News Analytics and Sentiment Analysis to Predict Stock Price Trends

Spandan Ghose Chowdhury, Soham Routh, Satyajit Chakrabarti

Department of Computer Science and Engineering
Institute of Engineering & Management
Kolkata, India

Abstract— Business news carries varied information of different companies. In this rapidly moving world the number of news sources present is uncountable, and it's humanly impossible to read and find all relevant information in the form of news to draw a conclusion timely to make an investment plan that returns maximum profit. In this paper, we have proposed a predictive model to predict sentiment around stock price. First the relevant real time news headlines and press-releases have been filtered from the large set of business news sources, and then they have been analyzed to predict the sentiment around companies. In order to find correlation between sentiment predicted from news and original stock price curve's ups and downs, we plot the sentiment of 15 odd companies over a period of 4 weeks. Our result shows an average accuracy score for identifying correct sentiment of around 70.1%. We also have plotted the errors of prediction for different companies which have brought out the RMSE and MAE of 30.3% and 30.04% respectively and an enhanced $F_1$ factor of 78.1%. The comparison between positive sentiment curve and stock price trends reveals 67% co-relation between them, which indicates towards existence of a semi-strong to strong efficient market hypothesis.

Keywords— stock price trends, prediction model, knowledge discovery, sentiment analysis, market trends, news analytics, efficient market hypothesis.

I. INTRODUCTION

In this closely connected world, we have access to all sorts of information. Hence, it’s easier for us to make informed choices about any topic. Our decisions constantly get influenced by the kind of news we come across on day to day basis. The sentiment towards a particular entity becomes the driving force to take a proper decision.

There is a strong yet complicated relation between the market and the information available in the form of news. The arrival of news at every moment changes the perception or sentiment towards a particular company or their adopted business strategies. These days due to the bliss of internet, the traders, and investors have constant access to the updated news, and the news constantly mould their sentiments and influences their decision to invest in a particular company. The news from standard and authentic sources helps them to rebalance their assets judiciously. This sentiment not only becomes a deciding factor while rebalancing funds, but also aids in the decision of risk controls, and hence the stock market reflects the sentiment.

Major news can significantly impact on traders’ sentiments [1] [2] [3] [4] [5] [6] yielding to a major change in the entire investment plan. Every investment, small or big, has significant impact on the growth of the company. When a trader having an enhanced positive sentiment finds a company worth investing, the company’s growth also gets boosted up. On the other hand, the trader having a negative perception about a particular doesn’t feel safe investing in a company, withdraws investments in order to avoid loss, and as a result of which the stock price of the company goes down drastically. [7] [8]

Considering how news impacts markets, Barber and Odean [9] note “significant news will often affect investors’ beliefs and portfolio goals heterogeneously, resulting in more investors trading than is usual” (high trading volume). It is well known that volume increases on days with information releases [10]. Important news frequently results in large positive or negative returns. Ryan and Taffler (2002) find for large firms a significant portion (65%) of large price changes and volume movements can be linked to publicly available news releases. Sometimes investors may find it difficult to interpret news resulting in high trading volume without significant price change. [11] [12]

News has always been an important source of information to build perception of market investments. As the volumes of news and news sources are increasing rapidly, it’s becoming impossible for an investor or even a group of investors to find out relevant news from the big chunk of news available. But, it’s important to make a rational choice of investing timely in order to make maximum profit out of the investment plan. And having this limitation, computation comes into the place which automatically extracts news from all possible news sources, filter and aggregate relevant ones, analyse them in order to give them real time sentiments.

In this paper we have collected, classified and analysed relevant real time news from widely accepted and trusted news sources available in the public domain to forecast the perception around a particular company which helps the investors and traders to come up with informed, rational investment plan timely which ensures maximized profit. In this study we also have tracked the sentiments of the companies over a period of 4 weeks in order to find out correlation between the moods predicted from the relevant news and the original stock price curve’s ups and downs.

