Stock Market Predictions Using Machine Learning Techniques

Nagapoojitha D N,
Student, 4th SEM , 1DS22CB034,
Department of Computer Science and Business Systems, Dayananda Sagar College of Engineering.

ABSTRACT

Accurately predicting stock market prices is vital in today's economy, leading researchers to explore novel approaches for forecasting. Recent studies have shown that historical stock data, search engine queries, and social mood from platforms like Twitter and news websites can predict future stock prices. Previous research often lacked comprehensive data, especially concerning social mood. This study presents an effective method to integrate multiple information sources to address this gap and enhance prediction accuracy. We utilized Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) models to analyse individual data sources. To further improve prediction accuracy, we employed an ensemble method combining Weighted Average and Differential Evolution techniques. The results yielded precise forecasts for one-day, seven-day, 15-day, and 30-day intervals, providing valuable insights for investors and helping companies gauge their future market performance.

Keywords-- Stock market prediction; Sentiment Analysis; Neural Networks; Long-short Term Memory Neural Networks, DJIA, Ensemble Method, Weighted Average

INTRODUCTION

Financial markets are intricate and highly dynamic systems that significantly impact the global economy. Various factors, including macroeconomic elements such as politics, natural and man-made disasters, market psychology, as well as direct influences like supply and demand, speculation, and expectations, heavily affect financial stock markets.
As globalization and financial market integration introduce additional complexities, relying solely on theoretical models to predict stock market movements becomes increasingly challenging and less feasible. Therefore, it is essential to consider these external factors.

Over the past two decades, stock market prediction has become a prominent research topic. Advances in computational power and the availability of big data have led researchers to leverage web resources for forecasting stock market trends. Machine learning and Natural Language Processing techniques are now employed to analyse stock market behaviour. Financial data, which can be in the form of text, videos, or images, is often sourced from web news, search engine queries, and social media.

In this context, we present an application designed to predict stock return prices for 30 companies listed in the Dow Jones Industrial Average (DJIA). Our approach integrates Twitter sentiment analysis, web news text analysis, search engine query data, historical stock prices, and various machine learning techniques to develop a

stock price return prediction algorithm. The algorithm's accuracy is validated using real stock market data.

Previous research has established a correlation between stock market data and these data sources, suggesting a strong potential for predicting stock return prices. These predictions can provide valuable insights into how market conditions affect a company's financial security, enabling more informed financial decisions that could potentially save millions of dollars.

LITERATURE REVIEW

Over the past two decades, increased commercialization has significantly raised investor engagement in the stock market. Technological advancements have provided investors with expanded market access and opportunities. According to research by S. Abdulsalam Sulaiman, S. Adewole, K. S. Jimoh, and R. G., the volume and speed of data collection have grown exponentially due to advancements in computer technology. This surge in data presents opportunities for those who can effectively mine and utilize the information within [4].

Consequently, there has been a growing focus on predicting stock market values. Research by David Enke and Suraphan Thawornwong [4] highlights a relationship between historical data and future stock market values, suggesting that data mining techniques can be instrumental in uncovering these predictive relationships. As a result, researchers have increasingly turned to Artificial Intelligence (AI) and data mining techniques for stock market prediction. A study [3] utilized daily stock movement data from the Nigerian Stock Exchange to build a database for predicting future values, employing regression analysis and moving averages to forecast stock market trends.

Further research by Huina Mao, Scott Counts, and Johan Bollen [5] emphasizes that stock market valuations are influenced by both new and hidden information, including behavioral and emotional factors such as social mood. Analyzing social mood has thus become a key task in stock market forecasting.

With the advent of advanced technology, global stakeholders have expressed their sentiments and ideas online. Research [6] explored whether daily tweet volumes can predict S&P 500 stock indicators. Another study examined the relationship between Twitter sentiment and DJIA, finding significant correlations. Research [2] demonstrated a notable correlation between stock price returns and Twitter sentiment, using financial econometrics methods to analyze abnormal price returns and predict stock price trends over a broader time series.

