Novel Method of Stock Price Prediction and Recommendation

Shereen Fathima A[1]

Undergraduate, B. Tech Computer Science and Engineering
B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai 48

Tasnim Tabasum R[2]

Undergraduate, B. Tech Computer Science and Engineering
B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai 48

Abstract*-- To be able to forecast the stock prices of a particular company as it might turn out the next day based on the implementation of the Machine learning algorithms and News data gathered from the New York Times API using Python. This methodology of stock prediction is to accurately predict the stock prices initially by implementing Machine learning and time series algorithm on the historical stock prices of the company; the second step is to predict the stock prices, using both the historical stock prices of the company and the extracted news data from NYT API, as a dependant variable for the stock price prediction for making informed decisions as the news data extracted from NYT is authentic and can be used as a dependant factor for stock data forecasting. Once the stock prices are predicted, the users are then provided with an option of Buying or Selling the stock using the Recommendation module. This project aims at being an all in one stock market aid project, helpful for the users in making sound financial decisions. The entire project is then deployed on the Flask environment as a web interface for better usability.

***Keywords*—Machine Learning, Python, Stock, Prediction**

I. INTRODUCTION

Data analytics is for inspecting raw data with the purpose of drawing conclusions with the records. It's far mostly utilized in stock market, many companies and corporations to make higher enterprise choices. It enables them to decide the effect on sales, profits, value reduction, new product and offerings.

Stock costs vary swiftly with the change in international market economy. The stock marketplace is a constantly converting indicator of world economic activity. There are limitless stocks which can be offered and sold each day, and these transactions decide market fees. In order to get an idea about the way the stock market is faring, there are stock indices that aggregate the overall performance of units of key stocks of various companies, along with the Dow Jones Industrial Average (DJIA), the S&P 500, which aggregate the overall performance of the stock markets and so forth. In view that these indices are representative of the overall economy, they may be also suffering from all styles of events inside the international, whether it be elections, army activity, profits announcements, new product launches, and so forth. As our primary supply of this form of facts, the news plays an exceedingly crucial role inside the determination of inventory marketplace pastime.

The trouble to hand here is trying to apply AI techniques to allow inventory market prediction. The usefulness of this lies within the fact that if you had the capability to reliably predict stock marketplace motion, you can buy monetary gadgets that wager in the marketplace, and in case your predictions are right, you stand to advantage cash. The enter output of this model at a high stage is information and different key signs.

The assessment metric for this mission will be the accuracy with which the version can expect the market going up or down the next day primarily based on visible statistics as of a given day. The dataset feeding that is a combination of NYT - sourced international news headlines and DJIA returns that became shared with the yahoo-finance and a few derived capabilities.

Inventory marketplace prediction stays a secretive and best few is familiar with it efficaciously. Very few humans, if any, are willing to share what a success techniques they have. A primary intention of this project is to add to the knowledge of inventory marketplace prediction. The wish is that with an extra understanding of how the market actions, traders might be better prepared to save you another financial crisis. The undertaking will evaluate some existing strategies from a rigorous scientific attitude and provide a quantitative evaluation of recent techniques.

It's far vital right here to outline the scope of the project. Even though important to any investor running inside the real international, no strive is made in this mission at portfolio management. Portfolio management is essentially a further step executed after an investor has made a prediction on which course any unique stock will flow. The investor can also choose to allocate budget throughout a variety of shares in such a manner to minimize his or her chance. A more commonplace technique would be for an investor to make investments across an extensive variety of stocks primarily based on some criteria he has determined on earlier than. These assignments will consciousness exclusively on predicting the day by day fashion (price motion) of character stocks. The assignment will make no attempt to finding out how plenty cash to allocate to every prediction. More so, the project will analyze the accuracies of those predictions.

