CHAPTER 4

SENTENCE LEVEL SENTIMENT SCORE (SSS) ANALYSIS USING RSS NEWS FEEDS AND STOCK LEVEL INDICATOR

4.1. INTRODUCTION OF RSS NEWS FEEDS

Opinion Mining is the method of identifying various opinions, emotions from a piece of text expressed by online web users. Their opinions or sentiments are categorized into three categories. They are positive, negative, or neutral. There are three levels of sentiment analysis, namely document level, sentence level and aspect level. From the three levels, the proposed work considers sentence level analysis for the identification of sentiments.

- RSS FEEDS

In order to identify the sentiments, this research work focuses on the collection of social media contents from Really Simple Syndication (RSS) news feeds for a particular period of time. Rich Site Summary or RSS feeds are generally in the form of XML format which changes its text contents regularly. RSS syndicates the news documents and web contents to the users. It pushes the latest headlines from the websites and downloads to the user’s personal computer for quick scanning of up to date news. In this research work, for the particular company the stock news contents are collected using RSS feeds which are stored in the text documents. A new model has been proposed to predict the sentiment scores of the RSS news feeds which creates an algorithm called sentence level sentiment score (SSS) algorithm.

In order to decide the stock market fall and rise conditions, there are three conditions available in the results of RSS feeds. If there is a progressive RSS news feed, then there will be positive effect on the stock market which also indicates that the share prices will improve in future.
Same way if there is negative RSS news feed, then there will be negative feedback on the stock market users and will cause negative effect on the stock holders and causes the stock prices to down. If there is no impact on the RSS news feeds or fact news released about the company, then opinion will be neutral. The sample screen shot of RSS news feed is shown in Figure 4.1.

![Sample screen shot of RSS news feed.](image)

**Figure: 4.1 Sample screen shot of RSS news feed.**

- **MATHEMATICAL MODEL**

  The vital role of the research work is the prediction of stock market values using sentiments of RSS news feeds along with the mathematical model. In general, mathematical model involves volume, price or open interest contract. Using the historical prices, technical analysis employs
indicators to predict future price movements. The mathematical model involves various stock level indicators. A few of them are Moving Average, Bollinger Bands, Moving Average Convergence or Divergence, Accumulation – Distribution, Stochastic Relative Strength Index (RSI). In order to compute the stock value, one of the technical factors of the stock market called historical prices is considered for the mathematical calculation part. In this approach, one of the stock level indicators called Moving Average (MA) is considered for the prediction of stock market value along with RSS news feed sentiments.

In general, there are two levels of common time periods for stocks markets. One is long-term support and the other is shorter-term support. Long term support considers the 50-days, 100-days and 200-days, whereas short term support considers 5 days, 10 days, 15 days and 21 days.

4.2. SYSTEM ARCHITECTURE OF SENTENCE LEVEL SENTIMENT SCORE ALGORITHM

The research work explores an automatic identification of stock news opinions using RSS news feeds and also predicts the movement of the stock market with the help of Moving Average stock level indicator.

- RSS STOCK NEWS FEED

In general, enormous stock market data are collected from the website for all stock related companies.

RSS news feeds are used to collect the stock market related news data from the web pages. These are in the format of XML document which consists of title, author, description, date, link etc. The description of the stock news is collected in the XML format. Here RSS news feed list is used to collect the XML contents available in the web pages.
The system architecture for Sentence level Sentiment Score (SSS) Algorithm of the proposed system is given in Figure 4.2.

Figure: 4.2 System Architecture for Sentence level Sentiment Score (SSS) Algorithm
• **RSS STOCK NEWS PREPROCESSOR**

RSS news stock pre-processor is used to remove the incomplete, duplicate, incorrect, improper data available in the documents. The unwanted information available in the document is removed using pre-processor. Hence, smoothening of noisy data, filling of missing value, identifying and removing the outliers are carried out in this module. After the stock news pre-processing stage, the cleaned data are passed to the next level.

• **RSS FEEDS SENTENCE SPLITTER**

RSS feeds sentence splitter is used to split the cleaned RSS news data into parsed sentences. All the news data available in the entire document is now separated as individual sentences and is processed further for sentiment analysis.

• **SENTENCE LEVEL SENTIMENT ANALYSER**

In the determination of the sentiments of the sentence level document, two important phases are involved. They are Part-of-Speech tagger and Dictionary based approach.

In the process of identifying the polarity of the sentence, it is divided into three types such as positive, negative, or neutral. Examples for positive sentences are the sentence that considers the positive sentiments like enthusiasm, kindness, happiness and such others. Same way negative sentences that consider the negative sentiments are of hate, violence, sadness, discrimination and such others. If the sentence implies no emotions, then it is classified as neutral sentence.
- **Part-of-speech—tagger**

  The other name of the Part-Of-Speech (POS) tagger is word-category disambiguation or grammatical tagging. Natural Language Processing (NLP) task uses the concept of POS tagger which does not require full constituency parse tree. It only requires the categories of word information. POS is a piece of software and assigns part of speech to each word of the readable text. POS assigns the following tagger such as noun, adjective, verb, and adverb to the word.

