International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Stock Value Prediction using Machine Learning

R. Aadith Narayan, Baby D. Dayana, B. Yagneshwaran, Vignesh Babu M.R, Kurmala Ajay Vamsi Krishna

Abstract*: *The prediction of stock value has always been esoteric to stock analysts and statisticians. However, the stock market banks on public investment for its survival which also accounts for its dynamic nature. Previous methods of stock value prediction involve implementation of applied statistics, machine learning, news feed extraction with a moderate prediction accuracy. The proposed system involves the best practices from previous attempts and also a new approach to stock value prediction which would have an improved prediction accuracy than previous systems. The proposed system is implemented using deep learning: LSTM (Long short-term memory) and RNN (Recurrent neural network) algorithms which act as the prediction model and thus helps in delivering accurate predictions for the future by analyzing the pattern of variable stock prices for a time period. A conjunctive system of a keyword extractor and a sentiment analyzer directed towards news articles hosted by Twitter would help indicate the current performance of the company whether optimistic or not. The usage of deep learning algorithms provides a more robust mechanism to predict stock prices. The sentiment analyzer indicting the performance of the company thus acting as an important asset for investors to understand the stability of the company during the long term. The proposed system holistically covers all the important parameters considerable for an investor to invest in a particular company. Also, the proposed system helps in eliminating the esoteric nature behind stock analysis and encourages the common investors with partial knowledge of finance to invest in the stock market.

Index Terms*: *Sentiment Analysis, LSTM, RNN, Machine Learning and Python.

I. INTRODUCTION

There has always been an implementation of a system as per convention or tradition. This is mainly attributing to the avoidance of risk and a trust factor people have in the accuracy and end result achieved by the traditional system. However, in an advancing economy, a revolutionary upgrade has to be compromised over the traditional system to cope up with the demand of the market. Such is the traditional stock prediction system which is still esoteric to stock analysts and statisticians. Stock has always been a go-to option for companies to gain investment from the public which enables them to commence new projects or maintain performance. The stock market which depends on public investment is swayed by the recommendations of these analysts and statisticians. Technical analysis has been an option for the public to understand the market using indicators which are statistical methods to indicate the share value for the future. However, this methodology is complex to understand due to its high learning curve. There has to exist a stock prediction system which indicates the rise and fall of stock value over a period of time. Most importantly, it should be commercially which would be accessible to investors all over the world to gain assistance from this system to invest in the right company at the right moment. There have been previous attempts at stock prediction but only limited to research which has been implemented. Some of which uses applied statistics, machine learning, technical analysis, Momentum, trading volume, moving average etc. However accurate these systems are, it still could not be proven fit for commercial deployment because of the accuracy and trust issues which followed these project implementations. Therefore, the proposed system aims to provide a system which is trust worthy and is accurate at the same time of the general public.

Revised Manuscript Received on 30 March 2019.
***** Correspondence Author
Baby D. Dayana*, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
R. Aadith Narayan, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
B. Yagneshwaran, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Vignesh Babu M.R, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
Kurmala Ajay Vamsi Krishna, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

II. RELATED WORK

[1] does the stock prediction based on the social sentiment. In this system the analysis is done on the basis of the news about the companies in social media, for this the system uses Yahoo finance and Twitter API to get the input datasets. Yahoo finance is used to get the stock data and Twitter API is used to analyze tweets. In order to use the API, the user needs to have a key to access the API. Once the access is granted, there is a need to preprocess the data that is retrieved, that is to refine the data by removing the recurrent and less important words so as to make it suitable for prediction. The sentiment analysis is applied to the data and for this, two classifiers are used: Naive Bayes and Support Vector Machine. Both the classifiers analyze to provide the same results. In order to handle large datasets, Natural Language Toolkit is used to filter out the important features and using this the data is generated. Once the data is generated, it is used to train the system for carrying out sentiment analysis and the generated result set is applied to predict stock value. [2] discusses about prediction of stock trends using moving average supported by news classification. First, the day to day stock data is retrieved on the basis of date and stock closing value and the retrieved data is stored.

