CHAPTER VI

CALIBRATE THE EFFECT ON THE STOCK PRICES BY

SENTIMENT ANALYSIS

6.1 INTRODUCTION

Every investor’s motive is to invest in a company with a good profile and grab huge profits in a short span of time. The investors would not want to invest in a company with high debt and suffer later, for example. No investor has any clue of what is going on in the company and hence, he cannot act in advance in case of any financial calamities.

In this scenario, if the investor is armed with the tool which gives an early prediction of the companies that are in trouble or are suffering from some financial crunch, it’s a breakthrough for the investors to take informed decisions. This tool can be utilized by the investor to verify the health of the company and thus he can opt out, if it is known that the company is going through some critical phase.

By now, it is clearly evident that forecasting the stock price plays a vital role in financial as well as academic research. As mentioned earlier, there are three schools of thought namely, technical, fundamental and sentiment analysis. The fundamental analysis is something to deal with the information of the company like the company’s assets, management etc. On the other hand, the technical analysis is based solely on the stock prices. The sentiment analysis came into existence after the researcher community proved that there is an impact of the news articles of a company on its stock price.

According to (Jekaterina et al., 2014), there was major drawback with fundamental analysis, the one, which comes from the followers of EMH. It is believed that the stock prices reveal all the needed information. As the technical analysis deals with the stock prices, the proposed system employs this approach. It was also proved that there is a correlation between the news articles and the stock prices of a company, thus making use of sentiment analysis is inevitable. To exploit the goodness of both worlds i.e., technical and sentiment analysis, we develop a hybrid model to predict the overall health of a company.
Technical Analysis is generally based on data visualization and some statistical tools. Technical analysis of a stock or security does not absolutely predict the future price of that stock but it helps the investors to anticipate the possible thing that can happen to the prices over the time. Technical indicators in the stock market provide insight to the investors to identify major turning points in the market.

6.2 MOVING AVERAGE INDICATOR IN TECHNICAL ANALYSIS

There are many technical indicators; some of them are Trending Indicators, Momentum Indicators, Volatility Indicators, Strength and Sentiment Indicators, Stock Market Indicators, Moving average indicators. Among these “Moving average indicators” was chosen for the proposed work. The moving average indicator gives the average value of a Stock’s price over a period of time. The moving average indicator can be calculated by considering mathematical analysis of a stock’s average value over a predetermined time period. It is a trend following or lagging indicator because it is based on the past prices.

An indicator of average value of a stock over a period of time is called moving average indicator (ClifDroke, 2001). A moving average can be calculated by using the mathematical analysis of the stock’s average price value over a period of time which is predetermined.

Whenever the stock’s price changes, the average price either goes up or comes down. It is the indicator that is mostly used in technical indicators. Various types of moving average indicators are the simple (arithmetic), triangular, exponential, and variable and weighted moving averages and these are calculated by using open, close, low, high and volume of a stock price.

Moving Averages can be calculated in very short time frame or long time as shown in the following Figure 6.1:
Table 6.1: Time frames for moving averages
(Source: Moving Averages Simplified by ClifDroke, page 14)

<table>
<thead>
<tr>
<th>Trend</th>
<th>Moving Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Short Term</td>
<td>5-13 days</td>
</tr>
<tr>
<td>Short Term</td>
<td>14-25 days</td>
</tr>
<tr>
<td>Minor Intermediate</td>
<td>26-49 days</td>
</tr>
<tr>
<td>Intermediate</td>
<td>50-100 days</td>
</tr>
<tr>
<td>Long Term</td>
<td>100-200 days</td>
</tr>
</tbody>
</table>

According to (Marcus C. Toms, 2011) systematic selling and buying of stock whenever the price of a stock crosses its average, is referred to as moving average trading. The prices move in trends and they follow uptrend or downtrend at any point of time. A trend can be defined as the general direction of the price of the stock or market. The length of a trend can categorized as short term, intermediate or long term.[Source: http://www.investopedia.com].

