MI - CO472

Stock Market Prediction using stock data and tweets

Prajval M
16CO234
Semester 8
9449642887
26prajval98@gmail.com

Abstract

Stock Market Prediction is used to predict the stock-value of a company in the future. Prediction of stock-value can be done using many machine-learning techniques. There can be many models that can be used to predict these stock-values and each model has its own advantages and disadvantages. This paper uses a machine-learning model created using a combination of Long short-term memory (LSTM) or Gated recurrent unit (GRU) networks and Neural Networks. Most models only use the stock data for the prediction but this model also uses Twitter data along with the stock data i.e. the high stock value of the stock value for predicting the stock-value of a company. This paper also compares the usage of only stock value versus the usage of stock-value as well as Twitter data and the LSTM model versus the GRU models.

Keywords

Stock Prediction, Sentiment Analysis, LSTM, GRU, machine learning, Recurrent Neural Networks (RNNs), neural networks, machine learning models

Introduction

Economic researchers assume stock market prices to follow a random process. The stock market prediction has to account for many variables. This paper tries to bring in the sentiment of stock traders into account. Social Media is influencing a lot of people nowadays. Twitter is where a lot of people share their thoughts on and everyone including common people, influencers, etc. share and discuss their thoughts on different issues. Hence tweets can have a high causal impact on stock prices. Leveraging this fact, this paper proposes a model to predict stock prices. Machine learning models trained on the right data set can predict future stock prices close to the actual value. Traders would be able to make the right call based on the predicted prices, so as to maximize their profit and incur minimal losses.

The stock value data set has been collected from Kaggle. The data is of time-series type and contains relevant stock information about the IBM stock. It consisted of seven columns: date, open, high, low, close, volume, and name. A new column of tweet sentiment scores is added.

Two types of neural networks were used for building the models, namely, GRU and LSTM networks. They both are subtypes of a broader class known as Recurrent Neural Networks (RNNs). RNNs are neural networks that process sequences of data and make predictions based on the recent

input. A classical LSTM unit contains an input gate, a cell state, an output gate, and a forget gate. The cell state acts as the memory of the unit and the gates act as regulators. The input gate gets the input from the previous LSTM output. The forget gate removes a part of the information from the cell state depending on the recent input x_t and the previous output, h_{t-1} made by the unrolled network. The input gate layer decides how the new cell state, C_t should be updated. The gates internally have an activation layer which is triggered when the input crosses a certain threshold. The activation function generally used is either the hyperbolic tangent or the sigmoid function. The output gate computes the value that might be required by the unit in the next time step. A GRU cell does not contain the output gate. A GRU is a simplified variant of the LSTM unit.

The paper discusses how this is better and different from the other papers discussed in the literature review. Then it discusses the data set itself. The next section is about the models used in this paper. This paper compares the models trained with and without the sentiment scores obtained from the tweets. Then the paper derives a conclusion based on the comparison of the results and finally suggests future work that can be done to improve stock-prediction.

Literature Review

A detailed literature review was done in order to understand different Recurrent Neural Network architectures and also different methodologies used in stock market prediction. Machine learning is not so uncommon in stock market prediction however different papers used different machine learning techniques.

Sepp Hochreiter and Jürgen Schmidhuber [1] introduced LSTMs in 1997. It solved the Vanishing-Gradient-Problem which was a problem in traditional RNNs. Since then LSTMs have evolved a lot. Initially, LSTMs had only 3 gates but in 1999 a 4th gate forget gate was introduced into the architecture.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio [2] introduced GRUs. GRUs are a type of RNNs similar to LSTMs but lack an output gate and hence only have 3 gates. Hence GRUs

Chouhan, Lokesh & Agarwal, Navanshu & Parmar, Ishita & Saxena, Sheirsh & Arora, Ridam & Gupta, Shikhin & Dhiman, Himanshu [3] paper uses LSTM model and regression model to predict stocks. They are however using only stock-data as a factor.

