CHAPTER 2

Literature Survey

2.1 Paper title: "Predicting Stock Movement using Sentimental Analysis of Twitter Feed".

2.1.1 Objective

The main objective of this paper is the proposed system going to use a sentiment tagged Twitter dataset of 1.5 million tweets, collected from Sentiment140 for sentiment classification. We have filtered the tweets by some regularly used and some unique parameters to convert them as close as possible to plain text. Then, we used Boosted Regression Tree classifier for predicting next day's stock movement with present day's tweets containing 'stock market', 'stocktwits', 'AAPL'.

2.1.2 Methodology

The methodologies are as follows

2.1.2.1 Data

Our dataset contains, as mentioned earlier, 1.5 million (1, 578, 614) hand-tagged tweets, collected through Sentiment140 API. The tweets are tagged '1' and '0' for being 'positive' and 'negative' respectively. Then we performed a random split over the dataset, to divide the dataset into training dataset, containing 1, 499, 337 tweets and testing dataset, containing 79, 277 tweets.

	Training	Testing
Positive	750, 527	39, 651
Negative	748, 810	39, 626

Table 2.1 Data distributions

2.1.2.2 Data Pre-processing

The tweets have gone through following pre- processing steps-

(i) All characters have been changed to lower case letters to avoid repetition of same words in feature vector.

(ii) All URL, starting with <http://>, <https://> and <www. > have been stripped off from thetext.

(iii) Consecutive whitespaces are replaced by single whitespace. Apart from that, more than two consecutive characters, symbols, punctuation marks have been replaced by just two ofthem.

(iv) Usernames (starts with <@>) are removed. Hashtags are kept stripping off the <#> sign. Emoticons are kept thesame.

(v) Apart from these, all retweet signs ('RT'), single (<'>) and double (<">) quote signs are stripped off. All non-ascii characters are removed too.

2.1.2.3 Validation

GraphLab Create has been used to work with this large amount of data. 5% of the training data has been used for validation. These results in table 4, came from Sentimet140 dataset, by applying different classification algorithms along with unigram, bigram, trigramapproach.

Model	Unigram		Bigram		Trigram	
	Train	Test	Train	Test	Train	Test
Logistic Regression	0.84	0.78	0.98	0.76	0.98	0.74
SVM	0.84	0.79	0.98	0.76	0.98	0.74
Decision Tree	0.64	0.64	0.52	0.52	0.52	0.52
Boosted Tree	0.55	0.55	0.60	0.60	0.55	0.54
Random	0.64	0.64	0.53	0.53	0.53	0.53

Forest						

Table 2.2 n-gram (n=1, 2, 3) model outputs

2.1.2.4 Stock Movement Prediction

From the results of sentiment analysis, we found that SVM worked best on our test data of Sentiment140, along with unigram feature. So, for sentiment analysis of the stock related tweets, we used SVM with unigram approach where Sentiment140 data set is our trainingdata.

Tweets containing 'stock market', 'stocktwits' were trained to predict general stock market points (DJIA) closing difference. Besides, tweets containing 'AAPL' were trained to predict Apple Inc. closing stock point difference. Tweets were ignored of the days when the stock market or Apple's stock-prices was closed.

2.1.2.5 SCORE OF 'STOCK MARKET', 'STOCKTWITS', 'AAPL' RELATED TWEETS

we have used marginal values that came as an output from running Support Vector Machine (SVM) on tweets containing 'stock market', 'stocktwits', 'AAPL'. As SVM is basically a classifying algorithm that draws the best hyperplane or margin to separate classes, marginal values are distance from the hyperplane to a data point.

We took the average marginal value of these tweets for each day as we had at most a thousand tweets for each day. Positive marginal value refers to positive sentiment and negative marginal value refers to negative sentiment.

2.1.2.6 STOCK INDEX VALUE PREDICTION USING BOOSTED REGRESSION TREE

We used the Boosted Regression Tree model for the stock index value prediction. Training set is data from January to August 2016 and testing was done on stock related data from September to December

2016. The marginal values of tweets containing 'stock market', 'stocktwits' are trained with DJIA closing price difference. Whereas, tweets that contain 'AAPL'; their marginal value is trained with APPLE Inc. closing stock pricedifference.

We trained our model to predict stock price difference of the next day. In the training data set average marginal value are of a day's tweets and their corresponding closing price difference is between that day and the next day. So that, after getting marginal value of tweets of present day, we can predict how much stock market will rise or fall the next day. In other words, for present day's stock value prediction we will need previous day's tweets average marginalvalue.

