Volume 8, No. 3, March – April 2017
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info
ISSN No. 0976-5697

Exponential Distribution model for Review Spam Detection

Siddu P. Algur, N. H. Ayachit and Jyoti G.Biradar*
Department of Computer Science
Rani Channamma University
Belagavi – 591156 , Karnataka, India

Abstract: Online reviews capture the testimonials of real people and help shape the decisions of other consumers. It has become very crucial for e-Commerce trades to empower their end customers to write reviews about the services that they have utilized. Such reviews provide vital sources of information on these products/stores. Review information is utilized by the future potential customers before deciding on purchase of new products or services. These opinions or reviews are also exploited by marketers to find out the drawbacks of their own products or services and alternatively to find the vital information related to their competitor's products or services. This in turn allows identifying weaknesses or strengths of the products/stores. Unfortunately, this significant usefulness of opinions has also raised the problem for spam, which contains forged positive or spiteful negative opinions. These reviews are written due to the financial gains associated with positive reviews, with often paid spam reviewers writing fake reviews to unjustly promote or demote certain products or businesses. Identifying such opinion spam reviews have become a challenge in opinion mining. Hence, in this work a novel approach exponential distribution model is used to find review spamicity. This method significantly outperforms several baselines and other methods.

Keywords: Reviews, review spam, multiple criteria, exponential distribution model

1. INTRODUCTION

As the technology changes for publicity, way to traditional marketing also changes as person-to-person communicate through online reviews. These online reviews are important to customers and to companies or vendors and are helpful for making decisions regarding quality of products and services. Companies and vendors use opinions to take a decision to improve the sales according to intelligent things done from other competitors, for marketing strategies, performance to services or product, for improvement. All reviews given by the customers or users are not true reviews. These reviews are given to promote or to demote the product. Product reviews are an increasingly important type of user generated online content since they offer valuable information that helps product designers better understand the needs and preferences of consumers, and in the meantime, influences potential consumers in their purchase decision making. The web contains a wealth of opinions about products, politicians, and more, which are expressed in newsgroup, posts, review sites, and elsewhere. As a result, the problem of opinion mining has seen increasing attention over the last decade [7]. It is now a common practice for e-Commerce web sites to enable their customers to write reviews of products that they have purchased. The reviews are then used by potential customers to find opinions of existing users before purchasing the products and also to identify problems in their products and/or to find competitive intelligence information about their competitor [12]. As internet has no quality control, anyone can write anything on the web, which results in many low quality reviews, and worse still review spam which is often biased and may mislead the customer affecting his buying decisions. Thus, it is very essential to have a mechanism which is capable of assessing the trustworthiness of reviews for proper decision making or for marketing intelligence. Trusted customer reviews are useful for both potential buyers and product manufacturers. It is more convenient and less time consuming for buyer to see at a glance feature by feature comparison of reviews written by most of the customers in taking buying decisions without getting biased and product manufacturer gets to know strengths and weaknesses of his/her own products and also that of the competitors, consumer preferences and interests by which profits could be maximized [4]. However, the intentions to all customers of users are not true for writing reviews. This concepts, changes the face of advertising to conventional, individual-to individual correspondence to online audits. These online audits are important to client and to organizations or sellers. Considering the increasing damage caused by review spam, it is a critical and urgent task to detect review spam. But, this is unsurprisingly difficult since it is hard to filter out, even manually, a spam review or capture spammer behaviour. The reason may be two-fold, the subjective nature of the reviews and the human-generated contents and patterns that disguise spam behaviour. The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events and aspects of parallel applications which can be easily taken into account. Hence, in the proposed work a novel approach exponential distribution model is used to find review spamicity by using reviews from multiple stores extracted from review website resellerratings.com.

Multiple criteria's are used to find spamicity of the reviews. Eight criteria's are identified and used namely word length score, number of reviews, review rating, Positive Review Length Difference (PRLD), Negative Review Length Difference (NRLD), ALL CAPS, advertisement link, reviewer names end with three numbers. The criteria values are normalized in the range 0-1, and these normalized eight criterion values of the reviews are further grouped into ten frequency values. For the ten frequency values, exponential curves are fitted for each criteria of the review dataset.

After fitting of exponential curve for each criterion, the frequency values reviews which are found above the exponential curve are suspected as spam review i.e the reviews which are above the predicated values of the exponential curve. Spamicity of the reviews is measured by considering number of reviews found for the frequency values which are above the exponential curve by total number of reviews for the store for the entire duration.

The rest of the paper is organized as follows: Section 2, introduces about the related work. Section 3, gives an overview of the proposed technique used to find review spamicity. Section 4, describes the working and experimental results for detecting review spamicity. And finally the section 5 presents conclusion and future work.

