



Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Information and Engineering

<http://www.elsevier.com/locate/fiae>



ORIGINAL ARTICLE

A Novel Fuzzy Document Based Information Retrieval Model for Forecasting



Partha Roy · Ramesh Kumar · Sanjay Sharma

Received: 25 August, 2015 / Revised: 12 January, 2017 /

Accepted: 3 May, 2017 /

Abstract Information retrieval systems are generally used to find documents that are most appropriate according to some query that comes dynamically from users. In this paper a novel Fuzzy Document based Information Retrieval Model (FDIRM) is proposed for the purpose of Stock Market Index forecasting. The novelty of proposed approach is a modified tf-idf scoring scheme to predict the future trend of the stock market index. The contribution of this paper has two dimensions, 1) In the proposed system the simple time series is converted to an enriched fuzzy linguistic time series with a unique approach of incorporating market sentiment related information along with the price and 2) A unique approach is followed while modeling the information retrieval (IR) system which converts a simple IR system into a forecasting system. From the performance comparison of FDIRM with standard benchmark models it can be affirmed that the proposed model has a potential of becoming a good forecasting model. The stock market data provided by Standard & Poor's CRISIL NSE Index 50 (CNX NIFTY-50 index) of National Stock Exchange of India (NSE) is used to experiment and validate the proposed model. The authentic data for validation and experimentation is obtained from <http://www.nseindia.com> which is the official website of NSE. A java program is under construction to implement the model in

Partha Roy (✉)

Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, Chhattisgarh-491001, India

email: patsroy@gmail.com

Ramesh Kumar

Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, Chhattisgarh-491001, India

Sanjay Sharma

Department of Applied Mathematics, Bhilai Institute of Technology, Durg, Chhattisgarh-491001, India

Peer review under responsibility of Fuzzy Information and Engineering Branch of the Operations Research Society of China.

© 2017 Fuzzy Information and Engineering Branch of the Operations Research Society of China. Hosting by Elsevier B.V. All rights reserved.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

<http://dx.doi.org/10.1016/j.fiae.2017.06.002>

real-time with graphical users' interface.

Keywords Japanese Candlesticks · Fuzzy logic · Tf-idf · Forecasting · Data mining · Information Retrieval

© 2017 Fuzzy Information and Engineering Branch of the Operations Research Society of China. Hosting by Elsevier B.V. All rights reserved.

1. Introduction

Predicting or forecasting is both an art as well as science. The process and outcome of the forecasting has long been a matter of research and still is in its childhood state. We can devise numerous ways of modeling a phenomena and predict its outcome, but there are no universal methods to model every phenomena. Modeling of linear systems is comparatively simpler than dynamical systems. Stock markets are completely chaotic and dynamic systems which are both time and sentiment driven. The time series generated through stock market data can only represent a financial time series of prices but cannot represent the overall sentiment of the market players who trade and invest in the stock markets. Hence modeling of stock market data is one of the toughest as it should incorporate not only data but market sentiment also.

The stock market data is a series of prices that are observed in a series of certain time intervals (minutes, hours, days or weeks etc). Data mining serves as a potential tool to build models that can use past behavior of price movement to predict the future. Fuzzy logic is another tool which can be used effectively to create models that can capture market sentiments. By adopting a hybrid approach of combining Time series, Data mining and Fuzzy Logic an efficacious system can be built to model the stock market price data that can not only give information about price but also the market sentiment or the mood of the market participants. Stock market prediction is one of the most researched and discussed fields due to its criticality in commercial applications and attractive benefits.

Forecasting in itself is intriguing and if money is involved then its interestingness increases many folds. Financial time series are the toughest to forecast as, the modeling of such time series governs the quality of results achieved. The same financial time series would fetch better results if it is modeled appropriately rather than taking the time series as it is.

Soft computing presents us with a wide variety of options to model any dynamic system as it is adapted from physical science. The problem solving through appropriate modeling of the observed system using soft computing and artificial intelligence is very effective. As these systems are intelligent and tolerant to imprecision and uncertainty, making them most adaptable to noisy realms. Soft computing encompasses three key areas of Probabilistic Reasoning, Neural Networks and Fuzzy Logic. The fuzzy logic area of soft computing is adopted in the proposed model. The property of fuzzy logic system to capture the market sentiment from the price, helped to build a linguistic time series that not only represents the actual time series but expose and extract a lot of hidden information from the same crisp time series.

The IR systems try to find the most appropriate and relevant documents depending upon the query. This quality of the IR systems helped to build a model that

would suggest the most appropriate future trend. A novel fuzzy document based information retrieval model (FDIRM) is proposed for the purpose of stock market index forecasting. In the proposed system the entire document corpus is generated by using a fuzzification process and the queries containing fuzzy terms would be processed by the proposed system to fetch the most appropriate document from document corpus.

The novelty of the approach followed here is that the trend is represented as a document and the query consists of the fuzzy linguistic terms representing the current state of the financial time series. This approach gives an entirely new dimension of looking at how traditional IR systems are used. The tf-idf scoring scheme is used to complete the task of forecasting.

The contribution of this paper has two dimensions, 1) In the proposed system the simple time series is converted to an enriched fuzzy linguistic time series with a unique approach of incorporating market sentiment related information along with the price and 2) A unique approach is followed while modeling the IR system which converts a simple IR system it into a forecasting system. Transaction data used to validate the proposed model is obtained from CNX NIFTY-50 index of NSE.

1.1. About Japanese Candlestick Theory

In stock market trading line and bar charts are mainly used to visualize the stock price action during trading sessions. A line chart or line graph displays information as a series of data points (generally the daily closing prices of a stock) connected by straight line segments. A bar chart consists of vertical lines (bars) drawn consecutively on the x-axis. The bars represent open, high, low and close prices of a stocks’ trading session. In both line and bar charts y-axis represents price values and x-axis depicts time or trading session. Japanese candlestick charts are created by combining information present in line and bar charts. According to the Japanese candlestick theory the value representing the difference between a trading sessions’ open and close values represent the body of a candle. The low and high values represent the extreme ends emerging from the body of a candle, called the wicks or shadows of the candle. Figure 1 illustrates a typical candlestick formation, here when the trading sessions’ close value is lower than the open value then the candle is filled with any dark color and if the close value is higher than the open value then the candle is filled with white color.

