News Impact on Stock Trend

Protim Dey, Nadia Nahar, B M Mainul Hossain

Institute of Information Technology, University of Dhaka, Dhaka 1000, Bangladesh

Abstract

Stock market trend can be predicted with the help of machine learning techniques. However, the stock market changes is uncertain. So it is very difficult and challenging to forecast stock price trend. The main goal of this paper is to implement a model for stock value trend prediction using share market news by machine learning techniques. Although this kind of work is implemented for the stock markets of various developed countries, it is not so common to observe such kind of analysis for the stock markets of underdeveloped countries. The model for this work is built on published stock data obtained from DSE (Dhaka Stock Exchange, Bangladesh), a representative stock market of an underdeveloped country. The empirical result reveals the effectiveness of Convolutional Neural Networks with LSTM model.

1. Introduction

Stock market is the important part of economy of the country and plays a vital role for the industrial and commercial instruments. Both investors and industries are involved in stock market and want to know the rise and fall of stocks over certain period of time. The stock market is one of the primary sources for a company to raise funds for business expansions. It is based on the concept of demand and supply. If the demand for a company's stock is higher, then the company share price increases and if the demand for company's stock is low then the company share price decreases. Therefore, it is significantly useful when the stock market

* Corresponding author.
E-mail address: email2ppd@gmail.com

trend for the day itself or the next day can be predicted. As there are no certain rules to estimate or predict the price of a stock, this work aims to find a way to predict the stock price trend.

Stock market have huge fluctuation. Investing in a good stock but at a bad time can make big economical loss, while investing in a stock at the right time can bear profits. Nowadays, financial investors do not properly understand which stocks to buy or which stocks to sell in order to make profit. This paper will reduce the problem, with suitable accuracy, faced in such real time scenario.

The aims of this work, named News Impact on Stock Trend, are as follows:
1) To identify the influencing factors of share market.
2) To identify the correlation between share market news and stock price movement from large set of data.
3) To predict the trend of share price.

The main goal of this paper is to generate an approximate forecasting output of each stock based on the previous historical data.

The remainder of this article is arranged as follows: beginning with this introduction and in the next section, related works are explained. The Materials and Methods are explained in Section 3 and Section 4 describes the experimental results of the proposed system. The conclusions are then described in the last section.

2. Background and Related Work

Curriculum In the last few decades, forecasting of stock returns has become an important field of research. Using Twitter messages as input to predict the stock price is a very popular approach [1-2]. Like this paper, some of the research done in this field focuses on the idea of combining stock prices with news headlines [3-5].

Kalyani et al. (2016) collected three years of Apple Inc. stock data. Additionally, they also collected news data of this company over the same time span. The news data has been collected from several news websites, such as Google news1 and yahoo finance2. First, they used a sentiment detection algorithm to detect the sentiment of news articles. The sentiment detection algorithm transferred the news article into a count of positive and negative words. These results were then used in various classification algorithms, such as SVM, Random Forest and Naive Bayes. The accuracy of these classifiers are achieved between 75% and 90% on an unseen dataset.

Kirange and Deshmukh (2016) used the same steps as the above research mentioned, but they used 10 years historical data of Indian companies. They also invoke sentiments of the news article as input parameters to various classification algorithms, such as SVM, KNN and Naive Bayes. The outcome results of this research was the accuracy between 47% and 75%. The problem with these papers is that they have used the sentiment of news article as input parameters to their proposed approaches where the system consider only the current state of news sentiment, but did not consider the previous news sentiment.

Another research by, Wilson and Sharda [6] studied a research using neural networks and classical multiple discriminant analysis, where neural networks performed significantly better.

Min and Lee [7] were doing prediction of bankruptcy using machine learning techniques such as Support Vector Machine, multiple discriminant analysis, logistic regression analysis, and three-layer fully connected back-propagation neural networks. The experimental results expressed that support vector machines outperformed than other. They used various financial influencing factors such as interest coverage ratio, ordinary income to total assets, Net income to stakeholders' equity, current liabilities ratio, etc. as input parameters and output results gained the accuracy of around 60%. By using neural networks for forecasting credit rating of the companies and achieving accuracy between 75% and 80% for the Taiwan and United States markets.

Atsalakis, G. S., Dimitrakakis, E. M. & Zopounidis [8] proposed WASP (Wave Analysis Stock Prediction) system based on the neurofuzzy architecture, which utilizes the Elliott Wave Theory. The proposed technique has been used to forecast the trend of the stock prices and the system performed significantly better.