2. RELATED WORK

In [9], the first attempt for identifying spam reviews was proposed. In this work, the authors have highlighted that there are two types of review spam, one is manipulated review, the review which will mislead the customer and another is non-review i.e. it is not giving any actual opinion about the product/store, it can be advertisement of a product/store. In [2], attempts are made to identify fake/spam reviews by using machine learning methods. They used various features of reviews and reviewer and discussed a framework of product review mining system. In [13], the authors consider reviewers behaviors by introducing a social graph connecting reviewers, their reviews and stores. They found out the reinforcement relations of reviewers trustiness, reviews honesty, and stores reliability to identify suspicious spammers. In [14], the author considers behavior-based approaches to find review spam. The indicative features of spam extracted from the metadata associated with user behaviour, review content and product profile are considered. The authors worked on 36 such features on a (pseudo) ground truth dataset, constructed by labeling the duplicate reviews in an Amazon dataset as fake reviews.

3. PROPOSED METHODOLOGY

In this section, a novel and effective technique namely, exponential distribution model is used to detect review spamicity. Review spamicity is the degree or measure of spam reviews identified from the given dataset of reviews.

The large scale reviews from the three stores namely Auto_parts_warehouse.com, Dhgate.com and Neweggs.com are considered to construct exponential distribution model. From these reviews of stores, eigth criteria's are identified and used. The criteria's values are normalized in the range of 0-1. The system architecture of the exponential distribution model is shown in the Figure1, and has the following components. (In the system model PoSR is percentage of spam reviews)

- Review extraction and pre-processing
- Construction of exponential distribution model
- Result analysis and KDD

3.1 Review extraction and pre-processing

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 1st January 2014 to 15th September 2015. From the extracted reviews, stop words are 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.

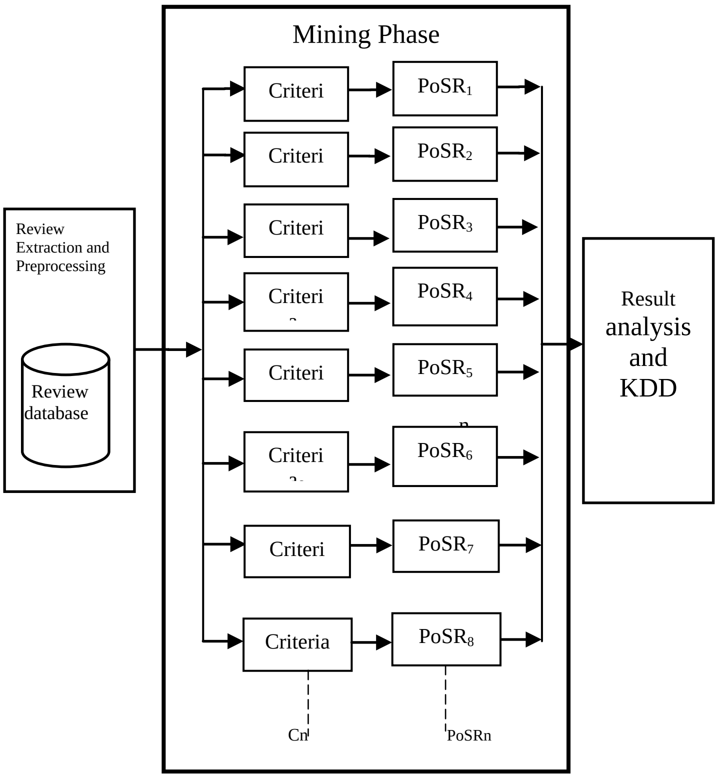

Figure 1: System model of the proposed approach

3.1.1 Identifying and defining multiple criteria

There are many criterias used to support detection of spamicity of reviews. For the proposed work, the eight criterias such as word length score, number of reviews, review rating, Positive Review Length Difference (PRLD), Negative Review Length Difference (NRLD). ALL CAPS, advertisement link, Reviewer names end with three numbers are identified and used.

The specifics of these criteria's are given below:

1. Word Length score: The length of the review is an indication to detect spam. As most spammer's writes long reviews to get reviewer attention. The review word length score can be obtained by dividing number of words in a review by maximum number of words of other reviews given by all reviewers [1].

2. Number of reviews: Number of reviews given by the reviewers varies day to day, thus to have a count of total number of reviews per day/week/month/year is essential [10, 11].Number of reviews are considered by normalizing the reviews per day month wise.

3. 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. Rating is regarded as reviewer's sentiment orientation. In the proposed work, the rating scale of the reviews given by the reviewers from the stores is 1 to 5.[10,11]. Review rating is considered by normalizing the ratings given by reviewers per day, month wise.

4. Positive Review Length Difference (PRLD): The positive text content of the review plays a vital role to identify suspicious / fraudulent reviews (spam reviews). Always long reviews that appear to be genuine is time consuming process, therefore, one expects that spam reviews might be shorter than normal. The positive review length difference is the average difference between the length of its positive reviews and the mean length [3].

5. Negative Review Length Difference (NRLD): The negative text content of the review also plays a vital role to identify suspicious / fraudulent reviews (spam reviews). The negative review length difference is the average difference between the length of its negative reviews and the mean length.

