2022 2nd International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 24-26, 2022

Stock Price Prediction Using News Sentiment Analysis

Lakshya
Software Engineering
Delhi Technological University
Delhi, India
lakshyad2879@gmail.com

Prateek
Software Engineering
Delhi Technological University
Delhi, India
prateeksingh181999@gmail.com

Dr. Divyasikha Sethia
Software Engineering
Delhi Technological University
Delhi, India
sethiadivya@gmail.com

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9847747

***Abstract*—In this era of digitization, tasks can be performed from anywhere in the world that previously required manual movement. It is the same for investing and trading in stocks. With the ease of investing and trading in stocks via the internet, a more extensive segment of society has started investing. The stock price depends on multiple factors such as politics, economics, war, society, and news sentiment. Therefore stocks are really hard to predict due to such vast dependencies. Stock markets are an important issue in the financial world. Prediction of stock prices during the global pandemic of Novel Coronavirus 2019 (COVID-19) can be very helpful to stakeholders. The attempt of predicting the stock prices have been made by previous researchers using sentimental news analysis through Support Vector Machine (SVM), Neural Network, and Naive Bayes. However, they have low accuracy, and some even claim that news is not a crucial governing factor for the stock price. This paper aims to predict the stock market prices through news sentimental analysis using techniques such as Long Short Term Memory and Artificial Neural Network against classifier models like Natural Language Toolkit, Valence Aware Dictionary for Sentiment Reasoning, Recurrent Neural Network for price prediction. S.Mohan [1] MAPE scores came out to be 1.17, 2.43 for RNN and RNN with news polarity for Facebook stock prices. Our results came out to be 1.21 and 1.94, slightly better results, thus showing optimism in the dependence of stock prices on the news.**

***Index Terms*—Deep Learning, Artificial Neural Network, Stock Market, Sentiment Analysis, Recurrent Neural Network (RNN), Valence Aware Dictionary for Sentiment Reasoning(VADER), Natural Language Toolkit (NLTK).**

I. Introduction

Stock prices being highly volatile and unpredictable makes it more important to get the most out of historical data to make reasonable predictions on the stock's price. Through the comfort of home, one can easily sell and buy stocks in real-time within seconds. The Stock market is very unpredictable, but with utmost study and statistics, a probabilistic approximation can be made concerning the nature of stock price and its magnitude. We have considered the stock's closing price each day as the news impact would be visible till the day's end. The stock price depends on many factors such as politics, economy, society, and trivial to trivial activities going across the world. Therefore, it is challenging to estimate the stock market due to a large dependency. Multiple efforts have been made to predict stock prices using news, which gives a positive upfront to show the relation between stock prices and news sentiment, whereas some researchers showed that these are nearly independent. S. Mohan et al. [1] attempted to predict stock prices using time series analysis and used ARIMA, Facebook prophet, and RNN-LSTM models with price, price and news polarity, price and text, and multivariate models with MAPE for relative error. And we applied multiple news sentiment polarities generating techniques and further used RNN with price, price with news polarity to exact stock price prediction.

The research objective is to make a well-quantified prediction of the nature of stock price and the magnitude of stock price depending on the news, the sentiment of news, and the category it belongs to.

Research has been done using multiple models Support Vector machines, Neural networks, and Naive Bayes. In all these previous attempts, news polarity generating techniques have been used to predict exact stock prices without considering various polarity generating techniques for sentimental analysis of news. Exact stock price prediction becomes error-prone when a fixed sentiment analysis technique is applied directly to predict the magnitude of stock price change. Therefore, this research paper proposes a novel approach to figure out the best-suited polarity generation technique so that the nature of stocks, whether increasing or decreasing, can be determined efficiently. Without considering the magnitude, the model focuses on predicting the nature of stock price, using binary classification through logistic regression and a Dense Neural Network (DNN). The model later uses particular polarity to predict actual stock price using Recurrent Neural network, a technique that is well suited for time series analysis.

