International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Predicting Stock Market Trends using Hybrid SVM Model and LSTM with Sentiment Determination using Natural Language Processing

shashank Singh, maaz Ahmad, aditya Bhattacharya, m. Azhagiri

Abstract: ***In the financial world, stock trading is one of the most crucial activities. Investors make educated guesses to predict stock market trends by analyzing news, studying the company history, industrial history and a lot of other data. successfully predicting the stock market trends and investing in the right shares at the right time can maximize the investor's profit or at least minimize the losses. Stock market price data is generated in huge volumes and is affected by various diverse factors. This work proposes two models to predict the stock market prices. The first model is an LSTM model that employs a backpropagation optimized LSTM network to forecast future stock prices. The second model is a hybrid model that combines an SVM model, KNN model and a Random Forest classifier using the Majority Voting algorithm to predict stock market trends. Both models have a sentiment analyzer to factor the news influencing the stock market using Natural Language Processing (NLP). The project aims to help investors who are new to the stock market and don't possess sufficient knowledge to make share investments as well the experienced investors by predicting stock market trends.***

Keywords: Stock Prediction, LSTM, SVM, KNN, Random Forest, Majority Voting, Sentiment Analysis, Natural Language Processing (NLP)

I. INTRODUCTION

The Indian stock market is an extremely complex system wherein large volumes of data are generated every second and change rapidly due to various diverse factors. Companies divide their ownership in the form of stocks. A stock is an investment that symbolizes ownership in a company. Purchasing a company's stock means purchasing ownership in that company. Hence the investors who invest in a company's stock become a part of the company's overall profit or loss. Hence by investing in the stock market one can efficiently increase one's net worth. Stock market is a platform for buying, selling and trading company stocks.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author
Shashank Singh*, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu
Maaz Ahmad, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu
Aditya Bhattacharya, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu
M. Azhagiri, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu

The stock market is non-linear, discontinuous and changes rapidly as it is affected by many diverse factors such as rumors in news, financial activities, political events, and various other mathematical factors. Buying the right company stock and selling it at a time when its value is higher than its original value results in making profits. When the trend of the market is successfully predicted, profits are made proportional to the investment. The challenge is to correctly predict the stock prices accurately to minimize loss and maximize profit.

Machine learning (ML) and Deep Learning (DL) can play a very important role in the field of finance. By using the various ML and DL algorithms, we can efficiently collect and visualize the huge volumes of stock market data and manipulate it effectively to precisely forecast the stock market trends. The popular traditionally used Machine Learning algorithms for predicting stock market trends are the various regressions like the Linear Regression and The Polynomial regression and Classifications such as the Random Forest classification, SVM classification, KNN algorithm. But the regression algorithms alone are unable to factor the various elements affecting the stock market and do not give accurate results. All the above-mentioned ML algorithms have been used in various studies and proved to give poor to moderate results and were unable to give good accuracy. A better approach to predict stock market trends more accurately and efficiently would be to create a Hybrid model that combines all of these popular ML classification algorithms and get a combined accurate prediction. In the domain of DL, RNN networks have been conventionally put to use to predict the stock market trends, but the RNN networks are not very efficient due to their problem of Vanishing Gradient. To put it in simple words, RNN retains information for only short intervals of time, i.e. if we need the information after a small duration of time, it may be available, but after a number of iterations, the original information gets lost. This limitation is overcome by using LSTM networks. LSTM networks have proved to be very efficient and versatile in predicting the stock market trends with very high accuracy. LSTM is a part of the Recurrent Neural Network structure but differs from the conventional RNN in the way it modifies its previous state data after every iteration ie. the original data is not lost after a few iterations. The prediction of the stock market trends involves not only factoring the historical and real-time data but also factoring the news that significantly affects the stock market.

The stock market is considerably affected by the political events and other news in the media. Sentiment Analysis is the process of determining the sentiment behind a piece of text and to determine the writer's attitude and emotion behind writing the text.

