Stock Recommendation based on Price Forecasting and Twitter Sentiment Analysis

Dheeraj Vurukuti
011810023
Washington State University

Kunal Sanjay Sanghvi
011809708
Washington State University

***Abstract*—Stock price forecasting may turn out to be a crucial and blooming area of financial engineering, especially since new techniques and approaches on this matter are gaining ground constantly. It is now widely believed that the behavior of stock prices can be associated with the public opinion expressed on social media because of the enormous level of constant use of these platforms in the modern period. The idea is to recognize patterns that confirm this correlation and utilize them to forecast how various stock prices will behave in the future. Without a doubt, when examined collectively, tweets can provide a satisfactory depiction of public sentiment, despite being unappealing individually. We develop a system that collects old tweets, processes them, and evaluates the effectiveness of various tweets using Sentiment Intensity Analyzer for providing a positive or negative sentiment on the tweet corpus. Then, we use Random Forest Regressor to predict the stock prices based on the previous data available. Seeing the outcome of the prediction as well as using the sentiment analysis of the tweets to predict we determine whether to buy or sell this stock. The final outcomes seem promising given the association we discovered between stock prices and tweet sentiment.**

***Index Terms*—Sentiment Intensity Analyzer, Random Forest Regressor, Linear Regression, Twitter Data.**

I. Introduction

Predicting the movement in the price of the stock market is in the interest of many individuals and organizations. A country's economy also heavily depends on its stock market. It is one of the most important chances for businesses and investors to invest. Stock exchange forecasting has proven to be a fruitful area for the application of sentiment analysis. This topic is undeniably the subject of intense research in modern times. A vast amount of knowledge, which contains information about many topics, is being transmitted online through various social media, with Twitter serving as a wonderful example with more than 400 million tweets sent each day. Though each tweet may not be noteworthy on its own, a large collection of them might yield information with valuable insight about the common opinion on a specific subject. topic. Forming trading strategies will benefit from determining the public's sentiment by retrieving online information from Twitter. Public opinion is arguably one of several elements that influence how accurately stock price fluctuations are predicted. Exchange rate prediction appears to be a random process over short time periods. Over a long period of time, the movement of the stock price often follows a linear curve. People tend to purchase equities whose prices are anticipated to increase in the near future.

The difficulties encountered in this domain are that it is frequently difficult to predict the trend of the stock, and sentiment analysis of tweets is dependent on a variety of factors.While researching different sentiment analyzers, we came up with the Sentiment Intensity Analyzer to work best for our project scope, and we also used two different methods to predict stock market prices, which are the Random Forest Regressor and Linear Regression, just to check which was best for the job.

The project is divided into two parts. The first part includes pre-downloaded Twitter data along with the stock prices to determine both the social sentiment of the stock and the prediction of the stock prices, respectively. The second part includes where we try to use live Twitter data using the developer options, but due to limited access, we are only able to extract a couple of hundred tweets, and then using live data for that stock, we are predicting the prices on a limited basis. In the end, using both analyses, we are trying to determine whether to buy or sell the stock.

II. LITERATURE SURVEY

A. Survey of securities market Prediction Using Machine Learning Approach

The prediction of the exchange has grown in importance in the modern world. Technical analysis is one of every method used, although these methods don't always produce reliable conclusions.[1]

B. Impact of monetary Ratios and Technical Analysis on Stock Price Prediction Using Random Forests

The use of machine learning and computing techniques to predict the costs of the stock is an increasing trend.[2]

C. Stock Market Prediction via Multi-Source

Learning from multiple instances Predicting the exchange accurately could be difficult, but the modern web has proven to be a great tool in making this process simpler.[3]

Identify applicable funding agency here. If none, delete this.