Kalra, Sneh & Prasad, Jay [4] and Mankar, Tejas & Hotchandani, Tushar & Madhwani, Manish & Chidrawar, Akshay & C S, Lifna [5] both the papers consider social factors for stock prediction. The earlier [4] does sentiment analysis on recent news articles and classifies it as positive or negative. Then it combines this with the stock data and uses a model containing Kth Nearest Neighbour (KNN), Support Vector Machine (SVM), Naive Bayes, and neural networks to predict

the stock-value. The latter [5] however uses tweets to predict using sentiment analysis and using KNN and SVM models to predict the stock-value.

Work Done

Stock Market Prediction has become very popular nowadays. New techniques are being developed to perform this task. Machine learning has proved to be highly efficient in doing this task successfully. However, the efficiency of this task can be improved further by improving the machine-learning models used to predict.

Stock Market Prediction Using Machine Learning [3] paper does not care into the fact that external factors like social media response have a very high impact on stock prices as the influencers influence a lot of people and hence buying trends change day-to-day. Also, social media effects are immediate and it is highly varying from time to time.

Efficacy of News Sentiment for Stock Market Prediction [4] and Stock Market Prediction based on Social Sentiments using Machine Learning [5] use social sentiments on stock prediction using news articles and tweets respectively but they use KNN and SVM for predicting the stock-value. As there is a time-series using an RNN model makes perfect sense. Also, they classify the sentiments as positive or negative. A better model would tell if the sentiment is how much positive or how much negative as two positive sentiments can not be considered equal and so is the case with any two negative sentiments e.g. The company is performing good is a positive sentiment but The company is performing excellent is a more positive sentiment than the previous.

Hence a model which uses sentiment analysis, as well as an RNN based model will be more suitable for stock-prediction and the paper will focus on one such model.

Concepts

Machine learning is a subpart of Artificial Intelligence (AI) which deals with making systems learn themselves in order to make decisions and recognize patterns with very little human interaction.

Stock-value prediction can use machine learning to predict the stock-values of the future. In this paper, the main concepts useful are sentiment analysis and recurrent neural network types LSTM and GRU based networks.

Sentiment analysis is useful in analyzing a product's success based on the feedback received from the customers on the product. Analyzing the customer is the same as analyzing the sentiment of the customer's feedback. This is where sentiment analysis comes into the picture. Hence sentiment analysis is identifying different sentiments in a text and categorizing based on score or categories like positive or negative or neutral. It involves tokenizing words, removing the stop and unnecessary words, and then finally classifying words in the text based on polarity. The words are

already known to the system and have been already classified. Finally, we end up with a score or label to mark the piece of text.

The next most important concept used in this paper is Recurrent Neural Networks (RNNs). In RNNs for every new input we give previously stored outputs also as input along with the new input i.e. there is a feedback always given. Hence RNNs are most suitable for use in sequential data. However, for RNNs it cannot hold information for more than like 4-5 iterations. This is because of the very fact of diminishing gradients found in Artificial Neural networks (ANNs) which use gradient descent for training and backpropagation. What happens is in every iteration a small gradient value is lost and after a lot of iterations the whole information is lost. This is called the vanishing gradient problem and it is a very common problem in large artificial neural networks especially in the RNNs. Hence LSTMs and GRUs are used which does not face this issue.

LSTMs were majorly developed to solve the vanishing gradient problem. A classical LSTM unit contains an input gate, a cell state, an output gate, and a forget gate. The cell state acts as the memory of the unit and the gates act as regulators. The input gate is used to give the input to the LSTM blocks and subsequently the output of previous LSTM blocks is given as the next input. The forget gate removes a part of the information from the cell state depending on the recent input x_t and the previous output, h_{t-1} made by the unrolled network. The input gate layer decides how the new cell state, C_t should be updated. The gates internally have an activation layer which is triggered when the input crosses a certain threshold. The activation function generally used is either the hyperbolic tangent or the sigmoid function. The output gate computes the value that might be required by the unit in the next time step. A GRU cell does not contain the output gate. A GRU is a simplified variant of the LSTM unit.