The rise of the Internet in the 1990s transformed the availability of news from print to immediate digital formats, allowing for real-time financial analysis [7]. An interdisciplinary field has emerged to enable computers to interpret news articles promptly and leverage them for financial gain.

An approach [8] investigated the correlation between RSS feed sentiments, tweets, and stock market values over specific periods. This model demonstrated a 20% improvement in forecast accuracy by predicting stock market indicators based on textual inputs.

Research by I. Bordino, S. Battiston, G. C., M. Cristelli, A. Ukkonen, and I. Weber [9] examined the prediction of stock market volumes using search engine queries. This research revealed that search engine traffic could forecast social event dynamics and trading volumes. However, it remains unclear if online user insights can

reliably predict financial market trends. The study found that query volumes for NASDAQ-100 stocks correlated with daily trading volumes, and query dynamics could anticipate trading peaks by more than one day. These findings contribute to the ongoing debate on identifying early warnings of financial risk based on online user activity.

METHODOLOGY

A. Data

This research focuses on predicting the closing stock price returns for the Dow 30 companies, using the Dow Jones Industrial Average Index (DJI) due to its stability. To predict stock prices, data from social media, web news, and search engine queries were collected from January 1, 2016, to July 31, 2019. Historical market data and live stock prices were obtained through the Yahoo Finance API.

Collecting historical Twitter data posed some challenges with the Twitter API, as tweets from the specified period needed to be downloaded. To address this, a Python-based web crawling application was developed, enabling the download of over 0.9 million tweets for each company. The collected tweets were in English, and searches were performed using the company's cash tag (e.g., $AAPL for Apple Inc.). Sentiment analysis was conducted on these tweets, and the average sentiment score for each company was calculated.

Web news headlines were sourced from Reuters. Both historical and current data were retrieved using a Python-based web crawler. News articles were filtered by using the company's cash tag.

For web search engine query data, Google Trends was utilized. Data was downloaded as the number of hits for specific keywords. The company name was used as the keyword within the finance and news categories to filter out irrelevant search results (e.g., excluding results related to apple fruit).

B. Procedure

- Data Preprocessing

The collected data required preprocessing to handle gaps and ensure consistency. Stock market data is unavailable during weekends, so null values were filled with the corresponding closing stock price from the previous Friday. For each company, multiple daily tweets were processed, and sentiment scores were calculated. The daily mean sentiment score and tweet volume were then computed. Sentiment polarity was defined as $s-$ or $s+$ representing the relative amount of negative or positive sentiment within a given time interval, where s representing the relative amount of negative or positive sentiment within a given time interval, where s is the daily sentiment value, t is time and n is the number of days. The sum of $s-$ or $s+$ was used to calculate the mean sentiment value for each day. Missing values in the web news dataset were addressed by calculating the mean sentiment score for available days and using linear interpolation to fill in gaps.

Sentiment analysis, a crucial part of this research, relies on the accuracy of sentiment scores derived from web news and tweets. This analysis was conducted using Python library TextBlob, a widely used tool for textual processing. Before applying TextBlob, textual data was cleaned using the NLTK English stop words corpus.

- Time Series Prediction

Given that stock prediction involves time series data, a Recurrent Neural Network (RNN) was employed. Various algorithms, including Long Short-Term Memory (LSTM) and ARIMA, were initially tested. LSTM demonstrated the best performance and was selected for this research. The LSTM network is a state-of-the-art RNN for time series prediction. Input variables for both multivariate and univariate time series forecasting included sentiment scores (*ts*), web news sentiment (*ns*), Google Trends hit volume (*gv*), and closing stock prices (*c*).