2.1.3 Architecture

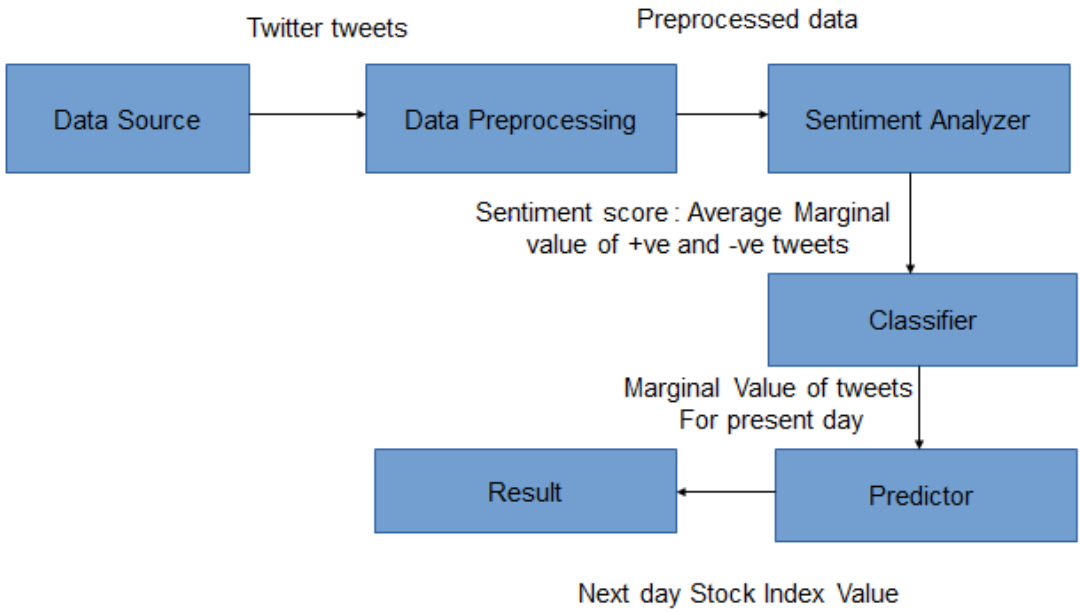

Fig 2.1 Architecture of proposed system

2.1.4 Limitations

- Extremely high and extremely low stock indexes are cannot be predicted using Boosted Regression Tree.

.

2.1.5 Results

Plotted both actual stock difference and predicted stock difference of testing period (September to December 2016).

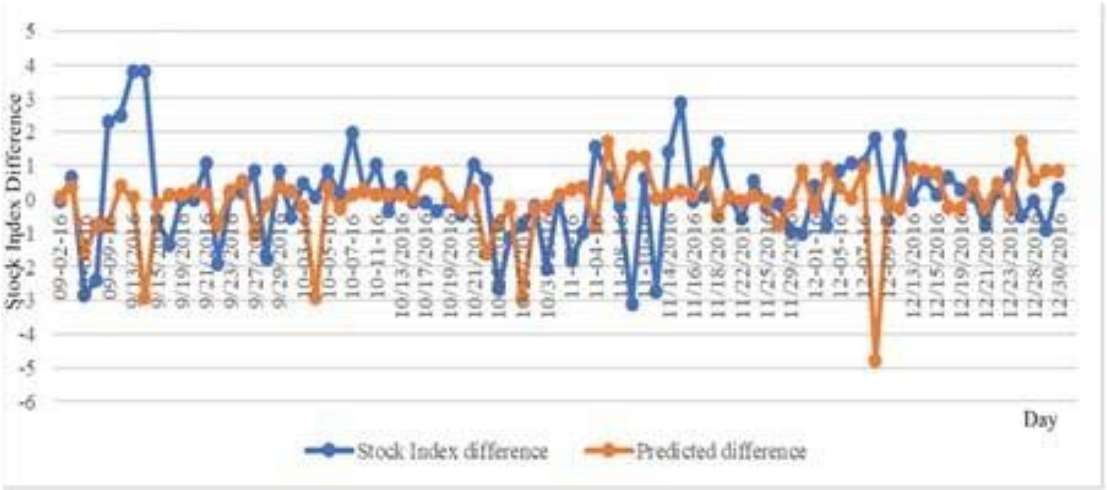

Fig 2.2 Company specific ('AAPL') closing stock price difference prediction using tweets containing 'AAPL' with Boosted Regression Tree

For fig.1, time is on the x-axis and difference of stock points are on the y-axis.

2.2 Paper title :"Using Sentimental Analysis in Prediction of StockMarket Investment".

2.2.1 Objective

The objective of paper is to make sentimental analysis is performed on the data extracted from Twitter and Stock Twits. The data is analyzed to compute the mood of user's comment. These comments are categorized into four category which are happy, up, down and rejected. The polarity index along with market data is supplied to an artificial neural network to predict the results.

2.2.2 Methodology

The methodology as follows

2.2.2.1 Artificial Neural Network

An Artificial neural network is a computational model which is based on the structure and functions like a biological neural network. Information that is supplied within a network affects it as it changes or learns with that information. An artificial neural network is a structure which is similar to a human brain. It contains highly interrelated structure of neurons. These neurons act as an input which activates the system. Thus an artificial neural network is mostly preferred by the researchers to use in problems which involves computational tasks, analysis, finding similarities and much more. Data used in our research work was primarily collected from stock dedicated website i.e Stock Twits. All the tweets were mined from Stock Twits on which sentimental analysis were performed. For the market data, yahoo acted as our means of source. The data extracted from yahoo consists five parameter as indexes which involves opening, closing, high, low and volume, out of which only first four were our target parameters.

2.2.2.2 Sentimental Analysis

In our work we have taken four parameters which are happy, up, down and rejected. The score of the data is calculated in the range of [0,1] . This score is added to csv sheet of each company and is given to an artificial neural network

to train and predict the closing value. The output prediction parameter is calculated through the daily return of investment method.