Japanese candlesticks present us with more than one dimension of understanding

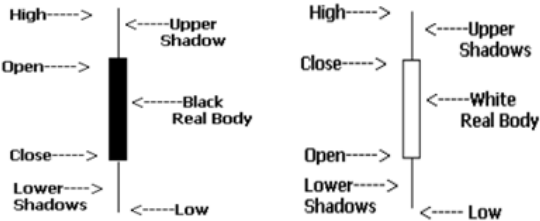


Fig. 1 Typical Japanese Candlestick Formation

the current market condition. The first dimension is the price, when close is higher

than the open then the market is moving upwards i.e., the buyers are outnumbering the sellers and the trend is bullish causing the market values to go up. Similarly, when close is lower than the open then the market is moving downwards i.e., the sellers are outnumbering the buyers and trend is bearish causing the market values to go down. The second dimension is the length of body in the candlestick formation, if the body length is very big then it means that the market sentiment is very strong, either bullish or bearish. If the body length is small then there exists some kind of uncertainty or indecision in the market and implies that the market may try to attain some direction in the coming trading sessions.

1.2. About Fuzzy Logic Theory

According to L.A. Zadeh [11] a fuzzy set A over a universe of discourse X is a set of pairs:

$$A = (x, \mu A(x)) : x \in X, \mu A(x) \in [0, 1],$$

where, $\mu A(x)$ is called the membership degree of the element x to the fuzzy set A . This degree ranges between the extremes 0 and 1:

$\mu A(x) = 0$ indicates that x in no way belongs to the fuzzy set A .

$\mu A(x) = 1$ indicates that x completely belongs to the fuzzy set A .

In the proposed model the concept of fuzzy logic is implemented, to model the approximate nature of the fuzzy candlestick time series as bullish, bearish, etc. and candlestick formations like Tiny, Very Small, Big, Very Big etc.

1.3. About Tf-idf Scheme

In information retrieval systems, the main intention is to retrieve that document which is most relevant to the query posed. The query and the documents both are a collection of terms. Terms are words that we use in our day to day speaking and writing. A scheme known as tf-idf (term frequency and inverse document frequency) is used to assign weights to the documents according to the query. The terms appearing in both query and documents are used as the basis of calculations done in this scheme. The document corpus or simply corpus is used to represent the collection of all the documents present for evaluation. A document would consist of lines of text and every line would consist of words and these words are known as terms. Similarly, every query would consist of a line of text that also would contain words or terms that is needed to be searched from the document corpus.

Term frequency $tf_{t,d}$ is the occurrence count of a query-term t that appears in the respective document d . The log frequency weight $\omega_{t,d}$ of the terms is simply the log of the term frequencies calculated for each term in the document. The normalized value of the log frequency weight $\omega_{t,d(norm)}$ is used in further calculations. The inverse document frequency idf_t is calculated by taking log of value achieved by dividing the total number of documents N by the df_t which is the document frequency of t in the specific document corpus. The normalized value of log frequency weight, $idf_{t(norm)}$, is used in further calculations. Normalized values are used for length normalization of the column vectors. Using the normalized vectors, the cosine similarity between the query vector and document vector is calculated. The final tf-idf score is the sum of the values in the column that represents the cosine similarity between the query vector

and document vector. The document which has the highest score is more relevant to the posed query, than others.

2. Background and Literature Review

E. F. Fama [3] introduced the Efficient Market Hypothesis and according to him the stock markets are random walks and previous prices cannot be used to predict future prices, however there are plenty of evidences proving that stock markets are predictable to a certain extent. According to A. Bagheri et al. [2] the investors and traders in the stock markets use two types of tools for forecasting, one is the fundamental analysis and second is technical analysis. Fundamental analysis uses information gathered from business and economic structure of the company and its related markets, to predict the future stock prices of the company. Technical analysis uses the information concealed in past stock prices to predict the future. Our approach is purely based on technical analysis.

Y. Zhang, L. Wu [13] proposed a novel approach of combining back-propagation neural network with an improved Bacterial Chemo-taxis Optimization (IBCO) for stock market data forecasting. Y. Hu et al. [5] proposed a hybrid approach by combining short term and long term trend following systems with extended classifier system for extraction of rules which selects stocks by different indicators. L. Wang et al. [9] proposed fuzzy time series for stock market prediction where the data is fuzzified to the cluster centers. H. Yu et al. [10] suggested that the selection of the representative features in creation of the rules is the governing factor for better forecasting results. L. Paulev, H. Jgou and L. Amsaleg [7] suggested that the existing information retrieval hashing schemes rely on structured quantizers which poorly fit the real data sets. They put forth a comparison of various space hashing functions. They concluded that for very large data sets query adaptive KLSH gives the highest recall for a fixed views selectivity. R. Salakhutdinov and G. Hinton [8], proposed a model describing a process of finding binary codes that can be used for fast document retrieval. The document is divided into layers and the lowest layer represents that word-count vector and highest layer constitutes the binary code learnt by the proposed system. They used back propagation neural networks for this purpose. W. Zhang, T. Yoshida and X. Tang [12] presented some experimental evaluations of indexing methods on text classification. Analyzed that, presently we don't have a standard measure to assess the semantic and statistical qualities of text.

Z. E. Attia, A. M. Gadallah and H. M. Hefny [1] proposed a linguistic based multi-view fuzzy ontology information retrieval model. Their proposed model allow users to define all their linguistic terms according to their subjective view which helps in retrieving documents according to their linguistic terms definitions not to our definitions. The resulted documents are ranked according to users' defined criteria. Y. Gupta, A. Saini and A. K. Saxena [4] proposed a new ranking function for information retrieval using fuzzy logic. The use of fuzzy logic increases the performance of the system. The fuzzy system incorporates term frequency, inverse document frequency and normalization. T. Korol [6] designed a fuzzy logic system that works by creating a knowledge-base containing fuzzy rules. The fuzzy rules are created by gathering experiences of various traders and investors. The author used 10 years of

gathered experience to generate a fuzzy rule base. The rules are formed on the basis of fundamental analysis done by the actual traders and investors. The literature review helped in the genesis of following concepts: i) It was found that forecasting is a complex process especially for financial time series. ii) The amount of information that a time series contains, if it is fully extracted, then only the forecasting algorithm can generate more accurate results. iii) The purpose of information retrieval schemes used at present are limited to assigning scores to the documents and identifying the most appropriate document from the corpus according to the query. So, they are not used for any kind of forecasting purposes.

