An LSTM neural network model for stock market data

Dragana Radojičić*
* MI SANU, Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia

draganar@mi.sanu.ac.rs

Abstract— **For each trading day, there is a huge number of trade events registered at the stock market, and thus a large volume of data is recorded. In this research, machine learning approaches are particularly employed in order to learn from stock market data. More precisely, in order to capture the time dependency between rows in the market data, the model based on the Long short-term memory (LSTM) network is employed.**

I. Introduction

Considering complex limit order book dynamics involving a number of trade events, both on the bid and ask side of the order book, it is a challenging task to model order book dynamics. In order to apply knowledge from the order book data, it is necessary to devise a sustainable plan for its use to enable work with a large amount of data. There are different types of neural networks, and a very important point is to choose the suitable artificial neural network (ANN) for the particular task. A recurrent neural network (RNN) belongs to a class of the neural network that remembers previous outputs, i.e. RNN remembers the past. Therefore, RNNs are conducive for sequential data, e.g. financial time series. The RNN has the ability to use previous outputs as inputs, which is crucial when the sequential data contains important information in the recent past about the present and the future. The Long Short-Term Memory (LSTM) is a variant of the RNN, which is developed to deal with the vanishing gradient problem. In this paper, LSTM neural network model that processes order book data is introduced. As an input for the LSTM network model, features extracted from the order book data are used, and to measure the performance of the model F1 score is utilized.

A. Motivation

Applying machine learning techniques to get insights about market behavior is attractive for researchers both from academia and from the industry. In [2] authors introduced a stock market model based on a long short-term memory network which is an associated deep recurrent neural network model with multiple inputs and multiple outputs. Two stock market models based on the long short-term memory neural network (LSTM), one with embedded layer and one with automatic encoder, are introduced and studied in [3]. Further, the comparison of those two models is implemented, and the study shows that the deep LSTM model with embedded layers performs better. In [4] authors studied stock market models based on the Gated Recurrent Unit (GRU) neural network topology, and furthermore, the performances of the model fed with different groups of features were cross compared. The precision of the algorithm of a LSTM model for stock prediction is studied in [5]. Especially, in [5] the authors studied how increasing the number of epochs improves the LSTM model performance. In [6] authors introduce a forecasting mechanism which using a leverage of a series of machine learning techniques, predicts stock market crash events. The goal of the strategy in this research is to assign a label (i.e. the value that shall be predicted for a given input) to each vector of the stock market limit order book data.

B. The Limit Order Book

The limit order book (LOB) is a trading database of limit orders. The main purpose of the LOB is to record all incoming and outgoing both sell and buy limit orders. Figure 1 depicts a snapshot of the MSFT (Microsoft Corporation stock) ticker order book. For the detailed explanation and notation of the mathematical concept of the limit order book object see Section II in [1]. The highest price on the bid side is called the best bid price, while the lowest price on the ask side is called the best ask price. For each price level there is a quantity called the volume, which represents the number of orders sitting at that price level. The mid-price is defined as the arithmetic average between best ask and best bid price, and the mid-price is very often used as a proxy of the real price of a considered asset. The tick is the minimum distance between two price levels, while the quote spread is the distance between best ask price and best bid price. Note that the bid-ask spread is very often equal to one tick, see [7]. The study of the limit order book dynamic is interesting for researchers from academia, but also for the researcher from the financial world, and for people working in hedge funds, investment banks. A new area of Artificial Intelligence has expanded opportunities for developing trading strategies. Many trades are occurring with fast trading speed, and in order to capture the stock market behavior, computer algorithms and machine learning approaches can be employed. Machine learning can be useful in modeling phenomena when the solution changes over time.

C. The database

LOBSTER[1](Limit Order Book System-Efficient Reconstructor) is a high-quality online limit order data that replicates the NASDAQ (National Association of Securities Dealers Automated Quotation) order book. For each company and for each captured timestamp during the trading day, LOBSTER contains information about the limit order book shape (prices and corresponding volumes up to the requested number of levels). Furthermore, for each event LOBSTER contains

[1] https://lobsterdata.com. Accessed: 2020-08-05

information of the type of the event (submission, execution, etc), time of the event (measured with milliseconds and up to nanoseconds precision), order ID, etc. Since there are a huge amount of events per day, it is hard to keep track of all information from the order book. Thus, it is necessary to perform a data transformation during which the relevant attributes from the order book are segregated.

The efficient LOB reconstructor for a particular LOBSTER database is presented in detail in Section 2 in [4] or [8]. With the data reconstructor various order book attributes are abstracted, such as number of orders awaiting execution at the level 3 on the ask side, the average number of executed trades during the 60 seconds time interval, the average number of orders sitting on the best ask price during the 60 seconds time interval, the average mid-price, etc. This wide set of derived attributes is further expanded with technical indicators such as bollinger bands, moving average, etc. The open-source library *ta-lib*[2] contains algorithmic implementation for various technical indicators, and it has an open API for various programming languages (C/C++, Java, Perl, Python and 100% Managed). After having all features extracted, the next step is to label data for training. This is done by using the RISK-REWARD algorithm which is presented in the Appendix in [4].