2.1.3 Architecture

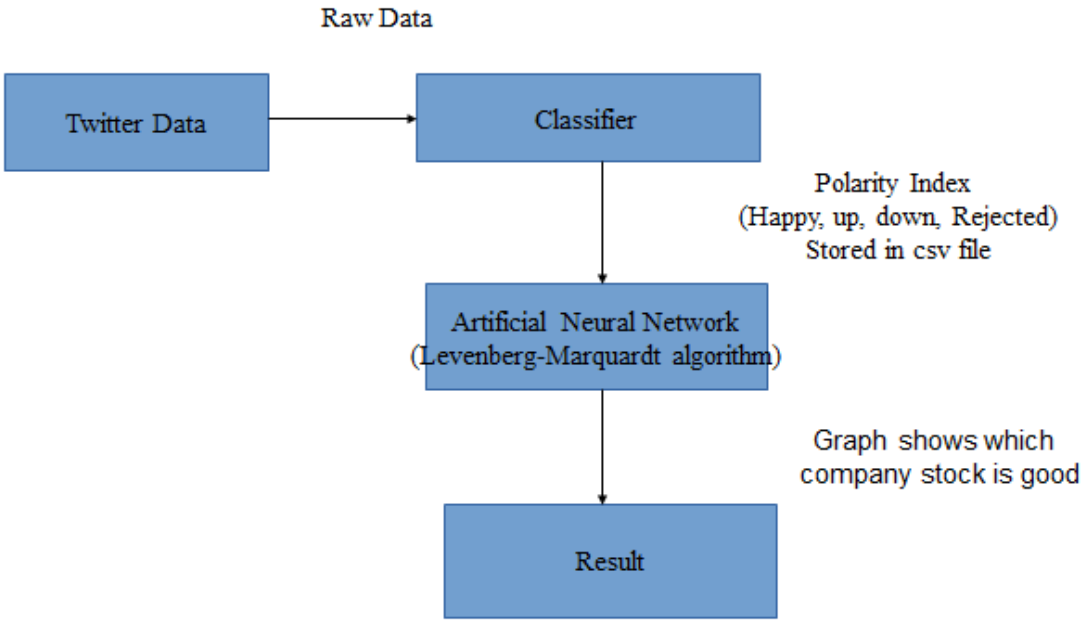

Figure 2.3 Architecture of the Proposed System

2.2.4 Limitations

- Shorter period stock prices gives better Output

2.2.5 Results

An artificial neural network is an effectual tool which helps in prediction and performing sentimental analysis because of its structure which represents the structure of human brain. The sentimental score with market values were provided to an artificial neural network to predict the future market value

Company Name	Parameters				Score
	Happy	Up	Down	Rejected	
Apple	0.54	0.56	0.35	0.37	0.46
Google	0.20	0.62	0.68	0.06	0.39
Microsoft	0.46	0.50	0.78	0.62	0.59
Oracle	0.44	0.30	0.42	0.60	0.44
Facebook	0.92	0.82	0.88	0.78	0.85

Table 2.3 SENTIMENTAL SCORE OBTAINED

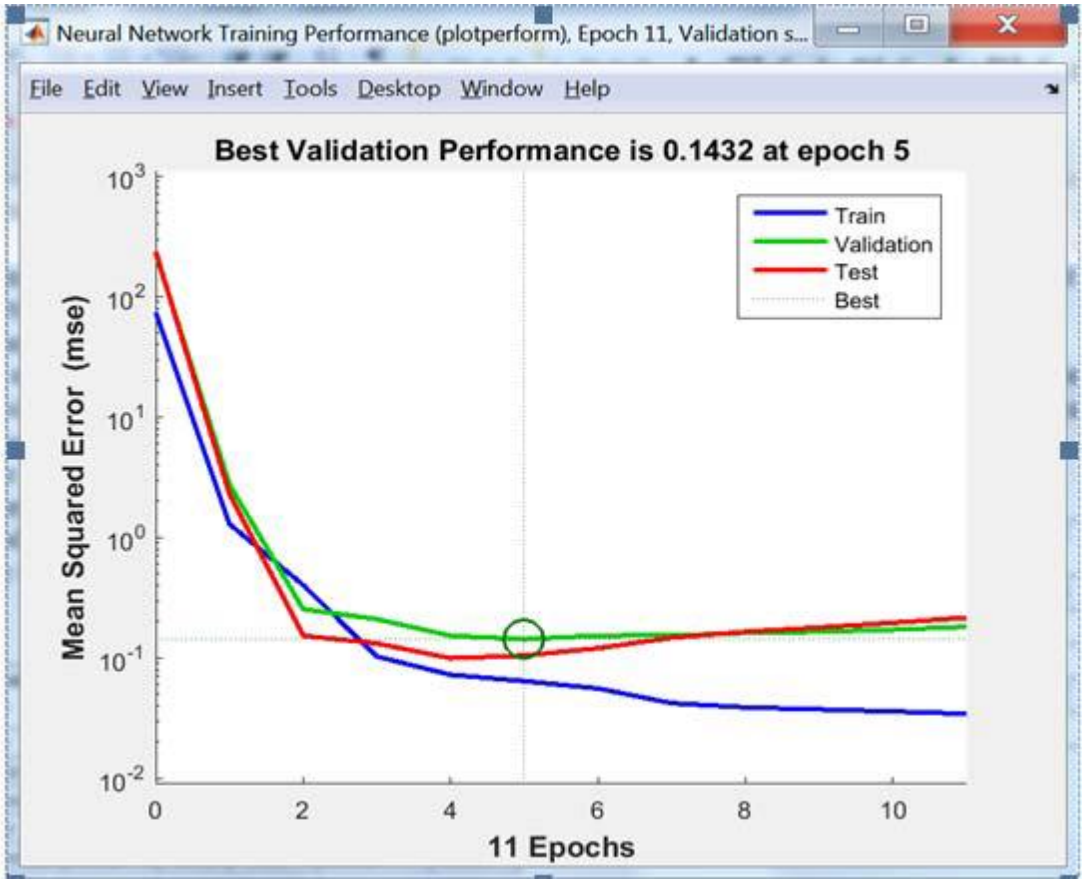

Figure 2.4 Implementation of Apple's data

Above figure indicates the implementation of Apple data with minimum mean square error.