3. Research Design

Following research design steps emerged from ideas generated through the literature review process:

i) The time series needs to be modified so that maximum possible information could be incorporated in it. Hence, maximum possible information represented using Japanese candlestick charts of the financial time series is to be fuzzified, because by using fuzzy logic the hidden information present in the candlestick charts, related to the market sentiments can be deciphered. Hence the proposed representation of financial time series is more information-rich than any other way of representation. ii) The information retrieval schemes have a latent property of predicting the most appropriate document based on the query posed, this latent property can be extracted out by modifying the information retrieval scheme so that it can be used as a forecasting tool.

Figure 2 represents the proposed methodology that is used while implementing the research process. The raw data is the Open, High, Low and Close of values of every day, together known in abbreviation as OHLC values. The OHLC values are again represented in the form of Japanese candlestick charts. The time series data is converted to fuzzy linguistic time series containing information enriched fuzzy time series elements.

The fuzzy information enriched time series is used to develop fuzzy document corpus, simultaneously fuzzy queries are also developed which would be used in the information retrieval process. The fuzzy query processing is done by using the tf-idf information retrieval scheme. Modifications are performed in the generation of documents and implementation of the tf-idf scheme resulting in a fuzzy document based information retrieval system. The results achieved through these processes; gives the forecasted output.

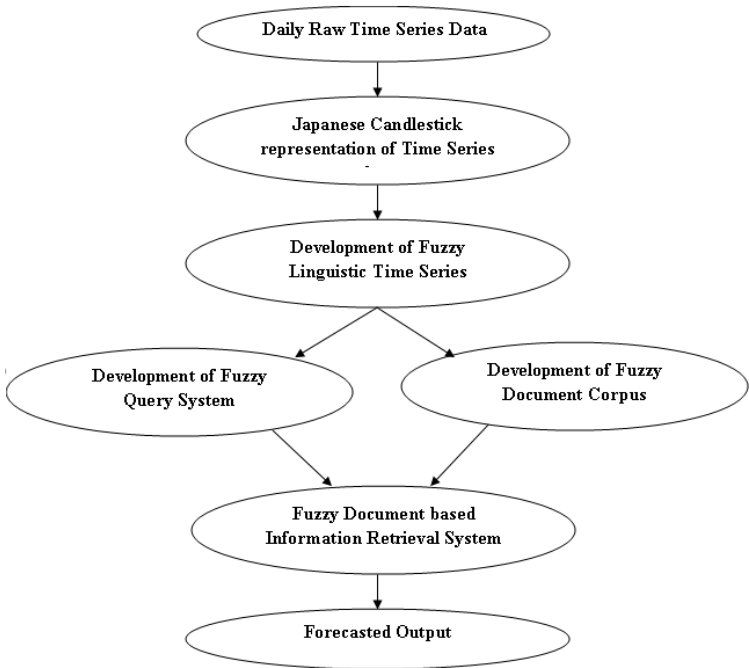


Fig. 2 Proposed methodology

4. Method

The method consists of three phases. In phase 1 the fuzzification of the stock market index time series data is done. In phase 2 the IR model is prepared by using a modified tf-idf approach and in phase 3 the modified tf-idf scheme is used to design queries which can be given to the proposed IR model for forecasting.

Figure 3 displays the candlestick chart in which the points p and p_1 are indicated. The information about these two points is mentioned in the methodology stated below. Following is a brief description of the development phases.

Phase 1: Fuzzification Process

1) Fuzzify the candlestick formations of daily observations of the stock market index time series by fuzzification of the attributes Upper Shadow (US), Body (BD), Lower Shadow (LS) and Candle Color (CC) for each day of observation (Figure 1). The information contained in US, BD, LS and CC are necessary to enrich the time series because the size of the ‘Upper Shadow’ represents the sentiment of buyers (also known as bulls) in the market who are trying to pull the values in the upward direction, the size of the ‘Lower Shadow’ represents the sentiment of sellers (also known as

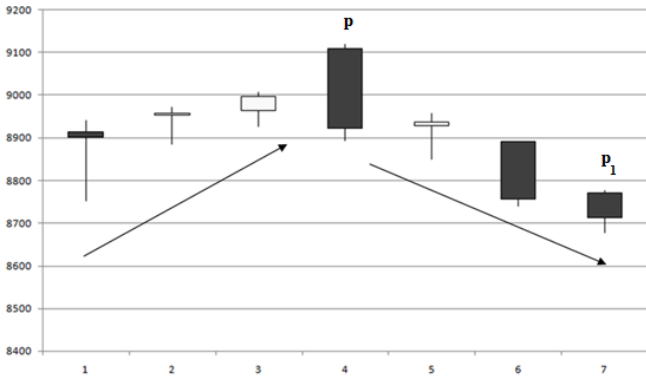


Fig. 3 Candlestick chart showing the points p and p_1 in the time series

bears) in the market who are trying to pull the values in the downward direction, the size of the 'Body' represents the intensity of the market sentiment and 'Candle Color' represents whether the sentiment is gaining strength or losing strength, so if the candle color is black then sellers are gaining on buyers (bearish sentiment is increasing) and if candle color is white then buyers are gaining on sellers (bullish sentiment is increasing).

2) Fuzzify the trend of closing values before and after a particular point p (Figure 3) in the time series into three fuzzy categories of trend namely BR (bearish: values going down), NT (neutral: values remaining range bound) and BL (bullish: values going up).

Phase 2: Information Retrieval using Modified Tf-idf Scoring Scheme

1) The trend formed after the point p will be any one of BR, NT or BL. These would form the three categories of documents BR, NT and BL. Every entry in the documents BR, NT and BL would again be considered as individual documents. This method is a unique approach and is different from the basic tf-idf scheme.

2) The terms in the IR system would constitute of two fuzzy terms, firstly, the trend (BR, NT or BL) formed till the point p and secondly, the fuzzy attributes (US,BD,LS,CC) of the candle formed at the day p in the time series.

