2019 International Conference on Advancements in Computing (ICAC)
December 5-6, 2019. Malabe, Sri Lanka

Stock Market Prediction Using Machine Learning Techniques

Naadun Sirimevan, I.G. U. H. Mamalgaha, Chandira Jayasekara, Y. S. Mayuran[1], and Chandimal Jayawardena[2]

Faculty of Computing, Sri Lanka Institute of Information Technology,

Malabe 10115, Sri Lanka.

bala.mayuran@my.sliit.lk[1], chandimal.j@sliit.lk[2]

***Abstract*— Predicting stock market prices is crucial subject at the present economy. Hence, the tendency of researchers towards new opportunities to predict the stock market has been increased. Researchers have found that, historical stock data and Search Engine Queries, social mood from user generated content in sources like Twitter, Web News has a predictive relationship to the future stock prices. Lack of information such as social mood was there in past studies and in this research, we discuss an effective method to analyze multiple information sources to fill the information gap and predict an accurate future value. For this, LSTM – RNN models were employed to analyze sperate sources and Ensembled method with Weighted Average and Differential Evolution technique were used for more accurate prediction of the stock prices. And highly accurate predictions were made to one-day, seven-days, 15-days and 30 days for the future. So that investors could gain an insight into what they are inventing for and the companies to track how well they will perform in the stock market.**

Keywords—Stock market prediction; Sentiment Analysis; Neural Networks; Long-short Term Memory Neural Networks, DJIA, Ensemble Method, Weigthed Average, Dow jones

I. INTRODUCTION

Financial markets are considered as sophisticated, highly dynamic markets declaring a strong impact on the global economy. Macroeconomic factors like politics, natural disasters, man-made-disasters, market psychology and other direct factors like supply and demand, speculation and expectation conduct a strong effect on financial stock markets.

Since globalization and the financial market integration intensifies further complexities [1], it is hardly accurate and feasible to predict stock market movements only using theories. As a result, above mentioned exogenous factors should be considered.

During the last two decades, the stock market prediction has been an active topic among researchers. With the rise of computational power and the availability of big data, researchers have tended to use web resources to predict stock market movements. Machine learning techniques, Natural Language Processing techniques can be used to predict the behavior of the stock markets. Necessary financial data can be in forms of text, videos, photos etc. Most of the researchers have used textual data extracted from web news, search engine queries and social media. What we present here is an application to predict stock return price of 30 stock companies from the Dow Jones Industrial Average (DJIA) index. We use the Twitter sentiment analysis, web news text analysis, search engine query hits, historical stock prices and various machine learning techniques for the stock price return prediction algorithm which validates the accuracy with real stock market data.

Since previous researchers have found that there is a correlation between stock market data and the above data sources, there is a strong possibility to predict the stock return prices. These predicted values can be utilized to get future insights on how the market affects the financial security of a company and make financial decisions that will save millions of dollars.

II. LITERATURE REVIEW

During the past two decades or so, due to the high Commercialization, there has been a huge improvement, high involvement of the investors in the stock market. Because of the development of technology, investors receive a considerable advantage of market scope enlargement.

Also, in a research conducted by S Abdulsalam Sulaiman Olaniyi, Adewole, Kayode S., Jimoh, R. G, they have mentioned that different kinds of data are being collected at a dramatic pace and the size of data generated and stored is growing in an increased rate due to the continuing development in computer technology. And there's a great opportunity to those who can retrieve the information hidden within that data [4].

So, much attention of researchers is being directed to stock market value prediction. As of [4], research by David Enke & Suraphan Thawornwong, mention that there is a relationship among the available past data and future stock market values. To retrieve and mine the relationships related to the predictions, data mining techniques can be used.

Therefore, researchers have been focused on Artificial Intelligence and Data Mining techniques. In the research [3], they have collected the daily movement of stock values between a time range from the daily official list of Nigerian Stock Exchange. And build a database to use the data to predict future values. They have built a data mining tool with the use of regression analysis of time series data which was used by moving average to predict future stock market values.

Furthermore, According to Huina Mao,Scott Counts,and Johan Bollen, in [5], the stock market valuations have a relationship with all the available new and hidden information such as behavioral and emotional factors like social mood. As a result of that, analyzing the social mood,

978-1-7281-4170-1/19/$31.00 ©2019 IEEE

has become an important and a key task in stock market forecast.