D. Stock Market Prediction: Using Historical Data Analysis

The process of making stock market predictions is full of uncertainty and subject to a variety of influences. As a result, the exchange plays a crucial role in business and finance.[4]

E. A Survey on securities market Prediction Using SVM

Recent research offers solid evidence that the bulk of predictive regression models performs poorly in tests of predictability outside of samples.[5]

F. Predicting Stock Price Direction Using Support Vector Machines

In an effort to outperform the market for themselves or their clients, financial institutions and retailers have developed a variety of proprietary models. However, nobody has consistently achieved greater than average levels of profitability.[6]

G. Prediction Method supported Support Vector Machines (SVM) and Independent Component Analysis (ICA)

Inside the work centers within the various financial organizations, the statistic prediction problem was studied. For exchange prediction, the SVM-ICA prediction model, which combines independent analysis and SVM, is suggested.[7]

III. Problem Statement

The aim of this project is to provide users with information on past, present, and future market trends and behaviors using Twitter sentiment analysis techniques, which require training on historical data to analyze data patterns. Finding the best model to forecast stock market decisions is the major goal of this project. We discovered that techniques were not properly utilized when taking into consideration the different strategies and variables that needed to be considered. We investigate and propose a more practical approach to forecasting stock movement and the decision of whether to buy or sell the stock. A machine-learning model to forecast the longevity of stock in a competitive market is also presented in this project. The stock market institutions will greatly benefit from the accurate stock forecast because it will provide them with practical information. The project's objective is to make a collective decision based on the stock market price prediction graph and the Twitter sentiment analysis for that stock. The Twitter API can be used to get a variety of tweets pertaining to different companies. There could be a lot of tweets that aren't used for prediction. To analyze the value, stock data can be retrieved using the Yahoo Finance API. The Vader Sentiment Analyzer was used to analyze the sentiment of the tweets for that company. This complete approach gives us a definition and certain boundaries for our project, which is further covered in problem setup.

IV. Problem Setup

The problem setup is divided into two parts:

The first part includes where we use a pkl file type of Twitter dataset consisting of Twitter data ranging from January 1, 2007 to December 31, 2016, which is 20 MB big. Here, we use the Vader sentiment analyzer to effectively derive the sentiment of the tweets for that stock. Here, we are using UAL (United Airlines Holdings Inc.) as a sample stock for the testing scenario. Also, we are using the stock prices for that period and predicting the stock price trend using the twitter sentiment as a main factor to determine whether to buy or sell the stock.

The second part includes where we use a live Twitter developer account (limited access version) to access live tweets of the stock, but due to limited access we were only able to get hold of 329 tweets, to be precise, every time we called the API, even though we gave a very large date range. We then used Yahoo Finance to pull up live data for that particular stock for the same setup date, and we fed it to a random forest regressor model to predict the prices. Since the data is quite sparse, we managed to get the sentiment analysis along with the graph of the predicted prices to determine whether to buy or sell the stock.

V. Solution Approach

So, the problem statement which we used is quite technical as there are an immense number of setups and constraints we had difficulty in extracting the data and also deciding which model to go for predicting the stock market prices among the following choices as follows:

1) Adaptive Boost (AdaBoost)
2) k-Nearest Neighbors (kNN)
3) Linear Regression (LR)
4) Artificial Neural Network (ANN)
5) Random Forest (RF)
6) Stochastic Gradient Descent (SGD)
7) Support Vector Machine (SVM)
8) Decision Trees (DT)

So, among the choices, seeing that the model is based on different parameters, we selected the two models of linear regression and random forest to predict stock prices as they fit the problem data and scenario with better accuracy.

NLP is one of the best ways to do the sentiment analysis of any text data, yet we used NLTK, which is a natural language toolkit for symbolic and statistical NLP scenarios, for Twitter sentiment analysis.

We use all the possible metrics to decide which models to go for based on some trial-and-error methods and reading the various literature survey papers mentioned above.

Coming to the first problem, extracting the Twitter data and stock prices of that stock, we cracked the part of getting the stock price data using the Yahoo Finance library, but accessing the tweet data was difficult because, in the live environment, due to restrictions or limits on accessing the tweets, we could not gather enough tweets to predict a good sentiment analysis. So, to solve this problem, we used an extracted Twitter data set for a particular testing stock that consisted of regular tweets from January 1st, 2007 until December 31st, 2016. This has a couple of million tweets to determine the sentiment analysis, and we extracted the stock prices for that date frame, and then we tackled the problem of not having enough Twitter data.

The disadvantage is that the Twitter data we used is backdated, which provides insight into past data rather than present data due to the limitation in accessing live tweets, which limits us to testing about predicting stock prices in live scenarios as we cannot provide sentiment analysis of live tweets. Here we believe it may fail because one of the goals is not met, but we believe that if we start getting live tweets as data, we will be able to make the prediction.