Phase 3: Forecasting using Modified Tf-idf Scoring Scheme

1) The query would constitute of two terms, firstly, the trend formed till any point p_1 (Figure 3) in the time series and secondly, the fuzzy attributes (US,BD,LS,CC) of the candle (price bar) formed at the day p_1 .

2) The tf-idf weights of the documents (BR,NT,BL) with respect to the terms in the query is calculated.

3) The document with the highest tf-idf score represents the most probable trend in the future that we can expect after the point p_1 in the time series.

Details of Every Phase

4.1. Phase1: Fuzzification Process

4.1.1. Fuzzification of the Candlestick Formations in the Time Series

To represent the attributes of each candlestick in the time series, five fuzzy linguistic terms TNY, VS, SM, BG and VBG are used. These terms mean tiny, very small, small, big and very big respectively. The fuzzy linguistic terms are used to individually represent the three attributes of every candlesticks Upper Shadow, Lower Shadow and Real Body. And a binary representation for the color of the candlestick (CC) as B and W to represent black and white colors respectively. For example, let there be a candlestick bar which has been analyzed by the proposed model and it generated the fuzzy representation of the candlestick bar as TNYTNYBGW, then it would be interpreted as the Upper Shadow is tiny (TNY), the Lower Shadow is tiny (TNY), the Real Body is big (BG) and the color of the candlestick is white (W). The fuzzy arithmetic for the above representations is as follows: Let,

X_i^j represents j value (Open, High, Low or Close values) on i^{th} day. Here j represents OP, HI, LO or CL values (which are Open, High, Low or Close values) respectively for the i^{th} day,

D_i^{jk} represents nonnegative distance between j (Open, High, Low or Close values) & k (Open, High, Low or Close values) values on the i^{th} day.

$$D_i^{jk} = |X_i^j - X_i^k|. \quad (1)$$

The color of the candlestick is determined by the difference between close and open, represented by Equation (2), where C_i is the color of the i^{th} candlestick.

$$C_i = \begin{cases} \text{Black, } X_i^{CL} < X_i^{OP}, \\ \text{White, } X_i^{CL} \geq X_i^{OP}. \end{cases} \quad (2)$$

The crisp value for the BD attribute of the i^{th} candlestick is represented as D_i^{OPCL} which is determined by Equation (3) and it is the non negative difference between the open and close values of each day,

$$D_i^{OPCL} = |X_i^{OP} - X_i^{CL}|. \quad (3)$$

The universe of discourse U is chosen as the collective average of the distance between the open and close values of every day in the considered range of consecutive observations. The Universe of discourse will be determined by AD^{OPCL} in Equation (4) representing the average of the difference between the open and close values for n consecutive observations. The difference of open and close values is taken because they represent crucial sentimental strength of the market direction. The value of n should be taken as required, for our experimentation the value of n is 7,

$$AD^{OPCL} = \left[\sum_{i=0}^{n-1} D_i^{OPCL} \right] / n. \quad (4)$$

The crisp value for the Upper Shadow attribute of the i^{th} candlestick represented as US_i is determined by Equation (5).

$$US_i = \begin{cases} D_i^{OPHI}, & C_i = \text{Black}, \\ D_i^{CLHI}, & C_i = \text{White}, \end{cases} \quad (5)$$

where,

$$D_i^{OPHI} = |X_i^{OP} - X_i^{HI}|, \quad (6)$$

$$D_i^{CLHI} = |X_i^{CL} - X_i^{HI}|. \quad (7)$$

The crisp value for the Lower Shadow attribute of the i^{th} candlestick represented as LS_i is determined by Equation (8).

$$LS_i = \begin{cases} D_i^{CLLO}, & C_i = \text{Black}, \\ D_i^{OPLO}, & C_i = \text{White}. \end{cases} \quad (8)$$

here,

$$D_i^{CLLO} = |X_i^{CL} - X_i^{LO}|, \quad (9)$$

$$D_i^{OPLO} = |X_i^{OP} - X_i^{LO}|. \quad (10)$$

To convert crisp values to fuzzy linguistic terms we use the following membership functions:

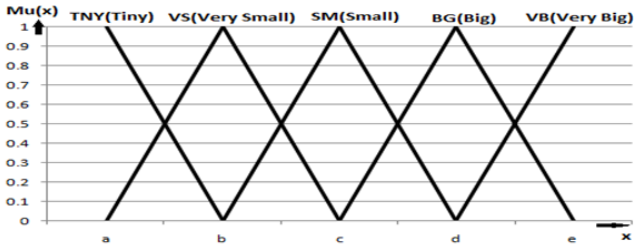


Fig. 4 Membership function

A graphical representation of the combination of Z function, triangular function and inverse Z function used as membership functions in our proposed system and is presented through Figure 4. The x-axis represents the crisp values represented as x which is any one of the candlestick attributes D_i^{OPCL} or US_i or LS_i at a time which is taken into consideration for generating fuzzy linguistic representations and y-axis represents the equivalent membership grades in the fuzzy linguistic categories namely

TNY, VS, SM, BG and VB which are realized by Equations (11) to (15). The values of a, b, c, d and e are taken as 15%, 30%, 45%, 60% and 75% of AD^{OPCL} respectively.

The mathematical representation of Figure 4 is as follows:

$$Mu_{TNY}(x) = \begin{cases} 1, & x \leq a, \\ \frac{b-x}{b-a}, & a < x < b, \\ 0, & x \geq b. \end{cases} \quad (11)$$

$$Mu_{VS}(x) = \begin{cases} 1, & x = b, \\ \frac{x-a}{b-a}, & a < x < b, \\ \frac{c-x}{c-b}, & b < x < c, \\ 0, & c < x \leq a. \end{cases} \quad (12)$$

$$Mu_{SM}(x) = \begin{cases} 1, & x = c, \\ \frac{x-b}{c-b}, & b < x < c, \\ \frac{d-x}{d-c}, & c < x < d, \\ 0, & d < x \leq c. \end{cases} \quad (13)$$

$$Mu_{BG}(x) = \begin{cases} 1, & x = d, \\ \frac{x-c}{d-c}, & c < x < d, \\ \frac{e-x}{e-d}, & d < x < e, \\ 0, & e < x \leq c. \end{cases} \quad (14)$$

$$Mu_{VB}(x) = \begin{cases} 1, & x \geq e, \\ \frac{x-d}{e-d}, & d < x < e, \\ 0, & x \leq d. \end{cases} \quad (15)$$

