# Quantitative Analysis of Equities using Machine Learning and Textual Analysis

**Aditya Bagla[1], Vaibhav Kudke[2], Sarthak Dongre[3], Surya Dhole[4], Tanuja Mulla[5]**

*[1-4]Dept. of Computer Engineering, JSPM's Narhe Technical Campus, Pune, Maharashtra. India*
*[5]Assistant Professor, Dept. of Computer Engineering, JSPM's Narhe Technical Campus, Pune, Maharashtra, India*

---

**ABSTRACT:** As the Stock Markets around the globe are constantly increasing in Volume, the participants are also increasing exponentially. This increase in market participants is a major factor which causes a lot of volatility in the prices of equities.

We are currently in the age where data is readily available, its use for the prediction of future stock prices and movement trends is becoming very popular. Thus we will be reviewing the most appropriate and efficient method to predict the stock movement with higher accuracy.

The stock price of a company is an important metric in finding out the growth or prosperity of the company, and many factors can affect these values. Various events taking place in a company and the news revolving around it can affect public sentiments and emotions regarding their investments in such company, which may have an effect on the trend of stock market prices. In time of market distress the reaction of stock prices to news is even more pronounced and the recent COVID-19 pandemic has proven that on multiple occasions. Performing successful stock market prediction is still a challenge. News articles are very useful and important in financial prediction, but currently no good method exists that can take these into consideration to provide better analysis of the financial market. Successfully predicting the future movement of the stock price will be a great tool for the investment institutions and will provide real-life solutions to the problems that stock investors face while taking decisions related to investments.

**MOTIVATION**

Time is money or even more valuable than money so instead of spending time on reading reports, charts and figuring out the future value of a company stock through the long process, we can let the highly developed and modern automated techniques in the field of Machine Learning, High Performance Computation, and Artificial Intelligence to do this task of Financial market analysis for us.

**Chapter 1 INTRODUCTION**

## 1.1 INFORMATION ON STOCK

We all have heard the word stock one way or the other. Particularly stock is related with the associates and companies which are commercialized and are to settling in the world of marketization. The other word used for stock is share which is prominently used in day to day life. People even term it as an investment plan and its something people see as a long term investment that secures and provides an abundant funds during the retirement age.

Buying a company stock is purchasing a small share of it. People invest on the same to get a long term benefit which they think is less value for now but has to potential to grow with the time. Its an investment that provides the long time run and deals with long time goals with the fair objectives. The value of share you invest today has to give you an yield of best tomorrow but its not the same.

Market is unpredictable so are the resources and the factors that are taken to drive it off or on the set. Its never been on the same level and the pattern of the same is still unpredictable till the time. Some closeness and prediction method had been derived and approximates values and the rough figures are generated hoping for the best but all of the resource can't be trusted and are still unpredictable in nature.

Knowing the market situation and researching on the same is the best way to find the reliability for which there are many agents who have taken the same as a profession and are making a fortune out of it. They predict and advise but the advisory cost and the charge is higher and the stock evaluation is never less the same.

Market is changing in an instantaneous rate even in a day there are many highs and lows in the market and having said the resources and the timing the external and internal agent. Stock is a fascinating resource to start with.

Stock in other term is defined as the fair share or ownership representation explaining the security measures and the agreement between two parties which are an individual and the company. Stock is there from the start and due to its tendency of uncertainty it has been a word of fancy. People researching on the same and implementing on the daily basis had made a fortune out of it. There are various agents available in market for making you understand and invest on the same and the charges of the same are hectic and insanely expensive.

The main resources for the company are the fund to carry out the daily work and create a profit out of it. In time of need for an higher budget estimation and to overgrow from the resources they need the finance and undergoing a finance loan for approval, passing and having one is hectic and the banks are vultures for which the interest rate is higher than the other form of investment hence limiting the margin of the product.

Stock is another way for company to collect revenue and boost up the production for the upper yield and to gain the most out of the business plan for the bigger pictures. This is found to be an effective way to invest and grow in the commercial field and a better alternative to tackle the financial crisis during the requirement.

For an investor its a risk phenomenon where they invest their saving and hope it brings back the return in higher yield. If the evaluation of the same increases then the stock evaluation and its price increases causing the financial gain to both the parties. In Indian Society it is even consider as a side point business and people believe it as a hand of luck.

When an individual purchases a company stock then they're referred as a shareholder and they will get a share out of the same as they have invested in their profit or the gain. A investor can sell and buy the stock as per their needs. They can share their stock to their respective or the other individuals where as there are many stock brokers available out in the firm playing with the same.

### 1.2 PROBLEM DEFINITION

Stock is an unpredictable curve that had been in picture ever since. Its essence had been ever long living and indulging. It had grown its popularity with respect to time. People are more fascinating and interested on the same then before times. Same for the case for the organization. Organization had created it as a better source of revenue generation rather than investing and taking a loan approval from the bank It's way efficient and less hectic from the firm point of view.

Stock is unpredictable and its been the same from the start. Its way of escalating and deescalating had been phenomenon and experiencing the same is the best integral part of it. It has its upper hand and flexibility with the changes that has the chances of uprising as well as crashing the whole market. Its easily defined in few words but making an essence and understanding the same is way more hectic and time consuming.

Simpler it sound complex are its phenomenon and integrating the same. It has its whole different sets of dependencies and integration from different agents which fluctuate the same in the market. Finding an accurate and getting the exact values out of the same is still unaligned and no particular model of the same is seen in the market value.

Finding the closest and getting an accurate proximate value out of such an unpredictability is a problem in itself. Merging of the data getting the best prediction to increase the efficiency alongside considering the different expects of the moderator is tough and we took the same in consideration and implemented with every aspect to generate the best out of the same and get a result that can be better interrupted and the efficiency remains the same with the value of different aspects of creating an impact of reducing the risk and influencing the same over the time period to gain the most out of it.

