An Effective Spam Content Detection in Web Pages Using H-BSSD Framework

Malathi. A
Assistant Professor, PG & Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamil Nadu, India

Amutha. B
M.Phil, Research Scholar, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamil Nadu, India

Abstract – In recent years, day-by-day increasing the network spam in online social media. World's each and every person has used social media for sharing messages, exchanging the files, advertising and posting the messages, official groups, chatting, social media courses, etc. The trackers or spammers to target this social media and blend our own messages called spam. The spam is a thing of unsolicited messages (unwanted messages, Advertisements). The spam can create spam reviews. So in this erroneous spam can be avoided by using Hybrid Buying and Sequence Spam Detection (HBSSD), Sequential Search Algorithm, Semantic Similarity Algorithm, GS Ranking Algorithm and Content Similarity Checking Algorithm. In this proposed concept is to detect the group spammers and provides better result by means of simple and effective methods. It analyzed the trusted reviews. The main goal of this proposed system confers a user-friendly interface for Net-Spam detection and increasing the performance of detection via high D dataset. In the proposed concept is very helpful for creating E-Com site with spam classification and spam content classification.

Index Terms – Network Spam Detection, Fake Spammer, Spam Review, OSNs.

1. INTRODUCTION

Online social media are interact to the people and the computer technologies provide facilities to the creativity and share the information, ideas, future information and other communication to share via networks. The lot of standalone and built in social media information currently available to introduce the common features. The people can share the opinion about the products or merchants in online communication, social media, forums, or directly post the comments or reviews in a various benefit systems provide by particular online businesses or maga-retailers like a Ebay or amazon or third parties like reseller rating.com, google+ Local, Bizrate etc. In the people can witten reviews using decision making system. And they send reviews for their motive like positive or negative and encourage or discourage them to selection of their service or product. And this feedback can be used to the providers to increase their product quality and to make better result and services. The online reviews can be useful for the providers and buyers. Lot of peoples can look at the reviews and then they can place the order. They can take a decision based on the online reviews it must be made cautiously. More business leaders must give incentives to who can write best review about their company merchandise or might pay who are give negative comments on the competitor's service and products. These fake reviews are examining the review spam and this can have a better impact in the online marketplace due to the significance of reviews.

The spam details gives negative review means that can be affected by the businesses due to the customers trust. This issue is serving enough to have attracted the attention of mainstream media and governments. The fake reviews are becoming a central problem on the websites. A review spam is spreading widely throughout an area and causing physical damage problems, increasing methods to help business and customers distinguish best reviews from fake ones is an important, but facing problems.

The review spam can be categories into three parts such as:

- Untruthful reviews
- Reviews on brands
- Non reviews

Untruthful reviews:

The untruthful reviews are the providers or owners gives money to get reviews. This review is mostly contain they undermine the integrity of the online review system.

Reviews on brands:

Where the contents are only based with product, things or services of the brand or the seller product and fail to review the product.

Non –Reviews:

This review is not related to their products or customers cannot send feedback of their product or services.

The review spam detection is a difficult task as it is challenge, if it not possible identifies the fake reviews and real reviews by manually watching them. As the customers can confidently identify the review is fake and which is authentic.

The sound and speed of online reviews are noted by purely visiting e-commerce and customer rating sites, such as Yelp and Amazon. There is better variation across the possible industry portion that is distinct from others for reviews (such as hotels, restaurants, e-commerce, home services, etc.). Along with the large number of languages that reviews are written in. Accuracy is a problem with online reviews, since the very great extent majority of reviews are without a label, which means it is not easily known whether the review is fake or not. As an extra factor, standard machine learning algorithms tend to separate into pieces as a result of a blow down and become not producing any significant when dealing with data of this size, which poses a problem when trying to make a formal application these algorithms for review spam detection. Thus, review spam detection is a major problem of online social media, as there are great in number challenges when recognizing and classifying various the reviews from disconnected sources.