6. Reviews in All Capital Letters (ALL CAPS) : Reviews with all capital letters are considered as spam. As these reviews are written generally to grab attention of the readers or for advertisements. Hence reviews with all capital letters should be ignored[5]. Here even the reviewers uses the "brand approved" version of the name (no normal person would write but a marketer would) like the names of the product in all caps. ALL CAPS reviews are calculated by considering no of All capital words review by total number of reviews in a day.

7. Reviewer names end with three numbers: The names of the common reviewers will be normally in alphabets as the names of the individuals. The names given in alphanumeric way as the names of the reviewers with more than two numbers at the end of the names of the reviewers will be suspected as spam reviews according to the consumerists.com. Reviewer names end with three numbers are calculated by considering number of Reviewer names end with three numbers reviews by total number of reviews in a day.

8. Advertisement link: The content of the reviews always plays a vital role to understand the feelings of the reviewers. In the content of the reviews a normal reviewer will hardly give an advertisement link or few links related to products/stores. The reviewers sometimes leave a link to their website or it is an advertisement. The reviewer in most cases writes the reviews just to leave the link and the review is useless. These reviews are considered as spam reviews. Even, according to consumerists.com reviews with the link/ advertisement link are suspected as spam reviews. Advertisement link are calculated by considering number of Advertisement link reviews by total number of reviews in a day.

To demonstrate the calculated results of the eigth criterias used, few examples, reviews of 1st January 2014, are taken from the store Auto_parts_warehouse.com. There are 31 reviews on this day. Among them, four review examples are taken, as given below:

1. "Orders was filled quickly and the arrived on time. Just one note on the product for the lug nuts when putting on the spare tire the lug nuts would flush up with the spare tire. There was at least 3/8 gap from the lug full extent to the whole on the rim. I know this is companies product side and auto parts warehouse doing just thought i would mention it".

In this review, there are total seventy one words, out of which thirty seven words are stop words (was, and, the, on, one, on, the, for, the, when, on, the, the, would, up, with, the ,there, was, at, least, from, the, full, to, the, whole, on, the, i, this, is, side, and, i, would, it) and the remaining words are,
"Orders filled quickly arrived time Just note product lug nuts putting spare tire lug nuts flush spare tire 3/8 gap lug extent rim know companies product auto parts warehouse doing just thought mention".

Similarly for the reviews,

2. "1/2/14 one item that I ordered still has not shipped".
There are total twelve words, out of which five are stop words (one, that, i, still, has), the remaining words are,
"1/2/14 item ordered not shipped".

3. "THE ORDER PROCESS WAS SIMPLE AND QUICK. THE ORDER COULD EASILY BE TRACKED AND WAS RECEIVED IN A TIMELY MANNER".
A total of twenty words are there, out of which ten are stop words

(THE, WAS, AND, THE, COULD, BE, AND, WAS, IN, A) the remaining words are "ORDER PROCESS SIMPLE QUICK ORDER EASILY TRACKED RECEIVED TIMELY MANNER".

4. "Great price and fast delivery at autoparts_warehouse.com".
A total of seven words, out of which two are stop words (and, at), and the remaining words are, "Great price fast delivery autoparts_warehouse.com".

Word Counter tool is used to count number of words in a review. Stop words are filtered from the review dataset The Table 1, shows structure of eigth criteria for a given review. The Table1 contains eigth criteria values for the four sample review contents. Similarly, eigth criteria values are computed for remaining reviews of three stores review dataset .

Table 1: Sample of calculated results of few reviews of the eight criteria's used from the store Auto parts warehouse.com

R.No	Review Content	Word length score	Number of reviews	Review rating	Positive review length difference (PRLD)	Negative review length difference (NRLD)	ALL CAPS	Reviewer names end with three numbers	Advertise ment link
1	Orders filled quickly arrived time Just note product lug nuts putting spare tire lug nuts flush spare tire 3/8 gap lug extent rim know companies product auto parts warehouse doing just thought mention.	0.81	0.07	1	0.40	0	0	0.32	0
2	1/2/14 item ordered not shipped	0.17	0.01	0.67	0	1	0	0	0
3	QUICK ORDER EASILY TRACKED RECEIVED TIMELY MANNER	0.24	0.02	1	0.02	0	0.03	0	0
4	Great price fast delivery autoparts_warehouse.com".	0.12	0.02	1	0.02	0	0	0	0.03

With respect to the first review content present in the Table 1, the specifics of eight criterias are calculated as follows:

1. a) Word length score value: Word length score is calculated by considering number of words in a review by maximum number of words of all the reviews in a day. The maximum word length score of 31 day reviews of 1st January 2014 is 42. And number of words in the review are 34.

$$\text{Word length score} = \frac{\text{no of words in a review}}{\text{maximum no of words among all the reviews in a day}}$$

= 34/42

= 0.81

b) No of reviews value: As on this day number of reviews are 33. Maximum number of reviews are 363 and minimum no of reviews are 10 for 31 days reviews of a day.