Contributions: The contributions of this paper are as follow:

- *Nature of stock price*
  Stocks are highly fluctuating and depend on various factors, and news is one of the significant factors. Therefore, this paper aims to solve the problem of whether the stock price is going to increase or decrease using news sentiment. This Binary Classification problem laid our primary focus on Natural Language Processing. Since there could be multiple ways to find the polarity of news so this paper tries to find the best technique that can generate polarity that best fits in exact stock prediction. Polarity

generation is done using three techniques: Natural language toolkit(NLTK a pre-trained model), Long short-term memory (LSTM), and Valence Aware Dictionary for Sentiment Reasoning (VADER) as shown in Fig. 1. For binary classification, the output feature depicting the nature of stock is 0 if the closing price is lower than the opening price and 1 if the closing price is higher the opening stock price and applying multiple machine learning techniques for binary classification. The binary classification techniques are as mentioned in Table I

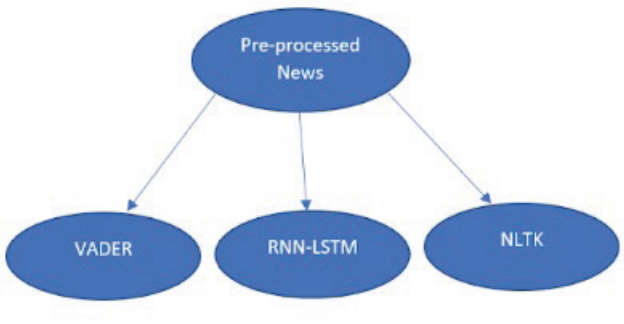

Fig. 1. Techniques for Polarity of news

TABLE I
MODELS

LR	Logistic regression with polarity (RNN-LSTM) as input
NN	Neural network with polarity (RNN-LSTM) as input
LR-n	Logistic regression with polarity (NLTK classifier) and as input
NN-n	Neural network with polarity (NLTK classifier) as input
LR-o	Logistic regression with polarity (Vader) as input
NN-o	Neural Network with polarity (VADER) as input
RNN -v	Recurrent Neural Network with VADER

- *Stock price Prediction.*
  The second target is to predict the stock price values as close as possible to the actual price using time series analysis and sentiment analysis of news. The polarity would be used of that technique that would perform best in the first problem statement. For exact stock price comparison:
  - RNN
  - RNN along with polarity

II. Related Work

S. Mohan et al. [1] predicted stock prices using time series analysis and used S&P's data set of 500 companies and 265463 news articles from February 2013 to March 2017. They removed duplicates and applied log transformation to the stock price on the data. They used ARIMA, Facebook prophet, and RNN-LSTM models with price, price and news polarity, price and text, and multivariate models with MAPE for relative error. Error measures by MAPE were 7.37, 14.08, 1.98, 1.12, 1.17, 5.93, respectively, for Apple stock prices.

A. Díaz et al. [2] observed positive stock response in case of "good news" and significant stock price decline in case of "bad news." Therefore, the news could determine the stock market price and behavior.

Yang Li et al. [3] proposed an ensemble learning technique and showed an increase in 33.4% accuracy, motivating further improvement in those techniques. John H. Boyd et al. [4] used SP 500 data set and showed the effect of unemployment news articles on Sentiment analysis.

M. Ballings et al. [5] studied the impact of multiple classifiers on stock market prediction and built their data set of 5767 European Companies. The authors applied seven techniques: Logistic Regression, Neural Networks, K-nearest neighbor, Support Vector Machines, Random Forest, AdaBoost, Kernel Factory, and the Area Under Curve (AUC)for accuracy. They achieved 0.66, 0.72, 0.72, 0.83, 0.90, 0.76, 0.79 accuracy respectively.

M. R. Vargas et al. [6] studied the impact of deep learning models on stock market prediction by comparing RNN and CNN with different inputs from data from SP 500 companies and newspaper articles. Their models used seven technical indicators derived from the financial newspaper title. The results showed that RNN with sentences as input has the best accuracy of 0.62 over the SP data set, and he also concluded that sentence embedding is better than word embedding; this was due to the sparsity of the data.

D. Shah et al. [7] shows the impact of news sentiment on stock market prediction, so they built a data set from the investing and money control website. They implemented a dictionary-based sentiment analysis model to determine the effect of the news on the stock price. They used an n-gram model and achieved 70.59% accuracy in predicting stock prices in the short term.

