CHAPTER VII

FUTURE PREDICTION OF STOCK PRICES OVER A TIMELINE
BASED ON SEQUENCE OF EVENTS

7.1 INTRODUCTION

To keep abreast with the economic revolutions made in their stock markets by the developed countries like U.S.A. Securities and Exchange Board of India (SEBI) also initiated some reforms. One such reform is the announcement of earnings, dividends, merger and buy back, to name a few. Some researchers proved that these announcements make an impact on the stock and this leads to a new concept called event study. In stock markets, stock prices are influenced by many factors which include news releases or news regarding national economy. These can be termed as events (Ng, A. et al., 2003).

An assumption, that the future stock price of the stock also depends on international political events, lead to incorporation of event knowledge in the prediction of stock market. Many researchers included the event information in establishing the stock market prediction. In the research fraternity, many claimed that there is actually a correlation between the stock price and the event information. Some researchers combined numerical analysis with event information to predict the future stock prices. The formulation of the problem of stock price forecasting as an activity and monitoring the correlation between the event information from the news articles and stock prices is evident in (Fawcett et. al., 1996).

There has been sufficient evidence that news articles influence the movements of the stock markets. The financial domain has been showing very poor performance in the stock market analysis, by considering all the news articles in spite of the fact whether they are relevant or not. The websites are publishing the relevant and non-relevant news articles about a company. The researcher will find it difficult to find which relevant news articles to be parsed and which are not. The work carried out on the stock prediction was based on the features that do not represent exact events.
The previous study was based solely on the shallow features such as bag of words which cannot represent the event i.e. information about the entity-relation and there is no event capturing. There is growing need of converting the quality information in the articles, such as corporate announcements etc., into the measure which can be quantified, by considering its tone which can be either negative or positive.

While many of the researchers have proved that, sentiment score in the news articles has an impact on the stock prices, none of them had represented significant shift in the thought about the event information in the articles and its impact on the stock price. Now-a-days, fortunately the researchers are armed with textual analysis methods by which one can easily find out which news articles are relevant.

As human decisions are affected by news events and the fluctuations in the stock prices are affected by human trading, it becomes valid to say that the stock market is influenced by the events. Figure 7.1 shows an event “Bounced cheque case: A nonbailable warrant issued against Vijaya Malya“ published on 13th Oct 2012. This is a negative event and its impact on the stock prices can be seen in the Figure 7.2.

With the advent of internet, enormous amount of event information is available that is influencing the investors’ community. The event information is updated in real time, though it takes some time lag to be appeared in the newspapers. Event can be represented as a tuple (Xiao et.al, 2014) $E = (O_1, O_2, P, T)$, where $P$ is an action, $O_1$ is an actor, $O_2$ is the object, $T$ is the timestamp. For example in the event news “Bounced cheque case: A nonbailable warrant issued against Vijaya Malya” published on 13th Oct 2012 can be modelled as $O_1=$Government, $O_2=$Kingfisher Airlines, $P=$Issue of warrant, $T=$13 October 2012).
Our work implements an approach which is based on IE platform i.e. information extraction platform (Feldman et.al., 2011). The advanced techniques provided by Open Information Extraction i.e. Open IE made the extraction of events from the websites. An Open IE technology is adapted for extracting events from the large scale news without manual intervention. These events are used for prediction of stock movements. The methodology employed in our work is different from the others. It uses dictionary based sentiment score and a methodology to find the relevant events of the firms. Employing the textual analysis that is sophisticated than simply employing the positive and negative count of words, may significantly improvise the results. An evidence of strong relationship can be shown if relevant news is identified and accurately evaluated.
The research on studying the impact of news articles proved that, in measuring the tone of the investor, negative word classifications are more effective (Tim Loughran et al., 2011). Our study is based on the movements of the stock of the companies in a time frame around the
time at which there is negative story about that company. Many of the researchers have taken a single event for example dividend announcements (Neetu et al., 2010), arrest of personnel etc., into account while predicting the stock market. Their systems have limited vocabulary and the term weights are not based on relevant market. Our system takes sequential events into account and study the behaviour of the stocks over a period. The major challenges would include

- Formulation of a model that would extract and combine data, historical stock prices and news articles reflecting events from different sources.
- How to extract the sentiment in the news articles.
- How to define the time frames.

Our system goes through the news articles and does the parsing out the meaning in the context of relevant events such as law suits, mergers & acquisitions, launch of new products etc. Initially many events were being considered but later augmented by events that contribute most to the impact on the stock prices. Considerable effort was made in writing the rules that map the news articles about a company with event and sentiment score.

