CHAPTER II

A REVIEW OF STOCK MARKET PREDICTION METHODS

2.1 INTRODUCTION

A survey is to be carried out to achieve the main objective that is described in the previous chapter which will be based on the content of relevant books, research papers and research theses. This survey provided me a greater insight into the stock market prediction methods and sentiment analysis. The stock market prediction need arises as the technology advances and to have the enhanced accuracy.

2.2 KNOWLEDGE DISCOVERY IN DATABASES (KDD)

The importance of the knowledge from the databases had been introduced in the workshop on KDD in 1989 (Frawley et. al., 1991). Knowledge discovery is the extraction of useful, unknown information from the data. This workshop on KDD started cultivation of many technologies in KDD. The data was being collected and stored for knowledge extraction. But the existing technology, methods and tools were not state-of-the-art level for handling rapidly growing data.

The KDD is considered to be the intersection of databases, artificial intelligence, pattern recognition, information retrieval and expert systems (Fayyad et al., 1996). Data Mining had been evolved as a step in KDD. Data mining refers to the process of producing the useful patterns, by applying the data analysis methods and algorithms. These methods and algorithms take computational efficiencies into consideration (Parker et al., 1998).

The various steps in the KDD process are shown in the Figure 2.1. In line with KDD, KDT (Knowledge Discovery in Text) showed its existence as well. According to the definition given by (Fayyad, 1996), KDT is the process of identifying the useful, novel and understandable patterns in the unstructured text data. The term KDT was also used by (Karanikas, 2002).
2.3 METHODS OF STOCK MARKET ANALYSIS

The investors use three main methods for analysing the stock market: Technical, Fundamental and Sentiment analysis of news articles.

2.3.1 Fundamental Analysis

Fundamental analysis (Abarbanell et al., 1997) is a way to evaluate the stock for predicting the stock price movement. It uses a method called “financial analysis” to achieve the same. The information that has to be taken into consideration, for analysis, includes the annual financial statements and reports of the company, its balance sheet, its health, its future prospects, industry comparisons, market environment and changes in the government policies etc. Fundamental analysis examines the firm’s financial statements to decide upon its worth, to
invest in the stock of the firm. Financial statements indicate cash flow, income as well as the balance sheets. This kind of information helps the investors to get some knowledge about the financial makeup of the company behind the stock.

The balance sheet shows the owner’s equity, the assets as well as the liabilities. Assets are the properties that the company owns which also have a potential to provide future value. It consists of the properties as well as the cash. Liabilities mostly comprise of the mortgages and the debts among a few others. Owner’s equity is termed as the amount of money raised by issuing the stocks to investors. Balance sheets show the investor, the way that company raises its money. Income statements do a similar job; they show the revenues and expenses of that company. These can be considered as the costs associated to run a business. The net income is calculated using the difference of the revenues and the expenses; this is essentially the earning of the company. The cash flow statements display how the company uses its cash for operations and making investments.

The stock market investor can use all these facts and figures to decide upon the feasibility to invest on that particular company. They can additionally use the ratios for further analysis like Price/Earning (P/E ratio), Price /Book value, Debt/Equity, Return on Equity, Current Ratio and Net Profit Margin.

The Price/Earnings Ratio or P/E ratio is a ratio used for valuing a company. The value is measured by its current share price, relative to its profit per share earnings. The Price/Earnings Ratio can be calculated as its share value in the market/ earning per share. It comes with a few limitations like the one when comparing the P/E ratios of different companies, another which lies in the calculation of the formula.

When a company’s stock’s market value is compared with its Price to book ratio it is known as Price to Book ratio. The division of its current closing price and the book value per share (of latest quarter) gives the P/B ratio.

\[
\frac{P}{B} \text{Ratio} = \frac{\text{Stock Price}}{\text{Total Assets} - \text{Intangible Assets and Liabilities}}
\]

To measure the financial leverage of a company one can use the debt ratio. The Debt Equity ratio is the division of the total liabilities of a company and equity of its stock holders. It shows the debt, the company uses to finance its assets relative to shareholder’s equity.
\[ \text{Debt – Equity Ratio} = \frac{\text{Total Liabilities}}{\text{Shareholders’ Equity}} \]

The return on equity measures the company’s profit by conceding the amount of profit it generates with the investment of the shareholders.

\[ \text{Return on Equity} = \frac{\text{Net Income}}{\text{Shareholders’ Equity}} \]

The current ratio which is also known as the liquidity ratio is a measure of the company’s capability to pay long term and short term obligations. This capability is measured by considering company’s total assets relative to its liabilities.

\[ \text{Current Ratio} = \frac{\text{Current Assets}}{\text{Current Liabilities}} \]

Net profit margin is the revenue which remains after deducting all the interest, expenses, tax and dividends from the total revenue.

\[ \text{Net Profit Margin} = \frac{\text{Total Revenue} - \text{Total Expenses}}{\text{Total Revenue}} \]

or

\[ \text{Net Profit Margin} = \frac{\text{Net Profit}}{\text{Total Revenue}} \]

### 2.3.2 Technical Analysis

The technical analysis, on the other hand, is a research on the stock prices in the stock market with the intent of making profits and/or investment decisions (YingziZhu et al., 2009). The technical analysis, when applied to the stock market, predicts the direction of the future stock prices based on their historic data. With a close examination of the previous price movements of a stock, the investor can predict the future price movements of that particular stock. But again, this forecasting too may not be 100% accurate but just like the weather forecasting; it gives the investors an overall picture of what likely is supposed to happen to the price of the stock. Technical analysis has a solid influence on the investor’s decision if he can or cannot,
safely bet on the stock. The stock can be bought when it’s running low and can be sold when it is at its peak. The technical analysis makes use of the price charts, studies the pricing patterns and makes use of certain formulae to bring in the future of the stock price.