This is totally based on Machine Learning Algorithm to proceed and provide an effective result. Getting the data, processing it and generating a forecast is the problem statement here.

### 1.3 PROJECT PURPOSE

Stock market prediction is a prediction system software that illuminate the risk that undergoes during the investment in stock market. It predicts the stock rates and its rate of exchange acknowledging the basic understanding and the statistical analysis in front of users.

Data is considered as the digital fuel that gives the possibilities of higher yearn and gives the upcoming terms. Knowledge is power and same holds correct with the stock. Stock is unpredictable and over-changing its dynamic in nature. The rise and fall of the same is uneven and can't be classified so easily. Dependencies of the same deals with flexible resources and the agents behind it.

Investment during a fiscal day determines the opening stock market for the next day. It has its dependencies and is total integration with the level of finances and revenue generation. The stock is tremendous and hectic in nature. The main theme of the project is to predict the turning curves and bring the predictability method and undergo the process and algorithms to conclude to a viable resource source.

Everything flows in a pattern. Pattern is the way of derivation and so holds true for the stock too. Stock in day to day life follows a pattern movement. Increase in some resource can increase the price of some whereas decrease the price rate for the others, The source and the outcome are derived on the polarity basis which can either be positive, neutral or an negative flow. Correlation of the given polarity is determined and an effective source and reliability is established.

This project helps in bridging the resources and empowering the people to know and trade the most out

of stock and understand the generation and the vulnerabilities that has to be seen and predicted. The enhancement of the same is done with the resource graph which makes a user to analyses the same and take the needs and important details before dealing and consider those things for the yield that the person is willing to invest on. Forecasting of the stock prediction is done by the available data source and the prediction is done for the upcoming period. The predictability itself is a challenge and that's the main purpose of the report.

### 1.4 PROJECT FEATURES

Features deals with the flexibilities and the top marks that one can present. The project was headed with the resource available and the most that the market demands and that is finance. Talking about finance and learning on the same gave an idea on the fiscal and stocks. So the featuring of the idea came with handling and automating the resource which other agents are making fortune out of it.

Knowledge is a bliss and learning is the curiosity whereas outcome is the expectation so the resource deals with the importation and extraction of multiple machine learning algorithms to learn, process and yield the result to derive and conclude a possible outcome set that is effective and generative in nature.

There are various models that outflows in market which are trying their best on creating a resource and give the predictability to most of it accurate but everything is not the same and the conclusion of the same are not ideal. The efficiency varies as the variation in the stock market and its prediction.

The project is purposed with the sole intent to make and undergo the following way of computing. The first deals with the data extraction that is done with clearing of data and its chunks from the database or the dataset. The second flow is the training from the source training is done and classified. During the same supervision is done and the last part is the generation of the yield which provides the result after computation of the same.

Salient features included are the Visualization and the prediction that gives a boost. Uses of different forecasting algorithm to forecast that holds true and are suffice in nature to yield to the positive resource source. Diving and initializing the expects that needs to be considered. Mitigating the risk factors to bridge and uplift the investment.

Analyzing and utilizing the same to support the live environment. Keep a track of progressive result and it's evaluation on day to day basis to find the flows and the level of integration. Automating for the ideas and making it most by using feasible algorithms which can undergo learning and implement the updates in itself to summon the efforts that one needs to take for the best.

### 1.5 MODULES DESCRIPTION

#### 1.5.1 DATA SET

This is the fundamental module before starting of the project. The dataset is a group of data that are mended together to show the data variations in a time span to undergo further estimation and the source of the resources and its outcome for the later time of evaluation. It generates the result optimization and gives a feasible time period to customize and get the flow to the derivation.

This increases and are used in the level of research and finding the best suitable resource out of the same the resources has to be finely estimated and derived for the best possible outcome and the finest the value become the better is the level of extraction and closure is the best yield values that needs to be considered.

#### 1.5.2 DATA ABSTRACTION

Abstraction is the finding of the resource to its best to categorized the above dataset and learning the best out of it. Abstraction of the data is the integral part to the flow. All the data are a huge set of chunks which on processing can limit the yield result and the computational mean too. Thus with the available resources the data yield had to be derivative.

Abstraction of the dataset is to customize the data set and finding the best suitable constraints to take into consideration and the unwanted resources are the dump which will be dumped and the supreme cluster is created with the valuable constrains and a pattern is needed to be derived from the same.

Data are cleared on this level for the beginning of the process. The valuable data are the set that brings the value to the data set for a better understanding and giving a better yield and production by evaluating the same.

This is a feature abstraction module to extract the featuring of the dataset. This is a feature model process where all the feasible resources are categorized and the same will be in use for the featuring.

#### 1.5.3 TRAINING DATASET

After the abstraction of the data and clustering of the same. The machine had to be trained for which the training data plays the

important role. There are thousands of machine learning algorithms that are into place and evolving with the same. The best to the practice of machine learning is to yield the result and the content to derive what's needed with the time frame.

This is a supervised learning form where the input are passed so that the system learns from the same. Various variants of inputs are passed which were stored in the dataset. Every resource is considered and taken into consideration. After considering the whole set of information and the resource the machine tries to learn from the passed dataset. The dataset has to be wide and versatile. After considering the learning it tries to integrate with the same type and flow like the same as the human mind and creates a pattern and the links between the same.

1.5.4 TEST DATASET

These are the sets of data that gives the result after learning from the data. This is the test generation with the output result. Results are generated in each phase of testing. This is also termed as the testing phase. Now a new set of datasets are passed which are deliberately like the training dataset and the efficiency of the same is calculated.

Over-Fitting of the dataset. Validation of the same with the effective constraints and hyper parameters are checked. This phase is training and the output is evaluated with the set of training. After each process of computation the set of data are trained and efficiency of the same is measured and is evaluated with the others.

