ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Multidimensional Time Series Based Review Spam Detection

Siddu P. Algur[1], Jyoti G. Biradar[2], Prashant Bhat[3]

Professor, Department of Computer Science, Rani Channamma University, Belagavi, Karnataka, India[1]

Research Scholar, Department of Computer Science, Rani Channamma University, Belagavi, Karnataka, India[2,3].

ABSTRACT: With the rapid growth of e-Commerce websites there is also an increase in consumers relies on user-generated online reviews when making purchase decisions. As online reviews play a crucial role in today's electronic commerce, these reviews are helpful for consumers to get more information about any store or product and are source of information for the potential customers before deciding to purchase a product. Unfortunately, the ease of posting content to the web, potentially anonymously, combined with the public's trust and growing reliance on opinions and other information found online, create opportunities and incentives for unscrupulous businesses to post deceptive opinion spam that are deliberately written, to sound trustworthy in order to deceive the customers. Conversely, due to the reason of fame, people try to game the system by opinion spamming i.e. writing fake reviews to promote or to demote some target products. Due to the pervasive spam reviews, customers can be misled to buy low-quality products, while decent stores can be defamed by malicious reviews. Hence a novel and effective technique is proposed in this work, which exploits the burstiness nature of the reviews to identify review spamicity based on multidimensional time series for the reviews extracted from review websiteresellerratings.com for the different stores. The experimental result significantly outperforms the state-of-the-art competitors.

KEYWORDS: Opinion mining, Review spam, Time series, Multidimension, Bursts.

I. INTRODUCTION

In today's day it's a growing trend that people rely on online reviews to make purchase decision. The rapid growth of social web with the advent of web2.0 has significantly contributed to the user generated data including opinions, reviews, comments, events and services. These opinions are helpful for both individuals and business organizations [1]. Prejudiced social media such as product reviews are now widely used by individuals and organizations for their decision making. Large amounts of online reviews, the valuable voice of the customer, benefit consumers and product designer. Online reviews and ratings about products and stores are essential parts in today's electronic commerce where they provide helpful information for potential customers. A product or store with a decent rating and a high proportion of positive reviews will attract more customers and larger amount of business, while a couple of negative reviews/ratings could substantially harm the reputation, leading to financial losses. Since there is no rule governing online reviews and ratings, some product providers or retailers are leveraging such public media to defame competitors and promote themselves unfairly, or even to cover the truth disclosed by genuine reviews [12]. With the rise of online business, consumers now-a-days are not only able to do their shopping online, but also they can leave reviews on their purchased products for other potential users. In the normal situation, reviews for a product arrive randomly. Yet, there are also areas (time periods) where the reviews for a product are bursts, in which there are sudden emphasis of reviews in these areas. Such areas are termed as review bursts. A review burst can be due to sudden increase of popularity of a product or because the product is under spam attack [3]. For example a product may get suddenly popular due to the advertisement given by a popular actor / actress or due to huge discount during a festival offer. Then a large number of customers may purchase the product and write reviews for that product in a short span of time. In this kind of bursts, most reviewers are likely to be non-spammers. In contrast, a number of fake reviews may be posted. Such possibilities lead to an important hypothesis about review bursts. A fundamental approach in text data mining is to extract meaningful structure from document streams that arrive continuously over time. Much of the world's supply of data is in the form of time series.

In the last decade, there has been an explosion of interest in mining time series data. A time series is a collection of observations made chronologically. The increasing use of time series data has initiated a great deal of research and

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

development attempts in the field of data mining. Time series is an important class of temporal data objects and it can be easily obtained from scientific and financial applications. Moreover time series data, which is characterized by its numerical and continuous nature, is always considered as a whole instead of individual numerical field.In this work, reviews from review website are extracted considering posting dates of the reviewers as a prime factor to find spamicity of reviews. As spam reviews undoubtedly reduce the quality of reviews, they may even mislead users to make wrong purchase decisions. There is also potential for vendors trying to influence their consumer's decisions by injecting deceptive product reviews online. Many efforts have been done recently to develop algorithms to detect such deceptive opinion spam. Review spam (or opinion spam) tries to mislead readers by composing untruthful views. Therefore, there is a great demand to detect spam reviews thoroughly on the web. Hence, a different approach is proposed in this work using multidimensional time series using four dimensions. Burstiness nature of reviews is identified from the different stores. Experiments are carried out to find review spamicity based on multidimensional time series construction from a review website resellarratings.com for the stores namely Auto_parts_warehouse.com, Dhgate.com and Neweggs.com stores and its results demonstrate the effectiveness of the proposed work. The present paper depicts about the trends of detection of review spamicity with respect to multidimensional time series construction. Section II,introduces about the related work. Section III,gives an overview of the proposed technique used to find review spamicity. Section IV,describes the working and experimental results for detecting review spam. And section V presents conclusion and future work.

II. RELATED WORK

Opinion mining and detection of deceptive reviews has become a vital research area in the field of product/store reviews. In [4], researchers focus on detecting disruptive review spam such as reviews with irrelevant texts by utilizing information such as statistical features from review texts, behaviours of review spammers, and relationships among reviewers. In [8], some behavioural patterns were designed to rank reviews and to find review spam. In [5] ,attempts are made for identifying manipulated reviews, in this work, they have highlighted that there are two types of review spam, one is manipulated review which will mislead the customer and another is non-review i.e. it is not giving any actual opinion about the product, it can be advertisement of product. In [7], attempts are made by using semi supervised manifold ranking algorithm to find opinion spam written to sound authentic and deliberately mislead readers. It will identify manipulated offerings on review portal. In [15], they focused on finding fake store reviewers using a graph-based method. In [6], different reviewing patterns were discovered by mining unexpected class association rules. In [9], the results were quite effective in detecting spam and non-spam reviews with text classification using *n*-gram features. Amazon Mechanical Turk was employed to crowd source fake hotel reviews by paying US$1 per review and anonymous online workers called Turkers to write fake reviews for some hotels. In [11], review spam detection is concerned with a problem of singleton review using time series pattern discovery by finding the correlation between rating and volume of singleton reviews. Our method aims at extracting the reviews from review website from three stores and to find review spamicity based on constructing multidimensional time series.

