Sentiment Analysis of Equities using Data Mining Techniques and Visualizing the Trends

Shradha Tulankar[1], Dr Rahul Athale[2], Sandeep Bhujbal[3]

[1]Department of Advanced Software and Computing Technologies
IGNOU – I[2]IT Centre of Excellence for Advanced Education and Research
Pune, Maharashtra 411 057, India

[2]Sunflower Information Technologies Pvt. Ltd Pune, Maharashtra 411 016, India

[3]Department of Advanced Software and Computing Technologies
IGNOU – I[2]IT Centre of Excellence for Advanced Education and Research
Pune, Maharashtra 411 057, India

Abstract

Markets reflect sensitively towards opinions and sentiments. Investors should be wary of the fact that external factors interact deeply with the share markets and mark their influence over the coming period. National tragedy, low economic growth, IPO releases, RBI decisions on interest rates, foreign matters and many more have an increasing impact on public mindset. Even inflationary markets show a good impact on public opinions and also country's future growth prospects. There is no such technique to beat markets outcome, yet technical analysis and fundamentalist approach give you sustained results. In this paper, we have prepared a model that can predict the current market trends along with accuracy measures based on sentiment analysis. Sentiment analysis is an emerging trend to judge the market highs and lows. Visualization is provided to help end users understand markets properly and decide on their investment strategy. This model uses manual as well as automated sentiment analysis. Research reports are also taken as a mode of comparative assessment for the automated analysis.

Keywords: *Sentiment analysis, sentiments, visualization, predicts, markets, accuracy.*

1.0 Introduction

Investors invest a lot of time in checking out which strategies will be beneficial for them. Lot of research is conducted to explain the investor's behavior, his risk perceptions, market approach, trends of investing, factors affecting his investment and behavior towards the current market trends. Decisions made by a person solely depend on the statistics of the current market and whether the market will be bullish or bearish in coming days. This helps in deciding one's investment strategy [9].

There is a chaos of public sentiments and opinions on market fluctuations. This can be widely seen on chat forums, blogs, message boards or public forums. Facebook and Twitter are major source of such data. Ray Chen and Marius Lazer in their paper have devised a model to investigate the relationship between Twitter feed content and stock market movement [1].

Robert P. Schumaker, Yulei Zhang and Chun-Neng Huang have built and tested the financial news article system that incorporated sentiment analysis techniques in its predictive arsenal [2]. E. Bennet, M. Selvam, Eva Esther Shalin Ebenezer analyses individual investors sentiment. This study also analyses the influence of stock specific factors on investors' sentiment [3].

This project is based on designing an approach to perform sentiment analysis on stocks. The concept involves web-mining, text mining inclusion of sentiment and opinion mining and data mining techniques.

It is a combined approach to model a technique that will analyze the user's comments and live data accessed from web; regarding the fall and rise of the markets on a daily basis. The visualization provides the positive, negative and neutral trend of markets. It is a comparative approach based on the predictions given by brokers and sentiment analysis performed on the live streaming data.

The results obtained from sentiment analysis are compared by visualization techniques with the closing price of EOD of stocks. Comparative study helps users to identify whether to Buy, Sell or Hold the stock.

2.0 Sentiment Analysis on Indian Markets

It would certainly not be a wrong thing to say that Indian markets are emotionally driven. Events like natural calamity, national tragedy, sudden economic growth or even financial budget announcements hold the power to churn out the market scenario. E. Bennet, M. Selvam, N. Vivek, and Eva Esther Shalin who undertook a study on Indian investment business by accessing the retail investors have pointed out that it is not the markets that don't behave neatly but also the individual decision makers who don't behave in accordance with the tenets of expected utility theory [4].

There are many obvious reasons to raise the speculations of how sentiment analysis will help in judging the markets. One of being that there are more number of people who stick to daily trading as compared to long term investment. As research shows; when a particular stock price falls because of news regarding that particular company many investors are averse at selling it at loss. However, there are many who still wait in anticipation that the price will increase and their profits will be gained [9].

Sometimes, technical and fundamental analysis advises to buy particular stocks but for all no reasons the stock is sold in markets at a bulk basis. A thorough research if conducted will also help us in predicting insider trading. In a manner adopting professional sentiment analysis techniques will only help to solve such weird activities.

Various studies have been conducted till date in other countries to predict the current trends. However, the research is still going and the rate of accuracy is also increasing.

