Tribhuvan University
Institute of Engineering
Central Campus, Pulchowk

A MAJOR PROJECT REPORT ON

Stock Market Analysis and Prediction

By:

Abin Shakya (070/BCT/503)

Anuj Pokhrel (070/BCT/507)

Ashuta Bhattarai (070/BCT/510)

Pinky Sitikhu (070/BCT/524)

A PROJECT WAS SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULLFILLMENT OF THE REQUIREMENT FOR THE BACHELOR'S DEGREE IN ELECTRONICS & COMMUNICATION / COMPUTER ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING
LALITPUR, NEPAL

November, 2017

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PULCHOWK CAMPUS
DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

The undersigned certify that they have read, and recommended to the Institute of Engineering for acceptance, a project report entitled "Stock Market Analysis and Prediction" submitted by Abin Shakya, Anuj Pokhrel, Ashuta Bhattarai and Pinky Sitikhu in partial fulfilment of the requirements for the Bachelor's degree in Electronics & Communication / Computer Engineering.

Supervisor,
Prof. Dr. Subarna Shakya,
Department of Electronics & Computer Engineering,
Institute of Engineering, Pulchowk Campus,
Tribhuvan University, Nepal

Internal Examiner,
Head of Department,
Department of Electronics & Computer Engineering,
Institute of Engineering, Pulchowk Campus,
Tribhuvan University, Nepal

Dr. Diwakar Raj Pant,
Head of Department,
Department of Electronics & Computer Engineering,
Institute of Engineering, Pulchowk Campus,
Tribhuvan University, Nepal

External Examiner,
....
....
....

DATE OF APPROVAL :

COPYRIGHT

ACKNOWLEDGMENT

This project is done as per the requirement of the 4th year Major Project in Bachelors in Computer Engineering, Institute of Engineering (IOE), Tribhuvan University, Nepal.

We would like to express our deepest gratitude to Prof. Dr. Subarna Shakya for supervising this project with full dedication, support and guidance. Without him, this project would have taken twice as long to complete.

We would like to thank the Department of Electronics and Computer Engineering, IOE for giving us this great opportunity to carry out a major project that will help us shape our career. Among all the members of the department, our special thanks goes to the Head of Department, Dr. Diwakar Raj Pant, and the Deputy Head of Department, Mrs. Bibha Sthapit and Mr. Dinesh Baniya Kshatri, for providing us the support and guidance for this major project.

We would like to acknowledge all the authors of the research papers that helped us understand the required concepts and algorithms better, and that we utilized to prepare this report.

Every attempt has been made to include each and every aspects of the project in this report so that the reader can clearly understand about our project. We would be pleased to get the feedback on this project.

Sincerely,
Abin Shakya (070/BCT/503)
Anuj Pokhrel (070/BCT/507)
Ashuta Bhattarai (070/BCT/510)
Pinky Sitikhu (070/BCT/524)

ABSTRACT

Stock price and stock index price forecasting system, used by investors and financial managers to describe the market and compare the return on specific investments, has been a topic of research for very long now. When in the stock market, there are more buyers than there are sellers, the price must adapt or no trades are made. This tends to drive the price upwards, increasing the market quotation at which investors can sell their shares, enticing investors who had previously not been interested in selling and vice versa. These demands and supplies are ever changing, resulting in highly-fluctuating, non-linear stock prices which poses a threat against the credibility of those prediction systems which only view the market from one perspective. For a reliable system, it is therefore important to explore the market on multiple grounds.

The project studies the stock market through Technical, Fundamental and News Sentimental terms and combines various artificial intelligence and data mining techniques in order to predict the stock movements. Under technical analysis, Artificial Neural Networks (ANNs) is employed to analyze the nonlinear relationships between the stock closed price and various technical indexes, and to capture the knowledge of trading signals that are hidden in historical data. The fundamental analysis involves thorough study of financial statements of companies which gives an insight on the company's future performance. Unlike technical analysis, it helps predicting stock price on a long run. In news analysis, we focus on understanding the news sentiment and its affects which may cause the investors to either buy or sell the shares based on positivity or negativity of the news.

Thus, through the circumstantial application of the above-mentioned analysis, the project proposes to predict the stock market in a more generalized manner so that, it can be useful for analysts as well as for general investors for day to day trading.

Keywords : *Stock Market Prediction, Artificial Neural Network, Technical Analysis, Fundamental Analysis, News-Sentimental Analysis*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ASCII	American Standard Code for Information Interchange
BOW	Bag of Words
CBOW	Continuous Bag of Words
DFD	Data Flow Diagram
DOECE	Department of Electronics and Computer Engineering
EMA	Exponential Moving Average
EMF	Efficient Market Hypothesis
IDE	Integrated Development Environment
IOE	Institute of Engineering
IR	Information Retrieval
*k***-NN**	*k* Nearest Neighbor
MAE	Mean Average Error
MLP	Multilayer Perceptron
MVC	Model–View–Controller
NASDAQ	National Association of Securities Dealers Automated Quotations
NB	Naive Bayes
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	Neural Network
NEPSE	Nepal Stock Exchange Limited
NYSE	New York Stock Exchange
PyPI	Python Package Index
RMS	Root Mean Square
ROE	Return On Equity
PE	Price to Earning

SMA	Simple Moving Average
TF-IDF	Term Frequency – Inverse Document Frequency
TU	Tribhuvan University
UML	Unified Modeling Language

1. INTRODUCTION

1.1. Background

A stock market is a network of economic transaction of stocks that represents a claim of ownership on businesses and companies. To own the stock of a certain company simply means to own a part of the company and claim a right to participate in its earnings. A shareholder is someone who invests his/her money in a corporation expecting certain profit in return. It is therefore very important to evaluate a company's performance before buying its stocks. The ownership of a shareholder towards a company depends upon the volume of share s/he holds, with respect to the volume of outstanding share. The stock market runs on the concept of supply and demand. The stock of a company with a better profile, well known products and high profits is more in demand than a company with an average profile, lesser known products and low profits and therefore, costs more. However, unlike the trade of foreign currency, the prices of companies in the stock market fluctuate heavily, in a minute-to-minute basis. That is because, every buying or selling activity results in a heavy distortion in the supply and demand cycle. Thus, a person earning profit in a stock transaction ultimately means a loss in someone else's transaction.

There are two types of market for stock transaction, **Primary Market and Secondary Market**. The primary market deals with new securities. It is the market for new long term equity capital, issued by the company directly to the investors for setting up new business or for expanding the existing business. [**?**] In a primary market, companies, governments or public sector institutions can raise funds through bond issues and corporations can raise capital through the sale of new stock through an initial public offering (IPO). It performs a crucial function of facilitating capital formation in the economy and therefore, is also referred to as New Issue Market (NIM).The secondary market is the market where previously issued financial stocks are bought and sold. The secondary market is required to be highly liquid so that, both the buyer and the seller can get the actual worth of return from the transaction. For this purpose, marketplace centralization is a must. Exchanges such as Nepal Stock Exchange (NEPSE), New York Stock Exchange (NYSE) provide a centralized and liquid platform to serve as a secondary marketplace for the investors.

[?] NEPSE is the one and only secondary marketplace in Nepal. Established before 1993 under the name 'Securities Exchange Center', NEPSE first opened its trading floor on 13th January, 1994. Since then, it has been working to fulfill its objective of imparting free marketability and liquidity to the government and corporate securities by facilitating transactions in its trading floor through stock brokers. As of now, more than 200 companies from Banking, Finance, Hydropower, Hotels, Development Banks Manufacturing and Co., Insurance and Trading sectors have registered in NEPSE. Stock market plays a vital role in a nation's economy. [?] A collapse in share prices has the potential to cause widespread economic disruption. The famous stock market crash of 1929 was a key factor in causing the great depression of the 1930s. Along with individual wealth, the stock market also has a huge impact on the well-being of financial corporations and firms. The ability of a company to perform well in the industrial sector is related directly with its ability to perform well in the stock market. If a company is going through a bearish trend, it may have to struggle financially. Moreover, the investors' trust in the company is also hampered in such cases. In turn, the overall performance of all industrial and financial firms associated with the stock market can have a huge impact in the overall economy of a country. Multinational companies with worldwide investors have an even bigger impact in the economy of individuals in multiple countries. In the same way, the stock market trends also affect the investments in other sectors like: forex, gold and land investments.

Hence, considering all the factors which affect and which are affected by the stock market, it can be said that the stock market prediction is a very sensitive topic. There are many theories such as Efficient Market Hypothesis (EMH) which claim that, outsmarting the market is impossible in all cases. Still, considerable efforts are being made in this sector using various artificial intelligence and deep learning techniques.

1.2. Objectives

The main objectives of this projects are as follows :

- To predict the stock market trend based on technical, fundamental and news-sentimental analysis

- To visualize the prediction results and daily trading prices in the form of interactive charts
- To compare the results and effectiveness of the algorithms: artificial neural network with backpropagation, knn and naive bayes.

1.3. Problem Statement

Theoretically, the stock market is said to be very difficult to predict, due to its dynamic and non-linear model. However, the investors and stock analysts have been trying to somehow predict the stock prices of a company, to increase the profit in buying and selling stocks. Appreciable efforts have also been made from academic researchers and enthusiasts in this field. However, identifying the pattern of such an uncertain system through simple calculations and mathematics results in poor accuracy with questionable reliability. The overall hit rates of these methodologies and models are generally too low to be practical for real-world application. Complex, dedicated systems and models are required which can take into consideration, the numerous factors that can affect the stock price of a company. For an instance, the intrinsic valuation of a company and its performance in the market till now are equally important factors in determining its future price. However, it is very difficult to know for certain which factor affects the most at the given time, and by how much. Therefore, the market should be analyzed under various influencing factors, the prime of which are: Technical factors, Fundamental factors and News-sentimental factors. In technical analysis, the prediction model is built considering a company's past performance in the stock market, which includes studying the past rise and fall trends, average traded volumes, bullish and bearish trend behaviors and so on. It is based on the assumption that history repeats itself and that future market directions can be determined by examining the way the market has behaved before. Thus, it is assumed that price trends and patterns exist that can be identified and utilized for predictions. In fundamental analysis, the worth of the company, its current profits, capital gains and the future profits plays a vital role in understanding its stock price behaviors. In news-sentiment analysis, the immediate effects of political, economic and stock related news in a company's stock prices is studied and applied.

Therefore, through circumstantial application of above mentioned analysis, this project presents a general and complete solution for stock prediction, which can be employed in the real world for gaining profit in the stock market.

1.4. Scope

The project aims to predict the stock trend movements of trading companies based on large volume of historical data collected from various sources. The historical data constitutes of a company's fundamental valuations, past trading prices and volumes, and past news features. The basic driving factors for choosing a prediction model is its effectiveness, applicability and accuracy of results. By using multiple analysis and prediction models, the project aims to compare the usability of each such models. On completion of this project, we aim to establish a highly reliable stock prediction system, which can be used by investors to decide when to buy or sell the stocks of a company in order to gain maximum profit. It is hoped that the project will be beneficial for the stakeholders including, researchers, business analysts, stock market enthusiasts and policy makers. The project is also focused on improving the trading experience of new investors who may or may not know much about the market behaviors.

2. LITERATURE REVIEW

2.1. Theory Details

Stock Market Prediction is a hot topic in data mining. Many analysts and researchers have done a lot of work in this field applying various data mining and statistical techniques to develop stock prediction models. The most popular methods used for stock market prediction are Dynamic time series model, Hidden Markov Model, Bayesian Classifiers, Artificial Neural Network, technical analysis, fundamental analysis and so on. All these methods work in different manner and work on different precision levels.

When predicting the future prices of Stock Market securities, there are several theories available. The first is Efficient Market Hypothesis (EMH). [**?**] In EMH, it is assumed that the price of a security reflects all of the information available and that everyone has some degree of access to the information. Fama's theory further breaks EMH into three forms: Weak, Semi-strong, and Strong. In Weak EMH, only historical information is embedded in the current price. The Semi-Strong form goes a step further by incorporating all historical and currently public information in the price. The Strong form includes historical, public, and private information, such as insider information, in the share price. From the tenets of EMH, it is believed that the market reacts instantaneously to any given news and that it is impossible to consistently outperform the market.

A different perspective on prediction comes from Random Walk Theory. [**?**] In this theory, Stock Market prediction is believed to be impossible where prices are determined randomly and outperforming the market is infeasible. Random Walk Theory has similar theoretical underpinnings to Semi-Strong EMH where all public information is assumed to be available to everyone. However, Random Walk Theory declares that even with such information, future prediction is ineffective.

It is from these theories that two distinct trading philosophies emerged; the fundamentalists and the technicians. In a fundamentalist trading philosophy, the price of a security can be determined through the nuts and bolts of financial numbers. These numbers are derived from the overall economy, the particular industry's sector, or most typically, from the company itself. Figures such as inflation, joblessness, return on equity (ROE), debt levels, and individual Price to Earnings (PE) ratios can all play a part in determining the price of a stock.

