Stock Market Prediction Using Data Mining

[1]Ruchi Desai, [2]Prof.Snehal Gandhi
[1]M.E., [2]M.Tech.
[1]Computer Department
[1]Sarvajanik College of Engineering and Technology, Surat, Gujarat, India

Abstract **- Data mining is well founded on the theory that the historic data holds the essential memory for predicting the future direction. This technology is designed to help investors discover hidden patterns from the historic data that have probable predictive capability in their investment decisions. The prediction of stock markets is regarded as a challenging task of financial time series prediction. Data analysis is one way of predicting if future stocks prices will increase or decrease. Also, it investigated various global events and their issues predicting on stock markets. The stock market can be viewed as a particular data mining problem. Text mining approach is also used for measuring the effect of real time news on stock. It uses different techniques and strategies to predict ups and downs in stock market. In this paper, we present a model that predicts the changes of stock trend by analyzing the influence of non- quantifiable information namely the news articles which are rich in information and superior to numeric data.**

Index Terms **- News articles, Stock market, Text mining**

I. INTRODUCTION

The rapid progress in digital data acquisition has led to the fast-growing amount of data stored in databases, data warehouses, or other kinds of data repositories. Although valuable information may be hiding behind the data, the overwhelming data volume makes it difficult for human beings to extract them without powerful tools. Easy and quick availability to news information was not possible until the beginning of the last decade. In this age of information, news is now easily accessible, as content providers and content locators such as online news services have sprouted on the World Wide Web. Continuous availability of more news articles in digital form, the latest developments in Natural Language Processing (NLP) and the availability of faster computers lead to the question how to extract more information out of news articles.

Financial analysts who invest in stock markets usually are not aware of the stock market behavior. They are facing the problem of stock trading as they do not know which stocks to buy and which to sell in order to gain more profits. All these users know that the progress of the stock market depends a lot on relevant news and they have to deal daily with vast amount of information. They have to analyze all the news that appears on newspapers, magazines and other textual resources. But analysis of such amount of financial news and articles in order to extract useful knowledge exceeds human capabilities. Text mining techniques can help them automatically extracting the useful knowledge out of textual resources.

We would develop a system which is able to use text mining techniques to model the reaction of the stock market to news articles and predict their reactions. By doing so, the investors are able to foresee the future behavior of their stocks when relevant news are released and act immediately upon them. As input we use real-time news articles and intra-day stock prices of some companies in Bombay Stock Exchange. The overall purpose of study can be summarized in the following research questions:

- How to predict the reaction of stock price trend using textual financial news?
- How data and text mining techniques help to generate this predictive model?

In order to investigate the impact of news on a stock trend movement, we have to make a prediction model.

II. BACKGROUND KNOWLEDGE

Knowledge Discovery in Text (KDT)

The term KDT is used to indicate the overall process of turning unstructured textual data into high level information and knowledge, while the term Text Mining is used for the step of the KDT process that deals with the extraction of patterns from textual data. By extending the definition of KDD, the following simple definition is given: Knowledge Discovery in Text (KDT) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in unstructured textual data.

Text Mining (TM) also known as text data mining is a step in the KDT process consisting of particular data mining and Natural Language Processing (NLP) algorithms that produces a particular enumeration of patterns over a set of unstructured textual data. There are various definitions and terminologies for text mining provided by different researchers. KDT is a multi-step process, which includes all the tasks from gathering of documents to the visualization and evaluation of the extracted information.

Influence of News Articles on Stock Market

Market and stock exchange news are special messages containing mainly economic and political information. Some of them are carrying information that is important for market prediction. There are various types of financial information sources on the Web which provide electronic versions of their daily issues. All these information sources contain global and regional political and economic news, citations from influential bankers and politicians, as well as recommendations from financial analysts.

Researchers confirm the reaction to news articles. They have shown that economic news always has a positive or negative effect on the number of traded stock. They used salient political and economic news as proxy for public information. They have found that both types of news have impact on measures of trading activity including return volatility, price volatility, number of shares traded, and trading frequency.

Klibanoff [2] investigate the relationship between closed-end country funds' prices and country-specific salient news. The news that occupies at least two columns wide on The New York Times front-page is considered as salient news. They have found that there is a positive relationship between trading volume and salient news. Mitchell and Mulherin [3] use the daily number of headlines reported by Dow Jones as a measure of public information. Using daily data on stock returns and trading volume, they find that market activity is affected by the arrival of news. They report that salient news has a positive impact on absolute price changes.

Berry and Howe [4], use the number of news released by Reuter's News Service measured in per unit of time as a proxy for public information. In contrast to Mitchell and Mulherin [3], they look into the impact of news on the intraday market activity. Their results suggest that there is a significant positive relationship between news arrivals and trading volume.