Data set Creation

The data set for the stock value prediction was obtained from Kaggle data set. The data set is of IBM company. The data set is a Comma Separated Values (CSV) file. It contains the following row headers: Date, Open, High, Low, Close, Volume and Name. Date tells on what date the stock value was measured. Open is the opening value of stock value for that date. High is the highest value of stock value for that date. Close is the closing value of stock value for that date. Volume is the number of shares sold. Name is the name of the company whose stock is being sold. The data set contains all this information from the year 2006 to the year 2017.

The data set is divided into testing and training data. In Total there are 3020 entries. All the entries from 2006 to 2016 are in testing data and all the entries of the year 2017 in training data. In total there are 2769 entries in the testing data set and 251 entries in the training data set.
For the twitter sentiment analysis, a score is obtained from the tweets based on dates. This data was added to the Comma Separated Values (CSV) file using a python script which edits the file. The python script uses tweepy to gather the tweets. Python script logs into the twitter using OAuth

API where a consumer key and secret provided by twitter to a user is used. For each date upto 100 top rated tweets with the hashtag #ibm is gathered from the previous day till that day and sentiment of each tweet is obtained using TextBlob. TextBlob is a tool for sentiment analysis. It tokenizes words, removing the stop and unnecessary words, and then finally classifies words in the text on the basis of polarity. The words are already defined in the TextBlob module and have been already classified. Finally, we end up with a score or label to mark the piece of text. This score is averaged among all the tweets and the final tweet score is obtained for that day. Its value is between 0 and 1. A new column Score is created which refers to the twitter score and is filled in with this value. This is done for all the 3020 rows i.e. for both the testing data set and training data set.

Methodology & Implementation

The implementation of this model was done in python 3. The models were done using machine learning library keras. For support numpy module was used for manipulating the data and to read the data set pandas module was used. For sentiment analysis the TextBlob module was used. Scikit learn python module is used for all the statistics that are done.

The major concept used to create the stock market prediction model is based on a combination of RNN networks which uses LSTM or GRU networks along with Sentiment Analysis. This combination of the two is a change in traditional machine learning approaches being followed in stock prediction which either uses LSTM models or Kth Nearest Neighbor, Naive Bayes etc. if done along with sentiment analysis.

Layer (type)	Output Shape	Param #	Connected to
stock_input (InputLayer)	(None, 60, 1)	0	-
RNN_17 (LSTM or GRU)	(None, 60, 50)	10400	stock_input[0][0]
dropout_17 (Dropout)	(None, 60, 50)	0	lstm_17[0][0]
RNN_18 (LSTM or GRU)	(None, 60, 50)	20200	dropout_17[0][0]
dropout_18 (Dropout)	(None, 60, 50)	0	lstm_18[0][0]
RNN_19 (LSTM or GRU)	(None, 60, 50)	20200	dropout_18[0][0]
dropout_19 (Dropout)	(None, 60, 50)	0	lstm_19[0][0]

RNN_20 (LSTM or GRU)	(None, 50)	20200	dropout_19[0][0]
dropout_20 (Dropout)	(None, 50)	0	lstm_20[0][0]
dense_11 (Dense	(None, 1)	51	dropout_20[0][0]
tweet_input (InputLayer)	(None, 1)	0	dense_11[0][0]
concatenate_3 (Concatenate)	(None, 2)	0	tweet_input[0][0]
dense_12 (Dense)	(None, 64)	192	concatenate_3[0][0]
dense_13 (Dense)	(None, 64)	4160	dense_12[0][0]
dense_14 (Dense)	(None, 64)	4160	dense_13[0][0]
main_output (Dense)	(None, 1)	65	dense_14[0][0]

Total params: 79,628
Trainable params: 79,628
Non trainable params: 0

Table 1. Details of the proposed model

The proposed model takes stock input as one of its input data. The shape of the input layer is (60) i.e. it takes in the previous 60 stock-data values to predict the next one. Then there are 4 layers of RNN (LSTM or GRU) networks and for each input sequence there are 50 LSTM units for the 1st 3 layers. Each of them also has a 20% i.e. 0.2 drop so as to avoid overfitting of data. But the 4th layer does not return any sequence and hence there are only 50 units overall. It also has a dropout of 20%. Next this layer is connected to a dense layer which provides an intermediate output which acts as an input to the next model which uses twitter sentiment score. If considering a model without sentimental analysis this output would predict the stock value.

