International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Spam Audit Detection Through Content Analysis

Uzma A Mujawar, Mallikarjun M Math

Abstract: ***Nowadays e-commerce has gained recognition in day to day life; hence network became tremendous source for gathering customer reviews/opinions by marketplace analyzers. The count of user reviews that merchandise receives is increasing at high velocity. Opinions being posted on social media differ significantly in superiority. The client needs to essentially go through all reviews regardless of their superiority and decide whether to purchase or not purchase the manufactured goods. The main difficulty in obtainable study on opinion assessment is that all opinions are considered regardless of the implication of each of them. Hence categorization of opinions depending on implication is an essential job. In this article, attempt is made for opinion assessment depending on its superiority, and help buyer make a proper buying judgment. A web mining technique that is narrative and effective is used to assess the consumer opinion for manufactured goods depending on marked allocation are anticipated. The superiority consideration of consumer reviews are classified as genuine, near duplicate, and duplicate opinion. It is carried out in three steps: (1) Recognize opinion regions to take out opinions. (2) Take out and separate features of reviews by quartile compute and assign weights to the features that belong to each group. (3) Consider the feature weights and group belongingness to assess the reviews. Investigational output demonstrates the usefulness of the proposed method which measures the quality of review and assesses it in view of that. The efficiency of client opinion summarization task is probably improved by recognizing and discarding irrelevant opinions.***

Keywords— review spam detection, feature extraction, feature comparison

I. INTRODUCTION

Internet became an integral part in our daily life. With the favors of web, individuals don't need to leave their home to purchase anything. These days buying items online is common as many do not have opportunity to hold up in queue to purchase. Everything has upsides and downsides, internet buying has own disadvantages. As purchasers can't ask about an item or evaluate before purchasing online, they go through reviews and then choose to purchase something. To improve the administration and items - merchants, retailers and specialist organizations gather client criticism as survey. Positive reviews can bring about remarkable benefit or glory to businesses and individuals. This motivates forces to "Opinion Spamming". Spammers support false reviews to advance items or depreciate administrations. Commonly there are two sorts of spam surveys. The type one comprises of the one that purposefully mislead readers or mechanized supposition mining frameworks which give unworthy positive assessments to objective items so as to advance them and additionally by giving uncalled for or then again not well arranged negative reviews to different items in request to destroy the reputation. The type two comprises of reviews which do not contain feelings on item. Though, reviews which has unconstructive feedback as right image of customer's opinion can't be categorized as fake. Hence, for dependable online audits; it became basic problem to recognize fake opinions.

Revised Manuscript Received on June 14, 2020.
* Correspondence Author
Uzma A Mujawar*, Computer Science and Engineering, KLS Gogte Institute of Technology,Belagavi,India,Email: aliyasimrn786@gmail.com
Dr. Mallikarjun M Math, Computer Science and Engineering, KLS GogteInstituteofTechnology,Belagavi,India.Email: mmmath@git.edu

II. LITERATURE REVIEW

J. K. Rout et al [1] tried best to summarize the overall issues as well as challenges for detection of fake reviews as well as fake reviewers. Finally, a framework has been proposed to deal with fake reviews. G.M.Shahariar et al [2], carried out a research that determined labeled data using supervised learning methods - an insufficiency for online review. This piece of writing focuses on detecting several deceiving content reviews. To gain this they used both labeled and unlabeled data and anticipated machine learning method to know fake opinion recognition. A method to accurately identify the appearance of spam reviews and detect the spam opinions by verifying the uniformity of opinion information amongst multiple review site. To show the accuracy of the detected results evaluative experiments were conducted, and comparison was made between the newly proposed method and previously proposed method. This was proposed by c.yoa et al [3] .Binary artificial colony with Naive Bayes for feature selection was proposed by SP.Rajamohana, Dr.K.Umamaheswari et al [4]. A Spam Review Recognition method was proposed by Patel et al [5], it was based on n-gram techniques along with feature selection and different ways to represent opinions. Gold-standard dataset was used to carry out experiment. According to investigators, NB classifier with bag-of words was the best performer when compared with others. Shashank et al. [6], attempted to detect spam opinions, then extract reviews that are expletive & offensive, that incorporate sentiment analysis. To detect review spam using hybrid approach was proposed by Istiaq et al. [7] first, duplicate reviews were detected, and then created a dataset that is hybrid by using active learning. Finally, to detect spam opinions a supervised approach was used. Wang P. et al. [8],

proposed semantic clustering that added an additional layer to the CNN architecture. Zhao et al. [9], proposed an improved four-layer OpCNN model that considered the Chinese word order problem. In the pooling layer, the k-max pooling method was used instead of the original pooling layer method which optimized the OpCNN model parameters. Tang et al. [10] for response classification used LSTM. This is done over four large datasets. LSTM is better than other classifiers for classifying the opinions as optimistic or pessimistic is shown in performance comparison.

