Spredictor: A Stock Analysis based on Sentiment, News and Historical Data

Luv Michael Surve
Dept. of Comp Science & Engineering Mangalam College of Engineering, Kottayam, India.

Vishnu Sivan
Dept. of Comp Science & Engineering Mangalam College of Engineering, Kottayam, India.

Yedukrishnan J
Dept. of Comp Science & Engineering Mangalam College of Engineering, Kottayam, India.

Yethi J Nair
Dept. of Comp Science & Engineering Mangalam College of Engineering, Kottayam, India.

Neena Joseph
Dept. of Comp Science & Engineering Mangalam College of Engineering, Kottayam, India.

***Abstract*— For numerous years, stock market analysis and prediction tools have been widely used, with various methodologies and models to accurately anticipate stock markets. The design and implementation of a technique for predicting stock market movements are presented in this project. Models are used in this method to determine whether or not to invest in a stock. To learn and forecast the future market trend, researchers used Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM). To forecast stock trends, sentiment analysis is undertaken for a certain stock using newsfeeds and tweets. Long-term and short-term predictions are provided by combining the two models.**

Keywords*— *Machine learning, LSTM model, Stock Market Prediction, Historical data, Sentimental analysis

I. INTRODUCTION

A stock market, also known as an equity market or a share market, is a gathering of buyers and sellers of stocks (also known as shares), which represent ownership claims on businesses. Stocks can be traded publicly or privately, such as shares of private companies sold to investors through equity crowdfunding platforms. The majority of investments are made with a strategy in mind. The goal of stock market prediction is to predict the future movement of a financial exchange's stock value. Investors will be able to make more money if they can accurately predict share price movement. Information overload, over-diversification, bad timing, and improper direction are some of the issues that new investors confront. Financial services organisations and corporations are embracing data analytics technologies to gain insights into stock market movements and make impactful decisions for their business operations. The stock market is extremely dynamic, as thousands of transactions and events occur every second throughout the world, influencing the market's numbers and figures. Investors can examine data using complex mathematical formulas and algorithms that are input into a computer using big data. The word "market sentiment" often known as "investor sentiment" refers to an investor's overall outlook or attitude toward a certain security or the entire financial market. The aggregate price trends reflect the market participants' optimism or pessimism. The results of [1] prove further that sentiment analysis is a key technique beneficial for efficient stock analysis for short-term predictions.

The sources of public opinions and sentiments include social networking websites such as twitter.com and financial news headlines, which too influence public sentiments about a stock. This study suggests using historical stock data and public attitudes about a particular stock from a variety of sources to train an LSTM model that can analyse stock prices and give findings that can be used to advise users on investment opportunities.

II. DATA ACQUISITION

The data used in this project consists of historical data, public opinions in the form of tweets, and news related to financials concerning particular stocks. The historical data is retrieved through Yahoo Finance libraries, which involves various fields such as closing and opening prices, daily volumes dealt and more. This involves collecting data for more than 10 years to get an effective understanding of trends. The twitter tweets in relation to a particular set of keywords which provide more insights about a particular stock are used with the help of the social networking service scraper (snscrape), which accesses twitter tweets through the twitter API without any restrictions or limits. The news feeds such as news headings, contents, and descriptions relating to a company's recent times are searched and retrieved through the news-api.

III. DATA PROCESSING

A. Historical Model

The two LSTM models involved in this project require their compatible feature sets, the first of which is

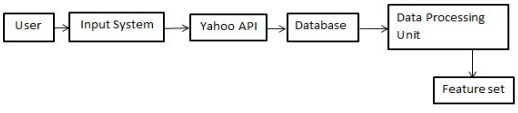

Figure 1. Feature set development for historical data model

it is more focused on analysing closing price trends based on historical data alone and requires training with the closing price data of stock datasets. The closing prices of datasets are isolated and split into training and testing sets in the partitions of sizes 70% and 30% respectively. The training is of the supervised type where the labels involve 60 days of closing prices to produce the 61st day's target value.

B. Sentimental Based Model

The second LSTM model, accounting for the impulsive events that may justify the movement of a particular stock trend, requires a feature set with labels for closing prices and sentiments regarding the stocks from retrieved tweets and news headings. The tweets and news headlines are cleaned to remove any symbols, redundant characters, irrelevant data, noise data, and others. The data is then processed through a sentiment-analyzer such as the VADER sentiment analyser. The historical data and sentiment data similar to the previous model are combined and reshaped to train the model. All values are normalised between 0 and 1 using the

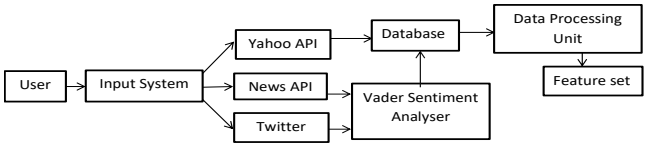

Figure 2. Feature set development for sentiment-based model.