Tsai and Wang [9] proposed a model where they used ensemble learning, composed of decision trees and artificial neural networks. They used Taiwanese stock market historical data for forecasting stock price. The evaluation of Decision Tree+Artificial Neural Network showed F-score 77%, but single algorithm showed up to 67%.

Kim and Han [10] proposed a genetic quantum algorithm which transform continuous input values into discrete ones. This model reduce the complexity of the feature selection.

Several variations of Neural Networks have been tried by the researchers to achieve improved results on forecasting [11-20]. New forecasting model on the basis of fuzzy systems have also been proposed [21-23]. Liang, Zhang et al. [24] proposed a model for predicting option prices using neural networks (NNs) and support vector regressions (SVRs), based on a two stage nonparametric method. In first stage they modified the improved conventional option pricing methods, to forecast the option prices. In second stage they employed the NNs and SVRs to increase the predicting accuracy. They used Hong Kong option market data. The experimental results demonstrated the ability of NNs and SVRs to improve forecast accuracy.

Kara, Boyacioglu et al. [25] proposed two efficient models based on artificial neural networks (ANN) and support vector machines (SVM) for predicting the direction of movement stock price. They used ten technical indicators as inputs parameters for predicting daily stock price of Istanbul Stock Exchange (ISE) National 100 Index. The outcomes demonstrated that the average performance of ANN model (75.74%) performs significantly better than the SVM model (71.52%).

Yavuz, Mehmet & Sakarya, ¸Sakir & Ozdemir, Necati & Karaoglan, Aslan. (2015) [26] proposed an ANN model for predicting daily and weekly returns of Borsa Istanbul (BIST)-100 Index during global ˙ crisis period (July 2007-December 2009). The results almost correctly gave the value for next day and next week with an accuracy margin error of less than 5% for unknown samples.

3. Materials and Methods

In this study, the proposed system can predict the stock's price movement of DSE for today, tomorrow and day after tomorrow by using Convolutional Neural Network (CNN) [29] with LSTM. During the training process, the network will learn to ignore any inputs that don't contribute to the output. In training phase, the input parameter is news headlines of stock and forecast the stock trend.

LSTM is a type of recurrent neural network that preserves its state. Generally the news headlines are text based and LSTM networks have already been proven successfully with text data for applications as NLP [27] and machine translation [28]. The architecture of the proposed system is shown in **Fig. 1**.

This model achieve a more precise approach compared to previous approaches. In the following, the detailed description of the different sections of proposed system is expressed.

A.   *Data Grabing*

The data used in this research work were historical daily stock prices and news. For this purpose, a data grabbing system is developed, which runs on every 5 minutes interval that scraps new[3] and stock price data [4] from Dhaka Stock Exchange (DSE), Bangladesh web site   and stored in SQL server database. The stock data consists of open price, low price, high price, close price, and volume traded. The open price is the opening price of the stock (PoS) at the start of the trading day, the low price represents the minimum PoS during the trading day, the high price represents the maximum PoS during the trading day, and the closing price indicates the PoS when the market closes.
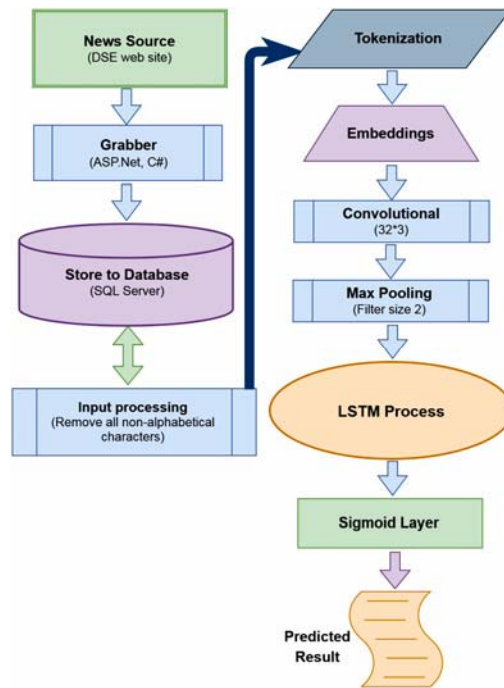
Fig. 1. Overall System Architecture

B. Input Processing

As shown in **Fig. 1** the headlines are used as input for the neural network. These headlines are preprocessed and organized as input parameter. Before invoke as input all preprocessed headlines are converted to lowercase characters to overcome double identical words with a different capitalization. For example, considering the news headline "- (Q1):   As per un-audited...". After processing the output will be "q1: as per un-audited".