In order to understand the sentiment behind the news and predict how it will affect the stock price, a Sentiment Analyzer must be employed to determine the sentiment behind the news and measure its polarity.

This work proposes a hybrid model combines the various ML classification algorithms viz SVM, KNN and Random Forest algorithms using the Majority Voting algorithm to predict stock market trends. The work also proposes a backpropagation optimized LSTM model for the stock market prediction. Both models factor a given company's historical stock price data as well as the news affecting the company stocks data to predict the company's future stock trends accurately and efficiently. Both models employ a Sentiment Analyzer to carry-out the sentiment analysis on the company news.

II. RELATED WORK

Over the years, many models and algorithms have been employed to predict stock market data. In the field of Machine Learning (ML), the most popular algorithm used is the Support Vector Machine (SVM), [1] the SVM has been widely used to predict the stock market trends since it is noise-tolerant and gives a decent accuracy. Other ML algorithm used is the K-Nearest Neighbour (KNN), [2] KNN has been used to predict stock market data to some extent but not with great accuracy. In the field of Deep Learning (DL), the most popular algorithms that have been used are the Artificial Neural Networks (ANN), [3] ANNs have been conventionally used to predict stock market trends but haven't proved to be very efficient due to their limitation of data-overfitting. The other very popular DL algorithm used is the Recurrent Neural Network (RNN), [4] RNN has been utilized to predict stock market data and gives descent to impressive accuracy depending on the number of epochs. The RNN hasn't proved to be very accurate in predicting complex stock market trends due to their Vanishing Gradient problem. Other approaches [5] attempted at predicting stock market data are based on the Sentiment Analysis of the news related to a company. By determining the sentiment behind a company news, we can determine how a particular event or news will affect the stock price of a company. The Sentiment Analysis [6] of a company news is implemented by doing the sentiment analysis of the Twitter news related to a company's stock. Several comparative studies have been done to compare the various individual ML and DL algorithms.

A. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification-cum-regression Supervised Machine Learning (ML) algorithm that can be employed for both classification and prediction problems. The algorithm takes every individual data entry as a point in a multi-dimensional space (which is determined by the number of features 'n') and the coordinates of the particle are its value. SVM algorithms employ mathematical functions called kernel functions for classification and prediction. The function of the kernel is to take input data and transform it into suitable classified form.

B. K-Nearest Neighbour Algorithm

The K-Nearest Neighbour (KNN) is a simple algorithm that can be employed both for classification and regression. It has proved to be more powerful and efficient than the conventional Supervised Learning algorithms like Logistic Regression and Random Forest algorithms. It takes a dataset with known categories and a new target dataset of an unknown category for classification. The algorithm finds that which known categories are the closest to the unknown target category (nearest neighbors) and tries to classify the unknown category. If the 'K' in the "K-Nearest Neighbours" is equal to the 1, then we will classify the target unknown category using the nearest neighbor category. Similarly, if K is equal to 'n', then we will use the 'n' nearest categories to classify the unknown category. In the case where the target unknown category is equidistant from two or more known categories then we count the votes from each category, and the category with the majority votes is considered the nearest neighbor and is used for classification.

C. Random Forest Algorithm

The Random Forest Algorithm is basically a collection of a large number of decorrelated decision trees and uses them to carry-out classification using Bagging technique. The algorithm takes in a sample dataset as its input and then divides this sample set into a number of sample subsets with random values. From each random subset, the algorithm creates a decision tree. So, from 'n' sample subsets the algorithm creates 'n' decision trees. All decision trees make a different classification for the main sample set. The algorithm then creates a ranking of these decision tree classifiers. The classification is made by using the Voting algorithm. The classification from each decision tree is counted as a vote and the classification with the majority votes is chosen by the algorithm. This majority chosen classification is selected as the main output sample classification.

D. Majority Voting Algorithm

Majority Voting algorithm combines various predictive or regressive models to give better results and higher accuracy. The first step is to create multiple classification or predictive models for different subsets of the same data. Run the various models and get a prediction output from each. Store the result of each model in a matrix or a multi-dimensional array. The result or prediction of every model is counted as a vote. Then the matrix is traversed and the instances of each prediction are counted. The final output is the prediction that received more than half of the total votes.