III. LITERATURE SURVEY

The first online system for predicting the opening prices of five stock indices (Dow Jones Industrial Average [Dow], Nikkei 225 [Nky], Financial Times 100 Index [Ftse], Hang Seng Index [His], and Singapore Straits Index [Sti]) was developed by Wuthrich [5]. In year 2004, Mittermayer [7] proposed a prediction system called NewsCATS (News Categorization and Trading System). It is a system for prediction of stock price trends for the time immediately after the publication of press releases. NewsCATS consists mainly of three components. The first component retrieves relevant information from press releases through the application of text preprocessing techniques. The second component sorts the press releases into predefined categories. Finally, appropriate trading strategies are derived by the third component by means of the earlier categorization.

In year 2005, Fung and Yu [6] with the corporation of another researcher, Lu, proposed another article named "The Prediction Power of Textual Information on Financial Markets". This is almost the same as the work they proposed in year 2002 with some amendment and modifications. They have recorded the intra-day prices of all the Hong Kong stock and for real-time news stories and more than 350,000 documents are collected.

In 2010, Shou-Hsiung Cheng [1] talks about Taiwan stock market. There are two elements trading Philosophies based on structured data, fundamental and Technical analysis in the stock market. Here first step is to filter out unstructured data and getting the hidden information in a well-defined format for easy decisions making for customers. The method used is to remove Tags, segmentation and to provide speech tags. The decision is made on basis of frequency of word in a particular sentence. The weighting technique is also provided for more accuracy as in case of low frequency important word is there. After this phrase document matrix is constructing for finalizing the decision. After this TO filter this phrase one rough set technique is used and two attributes are classified with this 1.desicion making 2.conditional. Here data sources are used according to some condition given. And outputs are measured at particular slots of time.

IV. PROPOSED WORK

Methodology for NLP module

To exactly predict the stock price is very complex task till the date. Here we are proposing to make a prediction based on news articles using one of the Text Mining concepts like sentiment analysis. We would like to make the prediction system for Indian Stock market. Implementation steps to be followed to make a prediction system are:

1. Gathering of news articles.
2. Perform sentiment analysis on news articles
3. Get Polarity of the text
4. Make a prediction based on current stock price and calculated polarity of the text.

To Get the News Articles

To collect the news articles R.S.S feed is the main source. As R.S.S feed is used for news article collection process. Here the Times of India's R.S.S feed is used for business and market related news. It will give results by retrieving top news of Indian stock market. We have to just specify the R.S.S feed address in our code.

To Perform Sentiment Analysis and Get Polarity of the text

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level - whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.

For sentiment analysis and calculating polarity of text two things are used:

1. POS tagger
2. SentiWordNet_3.0.0

POS Tagger

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. This software is a Java implementation of the log-linear part of-speech taggers.

Part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

POS-tagging algorithms fall into two distinctive groups: rule-based and stochastic. E. Brill's tagger, one of the first and widely used English POS-taggers, employs rule-based algorithms. Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times, and because some parts of speech are complex or unspoken. This is not rare in natural languages (as opposed to many artificial languages), a large percentage of word forms are ambiguous. In part-of-speech tagging by computer, it is typical to distinguish from 50 to 150 separate parts of speech for English. For example, NN for singular common nouns, NNS for plural common nouns, NP for singular proper nouns.

Several downloads are available. The basic download contains two trained tagger models for English. The POS tagger which I have used is developed by Stanford University natural language processing group [8]; it is licensed under the GNU general public license as **it is an open source**.

SentiWordNet_3.0.0

SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. In order to aid the extraction of opinions from text, recent research has tried to automatically determine the "PN-polarity" of subjective terms, i.e. identify whether a term that is a marker of opinionated content has a positive or a negative connotation. The method used to develop SENTIWORDNET is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification.

SENTIWORDNET is freely available for research purposes, and is endowed with a Web based graphical user interface. By downloading the SENTIWORDNET we get a 13Mb XML file which is having the whole English words and its weightage in floating point [9]. Aggregation of each word weightage results us the polarity of that sentence.

Methodology for Statistical parameter based module

A method of evaluating stocks by analyzing statistics generated by market activity, such as past prices and volume. Technical analysts do not attempt to measure a security's intrinsic value, but instead use charts and other tools to identify patterns that can suggest future activity.

Here to do prediction based on statistical parameter like past open & close value, we have collected past few years' data in terms of daily open and close price. We have also calculated the difference of daily open & close to compare the up down price points generally mentioned in news sentences.