In contrast, technical analysis depends on historical and time-series data. These strategists believe that market timing is critical and opportunities can be found through the careful averaging of historical price and volume movements and comparing them against current prices. Technicians also believe that there are certain high/low psychological price barriers such as support and resistance levels where opportunities may exist. They further reason that price movements are not totally random, however, technical analysis is considered to be more of an art form rather than a science and is subject to interpretation.

Both fundamentalists and technicians have developed certain techniques to predict prices from financial news articles. In one model that tested trading philosophies, [**?**] LeBaron et. al. posited much can be learned from a simulated stock market with simulated traders. In their work, simulated traders mimicked human trading activity. Because of their artificial nature, the decisions made by these simulated traders can be dissected to identify key nuggets of information that would otherwise be difficult to obtain. The simulated traders were programmed to follow a rule hierarchy when responding to changes in the market; in this case it was the introduction of relevant news articles and/or numeric data updates. Each simulated trader was then varied on the timing between the point of receiving the information and reacting to it. The results were startling and found that the length of reaction time dictated a preference of trading philosophy. Simulated traders that acted quickly formed technical strategies, while traders that possessed a longer waiting period formed fundamental strategies. It is believed that the technicians capitalized on the time lag by acting on information before the rest of the traders, which lent this research to support a weak ability to forecast the market for a brief period of time

In similar research on real stock data and financial news articles, [**?**] Gidofalvi gathered over 5,000 financial news articles concerning 12 stocks, and identified this brief duration of time to be a period of twenty minutes before and twenty minutes after a financial news article was released. Within this period of time, Gidofalvi demonstrated that there exists a weak ability to predict the direction of a security before the market corrects itself to equilibrium. One reason for the weak ability to forecast is because financial news articles are typically reprinted throughout the various news wire services. Gidofalvi posits that a stronger predictive ability may exist in isolating the first release of an article. Using this twenty-minute window of opportunity and an automated textual news parsing system, the possibility exists to capitalize on stock price movements before human traders can act.

Also Ralph Nelson Elliott developed the [?] Elliott wave theory in the late 1920s by discovering that stock markets, thought to behave in a somewhat chaotic manner, in fact traded in repetitive cycles. Elliott discovered that these market cycles resulted from investors' reactions to outside influences, or predominant of psychology of the masses at the time. He found that the upward and downward swings of the mass psychology always showed up in the same repetitive patterns, which were then divided further into patterns he termed "waves".

For the stock market prediction, Artificial Neural Network(ANN) has been considered the most efficient method. Inspired by neurosciences, ANNs have shown great potential in terms of recognizing patterns in nonlinear systems. Existing research suggests that ANN is an eminent model to predicting stock markets due to its dynamical characteristics. Even so, a common criticism of neural networks is that they require a large diversity of training for real-world operation. Moving average analysis and single exponential smoothing methods are frequently used in order to make stock analysis. The Nepal stock exchange (NEPSE) uses exponential smoothing in its website for this purpose. Moving averages work quite well in strong trending conditions, but often poorly in choppy or ranging conditions.

Under the assumption that the stock market could be predicted, there are some major cateogories of prediction methods: fundamental analysis, technical analysis and news analysis.

2.2. Related Work

Many algorithms of data mining have been proposed to predict stock price. Neural Network, Genetic Algorithm, Decision Tree and Fuzzy systems are widely used. In addition, pattern discovery is beneficial for stock market prediction and public sentiment is also related to predicting stock price. Projects have been done on predicting stock value based on Fuzzy logic.

There are a lot of software and web applications working with the similar concept. Nepal Sharemarket is a website that makes individual, comparative as well as in depth analysis on stock market companies and also forecasts their price on a chosen time basis. Another website, *stockforecasting.com* also makes stock prediction using neural networks and boasts of highest accuracy among all the stock-prediction applications. It is an American company and gives minute predictions of various international companies.

[**?**] *Iknowfirst.com* uses predictive forecast algorithm based on artificial intelligence and machine learning with elements of Artificial Neural Networks and Genetic Algorithms incorporated in it. Its system outputs the predicted trend as a number, positive or negative, along with the wave chart that predicts how the waves will overlap the trend. This helps the trader decide which direction to trade, at what point to enter the trade, and when to exit.

[**?**] MetaStock is a widely acclaimed software, the edge it has is excellent news service, expert advisors and system development, with a huge range of indicators and powerful scanning. It can provide excellent earlier shorter term signals of trend changes that allow the investor to fine tune the trade.

3. METHODOLOGY

3.1. Data Collection

Both analysis and prediction of stock market needed an extensive amount of data for better visualization and training. News data regarding stock market were required for news analysis, which were collected from *sharesansar.com* website via web crawling. Trading data of listed companies, for technical analysis were collected from *merolagani.com* website. Similarly, Sector wise data required for Nepal Stock Exchange Ltd.(NEPSE) index prediction were collected from *sharesansar.com*. Unfortunately, the data required for Fundamental Analysis were not available for crawling. As a result, required data were extracted manually from the web.

Aside from the static past data, a mechanism to update the recent changes was also required to update the database constantly. For this, a manual update module was built, which at the end of the day, updates the changes made throughout the day in the database.

3.2. Data Preprocessing

The trading data crawled from *Merolagani.com* and Sharesansar.com had numerous missing fields, which were filled up using interpolation technique to cover up the possible setbacks. Once the data is cleaned, it is stored in the database for future retrievals.

The training data on analysis showed high fluctuation, which needed some smoothing technique in order to feed it into our model for better results. Thus, Exponential Moving Average was used to reduce the data into suitable form.

For the news analysis, the news so collected are to be divided into feature set and label. The feature set were extracted using 'Bag of words representation'. The label were given to each vector as positive, negative based on whether the corresponding price increased or decreased. All the news were labeled accordingly to get a complete training data set.

3.3. Data analysis and visualization

Data Analysis in financial market involves two basic approaches and they are: Technical analysis and Fundamental analysis. *Technical analysis*, which involves detecting patterns in security prices, goes on the assumption that the price of a stock is like the price of everything else, is a matter of supply and demand. Technical analysis generates and interprets charts of the price and volume histories of stocks to predict movement in stock prices according to perceived trends. *Fundamental analysis*, which examines the earning potential of the company issuing a stock, goes on the assumption that a share of ownership of a company has an intrinsic value that is a function of the underlying value of the company as a whole. Fundamental analysis reports which shares are undervalued by the investor community and which are overvalued, then trust the market to make corrections.

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Data visualization was done with the help of charting library called AmCharts. Company wise data were shown in charts and features like comparison of stock data were also integrated.

3.4. Feature Selection

The data features that are used to train machine learning models have a huge influence on the performance that can be achieved. Irrelevant or less relevant selection of data features result in low performance during prediction. So, it is necessary to select the best possible features that best influence the result.

Benefits of performing feature selection before modeling the data are:

- Reduces Over-fitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.

Statistical tests can be used to select those features that have the strongest relationship with the output variable. The scikit-learn library provides the Select K-Best class that we used with a suite of different statistical tests to select a specific number of features.

3.5. Prediction

The system intends to predict the stock market using the available historical data. The prediction model is generated by manipulating the historical data by casting them through various artificial intelligence techniques. In general, stock market prediction can be done by analyzing the past stock trends with respect to fundamental, technical and news-sentiment analysis.

3.5.1. Prediction using Fundamental Analysis

i. Book Value and Market Value Comparison

[?] Understanding the difference between book value and market value is a simple yet fundamentally critical component of any attempt to analyze a company for investment. After all, when we invest in a share of stock or an entire business, we want to know we are paying a sensible price.

Book value literally means the value of the business according to its "books" or financial statements. In this case, book value is calculated from the balance sheet, and it is the difference between a company’s total assets and total liabilities. Note that this is also the term for shareholders’ equity. For example, if Company ABC has total assets of $100 million and total liabilities of 80 million, the book value of the company is 20 million. In a very broad sense, this means that if the company sold off its assets and paid down its liabilities, the equity value or net worth of the business, would be 20 million.

Market value is the value of a company according to the stock market. Market value is calculated by multiplying a company’s shares outstanding by its current market price. If Company ABC has 1 million shares outstanding and each share trades for 50, then the company’s market value is 50 million. Market value is most often the number analysts, newspapers and investors refer to when they mention the value of the business.

Book value simply implies the value of the company on its books, often referred to as accounting value. It's the accounting value once assets and liabilities have been accounted for by a company's auditors. Whether book value is an accurate assessment of a company's value is determined by stock market investors who buy and sell the stock. Market value has a more meaningful implication in the sense that it is the price we have to pay to own a part of the business regardless of what book value is stated.

As we can see from our fictitious example from Company ABC above, market value and book value differ substantially. In the actual financial markets, we will find that book value and market value differ the vast majority of the time. The difference between market value and book value can depend on various factors such as the company's industry, the nature of a company's assets and liabilities, and the company's specific attributes. There are three basic generalizations about the relationships between book value and market value.

1. **Book Value Greater Than Market Value :** The financial market values the company for less than its stated value or net worth. When this is the case, it's usually because the market has lost confidence in the ability of the company's assets to generate future profits and cash flows. In other words, the market doesn't believe that the company is worth the value on its books. Value investors often like to seek out companies in this category in hopes that the market perception turns out to be incorrect. After all, the market is giving us the opportunity to buy a business for less than its stated net worth.
2. **Market Value Greater Than Book Value :** The market assigns a higher value to the company due to the earnings power of the company's assets. Nearly all consistently profitable companies will have market values greater than book values.

In this method we calculate the Book value of a company, that is its Total Equity by Subtracting Liabilities from the assets.

Total Equity= Assets - Liabilities

Then we calculate Book value per share as follows :

Book value per share = Total Equity / Total Shares Outstanding

Thus we present user information about the Book value of the share and market value that can help in effective decision making.

3.5.2. Prediction using Technical Analysis

[?] The field of technical analysis is based on three assumptions:

- The market discounts everything
- Price moves in trends
- History tends to repeat itself

Technical analysis studies the trend of supply and demand within the market to determine what direction or trend will continue in the future. In other words, it attempts to understand the emotions in the market by studying the market itself, as opposed to its components.

1. Technical analysis using Artificial Neural Network

Artificial Neural Networks (ANNs) are simply inspired from biological neural networks that make up the networks of living neuron cells in animals. Computer systems excel the human brain in performing complex mathematical operations by thousands of times but, lack the human ability of logical reasoning and pattern recognition. The use of ANN allows computers to process data the same way the human brain processes a stimulus, providing them the ability to recognize patterns even in non-linear data such as that of stock market. For this process, the ANN is trained with historical data using supervised learning method. Once the training is completed, we move on to the testing phase, where the reliability, accuracy and efficiency of the training algorithm is tested. Once the ANN has passed the test, it can then be used for prediction. The artificial neural network model is displayed in Figure 3.1.

ANN is considered one of the most effective methods in predicting the stock market. Even within ANN, the Multi-Layer Perceptron (MLP) model is widely accepted for effective pattern recognition. [?] An MLP model is a class of feed-forward Artificial Neural Network that consists of at least three layers of nodes namely: Input layer, Hidden layer(s) and Output layer. MLP utilizes supervised learning technique and Backpropagation algorithm for training.