The two analyses are technical evaluation and fundamental analysis. Technical evaluation considers past charge and extent to be expecting the future trend where as fundamental evaluation. However, fundamental evaluation of a business entails studying its monetary records to get some insights. The efficacy of each technical and fundamental analysis is disputed by using the green-market hypothesis which states that inventory marketplace costs are essentially unpredictable. The motive of this mission is to be with a purpose to forecast the inventory fees as they could flip out based at the data accumulated by means of studying the feelings generated from information and to endorse the right shares to the buyers. Stock prices fluctuate unexpectedly with the change in world market economic system. There are more than one methodology to expect and endorse the stock price versions, but in this project, NY times' (NYT) information articles headlines is used as a dependant thing to be expecting the trade in inventory prices. New York Times archive application programming interface (API) is used to accumulate the information internet site articles records over the span of 10 years ranging from the 12 months 2008 till date March 2018. The preliminary procedure is in order to extract and download the inventory information of Dow Jones business common (DJIA) from the yahoo finance internet site, then extract the news articles from the New York Times API in the form of a JSON (JavaScript Object Notation) file. That has a ten year span. Upon the extraction, both these documents are merged as a single entity that's then utilized for the stock prediction and advice. The algorithms to be implemented on this challenge are the numerous machine learning algorithms and artificial intelligence algorithms that are SVR (Support Vector Regression) model, linear regression, ARIMA (Auto Regressive Integrated Moving Average) version, random forest graph and Naïve Bayes algorithm. Post the inventory records prediction, based totally on the variations, the advice of whether or not or not to buy the stock in phrases of purchase and promote values, the stock is suggested to the person, and that is then found to be beneficial for making knowledgeable selections on main inventory transactions. To make things easier and usable, all the processes are then deployed over the web using Flask Interface.

II. OBJECTIVE

The major objectives that led towards the development of the automated stock prediction system is to be able to forecast the stock prices of a particular company as it might turn out the next day based on the implementation of the Machine learning algorithms and News data gathered from the New York Times API using Python. This methodology of stock prediction is to accurately forecast the stock prices initially by implementing Machine learning

and time series algorithm on the historical stock prices of the company; the second step is to predict the stock prices, using both the historical stock prices of the company and the extracted news data from NYT API, as a dependant variable for the stock price prediction for making informed decisions as the news data extracted from NYT is authentic and can be used as a dependant factor for stock data forecasting. Once the stock prices are predicted, the users are then provided with an option of Buying or Selling the stock using the Recommendation module. This project aims at being an all in one stock market aid project, helpful for the users in making sound financial decisions. The entire project is then deployed on the Flask environment as a web interface for better usability. The Stock prices fluctuate rapidly with the change in world market economy. There are many ways and methodologies to forecast the stock prices and recommend the right stock , but in this project, The stock prices are initially simulated by using the historical stock prices at first, then the New York Times' (NYT) news articles headlines is used as a dependant factor to predict the change in stock prices as it is authentic and reliable means to predict the stock price fluctuations. NY Times Archive Application Programming Interface (API) is used to gather the news website articles data over the span of 10 years ranging from the year 2008 till date March 2018. The initial process is to be able to extract and download the stock data of any company from the Yahoo Finance website, then stock analysis is done by implementing the Machine learning algorithms SVR Model, Linear Regression and a time series ARIMA Model, then the stock data is pickled with the news articles from the New York Times API in the form of the JSON (JavaScript Object Notation) that has a 10 year span. Upon the extraction, both these files are merged as a single entity which is then utilized for the stock prediction and recommendation. The algorithms to be implemented in this project are the various Machine Learning algorithms and Artificial Intelligence algorithms which are Random Forest Graph and Naïve Bayes algorithm. Post the stock data prediction, based on the variations, the recommendation of whether or not to buy the stock in terms of BUY and SELL values, the stock is recommended to the user.

III. WORKING

The system consists of three main components: a data extraction/crawler, a simulation system, prediction and a recommendation in a web interface.

Data Extraction

The python libraries extract data from online sources. The data collected from websites and parsed information is stored in CSV files. The data extraction frequency and data sources are configurable.