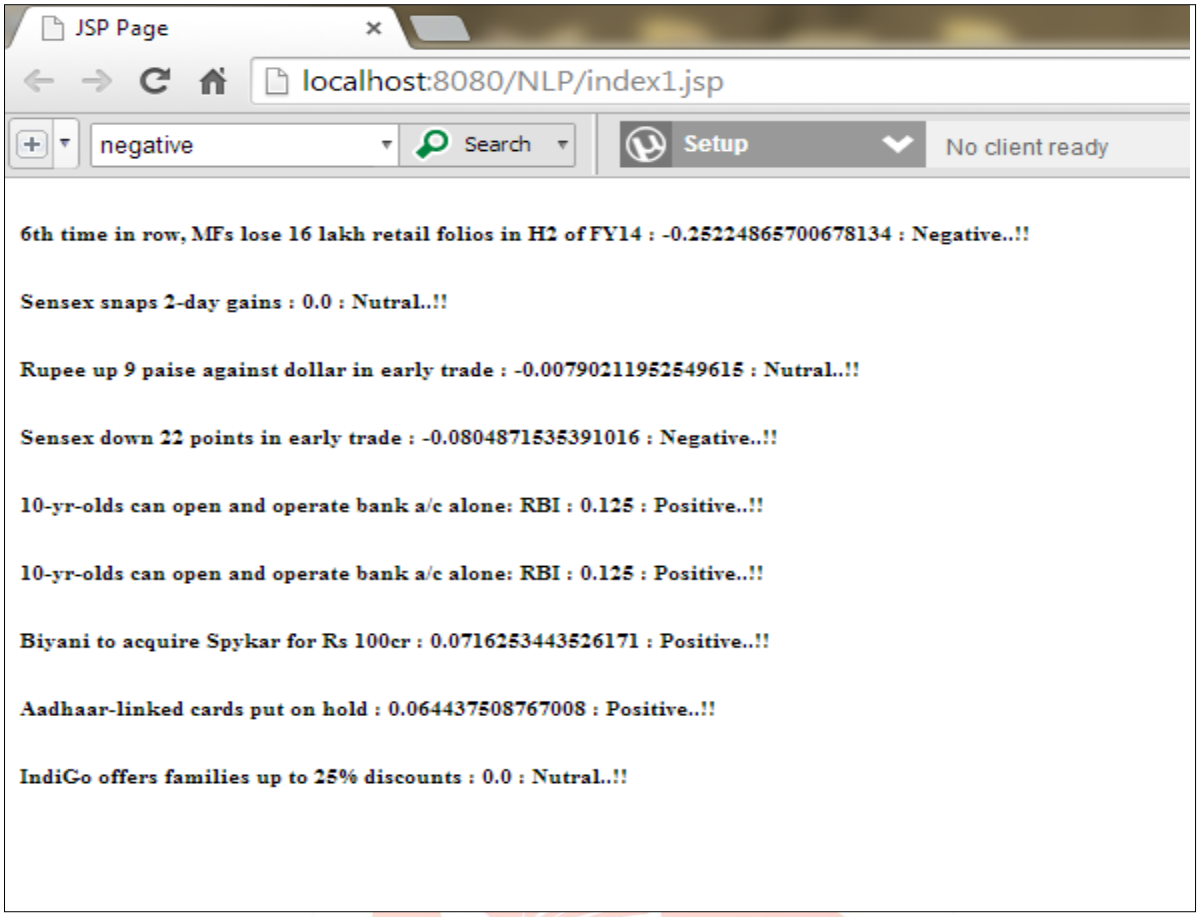
We can get past or historical data by different ways with different sources.

- Web sources
    - From business related news channels: e.g. Times Now, CNBC, Zee business etc.
    - From stock related corporate companies: BSE, NSE, etc.
    - From mediator companies: ShareKhan, Money Control, KARVY, Indiabulls etc.
- Manually from company's database engine