2.3 Paper title: "Market Impact Analysis via Sentimental Transfer Learning".

2.3.1 Objective

The objective of the proposed system is to propose sentimental transfer learning to transfer the knowledge learned from news-rich stocks that are within the same sector to the news-poor stocks. News articles of both kinds of stocks are mapped into the same feature space that is constructed by sentiment dimensions. New predictors are then trained in the sentimental space in contrast to the traditional ones. Experiments based on the data of Hong Kong stocks are conducted. From the early results, it could be seen that the proposed approach is convincing.

2.3.2 Methodology

The methodology as follows

2.3.2.1 Sentimental Transfer Learning

The purpose of the transfer learning is to transfer the knowledge from a source domain to a task in a target domain. Inthe survey by Pan and Yang, transfer learning approaches can be categorized into four groups, among which the idea of instance transfer technique is to reuse part of the instances in the source domain and help the learning task in the target domain, and the idea of feature transfer learning is to learn a feature representation to "better" represent the data in both the source domain and the target domain. In this paper, sentimental transfer learning is a combination of those two techniques:

1) It reuses the news articles of the news-rich stocks and
2) It constructs a sentimental feature space and maps the instances of both the news-rich and news-poor stocks into the same space.

2.3.3 Architecture

The work flow of the sentimental transfer learning is illustrated in Figure. In the first step, the news articles of the news-poor stocks are divided into two parts, the training part is mixed with the news of the news-rich stocks and the other part is left for testing. In the second step, words in the news articles are searched in the sentiment dictionary and projected onto sentimental dimensions if they have any affective

aspects. As the sentimental dimensions are fixed in the dictionary, each word can be represented by a sentiment feature vector of the same length. Thus, each news article can be represented by a sentiment feature vector by summing up all words' vectors. In the third step, all the preprocessed instances are fed into the machine learning model. After training and testing, the model will be evaluated in the final step.

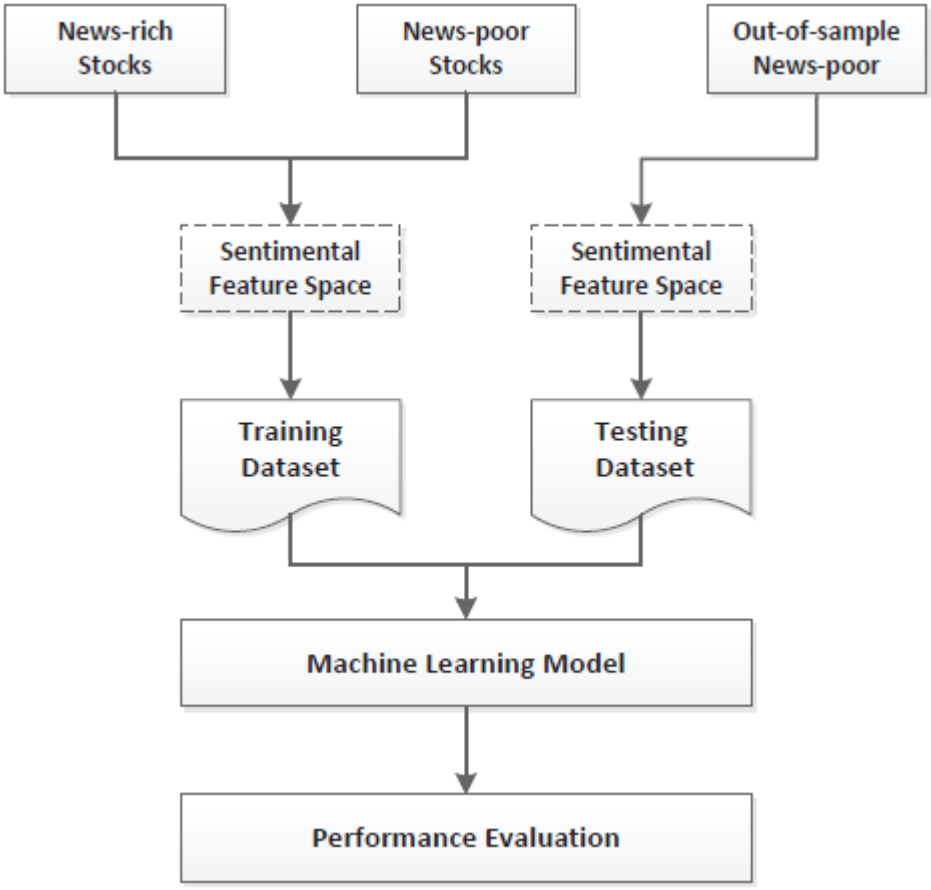

Figure 2.5: Workflow of sentimental transfering

2.3.4 Limitations

- Doesn't predict the stock trends.
- Only useful for transferring sentiments.

2.3.5 Results

A news archive from FINET2 is employed. The news archive contains both company-specific and market related news from Jan. 2003 to Mar. 2008. Each piece of news is tagged with a time stamp showing the time the news is released, which helps classify

news by dates. Stock codes of companies that are mentioned in the news are listed at the end of the article, which helps establish the mapping from the news articles to stocks and vice versa. The data set is split into three parts for different purposes: 1) from Jan. 2003 to Dec. 2005 is the training data set; 2) from Jan. 2006 to Dec. 2006 is the validation data set; and 3) from Jan. 2007 to Mar. 2008 is the independent testing data set.

2.4 Paper title: "Capital Market Forecasting By Using Sentimental Analysis".

2.4.1 Objective

The objective of this paper is to demystify the stock market to help the customers to get better idea of upcoming market trends, polarity of stocks and present market sentiment in regards to the company he is investing.