E. Long Short Term Memory (LSTM)

LSTMs are a part of the Recurrent Neural Networks but help overcome the Vanishing Gradient problem encountered in the conventional RNN. Similar to RNN, we have time steps in LSTM but there is an extra feature called "MEMORY" in LSTM for every time step. LSTM network is comprised of different memory blocks called cells. Two states are given as input from the previous cell to the next cell: the cell state and the hidden state. Three gates are used for managing information viz Forget Gate, Input Gate and, Output Gate. A forget gate removes unwanted information from the cell state.

The input gate appends new information to the cell state. It hence adds new information using the sigmoid function as well creates a vector of all the possible values suitable for addition to the current cell state.

III. DEVELOPED MODEL

This work proposes two models to predict stock market trends using a Hybrid model and the LSTM model. Both models use both the historical data of a company stock as well as the company sentiment and news affecting the company share price together to predict future trends of the company stock. For processing the historical dataset of a stock, the Hybrid model combines Support Vector Machine (SVM), K-Nearest Neighbours (KNN) and Random Forest Algorithm using Majority Voting Algorithm. The LSTM model employs the backpropagation optimized LSTM networks. The news is collected from various platforms like Twitter and is processed and classified using the Machine Learning technique called Natural Language Processing (NLP).

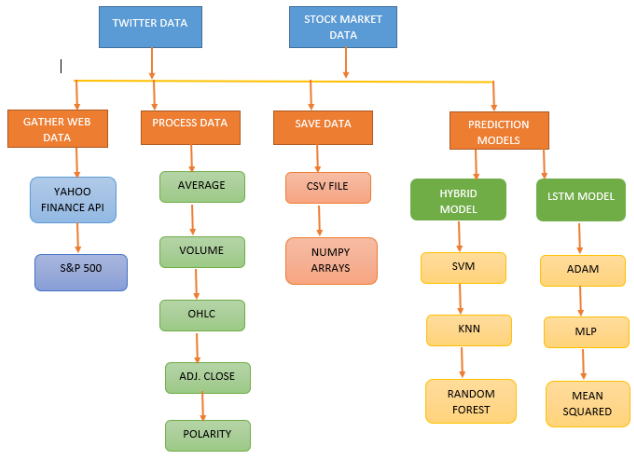

Fig. 1: Architecture Diagram Depicting the Developed Model

A. DATA COLLECTION

The ***Historical Stock Market Data*** is collected from the Yahoo Finance API. The Yahoo finance API is employed to gather historical stock price data of the required company.

The Real-Time Stock Trading Data is collected using the Yahoo Finance API and Alpha Vantage API from the NSE website. This real-time data is utilized to forecast the next-minute prices of company stock to assist real-time trading.

Sentiment Analysis Data is collected from Twitter using the Tweepy API. Crawler is also used to gather data from other Stock market news forums.

Table I: Sample Dataset for Apple Company Stock

DATE	*HIGH*	*LOW*	*OPEN*	*CLOSE*	*VOLUME*
12-09	226.42	222.86	224.8	223.09	32226700
13-09	220.70	217.09	220	218.75	39763300
14-09	220.13	217.56	217.73	219.9	21158100
15-09	220.82	219.12	219.96	220.7	18318700

B. DATA PREPROCESSING

The data gathered for Sentiment Analysis from Twitter and other news forums is processed using Natural Language Processing (NLP). The data that is in the form of text is split into words. These individual words are compared with a dictionary of words containing human-defined lexicons to categorize the words as positive, negative or neutral. Other special symbols and general emoticon symbols are also extracted from the data and are further used to classify a particular text from the data as positive, negative or neutral. The TextBlob library and Natural Language Processing Tool Kit (NLPTK) are used to determine the Polarity and Subjectivity of the particular texts. After the polarity and subjectivity are determined from a particular text these values are stored in a dataframe which is a multi-dimensional array.