A brief introduction about the methodology is as follows:

1. Random Forest Regressor:

Random Forest is an ensemble technique capable of handling both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as ”bagging.” In order to get the desired outcome, many decision trees are combined rather than relying simply on one decision tree.

Every decision tree has a significant variance, but when we mix them all in parallel, the variance is reduced since each decision tree is perfectly trained on the given sample data; thus, the output is dependent on numerous decision trees rather than just one. In a classification challenge, the result is determined by the majority voting classifier. The mean of each output constitutes the final output in a regression problem. This part is called ”aggregation.”

Multiple decision trees serve as the fundamental learning models in Random Forest. We create sample datasets for each model by randomly selecting rows and features from the dataset. This part is called ”bootstrap.”

2. Sentiment Analysis using Vader:

Sentiment analysis is a text analysis technique that finds polarity (such as a positive or negative sentiment) in the text, whether it be an entire document, paragraph, sentence, or clause.

Sentiment analysis uses computational methods to treat subjectivity in texts in order to measure the attitude, sentiments, evaluations, attitudes, and emotions of a speaker or writer.

Sentiment Analysis is difficult to perform because a text may contain multiple sentiments all at once.

2.1 Vader

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive or negative) and intensity (strength) of emotion.

It is available in the NLTK package and can be applied directly to unlabeled text data. A dictionary is used by the VADER sentiment analysis to translate lexical features into sentiment scores, which represent the degree of emotion. One can calculate the sentiment score of a text by multiplying the intensity of each word in the text.

VADER’s SentimentIntensityAnalyzer() takes in a string and returns a dictionary of scores in each of four categories:

1) negative
2) neutral
3) positive
4) compound (computed by normalizing the scores above)

Note: compound is computed by normalizing the scores above.

3. Linear Regression:

Linear Regression is a supervised machine learning model in which the model identifies the linear relationship between the dependent and independent variables by determining the best-fitting linear line between them.

More precisely, linear regression is used to determine the character and strength of the association between a dependent variable and a series of other independent variables. It helps create models to make predictions. Linear regression is commonly used for predictive analysis and modeling.

VI. Design and Implementation

Regression analysis is a statistical method that shows the relationship between two or more variables. The approach examines the relationship between a dependent variable and independent variables, typically represented in a graph. Linear Regression and Random Forest Regression are two examples of regression models utilized in this kind of arrangement.

Given that there are several actions taken by the companies, such as stock splits and bonuses, that affect stock prices, the adjusted close price is used as the training data for both the linear regression and the random forest regression models because it is the most recent updated data.

The segment of the project is divided as follows:

1. Data Collection:

Data collection may be a very basic module and the initial step towards the project. In general, it deals with compiling the appropriate dataset. The dataset that will be used for market prediction has to be filtered. By including more external data, data collection also enhances the dataset. The majority of our data is made up of back-dated stock prices and tweets about a chosen sample stock. We'll also attempt to use live, stock, and Twitter data simultaneously.

2. Pre-Processing:

Data is often typical, inconsistent, or incomplete, and typically contains many errors. Data processing, which entails changing data into a more coherent structure, The pre-processing of the data entails searching for missing values, looking for categorical values, dividing the data set into training and test sets, and ultimately performing feature scaling to narrow the range of variables so that they may be compared in common environments.

3. Training the Model:

Training the model is analogous to feeding information to the algorithm to touch up the test data. The training sets are to tune and fit the models. Since a model shouldn't be supported by unknown data, the test sets are left unmodified. The fit function, which uses the training data to generate a well-founded approximation of the model's performance, is part of the model training process. The purpose of tuning models is to precisely adjust the hyperparameters. The idea behind training a model is to use some initial values from the dataset to optimize any model parameters that need improvement. So, using the inputs from the test dataset, we use the predictions from the trained model. As a result, it is split into a 7:3 ratio, with 70% going to a training set and the remaining 30% going to a testing set of data. The main hyper parameter is the twitter sentiment analysis score which plays a very important role in predicting the stock market values.