The main contributions of the paper are as follows:

- Implementation of an efficient prediction model that scores sentiments from all relevant real time news available in over 15 news sources present in public domain.
This model classifies all of the relevant real time news from the large database of news.
This forecasts real time news sentiment that reflects stock price movement trends.
Compares and finds out a strong mean correlation of 67% between positive sentiment movement and original stock price curve of 15 companies over a period of a month, and hence proves the test of EMH.
The accuracy level of this prediction model is 70.1%. The prediction model has Mathew’s Correlation Coefficient of 0.4.

The rest of the paper is organized as follows. Section II outlines the existing work in this field of sentiment analysis and concept of Efficient Market Hypothesis. Section III proposes and describes our predictive model and further plots sentiment trend predicted from the model and compares with original stock prices movements. In next section i.e. section IV we evaluate the prediction model using different accuracy measurement techniques. In this section, we also compare and assess the co-relation of the sentiment trends of the proposed model with the original stock price movement. Finally in Section V, we conclude our paper and discuss the future work related to it.

II. LITERATURE REVIEW

A. Sentiment Analysis

Several systems have been built which attempt to quantify opinion from product reviews. Pang, Lee and Vaithyanathan [13] perform sentiment analysis of movie reviews. Their results show that the machine learning techniques perform better than simple counting methods. In another paper [14], they identified which sentences in a review are of subjective character to improve sentiment analysis. They did not make this distinction in their system, because they felt that both fact and opinion contribute to the public sentiment about news entities. [15] [16] [17]

Wilson, Wiebe, Hoffman [18] [19] presented a new approach that automatically identify the contextual polarity for a large set of sentence expressions, achieving results that are significantly better than baseline.

Vu Dung Nguyen, Blesson Varghese and Adam Barker [20] made an analysis of information retrieved from micro blogging services such as Twitter can provide valuable insight into public sentiment in a geographic region. This insight can be enriched by visualizing information in its geographic context. Two underlying approaches for sentiment analysis are dictionary based and machine learning. The former is popular for public sentiment analysis, and the latter has found limited use for aggregating public sentiment from Twitter data. The research presented in their paper aimed to widen the machine learning approach for aggregating public emotion. A framework for analysing and visualizing public sentiment from a Twitter corpus was developed. In the same paper, a dictionary-based approach and a machine learning approach were implemented within the framework and compared using one UK based case study – the royal birth of 2013.

B. Efficient Market Hypothesis

In the literature there is an existence of efficient market hypothesis (EMH) which states at any point of time and in a liquid market security prices fully reflect all available information. It further categorises market into three degrees: weak, semi-strong and strong, which addresses the inclusion of non-public information in market prices. [26] [27] [28] [29]

This theory contends that since markets are efficient and current prices reflect all information, attempts to outperform the market are essentially a game of chance rather than one of skill. In the weak form of EMH, future prices cannot be predicted by analysing prices from the past. The semi-strong form of EMH assumes that current stock prices adjust rapidly to the release of all new public information. It contends that security prices have factored in available market and non-market public information. The strong form of EMH assumes that current stock prices fully reflect all public and private information. It contends that market, non-market and inside information is all factored into security prices and that no one has monopolistic access to relevant information. It assumes a perfect market and concludes that excess returns are impossible to achieve consistently.

The research reported in this paper is motivated towards analysing public sentiment related to the stock of different companies in real-time and rapidly correlating it with the original stock price changes for a period of 4 weeks. In this paper we studied and analysed the news sources available in public domain that influences the market and the public
coconsciousness for a period of 4 weeks to study how the market moves on day to day basis with the available news using dictionary based approach. Public sentiment towards a particular company, for example, positive, negative or neutral, is understood by comparing news against lexicons from dictionaries. And these moods finally build the decision of selling, buying or holding the particular stock in order to have maximized profit out of every stock of the particular company. The positive mood pertaining to a company motivates people to invest more in a particular company over the others with the hope that the company where the person is going to invest is performing well, and the person will get profit in return. On the other hand, the negative and neutral mood channelizes the money to a company which has more positivity in the market place. Hence, the resultant positive mood finally becomes the index which determines the movements of stock prices of the particular company.