The forecasting models were defined as follows:

1. Twitter:
$$W1 \cdot ts(t-n) + W2 \cdot c(t-n) = c(t)$$

2. Web News:
$$W4 \cdot ns(t-n) + W5 \cdot c(t-n) = c(t)$$

3. Google Search Engine Query:
$$W6 \cdot gv(t-n) + W7 \cdot c(t-n) = c(t)$$

4. Stock Historical Data:
$$W8 \cdot c(t-n) = c(t)$$

Each output variable from these functions was used to fit both multivariate and univariate time series forecasting models using LSTM.

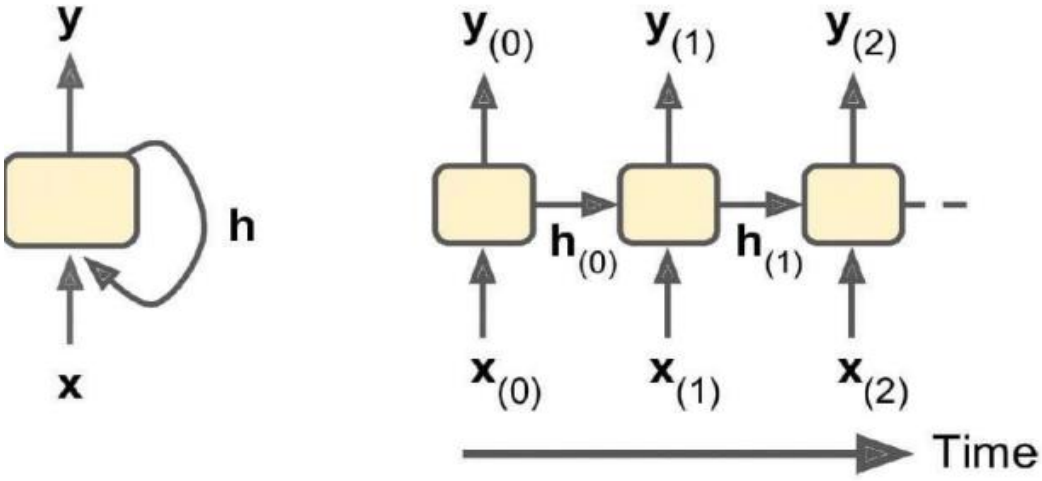

Figure 1

- Model Integration

Each model was fit with the relevant input data, and predictions were made using the most well-performing models, which were fine-tuned. To integrate these models and generate the final prediction, an ensemble methodology was employed. Ensemble methods combine several base models to produce a single optimal predictive model. In this scenario, a Weighted Average Ensemble was used. Initially, the Grid Search method was used to select weights for hyperparameters. However, due to high resource requirements and limited accuracy for decimal points, the Differential Evolution (DE) method was adopted. DE iteratively finds the optimal weights to maximize accuracy.

Weights were assigned to each model proportionally to their accuracy. The final prediction *P* was computed as follows:

$$Pt \cdot w1 + Pwn \cdot w2 + Pseq \cdot w3 = P$$

where *Pt, Pwn, and Pseq* are predictions based on Twitter sentiment, web news sentiment, and search engine query hits, respectively, with their corresponding weights *w1, w2, and w3.*

