International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Early Reviewers Prediction and Spammer Detection on E-Commerce Websites

Jayendra Kumar, Palakursha Sirisha

Abstract*: Reviews which are posted online play a vital part in present world as most of the customer's purchase items through an e-commerce website. Reviews which are posted on websites at an early stage known as early reviews, even though their contribution is very small their opinions determine new product's success and failure. Most of the spam reviews are written to improve their profit and promote their products and defame other products. In this system, the concentration is mainly on early reviews of the products and the products categories ranking on e-commerce websites i.e., Amazon. The analysis of reviews of product defines ratings of early reviewers' and helpfulness scores of them are probably influencing product promotion additionally this model is enhanced with ranking and spammer detection.

Keywords*: Early Reviewers, E-Commerce, Ranking Model, Spammer.

I. INTRODUCTION

Digital platforms are being used by many companies in order to promote their products. Users have got a platform such as E-commerce websites to share their experience of products by posting reviews after purchasing them on their websites. In order to purchase any product the customers mostly read the reviews of that particular product on their websites. Customers are buying the products based on the online reviews given to the products as they are very important for product rating and increasing product marketing and help users to rate the product also impact greatly on the products and their identification is useful for early promotion. Reading product reviews for purchasing a product became a habit to the customers. If there is a positive review then the chance of buying that product by customer is more compared to negative reviews. About 60% of overseas customers read reviews that are posted on online websites in order to buy a particular product. Online reviews help users to rate the product so as to increase product marketing. Reviews posted on websites at an early stage are known as early reviews are often a useful source of information especially the early reviews will have major influence on sales of product. Even though, early reviewers' contribution is very small, but their views might help in promoting or demoting the newly launched products [2], [3]. In specific, the lifetime of product is divided into three successive stages; they are early, majority and laggards. Early reviewers are being characterized described based on the behavior of their rating and score i.e., yes and no. Social networking links or communication channel are not noticed in most of the domains of applications. Early reviewer's prediction may not be done for communication channels and social networking links. This paper presents a way for modelling early reviewers' behavior and characterizes the process of adoption in Amazon review dataset. Mostly for a given product, sorting of reviewers is done based on the timestamp of reviews. According to Bell curve theory, the lifetime of product is being divided into three successive groups which are early, majority and laggards. Reviewers who post their reviews immediately after the product launch are known as early reviewers. The paper focuses on early reviewers' prediction and product categories ranking on e-commerce websites, which is different from the existing works. This paper main objective is to introduce a model for prediction to find early reviewers and assess whether a reviewer is a spammer or not and rank the categories of products.

A. Problem Definition

In e-commerce, before making an informed purchase choice most of the users will consider reviews posted on that websites. For buying a product, shoppers read online reviews. Product reviews (i.e. reviewed in starting phase of the product) will have major effect on future product sales. Early reviewers make just a small part of reviews, but the success or failure of new products and services can be confirmed by their opinions. Early reviewers are important for companies as their feedback can help enterprises in adapt marketing approaches. The problem of prediction of early reviewers from reviews posted online and detection of spammers is not consistent with existing social network structure methods or communication channels.

B. Objective

In this study, the related work is about studies carried out on review data mining. The paper focuses on early reviewers' prediction and ranking product categories on e-commerce websites, which is different from the existing works. The objective of this study is to develop and validate a predictive model to find early reviewers' prediction and ranking product categories on e-commerce websites, which is different from the existing works.

Manuscript published on November 30, 2019.
* Correspondence Author
Jayendra Kumar*, Computer Science and Engineering, Anurag Group of Institutions, Hyderabad, India. Email: jayendrakumar@cvsr.ac.in
Palakusha Sirisha, Computer Science and Engineering, Anurag Group of Institutions, Hyderabad, India. Email: sirishapalakusha@gmail.com

The objective of this study is to develop and validate a predictive model to find early reviewers and assess whether a reviewer is a spammer or not and rank the categories of products.

