CHAPTER 6

SENTENCE POLARITY CALCULATION (SPC) ANALYSIS USING RSS NEWS FEEDS, TWEETS AND NEWS CONTENTS ALONG WITH HYBRID MATHEMATICAL MODEL

4.1. INTRODUCTION OF SENTENCE POLARITY

It is a well-known fact that people’s mentality is highly influenced by the news trends. News biasness happens when the news media distorts the reporting and objectivity of news in any form. The next approach in the research is to achieve further improvement in the accuracy of prediction stock market can be achieved by adding one more sentimental factor called news bias. In this proposed research analysis, the stock market prediction is achieved by combining the sentiments of social media contents such as RSS news feeds, tweets and news bias with a hybrid mathematical model that uses three stock level indicators.

- NEWS BIAS

There are several dimensions available in the news bias that is interlinked with political, social and economic problems. News bias is a complex process which reflects the opinion of common people about topic or policies. In our day-to-day life, newspaper reading is very important for the people to know about latest headlines through Internet. The popularity of online social networks is so much that it gathers millions of users. Now a days, browsing and retrieving the news through online news websites is very easy and popular.

In order to determine whether news articles are fair and credible, the background information of news websites is necessary. News bias can be understood only by understanding the context of media industry as a whole. The main aim of the news bias is to provide the users or readers with
objective, unbiased, reliable and impartial news which also explores the selective omission, distortion of facts, lack of transparency and balanced reporting. Examples for distinct news topics are “Obama’s farewell speech”, “US election 2016 results”, “Indian Currency demonetization”. Sample RSS news feed is shown in Figure 6.1.

![Image of Obama's farewell speech]

Figure: 6.1 Example for news bias

- **HYBRID MATHEMATICAL MODEL**

  In order to forecast the stock market, along with the sentiment analysis, an add on model called hybrid mathematical model which uses the historical prices for the calculation of stock level indicators is used. Three stock level indicators, namely Moving Average, Moving Average Convergence or Divergence and Stochastic Relative Strength Index (RSI) are used. The combination of this hybrid mathematical model is called MMS engine. Historical prices are one of the technical factors that are used in the calculation of stock price prediction. Finally, the overall results of three stock level indicators form a hybrid mathematical model to check the performance of stock market analysis.
4.2. SYSTEM ARCHITECTURE OF SENTENCE POLARITY CALCULATION (SPC) ALGORITHM

The proposed system architecture for Sentence Polarity Calculation (SPC) Algorithm is given in Figure 6.2.

Figure: 6.2 System Architecture for Sentence Polarity Calculation (SPC) Algorithm
In Sentence Polarity Calculation (SPC) Algorithm, the data that are related to stock market are collected from the Web pages. The social media contents that are related to stock market are passed to each corresponding module.

This research is categorized into two tasks. One task is sentiment mining from social media factors of RSS stock news feeds, tweets and stock news bias. From the Web pages, for a given particular company RSS stock news feeds are accessed and stored in the document. The stored document is in the form of XML which contains title, author, description and date which is explained in section 1.3.

The next social media content is twitter which helps the users to express their views by means of communication through short messages. Twitter posts are called tweets that are accessed through online based on the stock company for the prediction of sentiments. Third social media content is news bias, which is added in this conventional method for further improvement of sentiment analysis that is carried for the forecasting of stock market prices.

Next is the social media content pre-processor, performs cleaning task such as removal of unwanted, incomplete, unreliable data. After performing the smoothening technique, the cleaned documents are given as input to the next module. Next module is social media content sentence parser where all the social media contents are divided into sentences and stored in the document.
**SENTIMENT POLARITY CALCULATION ALGORITHM**

The Sentiment Polarity Calculation Algorithm (SPC) is explained in Figure 6.3

```plaintext
for-each document in {social media contents} do
    [sentences] = extract_sentences (document)

for-each sentence in {sentences} do
    remove_punctuation (sentence)
    {words} = tokenize (sentence)
    {words} = {words} – {stopwords}
    {sentiments} = Φ

    for-each word in {words} do
        stem = stemmize (word)
        if stem in lexicon then
            {sentiments} = {sentiments} with lexicon
        check if the word = Noun/Adj/Verb
        Store the Word synset score value into Database.
        {sentiment_scores} = count ({sentiments})
        return {sentiment_scores}

        store the overall sentence {sentiment_scores} of
        each word into Database.
        calculate up to end of sentence.

    Else
        end the process
```