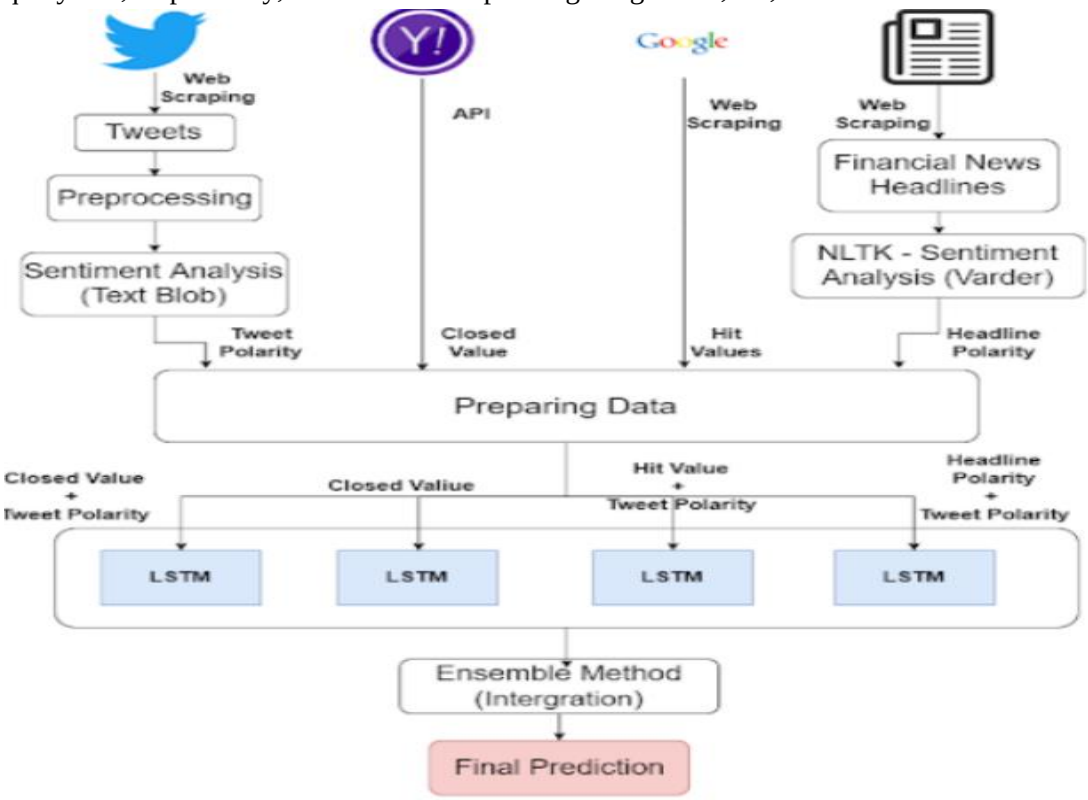

Figure 2: Model Integration

- System Implementation

The integrated components were hosted in an AWS EC2 server environment. A dashboard was created to visualize the statistics for end users, providing a comprehensive view of the predictions and underlying data.

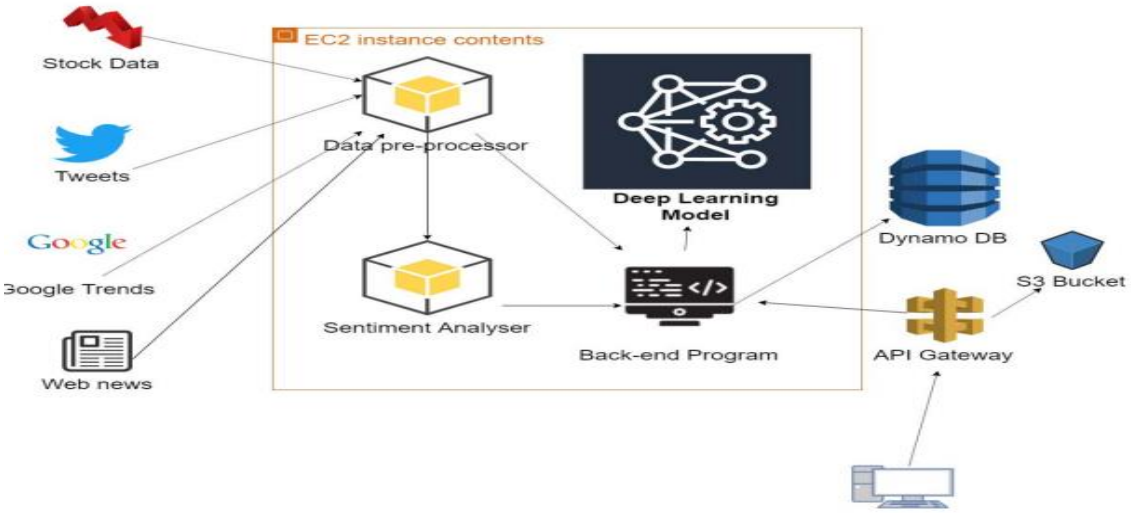

Figure 3: System Implementation

RESULT

The original stock price data of Dow 30 companies from Yahoo Finance was used as the ground truth for evaluating the predictions. The predicted stock values were compared with the actual values to assess accuracy. This section describes the stock market predictions based on Twitter sentiment, web news, search engine query hits, and the integrated model using the ensemble methodology.