The next layer takes the RNN (LSTM or GRU) output as input along with twitter sentiment score. Next layer is the concatenate layer which merges the RNN (LSTM or GRU) model and the sentiment analysis model. The next 3 layers is a densely connected neural network layer consisting of 64 neurons. The final layer is the output layer which predicts the next stock value. The output layer uses sigmoid function as the activation function.

Table 1. describes each layer, its input shape, number of parameters they have and what layer they are connected to. Fig 1. describes the model diagrammatically being discussed along with the input

shapes. Totally the model has 79,628 parameters all of which are trainable and are trained during the training phase.

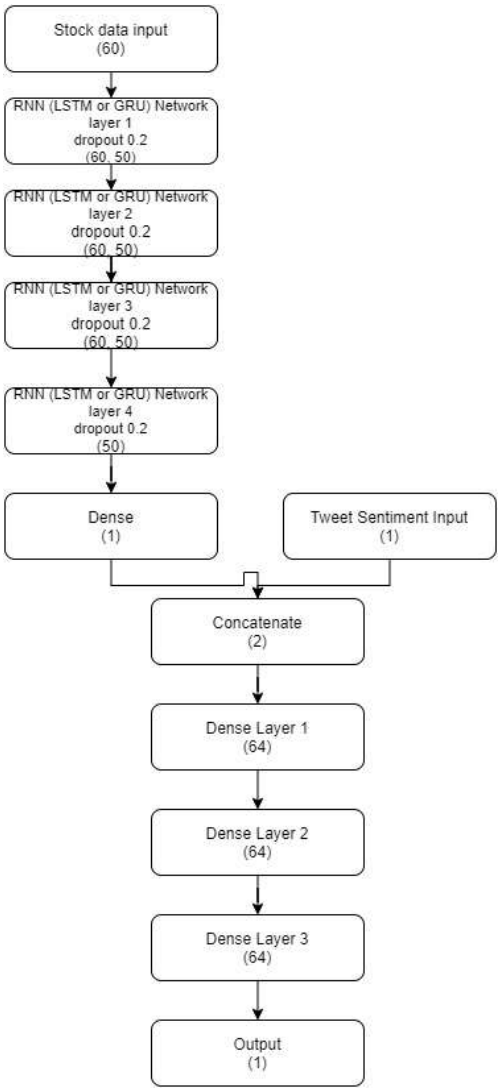

Fig 1. Proposed Stock Prediction Model

Training & Testing

For training the first 2769 rows of the twitter sentiment score added Comma Separated Values (CSV) file were used. We are using the high value of each stock-data to predict the high value of the stock value. Also twitter sentiment score for each of the previous day is being used to train. For each stock-value prediction the previous immediate 60 stock values are used as an input. Hence the training data set contains an array of array of 60 stock high values along with a sentiment score to predict the output stock high value. The model uses root mean square error for loss function and root mean square propagation (rmsprop) as optimizer. The training was done in 21 epochs for the LSTM based model and in 20 epochs for GRU based model but both of them are still comparable and a batch size of 32. This was the same for all the 4 models i.e. LSTM without sentiment analysis,

LSTM with sentiment analysis, GRU without sentiment analysis and GRU with sentiment analysis.

For the testing the last 311 rows of the twitter sentiment score added Comma Separated Values (CSV) file were used. This is because we need to predict 251 stock values and to predict each stock value prediction needs 60 immediate previous values. The model was tested for all the testing data.

Experimental Results

Graphs were plotted to visualize the predicted stock value to the real value and Root Mean Square Error was calculated.