**Figure: 6.3 Sentiment Polarity Calculation Algorithm**

Now synset score values of each sentence are passed to Sentiment Polarity Calculation Algorithm (SPC). For each sentence this algorithm performs the polarity calculation analysis.
Polarity Identification Algorithm for calculating the individual polarity of sentence is given in Figure 6.4.

```
Require: {sentiment_scores}, F_P, F_N;

{sentiment_scores} = from Sentiment Polarity Calculation Algorithm;
F_P: Final Positive;
F_N: Final Negative;
P: Positive;
N: Negative;

Function Polarity Identification ((sentiment_scores), F_P,F_N)

    If (sentiment_scores)_N == 0)
        Return finalSentimentScore (F_P)

    Else if (sentiment_scores)_P == 0)
        Return finalSentimentScore (F_N)

    Else
        {  
            If (F_P - F_N) > 0.1)
                Return finalSentimentScore (F_P)

            Else if (F_N - F_P) > 0.1)
                Return finalSentimentScore (F_N)

            Else
                {  
                    If (F_P + F_N) > 0)
                        Return finalSentimentScore (F_P)

                    Else if (F_P + F_N) < 0)
                        Return finalSentimentScore (F_N)

                    Else
                        Return 0
                }

        }

    finalSentimentScore(F)
    {
        Return F;
    }
```

Figure: 6.4 Polarity Identification Algorithm
Natural Language Processing Module

In Natural Language Processing (NLP) module, the inputs are given from sentence parser to NLP module for further processing of sentiment analysis. The first task is part-of-Speech (POS) tagger which carries bag of synsets for each part-of-speech such as verb, adverb, noun and adjective. Each synset carries synset score that is provided by dictionary based approach. If synset score is positive, then that sentence is said to be a positive sentence. Instead, if the synset score is negative, then that sentence is considered as a negative sentence. If the synset score is null, there are no changes in the sentence and it is considered as a neutral one.

Sentiment Polarity Algorithm is used to find sentiment score value of each synsets of the sentence. Here each social media document is passed inside the SPC algorithm for the calculation of their synset scores. All the synset values are stored in the database for final identification of polarity values. If the score ranges from 0.0 to 1.0, it is said to be a positive score; same way, if the score ranges from 0.0 to -1.0, then it is said to be a negative score. Finally, if the score is exactly 0.0 then there is no change and it is said to be neutral. In the proposed work an algorithm called Polarity Identification is designed and used to find the final polarity of the sentences that are available in the document. Based on the synset score calculation, the positivity, negativity or neutral score is identified.

Another parallel task that is performed on the sentiment analysis is hybrid mathematical model. Hybrid mathematical model is constructed using historical prices of stock companies. The required datasets for the calculation of mathematical model is accessed from the websites. These datasets are given as input to the Moving Average, Moving Average Convergence or Divergence and Stochastic Relative Strength Index (RSI) MMS engine. The stock prices are applied to this hybrid mathematical model which gives an idea to stock forecasters as to when to buy or sell their stocks.
The first stock level indicator is Moving Average indicator which is calculated with the help of the closing price for a specific time period and it is divided by total number of time intervals. In proposed research the common time intervals considered are 5-days, 10-days and 15-days.

Moving Average stock level indicator is represented in the equation (4.3)

The next stock level indicator is Moving Average Convergence or Divergence oscillator (MACD). This is the indicator used to find the difference between two Exponential Moving Averages (EMA). The common time intervals used in this indicator are 9-days, 12-days and 26-days.

This indicator shows negative when 12-day EMA is lower than 26-day EMA. This shows that negative gap is expanding between the faster MA and Slower MA. Also downward momentum is accelerating, and indicates fall in the stock trends. On the other hand, indicator shows positive, only when 12-day EMA is higher than 26-day EMA. This shows that positive gap is expanding between the faster MA and Slower MA. Also upward momentum is accelerating, and indicates rise in the stock trends. The formulae for MACD, Signal Line and MACD Histogram are represented in the equations (5.1),(5.2) and (5.3) respectively.

Third indicator used in MMS engine is Stochastic Relative Strength Index (RSI). The other name of RSI is oscillator and used to calculate the value between 0 and 1 which is then plotted as a line. Overbought and oversold conditions are identified using this indicator. The overall Stochastic RSI value calculation is given in equation (5.8).
4.3. IMPLEMENTATION AND ANALYSIS OF SPC ALGORITHM

From this research on prediction analysis, the forecasting of stock market involves sentiments of social media contents and hybrid mathematical model. The prediction analysis of all results will provide huge number of combinations in the output part. After optimizing the combinations of all results, the final prediction analysis is reduced to 15 significant predictions.

The overall prediction result of stock market analysis is given in Table 6.1.