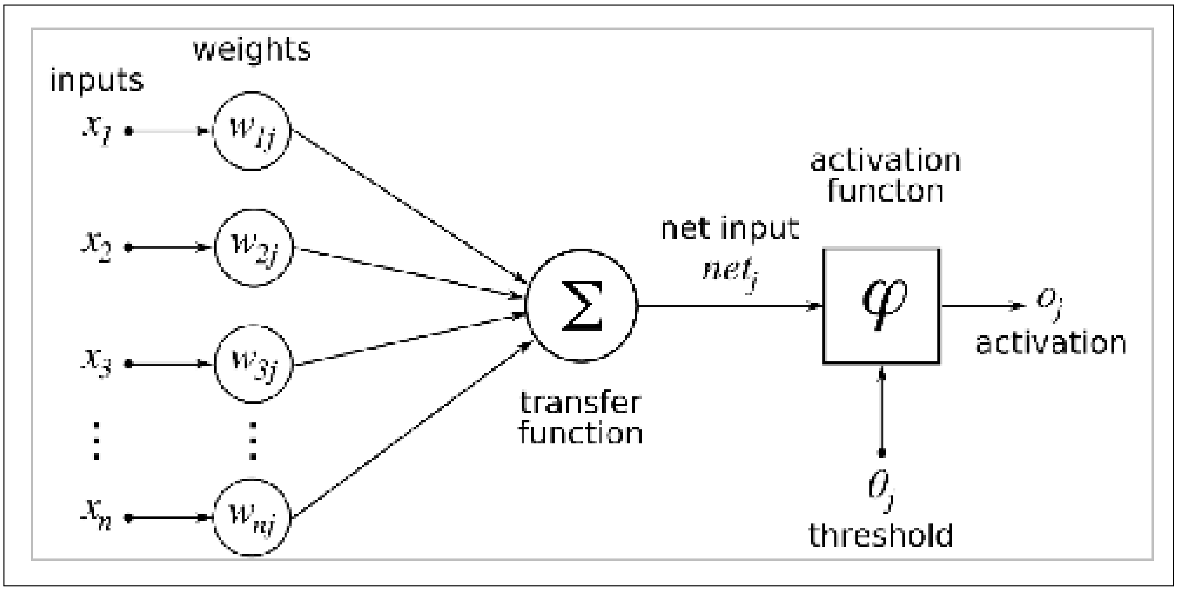


Figure 3.1: ANN Model

The artificial neurons shown in figure 3.1 have n number of inputs: $x_1, x_2, ..., x_n$ each of which is associated with a weight onto the connection line, denoted as $w_{1j}, w_{2j}, ..., w_{nj}$ respectively. The weights can be referred as synaptic weights as in a biological neural network. 'θ' represents the threshold and 'α' represents the activation function given by:

$\alpha = \sum_{k=1}^{n} w_{kj} * x_k + \theta$

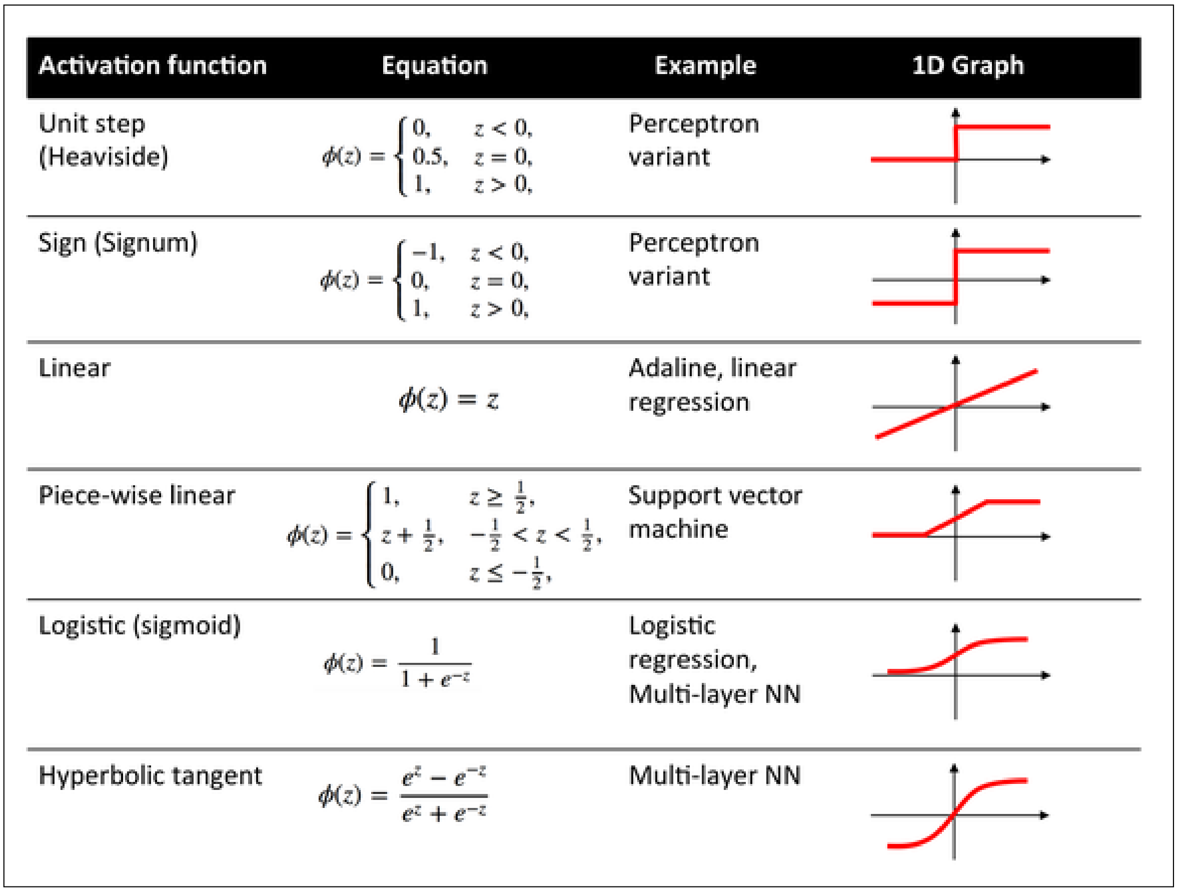
The output of the neuron, Oj, is a function of its activation given by:

$\mathrm{O}_j = f(\alpha)$

Several types of activation functions can be used, which are summarized in Figure 3.2:

The factors taken as input for the neural network are:

- Opening Price: the price of a company stock at the beginning of the trading day
- Closing Price: the price of a company stock at the end of the trading day
- High Price: the maximum price reached by a company stock in the entire trading day
- Low Price: the minimum price reached by a company stock in the entire trading day



Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \dfrac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \dfrac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multi-layer NN	

Figure 3.2: Table showing activation function

- Number of transactions: the total number of a company stock transactions within a trading day
- Traded Volume: the total volume of stocks traded within the entire trading day

Among the activation functions shown in figure 3.2, the Logistic (sigmoid) function is taken which best represents the non-linear feature of the dataset.

2. Backpropagation Algorithm

The chief objective of the Backpropagation Algorithm is to reduce the error function. [?] This algorithm falls into the general category of gradient descent algorithms, which intend to find the minima/maxima of a function by iteratively moving in the direction of the negative of the slope of the function to be minimized/maximized. This algorithm proceeds across the network, providing activation to each node until the output node is reached. Then, the weights are updated backwards, from the output layer towards the input layer until one epoch has been completed. The weights are updated according to the respective errors computed for each layer.

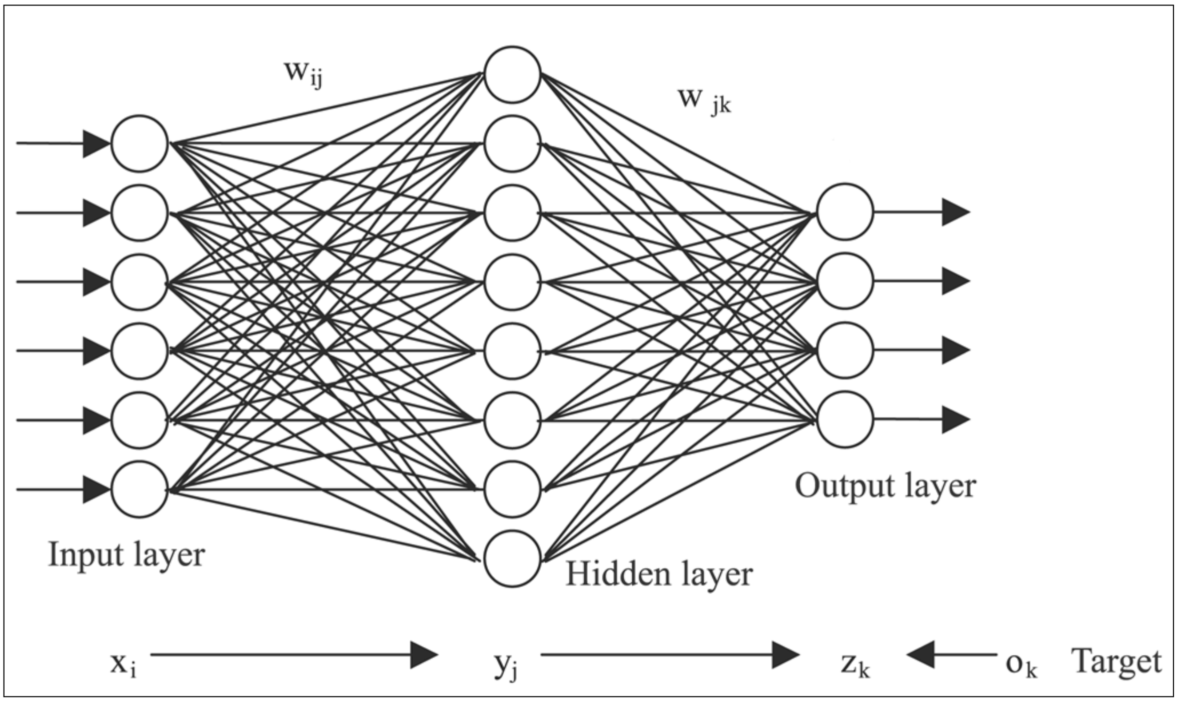


Figure 3.3: Backpropagation Algorithm

Given the figure 3.3, for k output units, if t_k signifies the target value, o_k signifies the actual output, α signifies the learning rate and z_{ink} signifies the activation function for the k^{th} output node, error (δ) is given by:

$\delta_k = (\mathrm{t}_k - o_k) * f'(z_{ink})$

Now, the weight correction for each output unit is given by: Δ w$_{jk}$ =α * δ_k * z$_j$(3.0)

Similarly, by propagating the delta term further back in the network, the input error in the hidden network is calculated.

$\delta_{inj} = \sum_{k=1}^{m} \delta_k * w_{jk}$ Where, m = number of neurons in the hidden layer

Now, error in the jth hidden unit is calculated by: $\delta_j = \delta_{inj}$ * $f\acute{}(x_{inj})$ Where x_{inj} signifies the activation function for the j^{th} hidden layer node.

Now, the weight correction is given by:

Δ w$_{ij}$ $=\alpha * \delta_j * X_i$, for hidden layer nodes

Δ w$_{jk}$ $=\alpha * \delta_k * Y_j$, for output layer nodes

Then, the weights for each neuron is updated with the new ones. Once an epoch has been completed, the average error for each training data is calculated. Usually the RMS error between the target value and actual outputs is computed for convergence. If the RMS error falls within the acceptable range, the training is completed, else, the whole process is repeated.

During backpropagation training, before entering the input factors in the neural network model, they are subjected to various data cleaning and pre-processing functions. The missing data are filled using interpolation techniques. The huge set of dynamic and heavily fluctuating dataset is made smooth for better calculation by taking exponential moving average (EMA) of a fixed set of data. The EMA is calculated using the formula:

EMA1 = (P* (2/(1+N) + [EMA0 * (1-(2/(1+N))])

Where, P represents price and EMA0 represents the EMA of previous calculation

Hence, the above algorithm can be used to train an Artificial Neural Network. The network, in general, can have an arbitrary number of hidden layers and an arbitrary number of hidden neurons in each layer. For practical reasons, ANNs implementing the backpropagation algorithm do not have too many layers, since the time for training the networks grows exponentially. The number of input layer neurons is decided by the number of input features in each pattern, and the number of output layer neurons is decided by the number of output features in the target values.

There are a few disadvantages associated with backpropagation learning as well:

- The convergence obtained from backpropagation learning is slow and not guaranteed.
- The result may generally converge to any local minimum on the error surface, since stochastic gradient descent exists on a surface which is not flat.
- Backpropagation learning requires input scaling or normalization.

- Backpropagation requires the activation function used by the neurons to be differentiable.

3. K- nearest neighbor

The k-nearest neighbors algorithm is a non-parametric method used for classification and regression. [?] It was used for classification of stock trends of individual companies in our project. The input consists of the k closest training examples in the feature space. The output in k-NN classification is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. The K-Nearest Neighbor is a simple lazy learner algorithm that stores all available data points and classifies new instances based on a similarity measure. It is defined by a set of objects known as examples for which the outcomes are known.

During the training phase the algorithm simply stores the data points including their class labels and all computation is deferred until the classification process. It is based on a principle that instances that are in close proximity to another have similar properties. Thus, to classify new unclassified instances, one simply has to look at their k-nearest neighbors, to figure out the classification label. The class membership can be defined by a majority vote of the k closest neighbors or the neighbors can be ranked and weighted according to their distance to the new instance. When new case of dependent values are given k-NN finds the outcomes by finding K examples that are closest in distance to the query point.

The choice of K is very important in building the model. [?] The k is an important factor that can influence the quality of predictions. For any problem, a small value of k will lead to large variance in predictions. On the other hand setting k to a large value may lead to large model bias. Thus, k should be set to a value large enough to minimize the probability of misclassification and small enough so that the nearest points are close enough to the query point.

The KNN model is displayed in the Figure 3.4.

The neighbors are taken from a set of objects for which the class is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. For this project the training set was created by making a table of six columns where the first five columns would list the difference in the closing price of that day to the previous day.