This technique is mainly useful for the investors who look for some short term investments (also known as short term trading). This is readily applicable to the stocks whose prices are affected by the pressure of supply and demand. The price being considered here could be low, high, open or the closing price of that particular stock. The time frame to be considered can be intraday, daily, weekly, monthly or yearly. In case of intraday, it can be in every 10 minutes, half hourly or hourly. Charles Dow served as one of the initial editors for the Wall Street Journal and his ideas formed the basis for technical analysis. His theory named “Dow Theory” laid the basic foundation for the technical analysis. The main principles put forth by Dow through his theory were:

- **Market Price Discounts Everything**

  Technical analysts take into consideration only the current price of the stock while ignoring the fundamental analysis of the company. They believe that the current price puts light on all the other information required, right from the summed up knowledge of the investors, analysts, traders, technical and fundamental analysts to various other people who could affect the company in some way. They nevertheless believe that all those factors which are fundamental, psychological and economic, are already included in the price and hence there isn’t any need to analyse the company’s fundamental information. This leads to the analysis of the price developments of a particular stock in the stock market, which can now be seen as a result of the supply and demand.

- **Prices Movements are not Totally Random/Price move in trends**

  The technical analysts believe that the price developments of a stock follow a particular trend. They say, if the stock price of a stock falls, it will continue falling and if the stock price of a stock rises, it’ll continue to do so. This indicates that after the establishment of a major trend, the future trend is most likely to follow a similar direction. If the price is rising, the market is said to be “bullish” or “bearish”, otherwise.

- **Historic trends usually repeat in the same patterns**
History repeats, here in the context of the technical analysis of the stock market. Every time the price of a stock moves in a particular direction, the investors react consistently in the same manner to the price movements of the stock. The analysts of technical analysis use price charts and the technical indicators for future prediction. These form the mathematical derivatives of the volumes and prices of the stock. The analysts believe that the investors recursively go back to the past behavior in the same environment which forms the patterns on the chart, the technical analysts exploit these patterns and trends to foresee the direction of the stock market in future.

Tony Plummer paraphrased Oscar Wilde in his book titled, The Psychology of Technical Analysis, stating that, a technical analyst knows the price of everything but the value of nothing. The following two factors are of utmost concern to the technical analysts: the present price, the price movements in the past. Of course, the main war is between the demand and supply of the stock of the company and the price is the result of that battle, which is also termed as the end product. The whole objective of the technical analysis is forecasting the future direction of the stock price of the company. Technical analysts focus only on what the price is, in contrast to the fundamental analysts who are concerned with the prices going up and down. As mentioned above, the technical analysis is based entirely on price and volume. The fields, which are of prime concern, are:

- **Open**: This is the price at which the stock trades immediately after the opening of an exchange on a given day. An opening price of the stock is a very important indicator of trading activity for short-term traders such as day traders. It is considered to be the consensus price, after it is possible for the traders to “sleep on that”.

- **High**: A high is the highest price that the stock traded during the day. At this point, there are fewer buyers and more sellers.

- **Low**: A low is the lowest price that the stock traded during the day. At this point, there are fewer sellers and more buyers.

- **Close**: Close refers to the final price of particular trading stock for that particular day. This is the price used most often by the analysts. Most of the technical analysts use the relationship between the closing and opening price of a stock for its analysis and this kind of relationship is used with candlestick charts for analysis.

- **Volume**: The volume could be the total number of shares traded during a day. The relation between the volume and prices often forms an important factor for analysis.
Technicians apply a variety of methods, involving various techniques and tools, one among which is using charts. The use of charts allows the technical analysts to identify the pricing patterns as well as the market trends of the financial markets and helps them try to exploit such patterns. Technicians make use of charts in order to search for certain archetypal price patterns, an example to which would be the well-known head and shoulders or double bottom/top reversal patterns, moving averages, study indicators etc.

Technical analysts use the various available market indicators like the mathematical transformation of the price which includes the up and down volume, decline/advance data etc. Such indicators help to qualify if an asset is trending, and if it is the probable direction of its continuation. The technical analysts also observe the relation, the price/volume indices and the market indicators have. For example, the moving average, the relative strength index and Moving Average Convergence and Divergence (MACD) etc. Following any one of the many available techniques which can do the technical analysis, adherence of various techniques such as candlestick charting, Elliott Wave theory or Dow Theory, may lead to ignorance of other approaches. Mostly, the traders trust the combination of more than one such technique for their analysis. While some other technical analysts take help of subjective judgment to decide which pattern, a particular instrument reflects at a given time and how to interpret that pattern. Some other technical analysts employ a carefully planned, systematic or a strictly mechanical approach in order to identify the pattern and use it for further interpretation.

**Technical analysis indicators**

Technical indicators, for their analysis, can be grouped into various types based on their way of presentation of the data as well as the type of market conditions, which suits them the best.

- **Trending Indicators:** As the name suggests, trending indicators are used the most in identifying and confirming the trends in the price. They also help identify certain points, which indicate the end of a trend, or emergence of a new trend.

- **Momentum indicators:** These indicators are most useful in order to detect certain shifts in the trading activity, which may cater to both trending and non-trending market conditions, where most of the time of the markets is spent. They are often called oscillators; these could be split further, based on a zero line or a neural center.
- **Volatility Indicators**: These indicators help to measure the degree of variation in the movement of the price, restricted to a given time period, as well its comparison to the history of price movements.

- **Sentiment and Strength Indicators**: These indicators are used in combination with the price-based indicators to give an idea about how a trader responds to a price activity or what they think about the future.

- **Stock market indicators**: These indicators show the readings which are particularly related to the trading of stocks and they also provide key insights of the possible price movements depending upon the actions and opinions of the traders.

### 2.3.3 Sentiment Analysis

Earlier, when the internet did not come into existence, people used to take advice from experts and go through newspapers to monitor the stocks they want to invest in. But in the most recent times they use internet for these activities. It has been observed that there is extremely large growth in the internet users. This is because of the availability of new devices, technologies which are affordable to most of the users.