**Table 6.1: Final Prediction analysis of hybrid mathematical model with sentiments of social media contents**

<table>
<thead>
<tr>
<th>Sentiment Analysis for RSS News Feeds</th>
<th>News Content Sentiments</th>
<th>Sentiment Analysis for Twitter</th>
<th>Stock level indicators Result</th>
<th>Final-Result Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>Negative</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
</tr>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
<td>Neutral</td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>Positive</td>
<td>Neutral</td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>Neutral</td>
<td>Negative</td>
<td>Negative</td>
<td>Neutral</td>
</tr>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
</tr>
<tr>
<td>Positive</td>
<td>Negative</td>
<td>Negative</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>Negative</td>
<td>Neutral</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>Neutral</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>Neutral</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Neutral</td>
<td>Neutral</td>
<td>Negative</td>
<td>Negative</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

4.4. EVALUATION AND RESULTS OF SPC ALGORITHM

In the experimental study, this research performed the prediction analysis for various companies like Arab Bank (ARBK) from Amman Stock Exchange, Infosys, TCS and Reliance. It should be noted that Infosys, TCS
and Reliance are the top notch IT companies in India. For the above companies, the sentiment datasets of RSS news feeds, tweets and news bias were collected from the year January 2016 to January 2018. For the same companies, hybrid mathematical model collected the historical prices from the year January 2016 to January 2018. The collection of dataset for sentiments of RSS news feeds, tweets, and historical prices is given in Table 6.2.

Table 6.2: Sources of dataset collection for various companies

<table>
<thead>
<tr>
<th>S. No</th>
<th>Company name</th>
<th>Methods</th>
<th>Arab Bank</th>
<th>Infosys</th>
<th>TCS</th>
<th>Reliance</th>
</tr>
</thead>
</table>

The above links were used to collect the stock values for all the companies. By correlating the results of the sentiment analysis and sensex
prices, the stock investors can take decision about when to buy or sell their stocks. News bias collection for all companies was done from various news websites that are given in Table 6.3.

**Table 6.3: List of various News Websites**

<table>
<thead>
<tr>
<th>Country Name</th>
<th>News Websites</th>
<th>URL Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. CNN</td>
<td><a href="http://www.cnn.com/">http://www.cnn.com/</a></td>
</tr>
<tr>
<td></td>
<td>3. USA Today</td>
<td><a href="http://www.usatoday.com/">http://www.usatoday.com/</a></td>
</tr>
<tr>
<td></td>
<td>10. Thomson Reuters</td>
<td><a href="https://in.reuters.com">https://in.reuters.com</a></td>
</tr>
<tr>
<td>United Kingdom</td>
<td>1. BBC</td>
<td><a href="http://www.bbc.co.uk/">http://www.bbc.co.uk/</a></td>
</tr>
<tr>
<td>India</td>
<td>1. The Hindu</td>
<td><a href="http://www.thehindu.com/">http://www.thehindu.com/</a></td>
</tr>
<tr>
<td></td>
<td>2. India Times</td>
<td><a href="http://www.indiatimes.com/">http://www.indiatimes.com/</a></td>
</tr>
<tr>
<td></td>
<td>3. Indian Express</td>
<td><a href="http://www.indianexpress.com/">http://www.indianexpress.com/</a></td>
</tr>
<tr>
<td></td>
<td>4. NDTV</td>
<td><a href="http://www.ndtv.com">http://www.ndtv.com</a></td>
</tr>
<tr>
<td></td>
<td>5. Economic Times</td>
<td><a href="https://economictimes.indiatimes.com">https://economictimes.indiatimes.com</a></td>
</tr>
</tbody>
</table>

Stock news for Arab bank was collected from https://www.ase.com.jo/en/news.

The result analysis was carried out for a specific period of time interval for both sentiment analysis and hybrid mathematical model. The final prediction accuracy will provide an idea to stock investors to take decision about their stocks.

Precision is defined as the total number of positive instances that are identified as correctly classified divided by total number of topics that are identified as correct. The formula for precision is given in equation (4.7).
For each individual stock level indicator the precision accuracy measure is obtained. The overall cumulative mean stock level indicators obtained are passed as input to the next level of prediction analysis.

The various notations used in the Table 6.4 are as follows:

Cumulative Mean SI. – Cumulative Mean Stock Level Indicators

SI – Stock level Indicators

MA – moving Average

MACD – Moving Average Convergence or Divergence

RSI – Relative Strength Index

Prediction accuracy levels of individual Stock level indicators are shown in Table 6.4.

