www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 6 June 2017, Page No. 21638-21649
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.16

Cooks Distance and Mahanabolis Distance Outlier Detection Methods to identify Review Spam

Siddu P. Algur[1], Jyoti G. Biradar[2]
[1]Department of Computer Science, Rani Channamma University, Belagavi-591156,
siddu_p_algur@hotmail.com
[2]Department of Computer Science, Rani Channamma University, Belagavi-591156,
jyoti.patil9131@gmail.com

Abstract: ***In the era of web 2.0, huge volumes of consumer reviews are posted to the internet every day. As the access to Internet has been so much easier, there is an increase in people using online applications more than ever. Online marketing, in fact, the whole e-commerce is getting enormous day by day if not in every minute. Online reviews play a very important role in this field and proved itself to be auspicious in terms of decision making from a customer's point of view. Customers use the reviews for deciding quality of products before purchasing them. Companies or vendors use opinions to take a decision to improve their sales according to intelligent things done by other competitors. However, all reviews given by customers or users are not true reviews. Manual approaches to detecting and analyzing fake reviews are not practical due to the problem of information overload. The design and development of automated methods of detecting fake reviews is a challenging research problem. The main reason is that fake reviews are specifically composed to mislead readers, so they may appear the same as legitimate reviews. As a result, discriminatory features that would enable individual reviews to be classified as spam or ham may not be available. The main contribution of this study is the design and instantiation of novel computational models for detecting fake reviews. Hence, a novel approach, distance based outlier detection methods namely Cooks distance and Mahanabolis distance is used to identify spam reviews. M***

Keywords: Reviews, Opinion spam, Cooks distance, Mahanabolis distance, Outlier detection

1. Introduction

Evaluative texts on the web have become a valuable source of opinions on products, services, events, individuals, etc. Recently, many researchers have studied such opinion sources as product reviews, forum posts, and blogs. The importance of online reviews for products and services is significant for today's businesses. Opinion sharing websites enable customers to post their opinions regarding purchased and utilized products. These posted reviews provide useful information for potential customers. In fact, it is quite helpful for a potential customer to read reviews of a product before making a purchase decision. Furthermore, business holders and manufacturers use product reviews not only to understand their customer's needs but also to determine the weaknesses of their product that they can customize and reshape the product to increase customer satisfaction and consequently increase sales [9]. In recent years, opinion-sharing websites are turning into a competitive arena for businesses. Unfortunately, there is an enormous drawback with most of the opinion sharing websites. These sites enable anyone from anywhere in the world to post reviews on products without any limits. This ease of posting allows manufacturers and organizations to hire spammers to post spurious positive reviews to promote or support their merchandise and sometimes unfair negative reviews to damage competitor's reputations and degrade their reliability. Most of these harmful fake reviews are not detectable by readers, due to their manipulated structure and their placement in the midst of truthful reviews. Opinion mining techniques are being employed to distinguish fake reviews from real opinions. One of the most significant issues in opinion mining is opinion-spam detection, which requires the researcher's attention more, than other issues because the trustworthiness of a review makes it valuable for various purposes. The difference between opinion spam and other forms of spam makes opinion-spam detection more challenging. An ordinary reader is able to detect almost all other types of spamming activities easily. However, it is very hard, if not impracticable, to identify fake reviews by manually reading the reviews [7]. Considering the easily accessibility of the reviews and the significant impacts to the retailers, there is an increasing incentive to manipulate the reviews, mostly profit driven. Without proper protection, spam reviews will cause gradual loss of credibility of the reviews and corrupt the entire online review systems eventually[16]. Therefore, review spam detection is considered as the first step towards securing the online review systems.

Outlier detection has become an important part of time series analysis. An outlier is a piece of data or observation that deviates drastically from the given norm or average of the data set. Outlier detection is the process of finding data objects with behaviours that are different from expectation or they are the few observations or records which appear to be inconsistent with the rest of the group of the sample and more effective on prediction values [3]. Outliers can be classified into three categories, namely global outliers, contextual (or conditional) outliers and collective outliers. In this work, outliers are identified using the four dimensions: average positive score, average negative score, average rating and average number of reviews identified from the reviews of stores Auto-parts_warehouse.com, Dhgate.com and neweggs.com extracted from review website resellerratings.com. The proposed work is categorized as contextual outlier which is also known as conditional outlier. As in the proposed work, using distance based outlier detection methods used in multivariate data namely, Cooks distance and Mahanabolis distance, a condition is stated, that the values above the cut-off i.e above the threshold values are considered as outliers. Cook's distance, is used in regression analysis to find influential outliers in a set of predictor variables. It is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis and is named after the American statistician R. Dennis Cook, who introduced the concept in 1977.The Mahalanobis distance is a measure of the distance between a point and a distribution , introduced by P. C. Mahalanobis in 1936. It is scale-invariant, and takes into account the correlations of the data set. Cooks and Mahanabolis Distance are used to find outliers in a set of predictor variables (independent variables) which are used to find spamicity of the reviews. Four dimensional values used are independent variables or predictors which are dependent on time/date for the specified duration. The conventional cut-off/threshold value used for Cooks distance is 4/n and for Mahanabolis distance the value is 18.47. The day/dates reviews found above the conventional cut-off/threshold value are outliers and these days reviews are suspected as spam reviews. Review spamicity is measured considering total number of spam reviews detected using outlier detection methods by total number of reviews of the stores for the duration of 623 days. The results show that the distance based outlier detection methods proposed in this work are effective in detecting outliers, and to find review spamicity measure for reviews of the stores. The present paper depicts about the trends of detection of review spamicity with respect to multidimensional time series. 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.