4.1.2. Fuzzification of the Trend of Closing Values Before and After a Particular Point p in the Time Series

In the proposed model the difference between the closing price at the observation day p (Figure 3) and closing price of 3rd day after p , as the measure for the market direction is considered. This difference is fuzzified in the following manner,

$$B_i = \begin{cases} Positive, & X_i^{CL} \geq X_{i-3}^{CL}, \\ Negative, & X_i^{CL} < X_{i-3}^{CL}. \end{cases} \quad (16)$$

$$M_i = \begin{cases} X_i^{CL} - X_{i-3}^{CL}, & B_i = positive, \\ X_{i-3}^{CL} - X_i^{CL}, & B_i = Negative. \end{cases} \quad (17)$$

Here, X_{i-3}^{CL} is the closing price on $i - 3^{rd}$ day from day p , where p is the i^{th} day of observation, X_i^{CL} is the closing price on the i^{th} day i.e., at point p . B_i is the representative of market bias, so if the difference between the closing prices of the day

p and three days after, comes to be a positive number then market bias is considered as positive as prices are climbing up or else negative. M_i represents the magnitude of market bias present after point p in the time series. This magnitude is converted to fuzzy linguistic market momentum categories FM_i by using the fuzzy rules R_1 to R_7 . Here, FM_i is the fuzzy value of the momentum recognized by the fuzzy rule and $FUZZY(x)$ is the function that converts the crisp value x in the input argument into equivalent fuzzy linguistic term using the Equations (11) to (15), which is depicted in Equation(18),

$$FUZZY(x) = \begin{cases} TNY, & Mu_{TNY}(x) = \max(Mu_{TNY}(x), Mu_{VS}(x), Mu_{SM}(x), Mu_{BG}(x), Mu_{VB}(x)), \\ VS, & Mu_{VS}(x) = \max(Mu_{TNY}(x), Mu_{VS}(x), Mu_{SM}(x), Mu_{BG}(x), Mu_{VB}(x)), \\ SM, & Mu_{SM}(x) = \max(Mu_{TNY}(x), Mu_{VS}(x), Mu_{SM}(x), Mu_{BG}(x), Mu_{VB}(x)), \\ BG, & Mu_{BG}(x) = \max(Mu_{TNY}(x), Mu_{VS}(x), Mu_{SM}(x), Mu_{BG}(x), Mu_{VB}(x)), \\ VB, & Mu_{VB}(x) = \max(Mu_{TNY}(x), Mu_{VS}(x), Mu_{SM}(x), Mu_{BG}(x), Mu_{VB}(x)). \end{cases} \quad (18)$$

The trend that the market has assumed or the sentiment in the market FM_i is represented using fuzzy linguistic terms namely *Extremely Bearish*, *Very Bearish*, *Bearish Neutral*, *Neutral*, *Bullish Neutral*, *Very Bullish* and *Extremely Bullish*. Here, *Bearish* word represents the situation where market sentiment is in selling mood and prices are going down, *Bullish* word represents the situation where market sentiment is in buying mood and prices are going up and *Neutral* word represents the situation where market sentiment is indecisive and prices are not moving in any particular direction. The adjectives *very* and *extremely* helps to represent the market sentiment to a higher degree of accuracy.

R_1 : IF ($B_i = \text{Positive OR } B_i = \text{Negative}$) and $FUZZY(M_i)$ is TNY THEN FM_i is Neutral,

R_2 : IF $B_i = \text{Positive}$ and $FUZZY(M_i)$ is VS THEN FM_i is Bullish Neutral,

R_3 : IF $B_i = \text{Negative}$ and $FUZZY(M_i)$ is VS THEN FM_i is Bearish Neutral,

R_4 : IF $B_i = \text{Positive}$ and $FUZZY(M_i)$ is BG THEN FM_i is Very Bullish,

R_5 : IF $B_i = \text{Negative}$ and $FUZZY(M_i)$ is BG THEN FM_i is Very Bearish,

R_6 : IF $B_i = \text{Positive}$ and $FUZZY(M_i)$ is VG THEN FM_i is Extremely Bullish,

R_7 : IF $B_i = \text{Negative}$ and $FUZZY(M_i)$ is VG THEN FM_i is Extremely Bearish.

Now the final market direction MD_i is set using the fuzzy rules R_8 to R_{10} .

R_8 : IF FM_i is Bearish Neutral or FM_i is Very Bearish or FM_i is Extremely Bearish THEN MD_i is 'BR',

R_9 : IF FM_i is Bullish Neutral or FM_i is Very Bullish or FM_i IS Extremely Bullish THEN MD_i is 'BL',

R_{10} : IF FM_i is Neutral THEN MD_i is 'NT'.

After evaluating markets' momentum FM_i the fuzzy rules R_8 to R_{10} are used to evaluate the final direction towards which the market is actually moving. The market direction is represented by MD_i which can have any one of the three values 'BL', 'NT' or 'BR', where 'BL' represents a bullish or an up trending market, 'NT' represents a neutral rangebound market and 'BR' represents a bearish or down trending market. Using the above mentioned approach the trend formed three days after point p will be fuzzified and will be represented as either 'BR', 'NT' or 'BL' linguistic terms. The fuzzy rule-base will be populated with the information regarding previous trend,

candlestick attributes of the p^{th} day and trend after p^{th} day. The information regarding the trend after p^{th} day would help us build the fuzzy document model in the IR system. The contents of the documents created with this model will be fuzzy rules.

4.2. Phase 2: Information Retrieval using Modified Tf-idf Scoring Scheme

Now that the fuzzification process has generated fuzzy rules and these fuzzy rules are stored in documents. The modified IR system would find the most appropriate and relevant document depending upon the query. This quality of the IR systems helped to build a model that would suggest the most appropriate future trend. The novel approach followed here is that the trend is represented as a document (containing fuzzy observations of the time series) and the query consists of the fuzzy linguistic terms that represent the current state of the financial time series, this approach is not present in the traditional tf-idf scheme and gives an entirely new dimension of looking at how IR systems are used.