In an initial phase, a large number of websites were studied and the ones most suitable for the project were identified. The following sections outline characteristics of each data source and list some examples.

Data Source	Type	URL
Yahoo Finance Historical Prices	Technical	Finance.yahoo.com
The New York Times	Technical	Nytimes.com

Table 1 Data Collection sources

New York Times

New York Times offers timely news and good coverage of the important news affecting stock trading. In contrast to many other websites, their company specific news archives are easily traversable and date back many years. This was an important criterion for the project, as stock market simulations require large historical datasets to be reliable. Furthermore, neither of these news websites relies heavily on JavaScript, thus simplifying the extraction task.

Yahoo Finance Historical Prices

Yahoo Finance consist of daily opening, high, low and closing prices and have been adjusted for stock splits and dividends. The more fine grained resolution of OpenTick (including minute-

frequency historical data) was more desirable, but was abandoned because of periods of missing prices and some price inconsistencies when compared to services like Google finance. Before storing data into the CSV, the news and analyst crawlers perform some preprocessing in order to extract the relevant information. The quotes crawler does not need this phase, as Yahoo's historical quotes are conveniently available in CSV format.

Preprocessing the News

The goal of the new preprocessing phase is to parse headlines and their exact timestamps from the raw data given in the form of pickle file. Below are some practical considerations that came up during the implementation of this phase.

Simulation system

The simulation system loads information from the CSV files and runs trading strategies such as sentiment analysis and prediction. The entire system will be developed using python.

The Yahoo Finance the stock data are extracted since it is an open source which can be accessed by all. The New York Times (NYT) is a newspaper of records where specifically the articles relevant to stock were extracted. The stock data extracted from the Yahoo Finance and New York Times are stock stored in local storage unit. This local storage unit does stock data analysis and sentiment analysis to predict and recommend it to investors. The SVR, ARIMA model and linear regression are the algorithms for analyzing stock data.

Support Vector Regression (SVR) model classifies the data to predict real values. Linear regression is used for classifying the single independent variable to predict the value of a dependent variable. Auto Regressive Integrated Moving Average (ARIMA) model is used for capturing the different standards of temporal and structural series as a time series data representation for stock data analysis.

Stock Data Analysis

The aim of the stock data analysis is to extract the stock data of any particular company of the set time period from the Yahoo Finance website. The analysis of the extracted stock data includes the stock data to be implemented in the machine learning algorithms SVR – Support Vector Regression and Linear Regression, along with a time series ARIMA – Auto Regressive Integrated Moving Average modeling to get the forecasted stock prices.

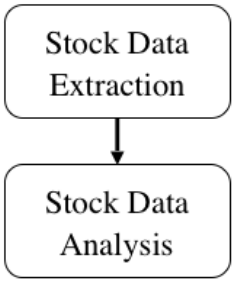

Fig 1 Stock Data Analysis

Sentiment Analysis

Once the stock data analysis is done, the news is extracted from the New York Times API, pickled along with the stock data of the particular company and then the NLTK package sentiment score analyzer is then implemented on the pickle file that sets a sentiment score for the news articles in the file.

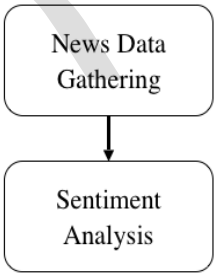

Fig 2 Sentiment Analysis Module

Stock Data Prediction

The sentiment analysis of the pickled file which is a combination of the stock data of a particular company along with the news articles is then assigned with the sentiment score using the NLTK package. Once the sentiment score is allotted to the news articles, it is then used for stock market prediction. The major objective of this is to

know if the stock market prices are actually affected by the news articles using the Random Forest and Naïve Bayes algorithms.