III. PROPOSED METHODOLOGY

In the proposed work, a novel and effective technique is used to detect review spamicity based on constructing multidimensional time series from the extracted reviews. The identified dimensions include positive word length score, negative word length score, review rating and no of reviews. The idea is based on extracting reviews from review website named resellarratings.com in customer reviews from different stores.
The various steps of the proposed method include:

- Review extractor
- Identifying Multidimensions
- Time series construction
- Review spamicity measure based on burstiness in reviews

3.1Review extractor

Reviews are extracted from review website www.resellarratings.com for the stores Auto_parts_warehouse.com, Dhgate.com and Neweggs.com using review exactor tool (import.io). From the extracted reviews, stop words are

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

filtered to improve efficiency and to reduce indexing file size of the reviews and are stored in raw review database for all the three stores separately.

3.2 Identifying Multidimensions

Multiple dimensions are used to support detection of spamicity of reviews like positive word length score, negative word length score, review word length score, average rating, review similarity spam score, rating similarity score, rating deviation score, average rating, total number of reviews, ratio of singleton reviews etc.,[2]. Among these dimensions, four dimensions are identified and are used in the proposed work; they are positive word length score, negative word length score, review rating and total number of reviews. These four dimensions are considered as with positive and negative word length score, the reviewers can express their opinion about the product/store with opinion bearing words in a review like ”beautiful”, “good” ,“poor” “bad” etc., which can be considered for spam detection. Review rating is identified, for instance rating is regarded as representation of reviewers sentiment orientation. Compared with rating score, the content of the reviews will represent more accurate sentiment of the reviewer. As genuine reviewer is expected to give ratings that are similar to other rating of user on the same product. The ratings of the spammer are different from these reasonable ratings. Spammer tends to give high rating in low quality product to promote that product/store and also give low rating in high quality product to damage that product/store reputation. The count of total number of reviews for a particular day/ week/ month/year always play a vital role in detection of spam reviews, hence number of reviews are identified in the proposed work. The four dimensions are identified based on the behaviour of the reviewers. The specifics of four dimensions are given below:

- Positive word length score: Let ‘TW’ be total words and ‘PosW’ be number of positive words in a review. The positive word length score can be obtained by dividing number of positive words in a review by total number of words in a review. The equation for positive word length score is given by Pos_score P = PosW/TW.
- Negative word length score: Let ‘TW’ be total words and ‘NegW’ be the number of negative words in a review. The negative word length score can be obtained by dividing number of negative words in a review by total number of words in a review The equation for negative word length score is given by Neg_score N = NegW/TW.
- Review rating: Rating is a grade or rank in the range 1 to 10 or 1 to 5, it’s an opinion given by the reviewers for a particular product/store. In the proposed work, the rating scale of the reviews given by the reviewers from the stores is One to Five (i.e. 1 to 5).Let ‘Rr’ be rating given by the reviewers.
- Number of reviews: Number of reviews given by the reviewers varies day to day; hence to have a count of total number of reviews per day/week/month/year is essential. Let ‘R’ be number of reviews for a store ‘S’.

Few examples, reviews are taken from a store Neweggs.com, 1. “Good customer support through chat”, there are total five words, out of which one is a stop word (through) and remaining four words are (good customer support chat). Similarly for the reviews 2. “I had a problem involving my bank stopping a payment due to suspicion of fraud” there are total fifteen words, eight are stop words (i , had , a , my , a , due to , of) remaining seven words are (problem involving bank stopping payment suspicion fraud) 3.“Overall service was good but communication delay”, there are total seven words two are stop words (was, but) and remaining five words are (overall service good communication delay) and 4. Generally decent prices - but buyer beware, many of the lower priced products/specials are items with a high percentage of poor reviews/performance. Service and return policies usually good but shipping can be a problem. If you need it fast order it elsewhere. Not unusual to take a week or more” it consists of fifty words, twenty three are stop words (but , many , of , the , are , with , a , of , and , but , can , be , a , if , you , it , elsewhere, not , to ,it, a , or , more) remaining twenty seven words are (Generally decent prices - buyer beware lower priced products/specials items high percentage poor reviews/performance service return policies usually good shipping problem need fast order unusual take week).Word Count tool is used to count number of words in a review. Stop words are filtered from the review dataset. Sentiment analyzer tool with Natural Language tool kit (NLTK) is used to find sentiment polarity i.e. positive and negative word length score.Table1.shows sample of calculated results of positive and negative word length score of the reviews from Neweggs.com store. Similarly, positive and negative word length score is calculated for remaining all the reviews of the stores Auto_parts_warehouse.com and Dhgate.com.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Table 1.Sample of positive and negative word length score of neweggs.com reviews

R. No	Review	Number of positive words	Number of negative words	+ve word score	-ve word score	Total words
1	Good customer support chat	2	0	0.50	0	4
2	problem involving bank stopping payment suspicion fraud	0	3	0	0.42	7
3	Overall service good communication delay	1	1	0.2	0.2	5
4	Generally decent prices - buyer beware lower priced products/specials items high percentage poor reviews/performance service return policies usually good shipping problem need fast order unusual take week	3	5	11.11	18.51	27

3.3 Time series construction

A time series is a sequence of numerical data points in successive order, usually occurring in uniform intervals. It's a sequence of numbers collected at regular intervals over a period of time. Time series analysis can be useful to check a given asset, security, economic variable, dimensions etc., changes over time or the way it can be compared to other variables over the same time period. In the proposed work, review spamicity detection approach is based on multidimensional time series construction. Let

$S = \{S_1, S_2, S_3 \ldots\ldots\ldots\ldots S_n\}$ be the collection of N stores.

$|S_1|$ be the number of reviews of store S_1,

$|S_2|$ be the number of reviews of store S_2 and so on. Then,

$S_1 = \{ r_{11}, r_{12}, r_{13}, \ldots\ldots r_{1i} \ldots\ldots r_{1n}\}$ be the number of reviews from first store S_1

$S_2 = \{ r_{21}, r_{22}, r_{23}, \ldots\ldots r_{2i} \ldots\ldots r_{2n}\}$ be the number of reviews from second store S_2 and

$S_3 = \{r_{31}, r_{32}, r_{33}, \ldots. r_{3i} \ldots\ldots r_{3n}\}$ be the number of reviews from third store S_3 and so on.