Various batches of the test is implemented to get to the level of accuracy and derive result to fetch and yield for the best performance and to be true to the effectiveness of the data which is not biased with any constrains available. This determines the efficiency of the system which is must for the predictions.

1.5.5 RESULT EVALUATION

This is the main part for any implementation of the project. Evaluation of the key point to the success. All the categorization of the work and the best to know the resource fundaments and again establishing the same to check the validity and the work flow and check on the output is must. The evaluation, utilization and implementation undergoes various level of extraction and evaluation.

The main theme is to provide and come up with the output with an accuracy that can be used and implemented. From the starting to the final the process is categorized, supervised and efficiency is check and the working is undergone. Testing is done and its evaluation are mended.

The process undergoes the same for various time and phase. Testing of the same undergoes sequential iteration for many more to meet up to the constituency. The remarks are to be noted and further work is done on the same with the implementation of the different aligned resources that are integrated with the available resources and its outcome.

After the evaluation and customization of the same the result is to be potted in a visible form and the best form of visibility is the graph. The Graph visualization is the best way of visualization that keeps the audience engaged for a long time. Derivation of the outcome is easily accessible and interpreted and the flow diagram is shown with the stock prediction that gives an upper hold to the appearance and shows the best level of the content.

After establishing a graph connectivity the customer or the user takes time to process the data and take that picture into consideration and can avail for the upcoming stock by investing in the same.

## Chapter 2 LITERATURE SURVEY

One of the integral parts to maintain the consistency is the literature survey. It's the crucial steps to be followed in the development process. The Software Development needs authenticity of the resources and the availability of the same. This part helps in discovering the content that been worked on and find the utilization and the implementation of the same in today's time. The key factor to the development is the economy and the strength of the product. Once the innovation of the same undergoes through the building phase the support and the resource flow is to be monitored and computed. This is also known as the Research phase where all the research is embedded and done to carry the flow.

### 2.1 MACHINE LEARNING

One of the finest word heard in today time is Machine Learning. Either it be at work or different places the machine learning has been an integral part of today's technology. Though its evolving and developing in a rapid rate and development and deployment of the same is still in progress. The machine learning itself had brought a random changes in today worlds because of which automation is in frame which was a mere existence in the

past.

It's an aspiring term in today's time. One of the moves that all the firm are interested into. It's a leading pillar for tomorrow leading the world to a better future of evolution where the customization and labor work can be reduce to half and the safety of the survival can be withheld to stand tall for the better utilization of human mind. Keeping that in picture it's been a hazard to many more in terms of irrespective field of interest. Since Machine is considered most efficient and the level of mistakes are kept at the minimum the level of work flow can be a work of hazard and further improvement on the same may create thousands sitting idle in home creating a larger impact on unemployment and livelihood. Which in other way is a threat to the society too.

ML is the abbreviation for Machine Learning. In other word it is making a human mind fitting inside a machine which uses the same to perform the task of thousands. Machine Learning deals with the higher aspects of learning techniques and algorithm which are highly aligned to make the work flow seamlessly effortless with the human tendency of doing work.

Algorithm of such are improvising in nature which learns by themselves and fit themselves in the world of impairment by getting the required data and adjusting with the same giving the effective results out of the same. ML is a subsidiary or the subset of an AI(Artificial Intelligence). It is a mathematical model where computation of the testcases plays the major role in driving of the results.

A wide level of machine learning architectures are implemented today to turn on the yield factor and make people life more efficient in terms of livelihood. Various use of such in Message Filtering like spams, Trash automation are automated and carried out by the same. Since the efficiency is way more than a human tendency. Multi-tasking and processing is also initiated by the same giving a dual output which a human can never ever possibly be able to.

Statistics is the major key role in driving the machine learning in figure. It deals with computation of statistics in a wide range view and processing the same to give a data driven output causing it to be more sensible. Not only to the same it optimizes the resources and the efficiency is unbitable and reliable in terms of any means.

Though its being evolved in nature but it has integrated itself well with the terms of computational and digitalization. Various computational fields like Data Mining, Statistical Analysis, Optimization of resources, Automation are a major part of it. Here the machine has the capacity to process the result on its own as same as the human bring. This process can be initiator as well as the derivable. The statistical flow is mainly reasonable with data driven pattern even the unstructured or the semi-structured data can be processed and approximate answer to the same can be derived. All the equations

are derived and the closest value to it's aligned field is found and the proximity is determined.

The various machine learning tools involved are as follows ;

### 2.1.1 SUPERVISED LEARNING

Supervised Learning deals with the supervision of the machine to derive the necessary input required. It's a mathematical model where the inputs and output of the same is already known and is passed to the machine to get expected output so that the efficiency is determined and this is the learning phase for the machine. Here the feeding and derivation of the same is measured.

Here the machines filters the inputs learns from the functional unit. Compute it and stores it into its memory for further process and if found a matching pattern it

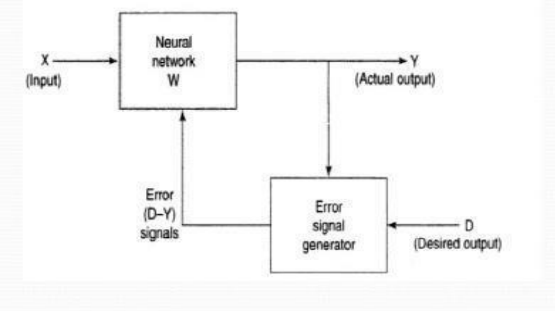



Fig 2.1: Supervised Learning

uses the same and learns from it and plot a result out of the same.

This is a dependent process. The machine totally depends on the user who has to feed the inputs and has to check the efficiency of the same and correct it with the flow of iteration. It's an ANN network. During the training phase vectors are taken into consideration.

Up in the above figure There's an input vector and the output vector. The input vector derives and gives an output flow of the output vector. If the error signal is generated then the iteration is undergone where as lacking of the same means the output field is derived and the output result is accurate and no modification needs to be undergone for same.