III. METHODOLOGY

This paper proposes a real time prediction model of sentiment analysis and news analytics. And also, over a period of 4 weeks, we compare and analyse our predicted sentiments with the market trends of 15 different companies namely – Wipro (WIPRO), Infosys (INFY), Unilever NV (UN), Tata Consultancy Services (TCS), International Business Machines (IBM), Tech Mahindra (TECHM), Reliance Industries Limited (RIL), Nestle India (NESTLEIND), General Motors (GM), Bharat Heavy Electicals Limited (BHEL), Google Inc (GOOG), Hindustan Motors (HINDMOTOR), Apple Inc (AAPL), Amazon (AMZN), IDBI Bank (IDBI) to find correlation between predicted sentiment curve with stock price trends and to test the Efficient Market Hypothesis in the market.

The prediction algorithm of this model to predict stock trends is driven by 3 indicators: Positive, Neutral and Negative.

The index of degree of positivity or the degree of negativity or the degree of neutrality is scored from the result of analysis of the available relevant real time news filtered and classified from a large stack of news collated from various popular news sources and press releases by the respective companies and the regular financial news sources – Wall Street Journal, Financial Times, Forbes, Reuters, CNBC, NDTV, Economic Times, Hindustan Times, Times of India, India Times, Telegraph, CNN Money, Market Watch, and Fortune.

B. News Sources

In this study the data have been collected from the press releases by the respective companies and the regular financial news sources – Wall Street Journal, Financial Times, Forbes, Reuters, CNBC, NDTV, Economic Times, Hindustan Times, Times of India, India Times, Telegraph, CNN Money, Market Watch, and Fortune.

C. Collector

The collector is responsible for collecting the corpus. In this study, the corpuses are the real time news headlines and press releases. The collector extracts the news headlines of the news RSS feeds which are stored into a database (T1) which nothing but an XML file and finally stores the collected corpus inside another database (T2). Everyday an average of around 2650 real time business news have been collected to do the sentiment analysis of the companies.

D. Keywords

The keywords are nothing but the words or triggers depending on which the news headlines or the press releases get classified. In our case the keywords will be the name, symbol of the company, the issues and phrases that concerns the company. The filtered news is the news relevant to the keywords. Keyword stays stored in a database (T6).

E. Classifier

The responsibility of the classifier is to trim the corpus offline. Parser trims the real time news headlines and press releases which are relevant to the particular organization and stores them to another database (T3) from the collected news stored inside database (T2). The classifier trims the news headlines on the basis of the. Output of the parsed files is used to analyse and estimate the indicators.

F. Analyser

The principal responsibility of the Analyser is to perform sentiment analysis on the parsed sentences which is the data stored in the T2 database. The approach that has been
followed in this paper is dictionary-based approach to analyse sentiment of the relevant news headlines in order to get the indicators. [30] [31] Dictionary based approach or statistic based approach has a significant high percentage of accuracy level and low percentage of tie level. [32] The analyser which brings out sentence-scores follows an algorithm which calculates cumulative sentiment scores of words present in each line of the parsed sentences. The algorithm which runs for every sentence to calculate its score is comprised of 3 steps – tokenizing, matching, and aggregating.

1) **Tokenizing:** The news headlines are tokenized using a lexical analyser using R. In T2 database, a vector of sentences has been gotten as a result of parsing, from where we had to get an array of scores which gets estimated in the later phase.

The algorithm to tokenize each sentence is as followed:

a) Cleaning up the sentence: The punctuations, control characters, and digits of the parsed sentence were cleaned up and stemmed.

b) Change sentence to lower case: Sentence then is converted into lower case.

c) Split the elements into character vector: The cleaned up lower cased sentence is split into words using.

d) Flatten lists: To produce a vector which contains all the atomic components which occur in the list, it gets flattened or unlisted.