And 'ts_1'be the time series corresponding to the reviews of store S_1, 'ts_2'be the time series corresponding to the reviews of store S_2 and 'ts_3'be the time series corresponding to the reviews of store S_3 and so on. Then, $T(S) = \{ts_1 \ldots\ldots ts_i \ldots\ldots ts_j\}$ where, $ts_i <= ts_j$ for all $1 <= i < j <= n_s$. After choosing the time windows size (denoted by Δt), the time interval under investigation (denoted by $I = [t_0, t_0 + T]$) can be divided into $M = T/\Delta t$ consecutive time windows or sub-intervals. Each time window of length't'contains reviews posted during that time window. Let I_n denote the n^{th} time window, hence $I_n = [t_0 + (n-1)\Delta t, t_0 + n\Delta t]$.

Given a time window I_n, the pos_score'P' and Avg pos_score f_1, the neg_score'N' and Avg neg_score f_2, the review rating 'R_r'and the Avg_rating f_3 and the no of reviews 'R' and Avg no of reviews f_4 is calculated [12], where

P = no of positive words score in a review / total words in the review

N = no of negative words in a review / total words in the review, and

R_r = review rating

R = total no of reviews in a day. Then,

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

$f_1(I_n) = \sum_{ts_j 2I_n} P(I_n)/R$

$f_2(I_n) = \sum_{ts_j 2I_n} N(I_n)/R$

$f_3(I_n) = \sum$ ratings (I_n)/no of ratings in a day

$f_4(I_n) = |R|$

Where $|R|$ denotes the cardinality of the set R. Given a store S_n, time interval $I = [t_0, t_0 + T]$ and time window size Δt, these aggregate functions represent a four dimensional time series and can be collectively represented by

$$Fs(I,\Delta t) = \begin{cases} f_1(1) \ldots\ldots f_1(N) \\ f_2(1) \ldots\ldots f_2(N) \\ f_3(1) \ldots\ldots f_3(N) \\ f_4(1) \ldots\ldots f_4(N) \end{cases}$$

where, $f_i(n)$ is a shorthand for $f_i(I_n)$, i= 1, 2, 3 and so on. In the resulting section, we drop the index on stores and let $F(I, \Delta t)$ denote the time series constructed for a certain store. The way we construct these time series can be generalized to handle spammers who write just a few reviews with similar ratings [12].

3.3.1Window size measure for Review spamicity detection

Given the review records of a store, multidimensional time series can be constructed with different time window sizes (resolutions). The window size may be set for 1day, 5days, 10days, 15days, 30days etc. If the window size is set too small, it may cause high false positive rate as the general trend of a time series would be buried in a large number of fluctuations [11, 12]. Initially the window size(resolution) was set for 7days in the proposed work, as we could not reach the desired resolution, such that the time when the spam attack can be easily pinpointed, experiments were carried with window size set to 15 days and 30 days time period to smooth out short-term punctuations using a larger window size (lower resolution).As with larger window size one would not be able to reveal more details i.e. exact time of the burst, in the proposed work the window size is set to 10 days.

3.3.2 Burst Detection

Bursts of reviews can be either due to sudden popularity of products or spam attacks. Regions of burst in a review of store is measured based on the average of the dimensions identified. Where three/four dimensions merge, that region (time period) is pin pointed as bursty pattern. And those reviews are suspected as spam reviews. Initially to identify regions of bursts, sliding window size was set with 15 days and 30 days. As larger window size (lower resolution) will not be able to reveal more details i.e. the exact time of the burst, the sliding window size 'W' of 10 days is used which is found to cover sufficiently well the bursts ranges in the store reviews. As smaller window size (higher resolution) can be used to reveal more details (the exact time of the burst).In the proposed work, window size of 10 days is used.
The procedure to detect bursts in reviews using multidimensional time series and identification of correlated abnormal patterns are used in ALGORITHM 1 and ALGORITHM 2 respectively.

ALGORITHM 1 Burst detection
//Burst detection in multidimensional time series
//Input : Multidimensional fitted curves $C= \{C_1, C_2, C_3, C_4\}$
//Output : Bursty template patterns, Correlated set of period T
For each dimension C_i do
// template V

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

// Find bursty pattern with respect to time point
a = Length of C_i
b = Length of V
for i = a-b+1 do
Normalize each curve
Find bursty points
End for
End for
Set time window size = 10
N ⟵ length (C)
for X= 1 to n-W+1 do T = T ⟶ U { [b,b+W-1]}
If |{ X ε L_i: i = 1,2,3,4 X ε [b_1,b+W-1]}|==4
End for
Return correlated period T

ALGORITHM 1., describes bursty pattern detection in a time series. For each block b on a fitted curve C, we first normalize it. Then, the number of matches between the template and the normalized block b is identified. By this way, the locations corresponding to bursty patterns in the time series can be found out. Taking the degree of burst into account, the number of matches in each block is considered by the range of values in that block.

ALGORITHM 2 Correlated abnormal patterns
// Correlated abnormal patterns in time series with fixed window size
//Input : Multidimensional curves C_1, C_2, C_3 and C_4
//Output: Correlated abnormal patterns with respect to time periods.
For each dimension C_i do
Time points of burst L_i = Abnormal pattern detection
End for
N= length (C_i)
W = 10
For a =1 to n-10+1 do
T = T U {[b, b+10-1]} if |{Xε L_i: i = 1,2,3,4 X ε [b, b+10-1]}| ==1
End for
Return set of periods (T)

ALGORITHM 2., describes correlated abnormal pattern in time series with window size 'W' set to 10 days. The algorithm takes multidimensional curves as input and returns correlated abnormal patterns with respect to time periods. The time window 'W' will be slide over the multidimensional curves and correlated abnormal patterns (if any) with respect to time period will be identified and returned as an algorithmic output.

