STOCK PRICE PREDICTION USING DEEP LEARNING AND SENTIMENT ANALYSIS

MAJOR PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the degree

Of

BACHELORS OF TECHNOLOGY

In

COMPUTER SCIENCE & ENGINEERING

By

ROHIT GUPTA	PARTH PATHAK	RISHABH MISHRA
160BTCCSE013	160BTCCSE003	160BTCCSE025

GUIDED BY

DR. ALPANA JIJJA

CSE DEPARTMENT

SCHOOL OF ENGINEERING AND TECHNOLOGY(SET)

ANSAL UNIVERSITY

2016 - 2020

DECLARATION

It is hereby certified that the work which is being presented in the B. Tech Major Project Report entitled **"STOCK PRICE PREDICTION USING DEEP LEARNING AND SENTIMENT ANALYSIS"** in partial fulfilment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY** and submitted in the **DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING – SCHOOL OF ENGINEERING AND TECHNOLOGY(SET) , ANSAL UNIVERSITY** is an authentic record of our own work carried out during a period from **June 2019 to November 2019** under the guidance of **Dr. Alpana Jijja, CSE Department.**

The matter presented in the B. Tech Major Project Report has not been submitted by us for the award of any other degree of this or any other Institute.

ROHIT GUPTA	PARTH PATHAK	RISHABH MISHRA
160BTCCSE013	160BTCCSE003	160BTCCSE025

CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge. They are permitted to appear in the External Major Project Examination.

Dr. Alpana Jijja

(Project Guide, CSE)

The B.Tech Major Project Viva-Voice Examination of **ROHIT GUPTA (160BTCCSE013), PARTH PATHAK (160BTCCSE003) & RISHABH MISHRA (160BTCCSE025)** has been held on **…….**

(Project Coordinator) **(Signature of External Examiner)**

ABSTRACT

The stock market price prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. In this paper, we propose a Machine Learning (ML) approach that will be trained from available stocks data, gain intelligence and then uses the acquired knowledge for accurate prediction. After the through research of various algorithms and their fitness for different problem domains, Recurrent Neural Network (RNN) was found to be the most practical consideration. Neural network models having the features and customize parameters makes it possible to implement wide number of features along with the cross validation sets.

We test a hypothesis based on the premise of behavioral economics, that the emotions and moods of individuals affect their decision making process, thus, leading to a direct correlation between" public sentiment" and "market sentiment". We perform sentiment analysis and use these moods and previous days' stock values to predict future stock movements.

This work will enable researchers in this field to know the current trend as well as help to inform their future research efforts.

The project attempts to predict whether a stock price sometimes in the future will be higher or lower than it is on a given day. We find a little predictive ability in the short run but definite predictive ability in the long run.

ACKNOWLEDGEMENT

Team effort together with precious words of encouragement and guidance makes daunting tasks achievable. It is a pleasure to acknowledge the direct and implied help we have received at various stages in the task of developing the project. It would not have been possible to develop such a project without the furtherance on part of numerous individuals. We find it impossible to express our thanks to each one of them in words, for it seems too trivial when compare to the profound encouragement that they have extended to us.

We are grateful to Dr. Alpana Jijja (**CSE Dept.)**, for having given us opportunity to do this project, which was of great interest to us.

We would again like to extend our gratitude to Dr. Alpana Jijja, for believing in us and providing motivation all through the development of this project. Without her guidance this project would not be such a success.

At last we thank the almighty, who had given the strength to complete this project on time. Finally, we would like to thank our parents, all friends and well-wishers for their valuable help and encouragement throughout the project.

ROHIT GUPTA	**PARTH PATHAK**	**RISHABH MISHRA**
160BTCCSE013	**160BTCCSE003**	**160BTCCSE025**

INDEX

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Motivation

Though there are many stock prediction models already available in the market but they all use a single static dataset of stock prices for training and hence predict stock prices. The motivation for this project is to build such a stock prediction mechanism which could be used for real-life stock prediction problems and take in consideration the requirements of its user without compromising on the accuracy. It should be able to predict stock prices for any date input by the user rather than just measuring the predictability of stocks of an organization using machine learning algorithms on a static dataset. Thus this project aims at dynamic gathering of data which could be used for training of stock prediction model according to the dates specified by the user and then make predictions about the stock prices using machine learning regression techniques. The main aim of this project is to create a user-friendly application where the user can specify the date for which he/she wants prediction of stock prices and get the corresponding results visualized in an easy to understand manner so they can be useful to even such users that are not from financial analytics or machine learning background. To make out for the only-mathematical predictions made by machine learning regression techniques, this project also includes sentiment analysis of news articles on the organization for which stock prices are to be predicted.

1.2 Literature Review

This paper talks about the World of Web Scraper, Web scraping is related to web indexing, whose task is to index information on the web with the help of a bot or web crawler [1] the Web Crawler to fetch the desired links, the data extractor to fetch the data from the links and storing that data into a csv file. The Python language is used for the implementation.

The data on the sites can be found in tables, articles, comments, nested in different HTML tags, etc. Gathering a large amount of data from the web is not an easy task, but it is a good way to collect information which can be used in further analyzes. [2] Đorđe Petrović and Ilja Stanišević deal with the process of web scraping data from different

locations on the Internet and their storage in a database, for the purpose of collecting and analyzing data of the used cars market.

Yahya Eru Cakra and Bayu Distiawan Trisedya in [3] predict the Indonesian stock market using simple sentiment analysis. Naive Bayes and Random Forest algorithm are used to classify tweet to calculate sentiment regarding a company. The results of sentiment analysis are used to predict the company stock price. They use linear regression method to build the prediction model.

In this paper [4], by applying linear regression for forecasting behavior of TCS data set, Dinesh Bhuriya, Girish Kaushal and Ashish Sharma prove that their proposed method is best to compare the other regression technique method and the stockholders can invest confidentially based on that.

M M. Goswami, C K. Bhensdadia, A P Ganatra have proposed [5] a novel model that tries to predict short term price fluctuation, using candlestick analysis. This is a proven technique used for short term prediction of stock price fluctuation and market timing since many years. Their approach has been hybrid that combines self-organizing map with case based reasoning to indemnify profitable patterns (candlestick) and predicting stock price fluctuation based on the pattern consequences.

This paper [6] proposes algorithms for the extraction of features from candlestick patterns for technical analysis of share indices. The significant features consist of: the direction of candlestick, the gap between CLOSE and OPEN price of two candlesticks, the body level of current and previous candlesticks, and the length of the candlesticks.

Hiransha M. et. al. showed the usage of four different Deep Learning architectures [7] for Stock Market Price Prediction thus, comparing the different models it was found that CNN performed the best.

RNN [8] takes input from two sources, one is from the present and the other from the past. Information from these two sources are used to decide how they react to the new set of data. This is done with the help of a feedback loop where output at each instant

memory**.** Each input sequence has plenty of information and this information are stored in the hidden state of recurrent networks.

LSTM [9] is a special type of RNN. These networks are proficient in learning about long-term dependencies. These networks are clearly designed to evade the long- term dependency problem, but remembering information for a long time period back is their normal behavior.

Murtaza Roondiwala et. al. in his [10] paper discusses the implementation of Recurrent Neural Networks along with Long Shot-term Memory. Further, Root Mean Square Error for analyzing the efficiency of the system where the error or the difference between the target and the obtained output value is minimized by using RMSE value is used.

Behavioral economics [11] tells that **e**motions can greatly affect individual behavior and decision- making**.** Bollen et. al. investigated whether measurements of collective mood states derived from large scale Twitter feeds are correlated to the value of the DJIA over time.

K.K. Suresh Kumar et. al. used prediction algorithms and functions to predict future share prices and compares their performance. The results from analysis showed that used isotonic regression function offers the ability to predict the stock prices more accurately than the other existing techniques [12].

As discussed in [13] the stock prediction can be done by using fundamental and technical analysis. The fundamental analysis assumes that the investors are more logical and stock price (current and future) depends on its intrinsic value. Technical analysis evaluates the stocks by analyzing statistics generated by market activity, past prices, and volume. It looks for peaks, bottoms, trends, patterns, and other factors affecting a stock's price movement

Aditya Bhardwaj et al. in [14] have demonstrated sentiment analysis for stock market by fetching Sensex and Nifty live server data values on different interval of time that

1.3 Objectives

i. **Collecting Data**: - Collecting historical data of a particular stock is required to carry out further analysis and prediction tasks. For that purpose, web scraping is used as a tool to collect information (historical data) of a stock**.** Stock data consists of the open, high, low and close price for a particular day.

ii. **Selection of stock**: - For simplicity purposes, shares of Alphabet Inc. has been chosen. (parent company of Google).

iii. **Visualization of the stock data**: - Stock Visualization is very important to get the complete and clear picture of how the stock of a particular company performing. For stock visualization various graphs like Time-Series graph, Scatter Plots, Box and Whisker Plots, and Candlestick-OHLC graph have been plotted.

iv. **Pre-processing:** - Data pre-processing is an important phase in which missing data is filled, some columns are scaled, some columns are encoded and many more. All this is done to apply various machine learning models on the pre-processed data.

v. **Prediction of Stock Price using linear regression**: - Feeding open price to the Linear Regression Model will predict the High price of stock for that day.

vi. **Prediction using Recurrent Neural Network (RNN)**: - Previous days data is utilized to predict the prices for next day and so on. This model is more realistic as stock prices very much depends on the previous days' prices.

    i. RNN without Long short term memory (LSTM)
    ii. RNN with Long short term memory(LSTM)

vii. **Sentimental Analysis**: - Sentimental analysis using Twitter and other news sources and incorporating this extra feature in both the models.

viii. **Developing Web application**: - A complete web application that can efficiently demonstrate the data visualization, predictions made by the two models and their respective accuracy and many more.

1.4 Summary

Stock market prediction is basically defined as trying to determine the stock value and offer a robust idea for the people to know and predict the market and the stock prices. It is generally presented using the quarterly financial ratio using the dataset. Thus, relying on a single dataset may not be sufficient for the prediction and can give a result which is inaccurate. Hence, there is a need to contemplatee towards the study of machine learning with various datasets integration to predict the market and the stock trends.

The stock market prediction process is filled with uncertainty and can be influenced by multiple factors. Therefore, the stock market plays an important role in business and finance.

There are a lot of complicated financial indicators and also the fluctuation of the stock market is highly violent. However, as the technology is getting advanced, the opportunity to gain a steady fortune from the stock market is increased and it also helps experts to find out the most informative indicators to make a better prediction. The prediction of the market value is of great importance to help in maximizing the profit of stock option purchase while keeping the risk low. The stock market prediction process is filled with uncertainty and can be influenced by multiple factors. Therefore, the stock market plays an important role in business and finance. The technical and fundamental analysis is done by sentimental analysis process. Social media data has a high impact due to its increased usage, and it can be helpful in predicting the trend of the stock market. Technical analysis is done using by applying machine learning algorithms on historical data of stock prices.