G. V. Atiigeri et al. [8] tried to classify stock prices from social media sentiment analysis, so they built a data set of different companies using Mozenda Web Crawler from twitter consisting of 3980 rows and news articles. Over the data, they removed stop words, URLs, and punctuation and then lemmatized the text. They used HDFS and logistic regression models and achieved 70

W. Chen et al. [9] conducted a study for time series analysis on stock prices and built a data set from the Sina Weibo website that was posted by blue verified accounts and newspapers, magazines. Recurrent Neural Network and logistic regression models are used to predict the result and error in the result is calculated by Mean absolute percentage error, Root Mean Square error and Mean Absolute Error.

Usman, Mehak et al. [10] conducted a study to classify the impact on Karachi Stock Exchange by traditional and social media. They gathered the data from relevant news and Twitter platforms over a period of three months. They implemented a dictionary for replacing the non-English words with their English meaning. The data was labeled with binary labels of positive and negative, describing the general sentiment of the

market of that closing day. They recorded the data from the market for the same period.

M. Dang et al. [11] studied the improvement method of stock market prediction using financial news. Using web crawler tools, they gathered the data from the vietstock, hsx, and hnx websites to collect 1884 different articles from May 01, 2014, to Apr. 30, 2015. They used a support vector machine to train over the news and finally achieved an accuracy of 73%.

C. Gondaliya et al. [12] for studying the impact of news on stock prices they gathered their news data from RSS Feed, Forum discussions, Twitter, and News portals. They compared six techniques: Decision Tree method, Random Forest method, Logistic Regression method, the Naïve Bayes method, Support Vector Machine method, and k-nearest neighbors, and achieved 78% accuracy by SVM and LR.

III. Dataset Description

A. Data Gathering

- Data is the core of every company and even research. It requires the stock price of 5 companies from 2013 to 2018: Apple, Amazon, Facebook, Tesla, Netflix. S and P 500 [1] Data set used for the stock price as S.Mohan et al. in stock price prediction.
- The data set comprises of scrapped 1,24,990 News articles with the headline category to which they belong for sentiment analysis. The Dataset has average of 112 news headlines per working day from the international news site. Considering the importance of sentiment and the category to which the news belongs, the category is also scrapped. Selenium scripting is used to scrap out news headlines and the new category to which it belongs, such as political, comedy, or any other category.
- For Binary classification, a dependent attribute named nature of stock is added, denoting (opening price - closing price) the nature of stock at a particular day, whether it is rising or dipping.
- The Closing stock price is used in our model for exact stock price prediction.

B. Data Preprocessing

Data Processing plays a critical role in our model results, so the first task is to club the stock price data with the news corresponding to the dates because stock market prices are not available for the day the stock market is closed. As we have decided to perform two types of model training.

- For Binary classification we added a new dependent attribute showing the difference in opening and closing price on a particular day. 0 denotes stock price downfall. 1 denotes a rise in stock prices. This would have been used in the binary classification of the nature of stock price.

[1]https://www.kaggle.com/camnugent/sandp500

Further One-hot encoding is applied on news categories.As these categories are without any ranks and are entirely independent of each other.
Simply closing stock prices were considered to predict exact stock prices.

- Data processing is involved in further steps for news sentiment analysis to find the polarity. The reprocessing involves tokenization, removing punctuation, stopping words, and finally stemming and lemmatization from the text.

IV. Research Methodology

A. RNN

A recurrent Neural Network is an ANN(Artificial Neural Network) in which nodes of adjacent layers are connected to form a direct or indirect graph that uses sequential or time-series data. RNN uses the previous layer's output as the input of the next layer as shown in Fig. 2.