The price can follow an uptrend or downtrend at each point of time. These common terms describe the flow of trading in the market. If a stock is showing an uptrend, the investors can gain profit by buying that stock that appreciates in value. In uptrend, the stock value consistently appreciates in value over a period of time, even though there are brief dips here and there. A particular stock price of a company, for instance started at say at Rs.255, its price reached say Rs.350, we say that there an uptrend for this stock, even though there were brief dips in the stock prices in between. An uptrend can range from hours to months and to years. A downtrend behaves exactly opposite to an uptrend. In downtrend, the stock price consistently depreciates over a time period which may have some brief raises. The investor can be benefitted if he follows these signals. Moving average trading is profitable if the changes in the level of the price are sufficient in between the sell and buy signals, otherwise it is loss making. The following Figure 6.1 shows an uptrend.
6.3 SENTIMENT ANALYSIS OF NEWS ARTICLES

“What you see is news, what you know is background, what you feel is opinion.” *Lester Markel, American journalist, 1894-1977.*

News is accessible to the people when it is publicly shared. In financial domain the news about the financial institutions, stocks and government policies pertaining the stock market
are reported. The information about the events happening in the companies is immediately reported. As a consequence of this news reports on a company, the opinions are formed among the investors so that they can take informed decision about their share in the stock of that company. The opinion expressed in the news is nothing but the sentiment, which plays an important role.

Sentiment analysis is the task of choosing the sentiment label for a given news article. It can also be considered as a classification task. All the inputs are considered independent of each other. The class labels are predefined. The Sentiment classification adapted for this work is a multi-class classifier in the sense, the news articles are classified as positive, negative or neutral.

Figure 6.2: Supervised classification framework.

As shown in the Figure 6.2, in supervised learning there are two phases, training phase and testing or prediction phase. In training phase feature extractor converts the input values into a feature set which contain the basic information for all the inputs, form the basis for
classification. A model can be generated by feeding the feature sets and labels in pairs into the machine learning algorithm. The feature extractor which was used in the training phase is also used to extract the features. These features are fed to the model that predicts the class labels for the inputs.

NBC classifier has been chosen for the classification task in this work. NBC is a supervised classifier which is based on the training corpus that consists of the inputs along with the correct class labels. It is a probabilistic classifier based on the Bayes theorem. It uses assumptions on the independence between the predictors. It is naïve since the assumptions rarely hold for the problems of real world. For NBC, the document is simply represented by a bag of words.

The Bayes’s rule can be applied as:

\[ P(c|d) = \frac{P(d|c)P(c)}{P(d)} \]

\[ C_{MAP} = \text{argmax}_c P(c|d) \]

Applying Bayes theorem, we get

\[ C_{MAP} = \text{argmax}_c \frac{P(d|c)P(c)}{P(d)} \]

After dropping the denominator

\[ C_{MAP} = \text{argmax}_c P(d|c)P(c) \]

Assuming that the document is represented by the features \( x_1, x_2, x_3, \ldots, x_n \)

\[ C_{MAP} = \text{argmax}_c \frac{P(x_1, x_2, x_3, \ldots, x_n|c)P(c)}{P(d)} \]

Every feature in a Naïve Bayes classifier is responsible for determining the label that should be assigned to a certain input value. In order to select a label for a given input value, the Naïve Bayes classifier initially calculates the prior probability of each label, which is
obtained using the frequency of each label in the training set. This leads to a probable estimate for each label. The input value is then assigned a label with the highest likelihood estimate.

**Algorithm** Stock market analysis by using technical and sentiment analysis

1. Choose the company.
2. Extract the stock prices through a period of one year of that company.
3. Calculate moving average \(m\) for a day which is the average value of previous 200 days and compare with that day’s closing price \(cp\).
   
   if \(m > cp\)
   
   then “it is Uptrend”
   
   else
   
   “it is Downtrend.