IV. Model Description

A. Historical Model

Both the models are sequential LSTM models with a "ReLu" activation function. The models include four LSTM layers and a dense layer. The LSTM layers include 50, 60, 80, and 120 units, respectively, with dropout values of 0.2, 0.3, 0.4, and 0.5, respectively. The return sequence is set to true for the first 3 layers to maintain the output sequence of the same length.

The models are compiled using the Adam optimizer and with a loss function of MSE. The first model is trained with 70% of the dataset and 50 epochs, whose feature set includes closing price and, upon validation, produces an RMSE value of 11.314 when trained and tested with the dataset from Apple Inc. spanning ten years.

(RMSE 11.941 - TSLA)

(RMSE 6.594 - TWTR)

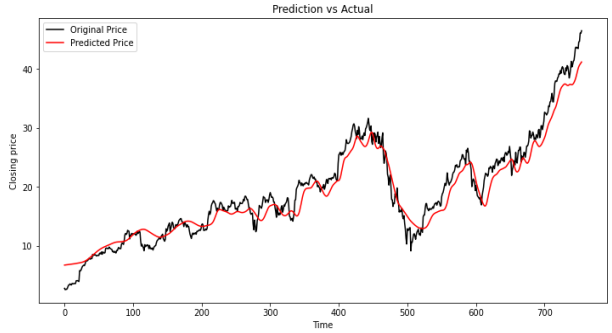

Figure 3. Prediction vs actual-AAPL

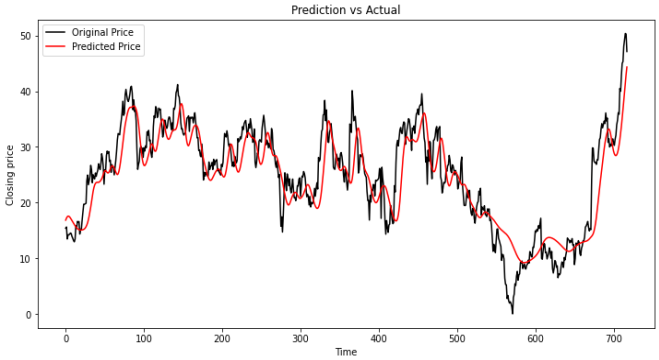

Figure 4. Prediction vs actual-TSLA

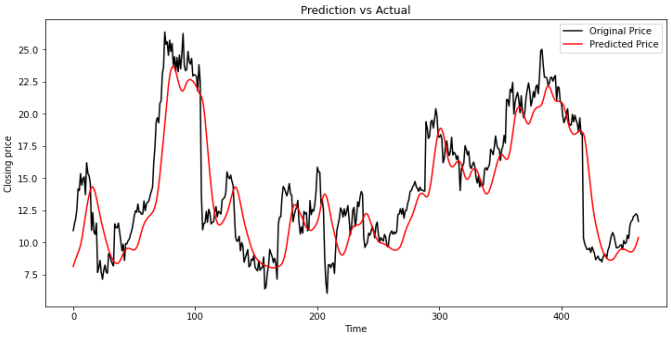

Figure 5. Prediction vs actual-TWTR

B. Sentiment Based Model

The second model is trained with 70% of the dataset with 100 epochs and its feature set includes closing price along with sentiments of a particular day. When trained and tested with 10 years of Tesla, Inc. datasets, it produces an RMSE value of 0.29.

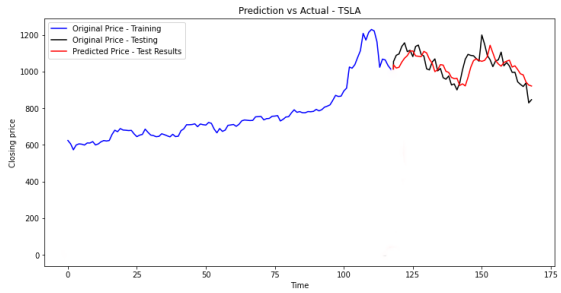

Figure 6. Prediction vs actual-TSLA

V. System Architecture

A. Input Manager

The system's Input management module is responsible for providing correct and compatible input feature sets to machine learning models in order to generate predictions.

It consists of two sub-modules:

1) Data Filter Unit

The data filter is responsible for filtering out unnecessary attributes from the input dataset and narrowing down the input dataset to the required attribute sets for further preprocessing activities.

2) Data Processing Unit

The data processing unit is responsible for various preprocessing activities such as data cleaning, normalization, data reshaping suitable for models, processing text data to sentiment values for sentiment-based models, and other necessary data processing.