3.0 Objective

This thesis concentrates on constructing a model which proves that stocks vulnerability can be predicted using sentiment analysis. We have conducted this research for Infosys stock. Initially we have collected the research reports published by brokers; namely Angel Broking and ICICI Direct Securities for the stock Infosys and month November, 2012 [7]. These reports were manually analyzed. It is not necessary that every time the predictions given by firms have always helped people to earn capital met their goals for all months.

This thesis conducts a survey of messages collected from the message boards[7][8] and applying sentiment analysis on the posts. Sentiment analysis can be done on wide source of data. Data can be in the form of messages, blogs, news corpus, research reports, etc. In this paper, we have used web mining to extract the messages from the websites. The chats that are used for analysis get updated continuously and hence it is very difficult to keep a track of page numbers while performing web mining.

Lastly, we extract the terms that determine positive, negative and hold tendency from the WordList. We have extracted these words manually. Classification, one of the data mining techniques is applied on the words generated to determine the rate of accuracy.

4.0 Constructing the Model

In this section we provide a thorough idea of how the model is built and the algorithm used to test the accuracy of sentiment analysis.

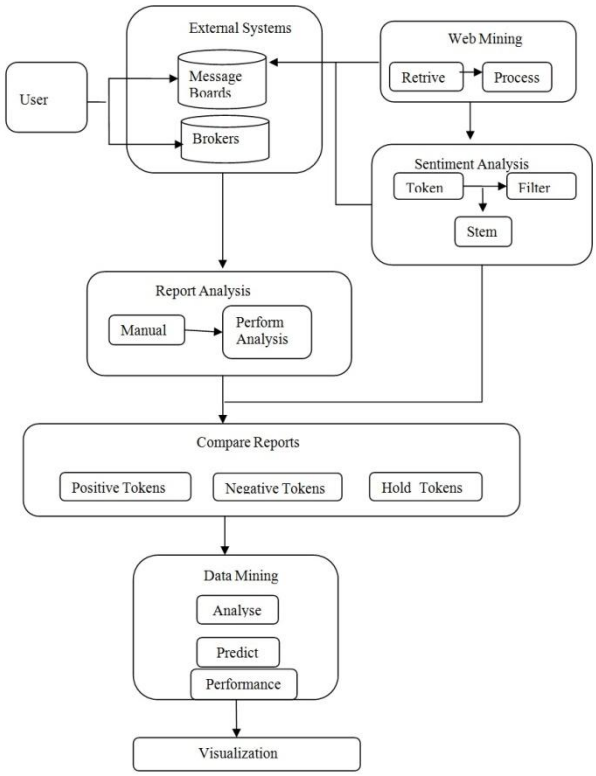

Fig 1: Proposed architecture of the model.

4.1 Data acquisition

The whole process is divided into three stages. Initially, the research is conducted on the data collected from the research reports. The research reports are thoroughly studied and verified manually.

Manual sentiment analysis is performed here to predict if the expectations have met their goals in real or not and if not then what are the specific reasons that went wrong in the analysis. Sentiment analysis is not bounded with the tools. It is mostly done manually and also with the help of tools. It cannot be restricted with a set of tools.

Secondly, we collect the data using web mining. Web mining allows you to look for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines and web spiders. Structure mining is used to examine data related to the structure of a particular Web site and usage mining is used to examine data related to a particular user's browser as well as data gathered by forms the user may have submitted during Web transactions [13].

It is important to verify that the data meets the requirements of business problem. When doing predictive modeling the data also needs to contain the desired outcome. In this case the data was collected on following points:-

- Reviews given by people on the message boards.
- The comments given after trading hours were given higher preference.
- The broker's suggestion for a specific stock was collected.

4.2 Manual analysis of Research reports

Manual sentiment analysis is performed on the research reports collected online. Both the brokers gave the advice to buy Infosys stock for that month and hold it as a longterm investment.

4.3 Extracting contents from URL

Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages [5].

This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. Web mining allows the data to be structured and used for text mining [5].

The data acquisition happens in the form of web links, the data being in unstructured format. The content needs to be converted into documents for further processing. Data preprocessing is a mandatory task as the downloaded data contains all type of scripts, advertisements, codes and unnecessary words.