As mentioned above, with the advent of internet and online trading in India, it has been observed that, there is an exponential increase in user generated data on the internet. The people started using various blogs, websites to express their opinions and emotions on various subjects, be it political, brands or movies etc. This huge amount of data on the Internet provided the researchers a chance to analyse this data to find patterns, interesting observations and other knowledge from the same. It was found that this knowledge could be used to take some decisions.

As a consequence of exponential growth of on-online trading, huge amounts of data pertaining to stocks are available. Stock related data that is available online can be thought of as two types: numerical data in the form of historical prices and textual information which contains the news articles available in the blogs or websites of news media. Stock prediction, in the earlier work, was based on historical data. Most of the researchers used fundamental analysis and technical predictors, based on the historical data, to predict the fall or raise in the future stock price. In the recent times, the online content in the form of news articles has become an important role player in the stock market prediction.
There’s always a quest in the investors regarding the news items and expert advices, so that they can predict the direction of a particular stock price, they are interested in.

This task needs analysis of information which is in the form of text. But manually going through the huge amount of data became a daunting task. The solution did lie in automatically digesting the gist of the news. The text analysis in the form of sentiment analysis which automatically extracts the gist i.e. whether the news article is positive, negative or neutral was required. The techniques used for text analysis can be classified into three main categories: Bag of Words Approach, List of keywords approach and sentiment analysis. The bag of words is used in information retrieval and natural language processing.

In the bag of words model, the news article is represented by a vector which consists of the weights of the words, the grammar and the order are not considered. Generally, it is used in document classification methods in which the classifier is trained by using the word count as a feature (Fung et al, 2002; Schumaker et al., 2009). In the second approach i.e., list of keywords approach, the researchers have a list of keywords or terms related to the news articles. (Satoru Takahashi, 2006) analysed analysts’ reports by using keyword information and confirmed that it plays an important role in predicting future direction of stock earnings.

Sentiment Analysis is the third method which is used by numerous researchers. There is always an emotion, called sentiment, attached with these news items. For instance, an expert in a financial blog may predict that the stock price of a particular company goes up. Here the sentiment is considered as positive. (Liu, 2010) gave a very convincing definition of sentiment analysis. It defines a model in which the unstructured texts are discussed as structured data. The model is a quintuple consisting of \((o_j, f_{jk}, oo_{ijk}, h_i, t_j)\). This means, for an object \(o_j\), we have a set of features \(f_{jk}\). The opinion orientations for this structure are \(oo_{ijk}\) and these can be tied to the root object or to specific features. These opinions can also have a specific orientation and strength (joy, anger, positive, sad, negative, etc.). A few non-subjective opinions can also be added as part of factual data if the structure is intended to completion (neutral opinion orientations). Hence, the model will contain a combination of factual data as well as opinions referred to an object along with its features. The opinion holder \(h_i\) can be considered if a person or organisation claims this opinion at a particular time \(t_j\).

Sentiment analysis (Bing Liu, 2012), can also be referred to as opinion mining in many cases, is a field of study which analyses attitude, opinion or emotions expressed by the people about a product, event, issue etc. It is not specific to a particular domain but comes with
different aliases like opinion mining, sentiment mining, review mining etc. As mentioned before, manual evaluation of extremely large volumes of data is time consuming and a very tedious job. In order to analyse these large volumes of data they applied sentiment analysis which automatically detects the sentiments expressed about a particular post, news item or article (Pang Lee and Vaithyanathan, 2002; B. Liu, 2010; Turney, 2002).

According to Efficient Markets Hypothesis (EMH) (Malkiel, Burton G, 2003), the stock price of a company reflects the whole information and that the new information causes a change in the stock price. The researchers of the stock market then started applying the sentiment analysis on the stock market data. The word sentiment showed its existence as the opinion of an investor (DeLong et al., 1990). As the research in financial market field progressed, it has shown its evolution. The sentiment which was the response to the news article about the company, corporate announcements etc., was attributed to the stock price movements.

Sentiment analysis is a research area where the sentiment of a news article, or a review is defined as the opinion or sentiment expressed in that news article or the review (Turney, 2002). Sentiment analysis is also referred to as opinion analysis (Lee et al., 2008). The terminology, history and information regarding these terms were in their work. They concluded that both terms are similar. We use the term sentiment analysis through this thesis. For analysing sentiment of a news article there exists three methods (Lee et al., 2008), Machine learning, linguistic and lexicon methods (Liu, B., 2012) showed that the sentiment can be detected at three levels namely document, sentence and entity. As the name suggests the document level sentiment is the sentiment of overall document and sentence level sentiment is the sentiment expressed in each sentiment. We consider the news article as a document and the sentiment is the overall sentiment of the whole news article.

The traces of applications of sentiment detecting methods are found in (Be Pang, 2002) and (Peter, 2002). Their applications include movie and product reviews. The task of sentiment analysis is to find out the sentiment polarity about an entity in the news item about a financial institution. The sentiment extracted can be either of the two classes of attributes i.e., positive and negative or three class attributes entity i.e. positive, negative and neutral. Moshe et al (2006) established the fact that learning sentiment polarity form positive and negative examples were not going to increase the classification accuracy.

Due to huge amount of data available in the websites and blogs an automated tool is required to classify the sentiment in the news articles published in those websites and blogs. The
machine learning techniques in broad perspective, for classifying sentiment, are explained in (Po Pang, 2002). Movie reviews were taken as data set and the techniques like SVM, NBC and Maximum Entropy were covered in their research.

(Be Pang et. al, 2005), was also one of the researchers to carry out classification of sentiment polarity over a multi-way scale. The rating inference problem was addressed and whether the review was positive or negative was determined by multi point scale (1 to 5 star rating). The approaches used for sentiment analysis can be separated as four different categories (Catherine et.al, 2013) i.e. statistical methods, keyword spotting, concept-based techniques, and lexical affinity.

The keyword spotting approach classifies the text present in the different groups based on the presence of certain words like sad, up, down, and happy etc. (Otttany et al.,1988, Vanipriya et al., 2011). The lexical affinity method is similar to the previous approach but a probability is added to the keywords listed. This method outperforms the previous methods. This thesis focuses only on the technical analysis and sentiment analysis.