Then the simple average is calculated for 5,10,20,50 days. For short term, the data for 5 and 10 day average is considered and for long term the data for 20 and 50 day average is considered and also distant value is calculated by subtracting short term value from long term value. The dataset is built using the data retrieved from online newspapers and then the feature processing is implemented which consists of 3 phases: the first is single discrete word feature where the entire news is described in a single word, the second one uses dictionary feature in order to apply label to a feature, the last process is representing the features as a word vector that is positive negative and neutral. These features are then implemented with Machine Learning using Artificial Neural Network. This system further classifies all the recent news in terms of sentiment value. Machine Learning using Artificial Neural Network is used to predict the Stock trend on the basis of three features. The moving average of the stock value, the total positive sentiment score and the total negative sentiment score.

[3] discusses about Keyword Extraction using Machine Learning. The system utilizes Machine learning to extract the keyword which can be further used to retrieve details based on the keyword extracted. For this, a known dataset is first fed into the system and then the dataset is prompted for preprocessing. The preprocessing is mainly carried out in order to refine the data that is to be analyzed by removing recurrent and meaningless words from the dataset and then the data is subjected for lexical analysis where the words are converted into tokens which can be further used for keyword extraction. The tokens are generated using NLP. Then the Keyword is extracted using Unsupervised Classification algorithm and are extracted on the basis of the similarity between the documents.

[4] details about Sentimental analysis on the basis of machine learning and semantic analysis. The dataset for the system is fetched from the twitter API. The fetched dataset is then preprocessed. This is done to refine the data by removing the recurring and useless word so as to make it easy for analysis. The improved dataset is further used for feature extraction and so as to determine the polarity of the news to determine the mindset of the public. Then the training of the system is carried out in order to improve the efficiency of the system. For these 4 techniques are used namely Naive Bayes, Maximum Entropy, Support Vector Machine and Semantic Analysis. The maximum entropy maximizes the entropy defined on the conditional probability distribution.

[5] elaborates about prediction model stock trend analysis. In this system the dataset is divided into two parts Training and Testing. In training, the mean value of the day to day change of stock value is calculated which would depict the rise and fall in price. A matrix is prepared so as to check if the stock value increases compared to the previous day. If it increases then it is defined as 1 else it is defined as 0. Then on the basis of the matrix data, the prediction is made.

III. METHODOLOGY

The proposed system aims to provide support to investors willing to invest money in the stock market. It especially aims to remove the esotericism surrounding the stock market and encourage laymen to take part in the proceedings. The proposed system consists of four modules: Data exploration, Feature engineering, Building the prediction model and visualization and Sentiment analysis. The program was implemented using Jupyter Notebook and Spyder. The programming language used was Python mainly because of its optimization features like Numpy which enables faster mathematical computations. The proposed system was trained to predict the stock value of the company Apple from February 2013 to February 2018.

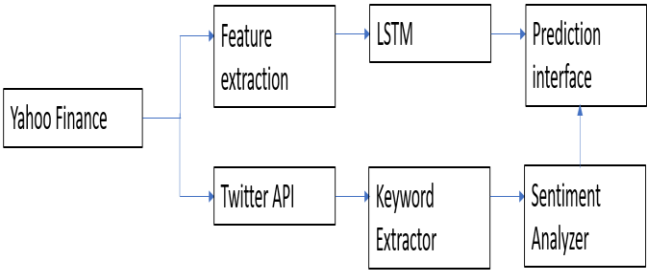


Fig 1: System Architecture

A. Data Exploration

The data exploration for the proposed system involves acquiring the dataset for training and testing. The dataset was acquired from Kaggle and downloaded. The dataset was checked for any redundancies which might affect the prediction result of the algorithm and acted upon. The features of the dataset were well understood and the size of the dataset was noticed. The dataset was uploaded then to the file depository of Jupyter notebook.