2. PROBLEM DEFINITION

In This paper (Huayi, 2017) have presented the concept namely, "Bimodal Distribution and co-bursting in review spam detection". In this concept mostly discovers the reviewers posting rates. That is the number of reviews written in some period of time. It follows the distribution pattern. This posting rate is called bimodal. But the co-bursting means multiple spammers group posting and view that posting to other members this is posted inset of products with short period time. The co-bursting network is based on some co-relations that are used to detect the groups of spammers. It collects the opinion spamming. It uses the Labeled Hidden Markov Model (LHMM) then captures the bimodal behavior for spammer or fake reviewer detection. It uses the classification and ranking methods. In this concept is to collect and analyzing the reviews by using the Frequent Itemset Mining (FIM). But in this concept based techniques is to use in the field of spam detection is holds some drawbacks are, computationally expensive, failure to captures the loosely connected sub-graphs. It does not need the co-reviewing and co-spamming.

This Paper (Huayi, 2015) has presented the conception is, "Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns". In this concept is detects fake secret algorithm for fake reviews. It introduces the large-scale analysis so easily recognize the fake spam. It follows the technical term is spatial and temporal features are used for the supervised learning opinion spam detection. It follows the Support Vector Machine (SVM) because an n-gram feature is available in this technique.

In this paper "Analysis the Graph-based Online Store Review Spammer Detection" is presented (Guan Wang, 2011) generally the online reviews is giving the useful and valuable pieces of information for products and services to the customers. In this scheme is mostly followed the online shopping websites and its related terms. In this spam detection is using the spam detecting techniques and methods are text similarity, linguistic features, and some rating patterns. In this logic is captured out the relationships among the reviewers and it gathers the stored review information's. The approaches are based used in this idea is to finds the difficult and complex spamming activities.

The paper "Finding Misleading the Opinion Spam by Any Stretch of the Imagination" is presented (Myle Ott, 2011). In this concept is mainly focusing the opinion spam and then it collects the website based consumer reviews. It organizes the three approaches for opinion spam detection. The spam is detected by using some effective techniques are referred to in this conceptual area called the spam detection with opinion spam. The text categorization has used the n-gram based classifiers, and in this classifier is included in the SVM. It differently uses the technique is psycholinguistic deception detection which solves the problems such as increased negative emotion and some psychological distancing and big problem scenario is genre identification. So these problem-based activities take by detecting the problem. The genre identification is classified into some types are imaginative and informative writing. The imaginative means the approximation reports. The information means some valuable information. In this concept follows some basic kinds of techniques are, genre identification, psycholinguistic deception detection, text categorization, and classifiers. Genre identification identifies the frequencies of each part-of-speech (POS). In the psycholinguistic deception detection is the same as the Linguistic Inquiry and Word Count (LIWC) software, it automatically analysis the text. But the LIWC is allowed the text classifier. The LIWC is depended and categorized by the following features are, linguistic processes act as function and it includes some features for text. It collects an average number of textual sentences and the rate of the misspelling and swearing, etc. Secondly, the psychological processes include all activities of the social, emotional, cognitive, perceptual and biological processes it slidely related to time or space. It also personal concerns any references (to work), leisure, money or cost, religion, etc. Fourthly, the spoken categories are primarily filtered and agreements the words. The text categorization is used the n-gram features and it includes the unigrams, bigrams+, trigrams+ and superscript+ that all indicate like features. In the classifier, the term is based on the naïve Bayes and SVM. The performance of SVM is separating the high-dimensional hyperplane between the two groups of data. In this combinational approach is detects the spam by using the n-gram features.