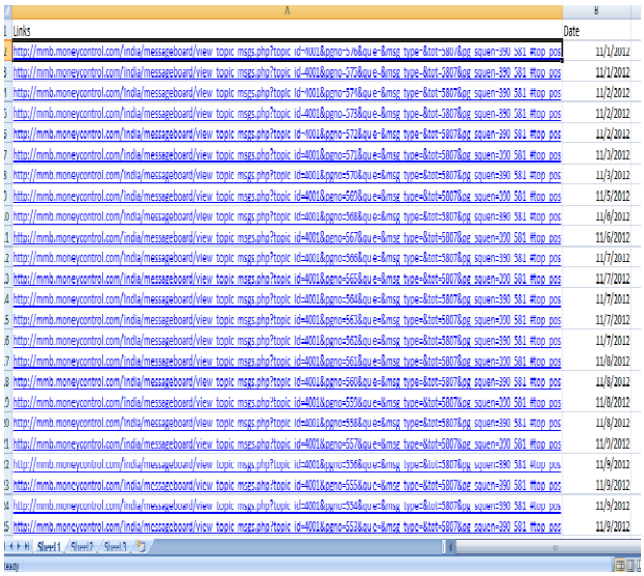

Fig 2: ExampleSet of web links extracted according to date.

In this paper, we have made use of RapidMiner 5.0 along with web mining, text mining and Information Extraction plugins. RapidMiner 5.0 is a open source data mining tool and is widely used for text mining purposes [6].

For performing web mining the data needs to be imported. In this case the data had to be collected from dynamic website so accessing the contents using URL was the best method.

4.4 Generating WordList

The frequency with which particular words are used in a text can tell us something meaningful both about that text and also about its author because their choice of words is seldom random [15]. After extraction of URL, processing of unstructured data is performed and WordList is generated.

Text mining is performed using the Process document operator in Rapid Miner [14]. The links converted into documents are the input to the process document operator. Process document operator applies Tokens, Filter Stopwords, Stemming, generate n-grams and Transform Cases. This operator uses one single TextObject as input for generating a term vector [11].

Transform Cases is used to after converting the words into Tokens. This operator transforms all characters in a document to either lower case or upper case, respectively. This operator is then connected with Filter Stopwords (English). This operator filters English stopwords from a document by removing every token which equals a stopword from the built-in stopword list [11]. Many comments contain regional languages and hence this forms to be impurity for the processing of documents.

Filter Stopwords(English) is connected to N-Grams operator. This operator creates all possible n-Grams of each token in a document. A character n-Gram is defined as a series of characters of length n. The n-Grams of a token generated by this operator consist of all series of characters of this token which have length n. If a token is shorter than the specified length n, the token itself is kept in the resulting document [11].

After Filtering Stemming operator is used. Stemming operator splits the text of a document into a sequence of tokens. There are several options how to split the points; either we may use all non-letter character, which are the default settings. This will result in tokens consisting of one single word, which is the most appropriate option before finally building the word vector [11].

For Stemming purpose we use WordNet dictionary [16]. WordNet is a built in extension in RapidMiner. But with every installation we need to update this plugin also with other files. It is stored in the directory in a repository location of RapidMiner. Porter stemming is conducted on the WordList obtained after WordNet.

This operator stems English words using the Porter stemming algorithm applying an iterative, rule-based replacement of word suffixes intending to reduce the length of the words until a minimum length is reached [11].

The resulting WordList obtained after processing the documents contains 2500 records and generated for 18 models.

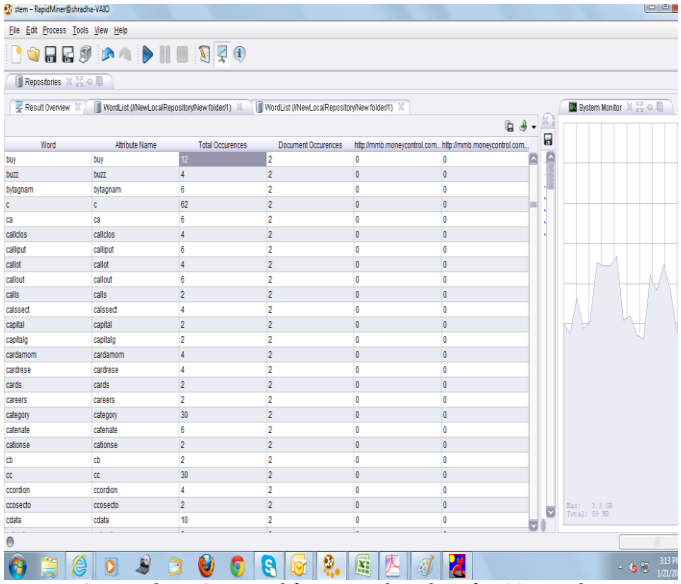

Fig 3:- WordList Generated for one trading day of 2500 records.