7.2 SENTIMENT ANALYSIS

After the news articles pertaining events are extracted the sentiment analysis is done to classify to check whether it is a negative event, positive event or neutral one. As mentioned in the previous chapters the main approaches for automatic extraction of sentiment are lexicon and machine learning approaches. The first approach is based on considering semantic orientation of the words in the news articles (Turney et al., 2002). In our system, we follow the first approach of sentiment analysis that uses a dictionary of words with polarity. In lexicon based approaches the dictionaries can be created manually (Tong., 2001) or automatically (Turney et al., 2003). Most of the researchers in the researcher community made use the generalized dictionaries for lexicon based sentiment analysis. The Loughran McDonald Dictionary is available, which contains the list of the positive and negative words pertaining financial domain. Our work has made use of this dictionary. Most of the researchers working on the prediction of stock market used this dictionary, but almost all of them extracted the news articles of the companies listed in the foreign stock markets like NYSE etc. Since we
focus on the companies listed in BSE, it was thought of building a dictionary manually in combination with LoughranMcDonald Dictionary so that it is customized accordingly, (YoosinKim et.al, 2014) manually built a dictionary for sentiment classification. The steps involved are shown in the figure.

![Figure 7.3: Steps involved in creating dictionary](image)

The general method of dictionary construction:

The method of dictionary construction is based on Jewnson’s inequality and likelihood estimation. The probability \( r_i \) conditioned for each document \( d_i \), where the probability of \( r_i \) conditioned on \( w_j \) is a multinomial distribution, and can be shown as:

\[
P\left( \frac{r_i}{d_i} \right) = \sum_{j=1}^{W} P\left( \frac{w_j}{d_i} \right) P\left( \frac{r_i}{w_j} \right) \quad \text{...........................................(1)}
\]
Where \( P\left( \frac{r_i}{w_j} \right) = \prod_{k=1}^{E} \left( \frac{p(\frac{e_k}{w_j})}{r_{ik}} \right) \)

\[
P\left( \frac{r_i}{d_i} \right) = \sum_{j=1}^{W} P\left( \frac{w_j}{d_i} \right) \prod_{k=1}^{E} \left( \frac{p(\frac{e_k}{w_j})}{r_{ik}} \right)
\]...........................(2)

The words in the document \( d_i \) are assumed to be independent.

Let \( \sigma_{ij} = P\left( \frac{w_j}{d_i} \right) \)

And \( \Theta_{ik} = P\left( \frac{e_k}{d_i} \right) \), let \( N \) be the number of documents, then log of likelihood over \( N \) can be defined as

\[
\log l = \log \left( \prod_{i=1}^{N} \sum_{j=1}^{W} \sigma_{ij} \sum_{k=1}^{E} r_{ik} \log \Theta_{jk} \right) = \sum_{i=1}^{N} \log \left( \sum_{j=1}^{W} \sigma_{ij} \prod_{k=1}^{E} \Theta_{jk} \right)...........................(3)
\]

Log-likelihood could be constructed, according to Jensen’s inequality, as follows:

\[
\log l \geq \sum_{i=1}^{N} \sum_{j=1}^{W} \sigma_{ij} \sum_{k=1}^{E} r_{ik} \log \Theta_{jk} + \lambda \left( \sum_{k=1}^{E} \Theta_{jk} - 1 \right)...........................(4)
\]

Since \( \sum_{k=1}^{E} \Theta_{jk} = 1 \) a Langrange multiplier can be added to the log-likelihood equation as shown below:

\[
l = \sum_{i=1}^{N} \sum_{j=1}^{W} \sigma_{ij} \sum_{k=1}^{E} r_{ik} \log \Theta_{jk} + \lambda \left( \sum_{k=1}^{E} \Theta_{jk} - 1 \right)...........................(5)
\]

Then by first order partial derivative of \( \Theta_{jk} \) can be calculated in order to maximize the likelihood.

\[
\frac{\partial l}{\partial \Theta_{jk}} = \sum_{i=1}^{N} \sigma_{ij} r_{ik} \log \Theta_{jk} + \lambda = \sum_{i=1}^{N} \sigma_{ij} r_{ik} \Theta_{jk} + \lambda = 0...........................(6)
\]

Hence \( \Theta_{jk} = \frac{-\sum_{i=1}^{N} \sigma_{ij} r_{ik}}{\lambda} \)...........................(7)

Since \( \sum_{k=1}^{E} \Theta_{jk} = 1 \),

\[
\lambda = -\sum_{k=1}^{E} \sum_{i=1}^{N} \sigma_{ij} r_{ik}...........................(8)
\]

By substituting the value of \( \lambda \) obtained in equation 8, in equation 7 we get,

\[
\Theta_{jk} = \frac{\sum_{i=1}^{N} \sigma_{ij} r_{ik}}{\sum_{k=1}^{E} \sum_{i=1}^{N} \sigma_{ij} r_{ik}}...........................(9)
\]

Therefore:

\[
P(e_k|w_j) = \frac{\sum_{i=1}^{N} p(w_j|d_i) r_{ik}}{\sum_{k=1}^{E} \sum_{i=1}^{N} p(w_j|d_i) r_{ik}}...........................(10)
\]
\( P(e_k | w_j) \) is the probability of the sentiment polarity \( e_k \) which is conditioned on the word \( w_j \) and this would become a basis for building the dictionary.