(Pak et al., 2010) did an exclusive research in using twitter feed as a source for sentiment classifiers. Emoticons played a key role in their study. They took help of the various emoticons to classify the tweets: Happy emoticons: “:-)”, “;)”, “=)”, “:D” etc. and sad emoticons: “:(“, “(“(, “(“,”(“ etc. They used TreeTagger on these tweets in order to tag all the words in each of the tweets and then they found out the grammatical patterns. The classifier thus presented, relied on the multinomial Naïve Bayes classifier which used POS-tags and N-gram as its key features.

2.4 SENTIMENT ANALYSIS METHODS

2.4.1 Feature selection for sentiment analysis

In the design of a sentiment analysis system, the feature selection holds key importance. Not choosing the features carefully, may lead to the classifier to not give accurate results. This calls for the three most popular techniques, POS tagging features, N-Grams and the syntax features. The use of these features either individually or in combination leads to an increase in the accuracy. As an example to this, the authors of a subjective text are usually described in first person whereas the verbs in the objective text are in third person. A simple past tense instead of a past participle is more prone to be used in Subjective texts. In contrast to the
positive set, the negative set consists of verbs more often in the past tense. This is because most of the authors express their negative sentiments about their loss or disappointment.

**POS tagging**

Stanford University developed a tool called POS tagger which is used to tag the parts of the speech in a file. An annotated text would be produced as an output after taking a file as an input that contains the reviews. Each of the terms in this annotated text contains the word and its corresponding POS. Each word of the review is tagged along with a grammatical feature with an aim of improving the accuracy and correctness which would help in finding useful patterns required for the classification.

The work on POS tagging can be seen in (Turney, 2004) that contains the patterns of the tags with a three word window (trigram). Adverbs, adjectives, single common nouns and plural common nouns were considered in his work.

**N-grams**

N-grams believe in using a combination of words as the required features. As an example, the combination of a single word “beautiful” could be a unigram for positive opinions, “I like” could be a bigram for positive opinions and “I don’t like” could be a trigram that does not show a negative opinion. Optimization on N-grams can be achieved via several techniques. For example, the N-grams that are very common but do not provide any information regarding the classification, could be avoided. N-grams feature a little amount of precision, when the negative words like not+like, doesn’t+work, etc, are attached to it. Now, choosing the correct dimension is also an important job. More generally, the bigrams have a good balance between the coverage (unigrams) and the ability to pick the sentiment expression patterns (trigrams). This has been proved by (Pak and Paroubek, 2010).

2.4.2 Syntax

In many cases, syntax is another feature which serves a great use case. For example, consider the sentence “My house is warm and cosy”. One can argue that the second adjective must be positive; this is because the first adjective is positive as well. This is particularly indicated by
the use of conjunction “and”. Such identification of certain connectors inside a piece of text can be used as valence shifters such as negation, intensifiers, and diminutive (Kennedy and Inkpen, 2006).

2.4.3 Term frequency
The sentiment can be categorized by considering the frequency of the appearance of a word in positive or negative contexts. This kind of a term frequency analysis model could be used to classify the text domain as well. This leads to the frequencies of the text to be useful in deciding whether a particular model can be chosen or not. A model based on such term frequencies is (Bakliwal et al., 2012), as an example.

The probability of being positive or negative can be defined by the model using the following formula:
\[ P_f = \frac{\text{Frequency in positive training set}}{\text{Frequency in negative training set}} \]
\[ P_p = P_f / (P_f + N_f) \]
\[ P_n = N_f / (P_f + N_f) \]
This type of model considers a training set to work perfectly, but they are greatly effective (around 87% of accuracy).

2.4.4 Noun identification and other wildcards
A good catch at increasing the accuracy is identifying the parts of speech that provide no subjectivity. Nouns are the usual example of such words which do not have a specific weight. Opinum (Bonev et al., 2012) forms an instance of such a solution. They made use of the named entity recognition algorithms to erase particular names from the training set. Converting “Bank BBVA is very bad” in “Bankbank-name is very bad” is exemplary. This technique can thus be used to isolate domain-independent negative words from domain-dependent negative words, and help develop cross-domain models.

2.4.5 Punctuation signs plus stop words removal
This process gives a boost to the classifications. It is very common to pick sentiments encoded in punctuation signs like exclamation marks or emoticons. Also, no real value is provided by the stop words. Stopwords are used in several cases that intend to connect different concepts and inculcate familiarity with a language, but this high frequency inherent in stop words suggests not using them in the classifications. When these stop words are removed the accuracy is usually incremented by one or two points as we have seen in (Bakliwal et al., 2012). It is good to include certain emoticons or the specific punctuation signs that encode the sentiments in the trainer, in order to gain more accuracy. Several opinions can be used to separate various opinions and keep them from mixing by using punctuation signs like commas or points.

2.4.6 Authority and Trust
In certain websites, some particular opinions hold more importance than others. For instance, eBay which is an online shopping site trusts the customers who have already sold something on a greater basis. This trend is aped in the reviews as well; certain lead writers have more influence compared to others if they happen to write a positive opinion about the product, the product thus gains an extra boost. (Ando et.al.,2012) contains a very creative paper which discusses the fact that how certain opinions serve a greater source of importance than others.

They formulated a dataset consisting of about 500 statements which included exclamation marks (!). The participants were then asked to choose the sentences, that they feel were more influential than others. This kind of analysis leads them to determine the key features of a particular hotel. They found out that the opinions which discuss about “room service”, “meal” and “scenery” hold more influence than others.
2.5 CLASSIFICATION TECHNIQUES

Figure 2.2: Techniques for Sentiment classification
(Source: Walaa et al. 2014)

Sentiment analysis classification falls broadly in two categories (Walaa et al. 2014) machine learning techniques and lexicon based techniques. Machine learning techniques use the standard machine learning algorithm for classification. It uses features for classification. The lexicon based techniques use lexicons which are pre compiled list of words. Machine learning techniques are of two types supervised and unsupervised techniques. The supervised classification techniques need training set of data that are already labelled whereas unsupervised techniques solve the domain dependency problem that reduces the need of training data. These are used when it is difficult to label the trained data.