2.4.2 Methodology

The methodology as follows

- Stock Market Prediction involves requirement of previous years of stock data of a particular company, which we have fulfilled by retrieving it from a finance website. After retrieval, this data is staged in R studio and converted into a dataframe. From the dataframe, appropriate column and rows are selectedfor running algorithms on it depending on amount of duration of data we want to take for running the algorithm.
- For predicting future price of stock by exponentially weighting the data (close price) is done by first selecting the number of days of data to be taken as input then giving the highest weight to the latest data and decreasing the weight to older ones with farthest data getting the lowest weight. The weights depend on algorithm and the data for all the days are first multiplied withtheir corresponding weights and then performing their average.
- For predicting future price of stock by simple averaging the data (closing price) is done by initially selecting the number of days of data to be taken as input and then adding all the data then performing their average.

- For predicting next day's polarity of a stock whether the stock price of a company will go up or down classification can beused. We have used KNN algorithm that stores all availablecases and classifies new cases based on a similarity measure. Ituses majority vote of its neighbors, with the case beingassigned to the class most common amongst its K nearest neighbors measured by a distance function.
- For predicting next day's polarity of a stock we have also usedSVM algorithm which uses machine learning and builds amodel that assigns new examples into one category or theother, making it a non – probabilistic binary linear classifier.
- For predicting market sentiment of the company, we have usedsentiment analysis which uses natural language processing andtext analysis to identify and extract required information fromour source of data.

2.4.4 Limitations

- Take too much time to give results as it considers whole 10 days of data.

2.4.5 Results

For instance, took the stock price for a company whose actualshare price on a particular day was **Rs. 105.80.** Then we testedthe algorithms on previous days of stock data and calculatedresults which are:

- For predicting future price of stock by exponentially weightingthe data, our algorithm result was **Rs. 106.47.**
- For predicting future price of stock by simply averaging thedata, our algorithm result was **Rs. 106.015.**
- Using KNN algorithm, our predicted result was a confusionmatrix with over 90% accuracy.
- Using SVM algorithm, our predicted result was a confusion matrix with close to 60% accuracy.
- The sentiment score was found to be **0.412**, which is a positive market sentiment of the company.

Hence, we predicted the company's next day close price, its polarity and the market sentiment of the company. All three of these information can be used by the investor

to get a better insight of next day's trends before buying or selling of company's stocks.

2.5 Paper title: "Sentiment Analysis for Effective Stock Market Prediction".

2.5.1 Objective

The objective of the proposed system claims is that the sentiment analysis of RSS news feeds has an impact on stock market values. Hence RSS news feed data are collected along with the stock market investment data for a period of time. Using our algorithm for sentiment analysis, the correlation between the stock market values and sentiments in RSS news feeds are established.

2.5.2 Methodology

The methodology as follows

2.5.2.1 RSS stock news feed

From the web pages, RSS feed reader reads the required news content such as title, description, date, author, link etc. in the format of XML document. This takes latest headlines from the stock related news website. The final information of description (sentence level) is grabbed after setting up the site parameter. Finally it parses the XML document from RSS feed list. This RSS feed helps to collect the stock market news as a dataset.

2.5.2.2 Pre-processing

In this process, it removes the incorrect, incomplete, improperly formatted, or duplicated data. Dirty data can cause confusion in the data set. Hence, this module cleans the data by filling missing values, smoothing the noisy data, identifying and removing the outliers. After pre-processing, the data are passed to the next module.

2.5.2.3 Sentence splitting module

The sentence splitting module is the one which splits the cleaned news data into parsed sentences. The parsed news data are collected in a text document for the testing

purpose. The document contains the RSS news data in the form of sentence by sentence.

2.1.2.4 Sentence level sentiment score (SSS) algorithm

SSS Algorithm is considered for finding the overall result where for every individual sentence the analysis approach is applied and finally their results are summarized to provide the overall result of the document.

In general, the score ranges from 0.0 to 1.0 and their sum is 1.0 for each synset. Here, initially the POS tagger is applied for each word and it specifies the tag as noun and adverb in such a manner which are equivalent to words. Assign the score value to each word of a sentence and find the sum of that sentence. If the score value of that sentence is positive then that sentence is considered as a positive sentence. If score value is negative then it is considered as negative sentence. If it is 0.0, then it is considered as a neutral sentence.

2.1.2.5 Stock market prediction

To predict the stock market, the results of both sentiment analysis and Sensex are combined and analysed. Table 1 shows the final result prediction technique for Stock market.

If Sentiment analysis results and Sensex Moving Average results are positive then Final prediction is positive. If both are negative, then result is also negative. Combinations of both will results into neutral.

Sentiment Analysis Result	Sensex-Moving Average Result	Final-Result Prediction
Positive	Positive	Positive
Positive	Negative	Neutral
Negative	Positive	Neutral
Negative	Negative	Negative