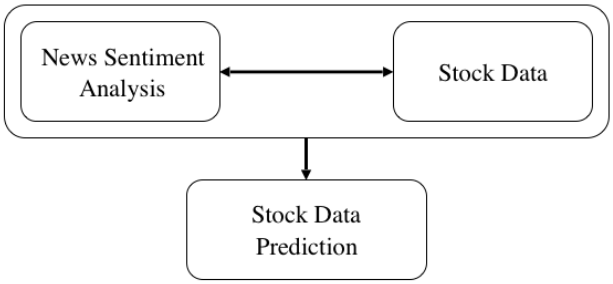

Fig 3 Stock data prediction module

Stock Recommendation

The predicted stock data recommends us to invest on the stocks which will give maximal profit or minimal profit or loss, whether to invest or to hold on. Based on the simulated values of stock prices, the artificial learning algorithm is then used to specify if the stock on the present day should be bought or sold i.e. being able to help the users make decisions that can be profitable.

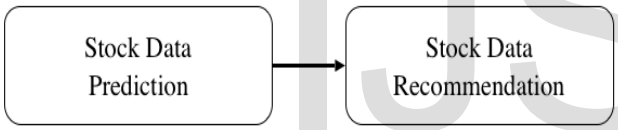

Fig 4 Stock recommendation module

Flask Deployment

The modules, all put together are then to be integrated with the Python Flask web interface for the users to access the data real time. The requirements to develop the Flask web interface are Python3, Virtual Environment, Pip Installer and a Flask development server

With the components, the Flask web interface that encapsulates all the functionalities that aid in stock price prediction and recommendation is developed. The website has access to all the stock related data that range from stock data download, visualize, sentiment score initialization, prediction and recommendation with a login page.

IV. EXISTING SYSTEM

Existing work on stock price prediction and recommendation based on news sentimental analysis does not generate information about large volume of data up to date and no string index value is obtained.

- Assumption
- Fundamental Analysis
- Technical analysis
- Social media extraction to predict stock prices

ASSUMPTION

There are several assumptions regarding the stock prediction and recommendation from the Yahoo Internet Interface which only assumes which are the best one for the consumers to invest based on the product values. But there are only ideas and there is no other platform for helping the investors.

FUNDAMENTAL ANALYSIS

The basic fundamental analysis that is done on the historical stock prices of the raw data that is by the weekly closing prices of the stock. Depending on how many have searched, willing to invest in stock market and filters. The complete history will be monitored and the stock price forecasting is analyzed. The basic analyses are carried out based on the investor's interests who want to invest.

TECHNICAL ANALYSIS

The technical analysis is the study and analysis of the internal data of the stock market, considering that all the economic, financial, political and psychological factors surrounding the sector which are incorporated into a single element - the share quotation. The people who study and observe the technical analysis of the stock prices - technical analysts, study the short-term changes of the shares' price,

starting with a study of the history of the quotations, within a set interval that ranges from 3 to 6 months, and assume that the past behavior will extend or reflect into the future. The technical analysis offers information about the possible future evolution of the stock market.

SOCIAL MEDIA

The social medias like Facebook, Twitter, YouTube and other online shopping websites where the reviews, ratings, comments based on the consumers and helps to extract stock prices to predict and invest. At times the reviews, ratings and comments may be paid and false in order to sell it.

V. PROPOSED SYSTEM

The first step in developing a project is to understand the objective which involves an understanding of the intent and essentials of a system to be developed. This comprehension is used as a problem description and a preparatory system to accomplish the expectations. The objective of our project is neither to build a system that makes billions nor to waste billions too. But the objective is to develop a system that finds the direction of change of stock price indices based on the co-relations between stock prices and help the investors in the stock market in taking a decision whether to buy/sell/hold a stock by providing the results in-terms of visualizations.

To overcome all the drawbacks of the existing stock rate prediction system, in proposed system we introduce the method of helping the investors to know about the up to date stock indices rise and fall through the website and when to invest at the right time. We propose sentiment analysis framework to crawler program that obtains the news headlines from the NY Times archive API. The output is fed to machine learning algorithm and the probable stock value is predicted. Neural network based recommendation system that suggests whether to buy, sell, hold the stock based on the situation.