There were always three methods to analyze and predict the stock market: financial, technical and sentiment.

i. Financial analysis evaluates past statements, reports and balances sheets and comparing it to prospects, the market and changes in government policy
ii. Technical analysis relies on the idea all factors which can influence the price are included in the current price of the stock and therefore no fundamental

information analysis is required. They believe that the prices move in trends and the same historic patterns

iii. Sentiment analysis relies on taking advice from experts and going through newspapers to monitor the stocks they'd like to invest in., going through text and data points to understand any changes that would move the price

1.4.1 Fin-Tech and AI

The benefits of using AI in financial services and operations by FinTech companies are extensive and recurring. From the marketing of their products and services to handling customer queries, AI along with Big Data and Predictive Analysis is quite instrumental in increasing the reach and scope of financial institutions across the world. Now, the entire loan process takes only a few hours, and if one is applying for a loan in the morning, the disbursal of loan money is taking place on the same date. Both speed and accuracy in the financial sector have increased manifold after the arrival of AI.

Adaptation to technology is the need of the hour, but offering products/services as per the needs of the market is the key to a successful business. That's why FinTech companies are moulding themselves as per the changing needs of the evolving market. They are working aggressively towards creating a more customer-friendly financial environment by removing the complexities in documentation and product delivery.

Unlike, the conventional banking firms that are still sticking to long paper works, FinTechs are growing speedily with the optimal utilization of AI, Big Data, and Predictive Intelligence. These technologies have become an integral part of FinTech firms, and their role is not limited to developing customized business solutions, they also play a pivotal role in protecting customers' personal information and providing security to financial assets of FinTech companies.

Previous academic literature has constrained sentiment analysis to relationships with equity returns without reference to underlying fundamentals.

Recurrent neural networks (RNN) have proved one of the most powerful models for processing sequential data.

Long Short-Term memory is one of the most successful RNNs architectures. LSTM introduces the memory cell, a unit of computation that replaces traditional artificial

to effectively associate memories and input remote in time, hence suit to grasp the structure of data dynamically over time with high prediction capacity.

Sentiment Analysis or Opinion Mining refers to the use of NLP, text analysis and computational linguistics to determine subjective information or the emotional state of the writer/subject/topic. It is commonly used in reviews which save businesses a lot of time from manually reading comments.

Just like Algorithmic Trading, Sentiment Analysis could also go very deep as a field. Besides just giving a positive/negative sentiment, it could understand how subjective a text is, the intensities of different emotions

CHAPTER 2: INTRODUCTION TO MACHINE LEARNING

2.1 Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task. Machine learning as illustrated in (Figure-2.1) is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

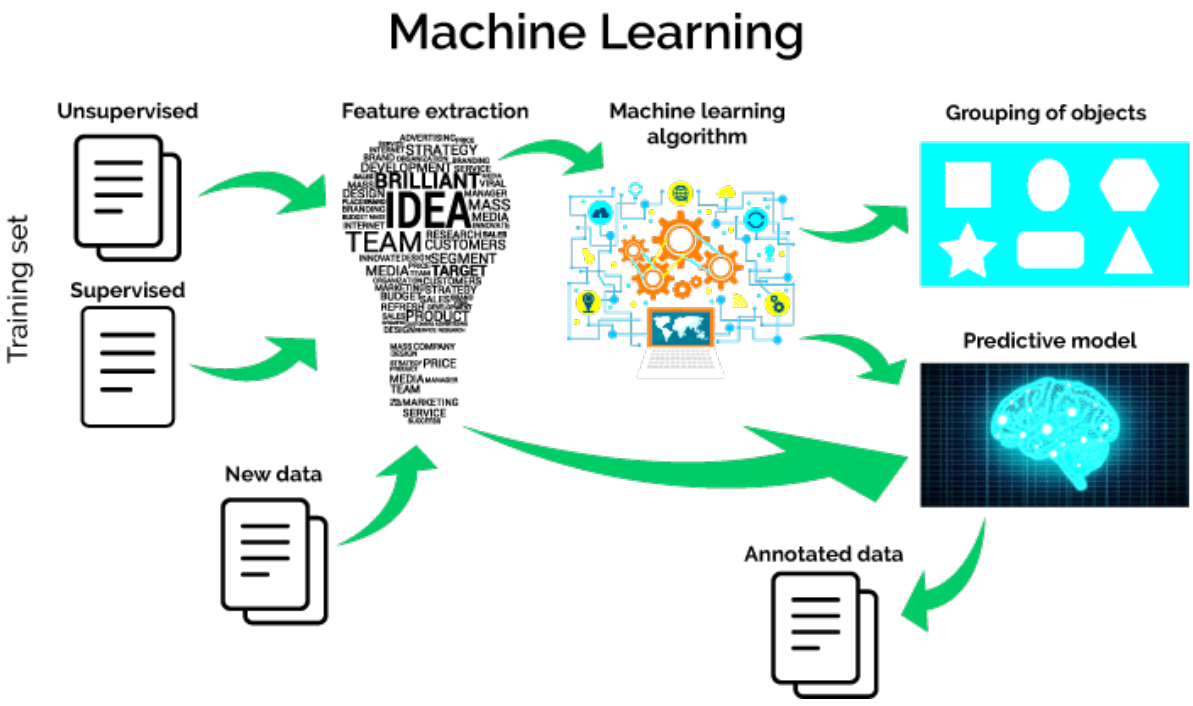

Figure - 2.1 Machine Learning

2.2 History of Machine Learning

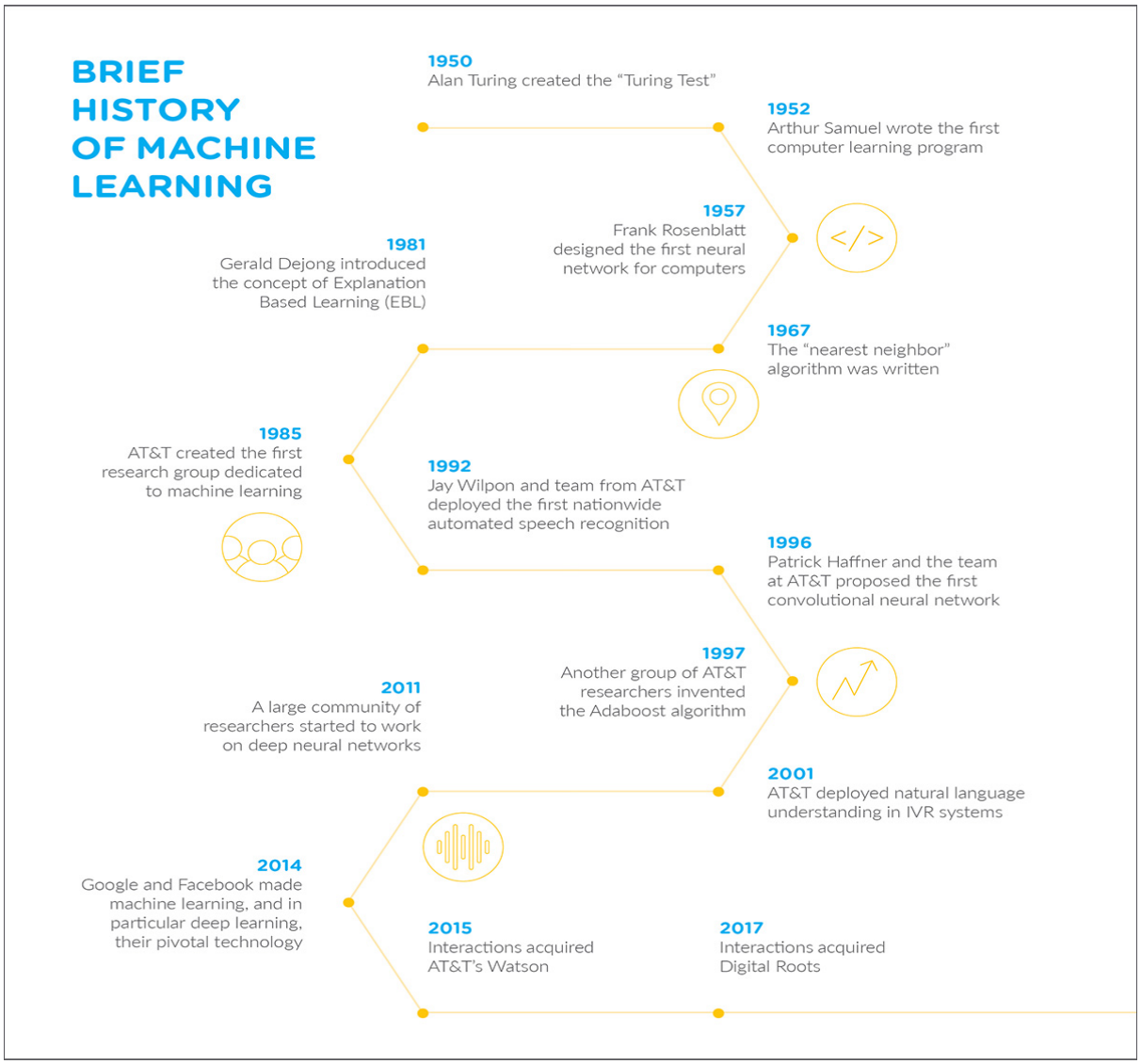

Figure - 2.2 History of Machine Learning

Arthur Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959 while at IBM. A representative book of the machine learning research during 1960s was the Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification. The interest of machine learning related to pattern recognition continued during 1970s, as described in the book of Duda and Hart in 1973. As a scientific endeavor, machine learning grew out of the quest for artificial intelligence. The

Evolution of Machine Learning is illustrated in (Figure- 2.2). Already in the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks"; these were mostly perceptrons and other models that were later found to be reinventions of the generalized linear models of statistics. Probabilistic reasoning was also employed, especially in automated medical diagnosis.

However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI and machine learning. Probabilistic systems were plagued by theoretical and practical problems of data acquisition and representation. By 1980, expert systems had come to dominate AI, and statistics was out of favor. Work on symbolic/knowledge-based learning did continue within AI, leading to inductive logic programming, but the more statistical line of research was now outside the field of AI proper, in pattern recognition and information retrieval. Neural networks research had been abandoned by AI and computer science around the same time. This line, too, was continued outside the AI/CS field, as "connectionism", by researchers from other disciplines including Hopfield, Rumelhart and Hinton. Their main success came in the mid-1980s with the reinvention of backpropagation.

Machine learning, reorganized as a separate field, started to flourish in the 1990s. The field changed its goal from achieving artificial intelligence to tackling solvable problems of a practical nature. It shifted focus away from the symbolic approaches it had inherited from AI, and toward methods and models borrowed from statistics and probability theory. It also benefited from the increasing availability of digitized information, and the ability to distribute it via the Internet.

2.3 Relationships to Other Fields

Machine learning is related to many other fields in its defining principles as well as delivery techniques. its most important relationships with other fields of study are.