Paper "Learning to Identify Review Spam" is presented (Fangtao, 2011). In this concept is uses the sentiment analysis and the opinion mining. In the concept main goal is to solve the problems are faked opinion or opinion spam oriented problem. It acts as a product review mining system. The review site is designed by any number of people may write the fake reviews so that is called review spam, that is it promotes the products or it de-frames the competitor's products. In a scenario using the machine learning algorithm and identifying the review spam and solves the problem. It first analyzes the various spam identification and observes the review spams. Typically, the machine learning algorithm provides the basic two types of methods are supervised and unsupervised learning. So it can provide two-views co-training, semi-supervised learning and it exploits the large or huge amount of unlabeled data. But the semi-supervised learning is accepting the label and unlabeled or both types of data is processed. Then the designed two-view co-training algorithm achieves the aim of the concept that is spam detection.

3. PROPOSED SYSTEM

The drawbacks, which are faced during existing system, can be eradicated by using the proposed research Hybrid Buying and Sequence Spam Detection (H-BSSD). The main objective of the proposed system is to provide a user-friendly interface to effectively detect Spam Detection in Social Network. The proposed system aims at increasing the detection performance through the high dimensional data set. The proposed system implements effective buying behavior analysis technique along with GSRank algorithm this help optimal way for identify spammers and the spam content and find out effective detect group spammers in product reviews. This can be effectively applied for high dimensional data set for increasing the detection performance. This paper implemented Sequential Search and Gs ranking Algorithm along with buying behavior analysis model to classify spammers and the spam content in online E-commerce site.

ADVANTAGES:

- Simple and Effective method to detect and group spammers and gives better result.
- This proposed system can be applied for any type of large dataset.
- To display only trusted reviews to the users.
- Automatic review block option available in this website
- This helps to the user to get appropriate reviews from the website about the product.

4. METHODOLOGIES

USER INTERFACE DESIGN

The purpose of this module is to provide a user interface, where the user/Seller can create their own account with needed information (username, password). Which is mainly created for provide a authentication for each individual user, who accessing the E-Commerce for some purpose. In this module, after authentication process users can purchase and post review they can view review. In case of seller thy can add them product along with complete details and they can view user who can buy their product. In this user interface initially start the admin side work in the implementation process. Admin must contain the unique username and password for authentication process to access in the system work. Admin is the major role to monitoring the user details and product details, purchase details and posting review details and spam detection details. Designing process need to implement the admin side menu details such as registered users details uploaded product details by the company, product complaint details and feedback details, major theme to detect the spam review posted by the customer details in this proposed system monitor the customer buying behavioral analysis and purchase details, proposed method compare the user profile details with product purchase details, trusted user allows to post their review about the purchased product details. Malicious user post split and store into the spam details. In this proposed system designed to show the secured and trusted product reviews only can view by the any customer in the web page. Spam detected content filter and hide from the web pages. HBSSD scheme effectively split the spam content from the group of datasets in the purchase details values.

PRODUCT SHARING

Product Sharing describes when E-commerce seller broadcast product complete web content on a E-commerce to their connections, groups, or specific individuals. One of the primary aims of product sharing strategies is to generate brand awareness and improve the product demand on E-commerce site. In this web page registered ecommerce product selling company initially registered into this proposed system. While registering the seller their profile details collected and stored into the particular field about the seller groups. Registered seller only can access their authentication for include the product into the WebPages. Online buyers are in search of finding products that satisfy their needs. They can easily check an online product with the one available at a shop nearest to their home or office. The web availability has made it that easy.

Online shopping world, the websites responsiveness of a product page is all about convenience for the end user. Customer will be able to write optimized review to provide to the needs, requirements, and common reading behavior of users. In this proposed method seller add their products into the common sales page system will display the product image and product cost and specification of the products and total positive reviews given by the pervious customer all these details collected from the datasets and generate the product details and display for the online purchase customer can easily identify the product quality and product description details. Product sharing

is the main objective for the e-commerce web pages admin and product sellers both are receive the valuable transaction by the customer through this product sharing method.

