International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

Separation of Genuine and Fake Users From e-Commerce Site Using Hybrid Learning Approach

Akanksha Singh[1], Prof. Dr. Abha Choubey[2], Prof. Dr.Siddharth Choubey[3]
[1, 2, 3]*Shri Shankaracharya group of Institution, Dept. of Computer Science and Engineering, Bhilai, Chhattisgarh, India*

Abstract: ***A mind blowing source of collecting the reviews on specific item is distinctive web electronic shopping destination where people share their reviews on items and their shopping learning. People may get past the wrong suppositions known as review spam. In this way, for this, it is central to recognize it by a couple of means. In this paper, presents methodologies for acknowledgment of spam user's account using feature extraction and discretization, in blend with Maximum Likelihood estimation. Our structure can perceive different spammers by knowing simply little course of action of spammer sets. Proposed strategy sufficiently picks vital features and produce features set to recognize the spammers.***
Keywords: Review spam, un-truthful reviews, openion spam, rating spam. Maximum Likelihood.

I. INTRODUCTION

Data is the new oil", in the age of big data, data becomes the invisible instrument for almost every business. Through data mining and intelligent decision based on data, enterprises can foresee the potential request of customers so as to complete focused on-line marketing with the end goal of benefit growth. Because of some reasons, such as commercial competition and a time limit of the customer, a lot of product's spam reviews, in any case, have showed up in the product reviews, resulting in the ineffectiveness and the mistake of the intelligent decision making that was based on data analysis [1].

Table I shows some selected mobile phone reviews from the Amazon website. For the mobile phone product's topic, reviews 1 and 2 are applicable to the topic, and review 1 has the highest pertinence than different reviews. Be that as it may, it is difficult to choose the importance between review 3 and the topic. Besides, reviews 4 and 5 are a piece of plagiarism of review 3, and review 6 is an advertisement. Only two of six reviews are significant to the mobile phone product's topic. Fake reviews can increase decision's making cost, as well as influence decision accuracy making.

To identify spam reviews, some scholars have done some related research by using the techniques of data mining and natural language processing [2-4]. What's more, several big data processing techniques, such as data cleaning [5, 6], data repair [7], and database query processing system [8], were also presented to manage raw data. Existing big data processing techniques, be that as it may, can't viably solve the spam reviews' concern. In the first place, each day a great deal of new reviews are composed on the corporate site by the reviewers. Data repair and successive data cleaning will prompt a surge in enterprise activity costs. Moreover, since we can't decide all reviews' credibility, it is unsuitable to embrace the database query that processes method to channel those spam reviews.

TABLE I. Some Selected Mobile Phone Reviews on MovieLens.Com.

S. No	Review
1	I feel so LUCKY to have found this used (phone to us & not used hard at all), phone on line from someone who upgraded and sold this one.
2	It came with Arabian keyboard :(
3	excellent product in perfect condition
4	excellent
5	excellent
6	Connecting People.

As of late, a few scholars suggested perceiving fake reviews based on the spammers' behavior features [9-11]. This technique can successfully solve the issue in which a reviewer writes a considerable measure of reviews in a short time [12]. The study of [9], in any case, has shown that 68% of the reviewers on Amazon just thought of one review [7], and the research of [12] has also shown that the singleton reviews' percentage is approximately 90%. The strategy for burst reviews in [10] does not work to identify spam reviews among singleton reviews. For singleton review spam assault discovery, [12] used the fleeting bend fitting and Longest Common Sub-sequence (LCS) algorithm to decide if a review is fake. While the proposed approach can viably recognize a segment of spam reviews, it still needs to be enhanced in several ways. At first, the customary LCS algorithm's opportunity effectiveness is low because it needs to play out countless processes. Besides, the conventional LCS algorithm does not recognize review 6 of Table 1 as a spam review.

So as to detect spam reviews all the more precisely, the paper first summarizes the spammers' some normal behavior properties and after that constructs a set of identification indicators of detecting spam reviews. When all is said in done, these spammers have some basic behaviors. For instance, some of them like to duplicate different reviews of the same product, at that point make a few of improvements or don't change, presenting as their own particular reviews. Another sort of fake reviewers is used to compose the product's some malicious reviews with the end goal of unfair commercial competition [13]. At that point, two algorithms are designed to distinguish similar reviews and applicable reviews, respectively. The proposed algorithm for identification of similar reviews improves the conventional LCS algorithm. The proposed algorithm for identification of pertinent reviews can acquire the pertinence consequently between a review and the topic of the product.

We have also led an investigation on product reviews crept from the Amazon website. The test results showed that our algorithms for detecting similar reviews not just have less execution time than the customary algorithm yet additionally recognize more spam reviews. Besides, as per our analysis results, we found that over half of these reviews are spam reviews. Thus, it is imperative to perceive and wipe out those spam reviews before enterprises settle on a business decision based on review analysis' result.