- **Dictionary based Approach**

  The unsupervised technique consists of various methods to determine the orientations of sentences. In this research work, dictionary based approach is used to predict the opinion of the words and their corresponding polarities. In order to determine the word sentiments, dictionary-based techniques use synonyms, antonyms and hierarchies in WordNet. This method assigns the synset of WordNet into three different numerical scores such as Obj(s), Pos(s) and Neg(s).

- **HISTORICAL STOCK MARKET DATA BASE**

  Another parallel task performed in the stock market prediction is mathematical model. In this module, the historical prices for a particular stock company are collected from websites. This acts as an input dataset for the prediction of mathematical model. This database carries huge amount of information over the years.

- **MOVING AVERAGE STOCK LEVEL INDICATOR**

  In this module, the historical prices are collected from the websites of the particular company and the Moving Average value for the stock values
are calculated for guiding the stock investors to take decision about buying or selling their stocks. Moving Average stock level indicator is a technical analysis tool which is used to compare the actual index data with its average taken over a period of time. There are different types of Moving Average indicator. They are Simple Moving Average (SMA), Exponential Moving Average (EMA) and Weighted Moving Average (WMA). In order to find the next day’s sensex price, the moving average stock level indicator is applied. The sensex is calculated using the equation (4.1).

\[
\text{Sensex} = \left( \frac{\text{sum of free float market capitalization}}{\text{Base Market Capital}} \right) \times 100 \tag{4.1}
\]

Free float market capitalization is represented by the equation (4.2).

Free Float Market capitalization = Share price \times (\text{shares outstanding} - \text{locked in shares}) 
\tag{4.2}

The sum of the closing prices is divided by number of time periods. The following equation (4.3) is used for the calculation of Moving Average method.

\[
F_t = \frac{A_{t-1} + A_{t-2} + A_{t-3} + \ldots + A_{t-n}}{n} \tag{4.3}
\]

\(F_t\) = Forecast for the coming period,
\(A_{t,n}\) = Actual occurrence in the past period for up to ‘n’ periods,
\(n\) = Number of periods to be averaged.

In this research \(n = 5\)-day or 10-day or 15-day.

The Moving Average computation of the proposed predictive system considers the following time period such as 5-days, 10-days and 15-days.
The comparative study of the research work is as follows. If 5-day value is higher than 10-day, and 10-day is higher than 15-day, then it indicates the results as positive and the sensex prices go up for the next day. If 5-day value is less than 10-day and, 10-day value is less than 15-day, then it indicates the result to be negative and the sensex prices will go down for the next day. If 5-day is higher than 10-day, and 10-day is less than 15-day, and also if the 5-day is less than 10-day, and 10-day is higher than 15-day, then it indicates the result to be neutral and there is no change in the sensex prices for the next day.

4.3. SENTENCE LEVEL SENTIMENT SCORE (SSS) ALGORITHM

An automatic predictive system is proposed to find the stock market forecasting by correlating the sentiments of RSS news feeds with Moving Average stock level indicator. To perform this task, an algorithm called Sentence level Sentiment Score (SSS) algorithm is designed and developed. Sentence level sentiment analyser is used to find the overall result of the individual sentence stored in the document. Here every individual sentence applies the concept of Part-of-Speech tagger and Dictionary based approach.

Dictionary based approach assigns each synset of WordNet into three sentiment numerical scores: Obj(s), Pos(s) and Neg(s). In general, each of the three scores ranges from 0.0 to 1.0, and their sum is 1.0. To each word POS tagger assigns the corresponding verb, adjective, noun in such a manner that they happen to be equivalent words [18]. If the synset score value ranges from 0.0 to 1.0 then that sentence is said to be a positive sentence; if the synset score ranges from -1.0 to 0.0, then that sentence is considered to be negative sentence. Finally, if the synset value is 0.0, it is said to be a neutral sentence.
Sequence of words is described in equation (4.4).

\[
\text{Sequence of words}(W) = W_1 + W_2 + \cdots + W_n
\]  \hspace{1cm} (4.4)

\(n\) = Number of words in the synset.