2. Related work

In [4], an early work on detecting review spammers which proposed scoring techniques for the spamicity degree of each reviewer is used and the detection methods are based on several predefined abnormalities indicators.The features weights were linearly combined towards a spamicity formula and computed empirically in order to maximize the value of the normalized discounted cumulative gain measure. In [5], the authors considered the triangular relationship among stores, reviewers and their reviews and proposed a heterogeneous graph based model, called the review graph, with 3 types of nodes, each type of node being characterized by a spamicity score inferred using the other 2 types. In [2], the authors tried to solve the problem of opinion spam resulted from group collaboration between multiple spammers. The method they proposed first extracts candidate groups of users using a frequent item set mining technique and experimented both with learning to rank methods, i.e. ranking of groups based on their spamicity score and with classification and logistic regression, using the labeled review data for training. In [6], they focused on detecting spammers that write reviews in short bursts. Classification was done using supervised learning, by employing human evaluation of the identified honest/deceptive reviewers. The authors relied on behavioral features to detect periods in time when review bursts per product coincided with reviewer burst. In [1],the authors built an unsupervised model called the Author Spamicity Model that aims to split the users into two clusters - truthful users and spammers. The novelty about this method is a posterior density analysis of each of the features used and is meant to validate the relevance of each model feature and also increase the knowledge on their expected values for truthful and fake reviews respectively. In [10],the authors made an interesting observation in their study, that the spammers caught by Yelps filter seems to have overdone faking tried to sound more genuine. And in their deceptive reviews, they tried to use words that appear in genuine reviews almost equally frequent, thus avoiding reusing the exact same words in their reviews. In [11], the authors ran a logistic regression classifier on a model trained on duplicate or near-duplicate reviews as positive training data, i.e. fake reviews, and the rest of the reviews they used as truthful reviews. They combined reviewer behavioral features with textual features and they aimed to demonstrate that the model could be generalized to detect non-duplicate review spam. In [15], the authors observed the normal flow of reviews is not correlated with the given ratings over time. Fake reviews come in bursts of either very high ratings, i.e. 5-stars, or very low ratings, i.e. 1-star, and aimed to detect time windows in which these abnormally correlated patterns appear.

3. Methodology

In the proposed work, multivariate outlier detection techniques namely, Cooks distance and Mahanabolis distance is used to identify spamicity of the reviews based on constructing multidimensional time series for the extracted reviews of the three stores namely Auto_parts_warehouse.com, Dhgate.com and Neweggs.com. From these reviews of stores, four dimensions [14] are identified and used. The dimensional values are normalized in the range 0-1. The length of the time window is chosen to be of one day. These dimensional values are assigned to the two multivariate outlier detection methods Cooks Distance and Mahanabolis distance, month wise, i.e for each month all the four dimensional values are given to detect outliers and spamicity of the reviews for the duration of 623 days from 1st January 2014 to 15th September 2015.

The various steps of the proposed method include:

- Review Extraction.
- Identifying dimensions.
- Time series construction.
- Review Spamicity
  - Outlier detection using two multivariate outlier detection methods namely,
    - i) Cooks distance and
    - ii) Mahanabolis distance
  - Identifying review spam and to measure spamicity of reviews

- **Review Extraction** :

Reviews are extracted from review website for the stores Auto_parts_warehouse.com, Dhgate.com and Neweggs.com using review exactor tool (import.io) and are stored in raw review database for all the three stores separately.

- **Identifying dimensions** :

There are various dimensions which are used to support detection of spamicity of reviews like review similarity spam score, rating similarity score, rating deviation score, positive word length score, negative word length score, review word length score, average rating , average rating, total number of reviews, ratio of singleton reviews etc. 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. The specific of these four dimensions has been discussed in [14].

- **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 In the proposed work, review spamicity detection approach is based on multidimensional time series construction. The description of time series has been given in the work [14].

- **Review spamicity:**

Review spamicity is the degree or measure of spam reviews identified from the given dataset of reviews. In the proposed work, detection of review spam and measure of review spamicity is given in two steps as follows:

➢ Outlier detection using two multivariate outlier detection methods
i) Cooks distance and
ii) Mahanabolis distance

➢ Identifying review spam days and to measure spamicity of the reviews.

➢ **Outlier detection using a multivariate outlier detection method namely Cooks distance**

Cooks distance is the distance based outlier detection method used to identify outliers in multivariate data. It is used in regression analysis, to find outliers in a set of predictor variables (independent variables) and also used to estimate the influence of a data points while performing a least square regression analysis [8]. In the proposed work, four dimensions used are independent variables or predictors which are dependent on time for the specified duration. As regression analysis is a statistical process for estimating the relationships among variables and includes many techniques for modeling and analyzing several variables, on the relationship between a dependent variable and one or more independent variables (or predictors), this approach is used. The formula used for cooks distance is

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{(p+1)\,\widehat{\sigma}^2}$$

Where,

Y_j = the prediction from the full regression model for observation j

$Y_{j(i)}$ = the prediction for observation j from a refitted regression

$\boldsymbol{\sigma^2}$ **=** estimator of the error variance.

p = predicted variables (independent variables)

As the formula can get quite cumbersome by hand, the formula is used in a software SPSS to detect outliers. The four dimensions average positive score, average negative score, average rating and average number of reviews are independent variables and date(time) is dependent variable used in the software for the specified duration. The time window chosen in this work is one day.

Few interpretations to detect outliers using Cooks distance are:

i. A general rule of thumb is that observations with a Cook's distance of more than 3 times the mean, is a possible outlier

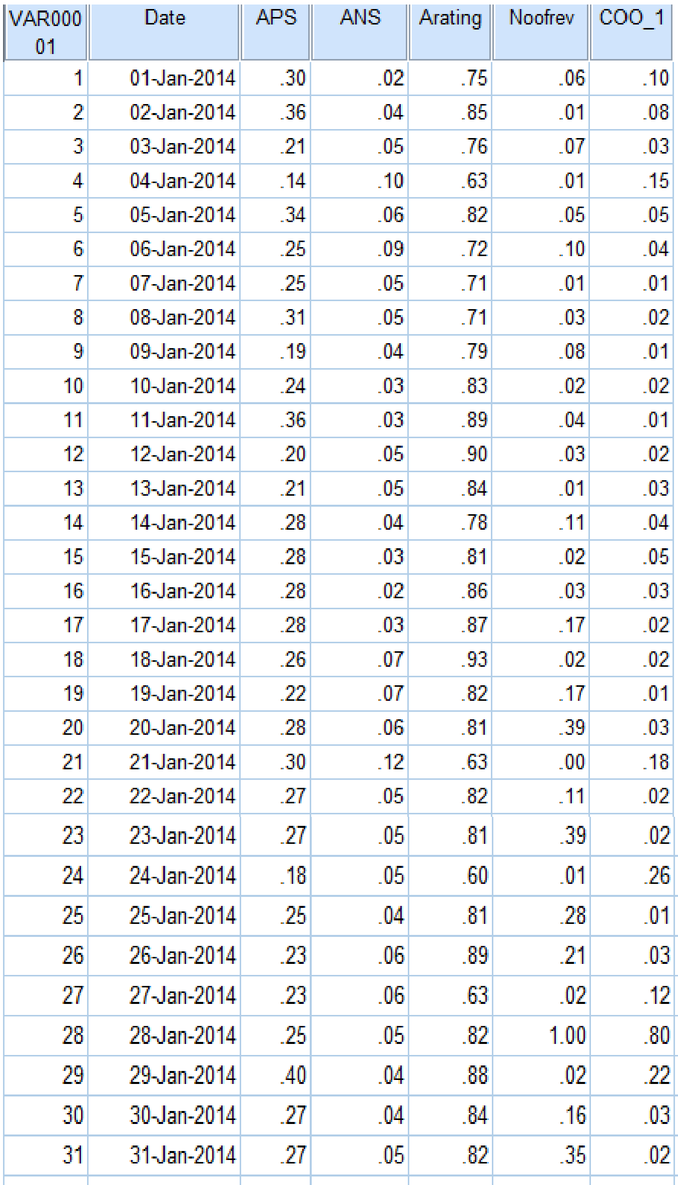
ii. An alternative interpretation is to investigate any point over 4/n, where n is the number of observations.