II. LITERATURE REVIEW

The literature review includes the work of Abhijit V. Banerjee [4] who made a study on "herd behavior" and analyzed a sequential choice version wherein each decision maker appears at the choices made via preceding decision makers in taking her very own selection. K. Zhang, Y. Cheng, W. Liao, and A. Choudhary introduced a approach for ranking products by exploring the sentiments of reviews and considering how a review's helpfulness votes and its posting date impact the product's ranking [5]. Gerardo Ocampo Diaz and Vincent Ng surveyed on several recommendations on computational modeling and prediction of evaluated helpfulness [6]. Ida Mel, Francesco Bonchi, Aristides Gionis introduced a webpage recommendation system that identifies new interesting pages early and recommending these pages to similar users can be done by analyzing early adopters browsing activity [7]. Due increase in opinion mining of reviews, the existing systems mainly rely on extraction of reviews that are positive and negative using natural language processing techniques No study reported that reviews are trustworthily which is important for opinion models. As there is no restriction in posting on web anything can be written by anyone thus, resulting in low level of reviews and cause of spammers who post spam reviews [6].

A gold-standard multi domain dataset is being used by the authors for detection of false opinions. By using crowdsourcing hotels, hospitals and restaurant reviews are being produced. Differentiation between truthful and deceptive text is being enabled using SAGE. K. Zhang, Y. Cheng, W. Liao, and A. Choudhary proposed a feature-based product ranking system in which reviews are categorized into four groups: positive subjective, negative subjective, positive comparative, and negative comparative. A pRank algorithm uses Amazon.com data for evaluation [5].

III. PROPOSED MODEL

To predict early reviewers, a completely unique approach is proposed which is considers as competitive game of posting of reviews. Solely the foremost competitive users will become the first reviewers of a product. Several pairwise comparisons are done between two players by using competition method.

In an exceedingly two-player competition, the leader can defeat the failure with associate degree early timestamp. From studies on distributed learning [8], a predictive ranking model is being proposed. In our proposed work we develop a model for prediction of reviews and ranking categories of product. The proposed model is developed in order to predict reviews that are posted at early stage of product release and ranking the categories of product. The Amazon dataset used here was made available by Dr. Julian McAuley from the UCSD which consists of reviews of product and their metadata. The reviews dataset consists of attributes such as user ID, review ID, category, and rating, helpfulness votes, and review text, review date for each review.

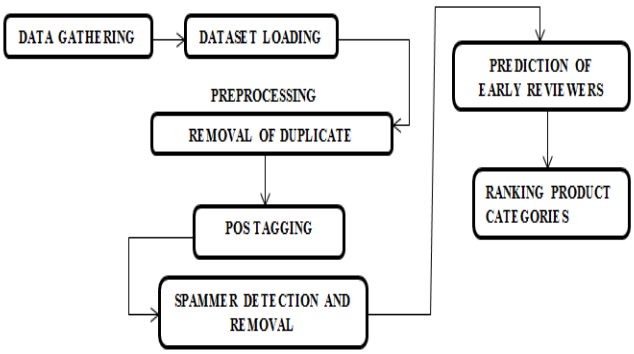

Fig. 1. Block Diagram of Proposed Model

IV. IMPLEMENTATION

In this system, At first preprocessing is done in next step the spammers detected and removed, later early reviewers are identified and predicted and next categories are being ranked. The block diagram of proposed model is as shown in figure 1. All the steps of proposed model are briefed in the below section.

A. Preprocessing

Initially all the duplicate reviews are being removed. Later, all the users who are not active and unfamiliar products are removed. The stop words and non-frequently used words are also removed from review text. Unwanted words are being removed from the review data. Later the review content is being analyzed whether it is positive or not by using sentiment analysis. This analysis of sentiments is useful for checking whether reviewers are genuine or not. In the next section the review spammers are being detected. Spam reviewers and their reviews are being ignored in further analysis.