Most investigational work is carried out using traditional methods .Still, attempts are made by researchers to progress the scope of detecting fake review accurately. The purpose of research is to identify fake opinions and apply machine learning methods to improve the fake recognition procedure with considerable output.

III. PROPOSED SYSTEM

The SCRD system model is shown in figure 3.1. The SCRD system works in two stages. First is the training phase during which it will extract the features from the multiple specifications for a single product. And in the second phase it will detect the reviews spam or non-spam. The input to system is URL of web pages containing reviews of product and output is the classified and summarized reviews.

The various components included in the proposed system model are as follows:

1. Multiple product specification extraction.
2. Distance calculation.
3. Specification clustering
4. Unit dictionary.
5. Customer review extraction.
6. Stop word removal.
7. Sentence splitter.
8. Bi-gram generator.
9. Generation of specification tree.
10. Feature extraction using specification tree.
11. Feature comparison.

Given the input, the system first extracts all the multiple specifications from the given web page for a particular product. Then clustering algorithms are run on the specifications and generate a specification tree. And then this specification tree is used to extract features from product reviews. And next it will extract customer review from the given web page, and extract the features using the specification tree, and stored in a unique list in database. Extract the two lists from the database and compare it, and calculate the quantity of spam or non-spam

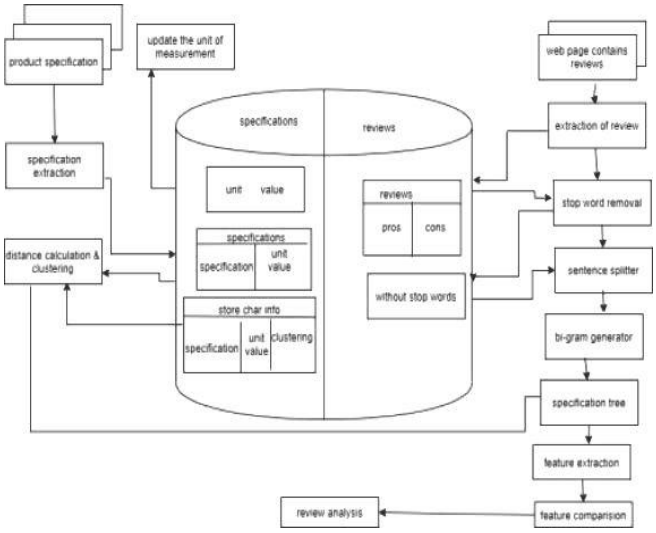

Fig 3.1: System model

A. Product Functions

This section deals with all the algorithms and procedures used for the implementation of the above mentioned **Spam customer review detection.**

Project analysis is done in the following steps. Using the following steps the spam review detection technique system is implemented and the reviews are classified based on the procedure and techniques used.

1) Multiple product Specification extraction

The system first retrieves the multiple product specification from the web page as shown in Fig 3.2, and stores into database. Every specification is preceded by a standard template <table id=product spec <table id="product spec"<td class="rkb" <td class="rkr" </table>.Store this template into a file. Search all html tags from the source code and store into Tag array. If the tag is equal to the template then extract specifications and store it into the respective table. And from each specifications stop words are removed, and stored under a database.

The algorithm used here is as follows:

Specification Extraction Algorithm

This procedure is used to extract all the specifications contained in a web page and store it in the raw review database.

Step 1: extract(source, array)
Step 2: if tag=template
Step 3: search html tags and store in sarray
Step 4: if tarray() equal tag, then Extract specifications
Step 5 :Store specifications in database
Step 7: end if
Step 8:loop

Fig 3.2: Specification extractor

2) Distance calculation

By using the "Levesnshtein algorithm" it will calculate the distance between the two features from the product specification, and next it generate distance matrix as shown in Fig 3.3.Using the random function 'four' numbers are generated and these numbers are unique values, using these values centroids are formed in the distance matrix. And these centroids must lie between the minimum and maximum value in the matrix, where the minimum value is to next 'Zero' value, and maximum value is to biggest value in the matrix.