II. LITERATURE SURVEY

Nitin Jindal [14], focused on review spam and spam detection. Three main types of spam were identified. Detection of such spam is done first by detecting duplicate reviews. We then detected type 2 and type 3 spam reviews by using supervised learning with manually labeled training examples. Results showed that the logistic regression model is highly effective. However, to detect type 1 spam reviews, the story is quite different because it is very hard to manually label training examples for type 1 spam. We presented an approach to use three kinds of duplicates, which are very likely to be spam, as positive training examples to build a classification model. The results are promising.

Siddu P. Algur [15], focused on a novel and effective technique for detecting the trustworthiness of customer reviews for a particular product based on the features of the product being commented by the reviewers. Spam reviews are been categorized as duplicate and near duplicate reviews and non-spam reviews as partially related and unique reviews. Results demonstrate the effectiveness of the proposed technique in detecting spam and non-spam reviews. The efficiency of the task of web based customer review spam detection can be enhanced by identifying and eliminating duplicate and near duplicate spam reviews, thereby providing a summary of the trusted reviews for customers to make buying decisions.

C.L. Lai [16], focused on the development of a novel computational methodology to combat online review spam. Our experimental results confirm that the KL divergence and the probabilistic language modeling based computational model is effective for the detection of untruthful reviews. Empowered by the proposed computational methods, our empirical study found that around 2% of the consumer reviews posted to a large e-Commerce site is spam.

RAYMOND Y. K. LAU [17], focused on the proposed models outperform other well-known baseline models in detecting fake reviews. To the best of our knowledge, the work discussed in this article represents the first successful attempt to apply text mining methods and semantic language models to the detection of fake consumer reviews. A managerial implication of our research is that firms can apply our design artifacts to monitor online consumer reviews to develop effective marketing or product design strategies based on genuine consumer feedback posted to the Internet.

III. METHODOLOGY

In this section, we introduce the proposed framework for classification of spammers and non-spammers. Some of the basic notation are presented below:

FS → Set of features from the dataset

N → set of users

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

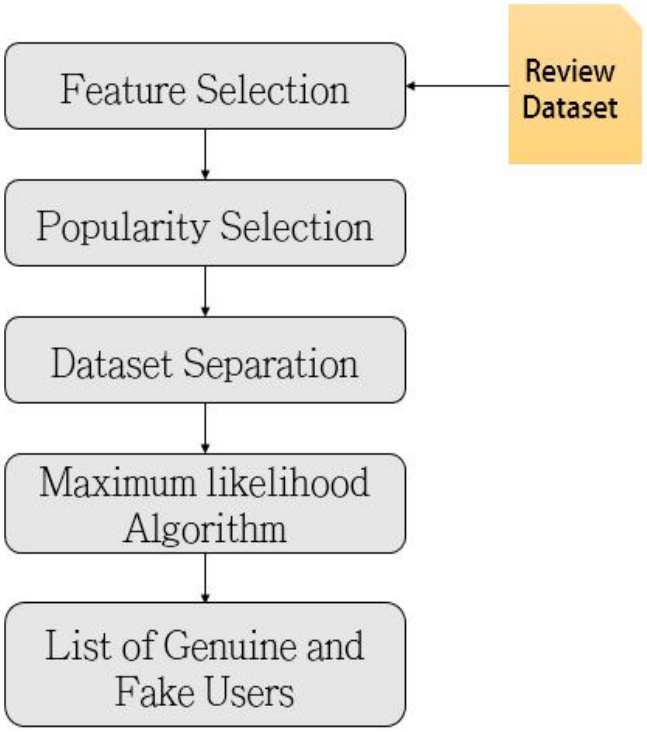

Fig. 2. Show the propsoed system acrhitecture

A. Feature Selection

In our proposed framework, the dataset which we have consider is amazon audit dataset. Amazon audit include dataset contains just numeric dataset. Demonstrating of the considerable number of highlights specifically isn't suggested for spammer location. Subsequently discretization is required. We discretize highlight f into set of v classifications. The classifications are:

1) DegSim
2) LengthVar
3) RDMA
4) FMTD
5) GFMV
6) TMF

These all are the features through which our algorithm runs and produce output.

B. Popularity Selection

We are interested in particular set of users which provides reviews not all. The selection of users which are very active and keeps on reviewing components are selected using popularity selection method.

C. Dataset Separation

The dataset is separated in two groups.

1) Genuine Users
2) Fake Users

Initially, we have selected genuine and fake users based on some manual approach. The approach is quite easy. Some people selects very small amount of data from the dataset by manually analyzing the dataset and there account's.

D. Maximum Likelihood Algorithm

In measurements, maximum likelihood estimation (MLE) is a technique for assessing the parameters of a factual model, given perceptions. MLE endeavors to discover the parameter values that amplify the probability work, given the perceptions. The resulting result is known as a maximum likelihood estimate, which is additionally shortened as MLE.