3.4 Review spam detection based on burstiness of reviews

In the proposed work, multiple dimensions are identified to support detection of review spamicity using bursty pattern detection method. The identified dimensions include average of positive word length score, negative word length score, review rating and total no of reviews. These dimensions are constructed from the reviews given by the reviewers for each store and the method to detect bursts of reviews is given in ALGORITHM1.This algorithm describes bursty pattern detection in a multidimensional time series fitted curve. As in the proposed work, four dimensions are taken the curves given are C_1...C_4. These dimensions are normalized in the range 0-1.The length of the time window in the time series construction is chosen to be 10 days. The locations corresponding to bursty patterns are considered by the number of matches found in the dimensions for duration of 10 days from the stores. Further, correlated abnormal patterns with respect to time series is given in ALGORITHM 2. This algorithm describes time points corresponding to the bursts in each dimension. A time window is reported if three/four dimensions have bursty pattern falling into the window. Considering the degree of bursts into account, the total number of bursts found from each store for 620 days is considered. The curves of multidimensional time series for the stores and the bursty patterns detected are shown in the Figures in experimental results.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

- *Procedure to identify spike (spam) reviews :*

Step1: Based on the four dimensions used in the proposed approach, bursty patterns of reviews are identified where three or four dimensions merge at a point of time as shown in Figure 4, Figure 5 and Figure 6., in the experimental results for the three stores Auto_parts_warehouse.com, Dhagte.com and Neweggs.com respectively.

Step2: Threshold values are set on each dimension based on the total number of burst (spike) reviews found and total number of reviews of the stores for duration of 620 days

Step3: Reviews found above the threshold value are suspected as spam reviews.

Step4: Among four dimensions at least three consecutive reviews above threshold value are considered.

Step5: Review spamicity is calculated based on the total number of reviews found above threshold value and total number of reviews of the stores for duration of 620 days.

IV. EXPERIMENTAL RESULTS

Experimental results are presented to demonstrate the effectiveness of the proposed method. The proposed technique is applied to see how effective it is in assessing the spamicity of the reviews. Experiments are carried from extracting reviews from review website resellerrating.com for the three stores Auto_parts_warehouse.com, Dhgate.com and Neweggs.com.The review website contains 49,49,284 reviews for 1,96,640 stores as on 15th September 2015. There are 27,522 reviews from Auto_parts_warehouse.com, 12,513 reviews from Dhgate.com and 3,281 reviews from Neweggs.com. A total of 43,316 reviews are taken from all the three stores.The main frame of the data consists of reviews, along with information about stores and reviewers. For each review following information is considered: reviewer's name, its rating (ranging from 1to 5), the posting date and content of the review. The detection of review spamicity is constructed on multidimensional time series analysis from the extracted reviews of stores based on the average of positive word length score, negative word length score, review ratings and number of reviews. Bursty pattern of the reviews is identified from the stores based on multidimensional burst detection method given in ALGORITHM 1. And time points corresponding to each dimension i.e correlated abnormal patterns with respect to time series is given in ALGORITHM 2. The length of the time window in time series constructed is chosen to be of 10 days. In the proposed work, where three or four dimensions merge at some point of time are considered instead of all the four dimensions merge at some point of time, as with four dimensions one could not be able to reveal the exact time of the burst, hence, in the proposed work three/four dimensions merge at a point of time is considered. The curves and bursty pattern of multidimensional time series is shown in Figure 1, Figure 2 and Figure 3 for the three stores Auto_parts_warehouse.com, Dhgate.com and Neweggs.com for duration of 620 days from 15th September 2015 to 1st January 2014 arranged in descending order. In the Figures below, values (Avg positive score ,Avg negative score, Avg rating and Avg number of reviews) of each dimension of the time series is plotted in dark points (up-per box –Average number of review, middle boxes–Average rating and average number of negative score, lower box –average number of positive score) are shown. The solid lines are the fitted curves. Sample of red vertical dash lines are used to highlight the bursty pattern detected in time series from the dimensions by the proposed approach.The remaining bursty patterns detected are shown in Figure 4, Figure 5 and Figure 6 for the stores Auto_parts_warehouse.com, Dhgate.com and Neweggs.com respectively.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

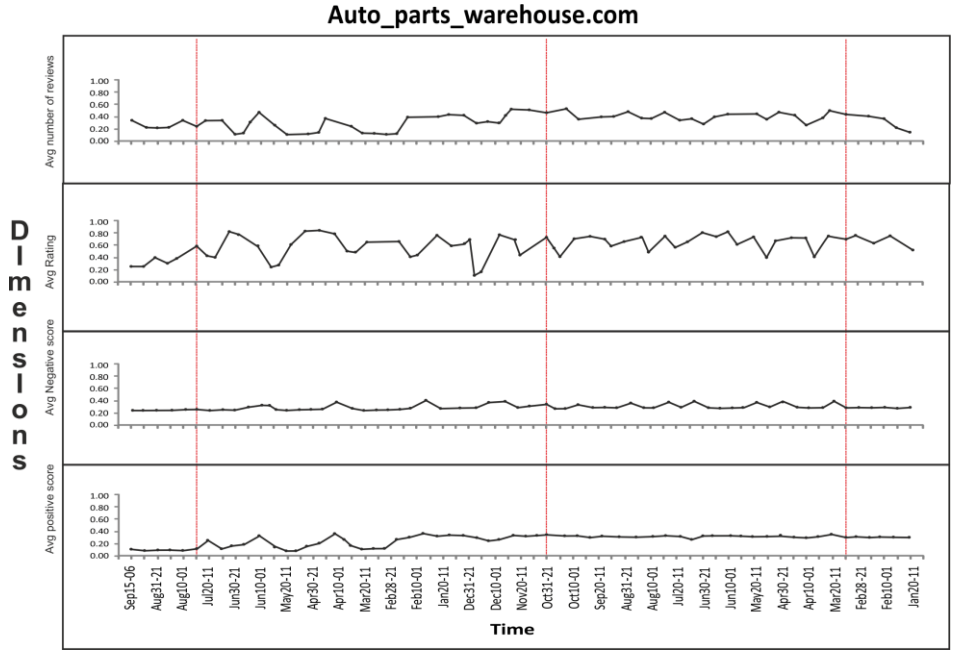

Figure 1. Curves and bursty patterns of multidimensional time series of Auto_parts_warehouse.com store