The supervised learning techniques, in turn, are categorized into decision tree based, linear, rule based and probabilistic classifiers. When given an input sample, a probabilistic classifier will be able to predict a probability distribution over a given set of classes, rather than predicting the most likely class it belongs to. These are also called generative classifiers. Decision tree based classification involves division of training that is based on the condition of the attribute value. Training data set is decomposed hierarchically (Quinlan et al., 1986). Rule based classifiers model the dataset, with some set of rules. The left hand side of the rule
is a condition on the feature set and class label is on the right hand side. The feature set is expressed in the disjunctive normal form. Linear classifiers classify an object by using its characteristics (feature vector). The classification decision is based on the characteristics’ linear combination. The probabilistic classifiers are further classified into NBC, ME whose concepts are discussed in the subsequent sections. The linear classifiers are categorized in to SVM and Neural Networks.

Lexicon based approaches can be either dictionary based or corpus based. The dictionary base technique finds sentiment seed words and uses the dictionary for antonyms and synonyms whereas the corpus based technique will have a seed list of sentiment words and finds sentiment words in the huge corpus. The technique uses both semantic and statistical methods. The semantic technique is based on various principles for computing words similarity and provides the opinion values directly. On the other hand the statistical techniques find the seed sentiment words or patterns that co-occur.

2.5.1 Naive Bayes

A Naive Bayes (Pang et al., 2002) classifier assumes that the there is no relation between a feature being present or absent with another such feature being present or absent, given the class variable. Out of the several Naïve Bayes classifiers, we discuss three of them here, the Multinomial, Binarized Multinomial and the Bernoulli Naïve Bayes classifiers. (Zhang, 2004) showed that even though the assumption made usually does not hold true, its analysis has proved the unreasonable efficacy of the Naïve Bayes classifiers.

2.5.2 Pointwise Mutual Information (PMI)

PMI does semantic analysis of huge amount of data without requiring large training sets. (Su et al., 2006) The World Wide Web helps in estimating the statistical dependency between each pole of semantic orientation and the phrase in question (using a search engine, for instance).
Table 1.1: Examples of words which are positive and negative

<table>
<thead>
<tr>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad, poor, nasty, wrong, unfortunate, inferior</td>
<td>Excellent, good, fortunate, positive, nice, superior, correct</td>
</tr>
</tbody>
</table>

The search engine is consulted to determine the number of co-occurrences with the paradigm word sets and these hits are used in the following formula:

$$SO(\text{Phrase}) = \log_2 \frac{\text{hits(phrase NEAR positive)} \cdot \text{hits(NEAR negative)}}{\text{hits(phrase NEAR negative)} \cdot \text{hits(NEAR positive)}}$$

Lastly, we determine the probability of certain words in a phrase as being positive or negative based on their frequencies in the World Wide Web.

2.5.3 Support Vector Machines.

Many researchers of sentiment analysis found Support Vector Machines to be very effective. When compared to Naïve Bayes Classifier in text categorization, SVM outperformed NBC (Joachims, 1998). In contrast to NBC and Maximum Entropy, SVM is a large margin classifier, rather than a probabilistic classifier. In case of two class label, the procedure for training is finding a hyperplane that is represented by the vector $\vec{w}$, which separates, not only document classes of the two classes, but for the margin, as large as possible. The corresponding constrained optimization search problem is:

Let $c_j$ be the class and $c_j \in \{1, -1\}$ (1 for positive and -1 for negative).

Let $d_j$ be the document it belongs to the class $c_j$.

The solution for the above search problem can be put as:

$$\vec{w} = -\Sigma \alpha_j c_j d_j, \quad \alpha_j > 0$$

where we can obtain the $\alpha_j$’s from the solutions of dual optimization problem. The documents $d_j$ whose $\alpha_j$ values are greater than 0 are nothing but support vectors. These are the document vectors that only contribute to the vector $\alpha_j$. The classification of test instances could be made by considering the side of $\vec{w}$, hyperplane they fall on.
2.5.4 Maximum Entropy

Many natural language processing applications are more effective if they use an alternative technique of Maximum entropy classification (MaxEnt, or ME, for short). (Berger et al., 1996). Nigam et al. (1999) established that, this outperforms Naïve Bayes at standard text classification, though this is not always the case.

Its estimate of \( P(c|d) \) follows the following exponential form:

\[
P_{\text{ME}}(c|d) := \frac{1}{Z(d)} \exp \left( \sum_{i} \lambda_{i,c} F_{i,c}(d, c) \right)
\]

Here, \( Z(d) \) refers to the normalization function, \( F_{i,c} \) refers to the features/class function for the feature \( f_i \) and class \( c \), defined as,

\[
F_{i,c}(d, c) = \begin{cases} 
1, & n_i(d) > 0 \text{ and } c = c' \\
0, & \text{otherwise}
\end{cases}
\]

For example, a certain feature/class function might fire start if the bigram “still hate” appears and also the hypothesized sentiment of the document is found to be negative. An important observation is that, unlike the Naïve Bayes classifier, the MaxEnt does not assume any relation between the features, and hence we could expect better performance in case of conditional assumptions not being met.

The \( \lambda_{i,c} \)'s corresponds to the feature-weight parameters. When the definition of \( P_{\text{ME}} \) is accurately inspected, it shows that a large \( \lambda_{i,c} \) meant that \( f_i \) is considered to be a strong indicator for class \( c \). In order to maximize the entropy of the induced distribution (hence the classifier’s name), the parameter values are so set, subject to the constraint, that values estimated for the class-feature function in accordance to the model and match with the expected value with respect to the training data.

