International Journal of Multidisciplinary and Current Research ISSN: 2321-3124
Research Article Available at: http://ijmcr.com

Building an Expert System based on Data Mining

Sagar Bhushan Gawde* **and Umesh Kulkarni**

Department of Computer, Vidyalankar Institute of Technology, Wadala East- 400037, Mumbai, India

Received 01 March 2019, Accepted 02 May 2019, Available online 04 May 2019, **Vol.7 (May/June 2019 issue)**

Abstract

A novel framework for predicting stock trends and making financial trading, decisions based on a combination of Data and Text Mining techniques. The prediction models of the proposed system extract data in text content of time-stamped web documents in addition to traditional numerical time series data, which is also available from the Web. The financial trading system based on model predictions uses three different trading strategies. In this work, our system is simulated and evaluated on real-world series of news stories and stocks data using Decision Tree Induction Algorithm. Performance is the predictive accuracy of the induced models and, more importantly, the profitability of each trading strategy using these predictions.

Keywords: *Expert system, Data mining etc.*

1. Introduction

The Efficient Market Hypothesis (EMH), [7] assumes that Stock rates are adequately reflected at every information at any given point in time. As the basis for growth and development of a modern economy, this means information or analysis can be expected to perform out. The market and that stock prices follow 'Random Walks' [9] where a change in stock price over time is purely random and statistically independent of the stock price in the past. However, to this day no one can explain the anomalies in the market, which can be utilized to assure some short term predictive power [6]. In making their forecasts, most financial specialists try to exploit the time gap of the market's adjustment to new information. They reduce their risk by combining both scientific (base future price predictions on past prices) and theoretical base predictions on real economic facts. Such as inflation, trading volume, organizational changes in the company, etc.) analysis strategies, which are mentioned by Gidofalvi [4]. To obtain the data required by both procedures, here refer to various publicly available resources like the stock market itself, the companies, news articles, etc. A somewhat new source for information in the late 20th and the 21st centuries is, of course, the Internet. To exploit this relatively new and additional tool supporting the forecasting task, we need to combine techniques from both time series data mining and web content mining.

In this work, we present a new system for analysis stock trends based on the combination of Data Mining and Web Content Mining techniques. New Financial Trading System which:

1) Creates a "melting pot" of numeric and textual data before running an induction algorithm,
2) Extracts automatically crucial phrases instead of using a prior expert list of phrases,
3) Eliminates the need for word independence assumption by using Decision Trees rather than Naive Bayes,
4) A new method Influence of news articles in the prediction task to dates

Equations

$$S = 1/3 * TF/N + 1/3(P/L * B/L) + 1/3 * AV \quad \text{(Eq.1)}$$

Where:

L: The time frame, in days, for the word dictionary.
B: The time window between the first and last occurrence of a word.
P: the number of days to the last occurrence of a word.
TF: the number of occurrences of a word during L (known as Term Frequency).
N: The number of words in the dictionary.
AV: The annualized volatility of the stock as calculated by

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2} \quad \text{(Eq.2)}$$

*Corresponding author's ORCID ID: 0000-0002-3906-7080
DOI: https://doi.org/10.14741/ijmcr/v.7.3.6

Where

u_i is the i^{th} observation of the stock price, and u bar is the mean of a stock price.

Whereas the proposed system uses

$$S = 1/3 * TF/N + 1/3(P/L * B/L) + 1/3 * AV \quad \text{(Eq.3)}$$

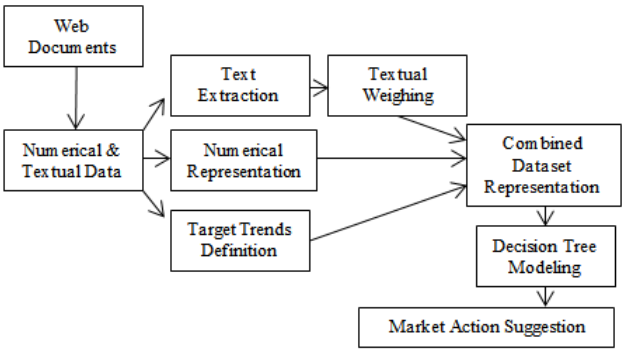

Fig. 1 Proposal with Mixed Numerical and Textual Data

Our system is such a way designed as a full cycle prediction system for stock trends according to past numeric values of the stocks as well as their related textual web articles. It goes through six steps, as shown in Fig. 1, which are:

- Step 1: Data Collection from the Web.
- Step 2: Feature Extraction.
- Step 3: Textual Weighting.
- Step 4: Combined Data-Set Construction.
- Step 5: Classification Model (Decision Tree) Induction.
- Step 6: Market Action Recommendation.

Data Collection from the web itself is a challenging task in itself. In all the standard web scrapping methods we worked through; we found that some programmatic code like HTML or script tag gets included into the scrap and adds to the text noise. Ultimately we found Selenium based automation with chrome to be useful to get exact text content from a website.

Next, we computed features like

L: The time frame, in days, for the word dictionary,
B: The time window between the first and last occurrence of a word,
P: Previous occurred nice of a word number of times, TF: the number of events of a word during L (known as Term Frequency), N: the number of words in the dictionary. These values are used to compute the score which will be used along with other technical benefits of the stock by the Classification Model to get the appropriate action recommendation.

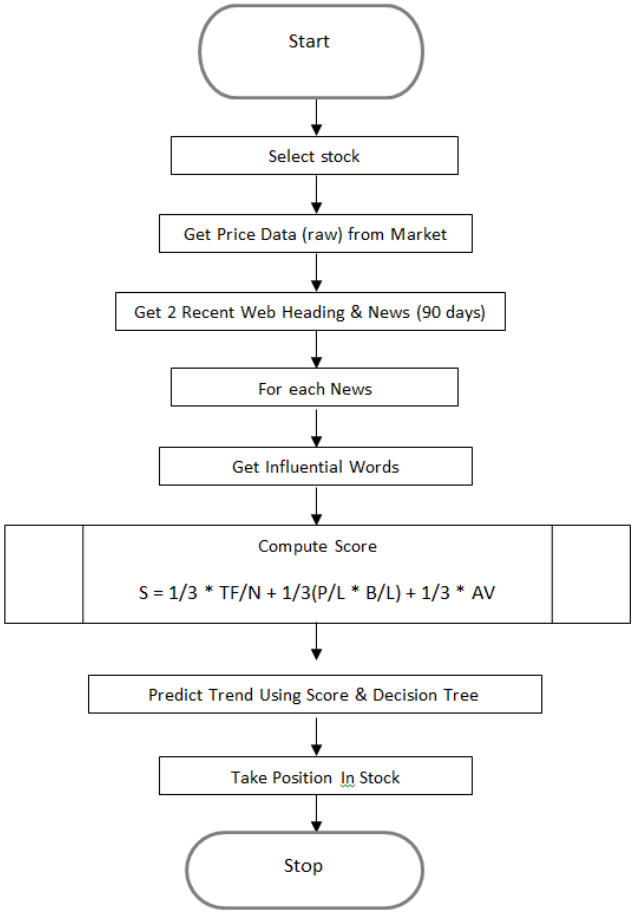

Fig.2 Flowchart

The methods to improve profit include

1. Combination of both numeric and textual data.
2. The use of an automatic text extraction mechanism instead of a predefined expert list.
3. The method of decision tree prediction model.
4. The purpose of smart trading strategies and techniques.

In the flowchart, we compute the score from the features calculated to take into account news articles related to the company. This score is in turn used by the decision tree model to extract the strategy to be followed for making a profit.