PRODUCT REVIEW SHARING

Review sharing is a process which refers an item you have purchased and used can be a great way to share useful information with other shoppers, promote products you love, or just build your writing portfolio. One can review almost any product in our proposed system and any one can view the review. This product review sharing become a major factor in business reputation and brand image due to the popularity of review websites. This effective review sharing which helps big part of people rely on available content in social media in their decisions more effectively. This method product review given by the customer side this proposed method monitors the customers buying behavior analysis method with the help of purchase details dataset, system will monitor the customer profile and compare with the purchase details. Incase post reviewing customer is an trusted customer system allows to post the review about their purchased product details, if the customer post review means system will check the customer already purchased the product, if customer details available in the purchase details system ready to receive the customer review content about the product.

BUYING BEHAVIOUR ANALYSIS

Buying Behavior is the decision processes and acts of people involved in buying and using products. Initially this process allows register user can log on this site after successful of user authentication they can purchase / post the Review for the particular E-commerce product. Purchase details include list of parameters such as who brought the product user name and purchase type, purchase id, date of purchase, product name, total quantity, amount etc In this all information to be maintains separately. After post the review system will check user id, product id in buying behavior analysis list and user content review similarity. Based on the similarity and behavior analysis automatically proposed model will find out spam and spammers.

CONTENT SIMILARITY ANALYSIS

User reviews may have multiple sentences along with each sentence May express different kind of meaning. Most of the time same content will be post by user frequently. Content spam is most popular type of Web page spam. So these identify the content similarity proposed and apply cosine similarity model. This model Content similarity is performed on the reviews given by same user. It offers cosine similarity to obtain similarity of two reviews. If the calculate user review cosine value is greater than 0.5 the review is considered to be spam review by user. Review deviation also be considered here if all users give positive review and one user gives negative review or vice versa, then that /review is considered to be spam initially calculate average of all the review and then check to what extent it deviates from the origin. If the review deviation values are greater than or equal to some threshold value, then the review is considered to be spam.

H-BSSD (Hybrid Buying and Sequence Spam Detection)

This chapter presents a discussion about the proposed methodology and the steps involved in the proposed work. Here the system specifies detailed steps about the spam detection.

DATASET PREPARATION AND PREPROCESS

Initially user has to upload review data this data first going to preprocessing unit. Data preprocessing is an important and critical step in the data mining process, and it has a huge impact on the success of a data mining project. The purpose of data preprocessing is to cleanse the dirty/noise data, extract and merge the data from different sources, and then transform and convert the data into a proper format. The purpose of it is to produce result that can be used to improve and optimize the spam content detection.

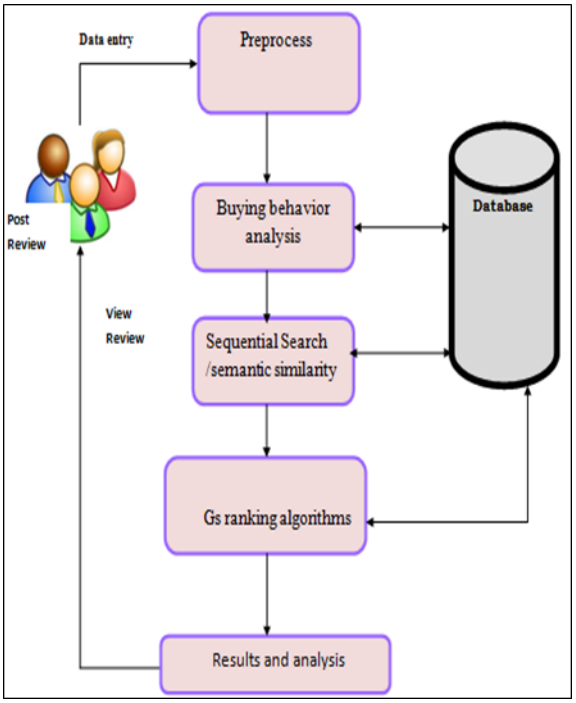


Figure 4.1 Overall Diagram

BUYING BEHAVIOR MODEL

Step 1:

Pattern=product _Id;

Step 2: user post the review for product using product Id

Step 3: compare with user purchase History

```
If (user_purchaseId== Pattern)
        {
                Post the review
        }
        else
        {
                spam content
        }
```

In this buying behavioral model algorithm we initially set the pattern for the product id, while user post the review system will check the user is already purchased this product, or else unauthorized user identification and spam review post method process happen system identifies then check the previous user posted review content and our proposed system analyze the similarity with the user posted previous content and current posting content, if the review is original system allows to show the review to all other users in the reviews page on the webpage.