No of Reviews = (no of reviews - minimum no of reviews) / (maximum no of reviews-minimum no of reviews)

By normalization formulae

= (33 – 10) / (363-10)

= 0.07

c) Review Rating value: The rating given by the reviewer is 5. Minimum review rating given by reviewers for this month is 2 and maximum is 5 (as rating is scale is 1 to 5) using normalization formula

Review Rating = (review rating - minimum rating given by reviewer) / (maximum review rating -minimum review rating)

= (5 – 2) / (5-2)

= 1

d) Positive Review Length Difference (PRLD) value : It is the average difference between the length of its positive reviews and the mean length

In this review example, no of words are 34, review length is 201,

Avg length= 34/201*100

= 16.91

And the word 'quickly' being a positive word in this review, its length is 7.

PRLD = (Avg length-poslength) /poslength

= (16.91-7) / 7

= 1.42

In this day review (of 31 reviews) maximum PRLD value is 3.55 and minimum value is 0. After normalizing using normalization formulae

PRLD values = (1.42-0) / (3.55-0)

= 0.40

e) Negative Review Length difference (NRLD) value is zero, as there are no negative words in that review.

f) ALL CAPS value is zero as this review is not with all the words capital.

g) Advertisement link is zero as in this review there is no advertisement link.

h) Reviewer name end with three numbers value: As there are 10 reviewers names end with three numbers on this day review. The reviewer name end with three numbers are calculated by considering number of reviewers name with three numbers by total number of reviews in a day.

Reviewer name end with three numbers =

$$\frac{\text{No of Reviewer names end with three numbers}}{\text{Total number of reviews in a day}}$$

= 10/31
= 0.32

Similarity, the calculation is performed for the remaining reviews example in the Table 1, for the criteria's word length score, number of reviews, review rating, Positive Review Length Difference (PRLD) and Reviewer name end with three numbers.

In the second review content example in the Table1,

2. The Negative Review Length Difference (NRLD) value: It is the average difference between the length of its negative review and the mean length

In this review example, no of words are 7, review length is 31,

Avg length = No of words/ review length*100

= 7/31*100
= 22.58

And the word 'not' being a negative word in this review, its length is 3.

NRLD = (Avg length-Neglength) / Neglength

= (22.58-3) / 3
= 6.53

On January 1st 2014, review (of 31 reviews) maximum NRLD value is 6.53 and minimum value is 1.38

After normalizing using normalization formula
NRLD values = (NRLD value -minimum value) / (maximum value- minimum value)
NRLD = (6.53-1.38) / (6.53-1.38)
= 1

In the third review content example in the Table 1

3. ALL CAPS value: ALL CAPS value is calculated by considering number of All CAPS value by total number of reviews in a day. As there is only one review with ALL CAPS and there are 31 reviews on January 1st 2014,

$$\text{ALL CAPS} = \frac{\text{No of ALL CAPS value}}{\text{Total number of reviews in a day}}$$

= 1/31
= 0.03

In the fourth review content example in the Table1

4. Advertisement link value: Advertisement link value is calculated by considering number of advertisement link value by total number of reviews in a day. As there is only one review with advertisement link and there are 31 reviews on January 1st 2014,

$$\text{Advertisement link} = \frac{\text{No of advertisement link value}}{\text{Total number of reviews in a day}}$$

= 1/31
= 0.03

In the fourth review content example in the Table1

Similarly, the eigth criteria's values are calculated for the remaining reviews of this store for entire duration and for the other two stores namely Dhgate.com and Neweggs.com respectively. Similarly, the eigth criteria's values are calculated for the remaining reviews of this store for entire duration and for the other two stores namely Dhgate.com and Neweggs.com respectively.

3.2 Construction of exponential distribution model

The procedure to construct exponential distribution model is divided into the following steps:

1. Construction of frequency distribution
2. Curve fitting for the frequency distribution
3. Identification of spamicity from the Exponential Curve fitting method

3.2.1 Construction of frequency distribution

The normalized eight criterion values of the reviews are grouped into ten frequency time band values. The corresponding occurrence of each band are noted which are called as frequencies. A sample of the frequency distribution table of one of the criteria named word length score of a store Auto_parts_warehouse is shown in the Figure 2. Similarly, frequency distributions are constructed for the remaining seven criteria of this store and for the remaining two stores namely Dhgate.com and neweggs.com.

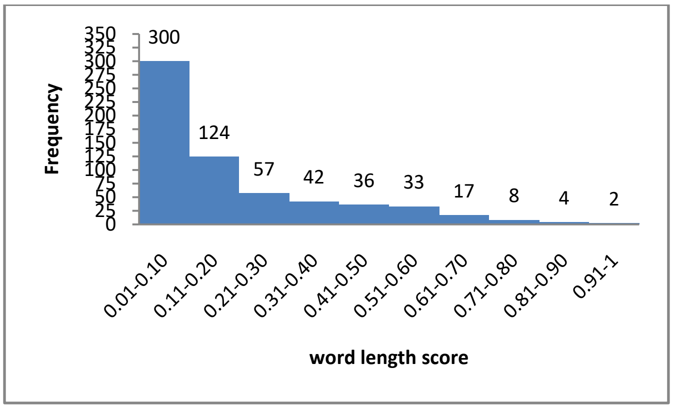

Figure 2. Frequency distribution values for ten bands of the criteria 'Word length score' for the store Auto_parts_warehouse.com