Positivity and Negativity of sentiment words are described in (4.5) and (4.6)

\[S \in S(W_{ve}) \]  \hspace{1cm} (4.5)

If \(W\) is positive sentiment word

\[S \in S(W_{ve}) \]  \hspace{1cm} (4.6)

If \(W\) is negative sentiment word

4.4. IMPLEMENTATION AND ANALYSIS OF SSS ALGORITHM

The proposed prediction algorithm is used to predict the stock market by combining the results of the RSS stock news by applying sentiment analysis concept and finding the Moving Average stock level indicator using the historical prices of the specific stock company. The prediction analysis of the stock market using the results of Sentiment and Sensex-Moving Average is shown in Table 4.1.

<table>
<thead>
<tr>
<th>Sentiment Analysis Result</th>
<th>Sensex-Moving Average Result</th>
<th>Final-Result Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>Negative</td>
<td>Neutral</td>
</tr>
<tr>
<td>Negative</td>
<td>Positive</td>
<td>Neutral</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Based on the above tabular column results, the final result prediction is considered as positive, only if both sentiment analysis and Sensex-Moving
Average results are positive. Same way, the final result prediction is considered as negative, only if both sentiment analysis and Sensex-Moving Average results are negative. Finally, the final result prediction is considered as neutral, if any one of sentiment analysis or Sensex-Moving Average results is positive or negative.

4.5. EVALUATION AND RESULTS OF SSS ALGORITHM

The experimental study has been conducted for the company Arab Bank (ARBK) from Amman Stock Exchange (ASE) to predict the stock market values. Amman Stock Exchange (ASE) is available in the oracle database which contains historical values of 230 companies that are listed in the exchange from the year 2000. In this research work, the dataset of the historical prices and RSS stock news feeds are collected for the year 2005 to 2007. The prediction analysis of this proposed work is compared with the prediction algorithms provided by the previous work which shows an improvement in their accuracy analysis. In order to compare the proposed research study with the previous work, the same dataset collected for the year 2005 to 2007 [101] is utilized to show the improvement in the prediction accuracy.

The dataset for the RSS news feed sentiments are collected for the company Arab Bank (ARBK) from http://investing.einnews.com/news/ase-stock and the historical prices of the same company for mathematical model are collected from http://www.marketstoday.net/markets/jordan/Historical-Prices/10/en/#. The Moving Average of ARBK Company is calculated for the month of April 2006. Same way, sentiment analysis of the same company is also performed for the month April 2006.

Finally, the results of both are combined and final stock market result is predicted according to the prediction strategy value given in Table 4.1. The sensex point - Moving Average calculation analysis is shown in Figure 4.3.
In the above Figure 4.3, various colors like blue, red and green show 5-day moving average, 10-day moving average and 15-day moving average respectively. From the graph it is inferred that the sensex values of 5-day moving average are higher than those of 10-day and the sensex values of 10-day moving average are higher than those of 15-day; then it shows positive result. If there is positive result, then the stock is also positive.

Precision and recall are the two measures used for the measurement of accuracy. Precision is the ratio of true positives among all retrieved instances and recall is the ratio of true positives among all positive instances.

In the returned document the number of the documents that are positive are known as Precision measure. The number of the returned documents that are correct are known as Prediction.
The formula for precision and recall value for instances are given in equations (4.7) and (4.8).

\[
\text{Precision} = \frac{TP}{(TP + FP)} \tag{4.7}
\]

\[
\text{Recall} = \frac{TP}{(TP + FN)} \tag{4.8}
\]

Accuracy of various classification algorithms are given in table 4.2. Also the accuracy comparisons of the previous and the proposed algorithms for Arab Bank Company are shown in Figure 4.4.

**Table 4.2: Accuracy of various classification algorithms**

<table>
<thead>
<tr>
<th>ARBK Company</th>
<th>ID3</th>
<th>C4.5</th>
<th>Moving Average</th>
<th>SSS Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Instances</td>
<td>499</td>
<td>499</td>
<td>499</td>
<td>499</td>
</tr>
<tr>
<td>Correctly Classified</td>
<td>233</td>
<td>237</td>
<td>321</td>
<td>391</td>
</tr>
<tr>
<td>Precision %</td>
<td>46.69</td>
<td>47.49</td>
<td>64.32</td>
<td>78.75</td>
</tr>
</tbody>
</table>

**Figure: 4.4 Over all accuracy comparison between the algorithms**
The precision measures of the previous and the proposed algorithms were compared for the company Arab Bank for the month of April 2006 in shown in Figure 4.4. The previous algorithms such as ID3 and C4.5 shows the precision values around 46 to 47%, whereas the proposed research work shows the precision value as 78.75% for Stock level indicator with sentiment analysis and without sentiment analysis shows 64.32%. This gives an improvement of 14.43% in result analysis.

The proposed algorithm called SSS algorithm is used to calculate the sentiments of news data and sensex point moving average data for the month of April 2006. Without sentiment value, Moving Average indicator gives the accuracy of 64.32%. Same way, with sentiment value, the Moving Average indicator produces 78.75% with an improvement of 14.43%.