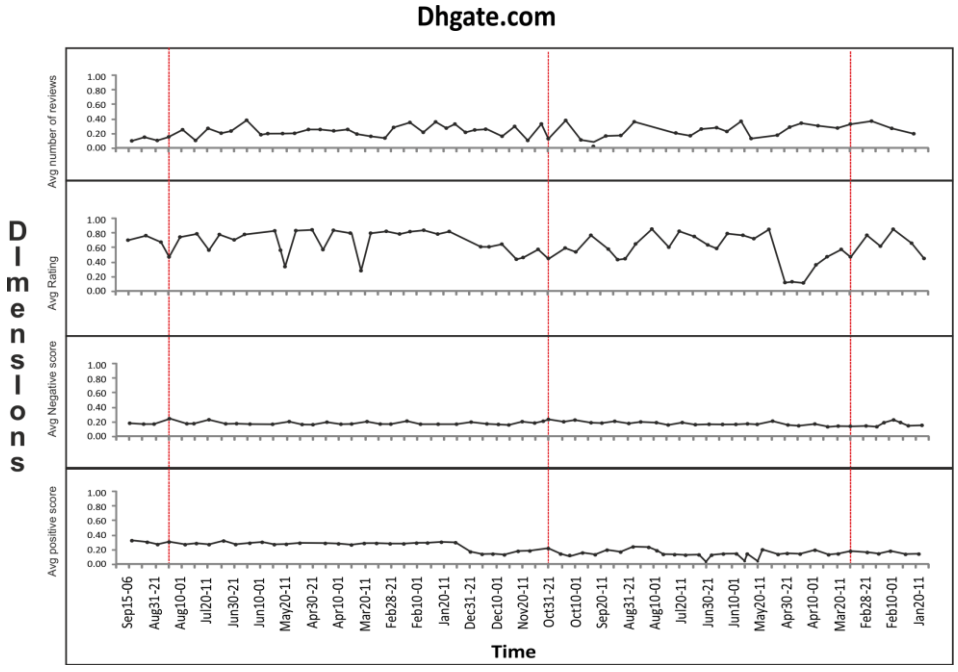

Figure 2. Curves and bursty patterns of multidimensional time series of Dhgate.com store

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

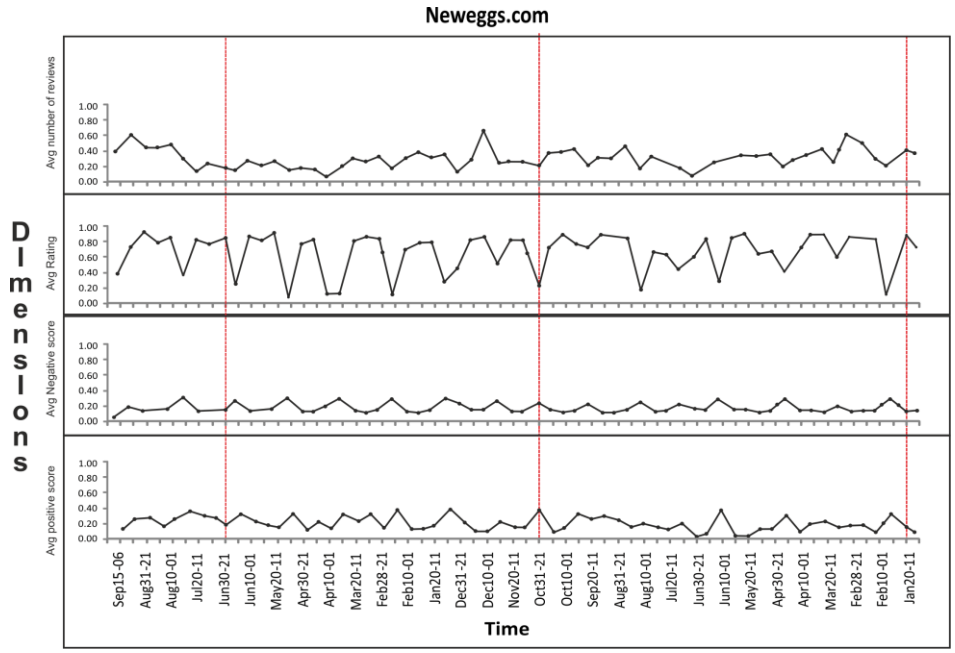

Figure 3. Curves and bursty pattern of multidimensional time series of Neweggs.com store

In the Figure 4, Figure 5 and Figure 6 all the bursty patterns detected of the three stores namely Auto_parts_warehouse.com, Dhgate.com and Neweggs.com stores are depicted. In Table 2, Table 3 and Table 4, reviews were bursty patterns are detected are given and in Table 2a, Table 3a and Table 4a reviews found above the threshold valuesare given for the duration of 10 days (window size 'W' is 10 days) for the three stores.

