown better results compared to the model that did include the sentiments, it is noteworthy that the latter tries to consider more factors like people's opinion when predicting something as volatile and unpredictable as stock prices. This paper aims at contrasting the two. Further improvement can be achieved by making use of ensemble deep learning methods which are proven to be better at throwing light at the complexities of an intricate pattern such as people's opinion and finding a different way to deal with the dearth of data when dealing with social media platforms in terms of the 30, 40 years of data usually needed to train a LSTM

network well. Tweets scraped can be filtered more specifically to study the individual effects of various parameters on the prediction.

References

1. Bharathi, S. and Geetha, A. (2017). Analysis for effective stock market prediction." International Journal of Intelligent Engineering and Systems, 146–154.
2. Bollen, J. and Mao., H. (2018). "Twitter mood as a stock market predictor." IEEE Computer, 44(6), 91–94.
3. Joshi, K., N., B. H., and Rao, J. "Stock trend prediction using news sentiment analysis.
4. Kordonis, J., Symeonidis, S., and Arampatzis, A. "Stock price forecasting via sentiment analysis on twitter.
5. M., H., E.A, G., Menon, V. K., and K.P., S. (2018). "Stock price forecasting via sentiment analysis on twitter." Procedia Computer Science, 132(6), 1351–1362.
6. Mittal, A. and Goel, A. "Stock prediction using twitter sentiment analysis.
7. Moghaddama, A. H., Moghaddamb, M. H., and Esfandyari, M. (2018). "Stock market index prediction using artificial neural network." Journal of Economics, Finance and Administrative Science, 21(6), 89–93.
8. Nelson, D. M. Q., Pereira, A. C. M., and de Oliveira, R. A. (2017). "Stock market's price movement prediction with lstm neural networks." - IJCNN.
9. Obthong, M., Tantisantiwong, N., Jeamwatthanacha, W., and Wills, G. "A survey on machine learning for stock price prediction: Algorithms and techniques.
10. Pang, X., Zhou, Y., Wang, P., Lin, W., and Chang, V. (2018). "An innovative neural network approach for stock market prediction." Springer Nature, (6).
11. Qiu, J., Wang, B., and Zhou, C. (2020). "Forecasting stock prices with long-short term memory neural network based on attention mechanism.
12. Roondiwala, M., Patel, H., and VarmBharathi, S. and Geetha, A. (2017). "Sentiment a, S. (2017). "Predicting stock prices using lstm." International Journal of Science and Research (IJSR).
13. Selvamuthu, D., Kumar, V., and Mishra, A. (2019). "Indian stock market prediction using artificial neural networks on tick data." Financial Innovation, 5–16.
14. Zhang, Cui, S., Xu, Y., Li, Q., and Li, T. (2017). "A novel data-driven stock price trend prediction system." Elsevier.
15. Shanmuganathan, V., Kalaivani, L., Kadry, S., Suresh, A., Robinson, Y. H., Lim, S. (2021). EECCRN: Energy Enhancement with CSS Approach Using Q-Learning and Coalition Game Modelling in CRN. Information Technology and Control, 50(1), 171-187. https://doi.org/10.5755/j01.itc.50.1.27494
16. Kumar, S.A.P., Nair, R.R., Kannan, E., Suresh, A., S. Raj Anand, "Intelligent Vehicle Parking System (IVPS) Using Wireless Sensor Networks". Wireless Personal Communications (2021). https://link.springer.com/article/10.1007/s11277-021-08360-z
17. Sajid, M.R., Muhammad, N., Zakaria, R. Ahmad Shahbaz, Syed Ahmad, SeifedineKadry, A. Suresh., "Nonclinical Features in Predictive Modeling of Cardiovascular Diseases: A Machine Learning Approach". Interdisciplinary Sciences: Computational Life Sciences (2021). https://doi.org/10.1007/s12539-021-00423-w
18. Kumar, S., Cengiz, K., Vimal, S. Suresh, A., "Energy Efficient Resource Migration Based Load Balance Mechanism for High Traffic Applications IoT". Wireless Personal Communications (2021). https://doi.org/10.1007/s11277-021-08269-7.
19. Arulananth, T.S., Balaji, L., Baskar, M. *et al.* PCA Based Dimensional Data Reduction and Segmentation for DICOM Images. *Neural Process Lett* (2020). November 2020. https://doi.org/10.1007/s11063-020-10391-9.
20. M .Baskar, J. Ramkumar,Ritik Rathore, Raghav Kabra, "A Deep Learning Based Approach for Automatic Detection of Bike Riders with No Helmet and Number Plate Recognition", International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1844-1854, ISSN: 2005-4238, April 2020.