Most of the work done in the event based sentiment analysis, the researchers assigned the sentiment to the news articles based on the rise or fall of the stock market. For example if the opening price on 12/12/2015 is 345.00 and closing price on the same day is 350.00, then all the news articles that are published on the same day are termed as positive. The sentiment extraction does not consider the semantic of the words; it assigns the score based on whether the stock of the company goes up or down. Our system is in contrast to this, consider the sentiment by considering the meaning of the word i.e. whether it is positive or negative by using dictionary approach.

The first and foremost step includes extraction of stock values of the companies listed in BSE. The companies’ related news articles are also extracted and analysed for sentiments. As it was proven fact that negative news has greater impact on the stock values, only the negative news articles reflecting some events are taken into consideration. The movements of stock values of the concerned companies in time intervals of one day, one week, one month, one year are studied after another event of that company takes place and a correlation between the price and the timeline of such stocks is made.

### 7.3 THE DATA COLLECTION

Our data set contains the closing prices of 3000 companies that are listed in Bombay Stock Exchange. We had collected eighteen lakh, end of the day price points. The time period considered was from the year 2007 to 2014. On the other hand about thirty million news articles were collected for the same period. We have also considered about a one and half lack insider trading disclosures.

Highlights are:

- 3000 stocks of BSE
- 18 lakh end of day price points
- From 2007 to 2014
- 30 million news articles
- 1.4 lakh insider trading disclosures
For every listed entity in BSE, we took price on a daily basis, mapped with news on a timeline. This could look like a simple Google news graph. The companies that have negative news, separated between 2 to 6 weeks are selected. The average prices of these stocks are mapped.

The historical data for all the companies listed in BSE (around 3000) are extracted from moneycontrol.com website. The code (scraper) is written to extract the each company’s open price, close price for the years 2007 to 2014. Approximately 18 lakh price points from BSE have been collected. Events from disclosure records (1.4 lakh) and content pieces collected from indiatimes.com, moneycontrol.com, sebi.com, watchoutinvestors.com, ecourts.gov.in and cibil.com (30 mil) are used as corpus for this system.

7.4 DATA PRE-PROCESSING

In our study, the main focus was to analyse the news articles of the stocks and to observe the relation between the stock price and news articles. We have manually labelled 5000 news articles. Manual labelling was required to test the dictionary that is built for sentiment detection.

The proposed system uses three approaches NBC, SVM and dictionary based system. In dictionary based system Loughran and McDonald Financial Sentiment Dictionaries are used for sentiment analysis as a starting point. We assign sentiment score for the words listed in the positive word lists and negative word lists in the standard dictionary as 1 since the dictionary has all the impact words that related to the financial domain.

5000 news articles have been selected and manually tagged for sentiments i.e. positive, negative and neutral. Each of the news article is represented as bag of words removing the stop words like “is, a” etc. The distinct words in each news articles are collected and its positive and negative probabilities are calculated in the following way:-

\[\text{Pos}_\text{prob} (\text{word}) = \frac{\text{No. of times it appeared in the positive articles}}{\text{Total no positive articles}}.\]

\[\text{Neg}_\text{prob} (\text{word}) = \frac{\text{No. of times it appeared in the negative articles}}{\text{Total no negative articles}}.\]

The probabilities obtained for the words are added to the standard dictionary. If the word to be added is already presented in the standard dictionary that word is discarded.
7.5 EXTRACT THE SENTIMENT

The Classifier Module classifies a news article as a positive or negative or neutral using Bag of Words approach. Each of the word except the stop word is compared with the words in the dictionary and its probability is collected. The sum of the probabilities of all the negative words and the sum of the probabilities of all the positive words are calculated. If \( \prod (\text{Pos}\_\text{prob}) > \prod (\text{Neg}\_\text{prob}) \), the article is considered as positive otherwise Negative. If the difference in the scores of \( \text{Pos}\_\text{prob} \) and \( \text{Neg}\_\text{prob} \) is less than 5%, then the document is treated to be neutral.

**Algorithm: Sentiment prediction of the events in the news articles**

Step1: Choose the significant event news articles.
Step2: Remove the stop words.
Step3: Extract the remaining words and calculate the probability
Step4: Assign the probability 1 to the words present in the dictionary
Step5: Add the words extracted in step 2 to the dictionary
Step6: Assign an appropriate label for each article as either positive or negative based on the formula
Step7: Randomly choose some articles and manually tag
Step8: Train a classifier on the labelled articles.
Step9: Predict sentiment of new future articles