Distance Matrix

	00	01	02	03	04	05	06	07	08	09	10	11	12	13
00	00	08	11	11	09	13	11	11	10	11	18	13	09	10
01	08	00	12	11	10	15	12	12	06	04	20	14	10	11
02	11	12	00	11	10	13	12	12	10	10	15	11	11	11
03	11	11	11	00	11	11	10	10	10	10	16	12	06	10
04	09	10	10	11	00	14	10	10	08	09	17	12	10	10
05	13	15	13	11	14	00	13	13	14	12	17	14	12	13
06	11	12	12	10	10	13	00	04	11	12	15	12	10	11
07	11	12	12	10	10	13	04	00	10	12	17	12	11	12
08	10	06	10	10	08	14	12	10	00	05	18	14	10	11
09	11	04	10	10	09	12	11	12	05	00	16	10	09	10
10	18	20	15	16	17	17	15	17	18	16	00	15	18	16
11	13	14	11	12	12	14	12	12	14	10	15	00	13	13
12	09	10	11	08	10	12	10	11	10	09	18	13	00	04
13	10	11	11	10	10	13	11	12	11	10	16	13	04	00
14	13	16	15	14	14	15	15	16	17	16	18	13	10	06
15	12	12	12	12	11	13	10	12	10	09	15	12	08	06
16	13	13	14	11	11	14	13	15	11	12	16	13	12	11
17	16	17	14	10	14	13	16	14	16	14	17	14	11	14
18	15	17	13	10	07	15	14	14	15	15	14	14	16	15
19	15	17	14	15	07	15	15	13	15	14	15	13	16	15
20	13	16	12	14	12	14	12	12	14	14	19	15	13	12
21	12	14	08	11	12	12	12	11	13	12	16	13	11	07
22	12	13	04	10	13	12	12	11	12	11	17	13	11	12
23	13	13	02	10	12	14	11	15	11	15	16	13	11	09
24	15	16	08	13	13	14	15	10	11	07	16	13	13	13

Fig 3.3: Distance Matrix

Levenshtein Distance Algorithm

This procedure findS the distance between the two strings.
Step 1: Calculate Distance between two strings(String s1,String s2)
Step 2: Initialize the two strings
Step 3: if s1(equals) to zero
return s1;
exit;
Step 4: end if
Step 5: if s2(equals)to zero
return s2;
exit;
Step 6: end if
Step 7: construct the matrix containing of 0--m rows and 0---n columns
Examine the each character from the two strings
if s1 [i] equals s2[j] the cost is 0
if s1 [i] does not equals to s2[j] the cost is 1
Set the cell d[i,j] of the matrix equal to minimum of
a)The cell immediately above plus 1:d[i-1,j]+1
b)The cell immediately to the left plus 1:d[i,j-1]+1
c)The cell diagonally above and to the left plus the cost :d[i-1,j+1]+cost.
Step 8: Iterate step 7 and complete matrix, the distance is found in cell d[n,m].
Step 9 : Loop

3) Centroid generation

From this module using the distance matrix four(4) centroid will be generated, by using the random function.

Centroid generation Algorithm

Step 1: Using random(min,max) function four centroids will be generated and stored in array.
Step 2: Check for duplicate centroids, if (found)
Step 3: regenerate four centroids.
Step 4: End if

4) Specification Clustering

In this module using the threshold value and centroid value specification clusters are formed as shown in Fig 3.4. From the distance matrix we find the min, max and centroid value, using this threshold can be calculated as follows.

Threshold value= (max distance/no of specifications)*k.

Where k is multiplication factor.

Using this threshold value clusters the features into four groups.

Specification clustering Algorithm

Step 1: Calculate the threshold value T= (max distance/no of specifications)*k, Where k is multiplication factor. k = (10*i); where i lie between 2 to 20;

Step 2: form clusters using, Cluster (threshold value, centroid value)

Step 3: Calculate distance from the centroid with other specification and compare it with threshold value, if the value lies between centroid and threshold value then put cluster 1.

Step 4: Repeat this process for other clusters.

Step 5: Loop.