Table 2.4 Sentiment and Sensex-Moving Average Final Result Prediction

2.5.3 Architecture

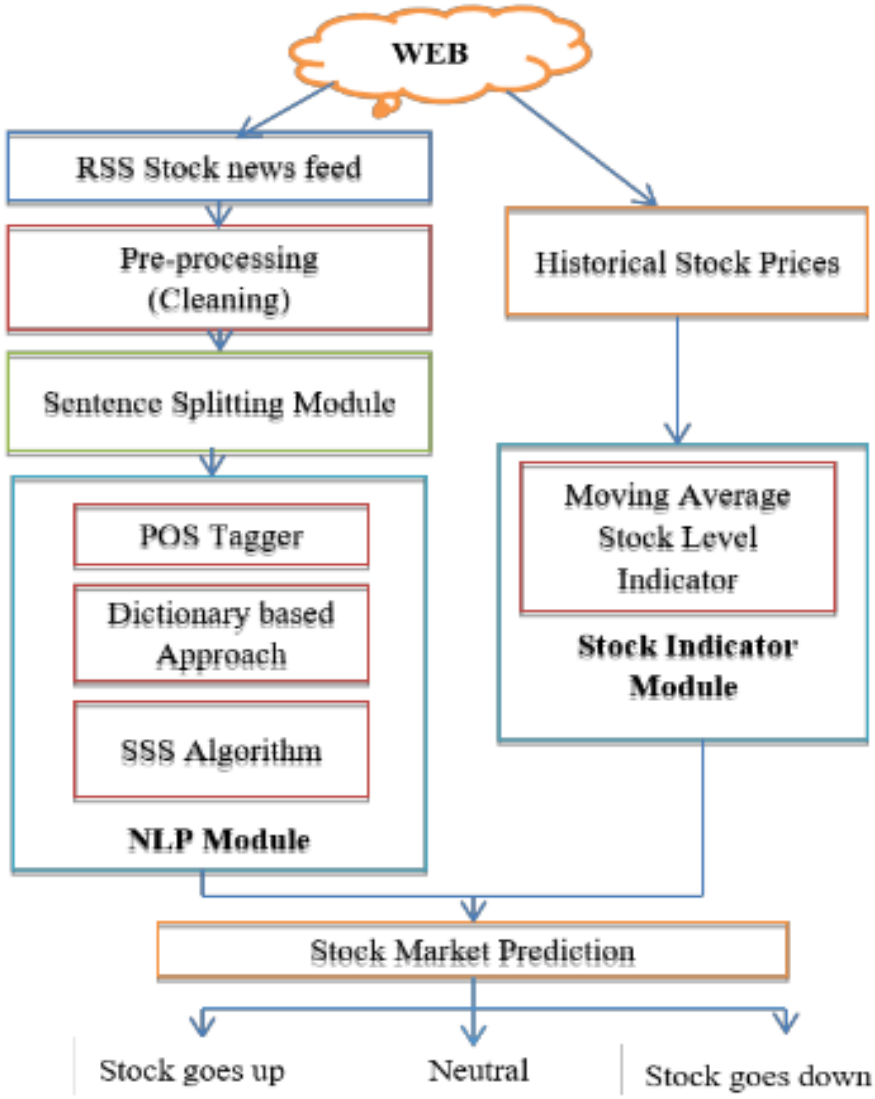

Figure 2.6: Architecture of Proposed system

The CIS detect the color change of the array strip and sends it to the MCU. After that, it transmits the result to the smartphone and Bluetooth, and generates a warning signal in the signal processing part according to the degree of gas detection. The array strip is attached to the slot next to the array reader and mounted in the reader.

2.5.4 Limitations

- Doesn’t consider real time data.

2.5.5 Results

In the experimental study the stock market forecasting is collected for the company ARBK from Amman Stock Exchange (ASE). The Oracle database of Amman Stock Exchange (ASE) contains the historical prices of the 230 companies listed in the exchange from the year 2000. The historical prices are collected from the year 2005 to 2007. The performance of the proposed algorithm is compared with the prediction of stock market using data mining techniques.

The sentiment RSS news feed for Arab Bank (ARBK) Company is collected from http://investing.einnews.com/news/ase-stock and then the sensex point for the same company is collected from http://www.marketstoday.net/markets/jordan/Historical-Prices/10/en/#. The moving average is calculated for the month of April 2006. Same way sentiment is also calculated for April 2006. Finally both the results are combined and the end result is predicted for stock market according.

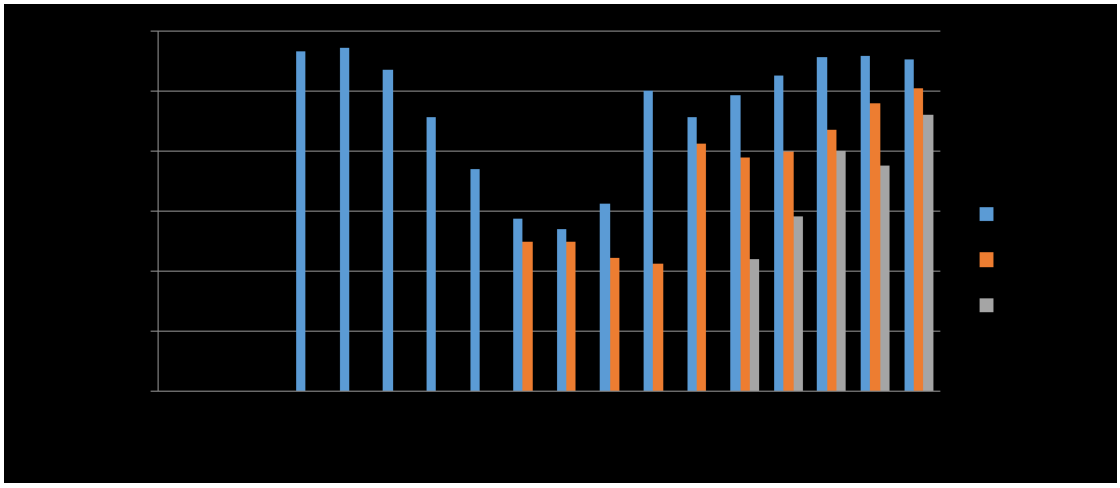

Figure 2.7 Moving Average for April 2006

The following Fig.6 (chart) shows sensex point –moving average calculation. In the below graph 5-day moving average, 10-day moving average and 15-day moving average are shown in blue color, red colour, and green color respectively.

2.6 "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis".