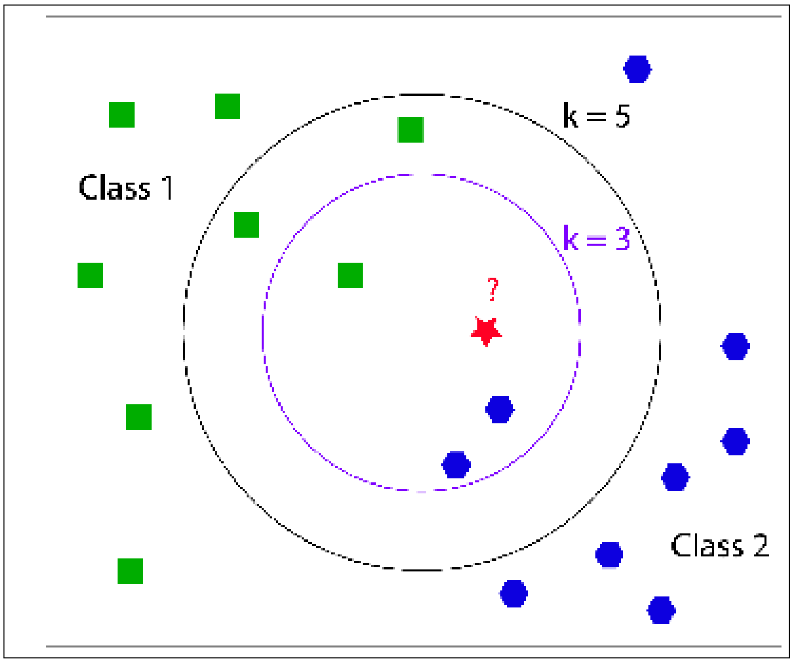


Figure 3.4: KNN model

The last column would however list either 1 or 0 based on whether the difference in closing price of that day to the previous day increased or decreased. Hundreds of these data would now create hundreds of classes for the test data to find a nearest neighbor from.

3.5.3. Prediction using News Sentiment Analysis

This system intends to predict the stock market based on the analysis of the past news headlines. Various natural language processing and machine learning techniques are combined in order to generate the classification model. Natural language toolkit (NLTK) was used for tokenizing and pre-processing the news headlines. The processed news were then trained with the classifier and a testing model was prepared. Naive bayes classifier was used for training and testing the data.

The training set of the data was prepared in two phases. In the first phase, the news headlines relating in any way to stock market from news websites *sharesansar.com* were taken for training set. The news headlines from May 2016 to June 2017 were taken for training set. The headlines were classified at first, on the basis of the next day's nepse index value. If the value of nepse index increased, then all the news from previous day were considered as positive news and vice versa. But the classification was not enough to generate good prediction hence further refinement was necessary. In second phase of classification, these classified news headlines were classified manually. Manual classification involved reading each news headlines and assigning it a sentiment tag: positive, negative or neutral. The second phase of the classification needed much research on the minute details.

Now a refined dataset was prepared for training the classifier. News headlines from past two days are collected as an unknown test dataset. The classifier predict the increment/decrement of NEPSE index by calculating the total number of positive, negative and neutral news headlines.

1. Text preprocessing

News headlines are unstructured data and these data needs preprocessing to input into classifier. For text preprocessing, sentence tokenizer and word tokenizer from NLTK modules were used. **Tokenizer** is used to divide strings into lists of substrings.**Sentence tokenizer** is used to find the list of sentences and **word tokenizer** is used to find the list of words in strings. After tokenizing, stop words were removed from the words. **Stop words** are the words that are to be filtered out before training the classifier. These are usually high frequency words that aren't giving any additional information to our labelling. In fact, they actually confuse the classifier. Example: is, the, at, etc.

2. Sentiment Detection

[**?**] For the sentiment detection of news articles, dictionary based approach is being followed which uses Bag of Word techniques for text mining. This method is based on the research of [**?**] J. Bean in his implementation of Twitter sentiment analysis for airline companies. From Bag of Word model, the features from news headlines are extracted and the features from model and the sentiment for each news headlines are inputed into naive bayes classifier for training the classifier. Then a test dataset is inputed to the classifier to predict the sentiments from the news headlines.

i. Bag of Words

The Bag of Words model is a simplifying representation used in natural language processing and information retrieval. In this model , a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. The Bag of Words model learns a vocabulary from all of the documents, then models each document by counting the number of times each word appears.

For example, consider the following two sentences:

Sentence 1: "The cat sat on the hat"

Sentence 2: "The dog ate the cat and the hat"

From these two sentences, our vocabulary is as follows: the, cat, sat, on, hat, dog, ate, and

To get the bags of words, the number of times each word occurs in each sentence is counted. In Sentence 1, "the" appears twice, and "cat", "sat", "on", and "hat" each appear once, so the feature vector for Sentence 1 is:

the, cat, sat, on, hat, dog, ate, and

Sentence 1: 2, 1, 1, 1, 1, 0, 0, 0

Similarly, the features for Sentence 2 are: 3, 1, 0, 0, 1, 1, 1, 1

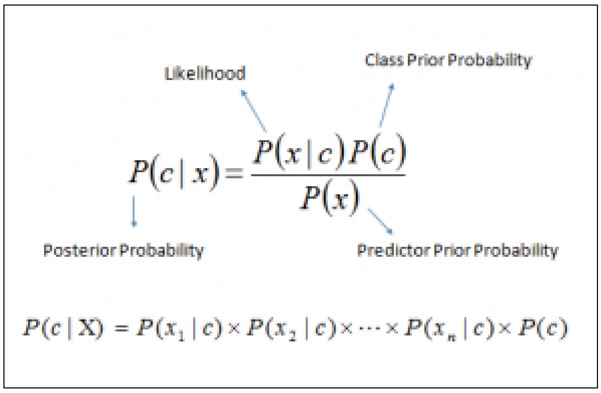
The feature extraction module from scikit-learn is used to create bag-of-words features.

ii. Naive Bayes Classifier

[**?**] Naive Bayes is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).

Look at the equation 3.5,

Where, P(c|x) is the posterior probability of class(c,target) give predictor(x, attributes).



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Figure 3.5: Equation of Bayes theorem

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

Pros and Cons of Naive Bayes Classifier :

Pros:

- It is easy and fast to predict class of test data set. It also perform well in multi-class prediction
- When assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Cons

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, the smoothing technique can be used. One of the simplest smoothing techniques is called Laplace estimation.

- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible to get a set of predictors which are completely independent.

Scikit learn (python library) also help to build a Naive Bayes model in Python. There are three types of Naive Bayes model under scikit learn library:

- **Gaussian :** It is used in classification and it assumes that features follow a normal distribution.
- **Multinomial :** It implements the naive Bayes algorithm for multinomially distributed data, and is one of the two classic naive Bayes variants used in text classification (where the data are typically represented as word vector counts, although tf-idf vectors are also known to work well in practice).
- **Bernoulli :** The binomial model is useful if your feature vectors are binary (i.e. zeros and ones). One application would be text classification with 'bag of words' model where the 1s & 0s are "word occurs in the document" and "word does not occur in the document" respectively.

4. Software development methodology

4.1. Software Development Life Cycle

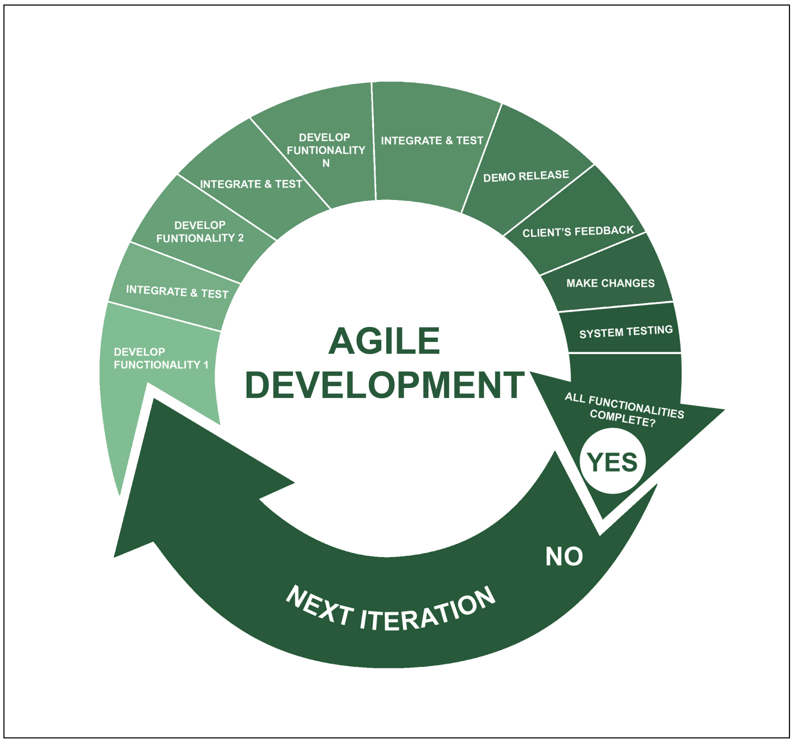


Figure 4.1: Agile Development Life Cycle

For the project Agile software development life cycle was followed. [?] The Agile software development life cycle is based upon the iterative and incremental process models, and focuses upon adaptability to changing product requirements and enhancing customer satisfaction through rapid delivery of working product features and client participation. Agile methods primarily focus upon breaking up the entire product into smaller, easily developable, "shippable" product features developed through "incremental" cycles known as "sprints". Agile methodology is an alternative to traditional project management, typically used in software development. Agile methodologies are an alternative to waterfall, or traditional sequential development. Agile development model is a type of Incremental model. Software is developed in incremental, rapid cycles. This results in small incremental releases with each release building on previous functionality. Each release is thoroughly tested to ensure software quality is maintained. Agile model is generally used for time critical applications.

The Agile Manifesto is based on twelve principles :

- Customer satisfaction by early and continuous delivery of valuable software
- Welcome changing requirements, even in late development
- Working software is delivered frequently (weeks rather than months)
- Close, daily cooperation between business people and developers
- Projects are built around motivated individuals, who should be trusted
- Face-to-face conversation is the best form of communication (co-location)
- Working software is the principal measure of progress
- Sustainable development, able to maintain a constant pace
- Continuous attention to technical excellence and good design
- Best architectures, requirements, and designs emerge from self-organizing teams
- Regularly, the team reflects on how to become more effective, and adjusts accordingly

The major characteristics of this model are listed below:

1. Self-organisation and motivation takes precedence over delegation of authority and following the "seniority" hierarchy. The Agile team has to collaborate and share ideas to develop the product "as a whole" unit i.e. each member should support a common vision.

2. Agile concentrates upon delivering sustained "working" product releases through product incremental cycles over documentation and working protocols. The main objective is to develop, and deliver, bug free product feature releases in a continuous and sustained manner until the entire product is developed.

3. Agile focuses upon incorporating dynamic changes in the product development cycle. Changes in the product features can be easily and effortlessly carried out by developing "user stories"- product functionality or features as defined in the product backlog.

4. Stakeholders and project owners "clear" the product features developed through the sprint cycles. A lot of time is saved through customer collaboration, and the project proceeds in a successful manner as the client always approves the development keeping in mind the current market trends.

The project is started by preparing the backlog. The backlog contains the modular decomposition of the overall system. Then the requirement specification and system model diagrams are prepared. Agile is an adaptive model which allows continuous changes in the system requirements as well the system model.

4.2. Requirement Analysis

The functional and non-functional requirements of this project are as listed below.

4.2.1. Functional Requirements

- The system shall examine the stocks of different companies based on the technical indicators, fundamental factors and news sentiments.
- The system shall predict the current value of a stock based on past values of that stock.
- The system shall predict the increase or decrease of the NEPSE index based on news analysis.
- The system shall predict the intrinsic worth of a company based on the fundamental indicators of the company.
- The system shall provide visualization of the stock market's data for individual companies of different sectors.

4.2.2. Non-Functional Requirements

- The system must predict stock market with acceptable accuracy.

- The data used for the prediction and analysis should be real and fault proof. Irregularities in data should be removed by using various techniques.
- The prediction system should be dynamic enough to easily adapt with the daily increase in the number of data.
- The visualization should be adaptive to dynamically add new data.