### 2.1.2 UNSUPERVISED LEARNING

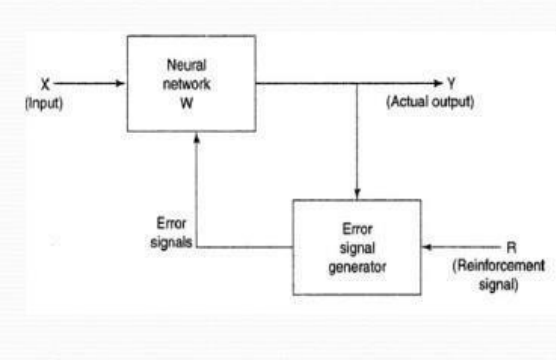
Unsupervised learning deals with learning by itself. It is also known as self-learning algorithm. Here only the input vector is known and passed. So the variance of the result deals with the input factors. Here the input factors are grouped and clustered. Cluster is the

main essence of this technique.

Test Data are passed and with the iteration of the same it learns from it derives itself more closer to the conclusion part. Labelled is missed in the data set and classification and categorization of the same had to be done my the machine itself. Cluster and Communalization is the main essence of it.

Fig 2.2: Unsupervised Learning

As described in the figure 2.2, In this ANN network when the input is processed by the function the output had to be self derived and to be matched with the cluster set to provide the result. If the result lacks the interpretation then it undergoes the iteration. All the data




sets are formed and combined in a cluster set for the effective uses of the same in further cases.

Feedbacks are not reciprocated in case of such it responds to commonalities. If the commonalities are found between the dataset then it applies the previous functionalities and derive the data. If not set then it learns and identifies for the others.

### 2.1.3 REINFORCEMENT LEARNING

In this type of learning a reinforced strategy is used. Its deals with blooming of the knowledge. It's neither Supervised nor Unsupervised form of learning. They use dynamic techniques for letting the user know the output and the derivation of the same.

In these sort of algorithm sets they don't assume the environmental set. These are even used in higher and complex mechanism finding likes genetic algorithm. They are widely in progress and implemented most in automation for the better efficiency of the establishment.

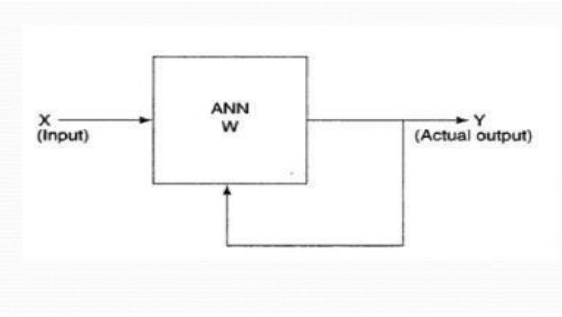



Fig 2.3: Reinforcement Learning

These algorithms are used in Games and Automation of the vehicle resources.

As described in the figure 2.3 the input vector is passed to a ANN model where the functionalities of the same are stored. If the accurate output is derived then a reward is given to the user making it go to the next level for further task of completion. If not then the Error signal is generated for the same. The accuracy level is calculated and passed down to the user stating the same.

The user sees the percentage of match and pass down and tries other keys of iteration to get the most out of it and complete the task to carry on the ladder of success. This is the same with the machine. Machine iterates the same and to the error signal an add on of reinforced signal is passed which the machine learn and iterates on the same to get closer to the actual results.

### 2.1.4 NATURAL LANGUAGE TOOLKIT

After the data collection, all articles are tokenized. This is done using the Natural Language Toolkit (NLTK), one of the leading platforms for building Python programs to work with human language data. Figure 2.4 displays the preprocessing. In general, tokenization means dividing a big quantity of text into smaller parts called tokens. Machine learning models need numeric data to be trained and make a prediction. Word tokenization becomes a crucial part of the text (string) to numeric data conversion. Following the tokenization, we can compute the "compounded sentiment score" using the BERT-based Financial Sentiment Index (Hiew et al., 2019). This textual-based sentiment index relies on BERT (Bidirectional Encoder Representations from Transformers) originally developed by Google and Devlin et al. (2018).

BERT is an open-source model that was pre-trained with millions of words from the entire Wikipedia corpus, employing a bidirectional Transformer encoder to predict masked words. The model performs two tasks. First, BERT randomly masks a fraction of words and predicts the words that have been masked-out.

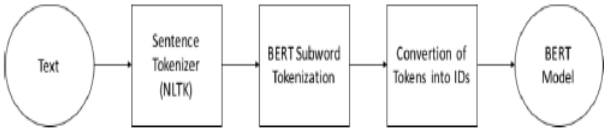



Fig 2.4: The preprocessing steps using the Natural Language Toolkit (NLTK)

Due to this pre-training approach, BERT has outperformed various state-of-the art NLP techniques. For instance, it achieves an accuracy of 94.9 percent in the completion of the Stanford Sentiment Treebank. Hiew et al. (2019) apply BERT for the stock market prediction and we use this pre-trained model for the sentiment analysis.

All tokenized sentences are classified as positive, negative or neutral along with the respective probabilities (logits). The sentiment is calculated as follows:

$$\text{Sentiment} = \text{Logit}_{\text{Positive}} - \text{Logit}_{\text{Negative}}$$

The Sentiment is therefore the probability that the sentence is positive minus the probability that the sentence is negative.