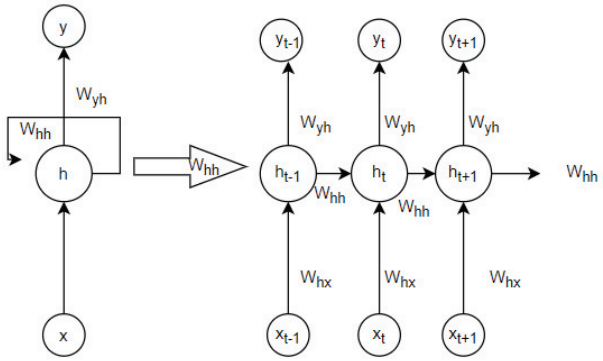

Fig. 2. Recurrent Neural Network

RNN takes a feedback loop to improve the result with each loop and provide better efficiency. It covers independent activation into dependent activation by giving all layers the same bias and weights.

B. Vader

Vader(Valance Aware Dictionary for Sentiment Reasoning) is a sentiment analysis model of the nltk package that gives both positive and negative polarity alongside its emotional strength. It provides the sentiment score by adding the score of individual words of the text. It focused on five major points Punctuation, Capitalization, Degree modifiers, Conjunctions, and Preceding Tri-gram.

C. DNN

A Deep Neural Network is an artificial neural network containing input and output layers and at least two hidden layers. Data flow in a deep neural network is always in the forward direction from the input to output layer without

backward propagation as shown in Fig. 3. It does not contain a loop like a Recurrent Neural Network.

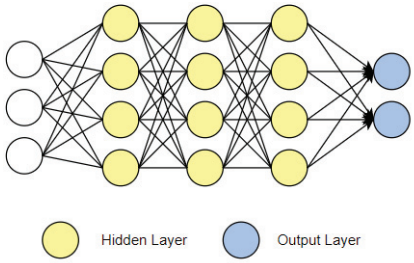

Fig. 3. DNN Architecture

D. Nltk

Natural Language Toolkit is a python based platform to process natural language. It contains text processing libraries that are very helpful with classification, tokenization, tagging, parsing, and sentiment reasoning. Architecture structure of Natural Language Toolkit is shown in fig 4.

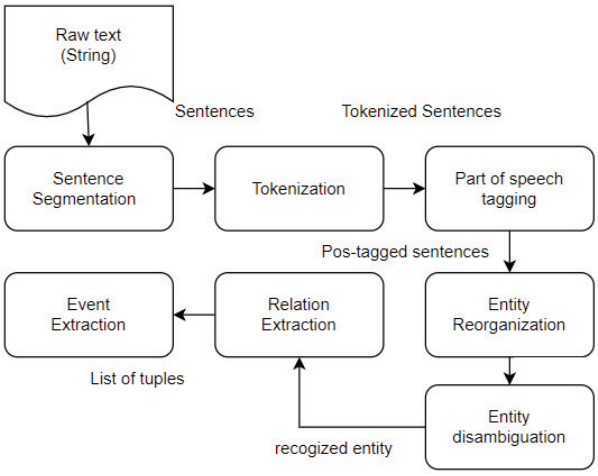

Fig. 4. NLTK Architecture

V. Performance Measures

A. Confusion Matrix

It is a table that describes or provide a brief overview of how good the model is working. It shows us quantitative values of True Positive(TP), False Positive(FP), False Negative(FN), True Negative(TN) as shown in Fig. 5 that can help in developing the notion about our model's performance.

	Actual Values	
Predicted Values	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True negative (TN)

Fig. 5. Confusion Matrix

B. Accuracy

It is the ratio of the correctly predicted test set instances to the total number of test set instances.

$$Accuracy = (TP+TN)/(TP+FN+FP+TN) \quad (1)$$

C. F1 Score

The F1 score can be said as the harmonic mean of the precision and recall, Higher the F1 score means higher performance by our model. Precision and recall values are used in order to calculate F1 score. The formula of F1 score is:

$$F1 = 2*(precision*recall)/(precision+recall) \quad (2)$$

D. Area Under the Curve (AUC)

The AUC means the area under curve of the ROC plotting.ROC means Receiver Operator characteristics. ROC curve is a curve between sensitivity and specificity. Sensitivity on y coordinate and specificity on x.AUC performance metric is best suited for bias dataset where other parameters like F1 score fail to figure out the appropriate outcome.