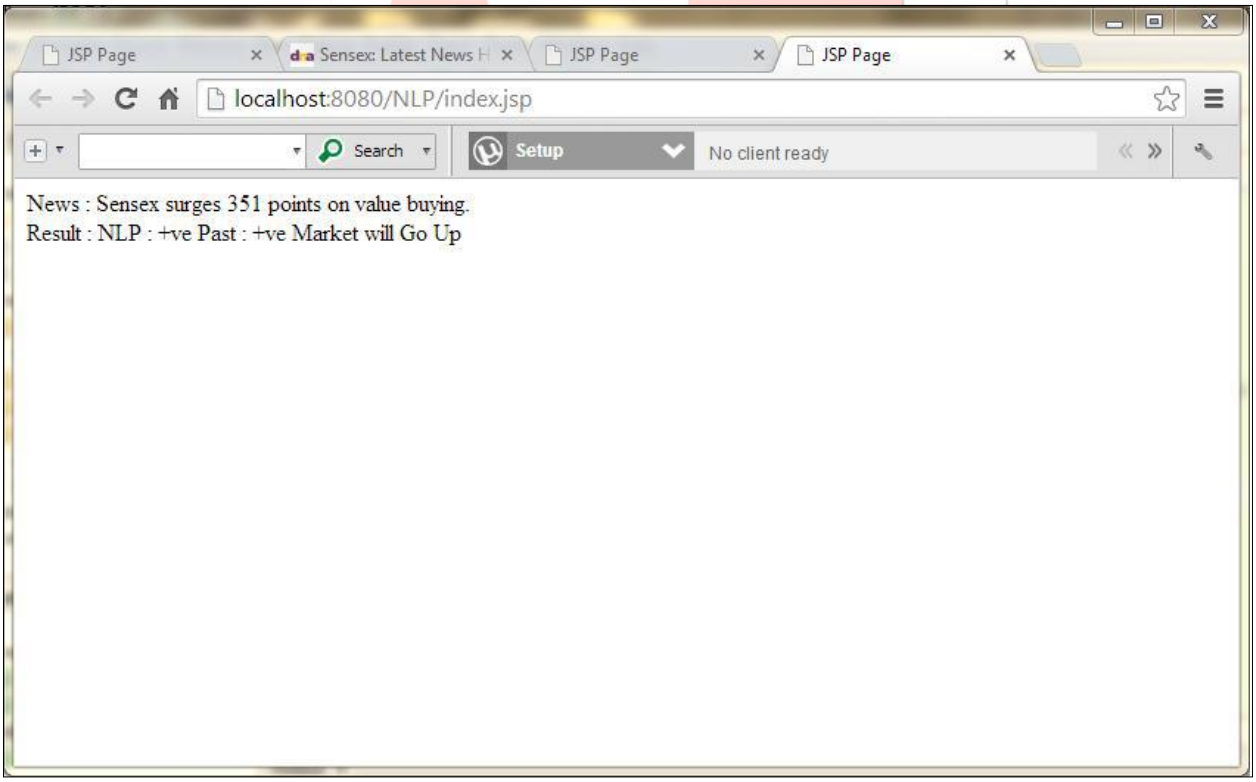
Here we are taking historical data from BSE India web site. The data which we have collected is of past 5 years for accurate results. In exact methodology in this module firstly we are fetching the numerical value of up down points given in news sentence, and taking the range of +50 to -50 to the given points. Then checking for how many time up down occurred in this range in past years' data. And checking its next day change was either positive or negative. Two counters are used one for positive change on next day and another for negative change. Result of prediction for this module is shown based on whether the positive counter or negative counter is greater. Final result is shown with the combination of result of NLP method and statistic method.

V. RESULTS

Result of NLP module



Final result of Statistical parameter based prediction module



VI. CONCLUSIONS

In Data Mining to predict stock market here we have created NLP based module & statistical parameter based module which results the sentence polarity & behavior compared to past year data.

By using this technique we get accurate & reliable prediction result which give consumer better solution for where to invest their valuable money. These modules evaluate the news sentences based on grammatical analysis and with the help of historical data also.

REFERENCES

[1] Shou-Hsiung Cheng, "Forecasting the Change of Intraday Stock Price by Using Text Mining News of Stock", IEEE, 2010.

[2] Klibanoff, P., Lamont, O., and Wizman, T.A., 1998. Investor Reaction to Salient News in Closed-end Country Funds. Journal of Finance, 53(2), pp.673-699.

[3] Mitchell, M.L., Mulherin, J.H., 1994. The Impact of Public Information on the Stock Market. Journal of Finance, 49(3), pp.923-950.

[4] Berry, T.D., Howe, K.M., 1994. "Public Information Arrival." Journal of Finance, 49(4), pp.1331–1346.

[5] Wuthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J., and Lam, W., Daily Stock Market Forecast from Textual Web Data. In IEEE International Conference on Systems, Man, and Cybernetics (San Diego, California, October 11-14, 1998). IEEE Press, Vol.3, pp.2720-2725.

[6] Fung G.P.C., Yu, J.X., and Lu, H., 2005. "The Predicting Power of Textual Information on Financial Markets." IEEE Intelligent Informatics Bulletin, 5(1), pp.1-10.

[7] M. Mittermayer, G. Knolmayer, "NewsCATS: A News Categorization And Trading System", IEEE Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 2006.

[8] "Pos Tagger", http://nlp.stanford.edu/software/tagger.shtml

[9] "Senti WordNet 3.0", http://sentiwordnet.isti.cnr.it