4. Baseline Approach:

Information is scored because of the application of a predictive model to a body of data. The Random Forest Algorithm is the method that seeks to process the dataset. We gain interesting results using the ensemble method of random forest, which is commonly used for classification as well as regression backed by educational models. Thus, the final module explains how the model's findings might aid in predicting the likelihood that a stock will rise or fall in value when supported by parameters. It also reveals a stock's or organization's weaknesses. Additionally, we employ Twitter sentiment analysis to support the study of whether to purchase or sell based on expected pricing.

The stepwise approach of the baseline approach is as follows: Step 1: Load the tweets data and the corresponding adjusted close price. Step 2: Sentiment analysis of the tweet is done first and stored in the data frame. Step 3: Stock Price data is split into train and test datasets. Step 4: Using the sentiment analysis score we are predicting the stock market prices. Step 5: All the years predicted and the actual stock prices graph is displayed where a main hyper parameter of twitter sentiment analysis is taken into consideration and it gives a fair accuracy.

VII. SOFTWARE REQUIREMENT

- Python
- IDE(Colab)
- Numpy
- Scikit-learn
- Pandas
- Matplotlib
- Pyplot
- Tweepy
- NLTK

VIII. RESULTS

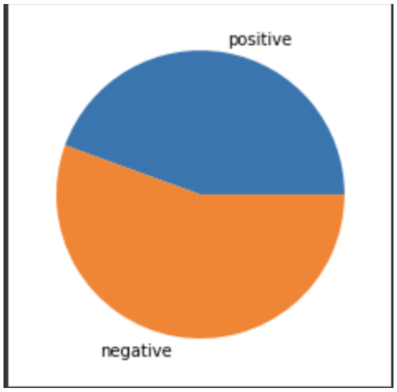

Fig. 1. Tweet Sentiment Analysis from January 2007 to December 2016 of the UAL dataset.The positive tweet percentage is 44.34 and the negative tweet percentage is 55.43