SEQUENTIAL SEARCH

Algorithm 1: Pseudo code of sequential search algorithm

Step 1: post Review

Ex: The product is good

Step 2: Convert sentences into words

The /product /is /good/

Step 3: Every word is going to compare with wordlist database.

Step 4: if word match with wordlist database automatic word score will be calculate the process will run continuously up to list end.

Step 5: Return review sentiments result.

Step 6: Read the sentiments result from the user review

Step 7: Compare, the result with the first element in the overall user review list.

Step 8: If both are matching, then store the count information display and continue the function.

Step 9: If both are not matching, then compare search element with the next element in the list.

Step10: Repeat steps 8 and 9 until the search element is compared with the last element in the list.

Step 11: Finally it display the overall sentiments result count along with total sentiments result review and terminate the function.

Sequential search is a very simple search algorithm. In this type of search, a sequential search is made over all items record list one by one. Every item is checked and if a match is found then that particular item is returned, otherwise the search continues till the end of the data collection of data set.

SEMANTIC SIMILARITY

Algorithm 2: Pseudo code of semantic similarity algorithm

1: for each review R1 in dataset do

2: Remove stop words list

3: Extract POSs (nouns, verbs and adjectives)

4: end for

For each business B do

6: for each reviews pair (R1i, R1j) ∈ B1 do

7: sim Ri, Rj∈B1

(R1i, R1j) = similarity measure

R1i, R1j∈B

(R1i, R1j)

Similarity measure is each measure out of {cosine, cosine pos non lemmatized, cosine pos lemmatized}

End for

9: for spam threshold T = 0.5, T <= 1, T+ =0.05 do

10: if sim(R1i, R1j) > T then

11: Mark R1i and R1j as spam

12: else

13: Mark R1i and R1j as truthful

14: end if

15: end for

16: end for

This algorithm m used to compute the pair wise similarity between reviews for a particular user. This algorithm analyzes the user previous review content with current review content and calculates the similarity score value of the similarity.

GS RANKING ALGORITHMS

Algorithm 3: Steps of GsRank algorithm

Input: Weight matrices WPG1, WMP1, and WGM1

Output: list of candidate spam groups

Rank the groups based on their activities to identify the spam

Step 1: Find MNR which means too much review in a day abnormal.

Step 2: Find RL Review length.

Initialize VG0 ← [0.5]|G| ; t←1;

Step 3: Iterate:

i. VP1 ← WPG VG (t-1) ; VM ← WMP VP1 ;

ii. VG 1← WGM VM; VM ← WGMT VG1;

iii. VP1 ← WMPT VM; VG1 (t) ← WPGT VP1;

iv. VG1 (t) ← VG1 (t)/ || VG (t) ||1;

Until || VG1 (t) – VG1 (t-1) ||∞ < δ

WPG- Weight per groups

WMP-Weight per maximum page contents

WGM-weigh per group messages

VP-Vector page

VG-Vector Group

in this gs rank algorithm we gives the input of weight matrices values, WPG means weighted per group of values in the review time. User posts the review this algorithm calculates the total content count and group the content values and gives the input for analyzing process. In the step 1 system will check the maximum number of reviews posted by the customer per day. In step 2 analyze the total review content length will be calculated.