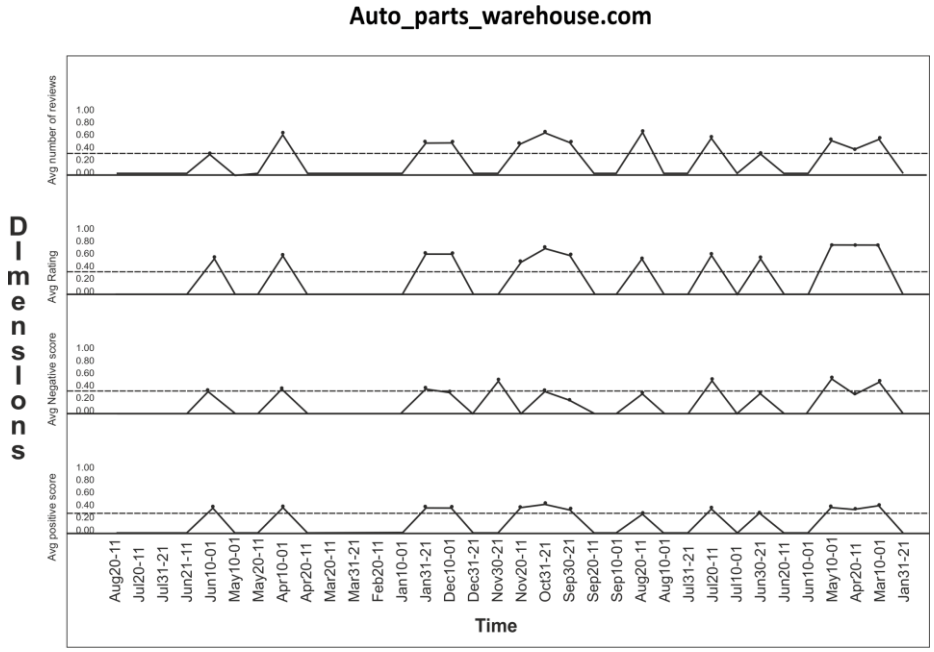

Figure 4. Bursty patterns detected in the store Auto_parts_warehouse.com

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

In Table 2, only the reviews where bursts (spikes) are deteted for duration of 10 days from the store Auto_parts_warehouse.com are specified.

Table 2. Outcome of bursty patterns (spikes) detected in Auto_parts_warehouse.com store

Date (10 days)	Jun10-01	Apr10-01	Jan31-21	Dec10-01	Nov30-21	Oct31-21	Sep30-21	Aug20-11	Jul20-11	Jun30-21	May10-01	Apr20-11	Mar10-01
No of Reviews	134	227	535	257	48	702	723	916	993	745	787	711	1312

Total number of reviews found, where bursty (spikes) pattern detected from the store Auto_parts_warehouse.com are 8090. Total number of reviews of this store for duration of 620 days are 27,522 reviews. Since, all the bursty pattern detected reviews are not to be considered as spam reviews, a threshold value is set for this store considering the number of spike reviews found (8090) and total number of reviews (27522) for duration of 620 days as 0.29. In Figure 4, a dotted line is drawn to each curve of the dimensions on the axis value 0.29. The reviews found above the threshold value are considered as spam reviews. In Table 2a., reviews found above the threshold value are specified and are considered as spam reviews.

Table 2a. Reviews found above the threshold value for Auto_parts_warehouse.com store

Date (10 days)	Apr10-01	Jan31-21	Dec10-01	Oct31-21	Sep30-21	Jul20-11	May10-01	Mar10-01
No of Reviews	227	535	257	702	723	993	787	1312

Total number of reviews found, above the threshold value from the store Auto_parts_warehouse.com are 5536. Total number of reviews of this store for duration of 620 days are 27,522 reviews. Hence, the review spamicity measure of Auto_parts_warehouse.com is 20%.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

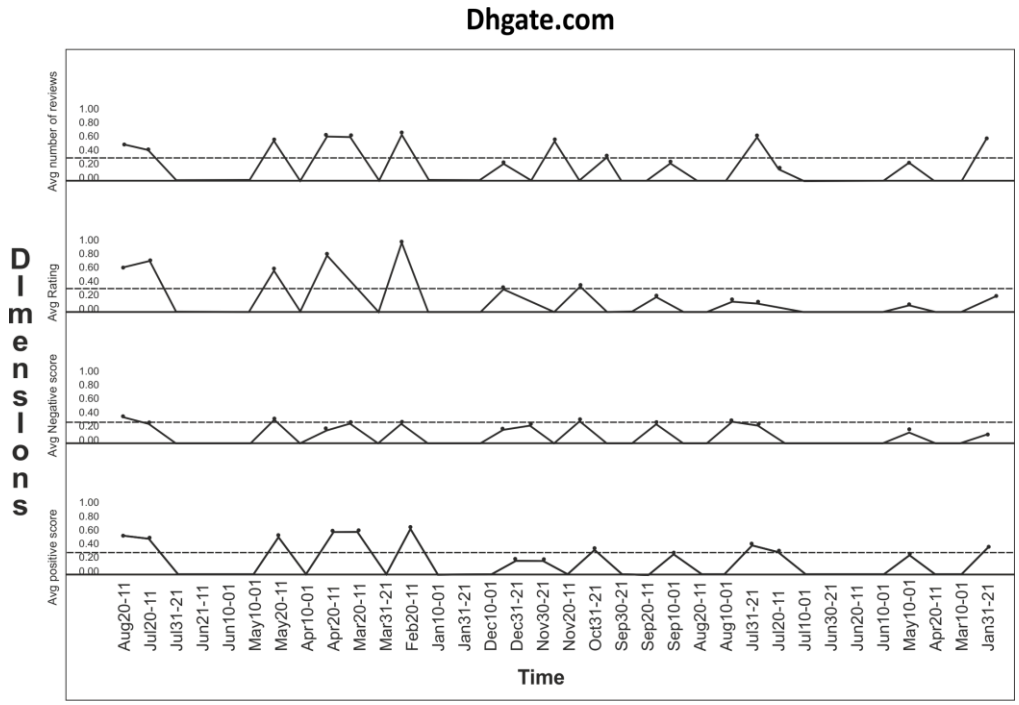

Figure 5. Bursty patterns detected in the store Dhgate.com

In Table 3, only the reviews where bursts (spikes) are deteted for duration of 10 days (window size 'W' is 10 days) from the store Dhgate.com are specified.

Table 3. Outcome of bursty patterns (spikes) detected in Dhgate.com store

Date (10 days)	Aug20-11	Jul20-11	May20-11	Apr20-11	Mar20-11	Feb20-11	Dec31-21	Nov20-11	Oct31-21	Sep10-01	Jul31-21	Jun20-11	May10-01	Jan31-21
No of Reviews	430	272	292	443	499	385	16	15	18	8	690	14	30	12

Total number of reviews found, where bursty (spikes) pattern detected from the store Dhgate.com are 3124. Total number of reviews of this store for duration of 620 days are 12,513 reviews. Since, all the bursty pattern detected reviews are not to be considered as spam reviews, a threshold value is fixed for this store considering the number of spike reviews found (3124) and total number of reviews for duration of 620 days as 0.24. In Figure 5, a dotted line is drawn to each curve of the dimensions on the axis value 0.24.The reviews found above the threshold value are considered as spam reviews. In Table 3a., reviews found above the threshold value are specified and are considered as spam reviews.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Table 3a. Reviews found above the threshold value for the store Dhgate.com store

Date (10 days)	Aug20-11	Jul20-11	May20-11	Apr20-11	Mar20-11	Feb20-11
No of Reviews	430	272	292	443	499	385

Total number of reviews found, above the threshold value from the store Dhgate.com are 2321. Total number of reviews of this store for duration of 620 days are 12,513 reviews. Hence,the review spamicity measure of Dhgate.com is 18%.