5.0 Predicting accuracy of model

Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior. For predicting accuracy we used Weka, open source data mining tool [18].

Initially after automated analysis, from the WordList we extracted the words or phrases that showed the positive, negative or hold effects in market. This dataset was used for prediction.

The confidence measures are derived by applying Random Forest algorithm on the sheet. We obtained an accuracy of 87.5% for positive words. This indicates a strong possibility to buy the stock. The reports in turn were showing a high trend for buying the stocks and showing an upward trend.

6.0 Visualization

The visual data exploration techniques provide a much higher degree of confidence in the findings of the exploration. Visualization for this project was done in Mathematica 8.0 [12]. The visualization generally describes the impact of sentiments on market which is compared with the EOD closing price.

Following is the visualization report for positive, negative and hold terms. For generating WordList binary term occurrences was used instead of TF-IDF. Sometimes it is possible that a set of words will show the sentimental tendency and be more appropriate to judge the outcome of markets.

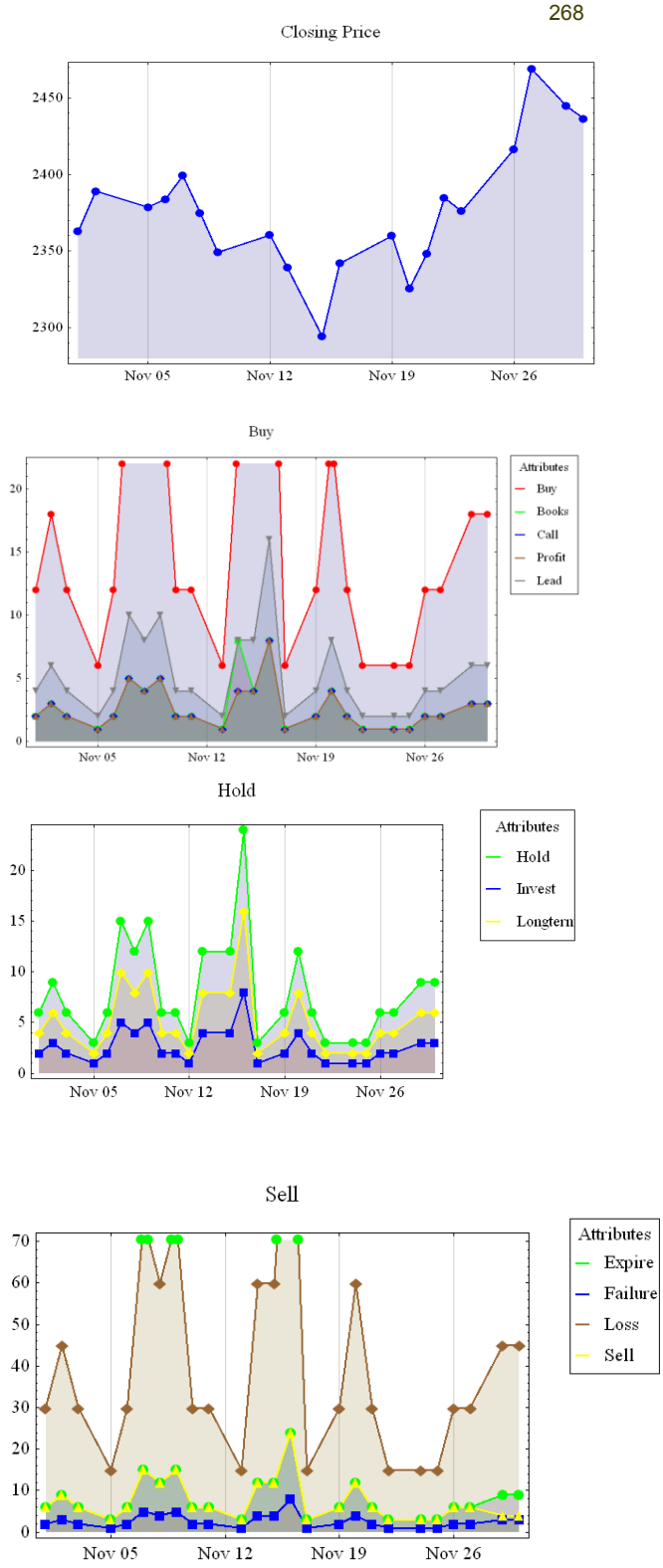

Fig 4:- Visualization graphs based on buy , sell and hold sentiment as compared with EOD closing price.