iii. Another alternative way is to find the potential outliers percentile value using the F-distribution. A percentile of over 50 indicates a highly influential point.

In this work, outlier detection using cooks distance is based on the second interpretation i.e the conventional cut-off point is 4/n. The day/dates reviews found above the conventional cut-off points are outliers and these day/dates reviews are suspected as spam reviews. Using four independent dimensional variable values and a dependent variable date/time in the cooks distance using SPSS, a column will appear in the data cells with the Cook's distance values (Cook's D values). With these Cooks D values, a graph is plotted, considering the Cooks distance values and time (month-wise). A sample of a month January 2014 of a store named Auto_parts_warehouse.com is shown in the Table 1. As the conventional cut-off is 4/n, the review day/dates found above 4/n (i.e 4/31) are suspected as spam reviews.

Review spamicity is measured by considering number of reviews found above the conventional cut-off points by total number of reviews for the particular month.

Table 1. A column with Cook's Distance value for a month January 2014 of a store Auto_parts_warehouse.com



VAR00001	Date	APS	ANS	Arating	Noofrev	COO_1
1	01-Jan-2014	.30	.02	.75	.06	.10
2	02-Jan-2014	.36	.04	.85	.01	.08
3	03-Jan-2014	.21	.05	.76	.07	.03
4	04-Jan-2014	.14	.10	.63	.01	.15
5	05-Jan-2014	.34	.06	.82	.05	.05
6	06-Jan-2014	.25	.09	.72	.10	.04
7	07-Jan-2014	.25	.05	.71	.01	.01
8	08-Jan-2014	.31	.05	.71	.03	.02
9	09-Jan-2014	.19	.04	.79	.08	.01
10	10-Jan-2014	.24	.03	.83	.02	.02
11	11-Jan-2014	.36	.03	.89	.04	.01
12	12-Jan-2014	.20	.05	.90	.03	.02
13	13-Jan-2014	.21	.05	.84	.01	.03
14	14-Jan-2014	.28	.04	.78	.11	.04
15	15-Jan-2014	.28	.03	.81	.02	.05
16	16-Jan-2014	.28	.02	.86	.03	.03
17	17-Jan-2014	.28	.03	.87	.17	.02
18	18-Jan-2014	.26	.07	.93	.02	.02
19	19-Jan-2014	.22	.07	.82	.17	.01
20	20-Jan-2014	.28	.06	.81	.39	.03
21	21-Jan-2014	.30	.12	.63	.00	.18
22	22-Jan-2014	.27	.05	.82	.11	.02
23	23-Jan-2014	.27	.05	.81	.39	.02
24	24-Jan-2014	.18	.05	.60	.01	.26
25	25-Jan-2014	.25	.04	.81	.28	.01
26	26-Jan-2014	.23	.06	.89	.21	.03
27	27-Jan-2014	.23	.06	.63	.02	.12
28	28-Jan-2014	.25	.05	.82	1.00	.80
29	29-Jan-2014	.40	.04	.88	.02	.22
30	30-Jan-2014	.27	.04	.84	.16	.03
31	31-Jan-2014	.27	.05	.82	.35	.02

With the Cook's Distance value appeared, a graph is plotted as shown in the Figure 1.

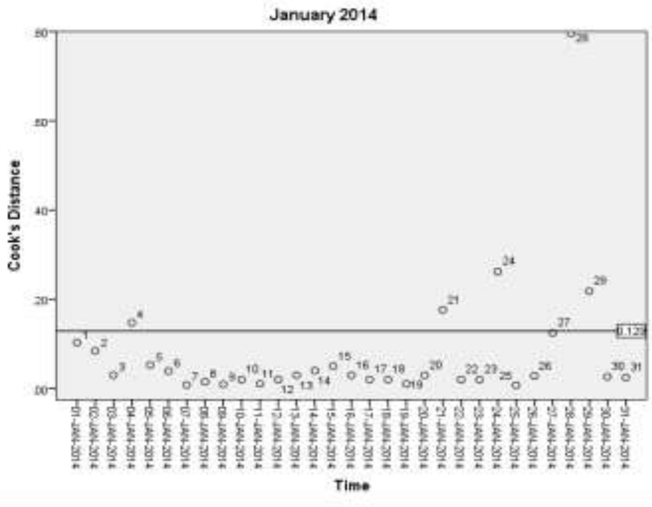


Figure 1. Spam reviews identified from Cooks distance value for January2014 of a store auto_parts _warehouse.com

From the Figure 1, the conventional cut-off value (threshold value) is 0.129. As in a month January, there are 31 days, and conventional cut-off is 4/n, (4/31) is 0.129. The review date/ days found above the cut-off/threshold value, are outliers and these reviews are suspected as spam reviews. Spamicity of the reviews is calculated by number of reviews found above the cut-off /threshold value by total number of reviews of the specified duration.

Similarly, cooks distance is calculated for all the remaining reviews of this store for the entire duration and with the two stores namely Dhgate.com and Neweggs.com respectively.

- ➢ **Outlier detection using a multivariate outlier detection method namely Mahanabolis distance**

Mahalanobis distance is the distance between a data point and a multivariate space's centroid (overall mean). Mahalanobis distances are based on both the mean and variance of the predictor variables, plus the covariance matrix of all the variables, and therefore take advantage of the covariance among variables[8].