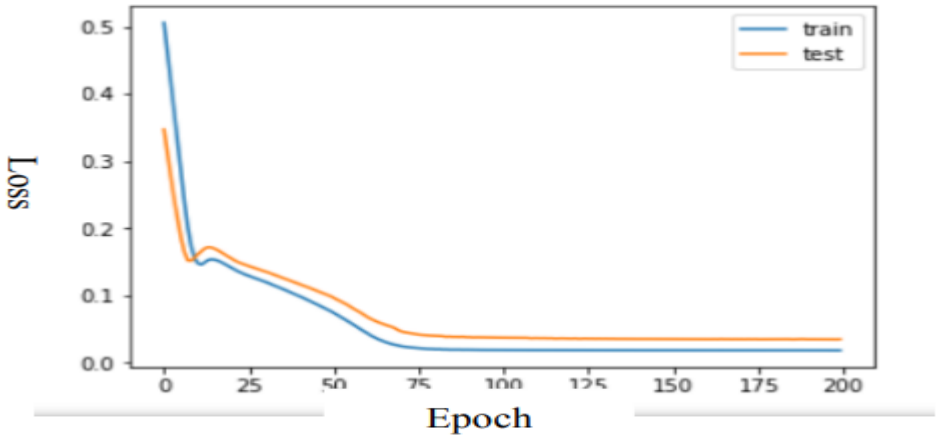

Figure 4: Model loss diagrams of twitter sentiment

Twitter Sentiment Analysis

The correlation between Twitter sentiment and stock market values for $AAPL was found to be 0.5. The model loss diagram for Twitter sentiment is shown below. The Twitter sentiment time series analysis forecasting for a 60-day period demonstrated a Root Mean Squared Error (RMSE) of 0.013 and an accuracy of 0.962.

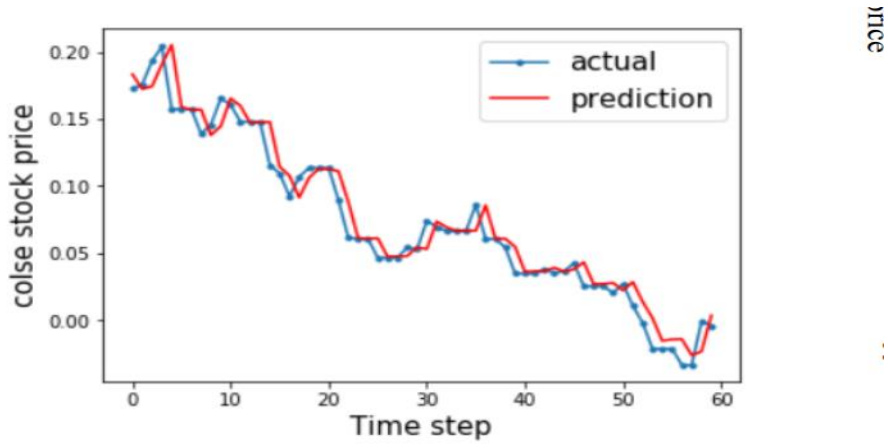

Figure 5: Twitter sentiment model prediction vs actual price prediction

Web News Sentiment Analysis

The correlation between web news sentiment and stock market values for $AAPL was found to be 0.0061 .