With the availability of advanced technology, these stocks have been addressed globally by the stakeholders expressing their emotions and the ideas. [6] studied whether the daily number of tweets predicts the S&P 500 stock indicators. Another research finds the contents of tweets. In a textual analysis approach to Twitter data, the authors find relations between the mood indicators and the DJIA.

[2] study shows that there is a significant correlation between stock price returns and twitter sentiments which is worth trying. They have found that there are a relatively low Pearson's correlation and Granger causality between the corresponding time series. Since they have used a methodology in financial econometrics called event study, they were able to analyze the abnormal price returns observed during external events. So, they were able to predict stock price returns of a broader time series.

With the rise and popularization of the Internet in the 1990s, the newspaper articles that were previously available overnight, printed on paper, began to be available as soon as possible, in digital format, at the speed that the market financial needs. Nevertheless, there is an interdisciplinary field of research has been created to allow computers to interpret news articles at the right time and generate profits on the financial markets.[7]

There is an approach [8], to find the correlation between sentiments of RSS (Really Simple Syndication) feeds and tweets, and the value of the stock market for a specific period. In that algorithm, the model formed is used to predict stock market indicators. A common framework is foreseen for finds a generic stock price prediction, the textual records are considered as inputs, and the price movements generate in the output. This proves the correlation between stock indicators and news feeds and RSS tweets and showed a 20% improvement in forecast accuracy.

In a research [9], by I Bordino, S Battiston, Guido C, M Cristelli, A Ukkonen, and Ir Weber, stock market volumes are predicted using search engine queries. And another important finding that search engine traffic can be used to trail and to foresee the dynamics of social events.

Several recent research studies used this method to stock market sentiment evaluations. Nevertheless, it's unclear if trends in financial markets can be foreseen by the collective insight of online users in the web. In the research [9], it is stated that trading volumes of stocks in NASDAQ-100 are correlated with daily volumes of queries relevant to the same stock. In discrete, query volumes foresee in many event peaks of trading by more than one day. Therefore, analysis is driven out on a specific dataset of queries, submitted to an important search engine, which lets on an investigation and the user behavior. [9] shows the query volume dynamics emerge from the collective but apparently uncoordinated activity of many users. These discoveries contribute to the debate on the identification of early warnings of financial risk, based on the activities of the www users.

III. METHODOLOGY

A. Data

The closed stock price return values of Dow 30 are the predicting variable in this research. DJI (Dow Jones Industrial Average Index) is used since it is a stable index. In order to predict the stock price values social media data, web news data and search engine query data are used from 2016-01-01 to 2019-07-31. Historical market data and live stock prices were gathered using Yahoo finance API.

Past Twitter data gathering became bit challenging with the Twitter API, because tweets from 2016 to 2019 July had to be downloaded. Hence python web crawling application is written to download past Tweets. So, downloading over 0.9 million tweets for each company was possible. Gathered Tweets were English, and the only keyword used to search Tweets was company Cash tag. e.g. $AAPL for Apple Inc. Twitter sentiment values are generated with each tweet and the mean sentiment score is calculated.

Web news headlines were gathered from https://www.reuters.com. Both past and live data were able to download using a web crawler application created using python. Cash tag was used to filter news. Google trend data were used for the web search engine query data. Google trends data were downloaded as number of hits for a keyword. To filter Google trends, the company name is used as the keyword specified under finance and news categories, such that other names related to the company name will be eliminated. E.g. to ignore apple fruit from the search results.

B. Procedure

Collected data needs to be preprocessed. Stock market were unavailable during the weekend. Hence null values were filled with the corresponding closed stock price of the Friday. There were multiple tweets per day for a company. All the tweets were preprocessed, and sentiment scores were calculated. Then sentiment mean score and the Twitter volume per day is calculated. s- or s+ is defined as the relative amount of polarity represented in a time interval.

$$s+(t) = \frac{1}{n}\sum_{i=1}^{n}(S_i \mid S_i > 0)|(t)$$

$$s^{-}(t) = \frac{1}{n}\sum_{i=1}^{n}(S_i \mid S_i < 0)|(t)$$

Where s is sentiment values per day, t is time and n is number of days.

Sum of s- and s+ is calculated as the mean sentiment value for a given day. There were missing values for some days in web news dataset. Sentiment mean score for a given day is calculated and missing values are filled with linear interpolation.