What here makes way for intuitive sense is the underlying philosophy being the provision to choose the model which makes the fewest possible assumptions about the data while not losing consistency. Our work uses ten iterations of the improved iterative scaling algorithm as suggested by (Della et al., 1997) for parameter training (this was proved to be sufficient for the convergence of the training data accuracy). It was used in combination with a Gaussian prior to prevent overfitting (Chen et al., 2000).
2.5.5 Neural Network for sentiment classification

A neuron is the basic unit of a Neural Network (NN) and a neural network is made up of a combination of many such neurons. The word frequency of the ith document is represented by $\bar{X}_i$ and this forms the input to the neurons. Each neuron is associated with a set of weights $A$ and these weights are used to calculate a function of its inputs $f(.)$. In a neural network, $p_i = A \cdot \bar{X}_i + b$ represents a linear function. For a binary classification problem, the sign of the prediction function gives the class label which is further denoted as $y_i$ (a class label of $X_i$). In case of non-linear boundaries, the multilayer neural networks are used. The task of approximation of the enclosed regions within a particular class uses these multiple layers to induce piecewise boundaries.

The neurons in the later layers are fed upon by output of the neurons from the earlier layers. As the errors have to go through back-propagation over different layers, the training process becomes more complex. (Ruiz et.al., 1999) showed several models of implementation of NNs for the textual data. (Moraes et. al., 2013) presents the document-level sentiment analysis through the empirical comparison between the Artificial Neural Networks (ANNs) and SVM. They chose to do such a comparison considering the fact that ANNs have attracted very little attention towards implementing SA whereas SVM has been a lot more successful in it. They have talked about the contexts, resulting models as well as the requirements, and they argue that both the approaches display a better accuracy in terms of classification. They showed that, barring some unbalanced data contexts, the ANN produced superior results compared to SVM.

After performing a test on three benchmark data sets viz. GPS, Books, Movie and Camera reviews from amazon.com, they proved that the experiment on movie reviews showed that SVM showed statistically significant difference than the ANN model. This gave a strong confirmation of the potential limitations that were present in both the models. These were very rarely mentioned in the SA literature. They showed that the computational effort by both the models i.e. ANN as well as SVM could be reduced by using the Information gain (this is computationally cheap feature selection method and this no way affects the accuracy of the classification results).
(Vane at. el., 2012) showed the use of SVM and NN for the classification of personal relationships in the biographical texts. The relations between two persons were marked as unknown, neutral or positive. They used the historical biographical information to describe people in a particular time frame, region and domain as their case study. They established the ability of their classifiers to label such relations above the majority class baseline score. They established that a training set that consists of relations, surrounding multiple persons, gave much more desirable results than compared to a training set that looks at one particular entity. They also showed that SVM and one layer NN (1-NN) algorithm gave the highest scores.

2.5.6 Hybrid Classifiers
A sentiment analysis system can be designed to use a combination of various classifiers, this forms another good approach. In some cases, the lesser quality of one classifier can be compensated by an improved quality of another one. (Das and Chen, 2007) devised a classifier which can extract sentiments from the stock message boards. It makes use of a mixture of various techniques, it uses a voting system to mark a word bullish (optimistic), bearish (pessimistic) or neutral (spam or neither bullish nor bearish). Overall classification, hence, depends obviously upon the majority vote amidst all the five classifiers (three out of five classifiers should agree on a particular message type). After carrying out such a process, if a majority could not be decided, the message is left unclassified. Hence, this technique improves the accuracy, though it reduces the number of the classified messages.

This thesis intends to study the impact of the sentiments in the news articles pertaining financial institutions listed in stock market, to analyse the overall health of a company. This is achieved by combining technical and sentiment analysis tools and building a system that would analyze and predict the variations in stock prices over a timeline, based on the sequence of events.

2.6 CORRELATION ANALYSIS BETWEEN THE STOCK PRICES AND SENTIMENTS

A snippet of, how exactly a dictionary approach looks like, was shown by (Wuthrich, et al, 1998). Their system could analyse overnight news which is nothing but, the news made at the time of financial markets being shut. Their dictionary consisted of different rows of words which were separated by a Boolean AND instruction, like “Bond and Strong”. The news articles could be categorized by the use of the dictionary. The sell/ buy instruction for the
index could be generated by counting the number of articles from each category. They argue to prove a claim of 21% advantage more, than compared to a trader who would play a guessing game on a uniform distribution.

A prototype (Wuthrich et al., 1998) was developed which analysed the news articles published in the websites in the night. The aim of this was to attempt to forecast the trend for one day, the five major equity indices, Straits Times, Dow Jones, FTSE, Nikkei, and Hang Seng. They employed three techniques namely neural networks, NBC and nearest neighbour. It was shown that nearest neighbour outperformed the neural network classification.

The paper which was most cited for the relationship between the stock prices and investor sentiment was (Otoo, 1999). The dataset used was the Wilshire 5000 total market index and University of Michigan’s Consumer Sentiment Index data. It was shown in her work that the increased values of equity boosted the sentiment. It was examined that whether both sentiment and prices of the stocks react to each other or some common factors made them to react each other. A VAR analysis conducted by her established that the shocks in the stock prices had significant effect on sentiments but there was no impact of sentiment on stock values.

An approach proposed by (Vijay Kumar Sagar et al., 1999) is similar to our work in which a combination of historical prices and news articles were used. Their aim was to show that the news articles, to an extent, influence the stock price. But the information extracted was limited pertaining to some limited words. TDNN was used for the prediction of the stock price in future and it was established that stock prediction would be better with the combination of news articles and historical prices than with only the historical prices. As it was mentioned earlier, a news article is analysed by using some automated tools to either two-way classify it into positive or negative or three-way classify it into neutral, positive or negative.

One more such system which analyses stuff based on a pre-compiled dictionary was brought into picture by (Peramunetilleke et al., 2002). Their model made an attempt to foresee the movements of targeted currencies. It assumed that the news articles which were published concurrently with a gain in a particular currency were positive, so was the case with negative articles and the drop. It limited its analysis to the headlines, as the headlines have a restricted usage of grammar and they give an accurate understanding of the whole plot. Their dictionary consisted of unigram words; it was these words that were extracted from the
headlines of the related news stories. All such words were stemmed and given a particular weight value. The unigrams were strategically assigned a positive or a negative polarity as mentioned above. The model categorized the news headlines using the terms in the dictionary and used such classified headlines to make a prediction about the future of the currency. The authors associated with the aforementioned model claim that almost all manual performance is correct nearly 50% times.