2) **Matching:** Since here in this paper, we aim at analysing it using the most used dictionary based approach; the next step is to match the words of each sentence with the dictionary. The dictionary has two sub dictionaries of positive and negative words.

a) Match the words of the sentence with the dictionary: For each sentence, each word gets matched with the dictionary.

b) Return TRUE or FALSE: If there’s a match between the word of the sentence and the word of the positive or negative dictionary it returns TRUE, otherwise FALSE.

c) Sentence Scoring: The total sentiment score is calculated using the following formula and the sentence score is stored into an array.

\[
\text{sentiment score} = \sum \text{positive matches} - \sum \text{negative matches}
\]

3) **Aggregating sentence scores:** Once each sentence gets a score, the number of sentences with same score gets counted and put along with the particular sentence score. It gets stored into a data frame with Number of sentences with a unique score and the unique score.

\[
\text{Number of sentences with a unique score} = \sum \text{sentences with the unique score}
\]

**G. Estimator:**

The estimator estimates the sentence wise scores and brings out the degree of positivity, negativity and neutrality in percentage. The estimator follows the following algorithm to estimate the indicators. The formulas that the estimator follows are:

\[
\text{Total Positive Indicator} = \sum \text{positive score} \times \text{number of sentences with a particular positive score}
\]

\[
\text{Total Negative Indicator} = \sum (-1) \times \text{negative score} \times \text{number of sentences with a particular negative score}
\]

\[
\text{Total Neutral Indicator} = \text{number of sentences with a sentence score 0.}
\]

\[
\text{Total Score} = \text{Total Positive Indicator + Total Negative Indicator + Total Neutral Indicator}
\]

\[
\text{Degree of Positivity} = \left(\frac{\text{Total Positive Indicator}}{\text{Total Score}}\right) \times 100\%
\]

\[
\text{Degree of Negativity} = \left(\frac{\text{Total Negative Indicator}}{\text{Total Score}}\right) \times 100\%
\]

\[
\text{Degree of Neutrality} = \left(\frac{\text{Total Neutrality}}{\text{Total Score}}\right) \times 100\%
\]

**H. Visualiser**

The result of predictive analysis of a particular time stamp is shown in a pie chart. The pie chart comprises of 3 segments: the degree of positivity, the degree of negativity and the degree of neutrality.

The predicted sentiments of effective positive index have been plotted for 15 companies over a period of 4 weeks. The graph of original stock market movement and the movement of predicted sentiment result have been compared in order to find out co-relation between the original curve and predicted curve and also to test efficient market hypothesis.

1) **Wipro:**

![Fig. 2. Curve generated from real time news sentiment of Wipro](image-url)

![Fig. 3. Original stock price curve of Wipro](image-url)
2) Infosys (25-03-2014 to 25-04-2014):

Fig. 4. Curve generated from real time news sentiment of Infosys

Fig. 5. Original Stock Price Curve of Infosys

3) Unilever (31-03-2013 to 25-04-2014):

Fig. 6. Curve generated from real time news sentiment of Unilever

Fig. 7. Original Stock Price Curve of Unilever

4) Tata Consultancy Services (25-3-2014 to 24-4-2014):

Fig. 8. Curve generated from real time news sentiment of TCS

Fig. 9. Original Stock Price Curve of TCS

5) International Business Machines (28-3-2014 to 25-4-2014):

Fig. 10. Curve generated from real time news sentiment of IBM

Fig. 11. Original stock price curve of IBM
6) **Tech Mahindra**

![Fig. 12. Curve generated from real time news sentiment of Tech Mahindra](image)

Fig. 12. Curve generated from real time news sentiment of Tech Mahindra

![Fig. 13. Original stock price curve of Tech Mahindra](image)

Fig. 13. Original stock price curve of Tech Mahindra

7) **Reliance Industries** (25-03-2014 to 25-04-2014):

![Fig. 14. Curve generated from real time news sentiment of RIL.](image)

Fig. 14. Curve generated from real time news sentiment of RIL.

![Fig. 15. Original curve of Reliance Industries Ltd.](image)

Fig. 15. Original curve of Reliance Industries Ltd.

8) **Nestle India**:  

![Fig. 16. Curve generated from real time news sentiment of Nestle India](image)

Fig. 16. Curve generated from real time news sentiment of Nestle India

![Fig. 17. Original curve of Nestle India.](image)

Fig. 17. Original curve of Nestle India.

9) **General Motors Company**:  

![Fig. 18. Curve generated from real time news sentiment of GM Company.](image)

Fig. 18. Curve generated from real time news sentiment of GM Company.

![Fig. 19. Original curve of GM Company.](image)

Fig. 19. Original curve of GM Company.
10) **Bharat Heavy Electronics Ltd. (BHEL):**

Fig. 20. Curve generated from real time news sentiment of BHEL.