II. The model

In general, there exists a sweet spot between the topology and the Machine Learning task. There are ways to automatically tune that (e.g. meta learning via potentially some genetic algorithms), but for the amount of data that has been processed for this research, it becomes hardly tractable. In general, when we put too shallow topology, we have too few weights to be trained and it doesn't capture the full variance. In general, when we put too shallow topology, we have too few weights to be trained and it doesn't capture the full variance. On the other hand, the noise or random fluctuations in the training data then the model overfits and fails to generalize. Thus, starting from something simple we slowly add up on the complexity of the network.

A. A neural network

An algorithm of an artificial neural network (ANN), very often called a neural network (NN), is inspired by the biological neural networks that are present in animal and human brains. NN is employed when the relationships between inputs and outputs is complex. The idea that stands behind artificial neural networks is that artificial intelligence can be exhibited by machines, in the sense that humans can give instructions to the machines and machines follow instructions. The job of the neural networks is to learn and detect the relationship between inputs and output during the training process. An Artificial Neuron consists of a collection of neurons, which are interconnected and perform calculations. Each artificial neuron has one or many input neurons, a computation unit that consists of a linear function and an activation function, and an output neuron (see Figure 2). To each interconnection between neurons, an associated Weight is assigned. With respect to the weight of the neuron, that neuron will more or less contribute to the overall picture of the output. In particular, higher weight means that the neuron input is more important, and smaller weight means that it is less important. An artificial neural network that has multiple hidden layers is called a deep neural network (DNN), see [9].

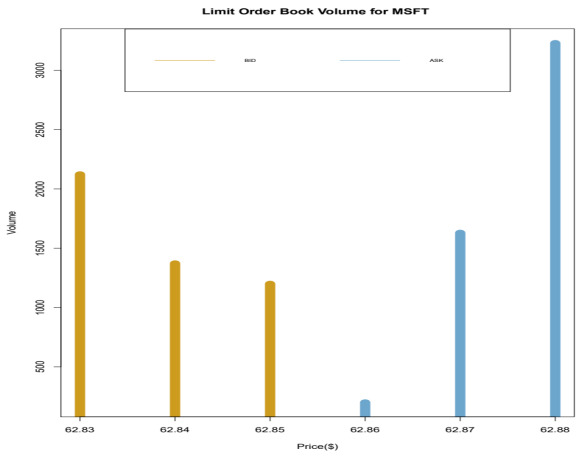

Figure 1. Snapshot of the NASDAQ limit order book for MSFT stock symbol for 3 levels. On the bid/ask side are placed outstanding buy/sell orders (gold/blue), and the best bid price is $62.85 with volume of 1300 shares, while the best ask price is $62.86 with volume of 200 shares.

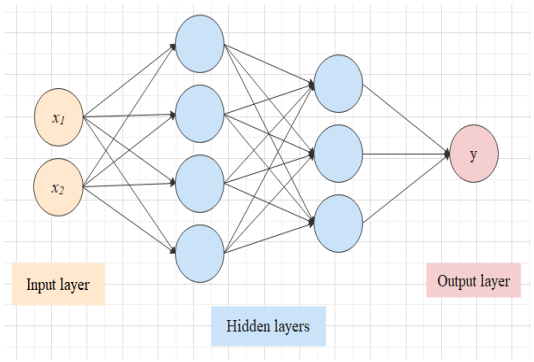

Figure 2. A simple artificial neural network with two hidden layers. This Figure is from the dissertation entitled "Stochastic modeling and statistical properties of the Limit Order Book", Dragana Radojicic

B. The Recurrent neural networks (RNNs)

The recurrent neural network (RNN) is a type of neural network that remembers the past. RNNs have the ability to take previous outputs as an input, and therefore they are very useful when the past has an influence on future decisions, e.g. when working with sequential data such as time series. Figure 3 depicts a simple RNN. The idea that stands behind RNN topology is firstly presented in [10].

Nowadays, RNNs have many modifications and a wide range of applications. When working with Vanilla RNN it

[2]https://ta-lib.org/. Accessed: 2020-08-05

is hard to capture long-term dependencies since there appears a vanishing gradient problem.

C. The Long short-term memory (LSTM) network

The LSTM is an artificial recurrent neural network (RNN) which is good at capturing the long-term dependency in sequence time data. It is a modification of traditional recurrent neural networks and it was initially introduced in [11]. The LSTM cell has the ability to choose which information to remember and which information to forget, and each LSTM cell contains a forget gate, input gate and output gate. One of the possible representations of a LSTM unit is depicted in Figure 4.

D. The Long short-term memory (LSTM) network model

In this research, the model based on the LSTM topology is employed. Precisely, the considered model consists of the following layers: the LSTM layer of 50 units, the flatten layer, and the dense layer. The LSTM processes the market data, and each market vector has a dimension of 100, i.e. each market data vector contains 100 extracted features. Further, the interim information acquired from are acquired "lookback" vectors. The flatten layer is employed to flatten the output from the LSTM layer so the dense layer can process data. Finally, the outcome of the dense layer is the probability that the market data vector should be labeled with 1 or with 0.