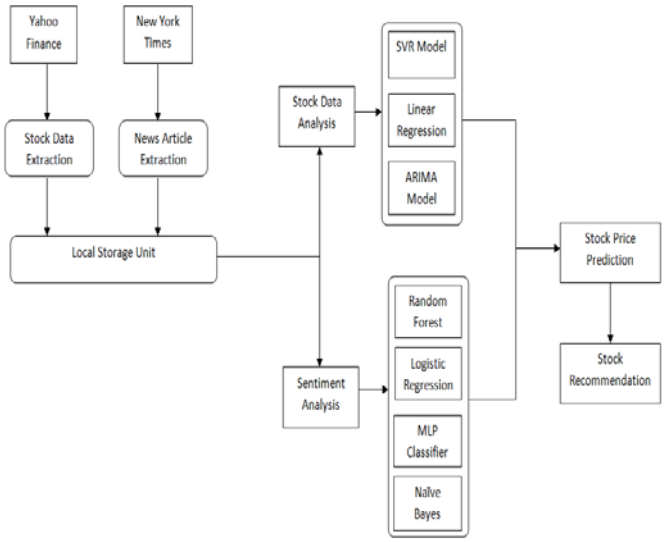

Fig 5 System Architecture

Data Gathering

In this project, there are two types of data extracted and gathered in the 10 year duration, from the year 2008 till 2018. The data gathered is of 10 year span to enhance the efficiency of the system and provide sufficient training data.

1. Stock indices: The stock indices are the historical stock prices of a particular company. The stock prices of each day, along with the highest, lowest price of the stock on a day, with the opening and closing value is obtained. The major focus is on the closing price of the stock on a given day. The stock price data is extracted from the Yahoo finance website. Initial stock analysis is done by only using the historical stock data.

2. News data: The second type in the stock prediction is the usage of a dependant variable – news articles. There are very few news articles and data available over the internet which is open for the public use. From our research the best openly available data that is accurate and viable source of news data which is also open source, and could be appropriately used in stock prediction was from

the NY Times Archive API. The news articles are downloaded in the form of JSON files.

Data Processing

News articles extracted from the NY Times archive API contain the data in the form of categories represented by sections. Some of the sections contain some irrelevant categories of articles, which are not related to stocks at all, such as Apartment listings, Movies, Biography, Jokes, Obituary, TV or radio schedules. Therefore, we remove such articles from the set of the extracted news data. The only sections of importance that are to be considered from the extracted news articles are: 'Tech', 'National', 'World', 'U.S.', 'Business', 'Politics', 'Shares', 'Opinion', 'Stock', 'Science', 'Health' and 'Foreign'. The news articles fro under these sections are considered for the sentiment analysis.

Pickling the stock data with the news articles:

After having separated the unwanted news articles, a single string is formed from concatenating all the articles headlines of a single day. After the news string of a single day is obtained, it is then pickled with the appropriate date (time series) and the stock market data of the particular company. This process of combining the news articles in string format along with stock data in the integer format is called Pickling of the file.

Sentiment Analysis

For the sentiment analysis, the Natural Language Toolkit (NLTK) package in python is used in this project. The major use of the NLTK package is for classifying emotions or behavior from the texts generated through natural language processing. VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Analyzer, which is a part of the NLTK package, is used to initialize a sentiment score for the text strings from the articles and specifies a sentiment score in terms of positive, negative and neutral scoring for the particular news data string.

Algorithms - Stock Data Analysis

1. Support Vector Regression (SVR)

The initial method of stock data analysis is by implementing the Support Vector Regression. It is a classification algorithm, but it is applied to predict real values rather than a class of values. SVR takes in the presence of non-linearity in the data and provides a proficient prediction model. There are 3 models in the SVR classification that are RBF, Linear and Polynomial model. It is used to check the predictability of the prices against the actual stock values. It does not forecast the prices but rather compares the actual data points to the predicted data points. The RBF model is found to be accurate compared to the other two models.