Tokenization is performed using **Keras Tokenizer Method** on processed news headlines data sets. This tokenization process works in two steps. First it updates internal vocabulary index based on word frequency. If providing something like, "Stock price is very difficult to predict" It will create a dictionary e.g. word_index ["Stock"] = 0; word_index["price"] = 1. It is word to index dictionary, therefore every word gets a unique integer value where lower integer means more frequent word. Second it transforms each word to a sequence of integers. So it basically takes each word in the text and transform a corresponding integer value from the word_index dictionary.

C. Embedding layer

The embedding layer of above architecture **Fig. 1**, which transfer the words by numeric indexes. Actually "Word embeddings" are a family of natural language processing techniques aiming at mapping semantic meaning into a geometric space. So, this layer vectorizes the news headlines (the "word2vec" technique). Let us consider a simple example of a training set consists of only two phrase:

"Stock price is volatile"
"Stock changes is uncertain"
After Tokenization our phrases could be rewritten as:
[0, 1, 2, 3]
[0, 4, 2, 5]

Once the model has been trained, the embedding layer with 168 top words size, gives a (6, 2) embedding vectors:

Table 1. Embedding vectors mapping

Index	Embedding
0	[2.1,2.0]
1	[0.1,3.0]
2	[1.0,3.1]
3	[0.3,2.1]
4	[0.7,1.7]
5	[0.5,2.3]]

According to these embeddings, our two training phrase will be represented as:
[[2.1,2.0], [0.1,3.0], [1.0,3.1], [0.3,2.1], [0.7,1.7], [0.5,2.3]]
So, embedding vectors gives the corresponding dimension for each word. Similar relational words are clustered together and they positioned closet each other, as Fig. 2. When a token (i.e. word) is given as input to the embedding layer, it returns the vector of that corresponding token.

In this work, for the word embeddings, top 2000 words in the corpus are used to convert the words to the proper index i.e. input_dim = 2000. All other words that were not present in the top 2000 words are set to an index of zero. Embedding vector length is used 32. The maximal size of the each input sequence is 100 and zero padded for smaller headlines.
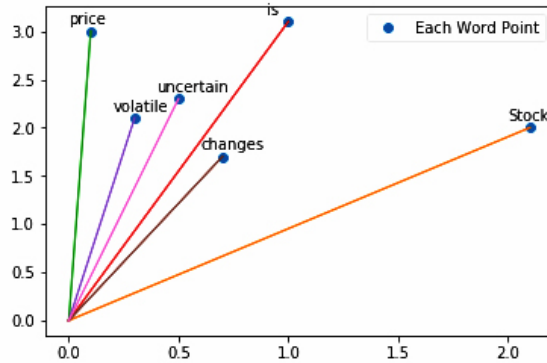

Fig. 2. Scatter chart for two training phrases

D.   Convolutional(CNN) and Max Pooling Layer

The output of embeddings layer are fed to a 1D convolutional layer with a kernel size of 3 and 32 filters. Adding a convolutional layer before a LSTM layer excel the mapping spatial structure of input data. The size of kernel has been chosen 3, because the surrounding less weighted words are filter out. So, a convolution

layer worked here is like a filter of a certain dimension. The final output of this computation forms a single integer element.

Next layer added in this proposed system is pooling or down sampling layer, which executes some mathematical operation over regions/patches. In this work, max-pooling is used.

Two of the most common pooling operations are:
1) Max- pooling.
2) Average-pooling.

Max-pooling [30] selects the maximum of the values in the input feature map region of each step and average-pooling the average value of the values in the region. The output in each step become a single scalar. Here max pooling layer is added after the convolutional layer to take the maximum weighted word.

E. LSTM Layer

The output of the max-pooling layer is used as input for the LSTM layer. LSTM is capable of learning long-term dependencies. Instead of neurons, LSTM networks have memory blocks that are connected through layers. It performs smarter than a classical neuron and a memory for recent sequences. LSTMs also have chain-like neural network layer.
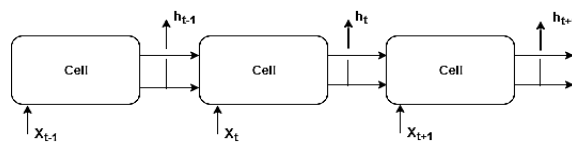

Fig. 3. Structure of an LSTM network.