The data gathered from the historical dataset is processed for missing data values or null values. Data Visualization is done by plotting graphs and plots for the given dataset using the Mathplotlib library. This data visualization gives an idea about the data and if any values are missing or null. The rows of data with missing values are either excluded from the dataset or Linear Regression is used to predict the missing values and null values. This data is then stored in a separate dataframe. The two dataframes are then combined to form a single multidimensional dataframe to be given as an input to the Hybrid and LSTM models.

The polarity values from the news sentiment analysis is added as a separate row in the historical dataset table. This makes the results of the news Sentiment Analysis an extra feature that will be used to train the Hybrid and LSTM models.

Table II: Apple Company Stock Input Dataset with the Polarities (POL.) Added as a Separate Column in the Dataset

DATE	*HIGH*	*LOW*	*OPEN*	*CLOSE*	*VOLUME*	*POL.*
12-09	226.42	222.86	224.8	223.09	32226700	0
13-09	220.70	217.09	220	218.75	39763300	0.0166
14-09	220.13	217.56	217.73	219.9	21158100	0
15-09	220.82	219.12	219.96	220.7	18318700	-0.184

IV. IMPLEMENTATION METHODOLOGY

The developed model is implemented in the following steps:

- Implementing Sentiment Analysis
- Creating the Hybrid Model
- Implementing LSTM Algorithm

A. Implementing Sentiment Analysis

The algorithm was implanted using the TextBlob Library for Sentiment Analysis in the following steps:

I. Register for Twitter API
II. Import the libraries and CSV files
III. Collect news and texts related to the particular company from Twitter using the API
IV. Using TextBlob determine the polarity and subjectivity of each news line
V. Store the data in a pandas dataframe as well as in a CSV file

First, we get a Twitter API key to collect data. Then using the "Tweepy" library we access the Twitter API.

Using the TextBlob library functions we determine the polarity and subjectivity values for each news line. These values are stored in a pandas dataframe and in an offline CSV file for further processing.
These values are added as a single row in the historical stock market dataset and this dataset will be given as further input to the SVM and LSTM models.

B. Creating the Hybrid Model

The Hybrid model combines an SVM classifier, a KNN classifier and a Random Forest classifier using the Majority Voting Ensemble classifier in python using Keras and TensorFlow backend in the following steps

I. Import the libraries and packages
II. Register for the Yahoo Finance API
III. Import the S&P 500 dataset
IV. Import the datasets containing historical data or take minute by minute data using the Yahoo Finance API
V. Create an SVM Classifier
VI. Create a KNN Classifier
VII. Create a Random Forest Classifier
VIII. Initialize an Ensemble Voting Classifier
IX. Pass the all the created Classifiers as parameters to the Voting Classifier
X. Get the combined predicted output of the Voting classifier
XI. Calculate the accuracy and confidence score

The various libraries and packages are installed. The Yahoo Finance API is used to gather company stock price historical data or minute by minute data. The S&P500 dataset is used to determine the company stock code. Create different SVM, KNN and Random Forest classifiers on the subsets of the same sample dataset. Create the Ensemble Voting Classifier to implement the Majority Voting algorithm to combine all the created classifiers.

C. Implementing LSTM Model

The LSTM Algorithm is implemented in Python using Keras and TensorFlow backend in the following steps:

I. Import the libraries and packages
II. Import the previously created dataset
III. Convert the dataset into a Numpy array
IV. Split the dataset into training and testing set
V. Define a scalar to normalize the data
VI. Define the various layers of the LSTM network
VII. Implement the LSTM function and compile all the created layers into one LSTM network
VIII. Feed the LSTM output to the final regression layer to get the final predicted output
IX. Calculate the prediction accuracy and confidence score

The above algorithm first imports the previously created dataset with historical data and Sentiment Analysis results. This data is stored in a Numpy array. This data is then split into training and test set. The values are normalized to scale them in a uniform range. The various layers of the LSTM network are created. This data is then fed into the three-layer LSTM and the LSTM output is fed into the final regression layer to get the output.

