2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)
Bangalore, India. Oct 7-9, 2022

Predicting Early Reviewers on E-Commerce Websites

Anil D * and Suresh S **
* Assistant Professor, CMR Institute of Technology, Bengaluru, India.
** Assistant Professor, Nagarjuna College of Engineering & Technology, Bengaluru, India.
anil.kumar@cmrit.ac.in

2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT) | 978-1-6654-6855-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/GCAT55367.2022.9972025

***Abstract*—E-Commerce websites have emerged in recent times for the users to share their opinion on the product received by them by posting reviews. Machine learning techniques are very useful for analyzing the customer reviews in e-commerce websites. These days, a large number of people look for computerized retailers, like Amazon and Yelp. The main objective of this work is to characterize & predicting early reviewers for increasing the product sales. Previous Studies shows how early reviewers rating's and their scores impact the product popularity. Subsequently, in this paper, with the use of begin and final time datasets along with product review time span. We have proposed Algorithm for predicting Early Reviewers using Margin Based Embedding Models and Product Embeddings. The outcomes showed the algorithm can predict early reviews and applied various classifiers for the same. Naive Bayes and SVM Classifier have been best performing among all.**

***Index Terms*—arly Review, E-Commerce, Product embeddings, Naive Bayes and Support Vector Machinearly Review, E-Commerce, Product embeddings, Naive Bayes and Support Vector MachineE**

I. Introduction

Nowadays because of e-commerce websites the user's are sharing their product or purchase experience by posting review comments in form of feedback, opinion or comments. Users will go through the review comments of that product before making a decision to purchase it. There is a report stating 71% of online shoppers globally will go through the review before going to product purchase. The reviews given by the early users of that product have a great impact on the product sales. The users who give reviews in the beginning of product sale are known as early reviewers. Although their reviews make a small contribution in reviews, their opinions will determine the product's success or failure. It is necessary to identify the early reviewers because their feedback can help the companies to improvise their product quality which in turn leads to the product success in future. Hence early reviewers are the attracters during the early or beginning stage of product launch.

Early reviewers have gained the attention of marketing practitioners who are interested in deriving the intention behind purchasing this product or customer satisfaction with the product. For instance Amazon which is one of the world largest e-commerce companies has arranged early reviewer programs to gather reviews of the products which have just launched or have no review yet. This program helps the online shoppers to know more about the product and make wise decisions in purchasing. There is another program called Amazon vine in which all trusted reviewers are invited to provide their valuable reviews for new items and pre-release items so that the following user's can make better decisions in choosing the product. [1].

From the above discussions it's clear that early reviews are necessary for product marketing. Here as an initiative, studies on behavioral characteristics of early reviewers are carried out by going through their posts in e-commerce platforms such as Amazon or Yelp. Effective analysis is performed to determine accurate prediction on early reviewers. Among these studies most of them are theoretical analysis and there is a deficiency of quantitative investigations. Rapid growth of social network and online social platforms and also huge availability of data packs, on these social networks diffusion studies of innovation has been carried out.

For any application domains, social networking and communication channels may go unnoticed hence present methods relying on these cannot be implemented to predict early reviewers from online reviews. To proceed further in analysis of early reviewer behavior, two large datasets are taken into consideration such as Amazon and Yelp. First foremost for a given product the reviewers are sorted based on their postings of review timestamps. Later product life time is divided into three stages: early, majority and laggards. A review post done by a user in the early days of a product is called an early reviewer. Here the main focus is on two tasks: one is to study the characteristics of an early reviewer by comparing with the majority and a laggard reviewer. Ratings behaviors, scores and correlation with reviews associated with product popularity are characterized. Another task is to predict the early reviewers of a product by learning prediction model [2].

Two important metrics are taken into consideration to analyze the characteristics of early reviewers. 1) Highest average rating score is given to the product by early reviewers. 2) Most useful reviews are shared by the early reviewers. The findings can be related to personality variables theories as highest average rating scores indicates a positive attitude towards the product and helpfulness reviews given by others are viewed as a measure of the opinion's leadership. The analysis also indicated that early reviewers ratings and scores measure the popularity of a product. These findings are explained further with help of herd behavior and this tells that people's decisions are dependent or influenced by others decisions.

978-1-6654-6855-8/22/$31.00 ©2022 IEEE

Early reviewers are predicted using a novel approach in which the process of posting reviews is viewed as a multiplayer completion game. Early reviewers are the users who are more competitive with respect to the product. This competition is further divided into multiple pairwise comparisons, here the winner beats the other by early timestamp. Here a margin-based embedding model is proposed where product and user are mapped into the same embedding space, and then couples of user's are determined based on the distance to the product representation.

Early studies told that individual decisions are influenced by other decisions and it's explained by herd behavior. An early review contains the evaluation performed by the previous adopters which will influence the product purchase by others. When there is consideration of previously given review comments on products by the customers in online shopping then there comes herd behavior. An early review has the evaluation done by early adopters which are the reference for product purchase. The importance is given to quantitative analyses of characteristics of early reviewers considering real world datasets.