Sentiment analysis is a crucial part in this research. Web news and tweets are depending upon the accuracy of the sentiment score. Sentiment analysis is performed using a python library called TextBlob. It is a widely using textual processing library. As prior to applying to the TextBlob, textual data is cleaned with using NLTK English stop word corpus.

Since stock prediction involves with a time series, Recurrent Neural Network (RNN) is used. Initially various types of algorithms like LSTM (Long Short-Term Memory), ARIMA method were employed. Among those, LSTM showed comparatively best performances. LSTM network is a state-of-the-art RNN for the time series prediction at the time. Sentiment score (t_s), Web news sentiment (n_s), Google trends hits volume (g_v), and closed stock price (c) were the input variables for the multivariate and univariate time series forecasting.

For Twitter,

$$W_1\, t_s(t\text{-}n) + W_2\, c(t\text{-}n) = c(t)$$

For Web News,

$$W_4\, n_s\, (t\text{-}n) + W_5\, c(t\text{-}n) = c(t)$$

For Google Search Engine Query,

$$W_6\, g_v\, (t\text{-}n) + W_7\, c(t\text{-}n) = c(t)$$

For Stock Historical Data,

$$W_8 c(t\text{-}n) = c(t)$$

Where Sentiment score (ts), Web news sentiment (ns), Google trends hits volume (gv), and closed stock price (c) were the input variables for the multivariate and univariate time series forecasting.

Each output variable of each function is used again to fit multivariate and univariate time series forecasting using LSTM.

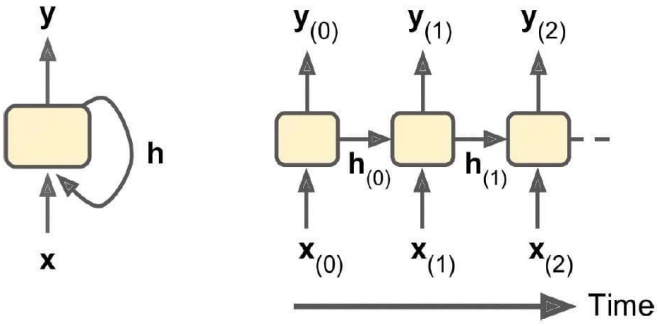

Figure 1

Each model is fit with relevant input data, prediction is performed with most good fit models which were fined tuned.

These models need to be integrated in order to find the final prediction. Ensemble methodology was used for the integration. [10] Ensemble methods is a machine learning technique that connects several base models in order to produce single optimal predictive model. In this scenario we have used Weighted Average Ensemble. To select weighs hyper parameters, Grid Search method was initially used. With the high requirement of resources and less accuracy for decimal points in Grid search method, Differential Evolution (DE) method has been used. DE runs iteratively and finds the weights for models to get a maximus accuracy

Weight is assigned to each model as directly proportional to the model accuracy. P_t, P_{wn}, P_{seq}, are twitter sentiment, web news and search engine query hits predictions as their respective weights are w_1, w_2 and w_3. If final prediction is P,

If final prediction is P,

$$P_t\, w_1 + P_{wn}\, w_2 + P_{seq}\, w_3 = P$$

Where Pt, Pwn, Pseq, are twitter sentiment, web news and search engine query hits predictions as their respective weights are w1, w2 and w3.

System design is modeled as in following diagram.