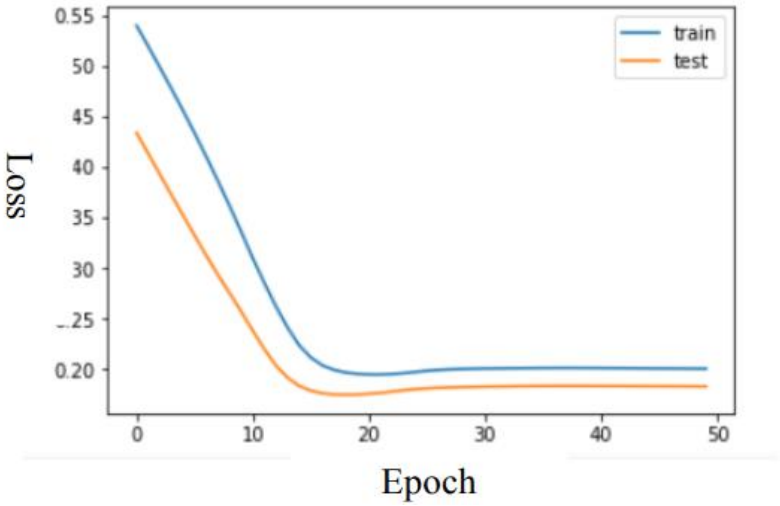

Figure 6: Model loss diagram of web news sentiment

The model loss diagram for web news sentiment is shown below. The web news sentiment time series analysis forecasting for a 250-day period demonstrated an RMSE of 0.013 and an accuracy of 0.961.

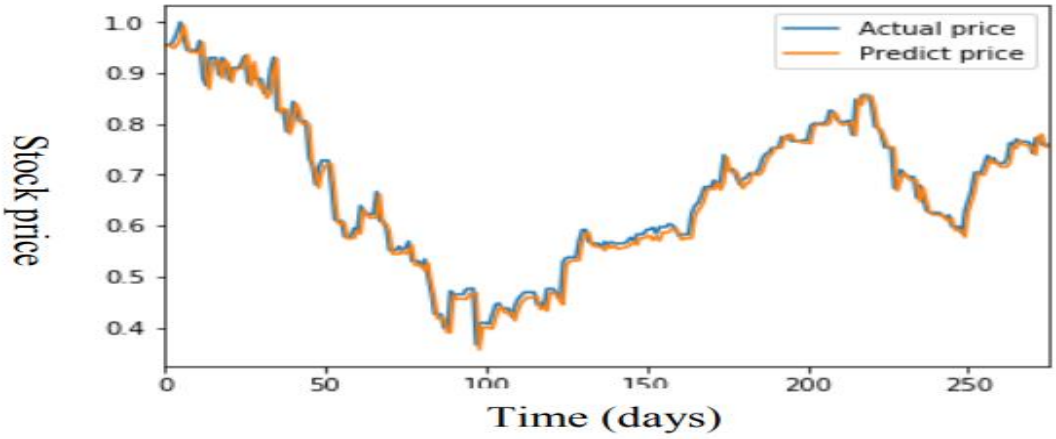

Figure 7: Web news predicted vs actual graph

Search Engine Query (SEQ) Analysis

The correlation between SEQ and stock market values for $AAPL was 0.45.