Here a traditional LSTM cell with the number of units set to 100, a hyperbolic tangent activation function and a hard sigmoid recurrent activation function is used. A GRU Cell [31] has also been used, but this did improve the network, but only worsened.

F. Sigmoid layer

A Dense with sigmoidal activation function has been used for this layer, because the final outputs will be a vector with a size of 1.

In this model a binary cross entropy loss function also used, since it is a binary classification problem. The Adam optimization function [32] is used in this model. The learning rate at the beginning is set to 0.001 with a learning rate decay of 0.1 when the validation loss did not improve any more for 5 epochs, then terminate from the epochs & model is stored. Accuracy of the model is also stored to CSV files for per stock.

So, the major steps involved in this network are as follows:
1) *Feed the preprocessed input data sample, compute the corresponding output.*
2) *The word embeddings are fed to a 1D convolutional layer.*
3) *A max pooling layer is added after the convolutional layer to make the filters location invariant.*
4) *The output filtered vector of the convolutional layer are used as input for the LSTM layer.*
5) *A sigmoidal activation function has been used for the output layer.*
6) *Using a binary cross entropy loss function for achieving best output.*
7) *Compute the error between the output(s) and the actual target(s).*
8) *IF accuracy_score > pre accuracy_score THEN go to Step 2 ELSE stop.*

4. Result Analysis

Different datasets has been used to train different models on. Each day stock data of each company are combined with a per day news headline were composed. For training, validation and test set, all datasets were split into 80-10-10 respectively. The implementation of the model has been done in Keras using a Tensorflow backend. This approach has been chosen because it can consider not only the current state, but also previous historical state. The dataset description and the experimental results are explained in this section.

A. Data Set

In this research, the closing price is chosen to represent the PoS to be modeled and predicted. This is because the closing price reflects all the activities of the stock of the day. The News data consists of PublishDate, Scrip, Title, Content, Source, NewsUrl. PublishDate is the release date of news. Scrip is the stock code of respective companies. Title, Content, Source and NewsUrl represents respectively, the news headline, full content of the news, scraping from which news source and news online url. In this research the Title is chosen for predicting the stock price trend. After gathering raw data, Stock and news data were combined together and stored in CSV files. During this step, News title also processed as described in Materials and Methods section. Thereafter, these processed news headlines dataset was further extended by including values which indicated the change of the company's stock at the day the news headline came out (PublishDate), one day later and two days later which are indicated by "today", "tomorrow" and "day_after_tomorrow". These three fields were populated and represented by binary values (0 or 1). When close price is greater than open price, then 1 otherwise 0 (close price is less than equal to open price). After that, all combined data were loaded from csv to import as pandas data frame and feed to this proposed model as input. So, the combined stock data consists of PublishDate, Scrip (Stock Name), Title, today, tomorrow, day_after_tomorrow, normalized_headline.

B. Predicted Results

After successfully completing the training phase, the proposed system can predict price trend for today, tomorrow and the day after tomorrow.

The accuracy results are given in Table 2. All the results are reported on the held out test set.

Table 2. Prediction Result

Company	Full Name	Today	Tomorrow	D.A. Tomorrow
ABBANK	AB Bank Limited	75.00%	91.67%	91.67%
SQURPHARMA	Square Pharmaceuticals Ltd.	70.1%	77.1%	78.1%
PRAGATILIF	Pragati Life Insurance Ltd.	68.6%	71.7%	73.8%
UNITEDAIR	United Airways (BD) Ltd.	76.92%	84.62%	69.23%
SOUTHEASTB	Southeast Bank Ltd.	41.67%	83.33%	87.50%
SPCL	Shahjibazar Power Co. Ltd.	92.86%	85.71%	85.71%

As it is seen that the proposed model performs better when the time progresses (Table 2) i.e. this model can predict the stock's price trend for tomorrow and the day after tomorrow overall better than today (except SPCL in the above table 2). For example, today, tomorrow and the day after tomorrow accuracy of "ABBANK" are respectively, 75%, 91.67% and 91.67%, where tomorrow and the day after tomorrow accuracy were greater than today. When test is run with specific model with specific day's data from combined data set, it gives output as decimal values. When the output less than 40, stock price will go down, for 40-70 price will remain same and if output greater than 70 price will be increased as it is shown in Table 3:

Table 3. Price Trend Ranges

Ranges	Price Trend
< 40	Negative
40 - 70	Neutral
70 <	Positive

5. Conclusion

The main goal of this research is the detailed investigation of the co-relation between stock price trend & the financial news. By incorporating only limited number of parameters, there is certain degree of accuracy. Since, there are many influencing parameters that directly affect stock market; each and every one of them cannot be taken into account. In this work, each day price is concerned with only one news. Dhaka Stock Exchange (DSE), Bangladesh web site is used only as source. As a part of the future work, potential improvement can be made to our data collection and analysis method. Multiple news collecting from multiple source and that can be used with each day stock price, i.e. multiple news for each day. The model could be trained on more companies instead of one specific company. A great improvement can be made by including technical indicators as input parameters. Since technical indicators directly influence the stock price & following it. Also the sentiment of a news headline can be included for improvement. This can help further improvement to the proposed system so that it can generates more accurate results.

References

[1] Bollen, J.; Mao, H.; Zeng, X.J. Twitter mood predicts the stock market. arXiv e-prints 2010, p. arXiv:1010.3003, [arXiv:cs.CE/1010.3003].

[2] Zhang, X.; Fuehres, H.; Gloor, P. Predicting Stock Market Indicators Through Twitter – "I Hope it is Not as Bad as I Fear. Procedia - Social and Behavioral Sciences 2011, 26, 55–62. doi:10.1016/j.sbspro.2011.10.562.

[3] Kalyani, J.; Bharathi, H.N.; Jyothi, R. Stock trend prediction using news sentiment analysis. ArXiv 2016, abs/1607.01958.

[4] Kirange, M.D.K.; Deshmukh, D.R.R. Sentiment Analysis of News Headlines for Stock Price Prediction. An International Journal of Advanced Computer Technology 2016, 5.

[5] Wong, F.M.F.; Liu, Z.; Chiang, M. Stock Market Prediction from WSJ: Text Mining via Sparse Matrix Factorization. CoRR 2014, abs/1406.7330, [1406.7330].

[6] Wilson, R.L.; Sharda, R. Bankruptcy prediction using neural networks. Decision Support Systems 1994, 11, 545 – 557. doi:https://doi.org/10.1016/0167-9236 (94)90024-8.

[7] H. Min, J.; Lee, Y.C. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications 2005, 28, 603–614. doi:10.1016/j.eswa.2004.12.008.

[8]   George S. Atsalakis, Emmanouil M. Dimitrakakis, C.D.Z. Elliott Wave Theory and neuro-fuzzy systems, in stock market prediction:The WASP system. Expert Systems with Applications 2011, 38.

[9]   Tsai, K.H.; Wang, J.C. External Technology Sourcing and Innovation Performance in LMT Sectors: An Analysis Based on the Taiwanese Technological Innovation Survey. Research Policy 2009, 38, 518–526. doi:10.1016/j.respol.2008.10.007.

[10] Han, K.H.; Kim, J.H. Genetic Quantum Algorithm and its Application to Combinatorial Optimization Problem 2003. 1354-1360.

[11] Kimoto, T.; Asakawa, K.; Yoda, M.; Takeoka, M. Stock market prediction system with modular neural networks. IJCNN, 1990.

[12] Kamijo, K.; Tanigawa, T. Stock price pattern recognition-a recurrent neural network approach. IJCNN, 1990, p. 215–221.

[13] Ayodele Ariyo Adebiyi, A.O.A.; Ayo, C.K. Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. Journal of Applied Mathematics 2014, 2014. doi:https://doi.org/10.1155/2014/614342.

[14] Guresen, E.; Kayakutlu, G.; Daim, T.U. Using Artificial Neural Network Models in Stock Market Index Prediction. Expert Syst. Appl. 2011, 38, 10389–10397. doi:10.1016/j.eswa.2011.02.068.

[15] Chen, A.S.; Leung, M.T.; Daouk, H. Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. Computers & OR 2003, 30, 901–923.

[16] Kuo, R.; Chen, C.; Hwang, Y. An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. Fuzzy Sets and Systems 2001, 118, 21 – 45. doi:https://doi.org/10.1016/S0165-0114 (98)00399-6.

[17] Roondiwala, M.; Patel, H.; Varma, S. Predicting Stock Prices Using LSTM. International Journal of Science and Research (IJSR) 2017, 6. doi:10.21275/ART20172755.