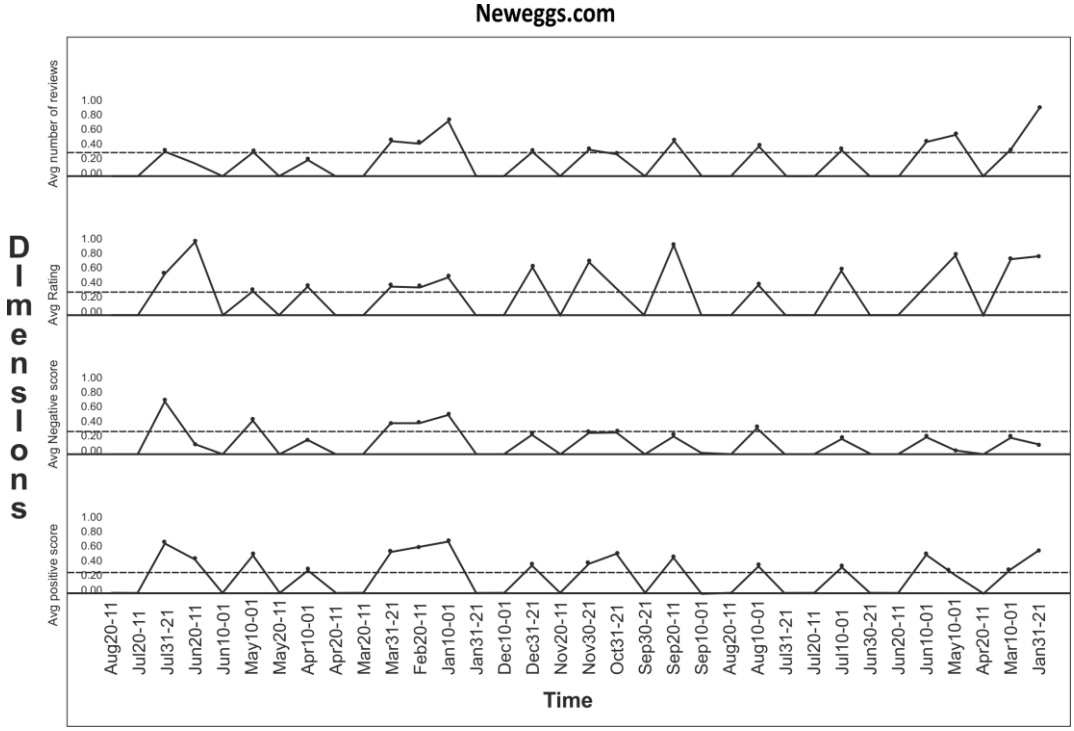

Figure 6. Bursty patterns detected in Neweggs.com store

In Table 4. Only the reviews where bursts (spikes) are deteted for duration of 10 days (window size 'W' is 10 days) from the store Neweggs.com are specified.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Table 4. Outcome of bursty patterns (spikes) detected in Neweggs.com store

Date (10 days)	Jul31-21	Jun21-11	May10-01	Apr10-01	Mar31-21	Feb20-11	Jan10-01	Dec31-21	Nov30-21	Oct31-21	Sep20-11	Aug10-01	Jul10-01	Jun10-01	May10-01	Mar10-01	Jan31-21
No of Reviews	70	44	59	41	45	63	54	46	73	67	44	28	41	48	33	61	61

Total number of reviews found, where bursty (spikes) pattern detected from the store Neweggs.com are 878. Total number of reviews of this store for duration of 620 days are 3,281reviews. Since, all the bursty pattern detected reviews are not to be considered as spam reviews,a threshold value is fixed for this store considering the number of spike reviews found (878) and total number of reviews for duration of 620 days(3281) as 0.26. In Figure 6, a dotted line is drawn to each curve of the dimensions on the axis value 0.26.The reviews found above the threshold value are considered as spam reviews. In Table 4a.,reviews found above the threshold value are specified and are considered as spam reviews.

Table 4a. Review found above the threshold value for Neweggs.com store

Date (10 days)	Jul31-21	Mar31-21	Feb20-11	Jan10-01	Sep20-11	Jun10-01	May10-01	Mar10-01	Jan31-21
No of Reviews	70	45	63	54	44	48	33	61	61

Total number of reviews found, above the threshold value from the store Neweggs.com are 479. Total number of reviews of this store for duration of 620 days are 3281 reviews. Hence, the review spamicity measure of Neweggs.com is 14%.

Therefore, the spam detection rates found are 20%, 18% and 14% for the stores Auto_parts_warehouse.com, Dhgate.com and Neweggs.com respectively. By observation, most of the burst discovered is indeed during the first and last week of the months. As reviews which are written early tend to get more reviewer attention and thus can have bigger impact on the sale of a product. The point of the review date is also an important factor because early reviews are more concentrated by users. For instance , some spammers tends to write reviews as soon as the product/store release because first review is very important for new product and first reviews are more concentrated by users. Hence, in the proposed work, these behaviours of reviewers are given vital importance to get accurate spamicity score. One can observe from the experimental results that there are large numbers of reviews belonging to non-spam category also. Hence, these reviews do not influence the buying decision of the customers and could be considered trustworthy as they provide genuine opinion on some or the other sentiment of the store and are often unbiased [10].