2.3.1 Artificial Intelligence

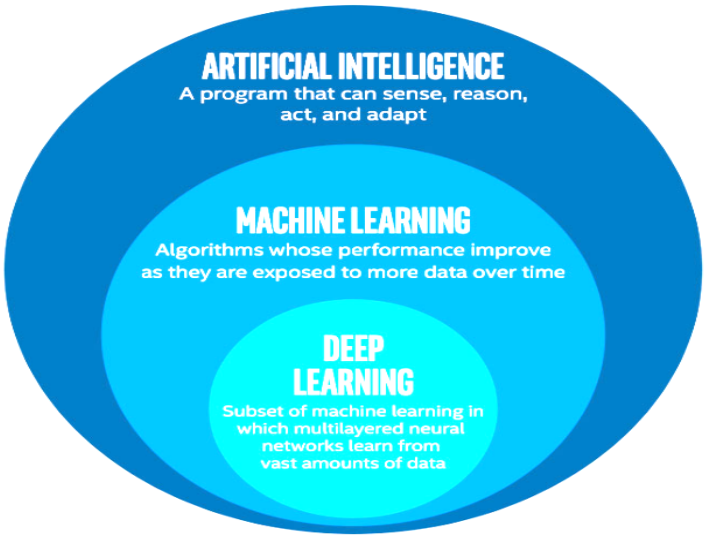

Figure - 2.3 Artificial Intelligence

Machine learning emerged as a separate field of study during scientists' pursuit of developing artificial intelligence (AI) in machines as shown in (Figure-2.3). They aimed to make machines learn from available data, using methods like neural networks, linear statistical models and probabilistic reasoning. However, by 1980s, expert systems were seen to be the right approach to achieve AI and all other approaches were dropped.

By 1990s, machine learning was restructured as a separate field and focused on solving practical problems rather than acquiring AI.

2.3.2 Computational Statistics

New principles of machine learning borrowed heavily from computational statistics. Two models – data model and algorithmic model – used in machine learning have their roots in computational statistics as illustrated in (Figure-2.4). Michael I Jordan, an American scientist and researcher in both machine learning and AI, has gone as far as to suggest that the overall field of machine learning and statistics be called data science.

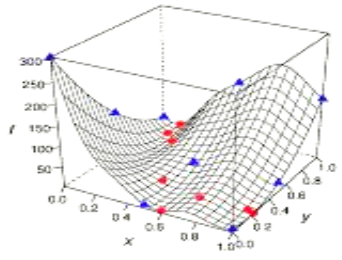

Figure - 2.4 Computational statistics

2.3.3 Probability

Machine learning utilizes probability theory for studying and exploring algorithms for pattern recognition, the building block of machine learning. These algorithms are then used to build models from available data to predict machine behavior in case of new inputs. Machine learning used loads of digitized information readily available through the Internet.

2.3.4 Data Mining

Machine learning and data mining both utilize similar methods to achieve different goals. Machine learning focuses on making predictions based on available data whereas data mining focuses on discovering unknown aspects of available data. Machine learning uses many techniques employed by data mining.

2.3.5 Mathematical Optimization

In simple terms, optimization can be defined as finding the best solution out of all solutions for any problem. In machine learning, predictions are made on the basis of available data. Mathematical optimization techniques can be used to deliver on the right theory to be used. Thus optimization and machine learning are closely related to each other as shown in (Figure -2.5).

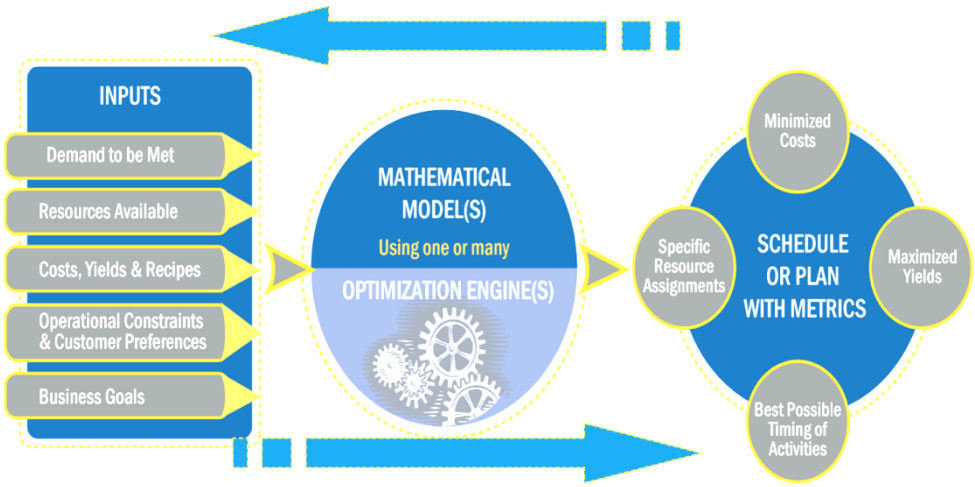

Figure - 2.5 Machine Learning Optimization

2.4 Different Machine Learning Approaches

The term machine learning, is used in a very general way and it refers to general techniques to extrapolate patterns from large sets or to the ability to make predictions on new data based on what is learnt by analyzing available known data. This is a very general and broad definition and it encompasses many different techniques. Machine learning techniques can be roughly divided into two large classes: Supervised and Unsupervised learning, though one more class is often added, and is referred to as Reinforcement Learning.

2.4.1 Supervised Learning

Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically supervised learning is a learning in which machine is taught or trained using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

2.4.2 Unsupervised Learning

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

2.4.3 Reinforcement Learning

Reinforcement learning is an area of Machine Learning. Reinforcement. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of training dataset, it is bound to learn from its experience.

2.5 Approaches Used in Our Project

2.5.1 Web Scrapping

Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

Implementing Web Scraping in Python with BeautifulSoup

There are mainly two ways to extract data from a website:

1. Use the API of the website (if it exists). For example, Facebook has the Facebook Graph API which allows retrieval of data posted on Facebook.
2. Access the HTML of the web page and extract useful information/data from it.

Steps involved in web scraping:

i. Send a HTTP request to the URL of the web page you want to access. The server responds to the request by returning the HTML content of the web page. For this task, third-party HTTP library for python requests will be used**.**

ii. After having access to HTML content, one task is left that is parsing the data. Since most of the HTML data is nested, it cannot be extracted simply through string processing. One needs a parser which can create a nested/tree structure of the HTML data**.**

    There are many HTML parser libraries available but the most advanced one is html5lib.

iii. Now, things needed to do is navigating and searching the parse tree that we created, i.e. tree traversal. For this task, another third-party python library will be used, Beautiful Soup. It is a Python library for pulling data out of HTML and XML files.

2.5.2 Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the

number of independent variables being used.

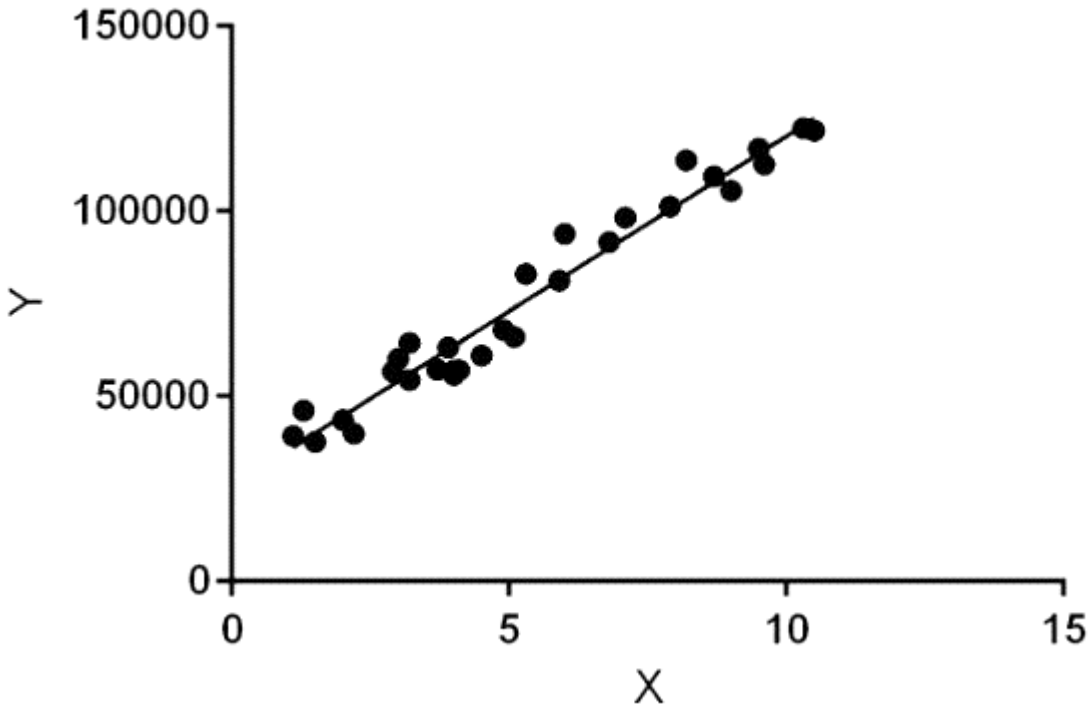

Figure - 2.6 Linear Regression

Linear regression as illustrated in (Figure-2.6) performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression as shown in (Figure-2.7):

$$y = \theta_1 + \theta_2 . x$$

Equation - 2.1 Linear Regression equation

While training the model it is given :

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.

θ_1: intercept

θ_2: coefficient of x

Once the best θ_1 and θ_2 values are calculated, and the best fit line is obtained. Then the model is ready for prediction, it will predict the value of y for the input value of x.

How to update θ_1 and θ_2 values to get the best fit line?

Cost Function (J):

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y). The Linear Regression Minimize Function is illustrated in (Figure 2.8).

$$minimize\ \frac{1}{n}\sum_{i=1}^{n}(pred_i\ -\ y_i)^2$$

Equation - 2.2 Linear Regression Minimize Function

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i\text{-}y_i)^2$$

Equation - 2.3 Linear Regression Cost Function

Cost function(J) as show in (Figure-2.9) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (prediction) and true y value (y).

Gradient Descent:

To update θ_1 and θ_2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ_1 and θ_2 values and then iteratively updating the values, reaching minimum cost.

2.5.3 Recurrent Neural Networks

A Recurrent Neural Network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

The term "recurrent neural network" as shown in (Figure-2.10) is used indiscriminately to refer to two broad classes of networks with a similar general structure, where one is finite impulse and the other is infinite impulse. Both classes of networks exhibit temporal dynamic behavior. A finite impulse recurrent network is a directed acyclic graph that can be unrolled and replaced with a strictly feed-forward neural network, while an infinite impulse recurrent network is a directed cyclic graph that cannot be unrolled.

Both finite impulse and infinite impulse recurrent networks can have additional stored state, and the storage can be under direct control by the neural network. The storage can also be replaced by another network or graph, if that incorporates time delays or has feedback loops. Such controlled states are referred to as gated state or gated memory, and are part of long short-term memory networks (LSTMs) and gated recurrent units.

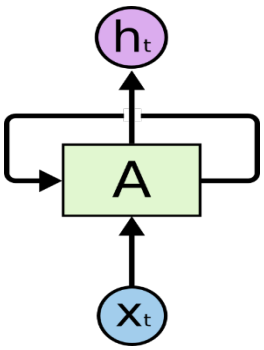

Figure - 2.10 Recurrent Neural Networks

2.5.3.1 How RNN Learn?

Artificial neural networks are created with interconnected data processing components that are loosely designed to function like the human brain. They are composed of layers of artificial neurons (network nodes) that have the capability to process input and forward output to other nodes in the network. The nodes are connected by edges or weights that influence a signal's strength and the network's ultimate output.

In some cases, artificial neural networks process information in a single direction from input to output. These "feed-forward" neural networks include convolutional neural networks that underpin image recognition systems. RNNs, on the other hand, can be layered to process information in two directions.