Table 6.4: Prediction accuracy level of individual Stock level indicators

<table>
<thead>
<tr>
<th>Company</th>
<th>SI</th>
<th>Total Instances</th>
<th>Correctly Classified</th>
<th>Precision %</th>
<th>Cumulative Mean SI. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arab Bank (ARBK)</td>
<td>MA</td>
<td>494</td>
<td>321</td>
<td>64.9</td>
<td>64.27</td>
</tr>
<tr>
<td></td>
<td>MACD</td>
<td>494</td>
<td>310</td>
<td>62.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>494</td>
<td>329</td>
<td>66.59</td>
<td></td>
</tr>
<tr>
<td>Infosys</td>
<td>MA</td>
<td>494</td>
<td>325</td>
<td>65.78</td>
<td>64.14</td>
</tr>
<tr>
<td></td>
<td>MACD</td>
<td>494</td>
<td>301</td>
<td>60.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>494</td>
<td>295</td>
<td>59.71</td>
<td></td>
</tr>
<tr>
<td>TCS</td>
<td>MA</td>
<td>494</td>
<td>316</td>
<td>63.96</td>
<td>64.75</td>
</tr>
<tr>
<td></td>
<td>MACD</td>
<td>494</td>
<td>308</td>
<td>62.34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>494</td>
<td>306</td>
<td>61.94</td>
<td></td>
</tr>
<tr>
<td>Reliance</td>
<td>MA</td>
<td>494</td>
<td>326</td>
<td>65.99</td>
<td>66.12</td>
</tr>
<tr>
<td></td>
<td>MACD</td>
<td>494</td>
<td>330</td>
<td>66.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>494</td>
<td>324</td>
<td>65.58</td>
<td></td>
</tr>
</tbody>
</table>
For various algorithms the accuracy prediction analysis is carried out and given in Table 6.5.

**Table 6.5: Accuracy prediction analysis for various algorithms**

<table>
<thead>
<tr>
<th>Algo.</th>
<th>Tot. Inst.</th>
<th>Arab Bank (ARBK)</th>
<th>Infosys</th>
<th>TCS</th>
<th>Reliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>494</td>
<td>321</td>
<td>64.97</td>
<td>314</td>
<td>63.52</td>
</tr>
<tr>
<td>SI</td>
<td>494</td>
<td>320</td>
<td>64.77</td>
<td>307</td>
<td>64.14</td>
</tr>
<tr>
<td>SSS Algo.</td>
<td>494</td>
<td>340</td>
<td>68.82</td>
<td>349</td>
<td>69.64</td>
</tr>
<tr>
<td>SMCS Algo.</td>
<td>494</td>
<td>352</td>
<td>71.25</td>
<td>368</td>
<td>73.49</td>
</tr>
<tr>
<td>CS Algo.</td>
<td>494</td>
<td>385</td>
<td>77.93</td>
<td>390</td>
<td>78.94</td>
</tr>
<tr>
<td>SPC Algo.</td>
<td>494</td>
<td>430</td>
<td>87.04</td>
<td>438</td>
<td>88.66</td>
</tr>
</tbody>
</table>

The various notations used in the Table 6.4 are as follows:

SI – Stock level Indicators

SSS Algo. – SSS Algorithm

SMCS Algo.– SMCS Algorithm

CS Algo.– CS algorithm

SPC Algo.– SPC Algorithm

Tot.Inst. – Total Instances

Corr.Clas.– Correctly Classified
Prec.%— Precision %

**SSS Algorithm** – Sentence level Sentiment Score algorithm is used for finding the Sentiments of RSS news feed along with mathematical model [17].

**SMCS Algorithm** – Social Media Contents Sentiment algorithm is used for analysing the sentiments of Tweets, RSS news feed, and hybrid mathematical model that uses three Stock level indicators [19].

**CS algorithm**– Contents sentiment algorithm is used for analysing the sentiments of news topics from the internet for the prediction of news bias.

**SPC Algorithm**– Sentence Polarity Calculation algorithm is used for analysing the sentiments of RSS news feed, Tweets and Biased News along with hybrid mathematical model that uses three Stock level indicators.

![Prediction Accuracy Graph](image)

Figure: 6.5 Comparison graph for various stock market companies using different sentiment analysis algorithms
For the companies considered, the comparison graphs for various algorithms were constructed and are shown in Figure 6.5. In this proposed analysis, 494 total instances are collected during the period of year 2016 - 2018. First, prediction SSS algorithm shows an improvement of 4% more than previous one. Second, SMCS algorithm shows an improvement of 6% over the previous method. Third SPC Algorithm shows an improvement of 10% over the previous method. This indicates that the proposed algorithm performs better than the previous one.

From this research work, it is seen that the impact of sentiments from social media contents such as RSS news feeds, tweets and news bias along with hybrid mathematical through three stock level indicators enhances the quality of stock prediction when compared with selected other previous research analysis.