V. CONCLUSION AND FUTURE WORK

In this work, a novel evaluation method, multidimensional time series is used to find review spamicity by using multiple stores. Four dimensions are identified namely, positive word length score, negative word length, review rating and number of reviews.Based on these dimensions, multidimensional time series is constructed. The length of the time window is of 10 days. Burst detection in multidimensional time series and Correlated abnormal patterns are given in ALGORITHM 1 and ALGORITHM 2 respectively. Experimental results of detecting review spamicity by using review website resellerratings.com for the stores Auto_parts_warehouse.com, Dhgate.com and Neweggs.com for the duration of 620 days from 15th September 2015 to 1st January 2014, demonstrate that the proposed method is effective in detecting review spamicity based on multidimensional time series. Review Spamicity detection using Outlier detection technique with multidimensional time series gives the scope for future work.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

REFERENCES

1. Bo,Pang.,and Lillian Lee., "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval Vol 2: No.1-2, pp.1–135, 2008.
2. Chan Myae Aye., and Kyaw May Oo., "Review spammer Detection by using Behaviours based scoring methods" International Conference on Advances in Engineering and Technology (ICAET) March 29-30, Singapore, 2014.
3. Geli Fei., Arjun Mukherjee., Bing Liu., Meichun Hsu., Malu Castellanos., and Riddhiman Ghosh "Exploiting Burstiness in Reviews for Review Spammer Detection" Proceedings of the Seventh International Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, AAAI, 2013.
4. Jindal Nitin., and Bing Liu..,. "Opinion spam and analysis".In Proceedings of the International Conference on Web search and web data mining, WSDM, New York, ACM ,pp.219–230., 2008.
5. Jindal Nitin., and Bing Liu., "Review Spam Detection", In proceedings of 16th International conference on world wide web,ACM, pp.1189- 1190, 2007.
6. Jindal Nitin., Bing Liu., and Lim E P., "Finding unusual review patterns using unexpected rules". In Proceedings of the 19th Conference on Information and Knowledge Management, ACM ,pp.1549–1552., 2010.
7. Jiwei Li., MyleOtt., and Claire Cardie, "Identifying Manipulated Offerings on Review Portals". In Proceedings of International Conference on Empirical Methods in Natural Language Processing,Seattle, Washington, USA, pp.1933–1942, 2013.
8. Lim E P., Nguyen V A.,, Jindal N., Liu B ., and Lauw H. W., "Detecting product review spammers using rating behaviors". In Proceedings of 19th International Conference on Information and Knowledge Management, CIKM,New York, ACM, pp.939–948. 2010.
9. Ott M., Choi Y., Cardie C., and Hancock J.T., "Finding Deceptive Opinion Spam by Any Stretch of the Imagination". In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics.Portland, Oregon, USA, pp.309–319., 2011.
10. Siddu P. Algur., Amit P.Patil., P.S Hiremath., and S. Shivashankar "Conceptual level Similarity Measure based Review Spam Detection" In IEEE International Conference on Signal and Image Processing . ISBN 978-1-4244-8594-9/10, 2010
11. Sihong Xie., Guan Wang, Shuyang Lin, Philip S. Yu "Review spam detection via time series pattern discovery", Proceedings of the 21st international conference companion on World Wide Web,ACM, pp.635-636, 2012.
12. SihongXie., GuanWang., Shuyang Lin., Philip SYu., "Review Spam Detection via Temporal Pattern Discovery "In Proceedings of the 18th ACM SIGKDD International conference on Knowledge discovery and data mining, KDD 'New York, NY, USA., ACM, pp. 823-831, 2012.
13. Siddu P. Algur. , and Jyoti G.Biradar., "Rating Consistency and Review Content based Multiple stores Review Spam Detection" In IEEE International Conference on Information Processing". ISBN 978-1-4673-7758-4/15, 2015.
14. Tak-chung Fu., "A review on time series data mining" Engineering Applications of Artificial Intelligence", Elsevier, 2011.
15. Wang G., Xie S., Liu B., and Yu P. S., "Review Graph Based Online Store Review Spammer Detection". In proceeding of the 11th IEEE International Conference on Data Mining, ICDM,Vancouver, Canada: IEEE., pp.1242-1247, 2011.

BIOGRAPHY

Dr. Siddu P. Algur is working as Professor, Dept. of Computer Science, Rani Channamma University, Belagavi, Karnataka, India. He received B.E. degree in Electrical and Electronics from Mysore University, Karnataka, India, in 1986. He received his M.E. degree in from NIT, Allahabad, India, in 1991.He obtained Ph.D. degree from the Department of P.G. Studies and Research in Computer Science at Gulbarga University, Gulbarga. He worked as Lecturer at KLE Society's College of Engineering and Technology and worked as Assistant Professor in the Department of Computer Science and Engineering at SDM College of Engineering and Technology, Dharwad. He was Professor, Dept. of Information Science and Engineering, BVBCET, Hubbli, before holding the present position. He was also Director, School of Mathematics and Computing Sciences, RCU, Belagavi. He was also Director, PG Programmes, RCU, Belagavi. His research interest includes Data Mining, Web Mining, Big Data and Information Retrieval from the web and Knowledge discovery techniques. He published more than 45 research papers in peer reviewed International Journals and chaired the sessions in many International conferences.(mail id : siddu_p_algur@hotmail.com)

Mrs. Jyoti. G.Biradar is pursuing Ph.D programme in Computer Science at Rani Channamma University Belagavi, Karnataka, India. She received MCA & M.Phil degrees from Indira Gandhi National Open University and Vinayak Missions University, India in 2005 and 2009 respectively. Her research interest are Data Mining, Text Mining, and Information Retrieval from the web and Knowledge discovery techniques, and published 05 research papers in International Journals. She has attended and participated in her research field in International and National Conferences and Workshops .(mail id : jyoti.patil9131@gmail.com)

Mr. Prashant Bhat is pursuing Ph.D programme in Computer Science at Rani Channamma University Belagavi, Karnataka, India. He received B.Sc and M.Sc (Computer Science) degrees from Karnataka University, Dharwad, Karnataka, India, in 2010 and 2012 respectively. His research interest includes Data Mining, Web Mining,web multimedia mining and Information Retrieval from the web and Knowledge discovery techniques, and published 20 research papers in International Journals. Also he has attended and participated in International and National Conferences and Workshops in his research field. (mail id : prashantrcu@gmail.com)