Like feed-forward neural networks, RNNs can process data from initial input to final output. Unlike feed-forward neural networks, RNNs use feedback loops such as Back-propagation Through Time or BPTT throughout the computational process to loop information back into the network. This connects inputs together and is what enables RNNs to process sequential and temporal data.

2.5.3.2 Vanishing and Exploding Gradients

The vanishing gradient problem is a difficulty found in training artificial neural networks with gradient-based learning methods and backpropagation. In such methods, each of the neural network's weights receives an update proportional to the partial derivative of the error function with respect to the current weight in each iteration of training. The problem is that in some cases, the gradient will be vanishingly small, effectively preventing the weight from changing its value. In the worst case, this may completely stop the neural network from further training. As one example of the problem cause, traditional activation functions such as the hyperbolic tangent function have gradients in the range (0, 1), and back-propagation computes gradients by the chain rule. This has the effect of multiplying n of these small numbers to compute gradients of the "front" layers in an n-layer network, meaning that the gradient (error signal) decreases exponentially with n while the front layers train very slowly.

2.5.3.3 Long Short-Term Memory(LSTM)

Long short-term memory (LSTM) as illustrated in (Figure-2.11) is a deep learning system that avoids the vanishing gradient problem. LSTM is normally augmented by recurrent gates called "forget" gates. LSTM prevents back-propagated errors from vanishing or exploding. Instead, errors can flow backwards through unlimited numbers of virtual layers unfolded in space. That is, LSTM can learn tasks that require memories of events that happened thousands or even millions of discrete time steps earlier. Problem-specific LSTM-like topologies can be evolved. LSTM works even given long delays between significant events and can handle signals that mix low and high frequency components. Many applications use stacks of LSTM RNNs and train them by Connectionist Temporal Classification (CTC) to find an RNN weight matrix that maximizes the probability of the label sequences in a training set, given the corresponding input sequences. CTC achieves both alignment and recognition. LSTM can learn to recognize context-sensitive languages unlike previous models based on hidden Markov models (HMM) and similar concepts.

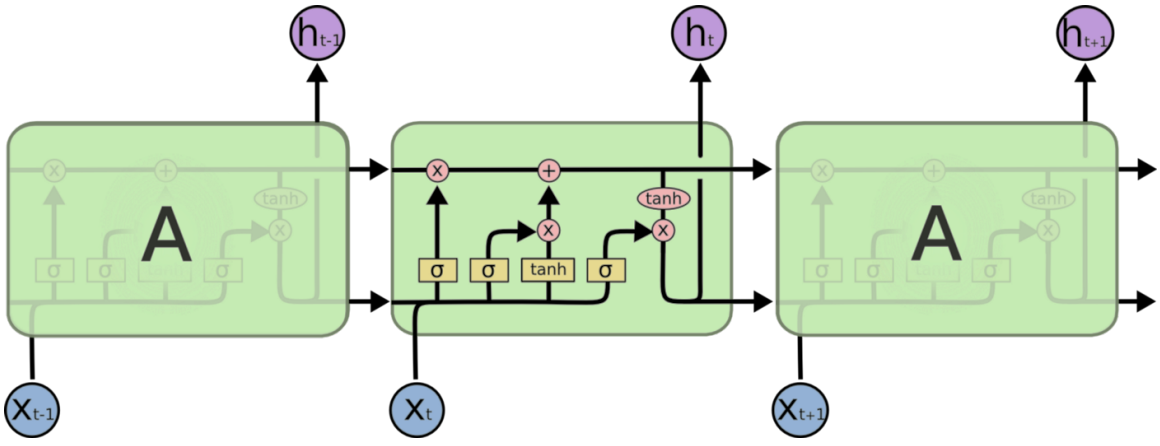

Figure - 2.11 Long Short-Term Memory Node

2.6 Sentimental Analysis

Sentiment analysis is the automated process that uses AI to identify positive, negative and neutral opinions from text. Sentiment analysis is widely used for getting insights from social media comments, survey responses, and product reviews, and making data-

driven decisions. In a world where users generate 2.5 quintillion bytes of data every day, sentiment analysis has become a key tool for making sense of that data.

Usually, besides identifying the opinion, these systems extract attributes of the expression:

i. Polarity: if the speaker expresses a positive or negative opinion,
ii. Subject: the thing that is being talked about,
iii. Opinion holder: the person, or entity that expresses the opinion.

Currently, sentiment analysis is a topic of great interest and development since it has many practical applications. Companies use sentiment analysis to automatically analyze survey responses, product reviews, social media comments, and the like to get valuable insights about their brands, product, and services. Sentiment Analysis Scope

Sentiment analysis can be applied at different levels of scope:

i. **Document level** sentiment analysis obtains the sentiment of a complete document or paragraph.
ii. **Sentence level** sentiment analysis obtains the sentiment of a single sentence.
iii. **Sub-sentence level** sentiment analysis obtains the sentiment of sub-expressions within a sentence.

2.6.1 Types of Sentiment Analysis

There are many types and flavours of sentiment analysis and SA tools range from systems that focus on polarity (positive, negative, neutral) to systems that detect feelings and emotions (angry, happy, sad, etc.) or identify intentions (interested or not interested). In the following section, the most important ones will be covered.

2.6.1.1 Fine-grained Sentiment Analysis

Sometimes you may be also interested in being more precise about the level of polarity of the opinion, so instead of just talking about positive, neutral, or negative opinions you could consider the following categories:

i. Very positive
ii. Positive

iv. Negative
v. Very negative

This is usually referred to as fine-grained sentiment analysis. This could be, for example, mapped onto a 5-star rating in a review, e.g.: Very Positive = 5 stars and Very Negative = 1 star. Some systems also provide different flavours of polarity by identifying if the positive or negative sentiment is associated with a particular feeling, such as, anger, sadness, or worries (i.e. negative feelings) or happiness, love, or enthusiasm (i.e. positive feelings).

2.6.1.2 Emotion Detection

Emotion detection aims at detecting emotions like, happiness, frustration, anger, sadness, and the like. Many emotion detection systems resort to lexicons (i.e. lists of words and the emotions they convey) or complex machine learning algorithms.
One of the downsides of resorting to lexicons is that the way people express their emotions varies a lot and so do the lexical items they use. Some words that would typically express anger might also express happiness. Aspect-based Sentiment Analysis Usually, when analyzing the sentiment in subjects, for example products, you might be interested in not only whether people are talking with a positive, neutral, or negative polarity about the product, but also which particular aspects or features of the product people talk about.

2.6.1.3 Intent Analysis

Intent analysis basically detects what people want to do with a text rather than what people say with that text.

2.6.1.4 Multilingual Sentiment Analysis

Multilingual sentiment analysis can be a difficult task. Usually, a lot of pre-processing

resources are available online (e.g. sentiment lexicons), but many others have to be created (e.g. translated corpora or noise detection algorithms). The use of the resources available requires a lot of coding experience and can take long to implement.

An alternative to that would be detecting language in texts automatically, then train a custom model for the language of your choice, and finally, perform the analysis. It's estimated that 80% of the world's data is unstructured and not organized in a pre-defined manner. Most of this comes from text data, like emails, support tickets, chats, social media, surveys, articles, and documents. These texts are usually difficult, time-consuming and expensive to analyze, understand, and sort through.

Sentiment analysis systems allows companies to make sense of this sea of unstructured text by automating business processes, getting actionable insights, and saving hours of manual data processing, in other words, by making teams more efficient.

Some of the advantages of sentiment analysis include the following:

i. **Scalability**:

   There's just too much data to process manually. Sentiment analysis allows to process data such as sorting through thousands of tweets, customer support conversations, or customer reviews at scale in an efficient and cost-effective way.

ii. **Real-time analysis**:

   sentiment analysis is used to identify critical information that allows situational awareness during specific scenarios in real-time. A sentiment analysis system can help you immediately identify these kinds of situations and take action.

iii. **Consistent criteria**:

   Humans don't observe clear criteria for evaluating the sentiment of a piece of text. It's estimated that different people only agree 60-65% of the times when judging the sentiment for a particular piece of text. It's a subjective task which is heavily influenced by personal experiences, thoughts, and beliefs. By using a centralized sentiment analysis system, companies can apply the same criteria to all of their data. This helps to reduce errors and improve data consistency

2.6.2 Sentiment Analysis Algorithms

There are many methods and algorithms to implement sentiment analysis systems, which can be classified as:

i. **Rule-based** systems that perform sentiment analysis based on a set of manually crafted rules.
ii. **Automatic** systems that rely on machine learning techniques to learn from data.
iii. **Hybrid** systems that combine both rule based and automatic approaches.

Precision, recall, and accuracy are standard metrics used to evaluate the performance of a classifier. Precision measures how many texts were predicted correctly as belonging to a given category out of all of the texts that were predicted (correctly and incorrectly) as belonging to the category.

Recall measures how many texts were predicted correctly as belonging to a given category out of all the texts that should have been predicted as belonging to the category. It is also known that the more data used to feed classifiers with, the better recall will be**.** Accuracy measures how many texts were predicted correctly (both as belonging to a category and not belonging to the category) out of all of the texts in the corpus. Most frequently, precision and recall are used to measure performance since accuracy alone does not say much about how good or bad a classifier is for a difficult task like analyzing sentiment, precision and recall levels are likely to be low at first. As you feed the classifier with more data, performance will improve. However, as is shown below, since annotated data is not likely to be accurate, the chances are that precision levels won't get too high. However, if you feed the classifier consistently tagged data, results are going to be as good as results can be for any other classification problem

2.6.2.1 Rule-Based Approaches

Usually, rule-based approaches define a set of rules in some kind of scripting language that identify subjectivity, polarity, or the subject of an opinion.

The rules may use a variety of inputs, such as the following:

i. Classic NLP techniques like stemming, tokenization, part of speech tagging and parsing.

ii. Other resources, such as lexicons (i.e. lists of words and expressions).

A basic example of a rule-based implementation would be the following:

i. Define two lists of polarized words (e.g. negative words such as bad, worst, ugly, etc and positive words such as good, best, beautiful, etc).
ii. Given a text:
    i. Count the number of positive words that appear in the text.
    ii. Count the number of negative words that appear in the text.
iii. If the number of positive word appearances is greater than the number of negative word appearances return a positive sentiment, conversely, return a negative sentiment. Otherwise, return neutral.

This system is very naïve since it doesn't take into account how words are combined in a sequence. A more advanced processing can be made, but these systems get very complex quickly. They can be very hard to maintain as new rules may be needed to add support for new expressions and vocabulary. Besides, adding new rules may have undesired outcomes as a result of the interaction with previous rules. As a result, these systems require important investments in manually tuning and maintaining the rules.

2.6.2.2 Automatic Approaches

Automatic methods, contrary to rule-based systems, don't rely on manually crafted rules, but on machine learning techniques. The sentiment analysis task is usually modeled as a classification problem where a classifier is fed with a text and returns the corresponding category, e.g. positive, negative, or neutral (in case polarity analysis is being performed).

Said machine learning classifier can usually be implemented with the following steps and components:

i. **The Training and Prediction Processes**

In the training process, our model learns to associate a particular input (i.e. a text) to the corresponding output (tag) based on the test samples used for training. The feature extractor transfers the text input into a feature vector. Pairs of feature vectors and tags are fed into the machine learning algorithm to generate a model.

