CHAPTER IV

METHODOLOGY

4.1 INTRODUCTION

The methodology is the strategy of the study of how to build a system. The details of how to build the system to accomplish the given objectives are specified in this chapter. This chapter gives research methods and the strategy implemented.

4.2 METHODOLOGY

The overall process of predicting the direction of the stock market consists of various steps which include data collection, text pre-processing and feature selection. In order to carry out the overall research, the programming is needed. R language, python, Java was used. The packages of those tools were used to implement the proposed algorithms.

The data needed for our problem is of two types. The historical stock prices and the news articles from which the sentiments are to be extracted. In contrast to the other systems, which used the static data, our system is based on the streaming data as well as the static data. The crawler crawls on the specified website and extract the news articles of the specified company for which the future direction of the stock is to be predicted. Since the stock prices have to be correlated with the news articles, the news articles have to be extracted along with the timestamps. These news articles are then served as the input to the sentiment analysis module.

Since many years, the researchers have been exploring avenues in sentiment analysis and they came up with many different algorithms for the classification of sentiment of the text. Every algorithm has some advantages and drawbacks. The choice of the algorithm for sentiment analysis may depend on the available datasets, domain and prior experience. Among the approaches, linguistic based, lexicon based and machine learning, one approach is to be chosen. After choosing the approach, one has to decide the appropriate algorithm in that approach.
4.3 WEB CRAWLER

The data available in the websites can be extracted automatically using a web crawler which can also be termed as a web robot or a web spider.

![Components of web crawler system](image)

**Figure 4.1: Components of web crawler system**

(Source: Aviral, 2014)

The web crawling algorithm takes an input of seed URL list and executes the following steps, repeatedly.

Step 1: A URL can be removed from the URL list.

Step 2: The concerned page is downloaded.

Step 3: The Relevance of the page downloaded is checked.

Step 4: If it contains any links, they are extracted.

Step 5: These links are added back in the list of URLs.

Step 6: After processing all URLs, the page that is most relevant is returned.
4.4 DATA SETS

The decision of what data set is being used for the research is very crucial. There are not readily available data sets for our system. Most of the data is streaming data that is mined and kept in the database. Our system needs two kinds of data. One pertaining to the historical stock values and the other data set that contain news articles that are published online. To study the correlation between the news articles and stock prices the data used is from a combination of two different sources: a dataset of the historical data and a corpus of news articles. The initial source of data, used for the extraction of the news articles is the moneycontrol website, which has a large reservoir of crucial news catering to the individual stocks. Moneycontrol is India’s leading financial information source. We have extracted historical values from http://ichart.finance.yahoo.com for the year of 2012 for the stock of Infosys in NIFTY. To predict the overall health of a company, the historical stock values are extracted for a period of one year from a website in.yahoo.finance.com. The data extracted is in a CSV format. This data is then placed into a database table which can then be queried and processed. For the news articles, the website moneycontrol.com has been used. To predict future prices based on the sequence of events, 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 each company’s open price, close prices for the years 2007 to 2014. Approximately 18 lakh price points from BSE have been collected. Events from disclosure records and content pieces collected from indiatimes.com, moneycontrol.com, sebi.com, watchoutinvestors.com, ecourts.gov.in, cibil.com are used as corpus for this system.

4.5 DATA PREPROCESSING

The accuracy of Sentiment Analysis can be improved by choosing the appropriate data pre-processing technique. This fact makes the data pre-processing step very crucial. Some of the news articles need specific pre-processing techniques apart from the standard pre-processing techniques since the content generated by the user community for example Twitter generated messages. Data pre-processing reduces the word space significantly, but there are chances of the loss of information also.

The steps involved in data pre-processing are:
i. Tokenization: A basic strategy could be to split the news articles into all non-alphanumeric characters. There may be a chance of information being lost, advanced techniques are needed for text tokenization. This step would be domain dependent.

ii. Dropping of the words which are common: A word which has no or little information value, that is common, can be identified (For eg: an, is, be and in etc.). The text space would be reduced significantly and only the data which contribute most for the sentiment identification remains in the text.

iii. Normalization: It is a process of creating the terms’ equivalence classes. Example INFOSYSS and INFY.

iv. Lemmatization and Stemming: News articles may have different word forms.