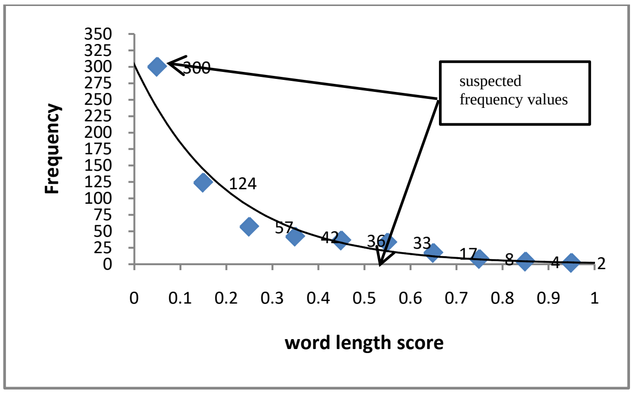

Figure 3. Sample of exponential curve for the frequency values of the criteria word length score of the store Auto_parts_warehouse.com

3.2.2 Construction of Exponential Curve fitting for the frequency distribution

The observation of the plots clearly indicates the frequency decreases as the value of the time band Increases non-linearly, which is an indicative of either a quadratic fit or an exponential fit of the data. If it would have been an quadratic fit, it is likely that the curve increases again for further increase in time axis which is unlikely in this case. Hence it is assumed that the data shall fit to the exponential and hence the exponential fit was fired to fit for the data.

In view of this, in all the cases the exponential curve was attempted by considering the equation -------- (1)

$$y = c + ae^{\alpha x} \quad \text{------- (1)}$$

Where c, a and α are constants. y is frequency values and x is frequency values with respect to time.

3.2.3 Identification of spamicity from the Exponential Curve

From the above constructed frequency distribution table, exponential curves are fitted for each criteria of the review dataset. It is found that the graph which is drawn by considering all the ten frequency values, seems to have steady (balanced) frequency values, to identify the number of data points above the predicted values of the exponential curve. Hence, in the proposed work, to measure review spamicity, the exponential curve is plotted for all the eight criteria's used for the three stores, considering all the ten frequency values (sample of which is shown in the Figure 3).

After fitting of exponential curve for each criterion, the frequency values found above the exponential curve are identified and the reviews found for these frequency values are suspected as spam reviews. In the Figure 3, the number of data points which are above the predicted values of the exponential curve are 300 and 33. Hence, these frequency values reviews are suspected as spam reviews. Spamicity of reviews is measured by considering number of reviews found for the frequency values which are above the exponential curve by total number of reviews of the store for the entire duration.

A sample of the exponential graph to identify spam reviews in the frequency values for the criteria word length score of the store Auto_parts_warehouse.com is shown in the Figure 3. Similarly exponential graphs are plotted to identify spam reviews for the remaining criteria's of this store and for the two stores namely Dhgate.com and Neweggs.com respectively.

3.3 Result analysis and KDD

This stage contains analysis part and knowledge discovery part of the system model. The analysis of exponential model based spamicity measures of review stores and comparisons of spamicity of review stores will be calculated effectively.

4. EXPERIMENTAL RESULTS

This section presents the experimental results to demonstrate the effectiveness of the proposed approach. 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. 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 based on constructing eight criteria's from the extracted reviews of the three stores. The criteria's values are normalized in the range 0-1 and are grouped into ten values (with class interval 0.01-0.10 to 0.91-1.0). The frequency of occurrence of the eight criteria's values are grouped and are given in a frequency distribution table. Further, for these frequency values of ten values, an exponential graph is plotted. The frequency values which are above the exponential curves i.e the number of data points which are above the predicated values of the exponential curve are considered, and these values reviews are suspected as spam reviews. Spamicity of the reviews is measured by considering number of reviews found for the frequency values which are above the exponential curve by total number of reviews for the store for the duration of 623 days.

In the Figure 4, the exponential graph is plotted for the ten frequency values of the criteria word length score for the store Auto_parts_warehouse.com.

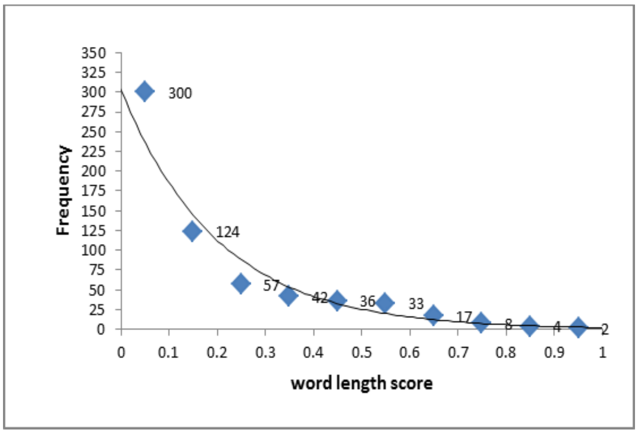

Figure 4. Criteria word length score's ten frequency values

From the Figure 4, the number of frequency values found above the exponential curve are two with the frequency values 300 and 33. The total number of reviews for the frequency values 300 and 33 are 5912 reviews. The total number of reviews for duration of 623 days of the store Auto_parts_warehouse.com are 27,522. The spamicity measure of the criteria word length score is 21.48%.