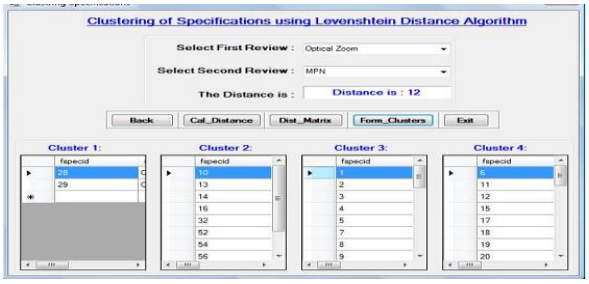

Fig 3.4:Specification clustering

5) Unit dictionary

In this function, it will add the units for all the features in the specification tree, and all the units are stored in unit dictionary.

6) Customer review extraction

Customer reviews are input to this function. The reviews extracted from the web pages are stored under database.

Source code of web pages containing reviews is input to review extractor. Every Pros is preceded by a standard template . Store this template into a file. Search all html tags from the source code and store into Tag array. If the tag is equal to the template then extract the pros and cons and store it into the respective table. Fig 5 shows the table where pros and cons extracted from reviews are stored in table.

Customer Review Extraction Algorithm

This procedure is used to extract all the reviews contained in a web page and store it in the raw review Database.

Step 1: extract (source, tarray)
Step 2: if tag=template
Step 3: search html tags and store in sarray
Step 4: if tarray () equal tag
Step 5: extract pros and cons
Step 6: store pros and cons in database
Step 7: end if
Step 8: loop

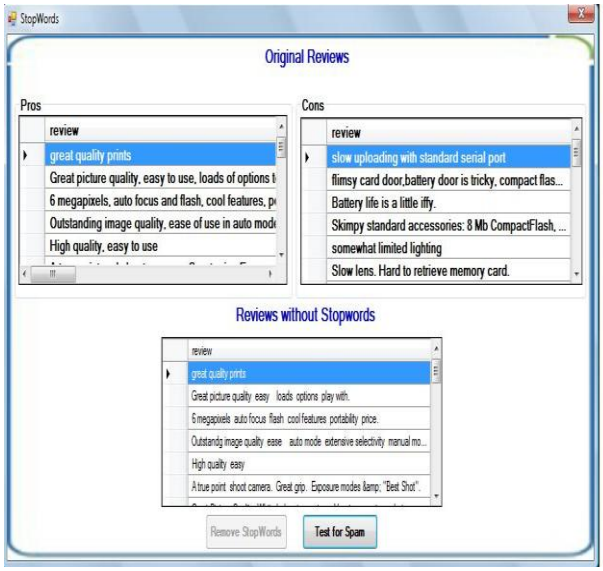

Fig 3.5: pros and cons stored in table

7) Stop word removal

Customer reviews are the input to this function, stop words are removed from this reviews and stored in database. Fig 3.6 represents the reviews without stop words .The stop words are removed from both specifications extracted and also from the reviews extracted.

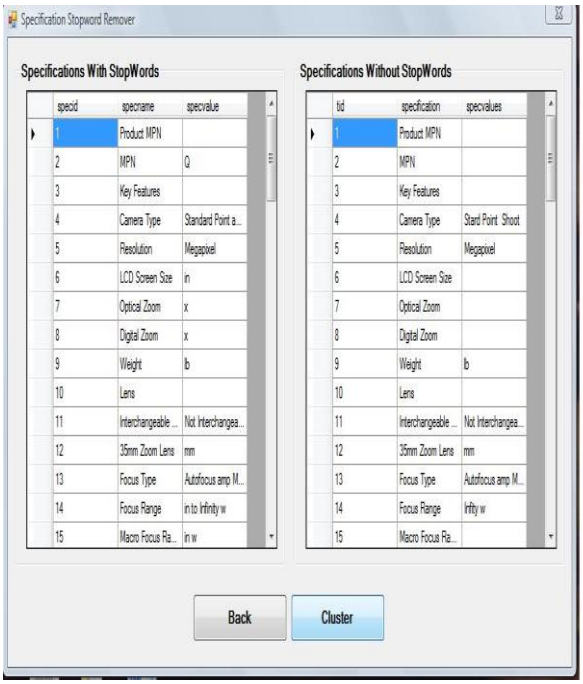

Fig 3.6 : Specifications without stop word

8) Sentence splitter

The stop word removed sentences are input to this function. And in this function sentence are split into words. For example: Easy pictures Compact Flash memory cards move pictures

Easy
Pictures
Compact
Flash

Memory
Cards. Etc

9) Bi-gram generator

In this function bi-gram are generated at word level, it is the combination of two word. The bi-grams generated are used further for generating specification tress. These specifications tree is further used for future comparison.