4.3. Backlog

The backlog of project is described in the table 4.1

4.4. System Design

4.4.1. System Architecture

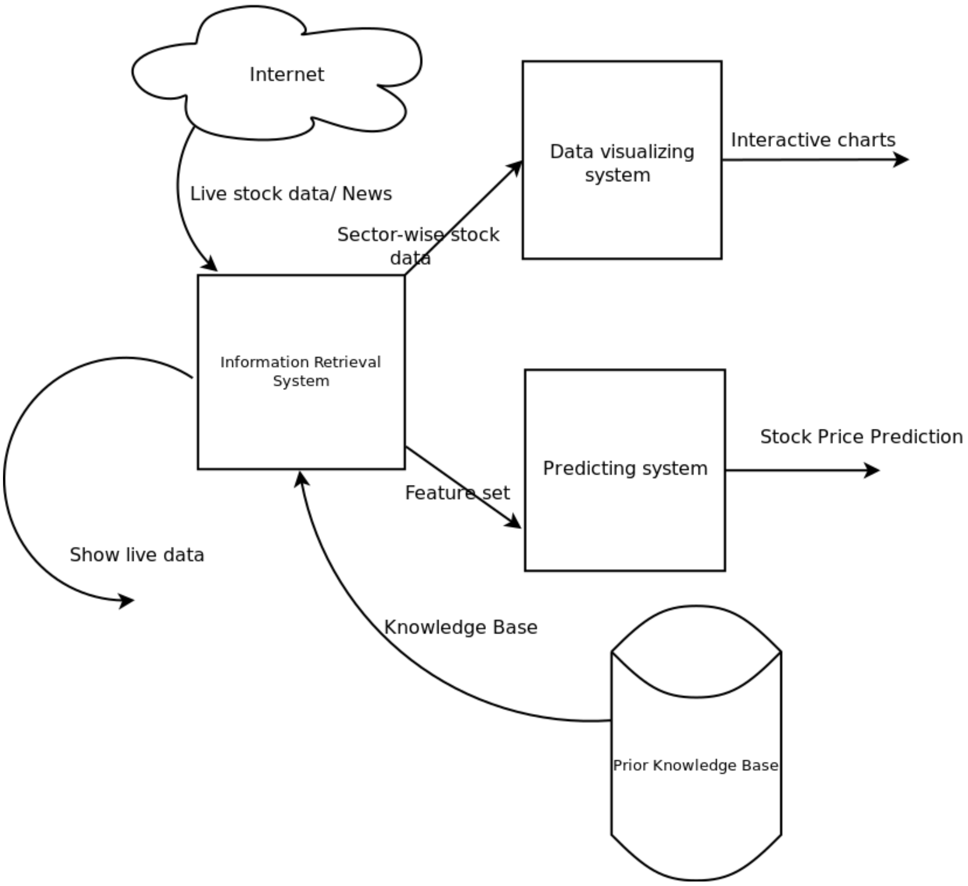


Figure 4.2: Overall system architecture

Story	Task	Time Estimation(Days)	Time Estimation(Hours)
As a user, I should have access to stock data	Crawl data from Merolagani, Sharesansar and NEPSE website	7	35
	Clean the data obtained from the web	1	5
	Design a data model and represent the data in proper format	7	35
	CRUD on the available data set	7	35
	Design simple User Interface to view the stock data	6	30
	Perform Tests	2	10
As a user, I should be able to visualize the stock data for various companies and sector, compare and analyze them	Represent data in proper format for visualization	3	15
	User Interface for data visualization	3	15
	Implementation of data visualization using Amcharts	7	35
	Design and write test cases	4	20
	Perform test	1	5
As a user, I should be able to view the fundamental background of a company	Represent the fundamental data in proper format	5	25
	User Interface for viewing fundamental data	4	20
	Integrate back-end and front-end	4	20
	Design a model of data for prediction	7	35
As a user I should be able to predict the company wise stock price	Design the prediction engine using ANN	25	125
	Implement the prediction engine	15	75
	Refine the ANN model	5	25
	Design the prediction engine using KNN	5	25
	Implement the prediction engine	2	10
	Test the prediction engine	2	10
As a user I should be able to predict the nepse index	Design a news data model for prediction	10	50
	Design a prediction using KNN	5	25
	Implement the prediction engine	2	10
	Test the prediction engine	2	10

Table 4.1: Backlog

4.4.2. UML Diagrams

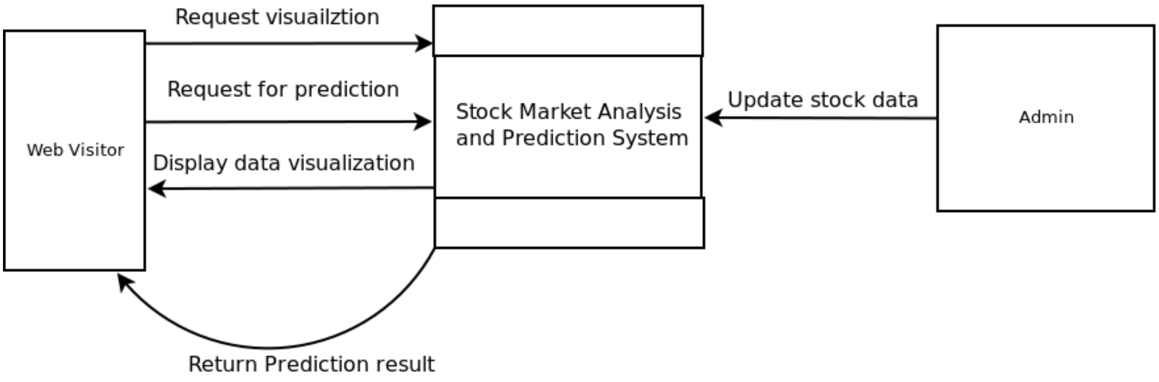


Figure 4.3: Level-1 dataflow diagram (DFD)

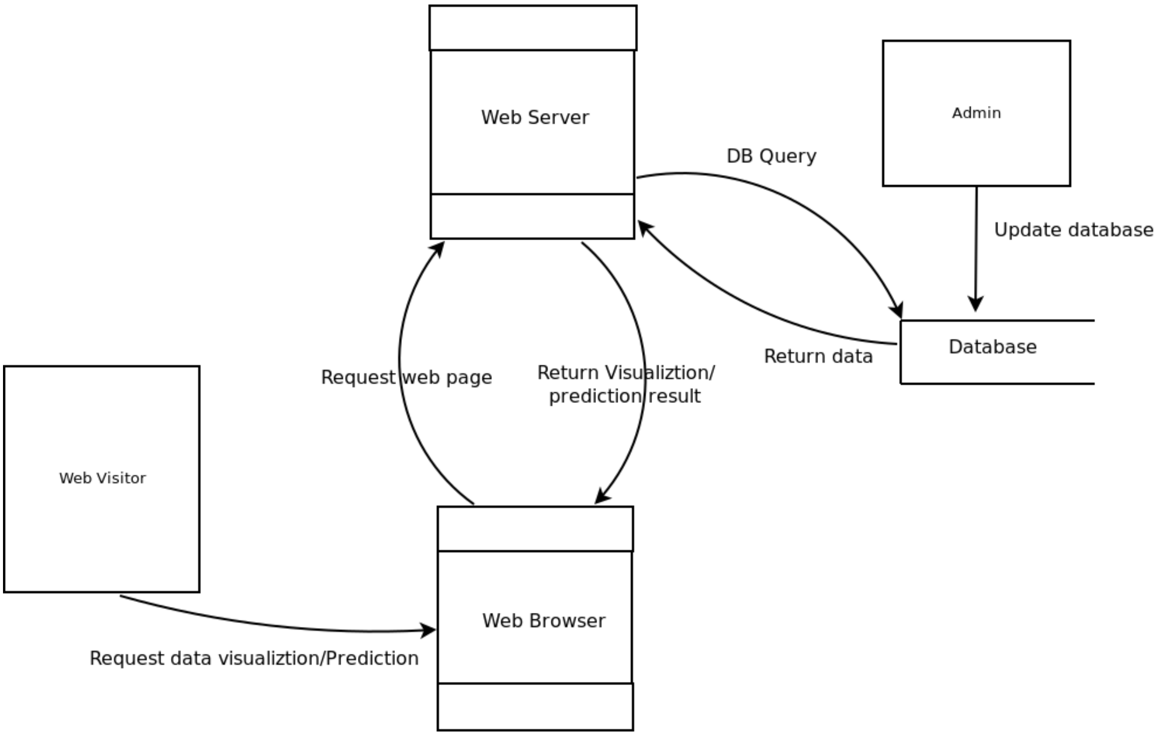


Figure 4.4: Level-2 dataflow diagram (DFD)