4. Extract the news articles of that company for a period of time.
5. Manually tag some of these news articles as positive, Negative or neutral.
6. Train the classifier (NBC has been used) on these set of data.
7. Predict the sentiment of news articles.
8. Extract the percentage of positive vs negative articles.
9. Combine Technical and sentiment analysis to analyze the company trend.
6.4 DATA ACQUISITION

6.4.1 Extraction of historical data

To train and evaluate the system, historical prices are needed. The historical stock values are extracted for a period of 1 year from the website in.yahoo.finance.com. The data extracted is in a CSV format. This data is then placed into a database table. This data can then be queried and processed. The crawler is implemented in PHP that crawls on to the website in.yahoo.finance.com.

6.4.2 Extraction of News articles (News Crawler)

Credibility and validity of news articles play a vital role before dissemination of news articles. The financial media reporters gather the information through the reliable sources and the same would be disseminated in the form of news article. The news articles that are published must ensure the trustworthiness of the news. The news article then create a common opinion among the market players which forms the basis for sentiment.

There are different media through which these news articles are disseminated, utter care needs to be taken in deciding the source from which the sentiment is to be derived from the news articles published in that source. The accuracy of the research results depends on the trustworthy sources. There are some official sites pertaining to the financial news articles which are trustworthy like Yahoofinance, moneycontrol etc. which guarantee the reliability of the news articles published. There are some blogs and social media where the general public opinion is expressed. Since the information from such sources cannot be validated, it is considered to go with the trustworthy sites.

According to Leinweber’s (2009), different forms of news can be classified as:

1. News: It refers to the news articles which are produced by the reliable sources and disseminated through newspapers (both online and offline), radio and TVs.

2. Pre-news: It is the source of data from which the news has been produced that can be derived from the primary information sources like corporate announcements, court documents, stock exchange etc.

3. Web sites and social media: They are the websites, blogs or social media like Facebook or twitter.
The news data which is the heart of our study uses third category. The initial step in sentiment analysis is the identification of the website containing the news articles of the chosen company. In the proposed model, the website “moneycontrol.com” has been chosen to extract the news articles that will be analysed. Once the site is chosen, the process of extraction begins, which includes the extraction of only the news articles over a period of 6 months while ignoring the various ads and other unnecessary contents from the link. Only the links to the news articles are taken into consideration and stored in a file.

6.5 DATA PRE-PROCESSING

All the URLs of the news articles for the selected company from the website www.moneycontrol.com are extracted and written into a file. This file is then used to open each URL separately and extract the news articles which it contains. The URLs that do not contain news articles are filtered and only the relevant URLs are written into the file that contains all the articles specified by the collected URLs and this file serves as an input to the Sentiment Analysis module. The role of text pre-processing in sentiment analysis (Haddi et.al.,2013) is a very important step since the accuracy of the classification is based heavily on the method chosen for the pre-processing. For data pre-processing the standard procedure of text pre-processing technique has been followed. The text in the news articles consists of the words that don’t contribute to the sentiment of that news article. For example, they may contain URLs which does not contribute. We need to remove such noise. The steps involved in pre-processing are tokenizing the text, stop words (is, was etc.) removal and n-gram construction. The meta data such as date and name of the companies are used based on which the news articles will be retrieved.

6.5.1 Tokenization

The news articles are split into words by considering the space between them. These are called the tokens. This concept is called the bag of words concept. All these tokens are used for training classifier.
6.5.2 Stop word removal

As mentioned before the stop words are to be removed since they become overhead for the system. NLTK (Natural Language Tool Kit) comes with a dictionary which contains the stop words. All the tokens are compared with these words in the dictionary; if the word is present in the dictionary it will be removed.

6.6 EXPERIMENTAL SETTING

As mentioned earlier, the common approaches for sentiment analysis are machine learning, lexicon based and linguistic approach (Thelwall et al., 2011). Machine learning algorithms can be supervised or unsupervised. For this proposed system, a supervised learning algorithm has been used. Standard supervised machine learning method requires news articles which are manually tagged as either positive or negative class. This dataset labelled is used as a training set for the unlabelled news articles. In our study we have used Naïve Byes Classification that comes under the machine learning category.