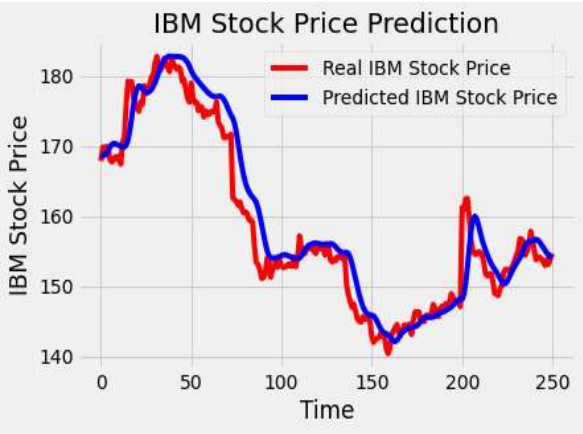

Fig 2. Stock Prediction using LSTM without Sentiment

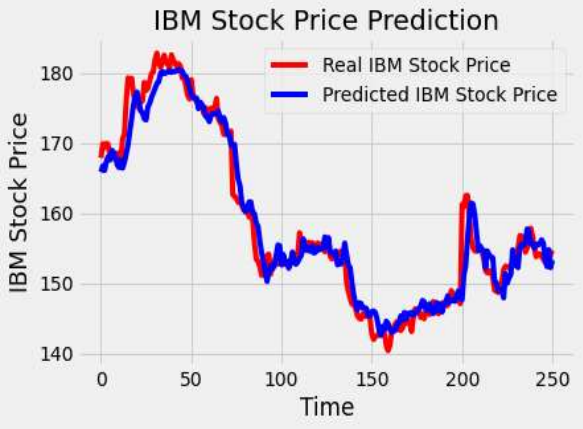

Fig 3. Stock Prediction using LSTM with Sentiment Analysis

The above two figures represent the graph of predicted vs the actual stock price of the company. Fig 2. represents the stock prediction of the LSTM model without sentiment analysis data. The line in the red represents Real IBM Stock Price and the line in blue represents Predicted IBM Stock Price. It can be seen that predicted values are very close to the real values and the Root Mean Square Error value is 4.89. In the right Fig 3. represents the stock prediction of the LSTM model with sentiment analysis data. Here also the line in the red represents Real IBM Stock Price and the line in blue represents Predicted IBM Stock Price. Here also it can be seen that predicted values are very close to the real values and the Root Mean Square Error value is 2.57. For the Root Mean Square Error lower the better and hence sentiment analysis has improved the efficiency of stock-value prediction. It can also be seen from the graph as Fig 2. has more gaps between real and predicted stock prices.

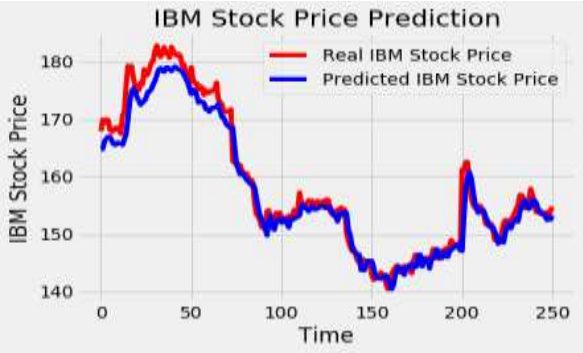

Fig 5. Stock Prediction using GRU with Sentiment Analysis

Fig 4. Stock Prediction using GRU without Sentiment Analysis

The above two figures represent the graph of predicted vs the actual stock price of the company. Fig 4. represents the stock prediction of the GRU model without sentiment analysis data. The line in the red represents Real IBM Stock Price and the line in blue represents Predicted IBM Stock Price. It can be seen that predicted values are very close to the real values and the Root Mean Square Error value is 4.68. In the right Fig 5. represents the stock prediction of the GRU model with sentiment analysis data. Here also the line in the red represents Real IBM Stock Price and the line in blue represents Predicted IBM Stock Price. Here also it can be seen that predicted values are very close to the real values and the Root Mean Square Error value is 2.42. For the Root Mean Square Error lower the better and hence sentiment analysis has improved the efficiency of stock-value prediction. It can also be seen from the graph as Fig 2. has more gaps between real and predicted stock prices.

Conclusion

The objective of this paper was to attempt to create a machine learning model for stock prediction using RNN networks (LSTM and GRU) and sentiment analysis with better accuracy and reliability as compared to the standard RNN models which use LSTM networks or GRU networks.