In the prediction process, the feature extractor is used to transform unseen text inputs into feature vectors. These feature vectors are then fed into the model, which generates predicted tags.

ii. Feature Extraction from Text

The first step in a machine learning text classifier is to transform the text into a numerical representation, usually a vector. Usually, each component of the vector represents the frequency of a word or expression in a predefined dictionary (e.g. a lexicon of polarized words). This process is known as feature extraction or text vectorization and the classical approach has been bag-of-words or bag-of-n grams with their frequency.

More recently, new feature extraction techniques have been applied based on word embedding (also known as word vectors). This kind of representations makes it possible for words with similar meaning to have a similar representation, which can improve the performance of classifiers.

iii. Classification Algorithms

The classification step usually involves a statistical model like Naïve Bayes, Logistic Regression, Support Vector Machines, or Neural Networks:

i. Naïve Bayes: a family of probabilistic algorithms that uses Bayes's Theorem to predict the category of a text.
ii. Linear Regression: a very well-known algorithm in statistics used to predict some value (Y) given a set of features (X).
iii. Support Vector Machines: a non-probabilistic model which uses a representation of text examples as points in a multidimensional space. These examples are mapped so that the examples of the different categories (sentiments) belong to distinct regions of that space. Then, new texts are mapped onto that same space and predicted to belong to a category based on which region they fall into.
iv. Deep Learning: a diverse set of algorithms that attempts to imitate how the human brain works by employing artificial neural networks to process data.

iv. Sentiment Analysis Metrics and Evaluation

There are many ways in which you can obtain performance metrics for evaluating a classifier and to understand how accurate a sentiment analysis model is. One of the most frequently used is known as cross-validation.

What cross-validation does is splitting the training data into a certain number of training folds (with 75% of the training data) and at the same number of testing folds (with 25% of the training data), use the training folds to train the classifier, and test it against the testing folds to obtain performance metrics (see below). The process is repeated multiple times and an average for each of the metrics is calculated.

If your testing set is always the same, you might be overfitting to that testing set, which means you might be adjusting your analysis to a given set of data so much that you might fail to analyze a different set. Cross-validation helps prevent that. The more data you have, the more folds you will be able to use.

2.6.2.3 Hybrid Approaches

The concept of hybrid methods is very intuitive: just combine the best of both worlds, the rule-based and the automatic ones. Usually, by combining both approaches, the methods can improve accuracy and precision.

2.6.3 Sentiment Analysis Challenges

Most of the work in sentiment analysis in recent years has been around developing more accurate sentiment classifiers by dealing with some of the main challenges and limitations in the field.

2.6.3.1 Subjectivity and Tone

The detection of subjective and objective texts is just as important as analyzing their tone. In fact, so called objective texts do not contain explicit sentiments.

2.6.3.2 Context and Polarity

All utterances are uttered at some point in time, in some place, by and to some people, you get the point. All utterances are uttered in context. Analyzing sentiment without context gets pretty difficult. However, machines cannot learn about contexts if they are not mentioned explicitly. One of the problems that arise from context is changes in polarity.

2.6.3.3 Irony and Sarcasm

Differences between literal and intended meaning (i.e. irony) and the more insulting version of irony (i.e. sarcasm) usually change positive sentiment into negative whereas negative or neutral sentiment might be changed to positive. However, detecting irony or sarcasm takes a good deal of analysis of the context in which the texts are produced and, therefore, are really difficult to detect automatically.

2.6.3.4 Defining Neutral

Defining what it meant to neutral is another challenge to tackle in order to perform accurate sentiment analysis. As in all classification problems, defining your categories -and, in this case, the neutral tag- is one of the most important parts of the problem. What you mean by neutral, positive, or negative does matter when you train sentiment analysis models. Since tagging data requires that tagging criteria be consistent, a good definition of the problem is a must.

2.6.4 Use Cases and Applications of Sentiment Analysis

i. Social Media Monitoring

In today's day and age, brands of all shapes and sizes have meaningful interactions with customers, leads, and even competition on social networks like Facebook, Twitter, and Instagram. Most marketing departments are already tuned into to online mentions as far

analysis on social media, anyone can get incredible insights into the *quality* of conversation that's happening around a brand.

ii. Brand Monitoring

Not only do brands have a wealth of information available on social media, but they also can look more broadly across the internet to see how people are talking about them online. Instead of focusing on specific social media platforms such as Facebook and Twitter, there can be target mentions in places like news, blogs, and forums, again, looking **a**t not just the volume of mentions, but also the quality of those mentions.

iii. Customer Feedback

Social media and brand monitoring provides immediate, unfiltered, invaluable information on customer sentiment. In a parallel vein run two other troves of insight – surveys and customer support interactions. Teams often look at their Net Promoter Score (NPS), but this analyses can also apply to any type of survey or communication channel that yields textual customer feedback.NPS surveys ask a few simple questions and use them to identify customers as promoters, passives, or detractors. The goal is to identify overall customer experience, and find ways to elevate all customers to "promoter" level, where they theoretically will buy more, stay longer, and refer other customers. Numerical survey data is easily aggregated and assessed, but **we** want that same ease with the "why" answers as well. A regular NPS score simply gives you a number, without the additional context of what it is about and why the score landed there. Sentiment analysis takes it that step further.

iv. Business Intelligence Build-up

Having insights-rich information eliminates the guesswork and execution of timely decisions. With the sentiment data about your established and the new products, it's easier to estimate your customer retention rate. Based on the reviews generated through sentiment analysis in business, you can always adjust to the present market situation

with automated insights. Business intelligence is all about staying dynamic throughout. Having the sentiments data gives you that liberty. If you develop a big idea, you can test it before bringing life to it. This is known as concept testing. Whether it is a new product, campaign or a new logo, just put it to concept testing and analyze the sentiments attached to it.

v. Market Research

Sentiment analysis empowers all kinds of market research and competitive analysis. Whether you're exploring a new market, anticipating future trends, or keeping an edge on the competition, sentiment analysis can make all the difference.

2.6.5 How sentiment analysis increase accuracy

While regression algorithms like linear regression and recurrent neural networks only try to establish a relationship between the feature values and the target value which might be a polynomial function, sentiment analysis is based out of opinion of general public and experts about the brand, therefore, adding the sentiment about the brand as a feature takes the curve of relationship between feature values and target values into the right direction. Sentiment analysis have been done on news articles published by various media firms on the internet. These include news about new product launches or strategic moves of the brand and since these are written by journalists or experts of the specific domain, they include their own opinion and expectations about of the brand or that particular move, hence it is a better indication of the future movements of stock prices. Sentiment Analysis Results of various social media platforms for brands and their products have been known to have correlation with the stock prices of the brands. Hence adding them to the feature set adds some real-life components to the stock prediction model apart from all other mathematical features. Better the correlation between the sentiment analysis results of the news articles and the movements of stock prices, better will be the predictions of future stock prices. Since the bag-of-words approach for sentiment analysis of news articles have been used, much high correlation could not be achieved, but it could be improved if a sentiment analysis model could be trained specifically for news articles. Unfortunately, due to lack of labelled news

dataset, that is, news articles labelled as positive, negative or neutral, it was not possible for us to build a sentiment analysis model specifically for news articles.

CHAPTER 3: APPROACH TOWARDS THE PROBLEM

3.1 Challenges

i. Dynamic stock data is chosen to go with rather than static, this makes data collection itself a challenge.

ii. To get dynamic data web scrapping is used as a tool but the major issue is that, web scrapping stock exchange website is not appropriate as the dates between which the stock data is needed are not edited in the URL of the stock exchange website.

iii. Deciding which model to use in the prediction.

iv. Parameter tweaking for RNN is major challenge to increase the accuracy of model.

v. Sentimental analysis is one of the most challenging task to do, first challenge is to get sentiment for each date present in the stock data-set which have been web scrapped. Secondly both twitter and Reddit APIs are not much efficient and capable of doing this so the only one option is to web scrap the sentiments from different news channel and newspaper articles. So for web scrapping and getting sentiments for each date, Google news has been used**.**

vi. Web scrapping Google news itself is a challenge because having zero knowledge about the links that are going to be web scrapped, generates the involvement of a good pattern matching and generalize way so that correct data can be taken out from different websites.

vii. Integrating sentimental analysis with the models selected and to showcase that machine learning models along with sentimental analysis will produce better prediction all together is a major challenge.

3.2 Project Road-Map

3.2.1 Data Collection and Web Scrapping

The first phase of the project is the data collection phase. Collecting data allows you to capture a record of past events so that data analysis can be used to find recurring patterns. From those patterns, you build predictive models using machine

are only as good as the data from which they are built, so good data collection practices are crucial to developing high-performing models. The data need to be error-free (garbage in, garbage out) and contain relevant information for the task at hand.
Instead of going with the traditional method of downloading data the opted method for dynamic data collection through web scrapping to make our project one step closer to a realistic stock market predictor.
Dynamic data collection simply means that mentioning a starting date and an ending date, and stock data between these two dates is web scrapped. So for this purpose, the financial yahoo website has been used to web scrap data.
For web scraping stock data Pandas library's read_html () function has been used which return list of dataframes from which desired dataframe is retrieved.

3.2.2 Data Visualization

The second phase is data visualization. Data visualization is a quick, easy way to convey concepts in a universal manner and it can be experimented with different scenarios by making slight adjustments. Data visualization also helps identify areas that need attention, e.g. outliers, which can later impact our machine learning model.

Time Series graph: - A time series plot is a graph where some measure of time is the unit on the x-axis. In fact, the x-axis is labelled as the time-axis. The y-axis is for the variable that is being measured. Data points are plotted and generally connected with straight lines, which allows for the analysis of the graph generated. So in this phase, days are plotted on the x-axis as a unit of time where 0th day represents the starting date and Nth day represents the ending date, and on the y-axis**,** open, high, low and close prices are plotted. (Figure-3.1) and (Figure-3.2) shows the Time Series Plot of Open, High, Low, Close Prices and Time Series Plot Between Open and High Prices respectively.