#### 2.1.5 RANDOM FOREST ALGORITHM

Random forest algorithm is being used for the stock market prediction. Since it has been termed as one of the easiest to use and flexible machine learning algorithm, it gives good accuracy in the prediction. This is usually used in the classification tasks. Because of the high volatility in the stock market, the task of predicting is quite challenging. In stock market prediction we use random forest classifier which has the same hyper-parameters as of a decision tree. The decision tool has a model similar to that of a tree. It takes the decision based on possible consequences, which includes variables like event outcome, resource cost, and utility. The random forest algorithm represents an algorithm where it randomly selects different observations and features to build several decision trees and then takes the aggregate of the several decision trees outcomes. The data is split into partitions based on the questions on a label or an attribute. The data set we use is historical stock price data, 80 % of data is used to train the machine and the rest 20 % to test the data. The basic approach of the supervised learning model is to learn the patterns and relationships in the data from the training set and then reproduce them for the test data.

### 2.2 TECHNICAL SURVEY

#### 2.2.1 SURVEY – I

Using Neural Network as a case learning based method, the results of the research work demonstrated that the news sentiments relevant to stock market can be used to improve the performance of the learning based prediction. How many days the effects of the news will last is also an interesting thing, which is worthy of further research.

#### 2.2.2 SURVEY – II

Among the most suitable algorithms for predicting the market price of stocks based on various data points from the historical data is the 'Random Forest Algorithm' along with 'Support Vector Machine'. Random Forest is one of the easiest to use and flexible machine learning algorithm, having good accuracy in prediction models. This algorithm is used in classification tasks usually.

#### 2.2.3 SURVEY – III

To evaluate the effectiveness of the model a comparison is made between the two techniques on five different sector companies using both Artificial Neural Network and Random Forest Model to predict closing prices. These predicted closing prices are subjected to Root Mean Square Error(RMSE), Mean Absolute Percentage Error(MAPE), and Mean Bias Error (MBE) for finding the final minimalized errors in the predicted price. The comparative analysis based on RMSE, MAPE, MBE values clearly indicate that Artificial Neural Network gives better prediction compared to Random Forest.

#### 2.2.4 SURVEY – IV

Neural Networks in finance are highly useful tools to analyze financial markets trends based on preprocessing and transforming a large amount of information into machine readable data And its main advantage is its wide variety of applications and the ability to process vast amount of information simultaneously, ignoring inclinations and biases to particular school of thoughts.

#### 2.2.5 SURVEY - V

The application of long short term memory networks and Random Forest model to forecast directional movements of stock prices is researched. Siami-Namini and Namin (2018) compare LSTM with an Autoregressive Integrated Moving Average (ARIMA) model. Multi feature setting consisting not only of the returns with respect to the closing prices, but also with respect to opening prices and intraday returns , outperform the single feature of Krauss et al. (2017) and Fischer and Krauss (2018), both with respect to Random Forest and LSTM.

### 2.3 EXISTING SYSTEM

As many have invested their time and effort in this world trade for getting it closer and more reliable to the

people for carrying out the resources and make their lifestyle more deliberate than the previous. In the past few years various strategies and the plans had been derived and deployed ever since it's continuation and the topic is still a point of research where people are coming up with ideas to solve.

Intelligence fascinates mankind and having one in machine and integrating on the same is the hot key of research. There are various people contributing on the same research.

All the learning system from the past are limited and are simplest in nature where learning of the simple algorithm for a computational mean is not enough which can even be done by human brain itself. The main motto of learning was limited and learning model was not efficient.

The existing models can't cope up with the vulnerabilities and remove the rarest information that they can't process causing it a major data loss which creates a problem in forecasting.

Observation is the integral part in the resource and prediction management. If the outcome can't be observed it's point of time estimation is compromised causing it less liable in market. Monitoring of the same is not possible in the existing system.

The existing system in stock market predictions are apparently biased because it considers only a source point for data source. Before the prediction of the data set a simple data retrieval should be generated and tested on the training data set which are more flexible and versatile in nature.

Loss of sights is a major problem in the existing system as the stock varies each days and the loss margin can be higher with respect to time. An initial instance is taken for prediction.

News articles are a huge factor affecting the prices of stocks on a scale of short or even long term period, and majority of the models in the market existing currently ignore this factor and just consider the historical price data as input, and process on the same.

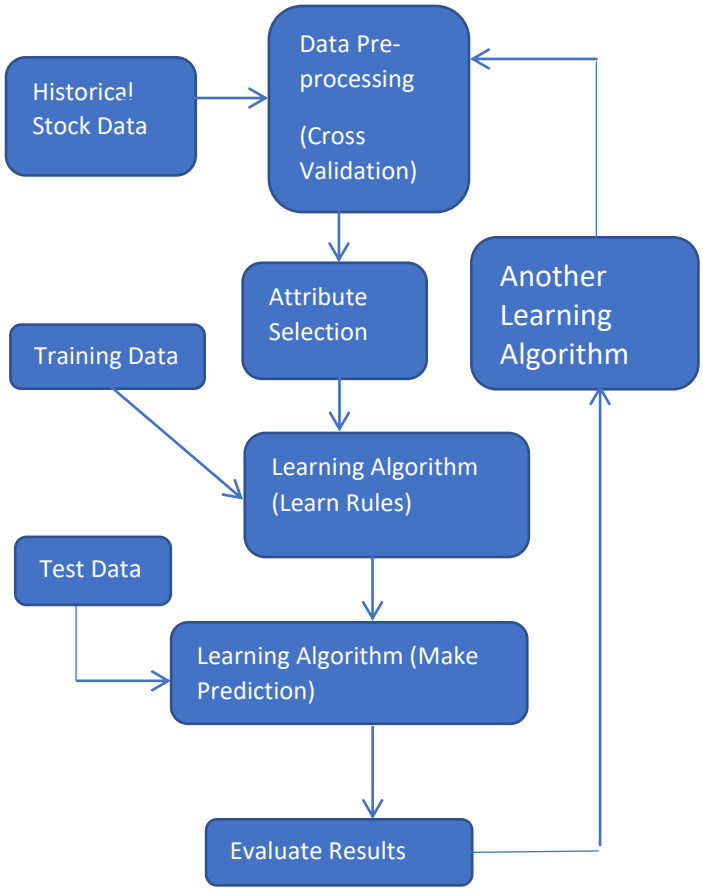



Fig 2.5: Existing System Architecture

## 2.4 PROPOSED SYSTEM

Stock is unpredicted and liberal in nature. The follow of the same is impressive and reluctant in nature. Finding the predictability and getting the nearest is the best hit goal for the same. The exact and accurate estimation of the same is never-less possible.