B. Feature Engineering

After the dataset is loaded in the file depository, the python program commences. The required packages namely numpy, pandas, matplotlib (pyplot), sklearn (Minmaxscaler), keras (Sequential, load_model, LSTM, Dense and Dropout) and os are imported. The dataset is read into a data frame using pandas. When analyzing the head of the data frame, it is concluded that only the “open” column which depicts the opening value of the Apple stock every day is required for prediction. The data frame holds 1259 features. Thus, the particular column is filtered and stored in the data frame which is accessed using a numpy array. The prediction model involves two sets of data: Training dataset and testing dataset. The training dataset helps to train the neural network by reducing the prediction loss rate with each prediction. The testing dataset is the data which is used to evaluate the accuracy of the prediction model. Both the datasets are initialized as numpy arrays. The training dataset is defined with 80% of the data frame while the testing dataset is defined with the rest 20% of the data frame. The training dataset contains 1007 features while the testing dataset contains 302 features. A minmax scaler is used to limit the value of the features between 0 and 1 as the prediction algorithm used is a regression algorithm. The values are standardized to identify any outliers present in the data. A function is then created which is aimed to create a dataset. In the function definition, the last 50 features are omitted to save it for the prediction model after. Numpy arrays are created and defined to house the training and testing datasets.

The function is called with training and testing arrays to assign it to the respective labels. The numpy arrays are reshaped into a three-dimensional array to fit in the LSTM layer.

C. Building the prediction model and visualization

The prediction model used is Long Short-Term Memory (LSTM). The prediction model is built by adding layers on top of another layer which forms a sequence. The LSTM is defined using Keras and the arguments are specified with respect to each layer. From the initial layer till the penultimate layer, the return sequence is set as true to carry the sequence over to the next layer. The input shape is defined with the shape of the training dataset. The dropout layers are also defined. At the final layer, the return sequence is set as false which marks the end of the LSTM layer sequence. The dense unit is defined as one because only a single value is expected to be predicted. As the proposed system is a regression problem, the loss function is set as mean squared error and an adam optimizer is used. Then the prediction model is trained up to 50 epochs. With each epoch, the loss gradient keeps decreasing which is a sign that the accuracy of the model is rising. The training specifications are stored in a h5 file. The prediction is then initiated using the model.predict function. The predicted value is inversely transformed to retrieve the original scale before the preprocessing. The 50 features which are now included to compare the prediction. The prediction is visualized using a line graph enabled by pyplot. The original values and predicted values are assigned unique color codes to differentiate between the two. The same process is repeated with the testing data and is visualized as a line graph to observe the accuracy of the model. Thus, the prediction is done for the dataset using LSTM which is then visualized.

D. Sentiment Analysis

The sentiment analyzer is used to identify the current perception of the public on the particular company, in this case Apple, to indicate whether there would be any significant movement of the stock price in the future. Stock price is hugely dependent on the performance of the company and the stability of the positions in the company. The sentiment analysis is carried out on twitter feeds which is one of the leading news and controversy source. The twitter API is accessed using the consumer key and consumer secret and also the access token and access token secret. The keyword is then taken as input and also the number of tweets to be analyzed is also taken as input because it is hugely dependent on the strength of the internet and the computer specifications. The package used to implement the keyword extraction is Tweepy and the package used for sentiment analysis is Textblob. Using tweepy, the twitter feeds are extracted and stored in a list. Using Textblob, the tweets are analyzed for the sentiment. Textblob returns a value between -1 to +1. The sentiment scale is defined as follows: 0- neutral, 0 to 0.3- weakly positive, 0.3 to 0.6 is positive, 0.6 to 1 is strongly positive, 0 to -0.3- weakly negative, -0.3 to -0.6- negative, -0.6 to -1- strongly negative. The decimal values are converted to a percentage and the average of all the percentages is deemed as the general report which is the effective sentiment. With the percentages, a pie chart is designed and displayed with unique color codes for each sentiment using pyplot package. Thus, the sentiment of the company is found and is indicated with the interface along the prediction which when analyzed can depict the current picture of the company on whether it is a good investment or not.