Experimental Results

Honest,Businessman,INDUSTRY,India,Wine,Cooler,and,Chest,Freezer,Market,Changing,Lifestyle,Consumption,Pattern,of,Consumers,Drives,Growth,finds,TMR,March,Views,Min,Read,ajinkya,tmrresearch,com,Share,This,Some,the,main,players,in,wine,cooler,chest,freezer,market,are,AB,Electrolux,Haier,Inc,Kieis,Ltd,Elan,Professional,Appliances,Pvt,Westinghouse,Electric,Corporation,Rockwell,Industries,The,Middleby,Western,Refrigeration,Private,Limited,Whirlpool,Williams,Such,Leading,Players,FedEx,UTi,Worldwide,Ryder,System,CEVA,Holdings,Deutsche,Bahn,Agility,Schneider,UPS,Expeditors,APL,SCIENCE,Autonomous,Underw

ater,Vehicle,projected,reach,USD,million,Bluefin,Robotics ,Saab,ECA,Group,About,author,VIEW,ALL,POSTS,Latest,Ne ws,Urban,Gas,Helium,Sales,Supply,Forecasts,COMMERCI AL,AND,SERVICE,MACHINERY,MANUFACTURING,MARKET, GLOBAL,TREND,SEGMENTATION,OPPORTUNITIES,FORECA ST,TO,HEALTH,Industrial,Institutional,Cleaning,Chemicals, register,forecast,Dunnage,Air,Bags,Expected,Witness,Sust ainable,over,Key,Cordstrap,Bates,Cargo,Pak,Stopak,Inc,Co pyright,Created,Meks,Powered,WordPress

Fig.3 Word collection GUI

The above screen shows the collection of keywords from the text content consisting of dated news articles from various websites related to the industry. Subsequently, features computed from these keywords. The news is collected from official sites only which listed in the list box.

From the experiments carried out on the web. Word selected for a selected stock for a specified period; this step includes Textual extraction from authentic sites
The chart of the selected stock can also analyse in this step.

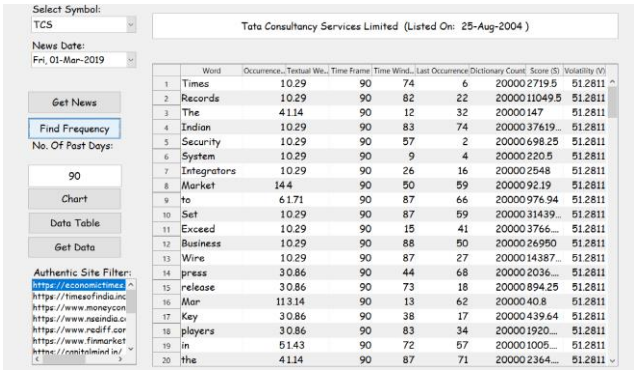

Fig.4 Frequency GUI

In this screen, we compute the features from the keywords collected from the news articles. We also calculate the score from the elements to be subsequently used by the decision tree model.

In this step, the occurrence of the words for a particular period has calculated. Helps us to find Textual weight and score for the same.

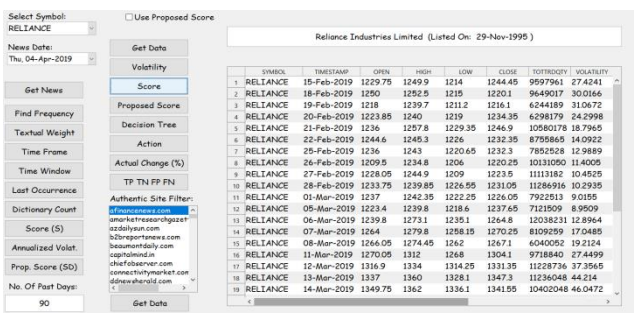

Fig.5 GUI - Data Assembling

Here we consider the technical data of 90 days period of the stock along with annualized volatility. This data s acquired from the above copy file provided by NSE India on its website for each share.