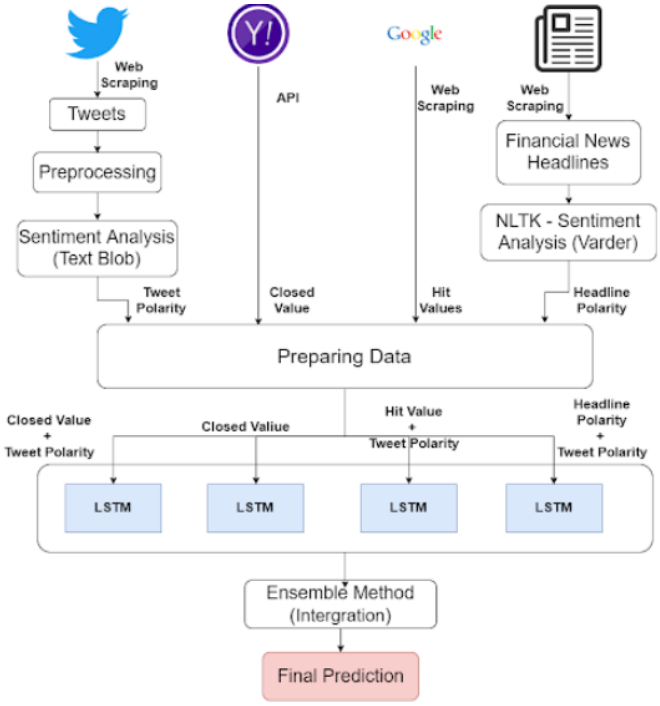

Figure 2: Application system design

As the final product, integrated components were hosted in AWS EC2 server environment. And a dashboard to visualize the statics for the end users.

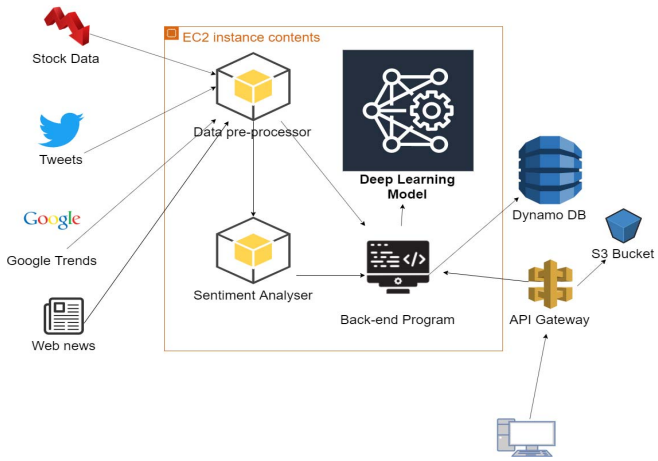

Figure 3: High level system architecture

IV. Result

Original stock price data of Dow 30 companies are used from Yahoo finance to find out the ground truth of the

prediction. The predicted stock values are compared with the actual values to evaluate the accuracy. Following section describes stock market predictions done with Twitter, web news, search engine query hits and the integrated model with ensemble methodology.

Twitter sentiment correlation with stock market values of $AAPL is 0.5.

Model loss diagrams of twitter sentiment is shown below.

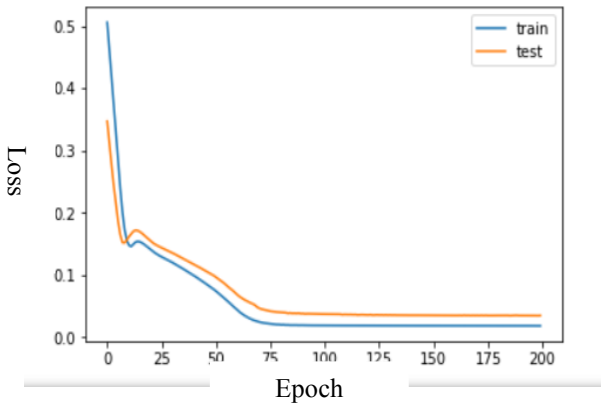

Figure 4 : Model loss diagrams of twitter sentiment

Twitter sentiment time series analysis forecasting for 60 days timestamp diagram is shown below. Its Root Mean Squared Error is 0.013. Its accuracy was 0.962.

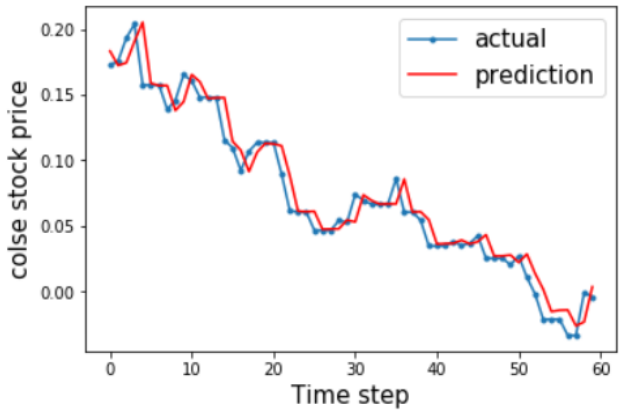

Figure 5: Twitter sentiment model predictin vs actual price returns

Web news sentiment correlation with stock market values of $AAPL is 0.0061.