6.6.1 Sentiment Analysis

A training set is created which contains manually tagged articles in a Comma Separated Value (CSV) format. The statistical learning method called Bayesian Algorithm has been used for classification. According this classification method, each word is weighted with the frequency in the article and this is categorized in to the specified category. The categories used are positive, negative and neutral. The classifier is trained with this set. After training, the file with the extracted reviews is passed as an input to the classifier which then predicts the sentiment of each article. The final output is a percentage of positive and negative articles which is represented graphically in a pie chart.

6.6.2 Technical Analysis

The 200- day moving average is a popular technical indicator which the investors use to analyse the price trends. It is nothing but the stock’s average closing price over the last 200 days (i.e. \( \frac{\text{stock price of day 1} + \text{stock price of day 2} \ldots \text{stock price of day 200}}{200} \)).
\[ MA = \frac{\sum_{i=1}^{200} SP_i}{200} \]

where \( MA \)=Moving Average and \( SP \)=Stock Price

It says about the market trend, by looking at the previous 200 trading days and taking the average of closing prices of 199 previous trading days.

\[ YAP = \frac{(CP \text{ of yesterday} + CP \text{ of 199 PTD})}{200} \]

where \( YAP \)=Yesterday’s Average Price, \( CP \)=Closing Price and \( PTD \)=Previous Closing Days.

Mark that point on a chart and make another dot by considering the closing price of yesterday. Again, calculate today’s average price by dropping a day off backend.

\[ TAP = \frac{(CP \text{ of today} + CP \text{ of 199 PTD})}{200} \]

Where \( TAP \)=Today’s Average Price

Again mark that point on a chart and make another dot by considering the closing price of today. This can be repeated every day and both the values can be plotted. It shows the direction of the market and measures the trend. Extremely high readings are a warning that the market may soon reverse to the down side whereas very low readings signify the reverse. The shorter the moving average, the sooner there will be a change in the market. It is used to determine whether the stock is technically healthy or not. The percentage of stocks above the 200- day moving average helps to determine the overall health of the market.

For understanding the moving average concept assume that we are using a 5-day moving average. A 5-day simple moving average can be calculated as sum of the closing stock prices of those prices divided by 5. The name is moving average since the averages move in the sense old data is scraped as new data becomes available.

Let the closing prices of a company be as given in the table 6.2:
Table 6.2 Closing Prices

<table>
<thead>
<tr>
<th>Date</th>
<th>Closing Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1/2015</td>
<td>100</td>
</tr>
<tr>
<td>2/1/2015</td>
<td>110</td>
</tr>
<tr>
<td>3/1/2015</td>
<td>120</td>
</tr>
<tr>
<td>4/1/2015</td>
<td>130</td>
</tr>
<tr>
<td>5/1/2015</td>
<td>140</td>
</tr>
<tr>
<td>6/1/2015</td>
<td>150</td>
</tr>
<tr>
<td>7/1/2015</td>
<td>160</td>
</tr>
</tbody>
</table>

The 5-day moving averages for the three days i.e. 5/1/2015, 6/1/2015, 7/1/2015 can be calculated as below.

5-day Moving Average on 5/1/2015: \((100+110+120+130+140)/5=120\)

5-day Moving Average on 6/1/2015: \((110+120+130+140+150)/5=130\)

5-day Moving Average on 7/1/2015: \((120+130+140+150+160)/5=140\)

6.6.3 Technical and Sentiment Analysis combined

The result of the previous two analyses has to be combined. The two percentages, i.e., the percentage of time that the closing price is higher than the 200-day moving average and the percentage of the positive news articles, are combined in a 60-40 ratio to give the final percentage.

The final percentage that is obtained from the previous step is then checked to give the final outlook for the company. Depending on the percentage we can say if the company has been
consistently seeing an uptrend and the news articles pertaining to the company are on the positive side.

If so, then it is safe to invest in the company as there is a positive outlook for the company. On the other hand, if the percentage signifies that the company has been seeing a downtrend, then it can be concluded that there a negative outlook and there is a risk in investing in the company. This helps the user take the correct decision while investing in a particular company.