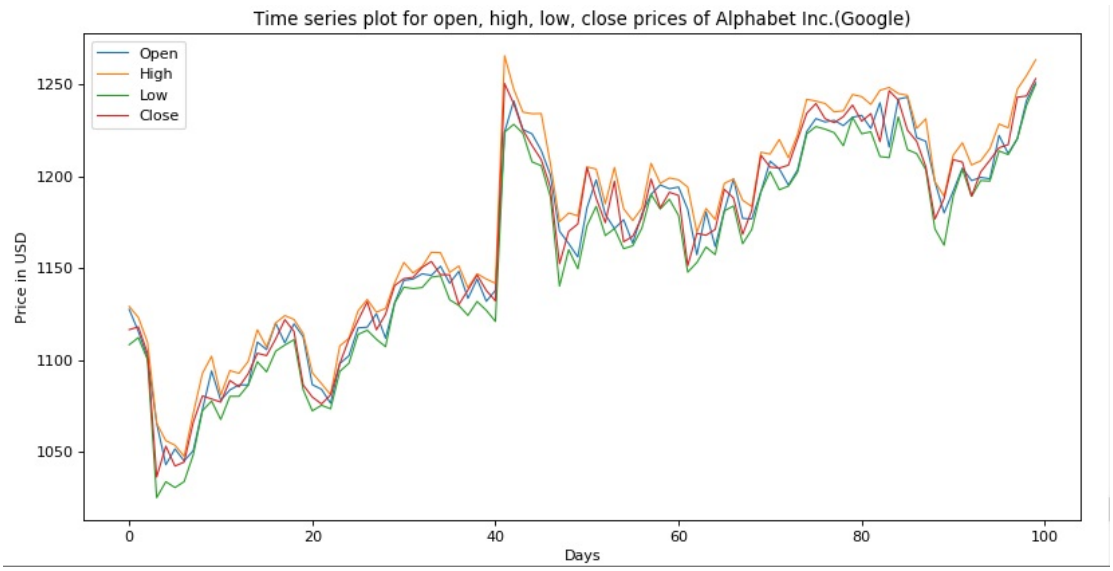

Figure - 3.1 Time Series Plot of Open, High, Low and Close Prices

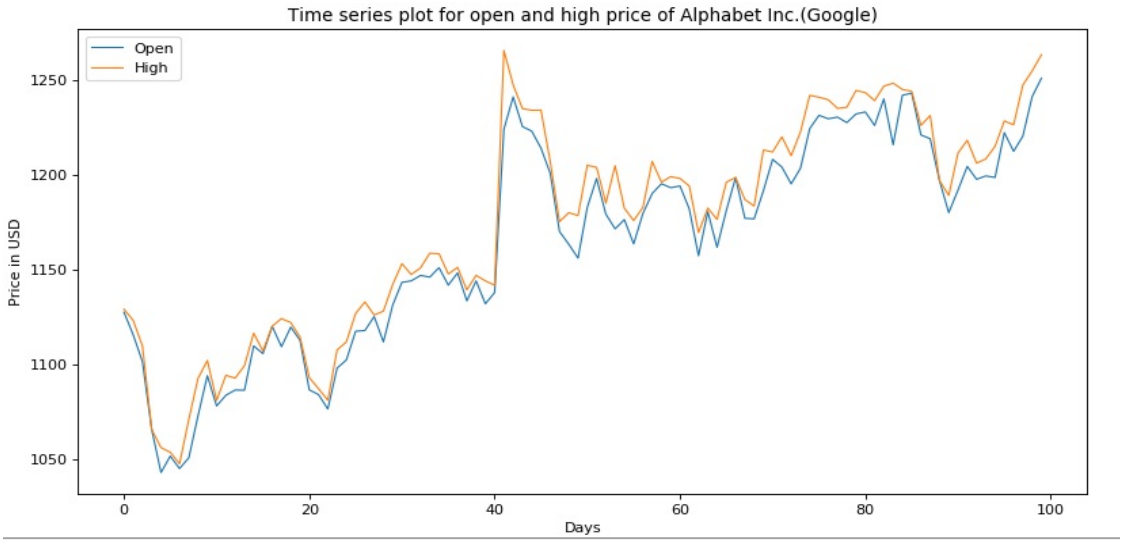

Figure - 3.2 Time Series Plot Between Open and High Prices

Scatter plot: - A scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis. Scatter plots are sometimes called correlation plots because they show how two variables are correlated. Scatter plot have been plotted between the open price and high price and also calculated the correlation between them. It is found that open and high prices are highly correlated with a positive correlation of 0.97 out of 1. The Scatter Plot Between Open and High Prices is illustrated in (Figure-3.3).

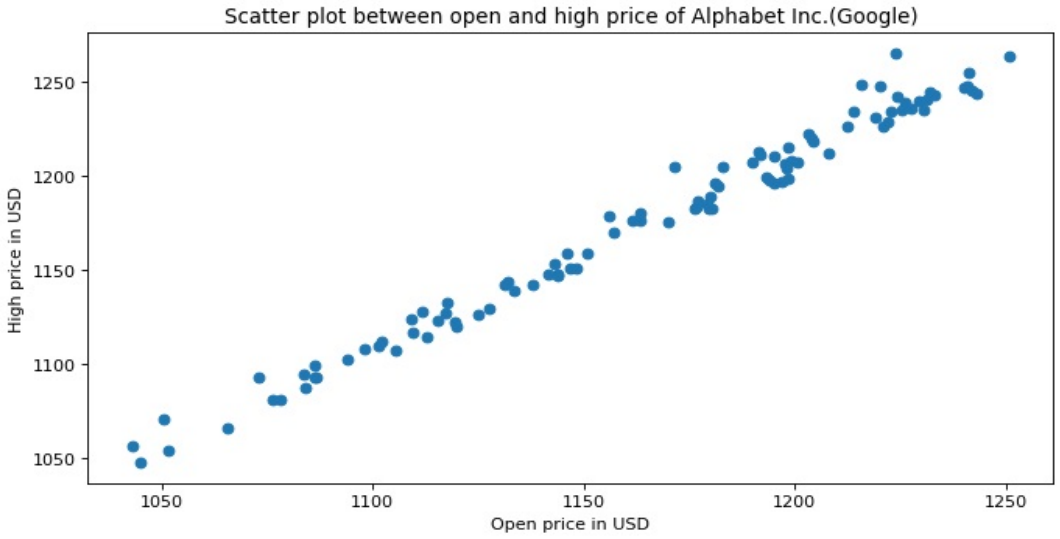

Figure - 3.3 Scatter Plot Between Open and High Prices

Box and whiskers plot: - A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum. Box and whiskers plot have been for the open**,** high, low, and close prices to get clear picture of the statistical aspect of the data. The Box and Whiskers Plot is shown in(Figure-3.4). The Box Plots of Open, High, Low and Close are illustrated in (Figure-3.5), (Figure-3.6), (Figure-3.7), (Figure-3.8) respectively.

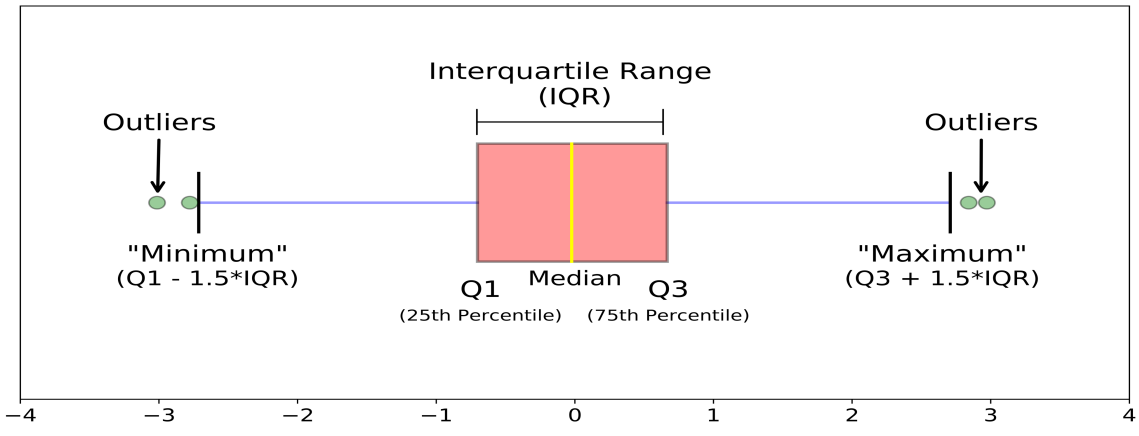

Figure - 3.4 Box and Whisker Plot

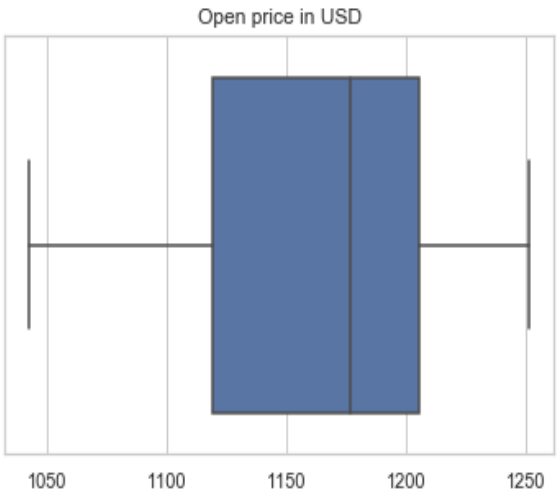

Figure - 3.5 Open Price Box Plot

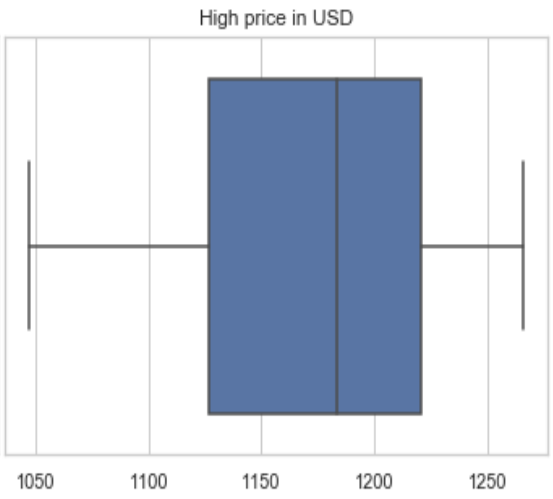

Figure - 3.6 High Price Box Plot

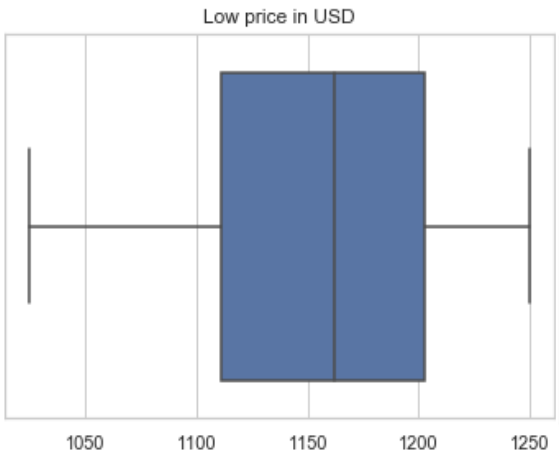

Figure - 3.7 Low Price Box Plot

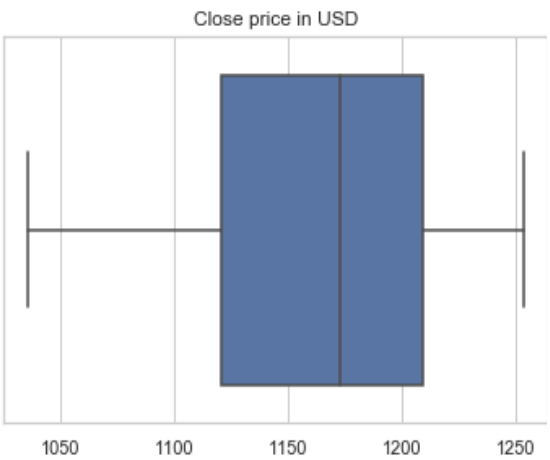

Figure - 3.8 Close Price Box Plot

Candlestick/OHLC graph: - Candlesticks are graphical representations of price movements for a given period. They are commonly formed by the opening, high, low, and closing prices of a financial instrument. If the opening price is above the closing price, then a filled (normally red or black) candlestick is drawn. If the closing price is above the opening price, then normally a green or a hollow candlestick (white with black outline) is shown. The filled or hollow portion of the candle is known as the body or real body and can be long, normal, or short depending on its proportion to the lines above or below it. The lines above and below, known as shadows, tails, or wicks represent the high and low price ranges within a specified period. However, not all candlesticks have shadows. Candlestick graph have been plotted for the open,