The tf-idf scoring scheme is used to complete the task of forecasting. The documents created in the modified IR system are 'BR', 'NT' and 'BL' and each contains fuzzy observations of the time series and each fuzzy observation is again considered as a document. The trend formed after the point p (Figure 3) will be any one of 'BR', 'NT' or 'BL'.

The constituents of 'BR' document would be all those fuzzy observations who have 'BR' as the trend after the point p in the time series as they would represent instances when market became Bearish after point p . The constituents of 'NT' document would be all those fuzzy observations who have 'NT' as the trend after the point p in the time series as they would represent instances when market became Neutral after point p . The constituents of 'BL' document would be all those fuzzy observations who have 'BL' as the trend after the point p in the time series as they would represent instances when the market became Bullish after point p .

The terms in the query would represent the trend ('BR', 'NT' or 'BL') formed till the point p_1 along with the fuzzy attributes ('US', 'BD', 'LS', 'CC') of the candlestick formed at the point p_1 in the time series. The query has two terms only. The first term would be the trend that was prevailing before point p_1 (Figure 3) and second term would be the attributes of the candlestick formed at point p_1 in the time series. The importance of the first term of the query is taking into consideration the prevailing trend till the point p_1 and second term would describe which type of candlestick formation took place at the point of observation, both these information would be necessary to forecast the future trend that might be forming after the observation point p_1 in the time series.

This treatment of query posed to the IR system is a unique approach, which is not presented in the traditional tf-idf scheme. So if a query is received then using the tf-idf technique we would calculate the scores and the document which gives the highest score would be the forecasted trend.

4.3. Phase3: Forecasting using Modified Tf-idf Scoring Scheme

4.3.1. The Data

For the experiments the CNX NIFTY-50 index daily data of the National Stock Exchange of India is used. Table 1(a) gives a snapshot of the data that was used. The range of data that was used started from 1-Jan-1997 to 25-Mar-2015.

Every row in Table 1(a) represents daily Open, High, Low and Close values of the NIFTY index. The data presented in Table 1(a) displays date in the first column in YYYYMMDD format, here YYYY represents year (ex. 2017), MM represents month i.e., from 01 to 12 for January to December respectively and DD represents day of the month i.e., from 01 to 31. Open value of the day is in second column. High value of the day is in third column. Low value of the day in fourth column and Close value of the day is in fifth column. From the data available through Table 1(a) fuzzy information is extracted from every row of the observations, using the methods presented in the previous sections. A snapshot of the fuzzy information generated by the proposed model is shown in Table 1(b).

Table1(a): Snapshot of CNX NIFTY-50 index daily date.

< Date >	< Open >	< High >	< Low >	< Close >
19970101	905.20	941.4	905.20	939.55
19970102	941.95	944.0	925.05	927.05
...
20150323	8591.55	8608.35	8540.55	8550.90
20150324	8537.05	8627.75	8535.85	8542.95
20150325	8568.90	8573.75	8516.55	8530.80

4.3.2 The Entire Document Corpus

The first column 'PrevTrnd' in Table 1(b) represents the previous trend that was prevailing three days before the observation point p (Figure 3) in the time series, the second column 'Candle' represents the attributes ('US', 'BD', 'LS', 'CC') of the candlestick formation that took place at point p and the third column 'FutTrnd' represents the trend that has formed three days after the observation point p . The fuzzy observations formed from time series data in Table 1(a) is converted to information-base or knowledge-base generated by the proposed model is represented through Table 1(b) and that would become the whole document corpus for the proposed IR system.

Every row in Table 1(b) is again considered as individual documents. Now that the entire document corpus is generated, it is then divided into three categories of documents namely 'BR', 'NT' and 'BL'. The 'BR' document would contain only those entries from the entire document corpus that are having 'FutTrnd' value as either Bearish, Bearish Neutral or Extremely Bearish. So, 'BR' document would contain

Table 1(b): Snapshot of fuzzy information generated by the proposed model.

Prev Trnd	Candle	Fut Trnd
Bullish	VSTNYTNYW	Extremely Bullish
Bullish	VSTNYTNYW	Extremely Bullish
...
Extremely Bearish	VSTNYTNYW	Extremely Bullish
Bearish	VSTNYTNYW	Extremely Bullish

those instances of the entire corpus whose future trend after point p were found to be Bearish in nature. Similarly, 'BL' document would contain only those entries from the entire document corpus that are having 'FutTrnd' as either Bullish, Bullish Neutral or Extremely Bullish. So, 'BL' document would contain those instances of the entire corpus whose future trend after point p were found to be Bullish in nature. And 'NT' document would contain only those entries from the entire document corpus that are having 'FutTrnd' as Neutral. So, 'NT' document would contain those instances of the entire corpus whose future trend after point p were found to be Neutral in nature. From this treatment three categories of documents are generated that represent the sentiment of the market and these would be helpful in forecasting the market sentiment.

4.3.3 The Query and Forecasting

Now that all the documents are in place, the query can be designed that can be given to the proposed system. The terms in the query would constitute the trend ('BR', 'NT' or 'BL') formed till the point p_1 (Figure 3) along with the fuzzy attributes ('US', 'BD', 'LS', 'CC') of the candle formed at the day p_1 in the time series. The query has two terms only. The first term would be the trend that was prevailing before point p_1 and second term would be the attributes of the candlestick formed at point p_1 in the time series.

The importance of the first term of the query is taking into consideration the prevailing trend till the point p_1 and second term would describe which type of candlestick formation took place at the point of observation, both these information would be necessary to forecast the future trend that might be forming after the observation point p_1 in the time series. So if a query is received then using the tf-idf technique we would calculate the scores and the document which gives the highest score would be the forecasted trend. For example, if we pose a query to the system with two terms, term 1 = "BL" and term 2 = "TNYTNYTNYW" then tf-idf scores would be calculated by the proposed system as shown in the Tables 1(c), 1(d) and 1(e) for the documents 'BR', 'NT' and 'BL' respectively.