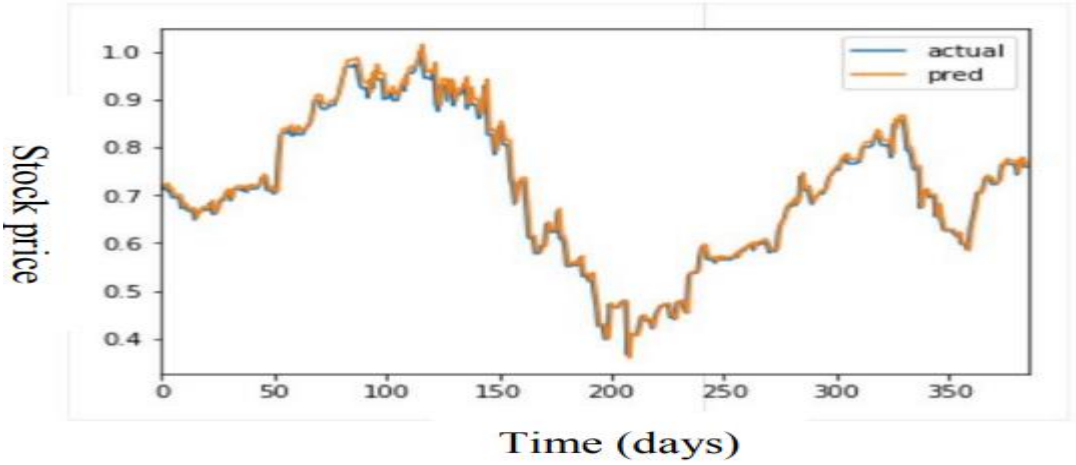

Figure 9 : SEQ predicted vs actual graph

The calculated weights for the Twitter, web news, and SEQ predicted models using the ensemble method were 0.4, 0.3, and 0.3, respectively. The best score for the ensemble model was 0.978.

The following table shows the actual stock price values and the predicted values by each model, as well as the final ensemble prediction values. An extended forecast was conducted using data from the past three days.

date	Twitter Predicted	Search engine predicted	Web news predicted	Ensemble prediction	Actual Values
6/20/2018	186.729370	185.980057	185.286896	186.152390	186.500000
6/21/2018	186.795837	186.206451	186.797684	186.797684	185.460007
6/22/2018	186.170914	185.792419	185.078552	185.733978	184.919998
6/23/2018	186.051971	185.226562	184.457062	185.414017	184.919998
6/24/2018	186.148560	185.363342	184.379349	185.440887	184.919998

Prediction Accuracy

For one-day predictions, the individual accuracies for the Twitter, web news, and SEQ models were 0.98, 0.97, and 0.96, respectively. The weights were 0.5, 0.3, and 0.2, resulting in a best score of 0.99 for the ensemble model.

For seven-day predictions, the individual accuracies for the Twitter, web news, and SEQ models were 0.91, 0.90, and 0.88, respectively. The weights were 0.6, 0.3, and 0.1, resulting in a best score of 0.9236 for the ensemble prediction model.

For 15-day predictions, the individual accuracies for the Twitter, web news, and SEQ models were 0.8559, 0.8220, and 0.8379, respectively. The weights were 0.7727, 0.2260, and 0.00013, resulting in a best score of 0.8391 for the ensemble prediction model.

For 30-day predictions, the individual accuracies for the Twitter, web news, and SEQ models were 0.6292, 0.6367, and 0.6702, respectively. The weights were 0.02, 0.1, and 0.97, resulting in a best score of 0.6702 for the ensemble prediction model.