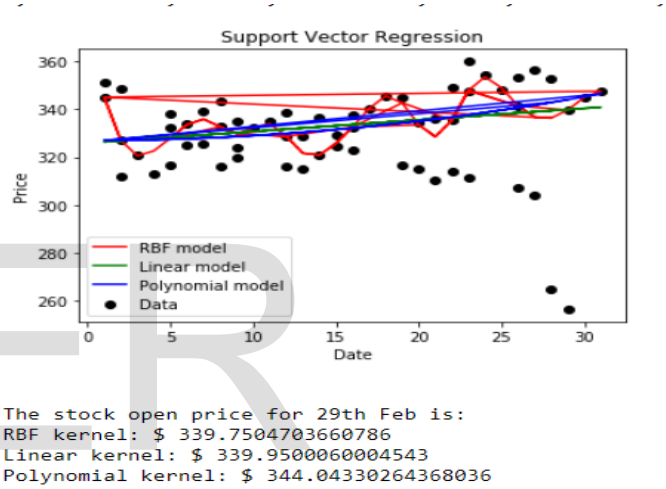

Fig 6 SVR Model

2. Linear Regression

The second method of stock data analysis is the implementation of the Linear Regression machine learning algorithm. It is used in situations where the data is of continuous nature. Linear Regression is implemented for predicting and recommending the best area to invest. This algorithm is for decision-making processes. On the implementation of this algorithm for Stock data prediction, the time period of stock forecast is specified and the future trend is simulated in a graphical format. This helps get a visual idea, whether the stock data goes uphill or in a negative curve. It particularly shows the probable future trend of the stock, as represented in the figure 7.

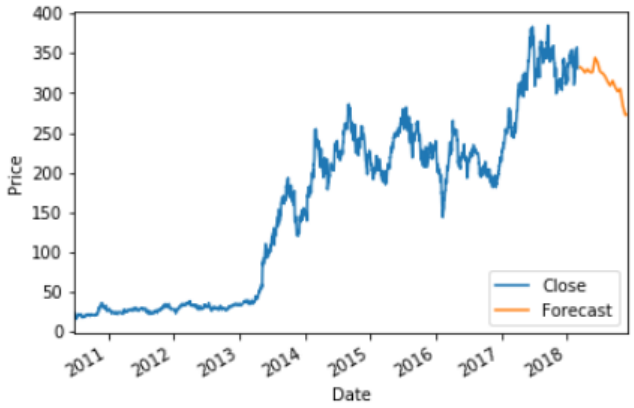

Fig 7 Linear Regression Prediction

3. Autoregressive Integrated Moving Average Model (ARIMA)

The third step in the stock data forecast is implementing a time series modeling methodology. To achieve this, ARIMA model forecasting is used. An ARIMA (Autoregressive Integrated Moving Average) model is a combination of statistical methodologies for analyzing and forecasting time series data. This model visually represents the stock data trend, seasonality and correlation level of the data. While the previous methodology represented a single line that denoted the probable growth of the stock curve, by using the ARIMA model, the confidence level along with predicted growth curve is denoted, which is called as the stop loss order. It shows the maximum range the stock prices might raise or fall as a method of correction, which gives more accuracy as it is backed with the confidence level in the visual representation, as shown in the fig 8.

Fig 8 ARIMA model prediction

Algorithms – Sentiment Analysis

The sentiment analysis of the pickled file which is a combination of the stock data of a particular company along with the news articles is then assigned with the sentiment score using the NLTK package. Once the sentiment score is allotted to the news articles, it is then used for stock market prediction. The major objective of this is to know if the stock market prices are actually affected by the news articles using the Random Forest and Naïve Bayes algorithms.