E. Mean Absolute percentage error (MAPE)

MAPE is a popular measure of prediction accuracy of the forecast system. It measures accuracy by calculating the average mean error of all time minus actual value divided by the actual value.

$$MAPE = \frac{100}{n}\sum\frac{y-y^{'}}{y} \quad (3)$$

The above equation shows the formula for MAPE in which n represents the number of data points, y' represents forecast value, y represents actual value. Multiplying by 100 converts it to percent.

VI. Results And analysis

The Data is limited to only five years; therefore, the holdout cross-validation technique was used. The results are shown in the Table II, III and IV are of performance measure F1 Score, Accuracy and AUC score respectively .

A. Binary Classification

TABLE II
F1-Score

	APPLE	TESLA	FACEBOOK	NETFLIX	AMAZON
LR	0.8212	0.8167	0.8722	0.9111	0.8789
NN	0.9127	0.9228	0.9158	0.9311	0.9269
LR-n	0.7514	0.9123	0.9222	0.7312	0.8138
NN-n	0.7676	0.9224	0.7656	0.8556	0.8977
LR-o	0.8524	0.8868	0.8272	0.9131	0.9098
NN-o	0.9127	0.9228	0.9158	0.9311	0.9269

TABLE III
Accuracy

	APPLE	TESLA	FACEBOOK	NETFLIX	AMAZON
LR	0.8212	0.8167	0.8722	0.9111	0.8789
NN	0.9127	0.9228	0.9158	0.9311	0.9269
LR-n	0.7514	0.9123	0.9222	0.7312	0.8138
NN-n	0.7676	0.9224	0.7656	0.8556	0.8977
LR-o	0.8524	0.8868	0.8272	0.9131	0.9098
NN-o	0.9127	0.9228	0.9158	0.9311	0.9269

TABLE IV
AUC

	APPLE	TESLA	FACEBOOK	NETFLIX	AMAZON
LR	0.8921	0.8469	0.9022	0.8021	0.8559
NN	0.8972	0.8978	0.9036	0.9148	0.9367
LR-n	0.8212	0.8167	0.8722	0.9111	0.8789
NN-n	0.9127	0.9228	0.9158	0.9311	0.9269
LR-o	0.9191	0.9112	0.8556	0.8663	0.9042
NN-o	0.9177	0.9298	0.9285	0.9131	0.9129

The stock price can be estimated using the news, and it is proportional to the sentiment of the news. Accuracy and the F1 were nearly the same as seen in Table II, III but gave a good glimpse of which model performed better. In Table IV, VADER polarity and the deep neural network gave the most satisfactory results with an average accuracy of 0.9218 overall for the five companies. Neural network with NLTK classifier and LR with VADER gave nearly the same results with an average accuracy of 0.9123 and 0.9127, respectively. Therefore VADER generated polarities are used in exact stock price prediction as it has the max AUC score. These show positive results and give more accurate information to predict the exact stock price.

B. Stock Price Prediction

Mean Absolute Percentage Error (MAPE) is used to calculate deviation from the original stock price. Higher MAPE shows lower model accuracy towards predicting the stock price value. As the VADER + ANN classifier gave best results, polarity generated through VADER was used in the stock price prediction.