V. RESULT AND ANALYSIS

The implementation methodology was carried out using the historical stock price datasets of companies from December 1999 to September 2019. The sentiment analysis results of the twitter news headlines related to the required company's stock were also combined with historical dataset to give our final sample input dataset.

A. Hybrid Model Result

The Hybrid model gave an impressive accuracy while predicting the daily stock prices but gave much better results in the real-time prediction of the stock market using minute-by-minute stock price data. The model gave the highest accuracy of 92.57% while predicting real-time next minute stock prices while it gave the highest accuracy of 89% while predicting for the daily prediction model. The major notable achievement of this model is its consistency in predicting future stock prices with great accuracy.

Fig. 2: Apple Stock Real Time Predicted Results using Hybrid Model

Figure 3: Amazon Stock Daily Prediction Results using Hybrid Model

B. LSTM Model Result

The LSTM model showed increased accuracy as we increase the number of epochs and reduce the batch size. LSTM gave high accuracy results of 86% while predicting daily stock price data for up to the next 7 days. Whereas the LSTM model performs poorly while predicting next-minute data with an accuracy score of 72%. The model performs excellently while forecasting market trends for up to 7 days. The drawback of this model is it gives inconsistent results with slight variations each time it is used.

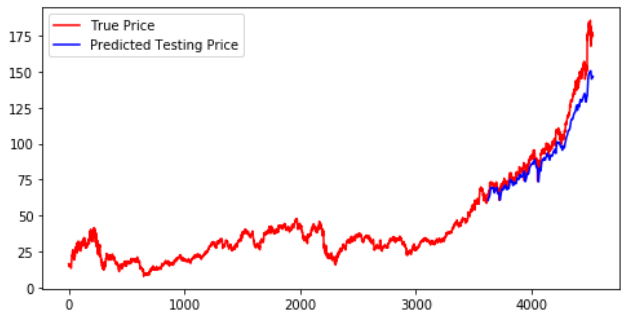

Fig. 4: Adobe Company Stock Daily Prediction using LSTM Model

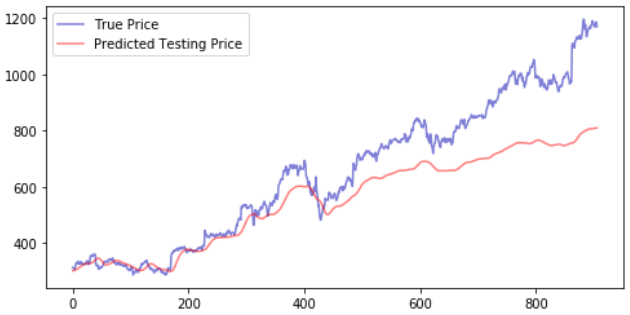

Fig. 5: Amazon Company Real-Time Stock Prediction using LSTM Model

Fig. 6: Reliance Company Daily Predicted Stock using LSTM Model

VI. CONCLUSION AND FUTURE WORK

The stock market is a very complex and volatile platform that is affected by various diverse factors and changes rapidly. Over the years many Machine Learning (ML) and Deep Learning (DL) algorithms have been used to predict the direction of the stock market. Many of these algorithms gave a descent prediction accuracy and made predictions based only on the historical stock market data. Another popular approach to predict the stock market trend is by carrying-out Sentiment Analysis on a company's news related to its stock to determine how the news will affect the stock market. This is implemented using Natural Language Processing (NLP) tools.

This work proposes two models to predict the stock market data. The first model is a Hybrid Model combining popular ML algorithms like SVM, KNN and Random Forest algorithms using the Majority Voting Algorithm. The second model utilizes backpropagation optimized LSTM networks to predict stock market trends. Both models have proven to be remarkable in predicting the direction of the stock market. The Hybrid model predicts the next-minute and daily stock prices with a very impressive accuracy and consistency. The LSTM model predicts stock prices accurately for the fore-coming days but gives decent accuracy in predicting from the real-time stock data. The major achievement of both models is that they use both the historical company stock price data as well as the news affecting the company stock prices to forecast the future stock prices of the given company.