IV. EXPERIMENTAL ANALYSIS

There was a need for a robust system which will predict the stock price with existing data. The proposed system is worthy of the recognition in both accuracy and performance. The prediction algorithm was implemented on Apples opening stock price for the years 2013-2018. Initially the algorithm was implemented for the last 50 values of the training dataset for which the line graph is as follows:

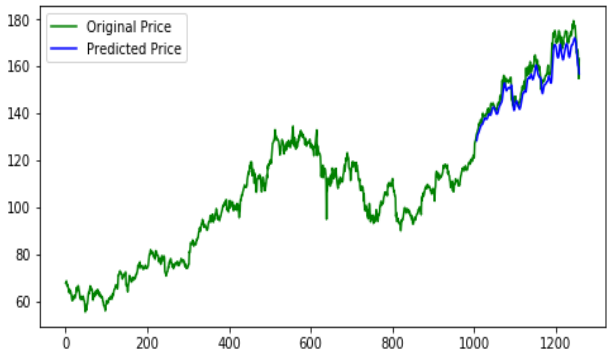


Fig 2: Line Graph of The Training Dataset

As displayed, the prediction pattern looks to follow the original price pattern accordingly. However, there are minor deviations from the price value which suggests that it is not completely accurate. The implementation of the trained prediction model on a testing dataset yielded a line graph as follows:

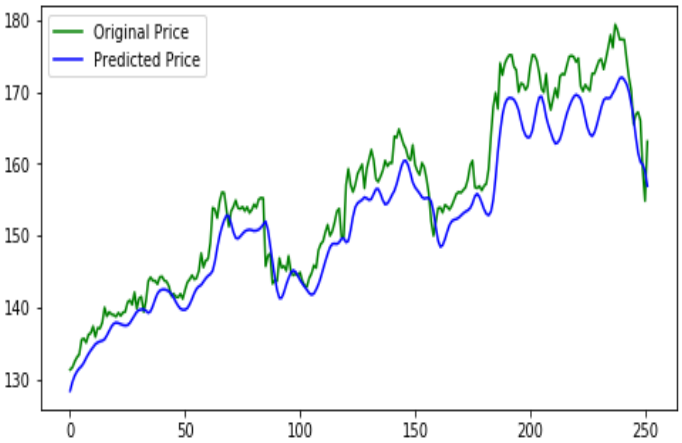


Fig 3: Line Graph of Testing Dataset

With reference to the line graph above, as expected, the prediction system is able to map the predicted price colored blue along with the original price colored green. Thus, the proposed system is capable of predicting the pattern of stock movements and also the future price value with accuracy if trained on a larger dataset. However, the prediction model only takes into account the previous price pattern over the past 5 years. Stock movement is also hugely dependent on the public perception and the performance of the company.

To gain a foresight to this, the sentiment analyzer is used. The following is a screenshot of the result provided by the sentiment analyzer for the company Apple depicted in the stock market as AAPL:

```
Enter Keyword/Tag to search about: AAPL

Enter how many tweets to search: 500
How people are reacting on AAPL by analyzing 500 tweets.

General Report:
Weakly Positive

Detailed Report:
23.40% people thought it was positive
15.60% people thought it was weakly positive
4.20% people thought it was strongly positive
8.00% people thought it was negative
4.40% people thought it was weakly negative
0.60% people thought it was strongly negative
43.80% people thought it was neutral
```