Contributions are as follows: 1) Early reviewers on e-commerce websites are characterized using two real world large datasets. 2) Early reviewer's characteristics and their effect on product popularization is quantitatively analyzed and this analysis supports theoretically conclusions derived from sociology and economics. 3) Review posting process is viewed as a multiplayer competition game and developed by embedding based ranking models for prediction of early reviewers. This model deals with cold start problems by incorporating side information of the product. 4) Intensive experiment using two large real world datasets i.e. Amazon and Yelp demonstrate the effectiveness of the proposed approach for prediction of early reviewers.

II. Literature Review

M. V. Sagvekar and P. Sharma [3] has presents study on information cascade have worried the preference of information in social sites influence the user decision making. For every time the people use product description and compare with many others products.

B. Baiju *et al.,* [4] has proposed to describe to using the Bayesian theorem perform hard simple decision behavioral. Assume the hard work design to expand the analysis using conformation functional magnetic resonance imaging (FMRI). They also presents to research on individual decision about the condition of independence and lack of independence in the faces of group force. The considerate of social influences will necessitate the learn to a broad range of conditions and of the interrelated operations of various psychological functions [5]. Rajesh M *et al.,* [6] has proposed a probability matrix related to the graphs. Unfinished observation under the imagination of probability matrix fulfill by the rows and columns by some permutation. Assume more than ten customer simultaneously purchase book online based on the possible fast or quickly basis on the book title is a keyword for searching book and also consider the good reviews and bad reviews. Only the book title is the keyword to search books, for customers there is no other preference to buying books[7]. Bai *et al.,* [8] has proposed brief introduction of the rational selection approaches, followed by an identification of numerous of the main criticisms of R.C.T and its conceptual and empirical boundaries. The rational choice theory it's very real time approach. Here to presents some includes reflection and table matching similarities and also variations between the mainstreams of R.C.T few of previous products of an rising selection concepts. Rational selection has been systematically criticized The reviews maps a small number of key ideas and assumptions underpinning the conceptual model and experimental applications of RCT. Sumathi *et al.,* [9] has proposed a new concept is psychological law, called the law of comparative judgment, it is special methodology to use measurement of psychological values. This in not only applicable for comparing judgment, its helps qualitative judgments, assumes that a1 and a2 are the psychological scalable values for compared to the proportion of judgment.

III. Proposed System

The main objective of problem statement is maintenance of all the positives of the existing system, while overcoming a few disadvantages are stated earlier. In online reviews based on the product sales in the time period. The previous year's very difficult to get product reviews, the product companies are manufacturing their products huge numbers, because the company does not find specifically which product is moving fast. And it is difficult to think of the end user, which products he's interested in. So that many problem facing companies without reviewing end user requirements.

To propose the following scheme to conclude whether a product's reviews time span is complete within our observation windows. The watching chart is described as the duration between the begin and final time datasets. Assume the Amazon datasets having products reviews rang from may 1996 to july 2014 and another one is yelp datasets rang form 2004 to 2017 [1][10]. For watching charts are 18 years and 13. Here propose mainly in product reviewers . Here new methods introduced to support the end users [11]. Such as product review time span is completely within the watching charts. In the time stamp divided in three category are the early, majority, and laggards.

A. Define fully Review on Time Span:

Here introducing the technique for leading gap and trailing gap. Provides the watching windows [s^{start}, s^{end}], the leading gap of a product p, represented by $\Delta_T^{(p)}$, its describes the time variations between the S1 and s^{start}, while trailing gap $\Delta_T^{(p)}$ of a product p is its describes the time variations between the Snp and s^{end}. The key ideology is at most interval between two consecutive reviews of a product p is lesser than both the leading and trailing gaps of product p.

[1]https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset/code

B. Estimating Product Lifetime:

A specific product to take a complete review time span as a duplicate measure of the lifetime. It should be noticed the time span derived from product reviews may not accurately align with the real products lifetime from the end-user point of view

C. Early Reviewer Identification:

The product complete lifetime ,to learn about to divide the product lifetime into various steps to finding the early reviewers. In the online shopping webs ports, having reviews processed users and product manufacturing industries to thing customer requirements to innovative products launch to the market. end-users are then categorize for that reason into five variety of groups, are innovators, early adopters, early majority, late majority and laggards.

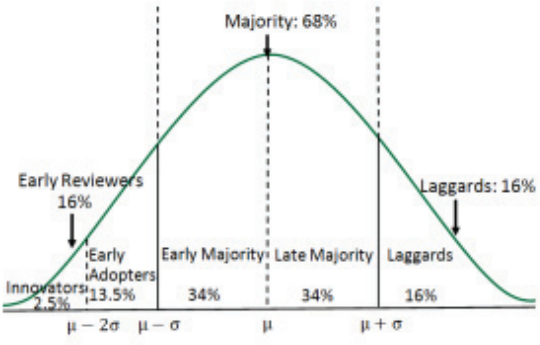

Fig. 1. Categorization Thresholds of three reviews.