The future-work for this project would be to create a hybrid model combing both the LSTM model and ML Hybrid model using a stacking algorithm to predict stock market trends with accuracy and efficiency consistently.

REFERENCES

1. Jui-Sheng Chou and Thi-Kha Nguyen, "Forward Forecast of Stock Price Using Sliding-window Metaheuristic-optimized Machine Learning Regression", DOI 10.1109/TII.2018.2794389, IEEE Transactions on Industrial Informatics
2. Min Wen, Ping Li, Lingfei Zhang, and Yan Chen, "Stock Market Trend Prediction Using High-Order Information of Time Series", date of publication February 26, 2019, date of current version March 18, 2019. 10.1109/ACCESS.2019.2901842
3. Yongsheng Ding, Lijun Cheng, Witold Pedrycz, and Kuangrong Hao, "Global Nonlinear Kernel Prediction for Large Data Set With a Particle Swarm-Optimized Interval Support Vector Regression", Ieee Transactions On Neural Networks And Learning Systems, Vol. 26, No. 10, October 2015
4. L. Minh Dang, Abolghasem Sadeghi-Niaraki, Huy D. Huynh, Kyungbok Min And Hyeonjoon Moon," Deep Learning Approach for Short-Term Stock Trends Prediction based on Two-stream Gated Recurrent Unit Network", DOI 10.1109/ACCESS.2018.2868970, IEEE Access
5. Lei Shi, Zhiyang Teng, Le Wang, Yue Zhang, and Alexander Binder, "DeepClue: Visual Interpretation of Text-based Deep Stock Prediction", DOI 10.1109/TKDE.2018.2854193, IEEE Transactions on Knowledge and Data Engineering
6. Guang Liu And Xiaojie Wang, "A Numerical-based Attention Method for Stock Market Prediction with Dual Information", 10.1109/ACCESS.2018.2886367, IEEE Access
7. Rashmi Sutkatti, Dr. D. A. Torse, "Stock Market Forecasting Techniques: A Survey", Volume: 06 Issue: 05 | May 2019, International Research Journal of Engineering and Technology (IRJET)
8. Divit Karmaini, Ruman Kazi, Ameya Nambisan, Aastha Shash, Vijaya Kamble, "Comparison of Predictive Algorithms: Backpropagation, SVM, LSTM and Kalman Filter for Stock Market", 10.1109/AICAI.2019.8701258, IEEE
9. Priyamvada, Rajesh Wadhvani, "Review on various models for time series forecasting", Inventive Computing and Informatics (ICICI) International Conference on, pp. 405-410, 2017.
10. Aparna Nayak, M. M. Manohara Pai∗ and Radhika M. Pai, "Prediction Models for Indian Stock Market", Elsevier, ScienceDirect

AUTHORS PROFILE

SHASHANK SINGH is currently pursuing B.Tech. Degree in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai. He has completed his Secondary Education from Amity International School, Mayur Vihar, Delhi. He has completed several projects in the field of Machine Learning and Deep Learning.

MAAZ AHMAD is studying Computer Science and Engineering in SRM Institute of Science and Technology, Chennai. He has completed his Secondary Education from DAV, Kabirmath Kandwara, Siwan, Bihar.

ADITYA BHATTACHARYA has completed his secondary education from Sri Chaitanya Institute, Vishakhapatnam, Andhra Pradesh. He is completing his Computer Science B.Tech. Degree from SRM Institute of Science and Technology, Chennai

M. AZHAGIRI has completed his B.E. Information Technology from Vinayaka Missions University and M.E. Computer Science and Engineering from Anna University and currently pursuing his P.hd. in St. Peter's Institute of Higher Education and Resource and working as Assistant Professor in Computer Science and Engineering at SRM Institute of Science and Technology, Chennai.