10) Generation of specification tree

From the distance matrix the four groups of clusters are formed. Using these clusters the specification tree will be generated, and each node in the specification tree represents a featres and these features are extracted from the database using the specification ID and reperesent in tree form.

Generation of specification tree Algorithm

In this module specifications are stored in tree form.

Step 1: Read the number of clusters into integer variable

Step 2: for (i=1 to number of clusters)

Step 3: read the centroid of clusters 'i' then make it as root of the tree.

Step 4: read the reaming member of the cluster, and find their distance, if distance lies between the centroid value and threshold value then add it to cluster1

Step 5: repeat the above steps for remaining three clusters.

Step 6: end if

Step 7: exit for

11) Feature extraction using specification tree

Features are extracted from the bi-gram list using the specification tree, and stored in a separate list. The lists generated in this step are used for feature comparison. Feature extraction is done using the specification tree which is generated using bi-grams.

Feature extraction Algorithm

Features are extracted from the Bi-gram list using a specification tree with preorder tree traversal method, and features are stored in a unique list.

Step 1: read the bi-gram list, then

Step 2: calculate the length of the list

Step 3: for (i=1 to length-1), then

Step 4: traversal the tree in preorder mechanism

Step 5: if root (equals) list of 'i', add to feature list.

Step 6: exit for

Step 7: end if

12) Feature comparison

In this function compare the two lists, which contain the features and calculating the percentage of the spam. If percentage of spam is hundred percent (100%) then, they are duplicate reviews, if the percentage of the spam is below hundred percent (100%) value and above the threshold value then they are near-duplicate reviews

IV. RESULT AND DISCUSSION

The quality assessment of a purchaser reviews are categorized as genuine, near duplicate, and duplicate opinions. The data from epinions.com is taken and the reviews are been assessed for spam reviews. The following figures represent the results of reviews obtained using the epinions.com dataset.

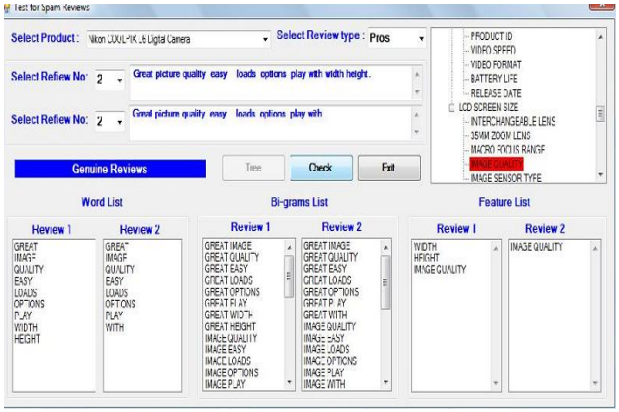

Fig 4.1: Result showing genuine reviews

The fig 4.1 represents the reviews that are genuine and hence helps the merchant sites to use thi review for their product manufacturing and economical profits.

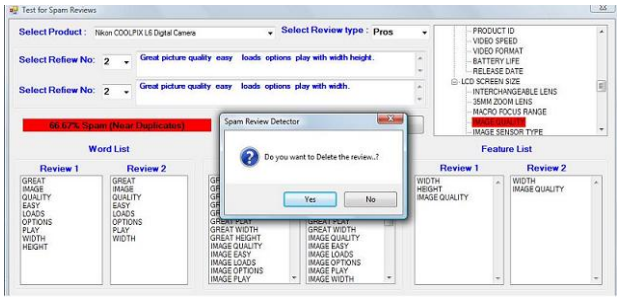

Fig 4.2: result showing near duplicate reviews

Fig 4.2 shows the reviews which are near duplicate whose percentage of spam is less than100% and above the threshold value. Here the product manufacturers can keep this review or can discard the review according to their convenient.