There are various constrains that in-fluctuate the pricing and the rate of stock. Those constrains had to be taken in consideration before jumping to the conclusion and report derivation.

As the reaction of stock prices to financial news are very sensitive in uncertain times of financial turmoil, it is necessary that we integrate that data in out model, So here we are looking at

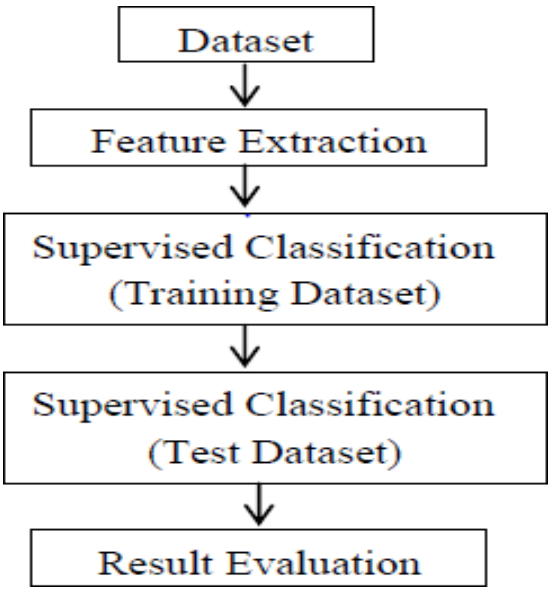



Fig 2.6 : System Flow

ways to take news data from financial news sites, and use the 'Natural Language Toolkit' to perform analysis of the news and extract sentimental data regarding the particular Stock.

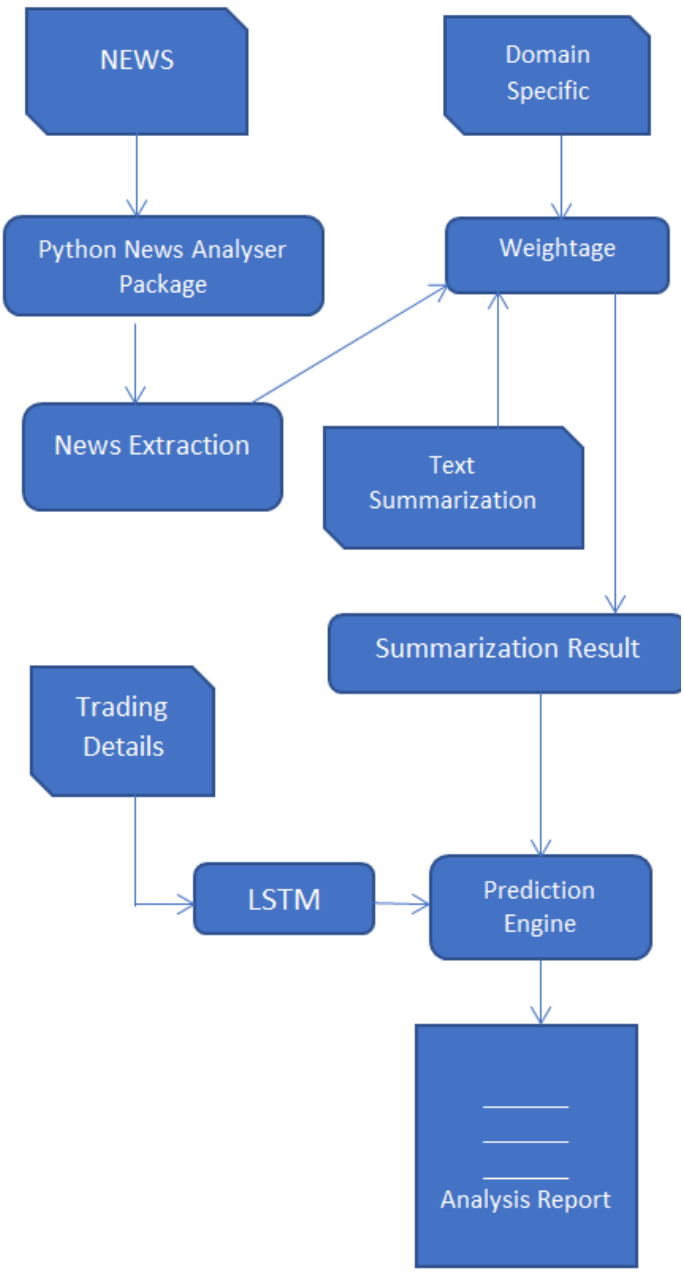



Fig 2.7 : Proposed System Architecture

Here as described in the figure above, the proposed system will have input from the datasets, these datasets are of two types, one containing the numerical data regarding the particular stock and its price, other data is the news converted in form of paragraphs and then into numeric data for textual analysis using natural language processing. The classification technique used here is supervised and the various techniques of machine level algorithms are implemented on the same.

Training Dataset are created for training the machine and the test cases are derived and implemented to carry out the visualization and the plotting's. The result generated are passed and visualized in the

graphical form.

### 2.5 SOFTWARE DESCRIPTION

#### 2.5.1 JUPYTER NOTEBOOK

Jupyter Notebook or so called IPython Notebook is an interactive web based computational mean for starting with Jupiter Notebook documents. The term notebook itself is a huge entity to represent the integration with different entity sets. JSON is the main document form from the same for the execution which follows the brief on the schema and the input and output means. It has high integration with several language set and has various flexibilities with the choices.