The fact that the consumer sentiment increases with stock is found by (Fisher et al., 2003). S&500 Index returns were used for their work. It was proved that there is a considerable link between investor’s confidence and stock returns but that is not statistically significant. It was established that the internet stock news content is related to the stock price (Xun Liang, 2005). But again, it was not that significant since they have used the volume of the news articles.

The usage of neural network was mentioned in the paper (Marc-André Mittermayer et al., 2006). Even the News Categorization and Trading System (NewsCATS) was dependent on a pre-compiled dictionary. NewsCats tried to predict the company’s share price by following and analysing a few press releases of that company. Even though their dictionary was not made publicly available, the authors still claim it to have been created manually. The methodology used to create the dictionary is also unknown. The dictionary allowed NewsCATS to seamlessly assign a pre-defined category to various press releases. The categories hence indicated the influence of the press release on the share price, positively or negatively. The authors of the NewsCATS system claim to prove a greater performance compared to a randomly bought stock by a regular trader.

A similar work can be found in (Jyotika et al., 2008). A heuristic approach called CoreEx was developed. This system extracts the relevant content from the news articles published online. The DOM tree of the page has been constructed. Every node has been scored, based on the number of links and the amount of the text, it contains.

A similar work can be found in (Tak-chung Fu, 2008). The sentiment analysis approach to extract sentiment polarity was proposed for the news articles in Chinese language. The three polarities i.e., positive, negative and neutral were taken in to consideration.
In the model proposed by (Long-Sheng et. al., 2009) four different semantic orientation indexes have been used as input neurons. The task is to identify the most critical semantic orientation that contributes to the sentiment classification. To achieve this, they have implemented structure pruning of neural network. Sum of absolute multiplication values are used to determine the important input node for the neural network. It is claimed that their proposed neural network based index outperforms the other traditional approaches.

Our extraction work of the content from the websites can be comparable with (Yan et al., 2010) who proposed an effective approach, called ECON which extracts the content form the websites, automatically. To represent the web page that contains the news, the system uses a DOM tree. It exploits the features of DOM tree. It finds the snippet node and then backtracks till the summary node. It was shown that ECON achieved highest accuracy for the news articles extraction.

Further empirical evidence can be obtained from the work (Caslav et al., 2011) that there is a correlation between the sentiment obtained from the news articles and future daily returns. All the news articles for a given company for one day are collected; the word vector is made and fed to the neural network which produces a text sentiment as an output. These text sentiments for a particular company for one day are averaged and aligned with the price of the same company on the same day. It was observed that this measure correlated with the actual stock price. A feed forward neural network with one input layer, two hidden layers with 16 and 8 neurons and one output layer was used.

A word-based technique for the extraction of sentiment from text was presented by (Maite at al., 2011). We extend the Semantic Orientation CALculator (SO-CAL) to other parts of speech, by piling up on the previous research which makes use of adjectives. This also marks the introduction of intensifiers which give a refined result for the negation approach. The results now obtained, represent a statistically-significant update to the earlier instantiations of the SO-CAL system. In addition, the work proves that a manually built up dictionary of words has a solid base for a lexicon-based approach. Such a dictionary is mandatory to gain fully from a SO-CAL system. A comparison of our dictionary with other, automatic or manual dictionaries showed that ours is a lot superior in its performance, in general.
This can be attributed to the fact that the criteria used by us to select/rank words is including lesser words and excluding words which cause ambiguity. Further, the use of Mechanical Turk Interface shows that these dictionary rankings go hand in hand with the personal judgements by humans. Essentially, the work proves that SO-CAL has a robust performance with respect to various types of reviews. This is a kind of domain-independence which is really complicated to be achieved using the available text classification methods.

The news articles were classified into either positive or negative by using artificial neural network (Anuj Sharma et al., 2012). Back Propagation Artificial Neural Network was proposed to classify the sentiment by using the combination of sentiment lexicons, info gain. One of the three techniques employed neural networks to predict the future stock price by combining time series analysis and sentiment analysis; it was evident in (Ding Tina et al, 2013).

(Sunil Kumar et al., 2014) used the sentiment analysis technique on the data obtained from social media and classified it by using the classification algorithm of machine learning. Thus classified data is then analysed to fix an overall mood of the comments. The comments are classified into four classes: hope, disappointment, happiness and sadness. The overall relative mood of all the classes in each day is taken as the input to the artificial neural network (ANN) which will be trained for the data of n days and their respective index value change, daily. This study is finally used in the vector prediction of Bombay Stock Exchange for the (n+1)th index value as suggested by (George, 1991). The study suggests that a function which maps input to output can be easily learnt by a neural network approach and encodes that output in the magnitude of weights in the connection of a network. The network consisted of a number of hidden units which lead to its viability. Increasing such hidden units gave rise to a higher performance up to a certain extent.

However, increasing the number of such hidden units beyond a limit made the performance of the model to go down. When the NN technique is compared with the MDA approach, the NN approach proves to be a significant improvement in the predictability of the stock price performance. Despite of the limitations of this approach, it could be proved that its use could improve the investor’s capability of decision making.
2.7. CALIBRATE THE EFFECT ON THE STOCK PRICES BY SENTIMENT ANALYSIS

(Brown et. al., 1989) proved that the value of technical analysis is boosted when used in a model where the prices are not fully revealed and traders have rational conjectures about the relation between signals and prices.(Neftci, 1991) proved that a few rules in technical analysis helped generate certain well-defined techniques for forecasting the stock prices, but it was shown that even the well-defined rules had no say in the prediction process, given the economic time series to be Gaussian. However, if the processes under consideration are non-linear, then the rules might capture some information. Tests showed that this may indeed be the case for the moving average rule.