The exponential graphs are plotted for the criteria's number of reviews, review rating, Positive Review Length Difference (PRLD), Negative Review Length Difference (NRLD), ALL CAPS and advertisement link for ten frequency values in the Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11 respectively And the frequency values found above the exponential curve, the total number of reviews for the frequency values and review spamicity measure are shown in the Table 2 for the store Auto_parts_warehouse.com for a total of 27,522 reviews for the entire duration.

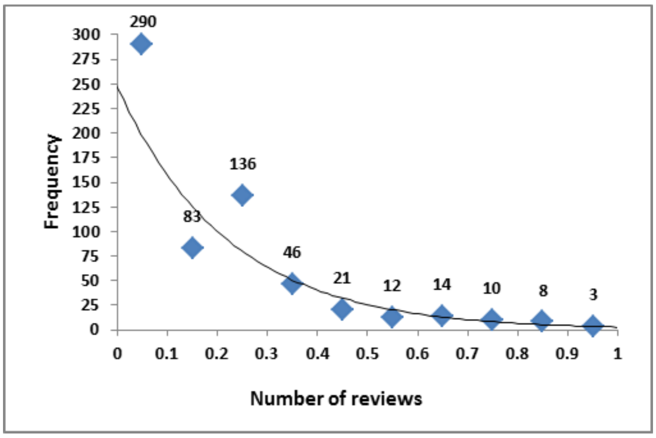

Figure 5 Criteria Number of reviews frequency values for ten values

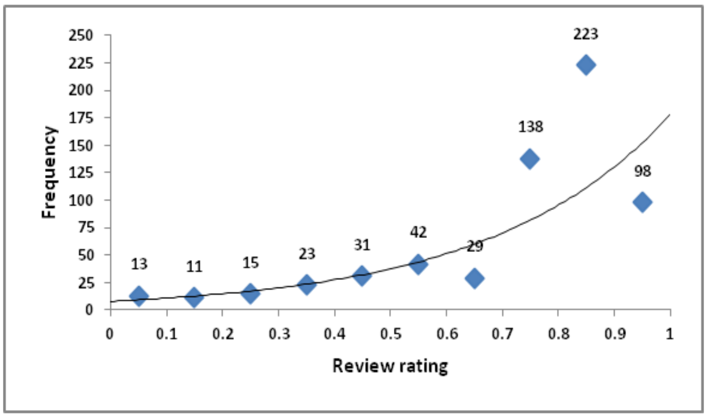

Figure 6. Criteria Review rating frequency values for ten values

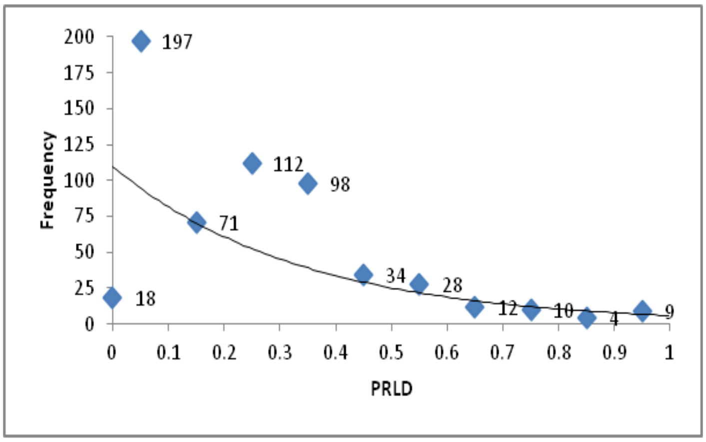

Figure 7. Criteria PRLD frequency values for ten values

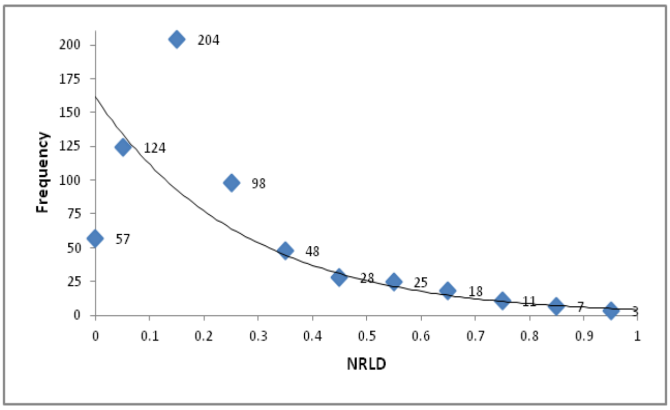

Figure 8 Criteria NRLD frequency values for ten values

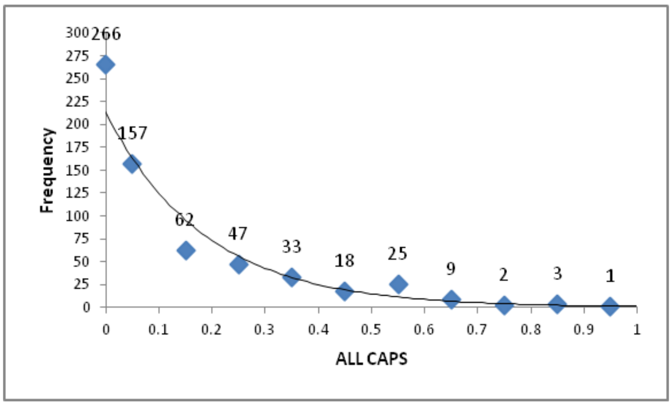

Figure 9. Criteria ALL CAPS frequency values for ten values

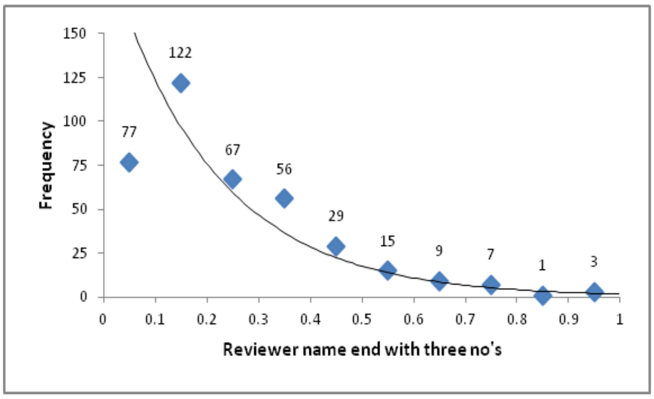

Figure 10. Criteria reviewer name end with three numbers frequency values for ten values

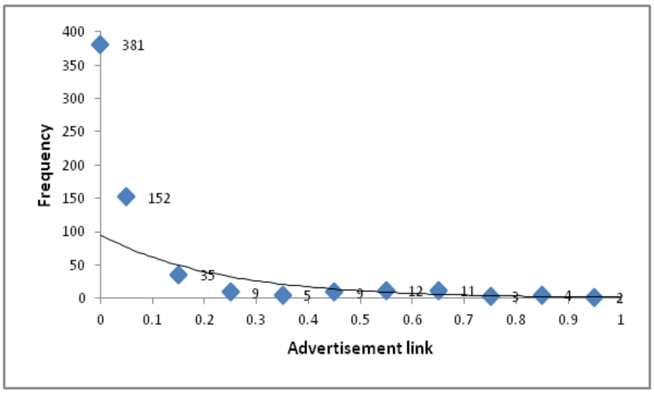

Figure 11. Criteria Advertisement link frequency values for ten values

Table 2. Review Spamicity measure of eight criteria's used, for the store Auto_parts_warehouse.com

Sl No	Names of Criteria	Number of frequency values above the exponential curve	Number of reviews found above the frequency values exponential curve	Review Spamicity measure (in %)
1.	Word length score	2	5912	21.48%
2.	Number of reviews	2	5819	21.14%
3.	Review rating	2	7849	28.52%
4.	PRLD	3	8851	32.16%
5.	NRLD	2	5578	20.27%
6.	ALL CAPS	1	3241	11.78%
7.	Reviewer names end with 3 no's	3	8304	30.17%
8.	Advertisement Link	1	2998	10.89%