The formula used for Mahanabolis distance is

$$Dj = (x - m)^T C^{-1} (x - m)$$

Where,

Dj = Mahalanobis distance
x = vector of data
m = vector of mean values of independent variables
T = Indicates vector should be transposed.
C^{-1} = Inverse covariance matrix of Independent variables

Using this standard formula, outliers are identified for the above mentioned four dimensional values which are considered as independent variables or predictors which are dependent on time/date (per day) for the specified duration.

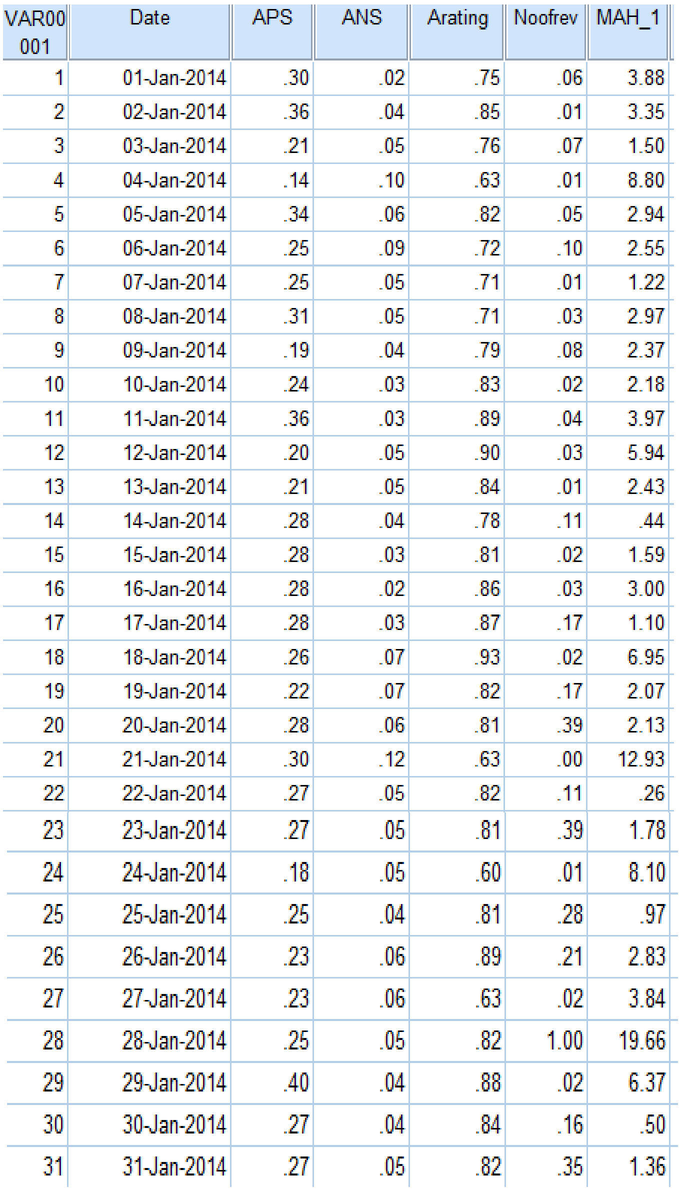
In Mahanabolis distance, critical values (threshold values) are assigned to the multivariate data. The list of the critical values assigned for different variants (attributes/ dimensions) are given in the Table 1:

Table 1: Critical value assigned to the number of attributes

No of attributes/ dimensions	Critical value
2	13.82
3	16.27
4	**18.47**
5	20.52
6	22.46
7	24.32
8	26.13
9	27.88
10	29.59

As in the proposed work, four dimensional values are used, the critical value (threshold is 18.47), the mahanabolis distance values found above 18.47 are outliers, and these outliers day/dates reviews are suspected as spam reviews. Using four independent dimensional variable values and a dependent variable date/time in the Mahanabolis distance using SPSS, a column will appear in the data cells with the Mahanabolis distance values. A sample of a month January 2014 of a store named Auto_parts_warehouse.com is shown in the Table 2.

Table 2. A column with Mahanabolis Distance value for a month January 2014 of a store Auto_parts_warehouse.com.



VAR00001	Date	APS	ANS	Arating	Noofrev	MAH_1
1	01-Jan-2014	.30	.02	.75	.06	3.88
2	02-Jan-2014	.36	.04	.85	.01	3.35
3	03-Jan-2014	.21	.05	.76	.07	1.50
4	04-Jan-2014	.14	.10	.63	.01	8.80
5	05-Jan-2014	.34	.06	.82	.05	2.94
6	06-Jan-2014	.25	.09	.72	.10	2.55
7	07-Jan-2014	.25	.05	.71	.01	1.22
8	08-Jan-2014	.31	.05	.71	.03	2.97
9	09-Jan-2014	.19	.04	.79	.08	2.37
10	10-Jan-2014	.24	.03	.83	.02	2.18
11	11-Jan-2014	.36	.03	.89	.04	3.97
12	12-Jan-2014	.20	.05	.90	.03	5.94
13	13-Jan-2014	.21	.05	.84	.01	2.43
14	14-Jan-2014	.28	.04	.78	.11	.44
15	15-Jan-2014	.28	.03	.81	.02	1.59
16	16-Jan-2014	.28	.02	.86	.03	3.00
17	17-Jan-2014	.28	.03	.87	.17	1.10
18	18-Jan-2014	.26	.07	.93	.02	6.95
19	19-Jan-2014	.22	.07	.82	.17	2.07
20	20-Jan-2014	.28	.06	.81	.39	2.13
21	21-Jan-2014	.30	.12	.63	.00	12.93
22	22-Jan-2014	.27	.05	.82	.11	.26
23	23-Jan-2014	.27	.05	.81	.39	1.78
24	24-Jan-2014	.18	.05	.60	.01	8.10
25	25-Jan-2014	.25	.04	.81	.28	.97
26	26-Jan-2014	.23	.06	.89	.21	2.83
27	27-Jan-2014	.23	.06	.63	.02	3.84
28	28-Jan-2014	.25	.05	.82	1.00	19.66
29	29-Jan-2014	.40	.04	.88	.02	6.37
30	30-Jan-2014	.27	.04	.84	.16	.50
31	31-Jan-2014	.27	.05	.82	.35	1.36

With the Mahanabolis distance value appeared, a graph is plotted as shown in the Figure 2