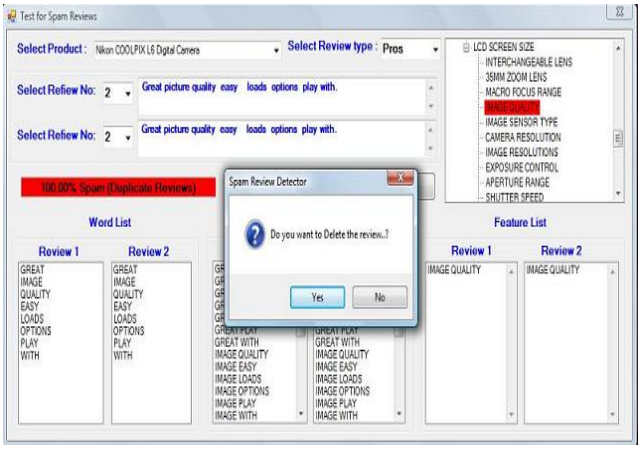

Fig 4.3: results showing duplicate reviews

Fig 4.3 shows the reviews that are duplicate, here the percentage of spam is 100% so these reviews can be directly discarded .And hence the owners of merchant site can protect their products from these kind of spam reviews.

Table 4.1: Table Showing Results of Different Reviews

SI.NO	Type of Review	Results
1	Genuine Reviews	Percentage of spam will be below threshold value
2	Duplicate Reviews	Percentage of spam will be 100%
3	Near Duplicate Reviews	Percentage of spam will be below 100% and above threshold value

V. CONCLUSION

In this paper, a web mining technique that is narrative and effective is used to assess the consumer opinion for manufactured goods depending on marked allocation is anticipated. The superiority assessment of a consumer reviews are categorized as genuine, near duplicate, and duplicate opinion. It is carried out in 3 steps: (1) Recognize opinion regions to take out opinions. (2) Take out and separate the features of reviews by quartile compute and assign weights to the features that belong to each group. (3) Consider the feature weights and group belongingness to assess the reviews. If percentage of spam is hundred percent (100%) then, they are duplicate reviews, otherwise if the percentage of the spam is below hundred percent (100%) value and above the threshold value then they are near-duplicate reviews. The investigational output concludes that there are considerable numbers of features that belong to irrelevant set. The opinions holding these characters can be unseen while opinion summarization, by optimizing the method of superiority evaluation.

ACKNOWLEDGMENT

The writers are thankful to consumers for their supportive remarks that enhanced the nature of article impressively.

REFERENCES

1. J. K. Rout, A. K. Dash and N. K. Ray, "A Framework for Fake Review Detection: Issues and Challenges," *2018 International Conference on Information Technology (ICIT)*, Bhubaneswar, India, 2018, pp.7-10.
2. G. M. Shahariar ; Swapnil Biswas ; Faiza Omar ; Faisal Muhammad Shah ; Samiha Binte Hassan, " Spam Review Detection Using Deep Learning," *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*
3. C. Yao, J. Wang and E. Kodama, "A Spam Review Detection Method by Verifying Consistency among Multiple Review Sites," *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/Smart City/DSS)*, Zhangjiajie, China, 2019, pp. 2825-2830.
4. SP.Rajamohana, Dr.K.Umamaheswari, " Feature selection using binary artificial bee colony for sentiment classification", *International Research Journal of Engineering and Technology, 2016.*
5. Patel, Rinki, Priyank Thakkar. "Opinion Spam Detection Using Feature Selection." *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on. IEEE, 2014*
6. Shashank Kumar Chauhan, ,Mahendra K Gurve and Anupam Goel. "Research on product review analysis and spam review detection". *In 4th International Conference on Signal Processing and Integrated Networks (SPIN), 2017*
7. M.N. Istiaq ,Abdullah All Kafi, , and Faisal Muhammad Shah." An ensemble approach to detect review spam using hybrid machine learning technique". *19th International Conference on Computer and Information Technology, Dhaka, December 2016*
8. Wang P., Xu J., Xu B., Liu C., Zhang H., Wang F., and Hao H. Semantic clustering and convolutional neural network for short text categorization. Proceedings of ACL, pages 352357, 2015
9. S. Zhao, Z. Xu, L. Liu, and M. Guo. Towards accurate deceptive opinion spam detection based on word order-preserving CNN. AvailableOnline:http://arxiv.org/abs/1711.09181, 2017
10. Duyu Tang, Bing Qin, and Ting Liu, Document modeling with gated recurrent neural network for sentiment classification. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, , Lisbon, Portugal, September, 2015.

AUTHORS PROFILE

Uzma A Mujawar, PG Scholar, persuing Mtech in KLS Gogte Institute of Technology, Belagavi..

Dr. Mallikarjun M Math, B.E (CSE), M.S (CSE), PhD, Professor, KLS Gogte Institute of Technology, Belagavi, published 10 papers in international journals, published 2 papers in international conference, has membership in LMISTE, MCSI.