Fig. 21. Original stock price curve of BHEL.

11) **Google Inc:**

Fig. 22. Curve generated from real time news sentiment of Google Inc.

Fig. 23. Original stock price curve of Google Inc.

12) **Hindustan Motors**

Fig. 24. Curve generated from real time news sentiment of Hindustan Motors.

Fig. 25. Original stock price curve of Hindustan Motors.

13) **Apple Inc:**

Fig. 26. Curve generated from real time news sentiment of Apple.

Fig. 27. Original stock price curve of Apple.
IV. EVALUATION

Evaluation of the result was done in order to measure accuracy of the prediction model and also to compare and find out the co-relation between predicted sentiment trends and original stock price movement, if at all any. [33]

In case of all the companies, the news got filtered without leaving out relevant news. The relevancy check of the filtered news from the collected set of news is 100%. The predictive model’s results were analysed in terms of Precision, Recall, Specificity, False Discovery Rate, Accuracy, F1 harmonic factor, Mathew’s Correlation Coefficient, Error in prediction, Mean Absolute Error, MAPE, Root Mean Square Error (RMSE), Coefficient of Variation of RMSE.

The outcomes have been formulated in a 2X2 contingency table or confusion matrix with the following components: True Positive (tp), False Positive (fp), True Negative (tn), and False Negative (fn) of the real time news headlines of different set of companies classified from news collected from the news sources.

<table>
<thead>
<tr>
<th>Confusion Matrix</th>
<th>Target</th>
<th>Positive</th>
<th>Negative</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Positive</td>
<td>330</td>
<td>144</td>
<td>69.62</td>
<td>71.35</td>
<td>70.1</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>53</td>
<td>132</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mathew’s Correlation Coefficient = 0.4  
F1 harmonic factor of this model is 78.1%
The graphs we generate from the predicted trends of sentiment are compared with the movements of original stock curve. And this comparison method yields out a mean correlation of 67.09%. This further passes the effective market hypothesis that says in any liquid market, at any point of time, stock prices reflect all information available in the public domain. This indicator can pass the time with an accuracy of 70.1%. The collected set of news is less positive sentiment will reduce the level of trust and hence the stock price curve will decline. If more positive sentiment becomes more trust grows and that influences a trader or investors choice of investment and participation. Less positive sentiment will reduce the level of trust and hence the stock price curve will decline.

Every time one runs the prediction model it, shows the current sentiment of the company calculated from relevant news headlines and press releases available at that point of time with an accuracy of 70.1%. The collected set of news is real time news; hence all available news gets collected by this model avoiding all sorts of bias or prejudice. That makes the study more perfect.

The study using the predictive model has collected news at the time of the market closure and then analyzed the sentiment on each company. This tracked sentiment of 4 weeks reveals a strong correlation between the original stock price curve and the curve of effective positive index generated from predictive news mining model that has been presented in this paper. This study reveals that there is a very strong correlation of about 67% between news sentiment and original stock price curve. This easily satisfies the test of effective market hypothesis and shows that the sentiment positively reflects on the stock price movement. A market with such a high index indicates towards existence of a semi-strong to a strong efficient market hypothesis.

Although the accuracy level of this predictive model is satisfactory, it can further be developed by incorporating more complex classifiers, and analysis techniques of machine learning or data mining. The future scope of this paper can be varied. We can compare the sentiment of several competitive companies and show how the sentiment around one company impacts on the other companies over the time. Also we can also study other factors ranging from social to legal issues to check whether at all they have any impact on overall market scenario. Using this proposed model, we can do several financial modelling like portfolio management, risk estimation models, and several other strategic modelling etc. where perception plays an important role.

ACKNOWLEDGMENT

We would like to thank Dr. Debika Bhattacharjee, HOD of Dept of CSE of IEM, Kolkata and other faculties for their support during the study of this paper.

REFERENCES

[9] Barber and Odean „All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors: The Handbook of News Analytics in Finance Chapter 7


[29] Basu,S., Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis


[33] The original stock price curves have been taken from Money Control (http://www.moneycontrol.com/) and Yahoo Finance (https://in.finance.yahoo.com/).