B. Spammer Detection

Using sentiment Analysis, review content is being categorized into positive and negative reviews. Most of the time the spam reviews' opinions will be extremely high or extremely low. Such type of review will affect the product reputation. The focus of this paper is to study the early adoption behaviors of genuine users. However, spam reviews count has gradually increased on websites of ecommerce, and it was found that about twelve to twenty percentage of reviews are earlier reviews and are frequently be written by spammers. Most of the spammers post their reviews in order to demote the product and organization and their reviews may lead to inaccurate results in our analysis. In order to avoid this removal of spam reviews and their reviewers is being done and to get accurate prediction the spamming methodology is used i.e., if same comment is written by different user id or same user writes different comments under same product then that user is considered as spammer, by using this method of spammers detection all the spam reviews are being removed.

C. Early Reviewers Identification and Prediction

In order to identify early reviews, we examine the process of division of life time of product into multiple categories. Their view process for customers is

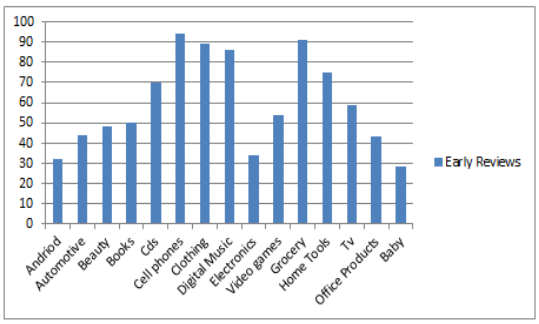

Fig. 2. Graphical representation of percentage of early reviews for different products

seen as adopting an innovation adoption of innovations on website of e-commerce. The time-consuming process is usually shown as bell curve and is categorized into five phases. Customers later are classified into innovators, early adopters, early majority and late majority, and laggards accordingly in five different groups. Rogers ' bell curve theory is being used for dividing product duration into five successive phase. Innovators are very less in count in the Amazon dataset which is used in this paper, so combining early adopters and innovators as early reviewers took place. Additionally, early majority and the late majority are also combined with the majority, later detecting and calculating early, majority and laggard reviews by producing graph to them. The early reviews being posted by Amazon users for several products are shown in Fig. 2. Probability distribution of early, majority and laggards is done and early reviewers are defined, by using these groups and the average of reviews' rating scores is being compared by the three groups. It is viewed that compared to other two groups early reviews are having more ratings so, the early reviewers may be helpful in promotion of product marketing.

D. Ranking the Categories of Product

After prediction of early reviewers, ranking the categories of product based on all the attributes is being carryout in this section. Finally best category and how it is useful to increase the sales of the website company. In order to rank the products, the total reviews and their quality matters a lot. Example, Books category has more rank then company should focus on that product category and increase that category to get maximum profit. Precision and recall metrics are used for checking ranking efficiency and by knowing the factors that affect product ranking allows you to better optimize your product sales. For ranking the categories of products, category ranking algorithm is used. The description of category ranking algorithm is as follows.

E. Category Ranking Algorithm

Reviews which have highest positive reviews and yes votes and has highest rating then they are ranked number one which means this particular category of product has good rank. The companies need to focus on such products in order to improve their business promotions and e-marketing. The graphical representation of category ranking of products is shown in Fig. 3.