In the Tables 1(c), 1(d) and 1(e), the columns represent the information about the documents 'BR', 'NT' and 'BL' respectively. The 'TERM' column represents the

The column header ‘TF-SQUARE’ represents the squared value of ‘TF-log’ value, $(\omega_{t,d})^2$, for the purpose of normalization. The ‘NORM-TF’ represents the normalized value of ‘TF-log’, $\omega_{t,d(norm)}$, for the purpose of length normalization of the column vector, by using their squared values in the TF-SQUARE column, using the following Equation (20),

$$\omega_{t,d(norm)} = \frac{\omega_{t,d}}{\sqrt{\sum_{i=1}^{|v|} (\omega_{t,d_i})^2}}. \quad (20)$$

The IDF column represents the inverse document frequency idf_t which is calculated by the following Equation (21),

$$idf_t = \log_{10} \left(\frac{N}{df_t} \right), \quad (21)$$

where, df_t is the document frequency of the term t in the specific document corpus and N is the total number of documents in the entire document corpus.

The column header ‘IDF-SQUARE’ represents the squared value of ‘IDF’ value, $(idf_t)^2$, for the purpose of normalization. The ‘NORM-IDF’ represents the normalized value of ‘IDF’, $idf_{t(norm)}$, for the purpose of length normalization of the column vector, by using their squared values in the ‘IDF-SQUARE’ column using the following Equation (22),

$$idf_{t(norm)} = \frac{idf_t}{\sqrt{\sum_{i=1}^{|v|} (idf_{t_i})^2}}. \quad (22)$$

The column ‘TF-IDF’ represents the product of the normalized weights of $tf_{t,d}$ and idf_t . The final ‘TF-IDFscore’ is the sum of the values in the column ‘TF-IDF’ that represents the cosine similarity between the query vector and document vector. So, the document (‘BR’, ‘NT’ or ‘BL’) having the highest ‘TF-IDFscore’ is the most relevant document that the IR system has given us and is the forecasted trend. In the above example the highest score of 1.00 was achieved by the document ‘NT’ whose details are given in Table 1(d), so the forecasted trend for above example is ‘NT’ i.e., Neutral.

4.3.4 Example

The concept discussed in the above sections is explained by using an example. Let us consider the following Table 1(f) which contains a sample data of seven consecutive day values of CNX Nifty-50 index. We will use the values of Table 1(f) for illustrating the examples. The first column indicates the i^{th} day of observation. The observations are daily observations without any gaps. The second column contains the dates on which the observations were taken and the date is represented in the format YYYYMMDD. The third, fourth, fifth and sixth columns contains the OPEN, HIGH, LOW and CLOSE value of each day respectively. The OPEN, HIGH, LOW

and CLOSE are the j^{th} values represented as OP, HI, LO and CL respectively. The seventh column represents the D_i^{OPCL} values calculated for each day and which is the non-negative difference $|X_i^{OP} - X_i^{CL}|$ of every day observation. Here, X_i^{OP} and X_i^{CL} are the OPEN and CLOSE values on the i^{th} day of observation respectively.

The data represented in Table 1(f) is depicted using the candlestick chart in Figure 5. In Figure 5 the observation point p is depicted which is the 4th day of observation and point p_1 is depicted which is the point beyond which the prediction is to done. In the proposed model every candle stick formation before p_1 is fuzzified and becomes the candidate for the creation of fuzzy rules. The fuzzified information before point p is used to develop the antecedent and after p but before p_1 is used to develop the consequent of each fuzzy rule. During actual practice when the fuzzy rule is created we shift the point p and p_1 to their respective next candlestick and the process of rule addition is done to the fuzzy rule base till all the observations are covered from the available data set.

Table 1(f): Seven consecutive day values of Nifty-50 index.

Day No.(i)	Date (YYYYMMDD)	OPEN (j=OP)	HIGH (j=HI)	LOW (j=LO)	CLOSE (j=CL)	$D_i^{OPCL} = X_i^{OP} - X_i^{CL} $
1	20150213	8741.5	8822.1	8729.65	8805.5	64
2	20150216	8831.4	8870.1	8793.4	8809.35	22.05
3	20150218	8811.55	8894.3	8808.9	8869.1	57.55
4	20150219	8883.05	8913.45	8794.45	8895.3	12.25
5	20150220	8895.5	8899.95	8816.3	8833.6	61.9
6	20150223	8856.85	8869	8736.1	8754.94	101.9
7	20150224	8772.9	8800.5	8726.75	8762.1	10.8

Let us consider X_4^{OP} and X_4^{CL} as two instances of values given in the Table 1(a). Here, the value of $i=4$ for both the instances which represents the 4th day of observation and $j='OP'$ which would represent the OPEN value on 4th day of observation which is equal to 8883.05 and $j='CL'$ value on the 4th day of observation which is equal to 8895.3 (highlighted in Table 1(f)).

From Equation (1), D_i^{jk} is calculated as the non-negative difference of X_4^{OP} and X_4^{CL} . Here, $i=4$, $j='OP'$ and $k='CL'$. Using Equation (1) we get $D_4^{OPCL} = |X_4^{OP} - X_4^{CL}|$. So, $D_4^{OPCL} = |8883.05 - 8895.3|$ or $D_4^{OPCL} = |-12.25|$ or $D_4^{OPCL} = 12.25$.

Now we determine the color of the candlestick formed at the i^{th} observation, for $i=4$, it is found that $X_4^{CL} > X_4^{OP}$, so the value of C4 using the Equation (2) evaluates to 'White', this implies that the candlestick color on the 4th day of observation in Table 1(f) is 'White'.

Using Equation (4) the universe of discourse U is calculated for the sample data in Table 1(f). The value AD^{OPCL} helps us to determine U as follows:

$$AD^{OPCL} = [D_1^{OPCL} + D_2^{OPCL} + D_3^{OPCL} + D_4^{OPCL} + D_5^{OPCL} + D_6^{OPCL} + D_7^{OPCL}]/7.$$

can be determined as the CLOSING value of the 4th day i.e., 8895.3 (highlighted in Table 1(f)). So, $X_4^{CL} > X_1^{CL}$ hence the value of B_4 would be Positive and the value of M_4 would be evaluated from Equation (17) as $M_4 = X_4^{CL} - X_1^{CL}$ or $M_4 = 8895.3 - 8805.5$ or $M_4 = 89.8$. So, for $i=4$ i.e. the data of the candlestick formed in the fourth day of observation from the Table 1(f), we calculated the Upper-Shadow value (US_4) as 18.15, the Lower-Shadow value (LS_4) as 88.6 and Candle-Body value (D_4^{OPCL}) as 12.25. The fuzzy values for US_4 , LS_4 and D_4^{OPCL} are calculated using the Equations (11) to (15) and are presented in the Table 1(g).