In this proposed system after successful product purchase by the trusted user, user can post their opinion about the product quality of service analyzed by the user, post the review in the post review page. In this page system initialize the buying behavioral model and sequence search process, system gather the post previously posted by this user, and verify with the current review content. Content similarity scheme will perform the similarity process and also the bad keywords also verified from the dataset. If users review maximum percentage will match with the bad keys and similarity content with the pervious review, ensure the review is spam and split it and only can view by the administrator of this application.

Content Similarity checking:

In this process system first confirm the user with buying behavioral model and compare with the pervious content and current review content

Step-1:

Check the user authentication,

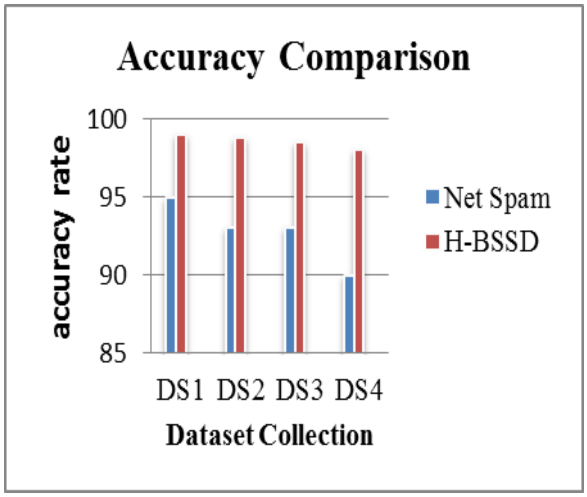
Compare with the previous review

Step-2:

```
while ((line = sr.ReadLine()) != null)
{
if (line.Length > MaxLineLength)
{
File.Type(include the line higher than (0) review messages,
MaxLineLength.ToString()));
}
_lines.Add(new TextLine(line));
}
```

5. RESULT AND ANALYSIS

The proposed techniques and protocols have been implemented in C#.net. The work proposes novel methods for effectively identify and solving the spammer and spam content in web pages. H-BSSD best mechanism for effectively detect Spam Detection in Social Network. Also it provides a comparative analysis of the proposed algorithm with the previous method. In GS RANKING algorithm provide best result when compared to previous algorithm. The primary goal is to determine classify the group spammers and spam content .This process improves efficient throughput and accuracy. In particular, comparing the results that would have been obtained applying proposed technique and show the result how its accuracy increased compared to previous algorithms. The performance of this proposed framework H-BSSD Scheme is compared with existing approaches. The figure below shows the results and comparison of the proposed system.



The table shows the performance comparison of the proposed method with other existing approaches based on the different metrics spam detection delay, accuracy, number of iterations.

Table 5.1 Comparison table

Metrics	Dataset	Net Spam	Proposed H-BSSD
	DS1(100)	95	99
Detection Accuracy (%)	DS2(150)	93	98.8
	DS3(200)	93	98.5
	DS4(250)	90	98

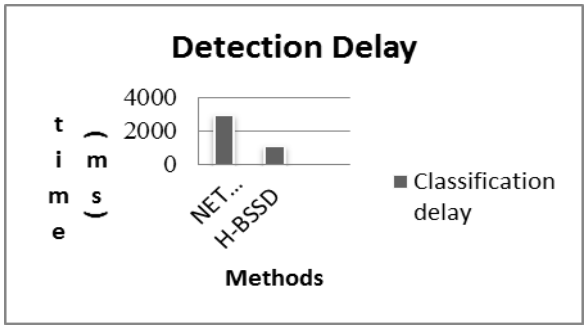
Table 5.2: detection Accuracy

	Net Spam	H-BSSD(GS Ranking)
Spam Detection Delay	2984.06	1108.08
Result Accuracy	90	99
time	70.68	49.69
Number of iterations	7.81	2.21

Spam detection Performance comparison of proposed H-BSSD with existing approaches based on Spam detection Result accuracyPerformance comparison of proposed H-BSSD with existing approaches based on Delay