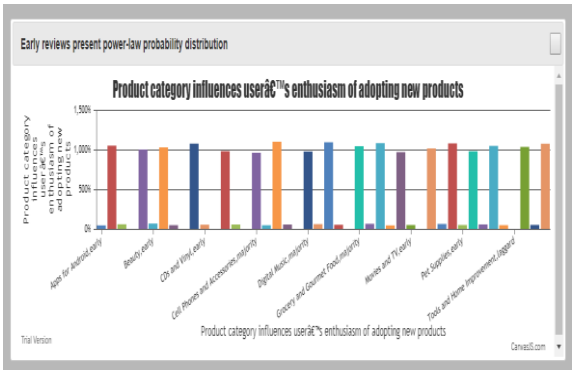

Fig. 3. Category ranking of products

V. ANALYSIS OF RESULTS

The early reviewers' prediction and category ranking results are being discussed in this section. The simple ranking used to rank reviewers which depends on the total count of reviews being posted is not best as it does not give accurate results. NER (Number of early reviewers) is improving over NR (Number of reviews), demonstrating that a customer who had previously been an early reviewer of products might have a chance of adopting newly launched products in future.

The proposed model gives significant progress than other concepts which are being discussed previously. Multiple attributes such as positive review count, voting score, ratings are being considered in order to give accurate ranking to the product categories. The below table I indicates the comparison of existing model with proposed model by measuring metrics such as accuracy, time of detection and iteration count.

Table-I: Comparison of Existing and Proposed Model

Metrics	Accuracy	Time of Detection	Iteration Count
Existing Model	91	71.09	53
Proposed Model	97	45.50	29

VI. CONCLUSION

In this paper, a unique task of prediction of early reviewer on real-life review dataset from Amazon is achieved. The experimental analysis of this study supports theoretic conclusions from sociology and economics. It has been reported that highest rating and helpfulness scores are being given to early reviewers. The experimental results also show that ratings of early reviewers' ratings and their received useful scores may have influence on popularity of product in future. In this study a competitive method is taken for modelling the process of posting reviews and developed a predictive ranking model for prediction of early reviewers and spammer detection and category ranking. In the current work review content is considered. Prediction of early reviews is performed and also sentiment analysis of review content is done to detect whether reviews are positive or negative. Ranking of categories of products is achieved using ranking methodologies.

In future, one can consider different datasets such as Netflix in order to carry out more perspective analysis. As of now, our focus is on prediction of early reviewers and detection of spammers in them and its removal, how product marketing can be improved using the predicted early reviewers is not being addressed which can be addressed in future.

REFERENCES

1. W. D. J. Salganik M J, Dodds P S, "Experimental study of inequality and unpredictability in an artificial cultural market," in ASONAM (2016) 529-532.
2. R. Peres, E. Muller, and V.Mahajan, "Innovation Diffusion and New Product Growth Models: A Critical Review and Research Directions," in International Journal of Research in Marketing (2010) 91-106.
3. L.A.Fort and J.W.Woodlock, "Early Prediction of Market Success for New Grocery Products," Journal of Marketing (1960) 31-38.
4. Banerjee: A simple model of herd behavior. In Quarterly Journal of Economics (1992)797–817.
5. K. Zhang, Y. Cheng, W. Liao, and A. Choudhary, "Mining millions of reviews: A technique to rank products based on importance of reviews," in Proc. 13th Int. Conf. Electron. Commerce (2011)1–8.
6. Gerardo Ocampo Diaz and Vincent Ng, "Modeling and Prediction of Online Product Review Helpfulness: A Survey," in Proceedings of 56th Annual Meeting of the Association for Computational Linguistics (Long Papers).Melbourne Australia (2018) 698-708.
7. Ida Mel, Francesco Bonchi, Aristides Gionis, "The early-adopter graph and its application to web-page recommendation," in CIKM (2012)1682–1686.
8. T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in ICLR (2013).

AUTHORS PROFILE

Jayendra Kumar is currently an Assistant Professor at Computer Science and Engineering Department, Anurag Group of Institutions, Hyderabad, India. He obtained M.Tech CSE from JNTU Hyderabad. His research interests are Data Mining, Internet of Things and Machine Learning. He is Life Member of Computer Society of India.

Palakusha Sirisha profile is currently pursuing her M.Tech in Anurag Group of Institutions, Hyderabad, India, after completing her B.Tech from Aurora's Technological and Research Institute. Her research interests include Data mining and Machine Learning.