For the LSTM based models it can be seen from the graphs and the Root Mean Square Error values that the model which took into account the twitter sentiment gave better predictions compared to the model which does not consider sentiment. For the GRU based models it can be seen from the graphs and the Root Mean Square Error values that the model which took into account the twitter sentiment gave better predictions compared to the model which does not consider sentiment. Hence for stock prediction sentiment analysis helps in improving the prediction and also RNN models can also be used in stock-prediction along with sentiment analysis successfully in order to give better results.

Even though GRU performed slightly better in this case, it can be seen that the performance difference is not that much. Hence LSTMs and GRUs perform almost the same. However, training

of GRU is easier as it does not have the fourth gate i.e. the output gate, hence there are less computation.

Future Work

A lot of work has been done on Stock-Prediction using Machine Learning models but still there is a lot of work left. Sentiment Analysis based on twitter data was done in this paper however. Instagram posts are influencing more than tweets nowadays hence a combination of Computer Vision and RNN can be used to develop a robust model along with using twitter data. These models can also be used in outlier detection in Stock Prediction e.g. fraud detection in Stock Prediction. This paper used completely a LSTM model or a completely a GRU model, a model which has a combination of GRU and LSTM network can be done and compared with this paper. GRUs were introduced very recently hence not much research has been done on them. Using GRUs for different applications and checking its performance. Comparing the performance, efficiency, and computation speed of LSTMs and GRUs.

References

1. Sepp Hochreiter and Jürgen Schmidhuber, " LONG SHORT-TERM MEMORY", 1997
2. Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", 2014
3. Chouhan, Lokesh & Agarwal, Navanshu & Parmar, Ishita & Saxena, Sheirsh & Arora, Ridam & Gupta, Shikhin & Dhiman, Himanshu, "Stock Market Prediction Using Machine Learning," ICSCCC, 2018.
4. Kalra, Sneh & Prasad, Jay, "Efficacy of News Sentiment for Stock Market Prediction," COMITCon, 2019
5. Mankar, Tejas & Hotchandani, Tushar & Madhwani, Manish & Chidrawar, Akshay & C S, Lifna, " Stock Market Prediction based on Social Sentiments using Machine Learning," 1-3 ICSCET, 2018
6. S. A. Bogle , W.D. Potter, "SentAMaL - A Sentiment Analysis Machine Learning Stock Predictive Model," in Proc. of Int. Conf. of Data Mining and Knowledge Engineering , 2015
7. V. S. Pagolu, K.N. Reddy, G. Panda , B. Majhi , "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements," Int. conf. on Signal Processing, Communication, Power and Embedded System(SCOPUS) 3-5 Oct. 2016
8. A. Mittal , A. Goel , "Stock Prediction Using Twitter Sentiment Analysis," Stanford University, 2012

9. V. Ingle, S. Deshmukh, “Hidden Markov Model Implementation for Prediction of Stock Prices with TF-IDF features ,” in Proc. of the Int. Conf. on Advances in Information Communication Technology & Computing, 12-13 Aug. 2016
10. K. A. Althelaya, E. M. El-Alfy and S. Mohammed, "Evaluation of bidirectional LSTM for short-and long-term stock market prediction," 2018 9th International Conference on Information and Communication Systems (ICICS), Irbid, 2018, pp. 151-156
11. T. Gao, Y. Chai and Y. Liu, "Applying long short term memory neural networks for predicting stock closing price," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, 2017, pp. 575-578
12. H. Gunduz, Z. Cataltepe and Y. Yaslan, "Stock market direction prediction using deep neural networks," 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, 2017, pp. 1-4
13. M. Billah, S. Waheed and A. Hanifa, "Stock market prediction using an improved training algorithm of neural network," 2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), Rajshahi, 2016, pp. 1-4
14. Sorto, Max, Cheryl Aasheim, and Hayden Wimmer. "Feeling The Stock Market: A Study in the Prediction of Financial Markets Based on News Sentiment." (2017)