Table 1(g): Fuzzy candlestick formation on the 4th day.

	US_4	LS_4	D_4^{OPCL}
x	18.15	88.6	12.25
$Mu_{TNY}(x)$	0	0	0.177
$Mu_{VS}(x)$	0.237	0	0.822
$Mu_{SM}(x)$	0.762	0	0
$Mu_{BG}(x)$	0	0	0
$Mu_{VB}(x)$	0	1	0
max	0.762	1	0.822
$FUZZY(x)$	SM	VB	VS
Candle colour(C_4)	White(W)		
Fuzzy Candlestick formation	SMVBVSW		

From Table 1(g) it is found that on the fourth day the fuzzy value of the candlestick is SMVBVSW which was achieved by concatenating the fuzzy values of US_4 , LS_4 and D_4^{OPCL} and C_4 . It means that the candlestick bar formed on the fourth day has a small upper shadow, a very big lower shadow, very small candle body and candle color is white. The above example shows the fuzzy calculation results for only one candlestick, similarly all the candlestick formations are fuzzified and the fuzzy rule-base is created. The fuzzy rule-base is then analyzed and phase-3 activities (mentioned above) are performed to achieve the forecasting results.

5. Results and Discussion

The results are tabulated in Tables 1(c), 1(d) and 1(e). Table 1(c) shows the tf-idf score calculated for the document classified as 'BR' is 0.87, Table 1(d) shows the tf-idf score calculated for the document classified as 'NT' is 1.00 and Table 1(e) shows the tf-idf score calculated for the document classified as 'BL' is 0.92. The greatest value came for the document classified as 'NT' and which suggests that the coming trend or the forecasted trend would be neutral.

The performance analysis of the proposed model is done by calculating the Root Mean Squared Error (RMSE). The RMSE (also called the root mean square devia-

tion, RMSD) is a measure frequently used to calculate the difference between values predicted by a model and the values actually observed from the environment from where the model is created. The individual differences so calculated are also called residuals, and the RMSE helps to aggregate these residuals into a single measure of predictive power. Lower values of RMSE relative to the number of observations suggest better predictability of the model. The RMSE of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error, where X_{obs} is observed values and X_{model} is modeled values at time/place i :

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (X_{obs_i} - X_{model_i})^2}{n}}. \quad (23)$$

By simply using RMSE, we can only judge the performance of those models whose total number of observations (sample size) used for evaluation are more or less similar, but it cannot be used in a raw form to compare the performance of those models whose sample size vary to a large extent, for that purpose we use RMSE% which is a normalized form of RMSE. To normalize the RMSE value with respect to total number of observations (sample size) we convert RMSE into RMSE%, this helps to compare the performance of various models whose total number of observations differ from each other.

$$RMSE\% = \frac{RMSE}{Total\ number\ of\ observations} \times 100. \quad (24)$$

Table 2: Performance comparison between FDIRM and other models.

Sr.No.	Model used	RMSE%
1	Holt-Winters triple exponential smoothing	59
2	SVM based Regression	59.36
3	Random Forest	47.94
4	Proposed Method (FDIRM)	1.72

A number of experiments were conducted by using the proposed model several times and it was found that around 57% times the proposed model could predict correctly. The RMSE value came around 1.03 and RMSE% came around 1.71% which is quite small compared to number of observations and hence this gives a strong indication that the proposed system seems to works very well in forecasting the future trend of the stock market index. In order to verify the efficiency of the proposed model, a number of experiments were performed with the same set of crisp values with other well-known algorithms through data-mining software WEKA 3.7.12. The contents of Table 2 displays the comparative of the already established benchmark models and the proposed models performance.

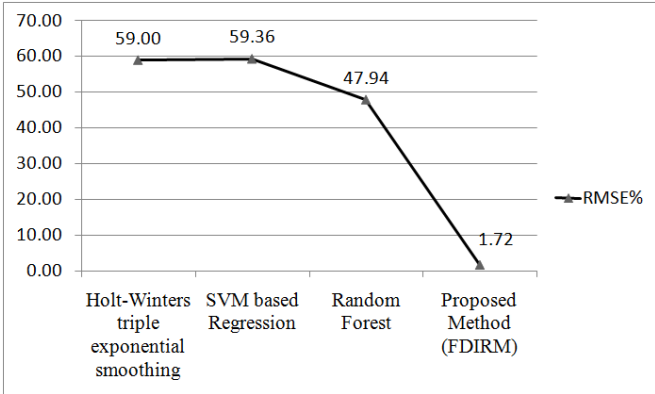


Fig. 6 Performance comparison between FDIRM and other models

The performance of the proposed model is compared with three other benchmark models namely, Holt-Winters with triple exponential smoothing, Support Vector Machine (SVM) based Regression and Random Forest on the basis of RMSE%. The comparative is tabulated in Table 2 and whose graphical depiction is presented in Figure 6. The proposed FDIRM method shows minimum RMSE% value which implies that, compared to others, the proposed models’ performance is quite high, which indicates that it performs more efficiently while forecasting as compared to other benchmark models.

6. Conclusion

In the proposed system the simple time series is converted into an enriched fuzzy linguistic time series with a unique approach of incorporating market sentiment related information along with the price. Another unique approach is followed while modeling the IR system which converts a simple IR system it into a forecasting system.

A number of experiments performed using the proposed model on CNX NIFTY-50 index values shows that the proposed FDIRM method has the minimum RMSE% value as compared to already established benchmark models, which implies that, compared to others, the proposed models performance is quite high and has high potential of becoming a good forecasting model. The same model was also tested on individual stocks of Indian stock market and reasonably good results were achieved. However, improvisation is underway for increasing the forecasting accuracy of the model by experimenting more on the fuzzy elements of the proposed model. A java program is under construction to implement the model in real-time with graphical user interface, so that it can be used by investors and traders as a decision support system.

Further research can be carried out by modifying the proposed model to perform the de-fuzzification process for predicting crisp futuristic values. To increase the accuracy of forecasting, elements of fundamental analysis could also be included in