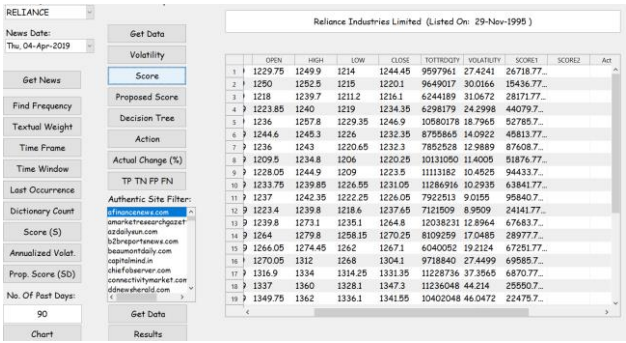

Fig.6 GUI - Score Computation

In this screen compute the score to be passed on to the decision tree model for each of the 90 days. The core data is based on the dated news articles related to the company for each of the 90 days.

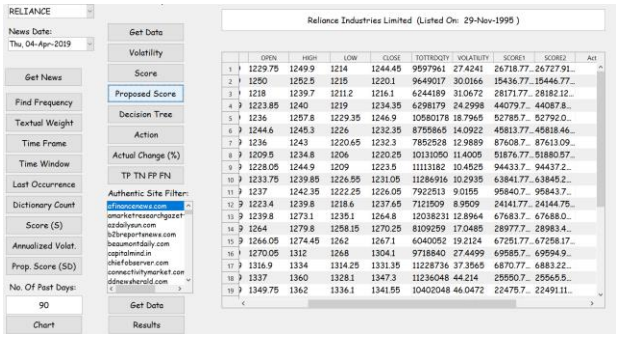

Fig.7 GUI - Proposed Score Computation

Here we compute the proposed score which takes into account an essential factor called as annualized volatility. Which is an indicator of motion in the stock price? If the volatility is high, there is a more upper movement in stock price which needs to be tapped in for profit.

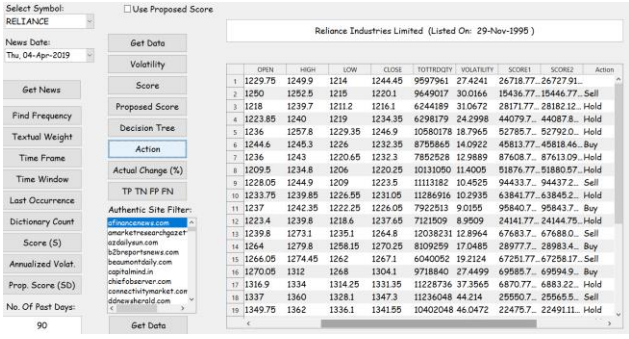

Fig.8 GUI – Action Strategy Computation

Here the computation of action, i.e. the strategy of trading to be followed is shown on each day for the stock

under consideration. Hold means we carry on with the approach adopted on the previous day.

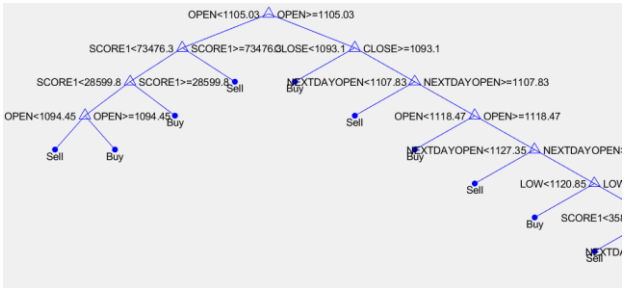

Fig.9 GUI – Decision Tree Model

This screen shows the decision tree constructed based on the technical data values and the computed score. It shows how an action decision arrives.

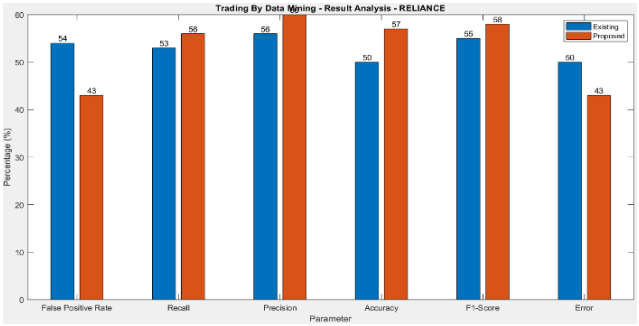

Fig.10 Result Analysis

Here we show the results of the proposed method based on the parameters of False Positive Rate, Recall, Precision, Accuracy, F1-Score and Error Per cent.