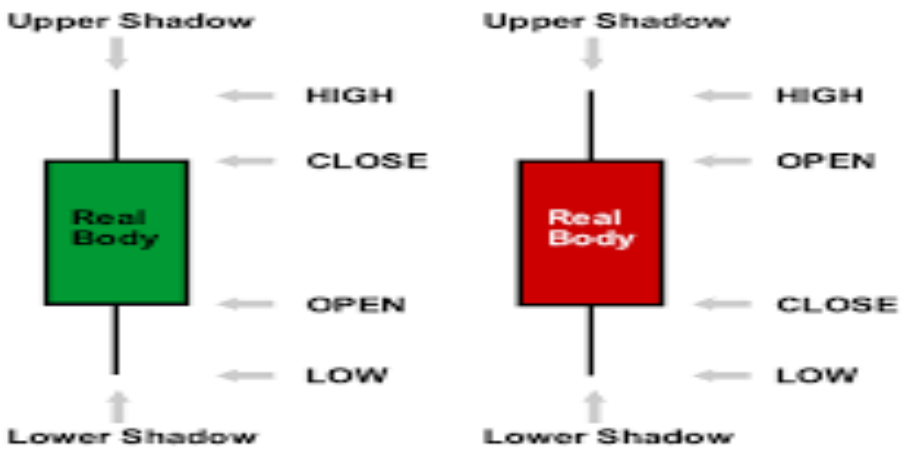

Figure - 3.9 Candlestick Chart

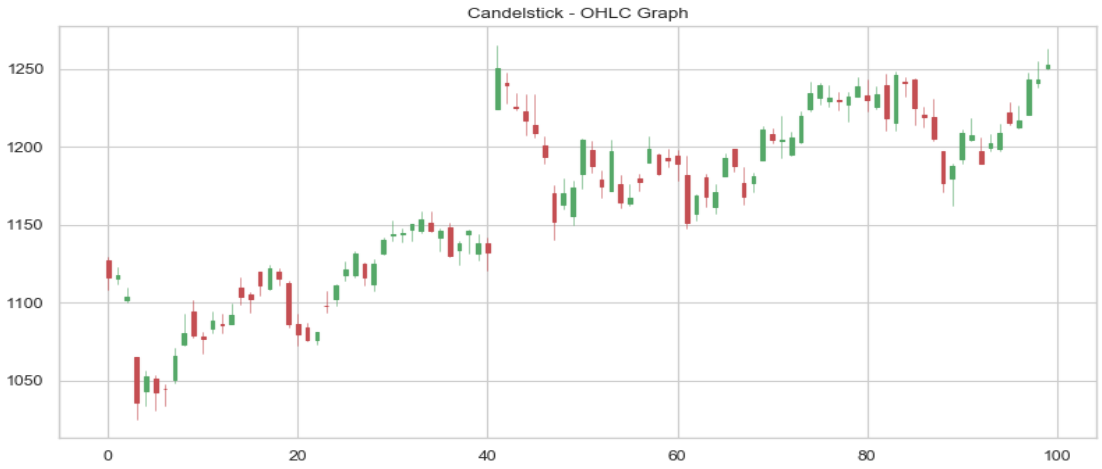

Figure - 3.10 Candlestick-OHLC Graph

After data visualization some amount of data preprocessing task is done in which unwanted or useless rows are removed from the data-set and make the data-set ready to apply any sort of machine learning model on it. (Figure-3.10) illustrates Candlestick-OHLC Graph

3.2.3 Applying Machine Learning Models

Third phase is applying Linear regression model on our pre-processed data-set.

Linear regression is a method used to model a relationship between a dependent variable (y), and an independent variable (x). With simple linear regression, there will

only be one independent variable x. There can be many independent variables which would fall under the category of multiple linear regression.

Multiple regression is used in our project where previous day open, high, low, close price have been used to predict the high price of the next day.

So the independent variable matrix (x) includes the previous day open, high, low and close price and the dependent variable (y) is the high price of the next day.

In Linear regression the stock data have been split in the ratio of 4:1 (80% training & 20% testing).

The next model used is the Recurrent Neural Network (RNN) in which Long Short Term Memory (LSTM) architecture is used.

In RNN, the previous 10 days stock data have been used to predict the high price for the 11^{th} day. Two hidden layers have been used**,** one input and one output or dense layer. The stock data have been split in the ratio of 77.78% training & 22.22% testing.

3.2.4 Sentimental Analysis

In the fourth phase of the project sentiments have been used as a tool to increase the accuracy of our prediction model to some extent. So for sentimental analysis there are several pre-existing options available like twitter sentimental analysis**,** Reddit sentimental analysis and many more. Sentimental analysis is one of the most challenging task to do, firstly sentiment needed to be fetched for each date present in the stock data-set which have been web scrapped**.** In Twitter sentimental analysis one can only access tweets within 7 consecutive days beyond which it is a paid service and also there is a high possibility that there is no tweet in majority of the dates present in the data-set, in Reddit sentimental analysis the top 1000 subeddits are fetched but here the date is not taken in consideration. So both twitter and Reddit APIs are not much efficient and capable of doing this so only option left is to web scrap the sentiments from different news channel and newspaper articles. So for web scrapping and getting sentiments for each date, Google news have been used.

Our sentimental analysis works in the following manner: -

i. Traversing the stock data data-set row wise and for each row the date is fetched.

ii. For this particular date Google news is web scrapped with the help of BeautifulSoup, so as to get links of different websites and newspaper articles.

iii. Now these links, associated with a particular date are web scrapped one by one and their sentiments is analyzed.

iv. Final Sentiment for a particular day is calculated by adding the sentiments of the links associated with a particular day and divide it by total number of links for that day.

For analysing sentiments, NLTK library was used. NLTK gives four output indicators for a given text string i.e. positive, negative, neutral and compound. So the compound value has been used as it depicts the overall sentiment for a given text string.

3.2.5 Integrating Sentimental Analysis with Machine Learning Models

Now the fifth and the last phase of the project is to integrate sentimental analysis with both of the models that are Linear regression and Recurrent neural network with LSTM**.** one extra column in the stock data-set with the name 'sentiments' which contains the sentiment for each day in the data-set has been added**.** Then the data-set is scaled (Standard scaling) and Linear regression model is applied. In linear regression our independent variable matrix (x) now includes the open, high, low and adjusted close prices, and sentiment for the previous day. Our dependent variable (y) is the high price for the next day.

Here also the methodology is simple open, high, low and adjusted close prices, and sentiment for the previous day were used to predict the high price for the next day.

Now the modified stock data-set which includes the sentiment value for each day is given as input to the RNN-LSTM model.

In RNN, the previous 7 days stock data have been used to predict the high price for the 8^{th} day. Two hidden layers**,** one input and one output or dense layer have been used.

The stock data have been split in ratio of 77.78% training & 22.22% testing.

3.2.6 Error calculation

The Sklearn mean_squared_error function has been used to calculate the error between the actual and predicted values. In statistics, the mean squared error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average

of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss.

The smaller the means squared error, the closer you are to finding the line of best fit. Depending on your data, it may be impossible to get a very small value for the mean squared error.

3.2.7 Limitations

i. In sentimental analysis, bag-of-words approach has been used due to which there is a lower level of co-relation between stock prices and sentiments.
ii. In web scrapping of Google news for sentimental analysis pattern matching and generalize algorithm to fetch articles from different websites have been used. But there are websites where this generalized algorithm fails and fetch wrong data for sentimental analysis like cookie policy, warnings etc.
iii. To avoid wrong data from passing on to the sentiment analyzer the length of the text string have been checked beforehand and if it is less than 200 that particular text string is not allowed to enter into sentiment analyzer and a neutral response is registered against that link. But if for a particular date having majority of links that can bypass the generalized algorithm then the total sentiment for that date would not be determined correctly

CHAPTER 4: RESULT & DISCUSSION

In this project two machine learning models are used: -

1. Linear Regression
2. Recurrent Neural Network with LSTM

Both these models are used in two configurations: -

1. Without incorporating sentimental analysis
2. Along with sentimental analysis

Both the models along with listed configurations have given nice results which can be witnessed from the following graphs: -

i. Linear Regression without sentimental analysis. The Result of Linear Regression is illustrated in(Figure-4.1).

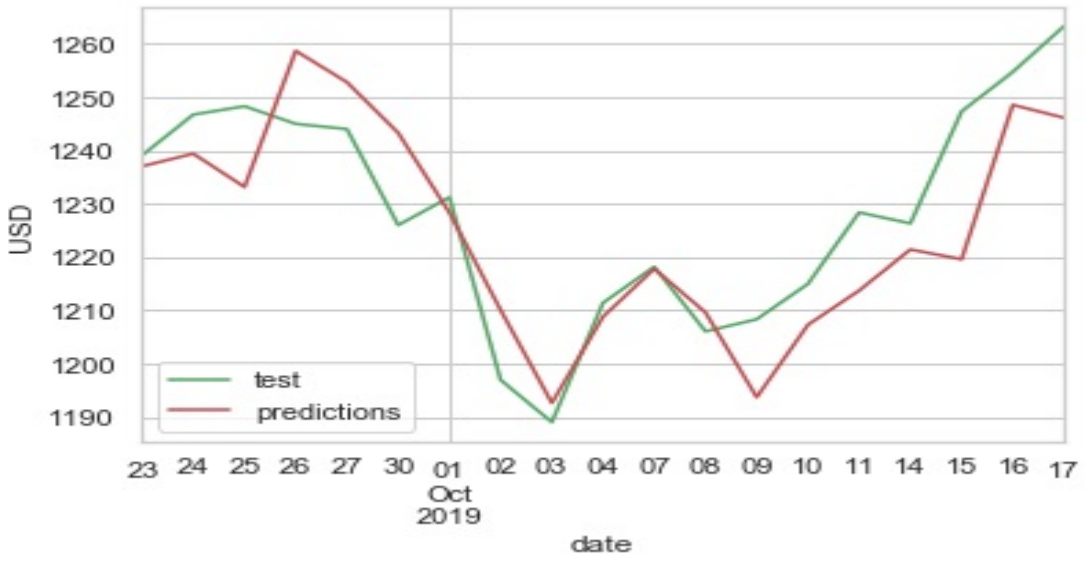

Figure - 4.1 Linear Regression Result

ii. Recurrent Neural Network without sentimental analysis. The Result of Recurrent Neural Network is illustrated in(Figure-4.2).