Model loss diagrams of web news sentiment is shown below.

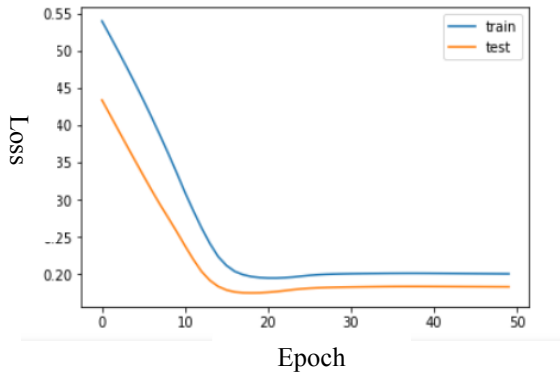

Figure 6: Model loss diagrams of web news sentiment

Web news sentiment time series analysis forecasting for 250 days timestamp diagram is shown below. Its Root Mean Squared Error is 0.013. Its accuracy was 0.961.

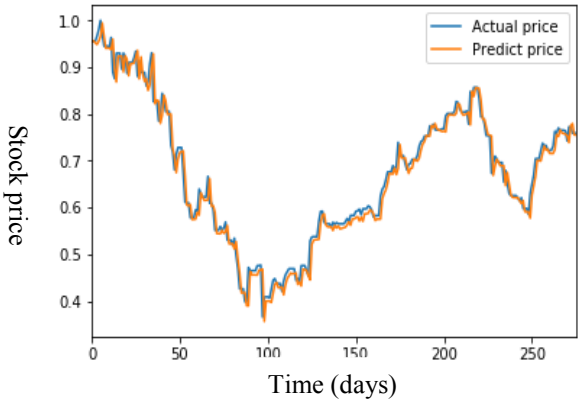

Figure 7: Web news predicted vs actual graph

SEQ correlation with stock market values of $AAPL is 0.45.

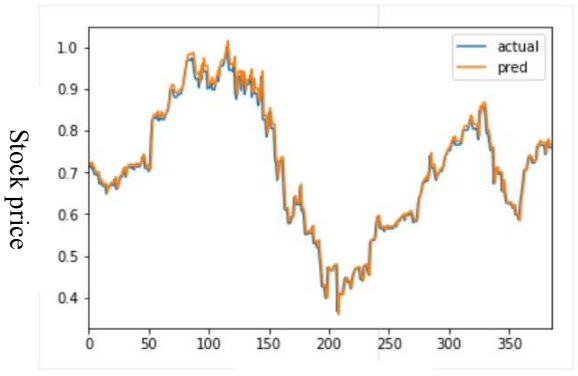

Figure 9: SEQ predicted vs actual graph

Calculated wieghts for Twitter, web news and SEQ predicted model using Ensembel method are respectively

0.4, 0.3 and 0.3. So the best score for the ensemble modle was 0.978.

Following table shows actual stock price values and predicted values by each models and final ensemble prediction values.

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

Figure 10: AAPL predicted prices

An extended forecasting was done by taking data for past three days.

Individual accuracies for the one-day prediction were 0.98, 0.97 and 0.96 respectively for twitter, web news and search engine query models. Weights were 0.5, 0.3 and 0.2 and the best score for the ensembled model was 0.99.

The individual accuracies for the seven-days prediction were 0.91, 0.90 and 0.88 respectively for twitter, web news and search engine query models. Weights were 0.6, 0.3 and 0.1 and the best score for the ensembled prediction model was 0.9236.

The individual accuracies for the 15-days prediction were 0.8559, 0.8220 and 0.8379 respectively for twitter, web news and search engine query models. Weights were 0.7727, 0.2260 and 0.00013 and the best score for the ensembled prediction model was 0.8391.

The individual accuracies for the 30-days prediction were 0.6292, 0.6367 and 0.6702 respectively for twitter, web news and search engine query models. Weights were 0.02, 0.1 and 0.9700 and the best score for the ensembled prediction model was 0.6702

Figure 11: Stock Market Prediction of Apple for next 15 days.

Figure 12: Previous Prices and Predicted Prices for Apple.

V. Conclution

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 find a much clear 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 C hange _Prediction_Using_News_Text_Mining. [Accessed: 03- Mar- 2019].

[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 Ana lysis 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.004 0014 [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-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f. [Accessed: 13- Sep- 2019]