Table 5.3: detection Delay

Metrics	Net Spam	Proposed H-BSSD
Spam Detection Delay	2984.06	2108.08

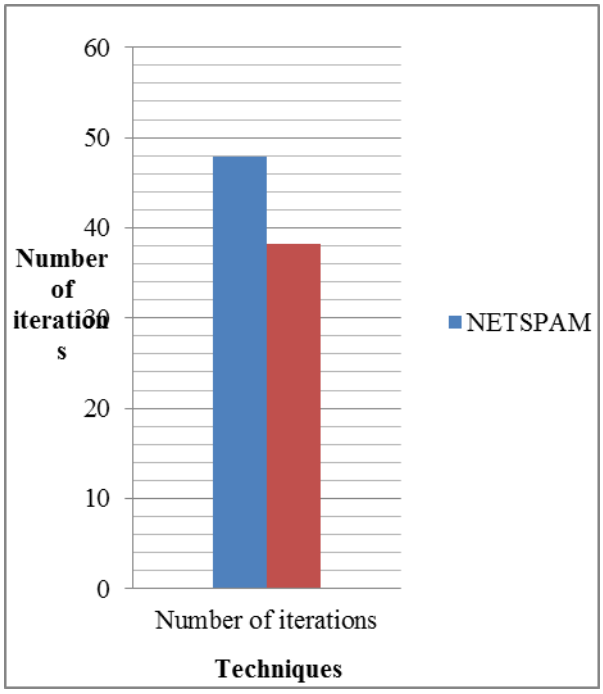


From the chart it shows the performance measure based on the Detection delay and the proposed approach took less time while comparing the other methods.

Table 5.4: Number of Iterations

Metrics	NETSPAM	Proposed H-BSSD
Number of iterations	47.81	38.21

Performance comparison of proposed H-BSSD with existing approaches NETSPAM based on number of iterations



From the chart it shows the performance measure based on the processing of the proposed approach H-BSSD took less number of iterations for the amount of datasets while comparing the existing method and the worst based on the iterations is NETSPAM.

6. CONCLUSION

The thesis handled the problem of analyzing and finding spammers and spam content in web pages. To solve the problem the thesis effectively proposes Hybrid Buying and Sequence Spam Detection (H-BSSD) models. The H-BSSD model can filter spam review in ecommerce site. proposed system implements effective buying behavior analysis technique along with GS Ranking algorithm this help optimal way for identify spammers and the spam content and find out effective detect group spammers in product reviews. This can

be effectively applied for high dimensional data set for increasing the detection performance. This paper implemented Sequential Search and Gs ranking Algorithm along with buying behavior analysis model to classify spammers and the spam content in online E-commerce site.

REFERENCES

[1] Fangtao Li, Minlie Huang, Yi Yang, Xiaoyan Zhu. Learning to Identify Review Spam". Li, Fangtao, et al. Learning to identify review spam. IJCAI Proceedings-International Joint Conference on Artificial Intelligence. Vol. 22. No. 3. (2015).

[2] Guan Wang, Sihong Xie, S. Yu. "Analysis the Graph based Online Store Review Spammer Detection". Data mining (icdm). 2011 ieee 11th international conference on. IEEE.(2011).

[3] Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Jidong Shao. Analyzing the opinion spam and detecting opinion spam on a large-scale dataset by the use of temporal and spatial patterns. ICWSM. (2015).

[4] Huayi Li, Geli Fei, Shuai Wang, Arjun Mukherjee, Jidong Shao. Bimodal Distribution and Co-Bursting in Review Spam Detection. "IC on World Wide Web", 26th proceedings. International World Wide Web Conferences Steering Committee. IC-International Conference. (2017).

[5] Myle Ott, Yejin Choi, Claire Cardie, Jeffrey T. Hancock. "Finding Misleading the Opinion Spam by Any Stretch of the Imagination". Annual Meeting of the Association for Computational Linguistics. Human Language Technologies-Volume 1. Proceedings of the 49th, Association for Computational Linguistics. (2011).