Fig.11 GUI – Result Analysis

Here we show the count of each of the parameters like True Positives, False Positives, True Negatives and False Negatives. Here Positive means Buy strategy suggested by the system and Negative means Sell.

Conclusion

Proposed method demonstrates a new method for the advance reading of stock movements and helping financial decisions in trading resting on a mix of financial information and news mining methods. The trend deciding factors of the proposed system make use of text material that dated present in web pages along with customary numeric price time series information, present on the Internet. The proposed trend deciding system makes use of more than one system thereby increasing chances of accurate decisions. The method is simulated and assessed based on actual time series data and news information.

References

[1]. J. Clerk Maxwl. S. Jacobs and C. P. Bean, "Fine particles, thin films, and exchange and T. Bollerslev, Generalized Autoregressive Conditional Heteroscedasticity," Journal of Econometrics, 31, 307-327, 1986.
[2]. T. Bollerslev," A Conditionally Heteroskedastic Time Series Model For Speculative Prices and Rates of Return," Review of Economics and Statistics, 69(3), 542-546, 1987.
[3]. E.F. Fama, Random Walks in Stock Market Prices, Financial Analysts Journal, September/ October 1965 (reprinted in January-February 1995).
[4]. G. Gid6falvi, 2001. Using News Articles to Predict Stock Price Movement, Online at, [http://citeseer.nj.nec.com/ 517027.html].
[5]. V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, Language Models for Financial News Recommendation, CIKM 2000, McLean, VA USA, ACM 2000.
[6]. M.A. Kaboudan, Genetic Programming Prediction of Stock Prices, Computational Economics 16: 207-236, 2000.
[7]. E.F. Fama, Long Term Returns and Behavioral Finance, Social Science Research Network.
[8]. M. Last, Y. Klein and A. Kandel, Knowledge Discovery in Time Series Databases, IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, Vol. 31 No. 1, February 2001.
[9]. Z. Bodie, A. Kane, A.J. Marcus, Investments, 4th Edition, McGraw Hill, 2001.
[10]. E. F. Fama, Efficient Capital Markets: A Review of Theory and Empirical Work, Journal of Finance, 25 (May 1970): 383-417.
[11]. R.A. Huagen, The New Finance: The Case Against Efficient Markets. Prentice-Hall, 1995.
[12]. R. Engle and T. Bollerslev, "Modelling the Persistence in Conditional Variances," Econometric Reviews, 5, 81 -87, 1986.
[13]. A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, Volume 31, No. 3, September 1999.
[14]. R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," In Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Newport Beach, CA, November 1997.
[15]. R. Kosala and H. Blockeel, "Web Mining Research: A Survey, SIGKDD Explorations," Volume 2, Issue 1.
[16]. O. Maimon, A. Kandel and M. Last, Knowledge Discovery and Data Mining, The Info-Fuzzy Network (IFN) Methodology, Norwell, MA: Kluwer, 2000.
[17]. D. Peramunetilleke, R.K. Wong, "Currency Exchange Rate Forecasting from News Headlines," Thirteenth Australasian Database Conference (ADC2002), Melbourne, Australia, Conferences in Information Technology, Vol. 5.
[18]. B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, W. Lam, "Daily Stock Market Forecast from Textual Web Data," In IEEE International Conference on Systems, Man. and Cybernetics, Volume: 3, Page(s): 2720 -2725, 1998.
[19]. L. Torgo, The TNT Financial Trading System: a midterm report, ECML-PKDD Workshop on Data Mining for Business, 2005.
[20]. J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman Publishers Inc., San Francisco, CA, 1993.
[21]. R. Landry Jr., R. Debreceny, G.L. Grey, "Grab Your Picks and Shovels! There's Gold in Your Data, Strategic Finance", January 2004, (85, 7).
[22]. Extractor DBI technologies (2003) [http://www.dbitech.com]