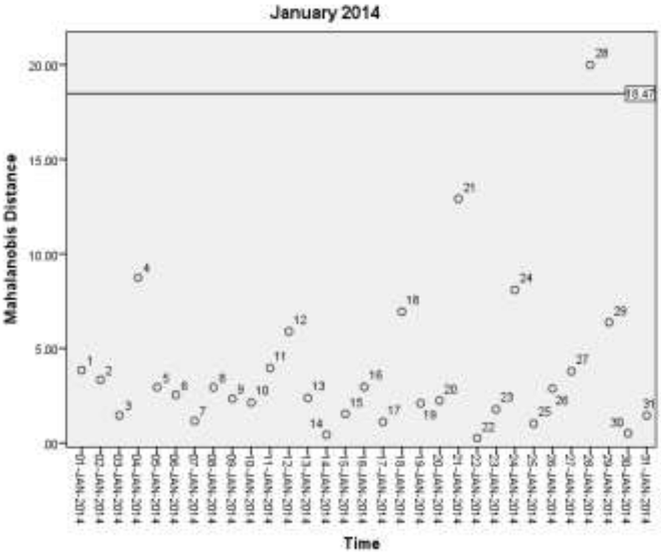


Figure 2. Spam reviews identified from Mahanabolis distance value for January 2014 of a store Auto_parts_warehouse.com

From the Figure 2, the critical value i.e the conventional cut-off value (threshold value) used is 18.47. Hence the review dates/days found above the critical value of Mahanabolis distance (18.47) are suspected as spam reviews. Spamicity of the reviews is calculated by number of reviews found above the threshold value by total number of reviews of the specified duration. Similarly, Mahanabolis distance is calculated, spam reviews are identified and spamicity of the reviews are measured for all the remaining reviews of this store month wise and with the two stores namely Dhgate.com and Neweggs.com for the entire duration.

- **Identifying review spam days and to measure spamicity of the reviews.**

Review spam days are identified using Cooks & Mahanabolis distance, which are distance based outlier detection techniques.The four dimensional values are independent variables which are dependent on the variable day/date (time). With these dependent and independent variables, using the standard formula for both Cooks and Mahanabolis distance, a graph is plotted with the cooks distance & Mahanabolis distance column values. The reviews found for those dates/days which are above the specified cut-off/ threshold values assigned to both the methods, are suspected as spam day reviews.

Spamicity of the reviews are measured by considering the number of spam days reviews found, by the total number of reviews of the stores for the entire duration.

In the proposed work, review spamicity is computed for the three stores namely Auto_parts_warehouse.com,Dhgate.com and Neweggs.com. The spamicity of these stores are computed by considering four dimensions which are mentioned earlier [13] .

4. Experimental results

Experimental results are obtained to demonstrate the effectiveness of the proposed method. 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 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 1 to 5), the posting date and content of the review. Detection of review spamicity is constructed on multidimensional time series analysis from the extracted reviews of the three stores based on the average of positive word length score, negative word length score, review ratings and number of reviews. These four dimensions, are considered as independent variables, and day/date (i.e time) as dependent variable. With these dimensional values, using SPSS, a column is generated for Cooks distance and Mahanabolis distance. A graph is plotted with the column values generated separately for both the outlier detection methods used. The conventional cut-off/ threshold for Cooks distance is 4/n and for Mahanabolis distance the value is 18.47.The days/dates found above the cut-off/threshold value are outliers, and these day/dates reviews are suspected as spam day reviews. For all the four dimensions, outliers detected dates/days are marked month wise from 1st January 2014 to 15th September 2015 .The size of the time window is set to one day. Spamicity of the reviews of the three stores are measured, considering the total number of spam reviews found using distance based outlier detection techniques Cooks and Mahanabolis distance by the total number of reviews of the stores for duration of 623 days.

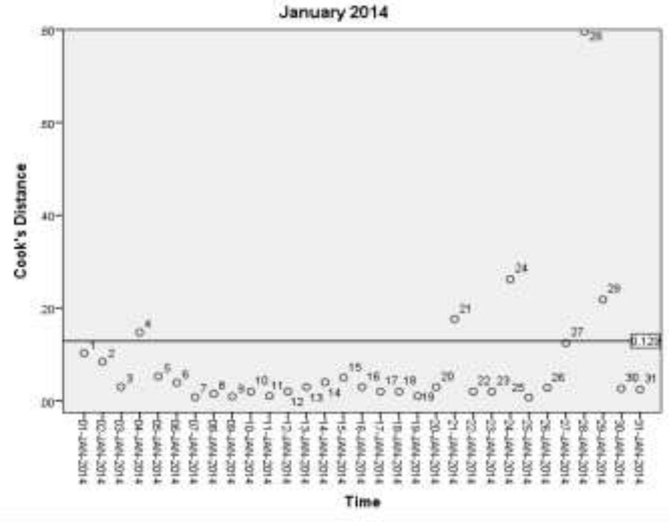


Figure 3. Spam reviews identified from Cooks distance value for January 2014 of a store Auto_parts_warehouse.com

From the Figure 3, the conventional cut-off value (threshold value) is 0.129. As in a month January, there are 31 days, and conventional cut-off value used in Cooks distance in this work is 4/n, (4/31 is 0.129). The review date/days found above the threshold value, are outliers and these days reviews are suspected as spam reviews. The dates 04-01-14, 21-01-14, 24-01-14, 28-01-14 and 29-01-14, are found above the cut-off value (0.129), these days/ dates reviews are suspected as spam reviews. The reviews found for these days are 425; these reviews are suspected as spam reviews. Review spamicity measure is calculated considering total number of spam reviews identified i.e, 425 reviews by total number of reviews of the store for the month January 2014 i.e 1701 reviews. The review spamicity measure for the month January 2014 of the store Auto_parts_warehouse is 24.98%.

Similarly, spamicity of the reviews for the store Auto_parts_warehouse.com is calculated for the remaining months for duration of 623 days is shown in the Table 3.

In the Figure 4, Spam reviews are identified using Cooks distance for a month January 2014 of the store Dhgate.com.