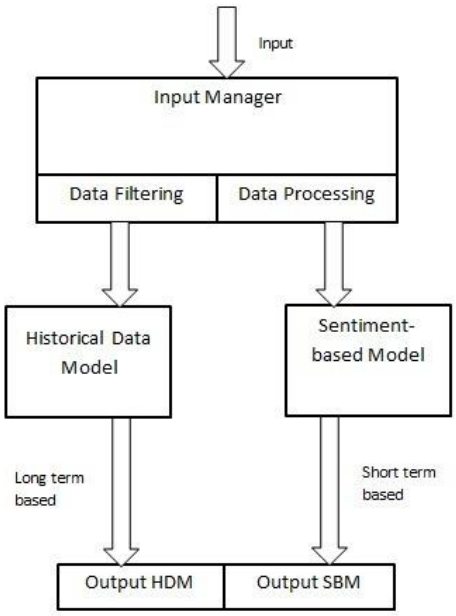

Figure 7. System Architecture

B. Historical Data Model

The historical data model is the one making use of LSTM networks and takes in the feature set with one feature "Closing price" of the shape (1,60,1), which is an array of closing price values of the past 60 days. The feature set is fed to the model, and it generates the prediction for the next day.

C. Sentimental Based Model

The sentiment-based model employs LSTM networks and includes two features: "Closing price and sentiments" of the shape (1,3,2), which is an array of closing price and sentiment values from the previous three days. The sentiment values must be generated by utilising the Vader-Sentiment analyzer to process tweets and news feeds over the previous three days. The model is fed the feature set, and it forecasts the next day's closing price.

D. Output

The sentiment-based model employs LSTM networks and includes two features: "Closing price and sentiments" of the shape (1,3,2), which is an array of closing price and sentiment values from the previous three days. The sentiment values must be generated by utilising the Vader-Sentiment analyzer to process tweets and news feeds over the previous three days. The model is fed the feature set, and it forecasts the next day's closing price.

VI. RESULT

The two LSTM models discussed in this paper are used to predict long-term and short-term trends, where the historical data model is more efficient in predicting long-term trends and the sentiment-based model is more suited for short-term trend predictions, which account for impulsive events.

VII. CONCLUSION

This paper proposes the use of two LSTM models that analyse historical data as well as public sentiment data to produce long-term and short-term forecasts that present the users with a calculated and statistical approach while investing in the stock market. The LSTM models have a memory element, unlike some artificial neural network models, which don't keep track of the entire dataset while training. The addition of sentiment parameters sourced from Twitter tweets and news headlines provides good recommendations while investing for short-term periods. The combined use of the suggested models will allow users to maintain short-term as well as long-term portfolios.

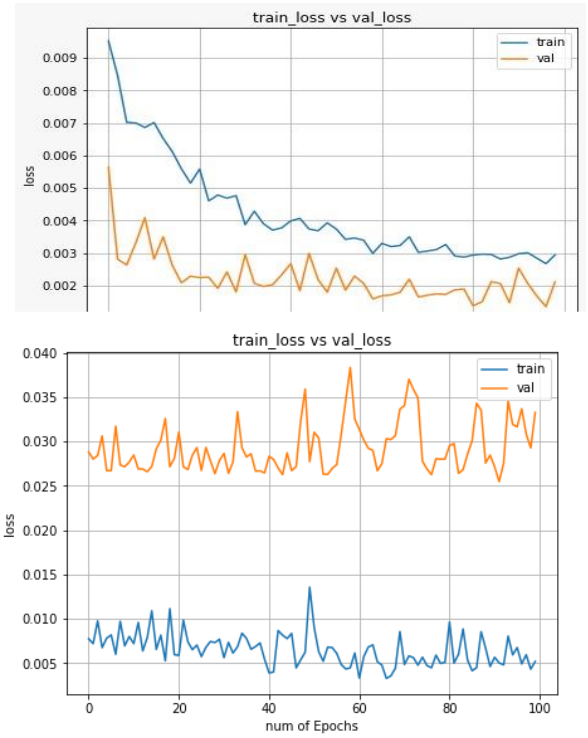

Figure 9. Sentiment based training loss vs validation loss

REFERENCES

[1] Yin Ni, Zeyu Su, Weiran Wang, Yuhang Ying, A novel stock evaluation index based on public opinion analysis, Procedia Computer Science, vol 147, 2019, pp.581-587

[2] Parmar et al., "Stock Market Prediction Using Machine Learning," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 574-576.

[3] T. Mankar, T. Hotchandani, M. Madhwani, A. Chidrawar and C. S. Lifna, "Stock Market Prediction based on Social Sentiments using Machine Learning," 2018 International Conference on Smart City and Emerging Technology (ICSCET), 2018, pp. 1-3.

[4] Sreyash Urlam,Bijit Ghosh,Dr.A.Suresh, "Stock Market Prediction Using LSTM and Sentiment Analysis" Turkish Journal of Computer and Mathematics Education,vol 12, 2021.

[5] Rouf, N.; Malik, M.B.; Arif, T.; Sharma, S.; Singh, S.; Aich, S.; Kim, H.-C. Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. Electronics 2021, 10, 2717.

[6] S. Liu, G. Liao and Y. Ding, "Stock transaction prediction modeling and analysis based on LSTM," 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2018, pp. 2787-2790, doi: 10.1109/ICIEA.2018.8398183..
[7] V. S. Pagolu, K. N. Reddy, G. Panda and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 2016, pp. 1345-1350, doi: 10.1109/SCOPES.2016.7955659..