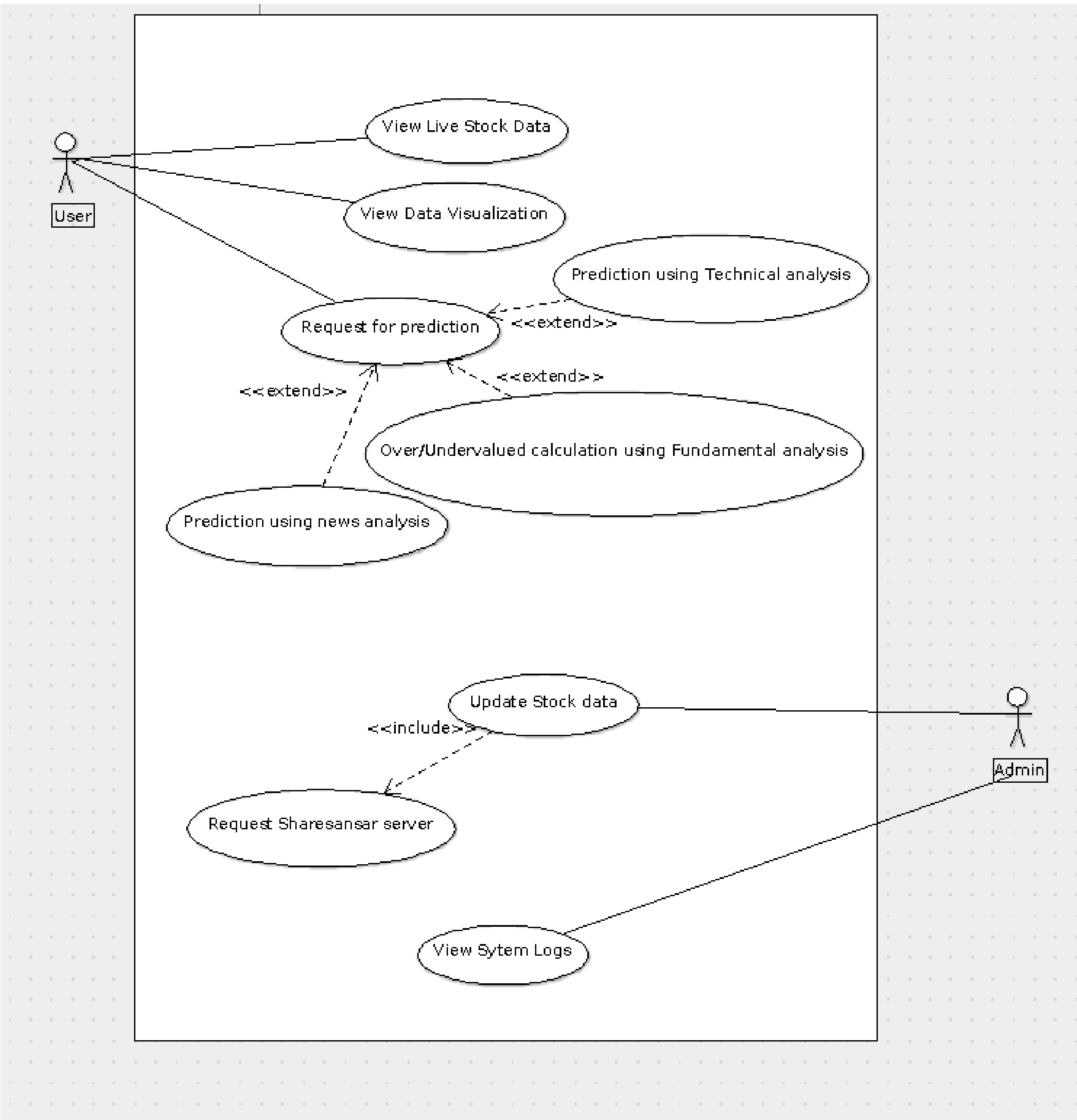


Figure 4.5: Use case diagram

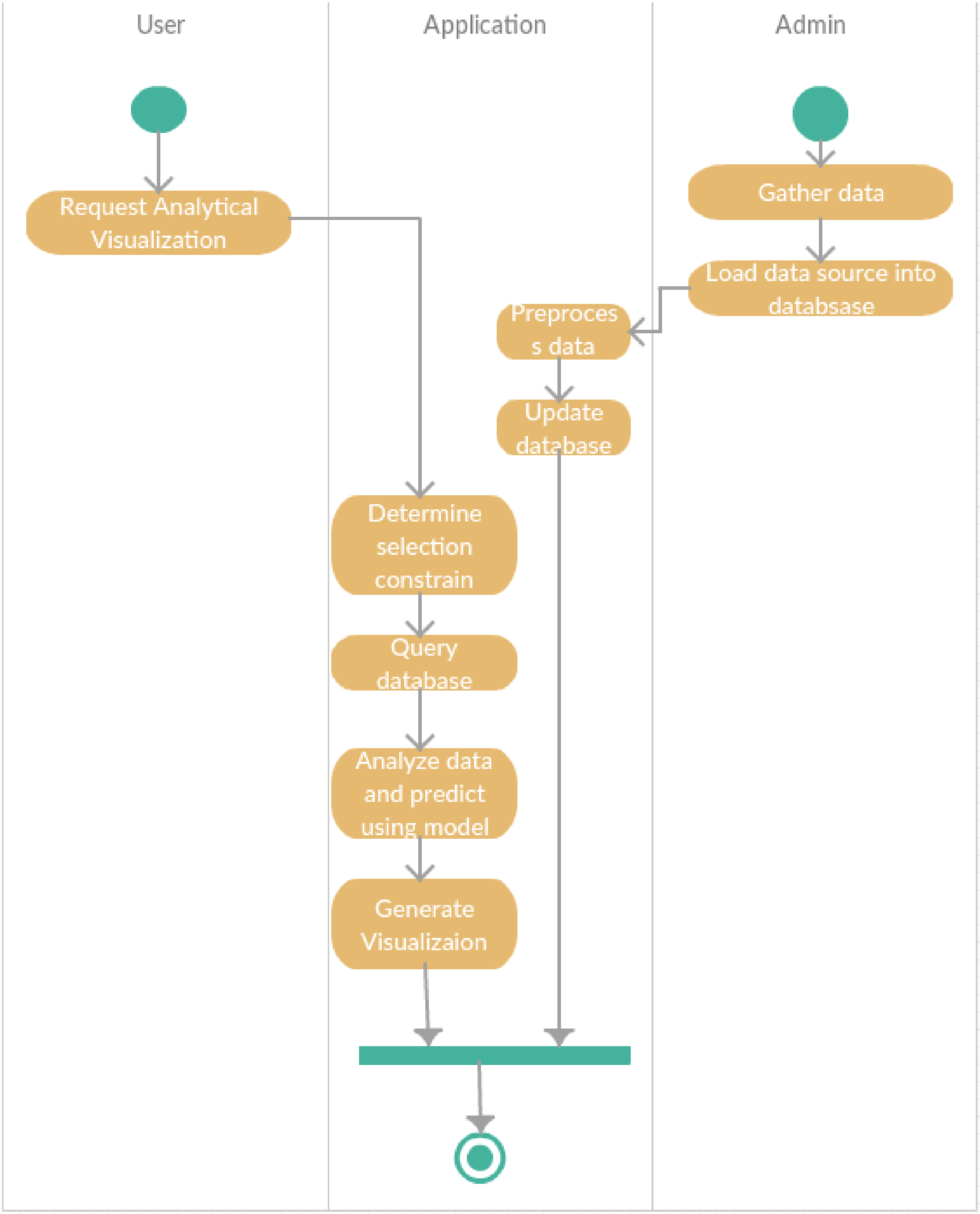

User
Application
Admin
Request Analytical Visualization
Gather data
Load data source into databsase
Preproces s data
Update database
Determine selection constrain
Query database
Analyze data and predict using model
Generate Visualizaion

Figure 4.6: Activity diagram

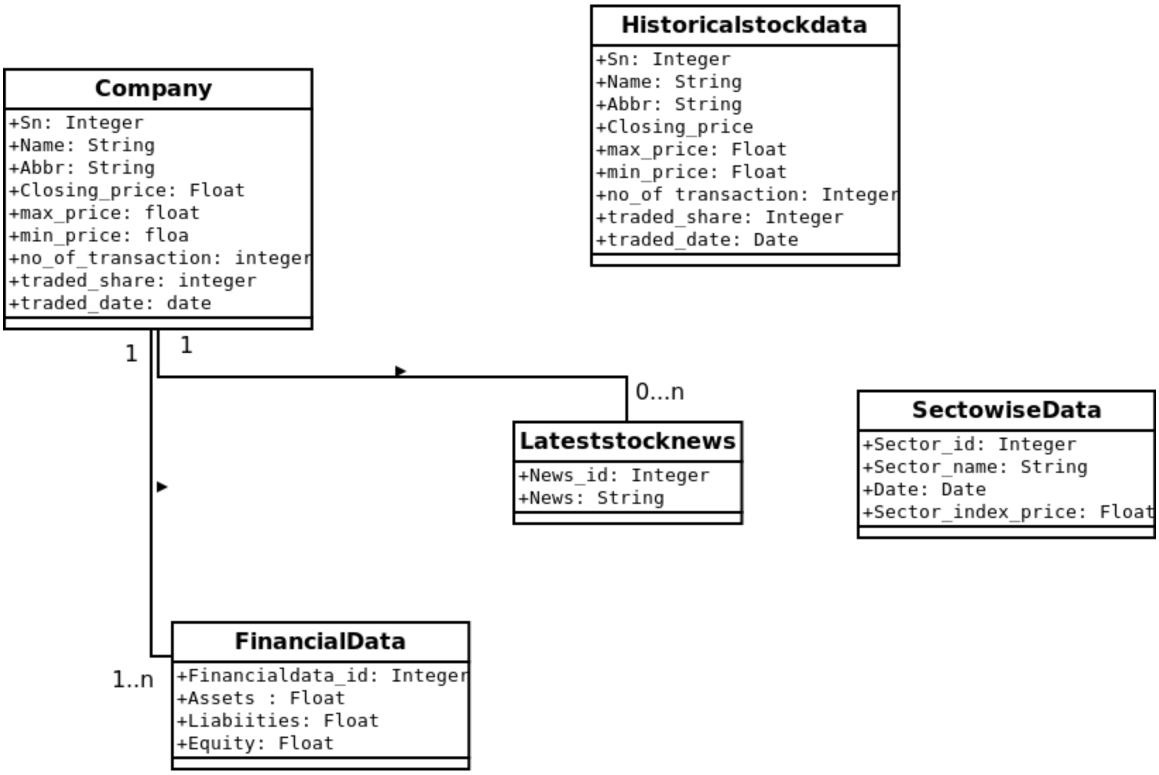


Figure 4.7: Class diagram

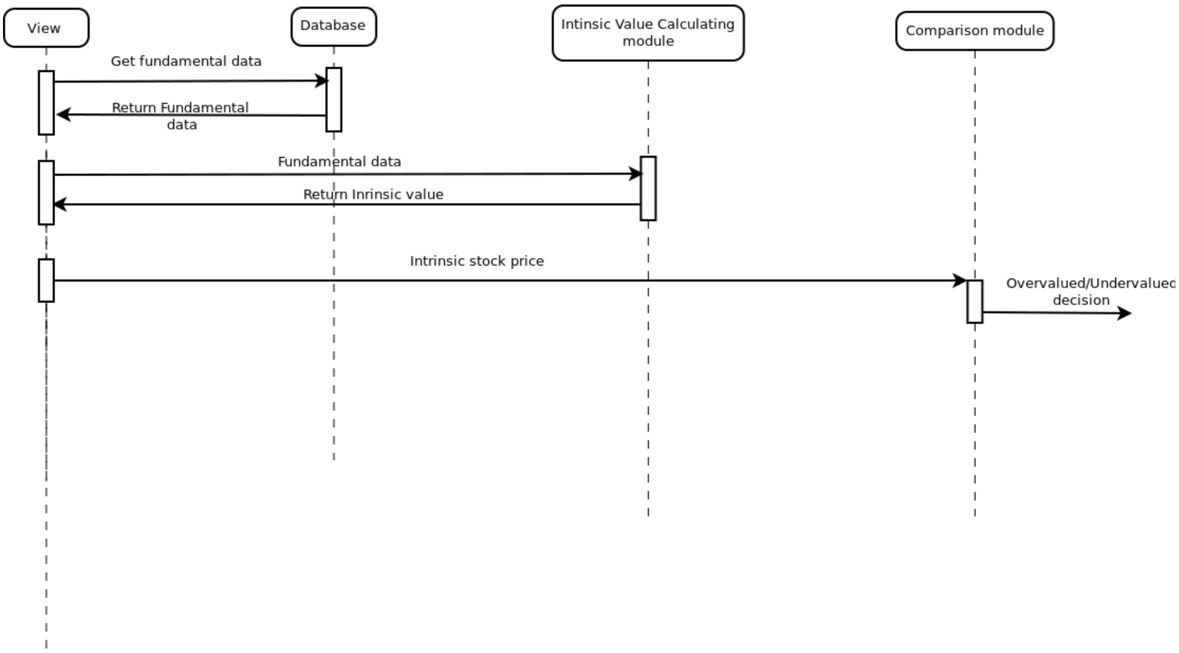


Figure 4.8: Sequence diagram for Fundamental Analysis

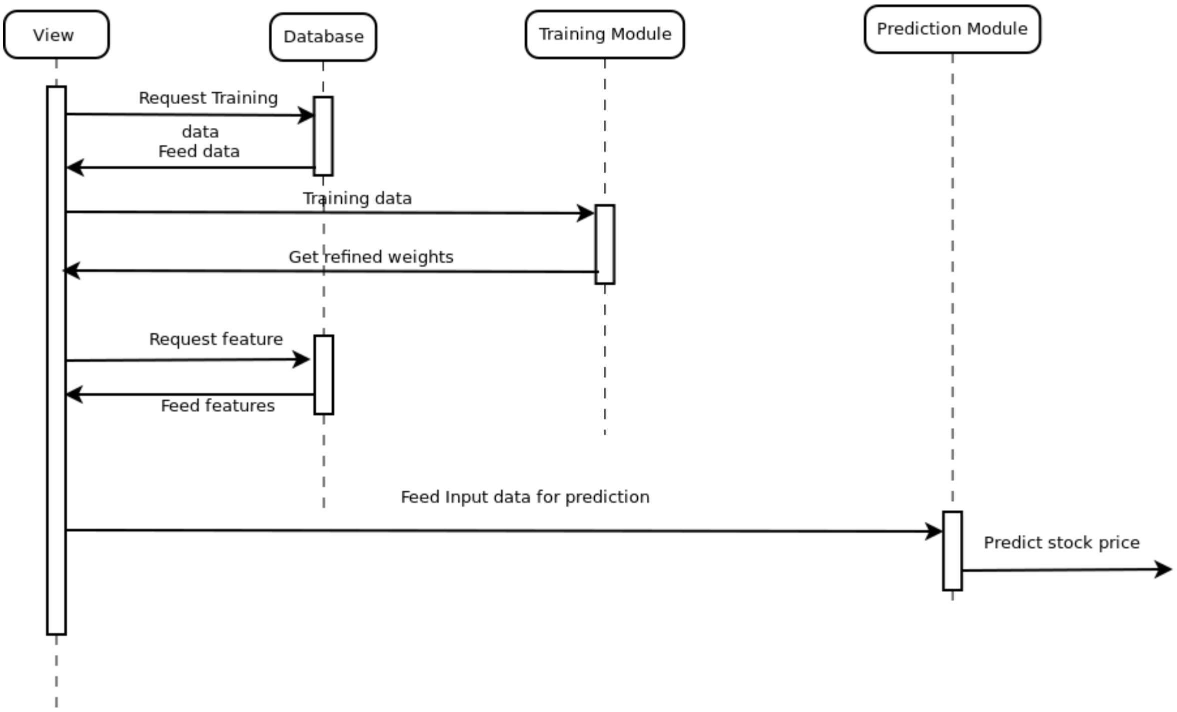


Figure 4.9: Sequence diagram For technical analysis using NN

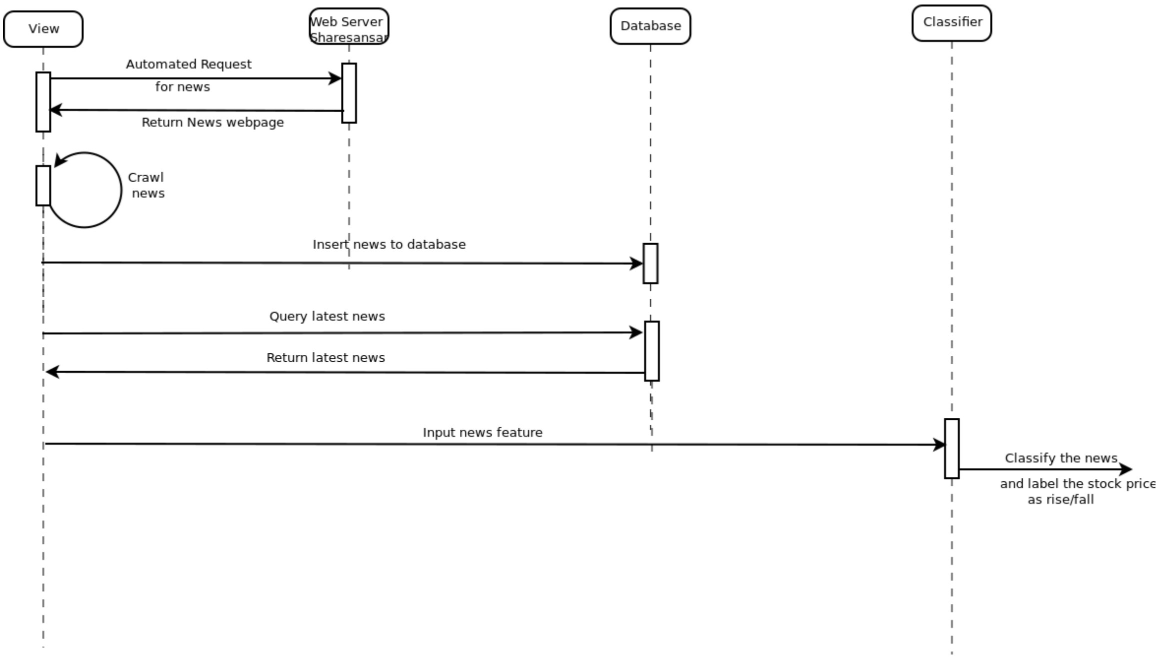


Figure 4.10: Sequence diagram for New Analysis

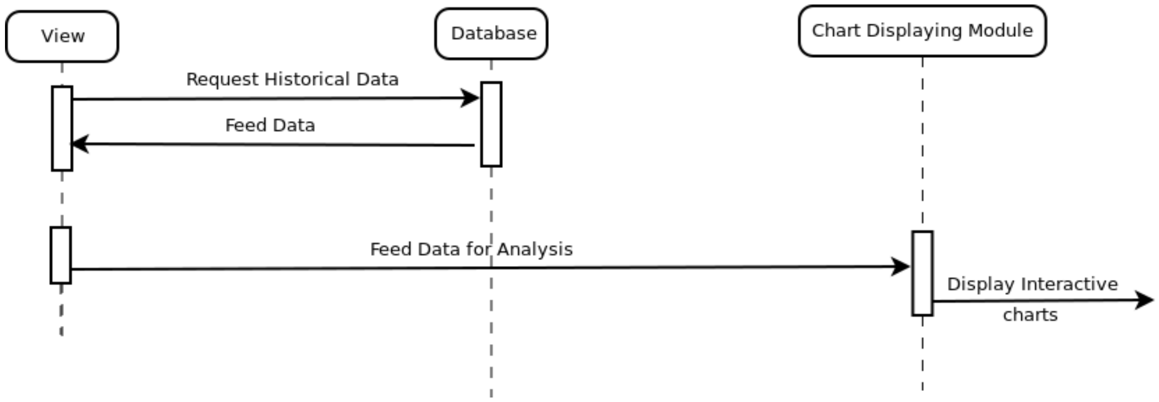


Figure 4.11: Sequence diagram for Data Visualization

4.5. Tools and Technique

The various tools and techniques used in this project are described below:

4.5.1. Python

The whole project is written in Python Programming Language. Various libraries of python are used in the project. Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985-1990. Like Perl, Python source code is also available under the GNU General Public License (GPL).

Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages. Python is processed at runtime by the interpreter. Python is Interactive. Users can actually sit at a Python prompt and interact with the interpreter directly to write programs. Python is Object-Oriented: Python supports Object-Oriented style or technique of programming that encapsulates code within objects. Python is a Beginner's Language: Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games. It supports functional and structured programming methods as well as OOP. It can be used as a scripting language or can be compiled to byte-code for building large applications. It provides very high-level dynamic data types and supports dynamic type checking. It supports automatic garbage collection. It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

4.5.2. Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. Pandas is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for multidimensional structured data sets. Python has long been great for data munging and preparation, but less so for data analysis and modeling. Pandas helps fill this gap, enabling users to carry out the entire data analysis workflow in Python without having to switch to a more domain specific language like R. Pandas modules uses objects to allow for data analysis at a fairly high performance rate in comparison to typical Python procedures. With it, users can easily read and write from and to CSV files, or even databases. From there, users can manipulate the data by columns, create new columns, and even base the new columns on other column data.

4.5.3. Scikit-learn

Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, naive bayes, gradient boosting, k-means and DBSCAN, and is designed to inter operate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn was initially developed by David Cournapeau as a Google summer of code project in 2007. Later Matthieu Brucher joined the project and started to use it as apart of his thesis work. In 2010 INRIA got involved and the first public release (v0.1 beta) was published in late January 2010. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. Some popular groups of models provided by scikit-learn include:

- Clustering: for grouping unlabeled data such as K-Means.
- Cross Validation: for estimating the performance of supervised models on unseen data.
- Dimensionality Reduction: for reducing the number of attributes in data for summarization, visualization and feature selection such as Principal component analysis.
- Feature extraction: for defining attributes in image and text data.
- Parameter Tuning: for getting the most out of supervised models.
- Supervised Models: a vast array not limited to generalized linear models, discriminate analysis, naive bayes, lazy methods, neural networks, support vector machines and decision trees.

4.5.4. NumPy

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases. NumPy is licensed under the BSD license, enabling reuse with few restrictions.

For scientific computing in Python, NumPy is the de facto standard. Therefore,heavy use of NumPy was used to implement our machine learning algorithms.

4.5.5. NLTK

NLTK (Natural Language Toolkit) is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project. NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

4.5.6. Django

Django is a free and open-source web framework, written in Python, "for perfectionists with deadline". It follows the model-view-controller (MVC) architectural pattern. It is maintained by the Django Software Foundation (DSF), an independent nonprofit organization. Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes re-usability and "pluggability" of components, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings files and data models. Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models. Some well-known sites that use Django include Pinterest, Instagram, Mozilla, The Washington Times, Disqus, The Public Broadcasting Service, BitBucket, and Nextdoor. Despite having its own nomenclature, such as naming the callable objects generating the HTTP responses "views", the core Django framework can be seen as an MVC architecture. It consists of an object-relational mapper (ORM) that mediates between data models (defined as Python classes) and a relational database ("Model"), a system for processing HTTP requests with a web templating system ("View"), and a regular-expression-based URL dispatcher ("Controller").

For developing a Django project, no special tools are necessary, since the source code can be edited with any conventional text editor. Nevertheless, editors specialized on computer programming can help increase the productivity of development, e.g., with features such as syntax highlighting. Since Django is written in Python, text editors which are aware of Python syntax are beneficial in this regard. Integrated development environments (IDE) add further functionality, such as debugging, refactoring, unit testing, etc. As with plain editors, IDEs with support for Python can be beneficial. Some IDEs that are specialized on Python additionally have integrated support for Django projects, so that using such an IDE when developing a Django project can help further increase productivity.

4.5.7. Git

Git was used as a version control system to collaborate among the team members. Git is a version control system that is used for software development and other version control tasks. As a distributed revision control system it is aimed at speed, data integrity, and support for distributed, non-linear workflows. Git was created by Linus Torvalds in 2005 for development of the Linux kernel, with other kernel developers contributing to its initial development. The Git feature that really makes it stand apart from nearly every other SCM out there is its branching model.

Git allows and encourages you to have multiple local branches that can be entirely independent of each other. The creation, merging, and deletion of those lines of development takes seconds.

4.5.8. PostgreSQL

PostgreSQL is a powerful, open source object-relational database system. It has more than 15 years of active development and a proven architecture that has earned it a strong reputation for reliability, data integrity, and correctness. PostgreSQL runs on all major operating systems, including Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64), and Windows. PostgreSQL (pronounced as post-gress-Q-L) is an open source relational database management system (RDBMS) developed by a worldwide team of volunteers. PostgreSQL is not controlled by any corporation or other private entity and the source code is available free of charge. It supports text, images, sounds, and video, and includes programming interfaces for C / C++, Java, Perl, Python, Ruby, Tcl and Open Database Connectivity (ODBC). PostgreSQL supports a large part of the SQL standard and offers many modern features like Complex SQL queries, SQL Sub-selects, Foreign keys, Trigger, Views, Transactions, Multiversion concurrency control (MVCC), Streaming Replication (as of 9.0), Hot Standby (as of 9.0).

4.5.9. AMCharts

AM Charts made it easy to display complex data visualizations. Combine various graph types on a single chart. Create clusters, or stacks, or clusters of stacks. Control the widths,open and close values, apply coloring based on value thresholds or changes, recalculate the values automatically. Use various value scales, including date and time. Those are just a few examples of what we can do. Some of the features of AM Charts are :

- **Interactive :** Zoom or pan serial charts, drill-down to other data levels, select slices, toggle graphs using legend, display HTML-rich contextual info, or draw trend lines directly on chart.
- **Responsive :** Resize your browser window, rotate the phone, watch the chart not just take the new shape, but adapt its contents and controls accommodate available space. Use full-fledged responsive features transparently, or write your own responsive rules.
- **Mobile-friendly :** It made extremely easy to control the charts using touch gestures. Zoom, pan, click the charts, without sacrificing the general responsiveness of the web page.
- **Dynamic :** Update data, size or just about any other configuration variable dynamically, without reloading the page. Add graphs, legends, titles, guides, bullets, or change colors, switch between 3D settings on the fly via well-documented API. Tap into chartś various events using custom handler functions.
- **Live-updated charts :** Update data every second to create ’live’ charts. Simulate just about any interaction using API function calls.

5. EXPERIMENTS AND RESULTS

5.1. Result from Prediction models

In this project, various algorithms were implemented to predict the stock market. Technical analysis was done for short term prediction of individual companies, using ANN classifier, ANN regression and KNN algorithms. Likewise, News-sentimental analysis was done for short term prediction of NEPSE index using Naive Bayes algorithm, whereas, under fundamental analysis, the intrinsic valuation of a company was calculated to determine whether the price of a stock is undervalued or overvalued. The algorithms used and their respective accuracies are summarized below:

Prediction Model	**Accuracy**
ANN classifier	72%
ANN Regression	92%
KNN classifier	65%
Naive Bayes Classifier	68%

Table 5.1: Result from prediction models

5.2. Result from Technical Analysis

5.2.1. Regression using ANN

In ANN regression, six factors were considered as input factors. They are: Open price, Close price, High price, Low price, Number of transactions, Traded volume. Comparison was done based on the number of hidden neurons. Comparison charts with actual and predicted values for NABIL bank, for 150 working days, using 20, 30 and 40 hidden neurons are shown below:

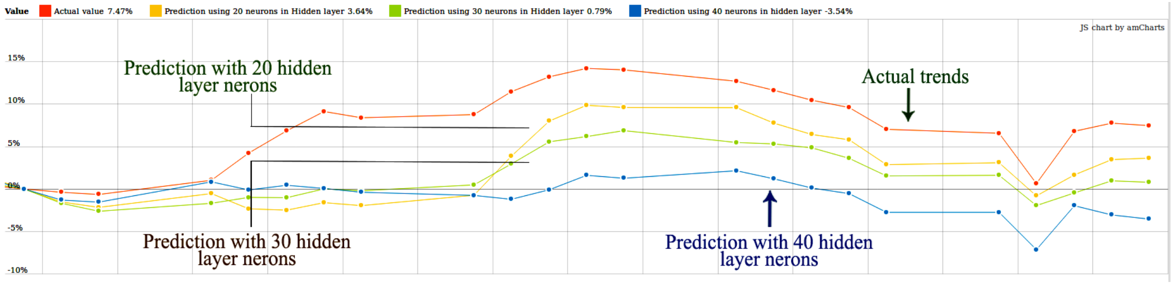


Figure 5.1: Result comparision for varying number of neurons in hidden layer

The prediction model with 30 hidden neurons shows better results.

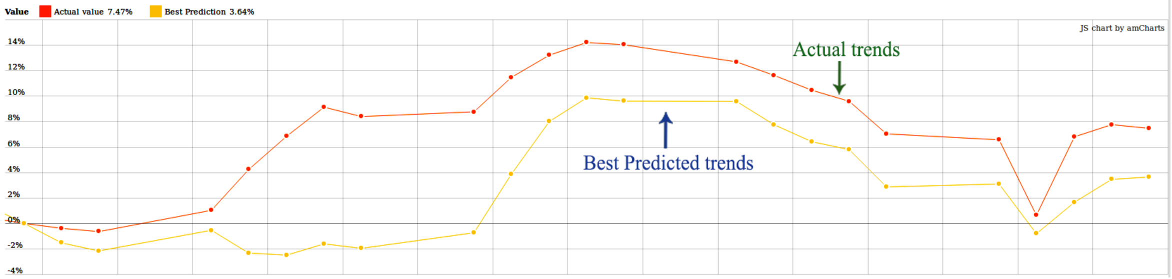


Figure 5.2: Result analysis showing 30 hidden layer neurons

5.2.2. Classification using ANN

In this model, the trend for the next day's stock price for a company is predicted. The chart below shows the actual trends and predicted trends for NABIL bank for 100 working days. The rise in waveform depicts the rise in stock price whereas the fall in waveform depicts the fall in stock price.