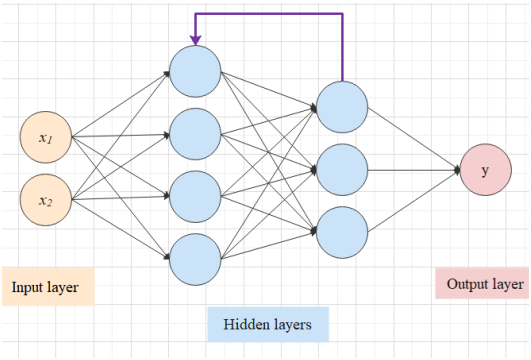

Figure 3. A recurrent neural network with two hidden layers. This Figure is from the dissertation entitled "Stochastic modeling and statistical properties of the Limit Order Book", Dragana Radojicic

III. The Set of Features

After choosing the neural network topology for this task, the focus is directed to features selection task. Some feature selection approaches are presented in [4] and in [12]. The aforementioned approaches are based on the assumption that features that have higher autocorrelation values or higher mutual information with the close price feature, are more powerful than the randomly selected features when it comes to labeling market data vectors.

IV. Conclusion

The developed model is tested with the real market data LOBSTER, that replicates the entire NASDAQ market stock exchange. Although stock market data is noisy and it is quite challenging to predict market behavior, neural networks have promising power to capture the dynamics of the market data. The future work can go into direction to cross compare the performances of the models based on different neural networks. Motivated by the multicriteria approach presented in [12], the PROMETHEE method can be employed to perform a comparison of different models.

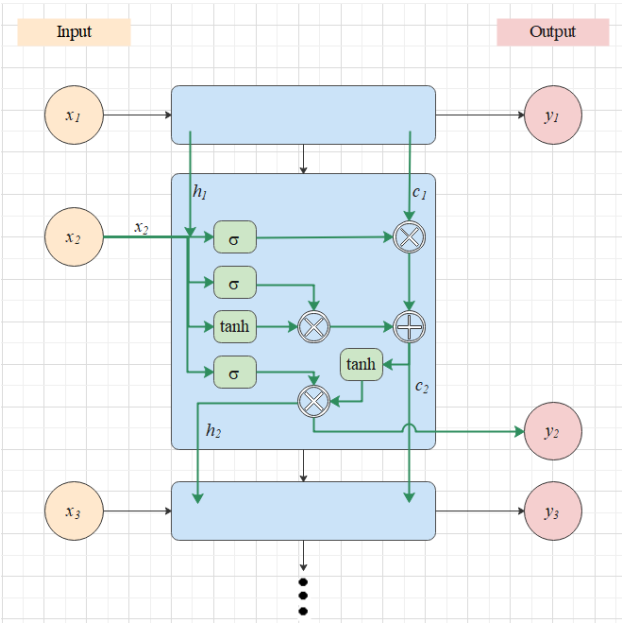

Figure 4. The representation of a LSTM unit.. This Figure is from the dissertation entitled "Stochastic modeling and statistical properties of the Limit Order Book", Dragana Radojicic

Acknowledgment

The author is thankful to Thorsten Rheinlander for his help in doing the LOBSTER data analysis and valuable and insightful suggestions.

References

[1] Gould, Martin D., Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. "Limit order books."*Quantitative Finance*13, no. 11 (2013): 1709-1742.

[2] Ding G, Qin L. Study on the prediction of stock price based on the associated network model of LSTM. International Journal of Machine Learning and Cybernetics. 2020 Jun;11(6):1307-17.

[3] Pang X, Zhou Y, Wang P, Lin W, Chang V. An innovative neural network approach for stock market prediction. The Journal of Supercomputing. 2020 Mar;76(3):2098-118.

[4] Radojičić, D., & Kredatus, S. (2020). The impact of stock market price Fourier transform analysis on the Gated Recurrent Unit classifier model. *Expert Systems with Applications*, *159*, 113565.

[5] Moghar A, Hamiche M. Stock market prediction using lstm recurrent neural network. Procedia Computer Science. 2020 Jan 1;170:1168-73.

[6] Chatzis SP, Siakoulis V, Petropoulos A, Stavroulakis E, Vlachogiannakis N. Forecasting stock market crisis events using deep and statistical machine learning techniques. Expert systems with applications. 2018 Dec 1;112:353-71.

[7] Dayri K, Rosenbaum M. Large tick assets: implicit spread and optimal tick size. Market Microstructure and Liquidity. 2015 Jun 4;1(01):1550003.

[8] Radojičić D, Kredatus S, Rheinländer T. An approach to reconstruction of data set via supervised and unsupervised learning. In2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI) 2018 Nov 21 (pp. 000053-000058). IEEE.

[9] Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015 Jan 1;61:85-117.

[10] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature. 1986 Oct;323(6088):533-6.

[11] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9, no. 8 (1997): 1735-1780.

[12] Radojičić, D., Radojičić, N., Kredatus, S.: A multicriteria optimization approach for the stock market feature selection. Computer Science and Information Systems, https://doi.org/doi.org/10.2298/CSIS200326044R