2.6.1 Objective

The main objective of this paper to analyze the Sentiment resulting from the product reviews using original techniques of text's search. These reviews can be classified as having a positive or negative feeling based on certain aspects in relation to a query based on terms. The proposed system to identify product reviews offered by Amazon.

2.6.2 Methodology

The proposed system takes a hybrid approach to solve the problem of classification and possesses the capabilities of Random Forest and SVM at the same time.

- Firstly, random forest is an ensemble learning method that construct a number of decision trees at randomly selected features and predict the class of a test instance by voting of the individual trees.
- Support Vector Machine revolves around the notion of a margin either side of a hyperplane that separates two classes. Maximizing the margin and with this way creating the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error. RF was not sensitive to input parameters, thus, we just used the default parameters for each classifier. The trained classifiers return scores between 0 and 1, these scores are then transformed to a binary state indicating 'negative' or 'positive'. For each combination, the existence of element is considered positive (P) or negative (N). The notation of TP indicates True Positives: number of examples predicted positive that are actually positive, FP indicates False Positives: number of examples predicted positive that are actually negative, TN indicates True Negatives: number of examples predicted negative that are actually negative and FN indicates False Negatives: number of examples predicted negative that are actually positive. The classification metrics considered for the sentiment analysis are Accuracy, Precision, Recall and F-Measure and these parameters are evaluated based on the calculated positivity and negativity of reviews by the proposed hybrid approach.

- The performance evaluation of classifiers is made according to the following formulas: Report of the true positives. It corresponds to:

$$TP\ Rate = \frac{TP}{TP+FN}$$

- It is thus the report between the number of positive instances classified well and the total number of elements which should be classified well. Report of the false positive one. He corresponds, symmetrically in the previous definition:

$$FP\ Rate = \frac{FP}{FP+TN}$$

- Accuracy is a common measure for the classification performance and it's proportional of correctly classified instances to the total number of instances, whereas the error rate uses incorrectly classified rather than correctly.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

- This quantity allows to group in a single number the performances of the classifier (for a given class) as regards Recall and the Precision:

$$F\text{-}Measure = 2 * \frac{Recall * Precision}{Recall+Precision}$$

2.6.3 Control Flow Diagram

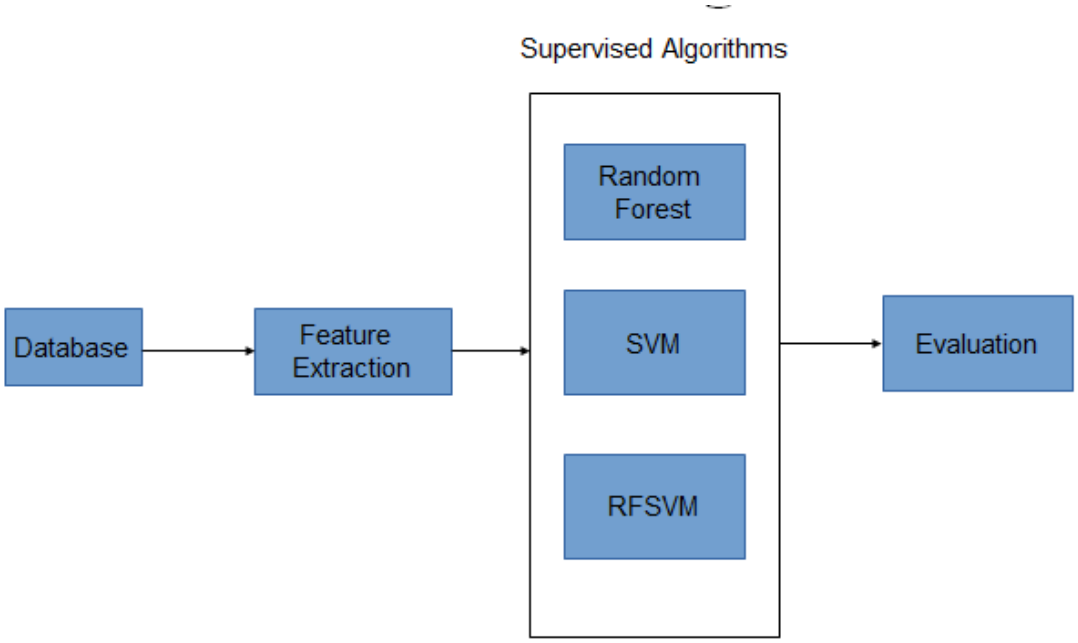

Figure 2.8 Control flow of system

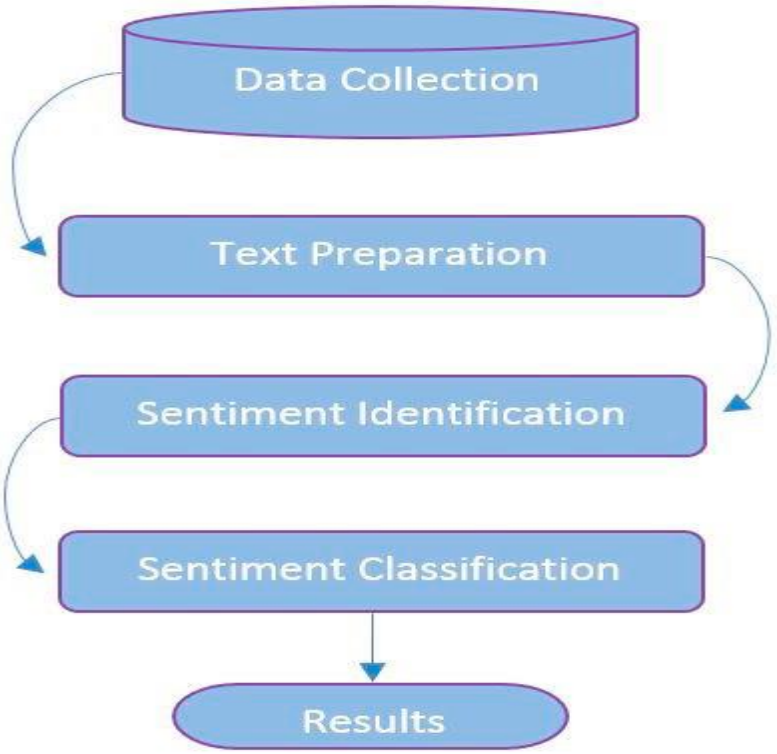

Figure 2.9 Sentiment analysis model

The sentiment analysis process is shown in figure 1. The text preparation step performs required text pre-processing and cleaning on the dataset which including removal of stop words. Sentiment identification step determines the sentiment of people expressed in the text and analyzes it. Finally, sentiment classification is conducted to get the results.

2.6.4 Limitations

- Doesn't give more weight to the real time reviews
    - Real time reviews provide more accurate information about product
    - Product may be performing better at present.

2.6.5 Results

The training and test data we used for this work were taken from the "Amazon" which contains 1000 instances divided into positive (500) and negative (500). In this article, Cross Validation method with fold value equal to 10 has been used for training and testing phases We will use some techniques that automatically extracts this data into positive or negative sentiments. By using the sentiment analysis, the customer can know the feedback about the product before making a purchase. Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product.

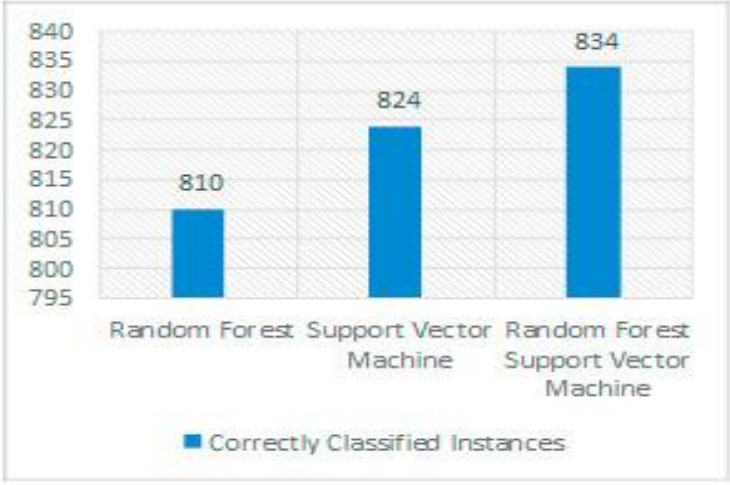

Figure 2.10.Number of correctly classified instances

From this table, it is represented that the accuracy computed in the case of proposed method (RFSVM) is better as compared to random forest and support vector machine.

2.7 Paper title: "Analysis of Various Sentiment Classification Techniques".

2.7.1 Objective

The objective of this paper is to present recent updates on papers related to classification of sentiment analysis of implemented various approaches and algorithms.

2.7.2 Methodology

Sentiment analysis process is as showed in below figure.

- Customers post their review in comment, forum or blog.
- These reviews are in form of unstructured data so first unstructured dataset is converted into structured form.
- Then extracts features from structured review using feature selection method then classification technique is applied on extracted features to classify them into its sentiment polarity that is namely either positive or negative.

2.7.3 Architecture

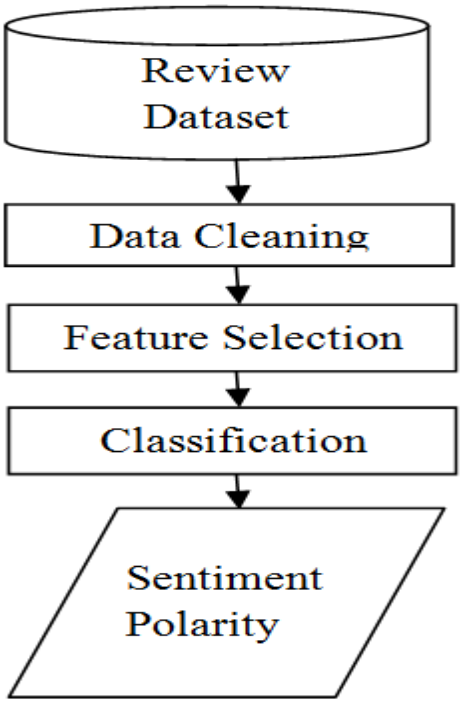

Fig 2.11 Architecture of Sentiment Analysis

2.7.5 Results

Unigram improves performance than other features but still accuracy is issue because classifier performs average classification on dataset and accuracy is comparatively low than topic based categorization. POS tagging is suggested to increase accuracy. Proper selection method is not used and some meanings conveyed are not captured. SVM performs better than Naïve Bayes. By using KNN, maximum entropy classifier and stochastic gradient classifier can improve accuracy than present SVM. Unigram performs very well. Various feature selection technique is applied but ensemble of feature selection can further improve accuracy and unigram with bag of word gives best accuracy. POS tagging identifies tagging of word and produces improved result. Ensemble of algorithms improves performance. Accuracy is improved by considering emotions as noisy label in twitter dataset and use of WordNet dictionary generates better result than SVM, Maximum entropy and Naïve Bayes. Accuracy can be still improved by doing careful feature selection and proper classification technique. Bag of word produce good accuracy compare to feature hashing but take more computational effort than feature hashing. Feature hashing takes less computational effort compare to bag of word but it is less accurate then bag of word.