TABLE V
MAPE

	APPLE	TESLA	FACEBOOK	NETFLIX	AMAZON
RNN	4.1	1.78	1.21	2.37	1.89
RNN-v	3.89	1.38	1.94	1.85	1.67

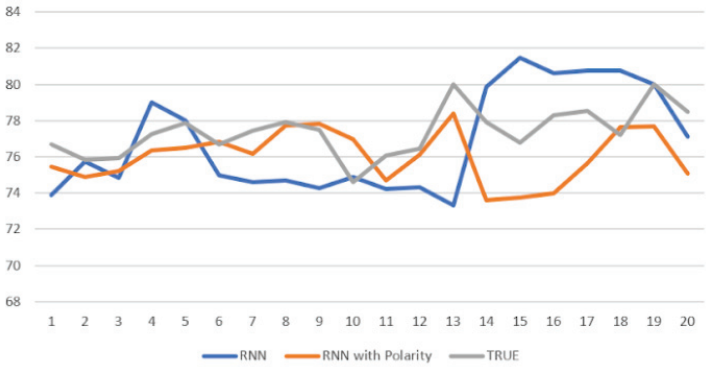

Fig. 6. Facebook stock Price

Smaller MAPE means the slightest deviation from the exact trajectory of the stock price. MAPE scores can be seen in Table V RNN and news polarity gives the best results with an average of 2.1 MAPE scores. Simple Time series analysis using RNN results were quite good with an average MAPE score of 2.4, but it did not trace the nature of the actual stock price, overlooking the magnitude as seen in Fig. 6. There are cases where it predicted to be rising but stock prices dipped. S.Mohan's MAPE score came out to be 1.17 and 2.43 for RNN and RNN with news polarity respectively for Facebook stock prices. Our results came out to be 1.21 and 1.94, slightly better results, thus showing optimism in the dependence of stock prices over the news. Overall The results obtained were :

- Among the three Binary Classification techniques, VADER along with ANN produces the best score. Denoting VADER polarities were more accurate when predicting the stock market.
- Deep Learning performed better as compared to logistic regression denoting that Deep Learning could be of immense use in further studies for stock market prediction.
- Simple RNN could not accurately predict the stock price, whereas RNN when clubbed with news polarity generated through VADER, is highly important in stock price prediction.

VII. Threats to Validity

A. Construct Validity

We have applied holdout cross-validation method. Applying other validation technique might have given some other results.

B. External Validity

On average, 112 news are used, which is still limited due to the number of news in each category. The news site portray their news differently and there are always multiple aspects of the same news.

C. Conclusion Validity

AUC being the best in class performance measure for imbalance dataset, but still there is always some scope of improvement. Thus it can be a threat to conclusion validity.

VIII. Conclusion and Future Work

This paper attempts to solve the stock market prediction problem and the expected price using the news sentiment. The results shows that though the stock market is complicated, news sentiment plays a significant role in governing its price. Further, along with news, multiple factors could be considered, such as the country's GDP increase rate and oil price. Correlation and multivariate models can also be used to predict stock prices. The company's future plan and quarterly stock prices will also be helpful in predicting the stock prices.

References

[1] S. Mohan et al. "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, Apr. 2019.

[2] A. Díaz et al. "Explanatory factors of the inflation news impact on stock returns by sector: The Spanish case," Research in International Business and Finance, vol. 23, no. 3. Elsevier BV, pp. 349–368, Sep. 2009

[3] Y. Li et al. "A novel ensemble deep learning model for stock prediction based on stock prices and news," International Journal of Data Science and Analytics. Springer Science and Business Media LLC, Sep. 17, 2021.

[4] J. H. Boyd et al. "The Stock Market's Reaction to Unemployment News: Why Bad News Is Usually Good for Stocks," The Journal of Finance, vol. 60, no. 2. Wiley, pp. 649–672, Mar. 02, 2005.

[5] M. Ballings et al. "Evaluating multiple classifiers for stock price direction prediction," Expert Systems with Applications, vol. 42, no. 20. Elsevier BV, pp. 7046–7056, Nov. 2015.

[6] M. R. Vargas et al. "Deep learning for stock market prediction from financial news articles," 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). IEEE, Jun. 2017.

[7] D. Shah et al. "Predicting the Effects of News Sentiments on the Stock Market," 2018 IEEE International Conference on Big Data (Big Data). IEEE, Dec. 2018.

[8] G. V. Attigeri et al. "Stock market prediction: A big data approach," TENCON 2015 - 2015 IEEE Region 10 Conference. IEEE, Nov. 2015.

[9] W. Chen et al. "Stock market prediction using neural network through the news on online social networks," 2017 International Smart Cities Conference (ISC2). IEEE, Sep. 2017.

[10] M. Usmani et al. "Stock market prediction using machine learning techniques," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS). IEEE, Aug. 2016

[11] M. Dang et al. "Improvement methods for stock market prediction using financial news articles," 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS). IEEE, Sep. 2016.

[12] C. Gondaliya et al. "Sentiment analysis and prediction of Indian stock market amid Covid-19 pandemic," IOP Conference Series: Materials Science and Engineering, vol. 1020. IOP Publishing, p. 012023, Jan. 16, 2021.