The extension used for the same is ".ipynb" which runs in this platform. It's an open-source software package with interactive communication means. It has it's open standards for the same. It's an open community best for budding programmers . The flexibility of the same is phenomenon and splendidly done the configuration and integration of the same is simplest and easy on hold so that no prior distortion is generated and the efficiency of the same is measured throughout any system of choice. It's the best software sets that been used across cross for designing and developing of the products and support wide help support.

Not only to that, it provides scalability in the code and the deployment of the same. Various Language can be changed and the project can be undertaken on the same. The created notebook files can be shared and stored in various means for further utilization. It supports cultivated and interactive output sets. Easily crossed over for graphing, plotting and visualizing of the elements.

Data Integration of the same is to its best. The integration of big data and it can process chunks of values in an approx. time which gives a better performance and the higher computational means. Various works on data like cleaning, cleansing, transforming modeling and visualizing can be done by the same.

## CHAPTER 3 REQUIREMENT ANALYSIS

### 3.1 FUNCTIONAL REQUIREMENTS

Functional requirements deals with the functionality of the software in the engineering view. The component flow and the structural flow of the same is enhanced and described by it.

The functional statement deals with the raw datasets that are categorized and learning from the same dataset. Later the datasets are categorized into clusters and the impairment of the same is checked for the efficiency purpose. After the dataset cleaning the data are cleansed and the machine learns and finds the pattern set for the same, it undergoes various iteration and produce output.

### 3.2 NON-FUNCTIONAL REQUIREMENTS

Non-functional requirement deals with the external factors which are non- functional in nature It is used for analysis purpose. Under the same the judgment of the operations are carried out for its performance. Stock is feasible and is ever changing so these extra effects and the requirements helps it to get the latest updates and integrate in a one go.

The non-functional requirements followed are its efficiency and hit gain ratio. The usability of the code for the further effectiveness and to implement and look for the security console. The System is reliable and the performance is maintained with the support of integration and portability of the same.

## CHAPTER 4 DESIGN

### 4.1 DESIGN GOALS

To make the project runs smoothly its required that we make plan and design some aspects like flowcharts and system architecture which are defined below.

#### 4.1.1 Data Collection

Data collection is one of the important and basic thing in this project. The right dataset must be provided to get robust results. Our data mainly consists of two parts, firstly the previous year or weeks stock prices. We will be taking and analyzing data from Yahoo Finance in the form of .csv file format. After that seeing the accuracy we will use the data in our model. And second is the news article data from finance news sites like Marketwatch, Reuters, NYTimes, Economic Times, Moneycontrol. This will be converted into numeric data using Natural Language tools so machine learning algorithms can perform analysis of reader's sentiments on these articles with the use of textual analysis.

#### 4.1.2 Data Pre processing

Human can understand any type of data but machine can't our model will also learn from scratch so it's better to make the data more machine readable. Raw data is usually inconsistent or incomplete. Data preprocessing involves checking missing values, converting the data in a specific format as per the requirement of the model, splitting the dataset and training the machine model, etc.

#### 4.1.3 Training Model

Similar to feeding somethings, machine/model should also learn by feeding and learning on data. The data set extracted from Yahoo Finance will be used to train the model. The training model uses a raw set of

data as the undefined dataset which is collected from the previous fiscal year and from the same dataset a refine view is presented which is seen as the desired output. For the refining of the dataset various algorithms are implemented to show the desired output.

### 4.2 SYSTEM ARCHITECTURE

The dataset we use for the proposed project is been taken from Kaggle. But, this data set is in raw format. The data set is a collection of valuation of stock market information about some companies. The initial step is to convert raw data into processed data. Which is done by feature extraction, since the raw data collected have multiple attributes but only some of those attributes are needed for the prediction. Feature extraction is a reduction process.

The structure, behavior and views of a system is given by structural model.

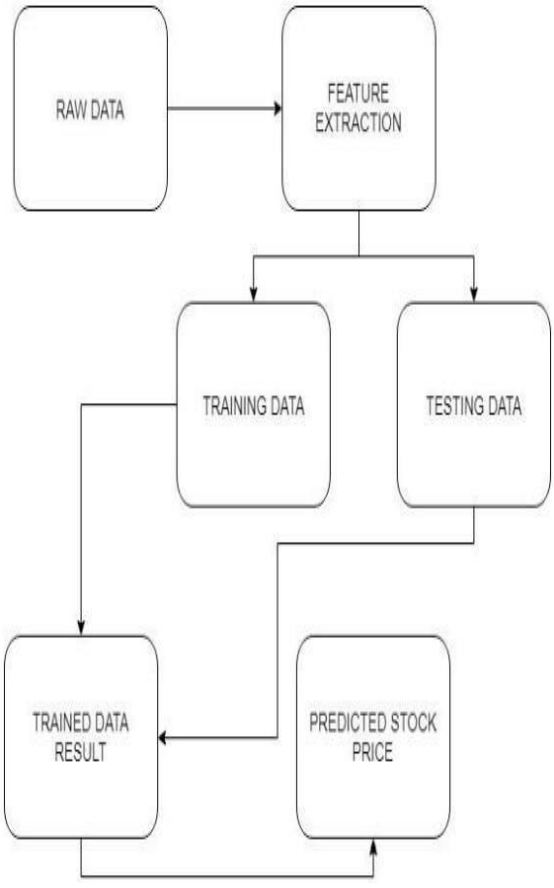



Fig 4.1: System Architecture

The above figure 4.1 gives the demonstration on the dataset extraction and redefining the raw dataset by categorizing into two phases of training and testing

From the given dataset a well modified categorization is extracted and a graph set is plotted to gain the required output which gives the stock prediction range.

### 4.3 Use case Diagram

A dynamic and behavioral diagram in UML is use case diagram. Use cases are basically set of actions, services which are used by system. To visualize the functionality requirement of the system this use case diagram are used. The internal and external events or party that may influence the system are also picturized. Use case diagram specify how the system acts on any action without worrying to know about the details how that functionality is achieved.