The employment of a technical predictor, the moving average, proved to be prudential by (Brock et al., 1992). One of the two technical indicators is the moving average indicator. They developed a system that produces buy and sell signals. It is proved that an investment on the buy signal generated annual returns of 12%. The system proposed by (Leigh et al., 2002) uses the neural networks and the genetic algorithms in combination with the technical analysis tools, for the stock market analysis. Their work claimed positive and excess results in comparison with the basic strategies. The machine learning techniques NBC, decision tree and bagging were employed by (Vivek Sehgal et al., 2007). The sentiments expressed by the experts were extracted to train the system; it learns the relation between the values of the stock and the sentiments. This learned system would then be used to predict the future stock values.

(Asif Ullah, 2008) compares the forecasting accuracies of various technical indicators like MA, MACD with the backpropagation neural network and the backpropagation neural network based on the genetic algorithm.

There are many technical indicators that rule the prediction of stock movements. The most widely used and popular indicator among them all is the moving average indicator (Yingzi Zhu et al., 2009). A theoretical justification was given by them on the use of moving average indicator as the technical indicator. It was shown that an investor can add a value to his investment by using moving average indicator as technical analysis indicator.
The model for predicting the stock by (Deng et al., 2011) was based on technical and sentiment analysis proved that their combination gives significant results than the technical analysis itself. Different time forms problem was encountered by the system developed by (Zhang et al, 2011). According to them, when the web mining is done for sentiment analysis, the present time different time forms. The accuracy they could achieve is approximately 70%.

A stock prediction model called MKL (Multiple Kernel Learning) was proposed which (Shangkun et al., 2011) combines the technical and sentiment analysis to predict the direction of stock prices. The data of news articles, time series data and comments were downloaded and news articles were analysed for the sentiment. SentiWordnet 3.0 was the tool used in their model for analysing the sentiment in the news articles. The features extracted from different sources were obtained and the stock price was predicted based on a MKL regression framework. It was proved that their MKL prediction model outperformed the baseline models.

The textual information from the annual reports of a firm can be used as one of the indicators for forecasting the stock value of that firm (Petr Hájek et al., 2013). Their study combined the quantitative information in the form of fundamental analysis and sentiment analysis on the information available in the annual reports. The major findings include the explanation of the variance in the future stock price using the change in the sentiment specifically in negative terms, in the annual reports. The effectiveness of keywords was explained in the system proposed by (Linhao Zhang, 2013). The correlation between the twitter sentiment and stock prices was studied first and later, the words that correlate to the stock changes were determined.
2.8 FUTURE PREDICTION OF STOCK PRICES OVER A TIMELINE BASED ON THE SEQUENCE OF EVENTS.

The study of stock market using event knowledge was performed by (Kohara et.al., 1997). Their proposed system used prior knowledge to gain the event knowledge form the newspaper information.

![Figure 2.3 Extracting Event Knowledge](Source: Kohara et.al., 1997)

According to the Figure 2.2, the prior knowledge is the information gained from the previous experience. Based on the prior knowledge, the decisions, that the stock price can increase or decrease, is taken. The authors chose the price of crude oil, interest rates and NYDJ average and closing prices as inputs to be fed to NN. The results showed that the prediction ability was improved by reducing error rate on the 5% level of significance. After a research on certain crucial issues of stock markets (Paul D. Yoo et al., 2005) observed that events such as international issues, political issues can show notable effect on the stock prices. It was established that for predicting the stock market prices, web can be regarded as an important source, since it contain latest information on the events. By applying the efficient web mining techniques high accuracy can be achieved in predicting the stock market in short span of time. The authors further suggested that the database based on the historical events can be built to weigh the events and these events can be incorporated in to the time series data.

To predict the stock market, (DiWuet al., 2008) modelled a process that is called Non Homogeneous Hidden Markov model. This model considers the extraction of data from heterogeneous sources. Their model mainly contain three elements the external event state sequence, observed state sequence and hidden market state which denoted the events happening within the stock market, current prices (rise/fall) and conceptual state that is invisible, respectively. The external state is measured from the news articles; the observed
state is measured by stock prices. Based on these two pieces of information and previous hidden market state information, the current hidden state information is determined to predict the stock movements. It was claimed that their method showed better performance than the existing approaches.

(Neetu Mehndiratta et al., 2010) proposed an event study method that attempts to correlate between stock prices and dividend announcements. According to their study, the investors could gain the profit post dividend announcement. The price reactions of the companies listed in BSE were examined for sixty days of announcement dates. This study provided evidence that showed, increase in the dividend lead to positive abnormal returns.

(Yu, 2013) has built a dictionary pertaining to stock market. This dictionary which is domain-specific, outperformed the other general dictionaries in predicting the stock market. The availability of huge information sourcing from various heterogeneous environments makes the investors to confront with the problem of information overload. (Uta Hellinger, 2009) provided some methods and tools to overcome this problem and help in filtering relevant information and decision making. Their methodology provides an event analysis and a sentiment analysis module. Based on whether the news article is representing an actual event or expectation of future event, it is processed by the event analysis module or the sentiment analysis module.

A system has been created to predict the movements of the stock prices on a given day based on time series data and sentiment analysis (Tina, 2011). The three research questions were raised and answered. Their study is concerned about the accurate prediction method among SVM, Logistic Regression or Neural Networks, whether the use of 1, 5, 10 or 30 days before the market day is helpful in prediction, whether adding sentiments helps in prediction. Their results proved that SVM outperformed the other mentioned techniques, 5-days prior market day performed better than other frames and the addition of sentiment analysis increased the performance moderately.

(Masaudet al. 2013) proposed a novel procedure for sentiment estimation for the prediction of movements of the stock prices. Twitter posts are used for sentiment analysis. It is claimed that high precision classifier can be built by automatically generating training data. The training data is based on the events which are related to the rise or fall of the stock prices. For
example all the tweets of a particular day are treated as positive if the stock price of the company embedded in that tweet, is raised. Within four months their trading strategy was able to outperform S&P 500 by about 20%.