DOI: 10.18535/ijecs/v6i6.16

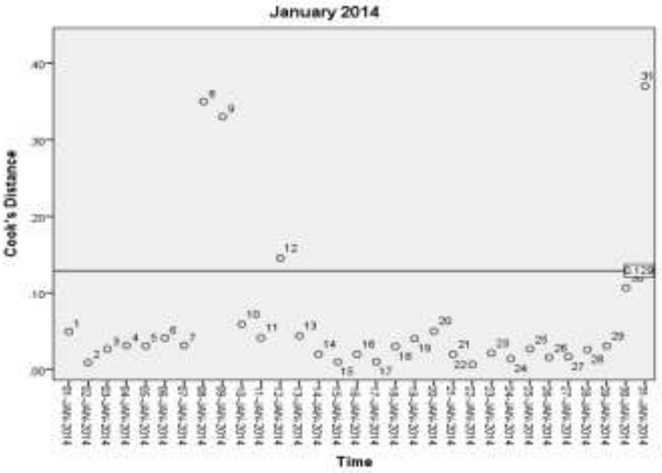


Figure 4. Spam reviews identified from Cooks distance value for January 2014 of a store Dhgate.com

From the Figure 4, the review date/days found above the cut-off threshold value (0.129) are outliers and these days reviews are suspected as spam reviews. The dates 08-01-14, 09-01-14, 12-01-14 and 31-01-14, are found above the cut-off/threshold value, these days/ dates reviews are suspected as spam reviews. The reviews found for these days are 19, these reviews are suspected as spam reviews. Review spamicity measure is calculated considering total number of spam reviews identified i.e, 19 reviews by total number of reviews of the store for the month January 2014 i.e 88 reviews. The review spamicity measure for the month January 2014 of the store Dhgate.com is 24.98%. Similarly, spamicity of the reviews for the store Dhgtae.com is calculated for the remaining months for duration of 623 days is shown in the Table 3.

In the Figure 5, Spam reviews are identified using Cooks distance for a month January 2014 of the store Neweggs.com.

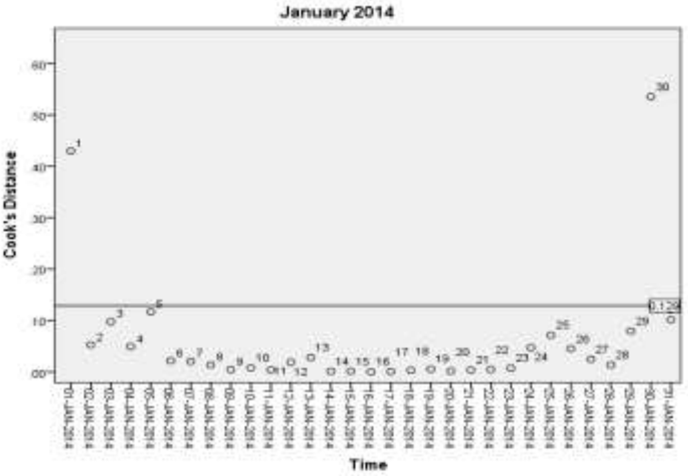


Figure 5. Spam reviews identified from Cooks distance value for January 2014 of a store Neweggs.com

From the Figure 5, the review date/days found above the cut-off threshold value (0.129) are outliers and these days reviews are suspected as spam reviews. The dates 01-01-14 and 30-01-14 are found above the cut-off/threshold value, these days/ dates reviews are suspected as spam reviews. The reviews found for these days are 48, these reviews are suspected as spam reviews. Review spamicity measure is calculated considering total number of spam reviews identified, i.e, 48 reviews by total number of reviews of the store for the month January 2014, i.e 175 reviews. The review spamicity measure for the month January 2014 of the store Neweggs.com is 27.43%. Similarly, spamicity of the reviews for the store Neweggs.com is calculated for the remaining months for duration of 623 days is shown in the Table 3.

Table 3. Comparative table of Total number of Reviews, Number of reviews detected as spam and Percentage of reviews detected as spam for the three stores using Cooks distance.

Names of Stores	**Auto_parts_warehouse. com**			**Dhgate.com**			**Neweggs.com**		
Date	Total number of Reviews	Number of reviews detected as Spam	% of reviews detected as spam	Total number of Reviews	Number of reviews detected as spam	% of reviews detected as spam	Total number of Reviews	Number of reviews detected as Spam	% of reviews detected as spam
Jan-14	1701	425	24.99	88	19	21.59	175	48	27.43
Feb-14	2536	679	26.77	62	17	27.42	198	52	26.26
Mar-14	2498	430	17.21	46	12	26.09	181	28	15.47
Apr-14	2488	648	26.05	42	7	16.67	135	31	22.96
May-14	2274	831	36.54	38	5	13.16	100	29	29.00
Jun-14	2249	391	17.39	40	9	22.50	115	24	20.87
Jul-14	2481	444	17.90	729	103	14.13	135	34	25.19
Aug-14	2302	578	25.11	453	94	20.75	125	26	20.80
Sep-14	2359	379	16.07	52	10	19.23	124	29	23.39
Oct-14	2207	398	18.03	46	15	32.61	170	34	20.00

Nov-14	1225	192	15.67	48	11	22.92	219	40	18.26
Dec-14	899	125	13.90	46	11	23.91	208	30	14.42
Jan-15	1154	155	13.43	1402	259	18.47	192	36	18.75
Feb-15	297	108	36.36	1220	241	19.75	160	17	10.63
Mar-15	102	30	29.41	1288	249	19.33	147	20	13.61
Apr-15	263	61	23.19	1303	269	20.64	132	28	21.21
May-15	189	91	48.15	1306	300	22.97	198	28	14.14
Jun-15	175	43	24.57	1252	324	25.88	139	22	15.83
Jul-15	43	16	37.21	1139	238	20.90	184	29	15.76
Aug-15	56	12	21.43	1424	293	20.58	192	22	11.46
Sep-15	24	7	29.17	489	77	15.75	52	11	21.15
Total no of Reviews/ No of reviews detected as spam / % of reviews detected as spam	**27522**	**6043**	**21.96%**	**12513**	**2563**	**20.48%**	**3281**	**618**	**18.84%**

As a result, the spamicity of the reviews measured using Cooks Distance outlier detection method are 21.96%, 20.48% and 18.84% for the store Auto_parts_warehouse.com, Dhgate.com and Neweggs.com respectively.