1. Random Forest

Random Forest algorithm is an ensemble learning method utilized for regression. The Random Forest algorithm operates by constructing a multitude of decision trees during the training time and the output obtained is the mean regression of the individual trees. When the pickled file containing the news data and the stock prices was implemented under the Random Forest algorithm, with the testing and training data segregated, the news articles scores and the stock price were visually represented in the form of a graph. It shows a positive correlation. This shows that any change in the news affects the stock prices.

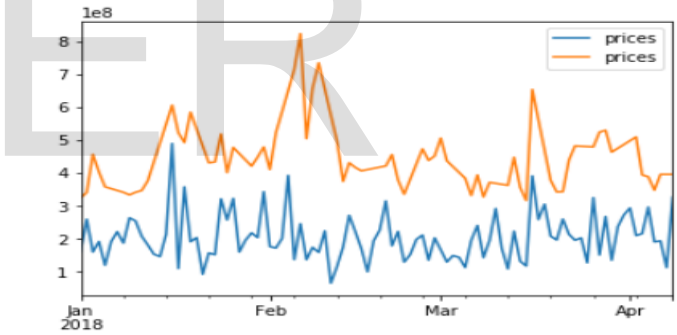

Fig 9 Random Forest Prediction

2. Naïve Bayes

To be able to predict the accuracy of the prediction, Naïve Bayes algorithm is used. It is to know the level of correctness between the predicted and the actual values. The Naïve Bayes learning scheme performs well on most classification tasks, and is often significantly more accurate than more sophisticated methods. It often assigns maximum probability to the correct class. This states that the

algorithm's performance might be limited to problems where the output is seemingly categorical in nature. In the implementation, the predicted prices are very much similar to the actual values that denote the algorithm's capability to predict the values accurately.

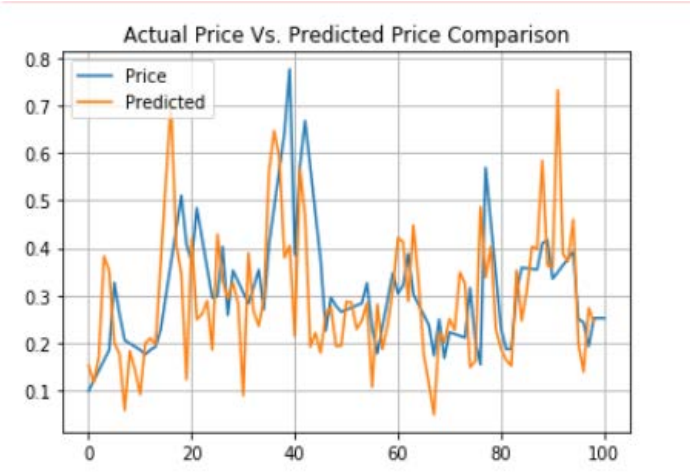

Fig 10 Naïve Bayes Prediction

Recommendation

The predicted stock data recommends us to invest on the stocks which will give maximal profit or minimal profit or loss, whether to invest or to hold on. Based on the simulated values of stock prices, the artificial learning algorithm is then used to specify if the stock on the present day should be bought or sold i.e. being able to help the users make decisions that can be profitable.

Out[1]:

Date	Price	Regime	Signal
2018-01-03	24922.679688	1.0	Buy
2018-01-08	25283.000000	-1.0	Sell
2018-01-09	25385.800781	1.0	Buy
2018-01-10	25369.130859	-1.0	Sell
2018-01-11	25574.730469	1.0	Buy
2018-01-16	25792.859375	-1.0	Sell
2018-01-17	26115.650391	1.0	Buy

Fig 11 Stock Recommendation

Flask web interface

The modules, all put together are integrated with the Python Flask web interface for the users to access the data real time. The requirements to develop the Flask web interface are:

- Python3
- Virtual Environment
- Pip
- Flask development server

With the components, the Flask web interface that encapsulates all the functionalities that aid in stock price prediction and recommendation is developed. The website has access to all the stock related data that range from stock data download, visualize, sentiment score initialization, prediction and recommendation with a login page.