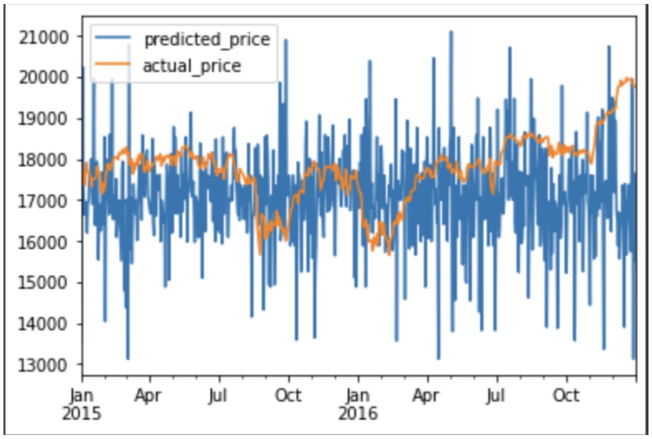

Fig. 2. Random Forest Regression for the UAL Test Data for Baseline Approach

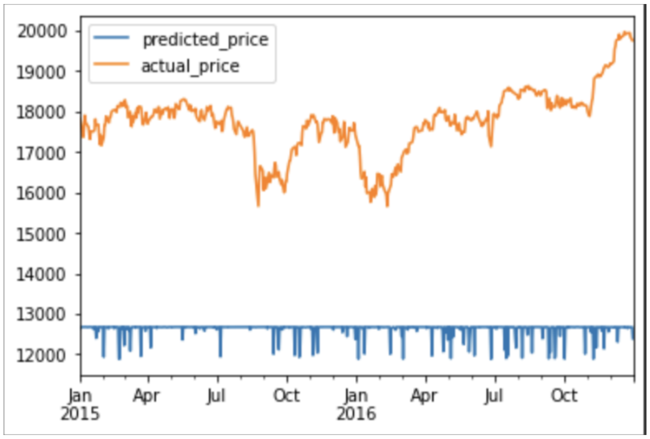

Fig. 3. Linear Regression for the UAL Test Data for Baseline Approach

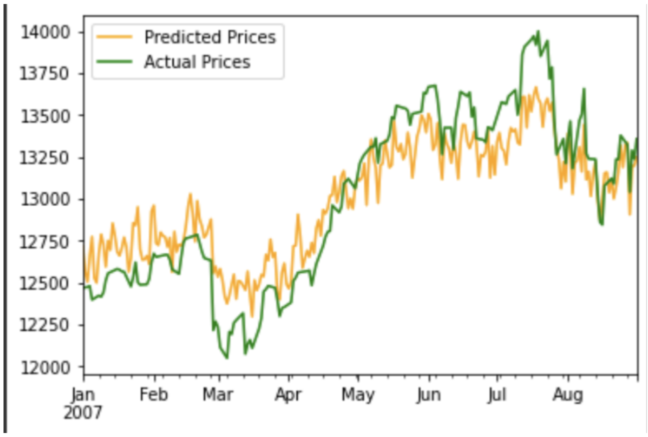

Fig. 4. Year 2007 Stock price prediction using Random Forest Regression

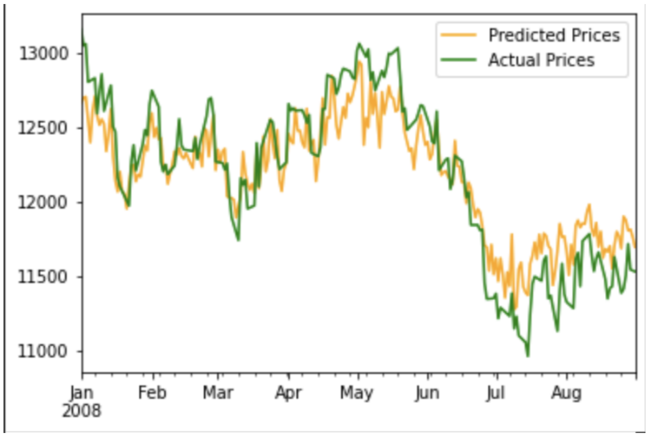

Fig. 5. Year 2008 Stock price prediction using Random Forest Regression

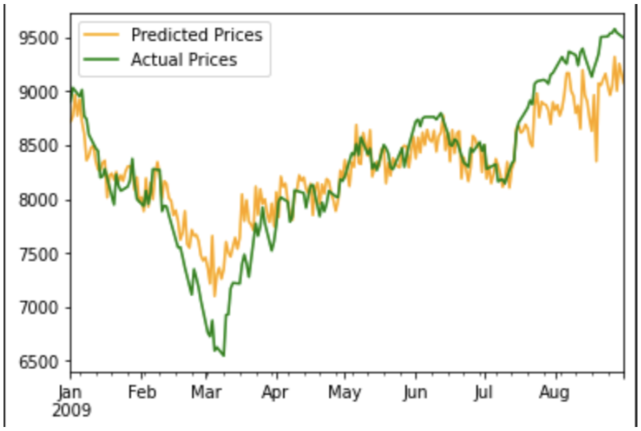

Fig. 6. Year 2009 Stock price prediction using Random Forest Regression

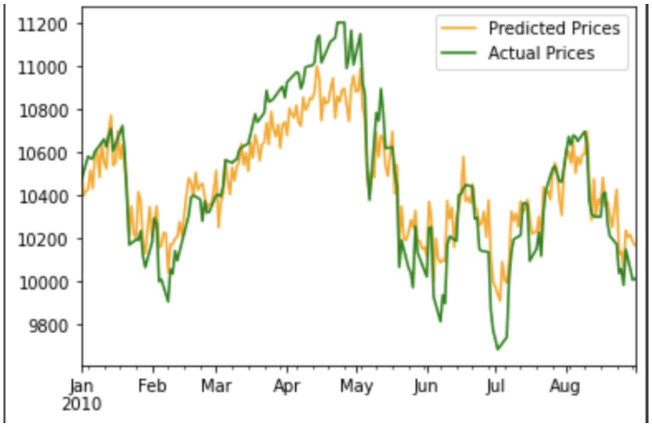

Fig. 7. Year 2010 Stock price prediction using Random Forest Regression

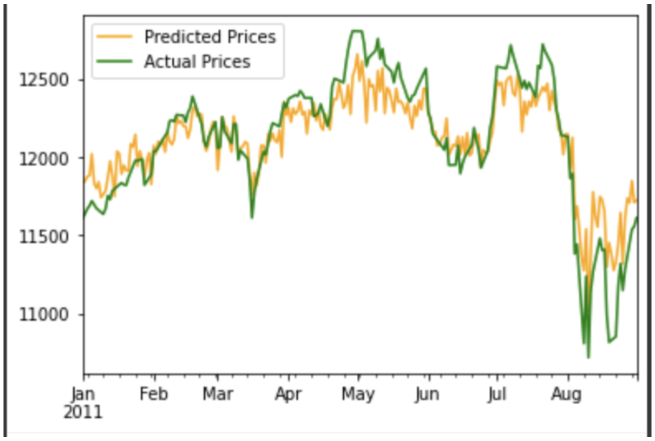

Fig. 8. Year 2011 Stock price prediction using Random Forest Regression

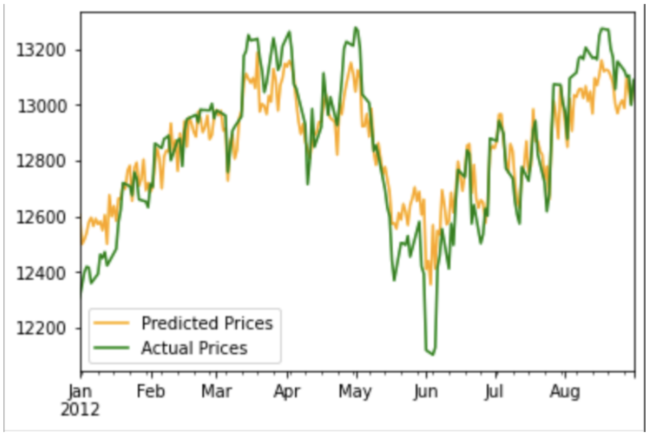

Fig. 9. Year 2012 Stock price prediction using Random Forest Regression

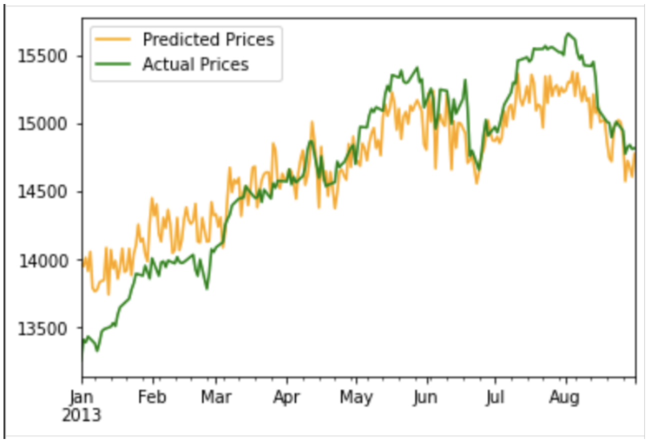

Fig. 10. Year 2013 Stock price prediction using Random Forest Regression