4.6 SENTIMENT ANALYSIS ALGORITHM

After the news articles are extracted the next step is to extract the sentiment which gives an overall opinion of whether that article is negative, positive or neutral. The sentiment analysis algorithm attempts to identify the opinion/sentiment that a news article may hold towards a financial company.

As shown in the Figure 4.1 the steps involved in Sentimental analysis are:

Step 1: Gathering data from internet is solely based on the (SOR) Subject of Reference (e.g. ICICI bank). We use web mining techniques (ex. crawler) to gather all web pages where the SOR is mentioned.

Step 2: Text Extraction can be done using several data mining or text mining techniques starting from simple 'keyword matching' to 'DOM structure mining' to 'neural networks' methods. The major challenge here is that web documents are highly unstructured and no single method can give 100% clean text extraction for all documents.
Step 3: Text Cleaning is mostly heuristic based and case specific. By this what we mean is to identify the unwanted portions in the extracted contents from Step 2 with respect to different kinds of web documents (e.g. News article, Blogs, Review, Micro Blogs etc.) and then write simple clean-up codes based on that learning, which will remove such unwanted portions with high accuracy.

Step 4: Once we have the data corpus of clean documents from the previous steps, it is then put to the various knowledge processing engines. This data can be analysed for feature analysis or business analytics or market research or consumer buzz trends or consumer sentiment analysis etc. depending on the need. Various techniques are used for each such purpose. For example, Inverse Document Frequency (TF-IDF) technique can be used to gather a pool of various features for the SOR, which in case of ICICI bank can get a pool as customer service, credit card, recovery agent, customer satisfaction etc. This is called feature analysis. From this pool one can determine things like how many consumers talk about recovery agent while talking about credit cards compared to how many consumers talk about customer satisfaction while talking about credit cards.
Step 4.1: Obviously, the above technique for Feature analysis will throw up few unwanted features in the pool which need to be removed from final analysis. This can be done by feature mapping from pool with keywords representing the various well-known features. These features can also be pre-defined by the SOR.

Step 5: Sentiment Analysis is done on the document to categorize as Positive, Negative or Neutral. Sentiment analysis can be done on the clean extracted web documents in two manners - manual rating or automated rating of such web documents. While manual rating is a near perfect method to do it, but it is a slow process when the volume of web documents is too high. Whereas automated system will be much faster method, but is bound to lack accuracy since it is effectively machine learning and deriving human sentiments through user generated content. Also, the language barrier is a major challenge for automated sentiment analysis.

Nevertheless, extensive research work on Natural Language Processing has addressed such challenges well and reasonably high performance machine learning techniques have evolved which can do sentiment analysis of web documents. The most efficient techniques are Naive Bayesian Classifier (NBC) and Support Vector Machines (SVM) Maximum Entropy. Apart from these the lexicon based techniques are also available. These learning algorithms require a learning corpus to first train on then from that training it can derive sentiment from web documents.

Step 5.1: Training NBC/SVM is fairly straightforward and well-studied technique, in which, a manually rated corpus of several thousands of web documents is generated, categorizing them as negative, positive or neutral for a given SOR. This corpus is then fed to the NBC/SVM engines which generate several measuring parameters for a given document to fall into one of the three categories [negative, positive or neutral].

Step 5.2: Once the NBC/SVM engines are trained, they can now be used to categorize rest of the web documents using the parameters generated by them. The accuracy definitely won't be 100% but several layers of training and tuning can increase and optimize the accuracy.
Once the sentiment is extracted these values would be combined with the historical values and the direction of the stock market is predicted. The techniques like Artificial Neural Networks and NBC are used for the prediction purposes, as mentioned already.

The artificial neural network uses five input neurons, two hidden layers and one output neuron. The historical prices and the sentiment score are the inputs to the system and the network learns from the training. It produces the next day price based on the input historical stock values and sentiment score.

In NBC, based on the probability of features, the classifier is trained and the classification of news articles as positive, negative or neutral is learnt.

4.7 PERFORMANCE EVALUATION METHODS

The main ingredient of our system is sentiment analysis of the news articles. The performance of the sentiment analysis algorithm can be measured by calculating its accuracy. Three other metrics are precision, recall and F-measure.