From the Table 2, review spamicity measure for the criteria's word length score, Number of reviews, Review Rating, Positive review length difference(PRLD), Negative review length difference(NRLD), Reviewer name with 3 no's are in the range 20 to 32 %. And the criteria's ALL CAPS and Advertisement link are in the range 11 to 12%, which are found comparatively low to other criterias.

Similarly, exponential graphs are plotted for all the eigth criteria's for the remaining two stores namely Dhgate.com and Neweggs.com. The frequency values found above the exponential curve, the total number of reviews for the frequency values and review spamicity measure are shown in the Table 3, for the store Dhgate.com for a total of 12,518 reviews for the entire duration. And the frequency values found above the exponential curve, the total number of reviews for the frequency values and review spamicity measure are shown in the Table 4, for the store Neweggs.com for a total of 3,218 reviews for the entire duration.

Table 3. Review Spamicity measure of eight criteria's used, for the store Dhgate.com

Sl No	Names of Criteria	Number of frequency values above the exponential curve	Number of reviews found above the frequency values exponential curve	Review Spamicity measure (in %)
1.	Word length score	2	2898	23.16 %
2.	Number of reviews	2	2984	23.85%
3.	Review rating	3	3711	29.66%
4.	PRLD	2	3116	24.90%
5.	NRLD	2	2319	18.53%
6.	ALL CAPS	1	1527	12.20%
7.	Reviewer names end with 3 no's	2	2539	20.29%
8.	Advertisement Link	1	1331	10.64%

From the Table 3, review spamicity measure for the criteria word length score, number of reviews, review rating, Positive Review Length Difference (PRLD), Negative Review Length Difference (NRLD), Reviewer name with 3 numbers are in the range 18 to 30%. And the criteria's ALL CAPS and Advertisement Link are in the range 10 to 12%, which are found comparatively low to other criterias.