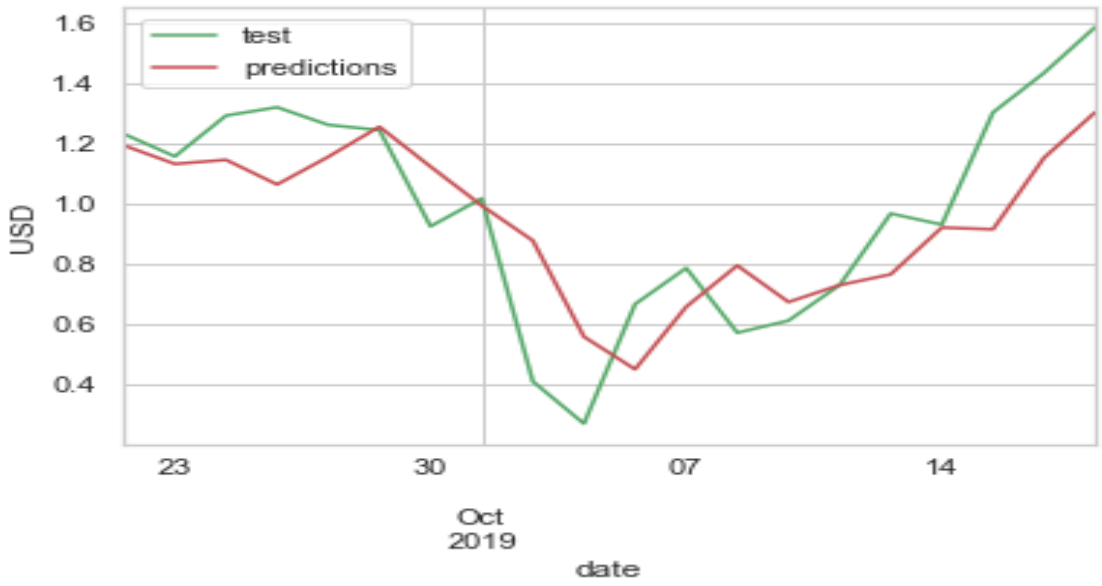

Figure - 4.2 Recurrent Neural Network Result

iii. Linear Regression with sentimental analysis The Result of Linear Regression with Sentiment Analysis is illustrated in(Figure-4.3).

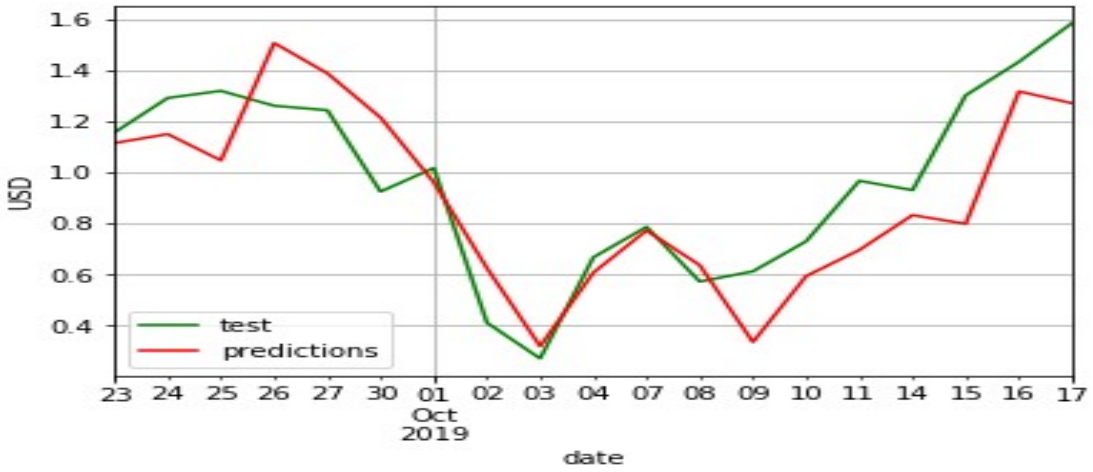

Figure - 4.3 Linear Regression with Sentimental analysis result

iv. Recurrent Neural Network with sentimental analysis. The Result of Recurrent Neural Network with Sentiment Analysis is illustrated in(Figure-4.4).

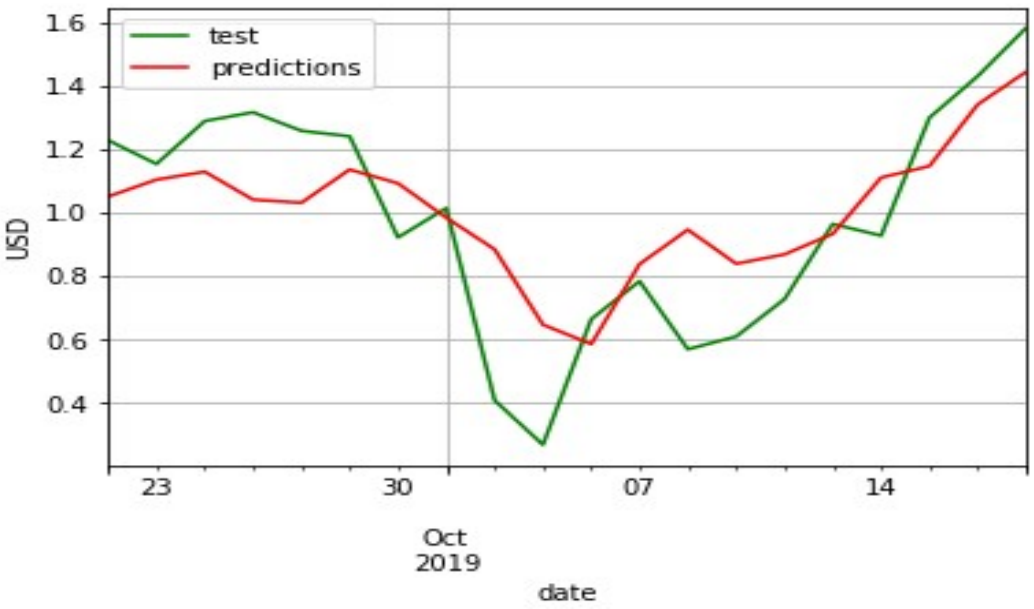

Figure - 4.4 RNN with Sentimental Analysis result

Graph in (Figure-4.4) depicts the actual high price of stock in green colour and our predicted high price in red colour.

4.1 Comparison between different models used in the project

Model	Mean Squared Error	
	Without Sentimental Analysis	**With Sentimental Analysis**
Linear Regression	0.04413	0.04554
Recurrent Neural Network (RNN)	0.04535	0.04534

Table - 4.1 Mean squared error for different ML models

CHAPTER 5: FUTURE SCOPE AND CONCLUSION

This project currently focuses on predicting stock prices of Alphabet Inc. and hence its scope may be expanded to include much more organizations in future and increase the usability.

The project is capable of predicting stock prices for only a few days in future for which the data of previous day is not available. The algorithm used may be modified to predict stock prices for much more days or even months in the future.

The project currently uses only the data of only previous 7-10 days to predict stock price values for a given day. Though stock prices might be dependent on broader history of the organization. Hence this might be taken into consideration in future enhancements

The current sentiment analysis model uses bag of words approach to predict the sentiment of news articles on organization. Thus a sentiment analysis model may be built in future specifically trained to predict sentiment news articles which would increase the accuracy of sentiment analysis.

Sentiment Analysis of tweets (Twitter posts) and subreddits (Reddit posts) were also tried but they did not turn out to be fruitful in the end as Twitter API has a weekly limit for fetching tweets for a tag in the unpaid version while Reddit API could retrieve only top thousand subreddits irrespective of their dates. These shortcomings might be overcome in the future and sentiment analysis for other valuable sources may also be incorporated.

The project may be deployed using a web application to provide a better user experience.

The idea of dynamically web scrapping stock data and corresponding news articles and applying regression using RNN and performing sentiment analysis of scrapped news articles has been successfully implemented**.** Though incorporating sentiment analysis into regression model introduce a little error in case of linear regression, it gave slightly better results in case of RNN. First linear regression was used to predict stock prices and then on using RNN errors were minimized**.**

REFERENCES

[1] Deepak Kumar Mahto and Lisha Singh, "A Dive into Web Scraper World.", in International Conference on Computing for Sustainable Global Development, INDIACom 2016, 16-18 March 2016, New Delhi, India, IEEE, 2016, pp 689-693.

[2] Ilja Stanišević and Đorđe Petrović, "Web Scrapping and Storing Data in a Database, A Case Study of the Used Cars Market.", in 25th Telecommunication Forum, TEFLOR 2017-Proceedings, 21-22 November 2017, Belgrade, Serbia, IEEE ,2016, pp 1-4.

[3] Yahya Eru Cakra and Bayu Distiawan Trisedya, "Stock Price Prediction Using Linear Regression Based on Sentiment Analysis.", in International Conference on Advanced Computer Science and Information Systems, ICACSIS 2015,10-11 October 2015, Depok, Indonesia, IEEE, 2015, pp 147-154.

[4] Dinesh Bhuriya, Girish Kaushal, Ashish Sharma and Upendra Singh, "Stock Market Predication Using a Linear Regression.", International Conference of Electronics, Communication and Aerospace Technology, ICECA 2017, 20-22 April 2017, Coimbatore, India, IEEE, 2017, pp 510-513.

[5] M M. Goswami, C K. Bhensdadia, A P Ganatra, "Candlestick Analysis Based Short Term Prediction of Stock Price Fluctuation Using SOM-CBR.", IEEE International Advance Computing Conference, IEEE IACC 2009,6-7 March 2009, Patiala India, IEEE, 2009, pp 1448-1452.

[6] Seksan Sangsawad, Chun Che Fung, "Extracting Significant Features Based on Candlestick Patterns Using Unsupervised Approach.", 2nd International Conference on Information Technology, INCIT 2017, 2-3 November 2017, Nakhonpathom, Thailand, IEEE, 2017, pp 1-5.

[7] Hiransha M., Gopala Krishnan E.A., Menon V.K., Soman K.P. (2018)." NSE Stock Market Prediction Using Deep-Learning Models." Procedia Computer Science, Volume 132, 2018, pp 1351-1362.

[8] Jia H. (2016).” Investigation into the Effectiveness of Long Short Term Memory Networks for Stock Price Prediction.” arXiv preprint arXiv :1603.07893, 2016, pp 1-6.

[9] Guresen E., Kayakutlu G., and Daim T. U. (2011).” Using Artificial Neural Network Models in Stock Market Index Prediction.”, Expert Systems with Applications: An International Journal Volume 38 Issue 8, 2011, pp 10389-10397.

[10] Roondiwala, Murtaza & Patel, Harshal & Varma, Shraddha., “Predicting Stock Prices Using LSTM.”, International Journal of Science and Research (IJSR) Volume 6 Issue 4,2017, pp 1754-1756.

[11] Bollen J,Mao H,Zeng X. “Twitter Mood Predicts the Stock Market.”, Journal of Computer Science Volume 2 Issue 1, 2011, pp 1-8.

[12] Suresh Kumar, K. K., Elango, N. M., “An Efficient Approach to Forecast Indian Stock Market Price and their Performance Analysis.”, International Journal of Computer Application, Volume 34 Issue 5, 2011, pp 44-49.

[13] J. G. Agrawal, Dr. V. S. Chourasia, Dr. A. K. Mittra , ” State-of-the-Art in Stock Prediction Techniques.”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Volume 2 Issue 4, 2013, pp 1360-1366.

[14] Aditya Bhardwaj, Yogendra Narayan, Vanraj, Pawan, Maitreyee Dutta.,” Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty.”, Procedia Computer Science, Volume 70, 2015, pp 85-91.