In the Figure 6, Spam reviews are identified using Mahanabolis distance for a month January 2014 of the store Auto_parts_warehouse.com

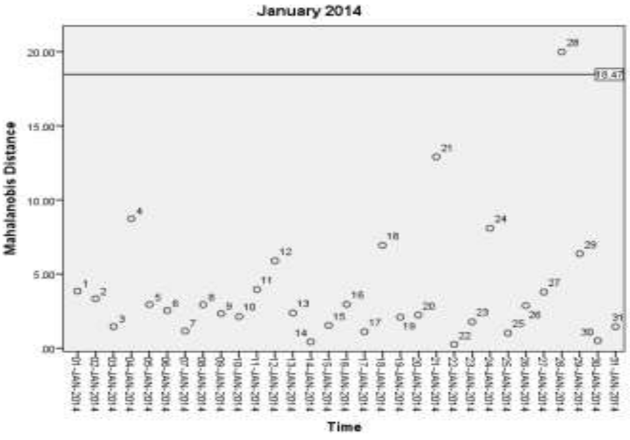


Figure 6. Spam reviews identified from Mahanabolis distance value for January 2014 of a store Auto_parts_warehouse.com

From the Figure 6, the review date/days found above the cut-off threshold value (18.47) are outliers and these days reviews are suspected as spam reviews. The date 28-01-14, is found above the cut-off/threshold value, these days/ dates reviews are suspected as spam reviews. The reviews found for these days are 363, these reviews are suspected as spam reviews. Review spamicity measure is calculated considering total number of spam reviews identified i.e, 363 reviews by total number of reviews of the store for the month January 2014 ,1701 reviews. The review spamicity measure for the month January 2014 of the store Auto_parts_warehouse.com is 21.34%

Similarly, spamicity of the reviews for the store Auto_parts_warehouse.com is calculated for the remaining months for duration of 623 days is shown in the Table 4.

In the Figure 7, Spam reviews are identified using Mahanabolis distance for a month January 2014 of the store Dhgate.com

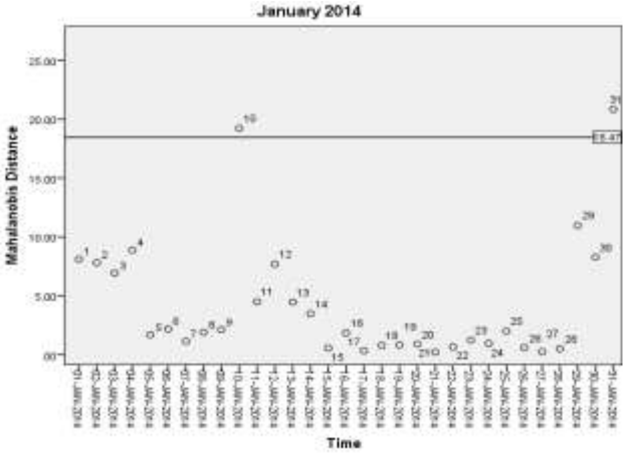


Figure 7. Spam reviews identified from Mahanabolis distance value for January 2014 of a store Dhgate.com

From the Figure 7, the review date/days found above the cut-off threshold value (18.47) are outliers and these days reviews are suspected as spam reviews. The two dates 10-01-14 & 31-01-14, are found above the cut-off/threshold value, these days/ dates reviews are suspected as spam reviews. The reviews found for these days are 15, these reviews are suspected as spam reviews. Review spamicity measure is calculated considering total number of spam reviews identified i.e, 15 reviews by total number of reviews of the store for the month January 2014, 88 reviews. The review spamicity measure for the month January 2014 of the store Dhgate.com is 17.05%

Similarly, spamicity of the reviews using Mahanabolis distance for the store Dhgate.com is calculated for the remaining months for duration of 623 days is shown in the Table 4.

In the Figure 8, Spam reviews are identified using Mahanabolis distance for a month January 2014 of the store Neweggs.com

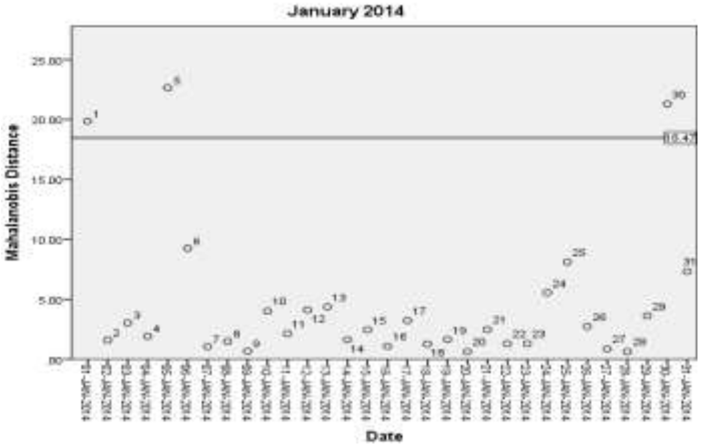


Figure 8. Spam reviews identified from Mahanabolis distance value for January 2014 of a store Neweggs.com

From the Figure 8, the review date/days found above the cut-off threshold value (18.47) are outliers and these days reviews are suspected as spam reviews. The three dates

01-01-14, 05-0-14 & 30-01-14, are found above the cut-off/threshold value, these days/ dates reviews are suspected as spam reviews. The reviews found for these days are 36, these reviews are suspected as spam reviews. Review spamicity measure is calculated considering total number of spam reviews identified, i.e, 36 reviews by total number of reviews of the store for the month January 2014, i.e 175 reviews. The review spamicity measure for the month January 2014 of the store Neweggs.com is 20.57%.Similarly, spamicity of the reviews using Mahanabolis distance for the store Neweggs.com is calculated for the remaining months for duration of 623 days is shown in the Table 4.

Table 4. Comparative table of Total number of Reviews, Number of reviews detected as spam and Percentage of reviews detected as spam for the three stores using Mahanabolis distance.