For the project we have created the below mentioned use case diagram.

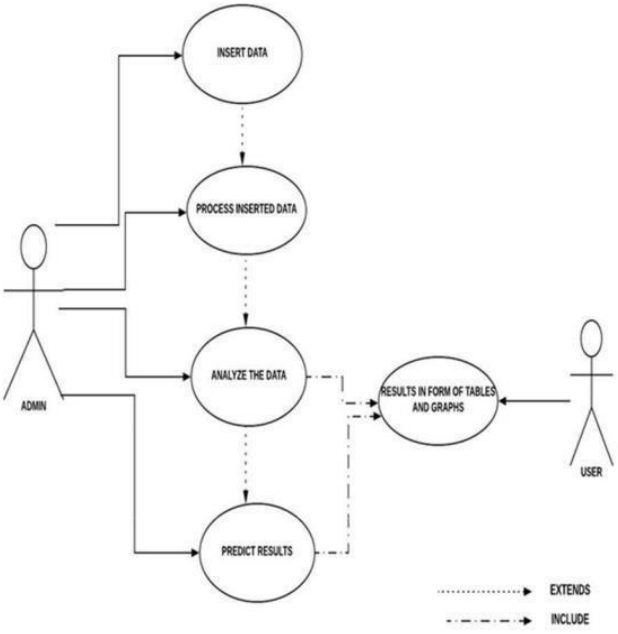



Fig 4.2: Use Case Diagram

The above figure 4.2 shows the use-case diagram of the entitled project and it's flow. From the diagram it's seen that the user gives the raw dataset as input and with the flow of the input in the system.

The system evaluates and process the dataset train itself with the provided dataset and extract the meaningful dataset to process and refine the cluster data and from the given cluster of the data, the plotting of the data values are shown and with the given range the system plots the data gives a figurative output as prediction and display the same as the refined output in the display screen.

### 4.4 Data Flow Diagram


Fig 4.3 Data Flow Diagram

In the above fig 4.3 we are taking a company fetching the data of the company from the panda's data-reader library then we are plotting the data, then we train the data to predict the stock for certain number of days. In this way data is flowing in our system.

## CONCLUSION & FUTURE WORK

In conclusion, Stock prices are a very volatile and unpredictable mechanism consisting of a complex system, with variable dependencies and a constantly changing curve turning the prices of stocks constantly.

Thus the evaluation and analysis of this requires a robust system which contains a huge input, various functionalities, packages and the use of high level Machine Learning computational algorithms. These algorithms, and functions need to work in an integrated way to give a highly accurate prediction, and should be flexible to analyse and consider changes in input data, regarding the news releases, and changes in price data.

In this paper various high level machine learning and artificial intelligence algorithms are discussed and implemented to give a highly accurate output prediction, which is identical to the real events to a large extent. In this process the raw data taken is the historical trading prices of the company's stocks, and the news articles from a news site of a certain time period. The relation between the stock prices and their reactions to regularly updating news is also considered and taken into account in order to increase the accuracy of the model's prediction with respect to real life events.

In all this process it is found out that, although complex and intricate, the prediction of future stock prices is possible that too with a good accuracy, and the tools helpful in this are Machine learning tool. Specificially, the Random Forest classifier algorithm is found out to be a majorly useful tool, as it is easy to implement, Artificial Neural Network is regarded to give good accuracy in the output. Various other algorithms and tools like Autoregressive integrated moving average (ARIMA), Long Short time series(LSTM), times series, mean square error, etc are also proven as a very useful part of the model. And apart from that the use of Natural language processing for the textual analysis of news articles to derive the sentiments around a particular stock proves as an helpful extra factor to give a rough idea about the future movement of the stock price.

Overall the flow of processes involved is firstly collecting the data, Pre-processing it and getting it ready to be given as input to training and testing models, then training the data, testing the data with the trained model and plotting the output to convey its accuracy.

To further increase the accuracy of the model to depict the real time events identically, various ways can be figured out to exploit the news article data and historical data in more efficient manner, so as to predict exactly what the technical and numerical output will be and how the sentiments of the investors after reading news articles will change regarding that particular stock affecting its prices in the near future.

## REFERENCES

[1] https://www.investopedia.com/terms/q/quantitativeanalysis.asp
[2] https://thecleverprogrammer.com/2021/06/21/microsoft-stock-price-prediction-with-machine-learning/
**[3]** Zhaoxia Wang , "Stock Market Prediction Analysis by Incorporating Social and News Opinion and Sentiments", November 2018
**[4]** K. Hiba Sadia, Aditya Sharma, Adarrsh Paul, Sarmistha Padhi, Saurav Sanyal, "Stock Market Prediction using Machine Learning Algorithms", April 2019, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-4
**[5]** Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar, "Stock Closing Price Prediction using Machine Learning Techniques", International Conference on Computational Intelligence and Data Science (ICCIDS 2019).
**[6]** Thierry Warin, Aleksandar Stojkov, "Machine Learning in Finance", 2 July 2021, Journal of Risks and Financial Management.
**[7]** Pushpendu Ghosh, Ariel Neufeld, Jajati Keshari Sahoo, "Forecasting directional movements of stock prices for intraday trading using LSTM and Random Forest", 30 June 2021, [a]Department of Computer Science and Information system, BITS Pilani K.K. Birla Goa Campus, India, [b]Divison of Mathematical Sciences, Nanyang Technological University, Singapore.

**[8]** Michele Costola, Michael Nofer, Oliver Hinz, Loriana Pelizzon, "Machine Learning Sentimen analysis, COVID-19 news and Stock Market reactions", 11 September 2020, Leibniz Institute for Financial Research SAFE.

**[9]** Shangxuan Han, "Stock Prediction with Random Forest and Long Short Term Memory", Fall 2019, Iowa State University, Ames, Iowa.