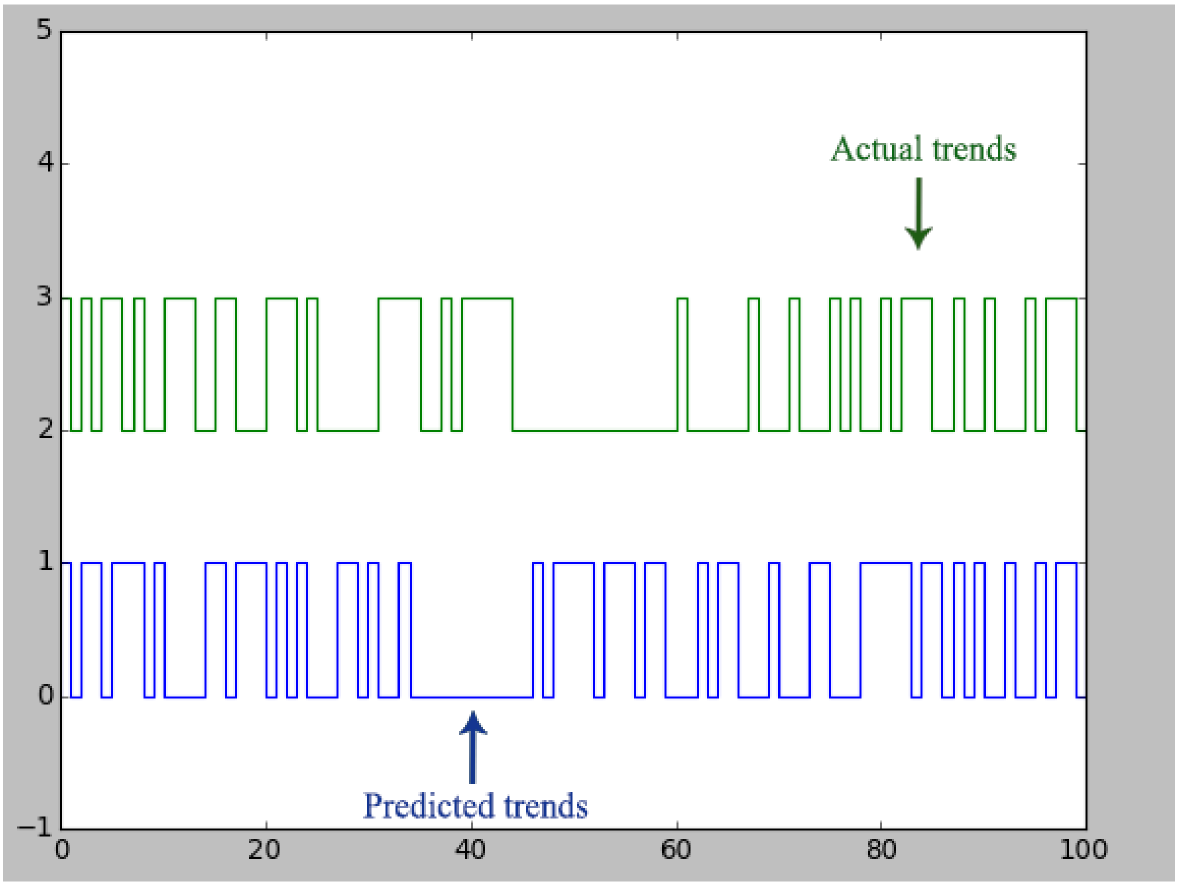


Figure 5.3: Result analysis of classification using ANN

5.2.3. Classification using KNN

In this model, the trend for the next day's stock price for company is predicted. The chart below shows the actual trends and the predicted trends for NABIL bank for 100 working days. The rise in waveform depicts the rise in stock price whereas the fall in waveform depicts the fall in stock price.

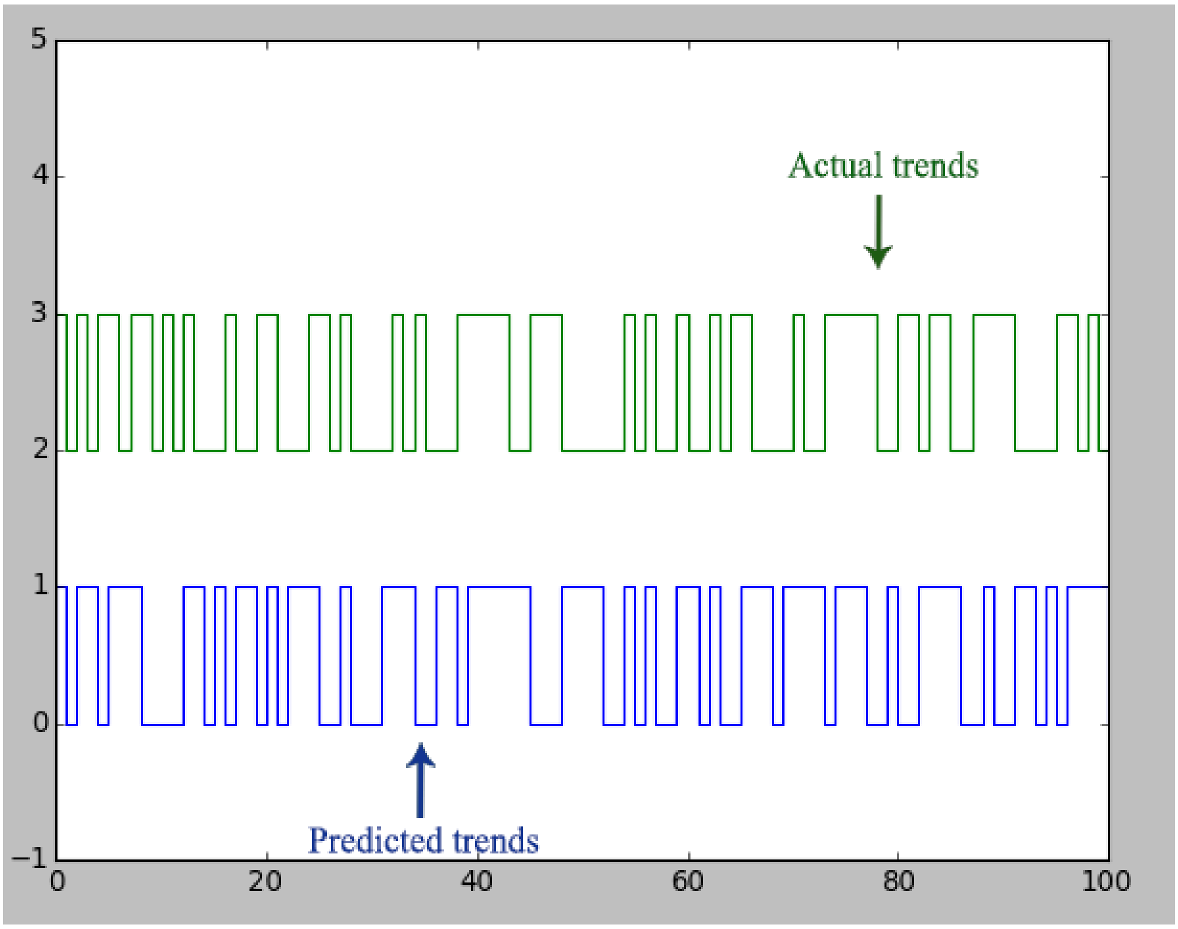


Figure 5.4: Result analysis of classification using KNN

5.3. Result from News Sentiment Analysis

5.3.1. Classification using News-sentiment

In this model, the trend for NEPSE index is predicted. The chart below shows the actual NEPSE trend and the predicted NEPSE trend taken over a period of 100 working days. The rise in waveform depicts the rise in index whereas the fall in waveform depicts the fall in index.

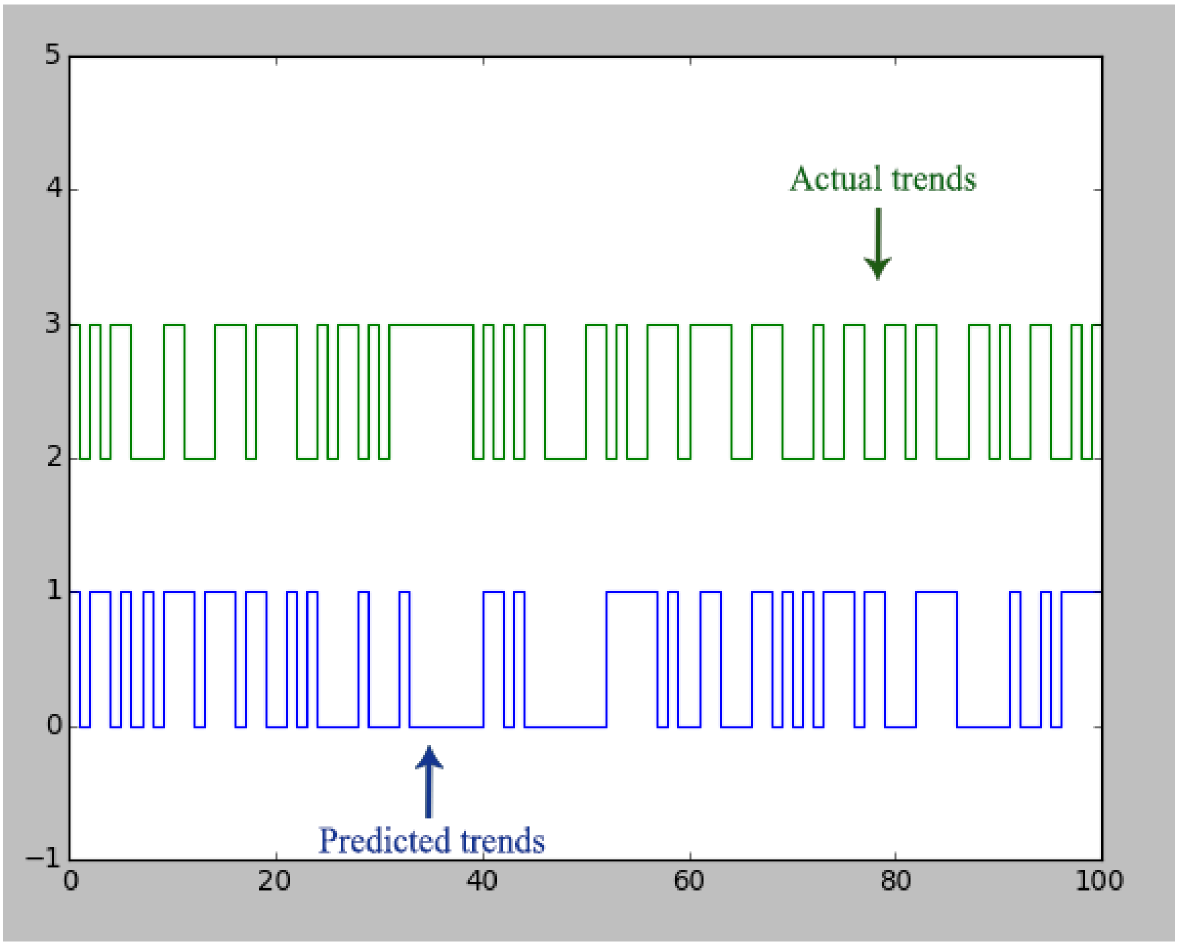


Figure 5.5: Result analysis of classification using news-sentiment

5.3.2. Result from Confusion Matrix

The table 5.2 shows confusion matrix for Naive Bayes algorithm implemented for news analysis. The matrix was created using the news collected over 60 days.

	Predicted High	**Predicted Low**
Actual High	True Positive = 27	False Negative = 9
Actual Low	False Positive = 10	True Negative = 14

Table 5.2: Confusion Matrix for Naive Bayes Classifier

In table 5.3, results obtained from the confusion matrix are analyzed using various terminologies.

Terminology	**Values**
Sensitivity(True positive rate)	0.7500
Specificity(True negative rate)	0.5833
Precision(Positive predictive value)	0.7297
Recall(Negative predictive value)	0.6086
Matthews correlation coefficient	0.3358
F1 score	0.7397
Bookmaker informedness	0.3333
Markedness	0.3383
Accuracy	68.3300%

Table 5.3: Result analysis of News Sentiment

6. LIMITATIONS

In this project, the prediction of stock market trends and stock market index is done using various artificial intelligence and data mining techniques. The maximum accuracy obtained falls within acceptable range, still, there are a few limitations of this project which are summarized below:

1. **Effects of stock dividend in future price of a company:** When a company announces a stock dividend, its stock price tends to decrease by certain amount right after the dividend interval is closed. That effect is not accounted in the prediction model.
2. **Effect of seasonal increase/decrease:** The stock prices tend to increase/decrease in certain months of the year. These seasonal characteristics are not incorporated in the prediction model.
3. **Sector wise analysis:** The companies trading in stock market belong to different sectors like: banks, finances, insurance companies etc. While predicting prices of a company, the values of the sector it belongs to also plays a vital role. Since, the sector-wise data are available only in small number, its effects could not be studied while building our prediction model.

7. CONCLUSION

In a nutshell, this project predicts the stock movements using various artificial intelligence and datamining techniques. In the project, the market is analyzed based on technical and fundamental factors to obtain short term and long term prediction, respectively. News-sentimental factors are analyzed to predict the NEPSE index trend. The stock trends are then shown graphically for visual representation. The predictions were made using various algorithms and models and the results were compared to each other. The project achieved 92% accuracy using artificial neural network considering five input factors- opening price, closing price, high price, number of traded shares, and traded volumes.

As future enhancement, the following points will be considered:

1. Incorporate the effects of seasonal increase/decrease of stock market
2. Incorporate the factors concerning sector-wise analysis in prediction model
3. Strengthen the prediction model to incorporate the effects of share dividend (like bonus, right share and cash dividend) on the future of the company stock.
4. Strengthen the fundamental analysis model to predict actual trends over specific and long period (e.g. 30% rise in stock price of XYZ company in 6 months is predicted)
5. Improve visualization systems to make it more interactive and easy to use.