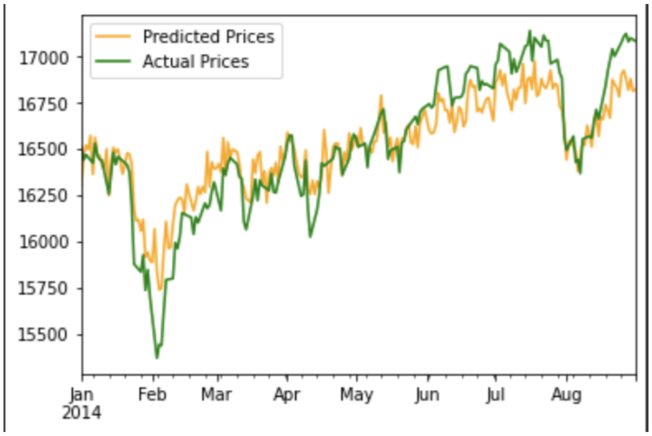

Fig. 11. Year 2014 Stock price prediction using Random Forest Regression

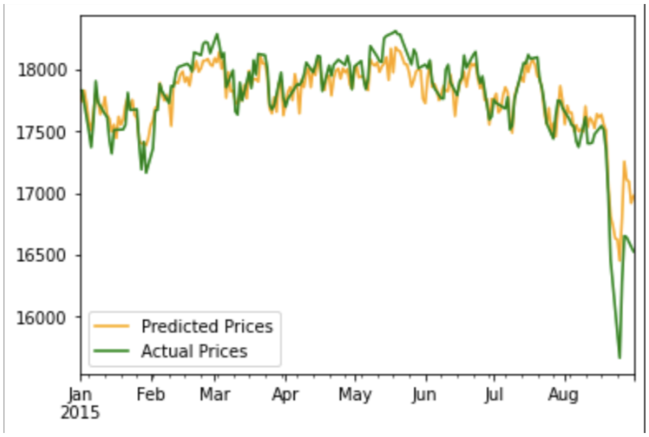

Fig. 12. Year 2015 Stock price prediction using Random Forest Regression

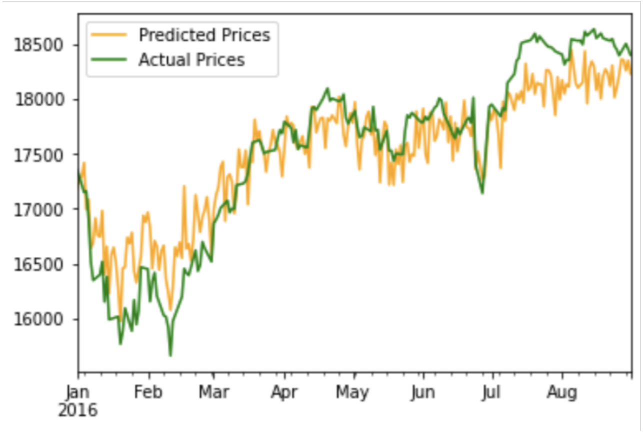

Fig. 13. Year 2016 Stock price prediction using Random Forest Regression

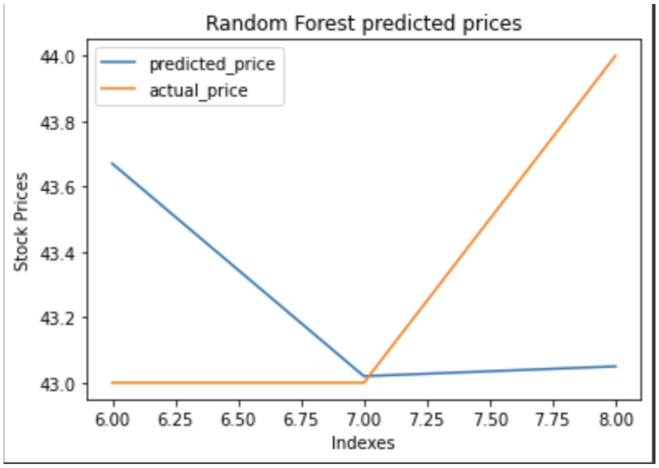

Fig. 14. Random Forest Predicted Price using live Twitter sentiment analysis

IX. CONCLUSION

By comparing the accuracy of the different algorithms, we discovered that the random forest algorithm is the most effective algorithm for predicting the value of a stock backed by numerous data points from historical data. Since the algorithm has been selected after being evaluated on a sample of historical data and has been trained on a sizable collection of historical data, brokers and investors will find it useful for investing money in the securities market. Our findings indicate that shifts in public sentiment can have an impact on the market, suggesting that there is a good probability that we will be able to predict the stock exchange.

REFERENCES

[1] Ashish Sharma, Dinesh Bhuriya, Upendra Singh. ”Survey of exchange Prediction Using Machine Learning Approach”,ICECA,2017.

[2] Loke. K.S. “Impact of monetary Ratios And Technical Analysis On Stock Price Prediction Using Random Forests”, IEEE,2017.

[3] Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, “Stock Market . Prediction via Multi-Source MultipleInstanceLearning.” IEEE 2018.

[4] VivekKanade, BhausahebDevikar, SayaliPhadatare, PranaliMunde, ShubhangiSonone. “Stock Market Prediction: Using Historical Data Analysis”, IJARCSSE 2017.

[5] Jabaseeli, A. Nisha, and E. Kirubakaran. “A Survey on Sentiment Analysis of (Product) Reviews.” International Journal of Computer Applications 47.11, 201

[6] Agarwal, Apoorv, et al. “Sentiment analysis of twitter data.” Proceedings of the Workshop on Languages in Social Media. Association for linguistics, 2011.

[7] a. Mittal and a. Goel. “Stock Prediction Using Twitter Sentiment Analysis.” Tomx.Inf. Elte.Hu, (June), 2012.