Names of Stores	**Auto_parts_warehouse. com**			**Dhgate.com**			**Neweggs.com**		
Date	Total number of Reviews	Number of reviews detected as Spam	% of reviews detected as spam	Total number of Reviews	Number of reviews detected as spam	% of reviews detected as spam	Total number of Reviews	Number of reviews detected as Spam	% of reviews detected as spam
Jan-14	1701	363	21.34	88	15	17.05	175	36	20.57
Feb-14	2536	560	22.08	62	14	22.58	198	42	21.21
Mar-14	2498	397	15.89	46	10	21.74	181	24	13.26
Apr-14	2488	555	22.31	42	5	11.90	135	22	16.30
May-14	2274	710	31.22	38	4	10.53	100	21	21.00
Jun-14	2249	330	14.67	40	6	15.00	115	18	15.65
Jul-14	2481	344	13.87	729	90	12.35	135	25	18.52
Aug-14	2302	479	20.81	453	78	17.22	125	18	14.40
Sep-14	2359	315	13.35	52	7	13.46	124	24	19.35
Oct-14	2207	269	12.19	46	14	30.43	170	27	15.88
Nov-14	1225	149	12.16	48	9	18.75	219	34	15.53
Dec-14	899	97	10.79	46	10	21.74	208	28	13.46
Jan-15	1154	123	10.66	1402	233	16.62	192	32	16.67
Feb-15	297	93	31.31	1220	208	17.05	160	15	9.38
Mar-15	102	22	21.57	1288	236	18.32	147	16	10.88
Apr-15	263	52	19.77	1303	228	17.50	132	23	17.42
May-15	189	60	31.75	1306	249	19.07	198	17	8.59
Jun-15	175	28	16	1252	299	23.88	139	18	12.95
Jul-15	43	14	32.56	1139	229	20.11	184	24	13.04
Aug-15	56	9	16.071	1424	242	16.99	192	17	8.85
Sep-15	24	5	20.83	489	63	12.88	52	10	19.23
Total no of Reviews/ No of reviews detected as spam / % of reviews detected as spam	**27522**	**4974**	**18.07 %**	**12513**	**2249**	**17.97%**	**3281**	**491**	**14.96%**

As a result, the spamicity of the reviews measured using Mahanabolis distance outlier detection method are 18.07%, 17.97% and 14.96% for the store Auto_parts_warehouse.com, Dhgate.com and Neweggs.com respectively. From the experimental results, one can observe that, if the number of reviews are more, spamicity of the reviews will be less. And if the numbers of reviews are less, spamicity of the reviews will be more [13]. There are even large numbers of non-spam reviews also. Hence, these reviews do not influence the buying decision of the customers and could be regarded as trustworthy as they provide genuine opinion on some or the other sentiment of the store and are often unbiased [12].

5. Conclusion and future work

In this work, a novel evaluation methods, distance based outlier detection methods, namely Cooks distance and Mahanabolis distance is used to find review spamicity using three stores. Four dimensions are identified and used as 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 chosen to be of one day. Four dimensional values used are independent variables or predictors which are dependent on time/date for the specified duration. The conventional cut-off/threshold value used for Cooks distance is 4/n and for Mahanabolis distance the value is 18.47. The day/dates reviews found above the conventional cut-off/threshold value are outliers and these days reviews are suspected as spam reviews. Spamicity of reviews are calculated by considering number of spam reviews identified by total number of reviews of the stores for the entire duration. 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 623 days from 1st January 2014 to 15th September 2015 demonstrates that the proposed outlier detection methods are effective in detecting review spamicity.

References

[1] Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos and Riddhiman Ghosh. “Spotting opinion spammers using behavioral footprints”. In Proceedings of the 19th ACM SIGKDD International conference on Knowledge discovery and data mining, pp. 632–640. ACM, 2013.

[2] Arjun Mukherjee, Bing Liu and Natalie Glance. “Spotting fake reviewer groups in consumer reviews”. In Proceedings of the 21st International conference on World Wide Web, pp. 191–200. ACM, 2012

[3] Diethelm W¨urtz1, Yohan Chalabi, and Ladislav Luksan “Parameter Estimation of ARMA Models with GARCH/APARCH Errors An R and SPlus Software Implementation “ Journal of Statistical Software, Volume 5, Issue 5.,pp.1-41, 2015.

[4] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. “Detecting product review spammers using rating behaviors”. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp.939–948. ACM, 2010.

[5] Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. “Learning to identify review spam”. In Proceedings of International Joint Conference on Artificial Intelligence, (IJCAI) volume 22, pp. 2488-2490, 2011.

[6] Geli Fei, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. “Exploiting burstiness in reviews for review spammer detection”. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM) pp.175-184, 2013.

[7] Heydari Atefeh, Mohammad Ali Tavakoli, Naomie Salim, and Zahra Heydari. "Detection of review spam: A survey." Expert Systems with Applications 42, no.7 pp.3634-3642,2015.

[8] K. Senthamarai Kannan and K. Manoj, ” Outlier Detection in Multivariate Data”, Applied Mathematical Sciences, Vol. 9, no. 47, pp. 2317 – 2324, 2015.

[9] Mohammadali Tavakoli, Atefeh Heydari, Zuriati Ismail, Naomie Salim, “A Framework for Review Spam Detection Research“, International Journal of Computer Electrical Automation Control and Information Engineering (IJCEACIE) Vol 10, Issue 1, pp.67-71, 2016

[10] Myle Ott, Claire Cardie, and Jeffrey T Hancock. “Negative deceptive opinion spam”. In proceedings of NAACL- HLT, pp. 497–501, Atlanta, Georgia, June 9-14,2013.

[11] Nitin Jindal and Bing Liu. “Opinion spam and analysis”. In Proceedings of the International Conference on Web Search and Data Mining, pp. 219–230, ACM, 2008.

[12] 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, pp.416-423, 2010.

[13] Siddu P. Algur, Jyoti G.Biradar, Prashant Bhat.,“ GARCH(1,1) Outlier detection technique for review spam detection” International Journal of Emerging Trends and Technology in Computer Science (IJETTCS).,Vol 5,Issue 6, ISSN:2278-6856, pp.6-15, Nov-Dec 2016.

[14] Siddu P. Algur, Jyoti G. Biradar and Prashant Bhat “Multidimensional Time Series Based Review Spam

Detection" International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE) "Vol.4, Issue 6, pp.11761-11774., June 2016.

[15] Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. "Distributional footprints of deceptive product reviews". In proceedings of sixth International AAAI conference on weblogs and social media (ICWSM), pp.98-105, 2012.

[16] Yingying Ma and Fengjun Li, "Detecting Review Spam:Challenges and Opportunities", In proceedings of eighth International Conference on Collaborative Computing: Networking Applications and Work sharing, Pittsburgh, PA, United States, October 14-17, 2012.

Authors Profile

Prof. Siddu P. Algur is a Registrar, 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 55 research papers in peer reviewed International Journals and chaired the sessions in many International conferences

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 09 research papers in International Journals. She has attended and participated in her research field in International, National Conferences and Workshops and is recipient of Best paper award in International Conference (ICSNS-2015) .