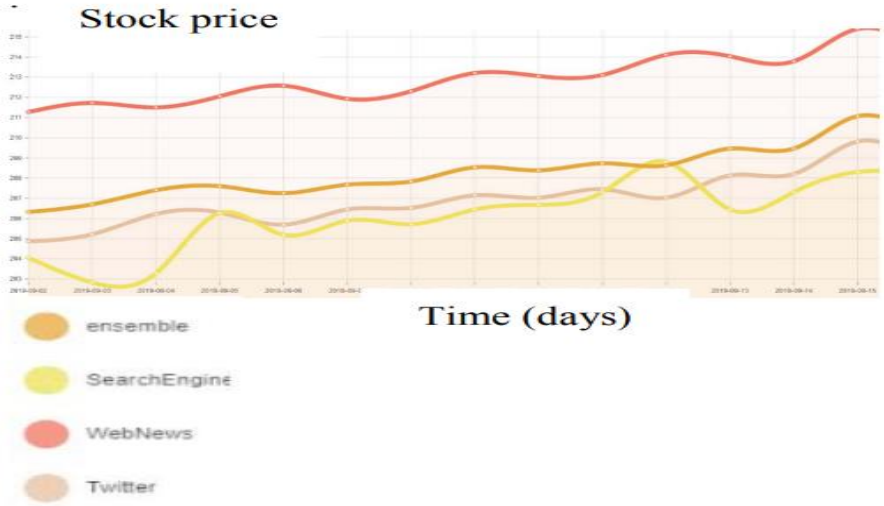

Figure 11: Stock Market Prediction

CONCLUSION

The utmost intension of this study is to forecast stock market values of Dow jones 30 by using four main components which were discussed throughout this paper. As mentioned, all the necessary data were retrieved and preprocess to fit different RNN models. Even though resulted correlations were relatively low, integrated model showed a success throughout this research. And paved the way to discover a much clearer path towards an accurate model. Modal weights were changed with the prediction time range. When the time range increases the weights for the twitter sentiment analysis and web news analysis modals were decreased and when the prediction time range decreases twitter analysis modal shows increase of weight and it can be concluded that the effect from twitter and web news data sources are diminishing for long term predictions and twitter is good for short term prediction. An extended study will build a finance-based sentiment corpus instead of using text blob sentiment analysis. As for future study, beta value for stock market data will be included for feature inputs. Furthermore, macroeconomic variables like exchange rate and gold rate will be studied.

REFERENCES

[1] Dassanayake, W. and Jayawardena, C. (2017). Determinants of stock market index movements: Evidence from the New Zealand stock market. [online] https://www.researchgate.net. Available at: https://www.researchgate.net/publication/314668127_Determinants_of_stock_market_index_movements_Evidence_from_New_Zealand_stock_market [Accessed 2 Jan. 2019].

[2] Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. and Mozetič, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. [online] https://www.researchgate.net. Available at: https://www.researchgate.net/publication/282049046_The_Effects_of_Twitter_Sentiment_on_Stock_Price_Returns [Accessed 20 Dec. 2018].I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[3] Sulaiman Olaniyi, A., Adewole, K. and Jimoh, R. (2011). Stock Trend Prediction Using Regression Analysis – A Data Mining Approach. [online] https://www.researchgate.net. Available at: https://www.researchgate.net/publication/277409163_Stock_Trend_Prediction_Using_Regression_Analysis_-_A_Data_Mining_Approach [Accessed 11 Jan. 2019].

[4] Enke, D., Thawornwong, S. (2005) "The use of data mining and neural networks for forecasting stock market returns", Expert Systems with Applications, 29, pp. 927-940

[5] Mao, H., Counts, S. and Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. [online] Arxiv.org. Available at: https://arxiv.org/pdf/1112.1051.pdf [Accessed 11 Jan. 2019].

[6] Mao Y, Wei W, Wang B, Liu B. Correlating S&P 500 stocks with Twitter data. In: Proc. 1st ACM Intl. Workshop on Hot Topics on Interdisciplinary Social Networks Research; 2012. p. 69–72.

[7]]M. Beckmann, "STOCK PRICE CHANGE PREDICTION USING NEWS TEXT MINING", 2017. [Online]. Available: https://www.researchgate.net/publication/313473231_Stock_Price_Change_Prediction_Using_News_Text_Mining. [Accessed: 03- Mar2019].

[8] S. Bharathi1*, A. Geetha1 and R. Sathiynarayanan1, "Sentiment Analysis of Twitter and RSS News Feeds and Its Impact on Stock Market Prediction", 2017. [Online]. Available: https://www.researchgate.net/publication/320694083_Sentiment_Analysis_of_Twitter_and_RSS_News_Feeds_and_Its_Impact_on_Stock_Market_Prediction. [Accessed: 05- Mar- 2019].

[9] Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A. and Weber, I. (2012). Web Search Queries Can Predict Stock Market Volumes. [online] Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0040014 [Accessed 3 Mar. 2019].

[10] "Ensemble Methods in Machine Learning: What are They and Why Use Them?", Medium, 2019. [Online]. Available: https://towardsdatascience.com/ensemble-methods-in-machinelearning-what-are-they-and-why-use-them-68ec3f9fef5f. [Accessed: 13- Sep- 2019]