IV. RESULTS

We have performed explore by taking Amazon survey dataset. It is of numerical kind as it were. Preview of dataset is exhibited in fig.3. The trial is led to foresee the measure of spam and non-spam clients display in a specific site which surges spam messages while checking on product.


```
{
    "reviewerID": "196",
    "asin": "242",
    "overall": 3,
    "reviewTime": "881250949"
}
{
    "reviewerID": "186",
    "asin": "302",
    "overall": 3,
    "reviewTime": "891717742"
```


Fig. 3. Shows the snapshot of Amazon Review Dataset

The fig.4. Presents the yield of spammers with their ID which are anticipated to be a spammer. Our calculation productively takes the arrangement of unlabeled clients list and their reviews and order them in the spammer and non-spammer class. Fig. 5. Demonstrates the quantity of spam and non-spam clients from the arrangement of unlabeled set U.

```
User with ID - 1 is Blocked
User with ID - 6 is Blocked
User with ID - 7 is Blocked
User with ID - 10 is Blocked
User with ID - 11 is Blocked
User with ID - 13 is Blocked
User with ID - 18 is Blocked
User with ID - 40 is Blocked
User with ID - 41 is Blocked
User with ID - 54 is Blocked
User with ID - 56 is Blocked
User with ID - 57 is Blocked
User with ID - 58 is Blocked
User with ID - 60 is Blocked
User with ID - 74 is Blocked
User with ID - 77 is Blocked
User with ID - 80 is Blocked
User with ID - 81 is Blocked
User with ID - 83 is Blocked
```

Fig. 4. Shows the snapshot of blocked users

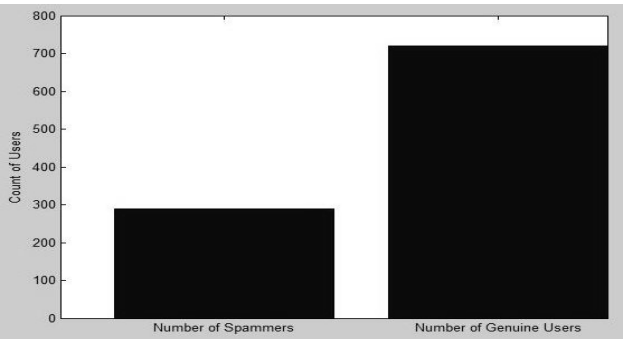

Fig. 5. Shows the snapshot of number of spammer and non-spammers from the U set

V. CONCLUSION

This paper basically give the response for the issue of singleton review spam area, which is both trying and essential to understand. This paper presents feature extraction and decision system through which the apropos features are picked used for gathering of spam reviews. The examination is coordinated on the MATLAB programming. The Maximum Likelihood figuring is considered and evaluated the execution of our framework. Our structure close 300 spam customers out of 1000+ customers. The rest 700 customers are recognized as genuine customer's shows in fig.5.

REFERENCES

[1] J. Malbon, "Taking fake online consumer reviews seriously," Journal of Consumer Policy, vol. 36, pp. 139-157, 2013.

[2] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519-528.

[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168-177.

[4] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Natural language processing and text mining, ed: Springer, 2007, pp. 9-28.

[5] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., vol. 23, pp. 3-13, 2000.

[6] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of “big data” on cloud computing: Review and open research issues," Information Systems, vol. 47, pp. 98-115, 2015.

[7] W. Fan, S. Ma, N. Tang, and W. Yu, "Interaction between record matching and data repairing," Journal of Data and Information Quality (JDIQ), vol. 4, p. 16, 2014.

[8] S. Razniewski and W. Nutt, "Completeness of queries over incomplete databases," Proc. VLDB Endow, vol. 4, pp. 749-760, 2011.

[9] N. Jindal and B. Liu, "Opinion spam and analysis," in Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008, pp. 219-230.

[10] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in Proceedings of the 21st international conference on World Wide Web, 2012, pp. 191-200.

[11] M. Jiang, P. Cui, and C. Faloutsos, "Suspicious behavior detection: Current trends and future directions," IEEE Intelligent Systems, vol. 31, pp. 31-39, 2016.

[12] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 823-831.

[13] A. Heydari, M. ali Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," Expert Systems with Applications, vol. 42, pp. 3634-3642, 2015.

[14] Nitin Jindal and Bing Liu, “Analyzing and Detecting Review Spam”, Seventh IEEE International Conference on Data Mining 2007.

[15] Siddu P. Algur, Amit P.Patil, P.S Hiremath, S. Shivashankar, “Conceptual level Similarity Measure based Review Spam Detection”, 2010 IEEE.

[16] C.L. Lai, K.Q. Xu, Raymond Y.K. Lau, Y. li, L. Jing “Toward A Language Modeling Approach for Consumer Review Spam Detection”, International Conference on E-Business Engineering 2010.

[17] RAYMOND Y. K. LAU, S. Y. LIAO, RON CHIWAI KWOK, KAIQUAN XU, YUNQING XIA, YUEFENG LI, “Text Mining and Probabilistic Language Modeling for Online Review Spam Detection”, ACM Trans. Manag. Inform. Syst. 2, 4, Article 25, December 2011