Fig 4: Report of the Sentiment Analysis

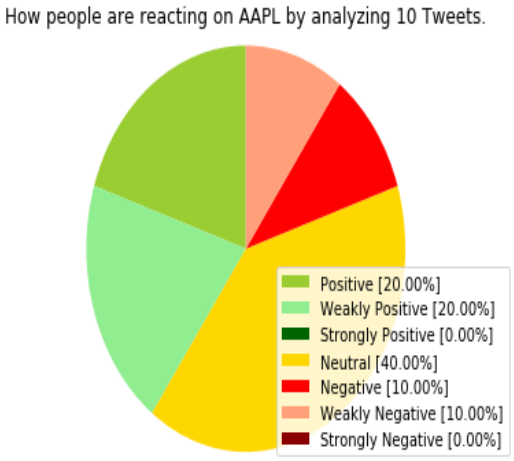


Fig 5: Pie Chart of Sentiment Analysis

The sentiment for Apple is weakly positive which means although there is no immediate risk, there will be risk in the future. Thus, an overall idea of the company's stock price in the future and the current state of the company is provided by the proposed system which would offer assistance for investors to invest in the company.

V. FUTURE ENHANCEMENT

The proposed system is an optimistic attempt at predicting stock prices by considering most of the factors which might affect the stock value pattern. Even though this system is accurate, stock prices are highly dynamic. The usage of deep learning has proved to be effective in analyzing the stock patterns effectively which plays a major role in the prediction phase. With a better system which has the capability of analyzing huge datasets, the prediction model can be trained so as to make it extremely robust in predicting the stock value. With time, better deep learning models can be employed so as to optimize the prediction mechanism and include more factors like indicators which can add resilience to the system.

VI. CONCLUSION

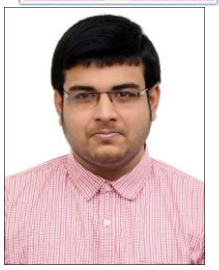
Stock market prediction has always been a very esoteric domain restricted to stock analysts and statisticians. When considering the stock market, prediction is more of an indication towards the movement of stock in the future. Our implementation of LSTM with Sentiment Analysis proved to be a very efficient and feasible model in predicting the stock price. The proposed system's usage of LSTM and RNN has the best accuracy in terms of predicting stock value. The sentiment analyzer has proved to be an important tool to earn confidence to invest in a company for the long term. Even though this system is a foundation to even more developments to come, it has proved to be efficient to the current market. The unpredictability of the stock market is what makes the study of stock so interesting and also acts as a motivation for us to implement this system. The proposed system is just the beginning for many developments to come.

REFERENCES

1. Tejas Mankar , Tushar Hotchandani , Manish Madhwani and Akshay Chidrawar. "Stock Market Prediction based on Social Sentiments using Machine learning", 2018 Mankar.
2. Stefen Lauren and Dra.Harlili. "Stock Trend Prediction Using Simple Moving Average Supported by News Classification", 2014 International Conference of Advanced Informatics Concept, Theory and Application.
3. Bhavneet Kaur and Dr.Sushma Jain. "Keywords Extraction Using Machine Learning Approaches", 2017 Kaur.
4. Dr.Devpriya Soni , Sparsh Agarwal , Tushar Agarwal , Pooshan Arora , Kopal Gupta. "Optimized Prediction Model For Stock Market Trend Analysis", 2018 Eleventh International Conference on Contemporary Computing.
5. Geetika Gautam and Divakar Yadav. "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis".

AUTHORS PROFILE

Mrs. Baby D. Dayana is an assistant professor at the Department of Computer Science, SRM Institute of Science and Technology, Ramapuram.



R. Aadith Narayan is a student at the Department of Computer Science, SRM Institute of Science and Technology, Ramapuram.

B. Yagneshwaran is a student at the Department of Computer Science, SRM Institute of Science and Technology, Ramapuram.

Vignesh Babu M.R is a student at the Department of Computer Science, SRM Institute of Science and Technology, Ramapuram.

Kurmala Ajay Vamsi Krishna is a student at the Department of Computer Science, SRM Institute of Science and Technology, Ramapuram.