7.0 Conclusion

The stock has listed the highest number of document pages on 16th November 2012 and hence the profit during this day was highest. Similarly, on 23rd November 2012 it recorded very less documents. This accounted to the fact that the outside controversies prevailing during those days were

highest which had its impacts on Infosys stock during that tenure.

However, for comparative measures the predictions given by brokers are similar to those achieved by sentiment analysis. Hence it can be concluded that the stock had to be bought for that month.

Brokers had given accurate measures to buy the stock for the month of November 2012 and it will reach the level of 2400. The ending price of stock at 30th November 2012 clearly shows that it has recorded profit during that month and the stock can be recorded to positive trend.

Sentiment analysis is used for showing positive, negative and holds trends in the markets. The graphs when compared can be devised easily to show that the positive comments have higher trends in the graph.

In conclusion, the above model can be used for other stocks and hence it can be proved that sentiment analysis can be used to predict buy, hold or sell trends in the markets.

8.0 References

[1] Ray Chen, Marius Lazer, "Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement".
[2] Robert P. Schumaker, Yulei Zhang and Chun-Neng Huang, "Sentiment Analysis of Financial News Articles".
[3] E. Bennet, M. Selvam, Eva Esther Shalin Ebenezer, "Investors' Sentiment influenced by Stock Specific Factors".
[4] E. Bennet, M. Selvam, N. Vivek, and Eva Esther Shalin, "The impact of investors' sentiment on the equity market: Evidence from Indian stock market" , 18 Oct. 2011.
[5] Web DataMining, "Web Data Mining .net", [online] available at http://www.web-datamining.net/content/, 6 April 2013.
[6] Webpedia, "RAPID MINER", [online] available at http://en.webpedia.com/RapidMiner, 15 Mar. 2013.
[7] e-Eighteen.com Ltd, "moneycontrol.com", [online] available at http://mmb.moneycontrol.com/india/ messageboard/view_topic_msgs.php? topic_id=4001& pgno=242&que=&msg_type= &tot=4603&pg_squen=141_461_ #top_pos, 8 January 2013.
[8] IRIS Business Services Limited, "myiris", [online] available at http://myiris.com/newsCentre/ storyShow.php?fileR=20121025111228199&dir=2012 /10/25, 12 January 2013.
[9] Investopedia US, A Division of ValueClick, Inc., "Investopedia" "http://www.investopedia.com /articles/05/032905.asp",05 March 2013.
[10] Michael Wurst, Ingo Mierswa , "The Word Vector Tool and the RapidMiner Text Plugin", Version 5.2, March 14 2009.
[11] Rapid Miner 5.2 User Manual, available online at http://docs.rapid-i.com/files/rapidminer/rapidminer-5.0-manual-english_v1.0.pdf, Copyright 2010.
[12] Stephan Wolfram, "Mathematica Virtual Book", ISBN 1-57955-022-3, 2010.
[13] Types of web mining, "http://searchcrm.techtarget.com/definition/Web-mining", 15 March 2013.
[14] Vancouver Data Blogs for Text Mining; "http://vancouverdata.blogspot.in/2010/11/text-analytics-with-rapidminer-loading.html", 12 March 2013.
[15] WordList ; "http://www.ashgate.com/isbn/9780754672401", 15 April 2013.
[16] WordNet; "http://wordnet.princeton.edu/", 19 April 2013.
[17] Predictive analytics; "http://searchcrm.techtarget.com/definition/predictive-analytics", 05 April 2013 .
[18] Weka; "http://www.cs.waikato.ac.nz/ml/weka/", 20 April 2013.

First Author: She is pursuing **M.Tech** in Advanced Information Technology with specialization in Software Technologies from IGNOU – I²IT Centre of Excellence for Advanced Education and Research, Pune, India. She has pursued **B.E** Computer Science from University of Mumbai. Her research interests include Data Mining, Business Intelligence and Text Mining.

Second Author: He pursued **Ph.D.** from University of Linz Austria. He is Vice President **Sunflower Information Technologies Pvt. Ltd** Pune, India. He was faculty(Adjunct) in IGNOU – I²IT Centre of Excellence for Advanced Education and Research, Pune, India. His research interests include Mathematics, Business Intelligence, Parallel computing and Cloud computing.

Third Author: **Sandeep Bhujbal** is **Sr. Research Associate** in Advanced Software and Computing Technologies department of IGNOU – I²IT Centre of Excellence for Advanced Education and Research, Pune, India. He has pursued **M.C.S** from University of Pune. His research interests include Operating systems, Compiler construction, Programming languages and Cloud computing.