VI. CONCLUSION

Finding and simulating the future trend and the movement of a stock for a particular company may be a seemingly difficult task to accomplish, as the outcome of the stock prices depend on a number of factors. From this project, it is clear that stock prices and news articles are correlated, that is; news articles and stock prices are positively correlated. As it is clear that both news and stock prices are associated with each other, and are also authentic, news data is kept as a dependant variable to predict the outcome of the stock prices. So, we totally study this relationship and determine that stock trend may be foretold by the news articles and former stock price history.

As news articles capture sentiment regarding this market, we have a tendency to modify this sentiment detection and supported the words within the news articles; we will get associate overall news polarity. If the news is positive, then we will state that this news impact is uphill within the market, thus additional possibilities of stock worth go high. And if the news is negative, then it should impact the stock worth to travel down in trend.

This project demonstrates a systematic method of stock value prediction that initially starts with a machine learning algorithm SVR model to predict the accuracy of the existing prices with the modeled prices. Linear Regression is used to predict the future stock trend; the third step is to implement a time series ARIMA Model to calculate the trend along with the confidence level with the stop loss order to represent a safe limit of investment.

The stock prices are pickled with the news data gathered and the machine learning algorithms like Random Forest and Naïve Bayes algorithm are applied for stock market value prediction. Then the users are given the option of whether to buy or sell the particular stock, represented by the recommendation module. This expected and counseled stock information is made in an exceedingly full-fledged setting in the form of a web interface using Flask, for enhancing the usability and for making sound financial decisions.

VII. REFERENCE

[1] Nicolas Pröllochs, Stefan Feuerriegel, Dirk Neumann, "Negation scope detection in sentiment analysis: Decision support for news-driven trading", Decision Support Systems, vol. 29, no.2, pp. 38-47, 2016.

[2] Xiaodong Li, Haoran Xie, Li Chen, "News Impact on Stock Price Return Via Sentiment Analysis", Knowledge Based System, vol. 52, no.4, pp. 54-63, 2016.

[3] Sahar Sohangir, Dingding Wang, Anna Pomeranets, Taghi M. Khoshgoftaar, "Big Data: Deep Learning for financial sentiment analysis", Semantic Computing, vol. 73, no. 8, pp. 56-60, 2018.

[4] Yauheniya Shynkevich, Sonya Coleman, T.M McGinnity, "Predicting Stock Price Movement Based on Different Categories of News Articles", Computational Intelligence, vol. 978, no. 15, pp. 703-710, 2017.

[5] Kalyani Joshi, Prof. Bharathi, "Stock Trend Prediction using News Sentiment Analysis", vol. 08, no. 03, pp. 67-76, 2016.

[6] R. Yamini Nivetha, C. Dhaya, "Developing a Prediction Model for Stock Analysis", Technical Advancements in Computers And Communications, vol. 998, no.17 , pp. 1-3, 2017.

[7] Xiongwen Pang, Yanqiang Zhou, Pan Wang, Weiwei Lin, Victor Chang, "An innovative neural network approach for stock market prediction", Supercomputing, vol. 228, no. 228, pp. 1-21, 2018.

[8] Mehek Usmani, Syed Hasan Adil, "Stock Market Prediction using Machine Learning Techniques", Computer and Information Sciences, vol. 978 , no. 07, pp. 322-327, 2017.

[9] Ayman E.Khedr, S. E. Salama, "Predicting Stock Market Behaviour using Data Mining Techniques and News Sentiment Analysis", vol. 09, no. 07, pp. 22-30, 2017.

[10] Yefeng Ruan, Arjan Durresi , Lina Alfantoukh, "Using Twitter trust network for stock market analysis", Knowledge Based Systems, vol. 145, no. 43, pp. 207-210, 2018.