[18] Mahanta, R.; Pandey, T.N.; Jagadev, A.K.; Dehuri, S. Optimized Radial Basis Functional neural network for stock index prediction. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) 2016, pp. 1252–1257.

[19] Dhiraj Mundada, Gaurav Chhaparwal, S.C.T.B. Stock Value Prediction System. International Journal on Recent and Innovation Trends in Computing and Communication 2015, 3, 2217 – 2219.

[20] Saad, E.W.; Prokhorov, D.V.; Wunsch, D.C. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. IEEE Transactions on Neural Networks 1998, 9, 1456–1470. doi:10.1109/72.728395.

[21] Qiu, W.; Liu, X.; Wang, L. Forecasting shanghai composite index based on fuzzy time series and improved C-fuzzy decision trees. Expert Syst. Appl. 2012, 39, 7680–7689.

[22] Atsalakis, G.; Valavanis, K. Forecasting stock market short-term trends using a neuro-fuzzy based methodology. Expert Systems with Applications 2009, 36, 10696–10707. doi:10.1016/j.eswa.2009.02.043.

[23] Wong, F.; Wang, P.; Goh, T.; Quek, B. Fuzzy Neural Systems for Stock Selection. Financial Analysts Journal 1992, 48, 47–52, [https://doi.org/10.2469/faj.v48.n1.47]. doi:10.2469/faj.v48.n1.47. Version July 16, 2019 submitted to Data 10 of 10

[24] Liang, X.; Zhang, H.; Xiao, J.; Chen, Y. Improving Option Price Forecasts with Neural Networks and Support Vector Regressions. Neurocomput. 2009, 72, 3055–3065. doi:10.1016/j.neucom.2009.03.015.

[25] Kara, Y.; Boyacioglu, M.; Baykan, O. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert Systems with Applications 2011, 38, 5311–5319. doi:10.1016/j.eswa.2010.10.027.

[26] SAKARYA, S.; YAVUZ, M.; KARAOGLAN, A.D.; ÖZDEMIR, N. Stock Market Index Prediction with Neural Network during Financial Crises: A Review on Bist-100. Financial Risk and Management Reviews 2015, 1, 53–67. doi:10.18488/journal.89/2015.

[27] Sundermeyer, M.; Schlüter, R.; Ney, H. LSTM Neural Networks for Language Modeling. 2012.

[28] Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase

Representations using RNN Encoder-Decoder for Statistical Machine Translation. CoRR 2014, abs/1406.1078, [1406.1078].

[29] Patterson, J.; Gibson, A. Deep Learning: A Practitioner's Approach; O'Reilly: Beijing, 2017.

[30] Long. Convolutional Neural Network.https://medium.com /@Aj.Cheng/convolutional-neural-network-d9f69e473feb. [Online], [Accessed 2019-06-25].

[31] Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Gated Feedback Recurrent Neural Networks. Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37. JMLR.org, 2015, ICML'15, pp. 2067–2075.

[32] Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Authors' Profiles

Partho Protim Dey received Master in Information Technology (MIT) from Institute of Information Technology (IIT), University of Dhaka, Bangladesh. Before that, he received B. Sc. Honors in Computer Science from National University Bangladesh and Currently he working as a Sr. System Analysts in a financial software company. Here, he works on various innovative project & big data analysis. His activities currently focus on various machine learning techniques & applied those in real life.

Nadia Nahar is a Lecturer at the Institute of Information Technology (IIT), University of Dhaka, Bangladesh. She pursued her Master of Science in Software Engineering (MSSE) and Bachelor of Science in Software Engineering (BSSE) from the same institution. She was the gold medalist for attaining top score in her class. As a student, her efforts have earned awards from different national and international software and programming competitions, project showcasing as well as publications in various international conferences.
She has the experiences of working both in industry and academia. Her core areas of interest are software engineering, web technologies, systems and security. She is an active researcher at Distributed Systems and Software Engineering (DSSE) group having supervision and research experiences in these areas.

Dr. B. M. Mainul Hossain is Associate Professor at the Institute of Information Technology (IIT), University of Dhaka, Bangladesh. He received his Ph.D. degree in computer science from University of Illinois at Chicago, USA. Before that, he earned his Bachelor of Science and Masters degrees from the department of Computer Science & Engineering, University of Dhaka, Bangladesh. He has the experiences of working both in industry and academia. He worked as a Software Engineer at Microsoft Corporation (Redmond, USA) & Accenture Technology Lab (Chicago & California). His core areas of interest are Machine learning, Cyber Security, Software Testing.