Table 4. Review Spamicity measure of eight criteria's used, for the store Neweggs.com

Sl No	Names of Criteria	Number of frequency values above the exponential curve	Number of reviews found above the frequency values exponential curve	Review Spamicity measure (in %)
1.	Word length score	3	1073	32.70
2.	Number of reviews	2	899	27.40
3.	Review rating	2	916	27.92
4.	PRLD	2	768	23.41
5.	NRLD	1	412	12.56
6.	ALL CAPS	1	397	12.10
7.	Reviewer names end with 3 no's	2	646	19.69
8.	Advertisement Link	1	366	11.16

From the Table 4, review spamicity measure for the criteria's word length score, Number of reviews, Review Rating, Positive Review Length Difference(PRLD), Reviewer name with 3 no's are in the range 19 to 33%. And the criteria's Negative Review Length Difference (NRLD), ALL CAPS and Advertisement Link are in the range 11 to 13%, which are found comparatively low to other criterias review spamicity measure.

From the experimental observations, the review spamicity measure for the criteria's like word length score, number of reviews, review rating, Positive Review Length Difference (PRLD), Negative Review Length Difference (NRLD), reviewer name with 3 numbers are found consistent in the range 20% to 32%. Hence these criteria's can be considered as sturdy (strong) criterias. And the review spamicity measure for the criterias ALL CAPS, advertisement link are found in the range 10% to 13%, which are comparatively low to the other criterias are considered to be frail (weak) criteria's.

5. CONCLUSION & FUTURE WORK

In this work, a novel evaluation method, exponential distribution model is used to find review spamicity from the eight criteria's. Eigth criteria's identified are namely, word length score, number of reviews, review rating, Positive Review Length Difference (PRLD), Negative Review Length Difference (NRLD), ALL CAPS, advertisement link and reviewer names end with three numbers. The criterion values are normalized and grouped into frequency values. Further, based on the frequency values frequency distribution table is constructed and exponential curves are fitted for each criteria of the review dataset. After fitting of exponential curve for each criterion, the predicated values which are above the exponential curve are identified and these frequency values reviews are suspected as spam reviews. Review spamicity is measured by considering number of spam reviews identified by total number of reviews for the entire duration. Experimental results of detecting review spamicity for the reviews by using review website resellerratings.com for the stores Auto_parts_warehouse.com, Dhgate.com and Neweggs.com for a specific period demonstrates the proposed method is effective in detecting review spamicity based on exponential distribution model. Comparing the spamicity measure used in this work with the ground truth table gives scope for future work.

6. REFERENCES

[1]. 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), Singapore, pp. 350-355.,March 29-30, 2014.

[2]. Fangtao Li, Minlie Huang, Yi Yang and Xiaoyan Zhu "Learning to Identify Review Spam", Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. pp.2488-2493, 2010

[3]. Guangyu Wu, Derek Greene, Pádraig Cunningham " Merging multiple Criteria to find suspicious reviews" ACM 978-1-60558-906-0/10/09, pp.241-244, 2010.

[4]. K. Dave, S. Lawrence & D. Pennock. "Mining the peanut gallery: opinion extraction and semantic classification of product reviews". pp. 451-460, ACM, 2003.

[5]. Kuldeep Sharma, King-Ip Lin " Review Spam Detector with Rating Consistency Check" (ACMSE'13), Savannah, GA, USA , pp.34-39, ACM, April 4-6, 2013.

[6]. Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey Hancock. Finding Deceptive Opinion Spam by Any Stretch of the Imagination.: pp. 309-319 , ACL 2011.

[7]. 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.

[8]. 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.

[9]. Nitin Jindal and Bing Liu., "Review Spam Detection", In proceedings of 16th International conference on world wide web, ACM, pp.1189- 1190, 2007.

[10]. 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.

[11]. Siddu P. Algur, Jyoti G.Biradar, Prashant Bhat "Multidimensional Time series Based Review Spam Detection, International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE). ISSN : 2320-9801. pp.11761-11774., Vol.4,Issue 6, June 2016.

[12]. Siddu P.Algur, Amit P.Patil, P.S Hiremath, S.Shivashankar "Web based customer review spam detection using conceptual level similarity measure and shingling technique" International Journal of Innovative Research in Science and Techniques (IJIRST), pp. 1-9 , January-June 2011

[13]. Sushant Kokate., Bharat Tidke., "Fake Review and Brand Spam Detection using J48" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , pp.3523-3526, 2015

[14]. Wang G, Xie S, Liu B, and S.Yu P. Identify online store review spammers via social review graph. Transactions on Intelligent Systems and Technology, ACM, Vol. 3, No. 4, Article 61, pp.1-21,September 2012.