The Figure 1 shows an example of instance on the stage division by Roger's theory for The x-axis represents the adoption time. Example if product is uploaded on 2010 Jan and users writes the reviews on product, we call them has early reviewers, and for the same product if the user writes the reviews on 2011 Dec, we call them as laggards. If time spam falls within the probability range we call them early reviews, laggards will depend on the time frame according to what we fix by using Rogers theory. Using the categorization

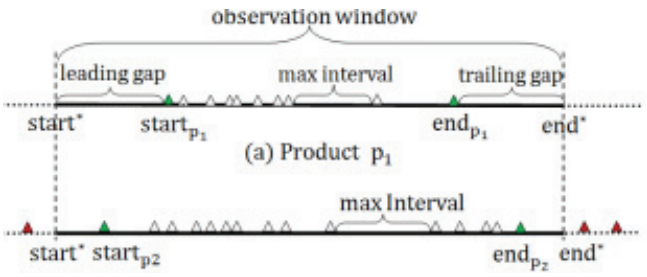

Fig. 2. Review time span of Product

thresholds listed in Figure 2, shows the overall Product reviews Time span of taken data set, which refers to the time span between the First and last received reviewers for a product, by considering observation Window, Based on the time frame from 1990 to 2000 data which contain in amazon catalog, firstly by the given observation by introducing the concept of leading gap and trailing gap, the graph is plotted to determine the complete review time spam. $[s^{start}]$ the data which come before this is leading gap and $[s^{end}]$ the data which come after this trailing gap.

IV. Data Collection and Preparation

The system shows the architecture diagram for predicting the early reviews on the products in e-commerce websites. The architecture diagram first data is collected from amazon website and data is cleaned in the data preprocessing stage. Data acquisition: Collect data from Amazon or flipchart online

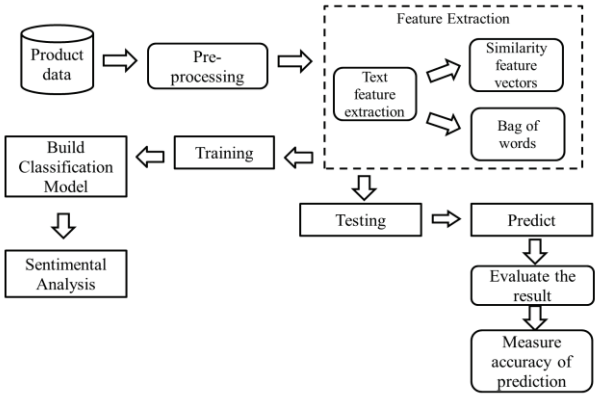

Fig. 3. Architecture of product review prediction

resources. Full data is available at [2] . For starters, we focus on databases for specific product categories, in order to have a reasonable size. Once on the page, We have to download the reviews for the "Musical Instruments" and "Baby" categories. We have also worked on the reviews of the "Movies and TV" category, but the size of the file and the calculation time make it much less easy to use. From the amazon product repository by importing baby-zip file which is in json format and by creating widget option for multiple dropdown menu dynamically taken for 3 category

The Figure 3 shows the features are extracted from the reviews and data transformation is done and different machine learning algorithms are modeled and analysis is done on early review prediction. Raw information transfers to the preprocessing steps in the pre processing methods handle the separates the information for feature extraction. Having main three sub categories, after preprocessing pass to the text feature extraction. This phase divides the similarity feature vectors and bag of words this two phases finding duplicate reviews and mistakes. Finished the process. Then pass the data into training. The training data train the algorithm and processing data going to next phase for build classification model, this model classified the data then pass to the next step for the sentimental analysis, after classified data processed in sentimentally analyzing. Find the similarities words and bag

[2]https://jmcauley.ucsd.edu/data/amazon/

of words correct then its send to the testing data in testing phases, finished the testing then predicting data then evaluating the outcome results. Then measure the accuracy of prediction of data.

V. IMPLEMENTATION

In our work Algorithms to predict product popularity in the early stage of customer reviews is being discussed.

A. Algorithm to Predict Early Reviewers

One of the many criteria for product popularity is the early reviews given by the customers of that product and that's what is explained so far. Once product is launched to market there can be the prediction for it's reviewers who are called as early reviewers who will give reviews in the early stage of the product. Hence early reviewers can be predicted, first foremost for the early stage of product identifying the early reviews is going to help the company to promote their product and also manage, monitor product for better marketing. And early revisers are the real adapters of products which will help in the product sale, their reviews will directly impact the product purchase.

1) A Margin based Embedding Models for Predicting Early Reviewers: In this work, two applicant users u and u0 of given product p are considered and in- between them partial order is modeled. Therefore the total order ranking problem can be casted into a pairwise comparison problem. Embedding model is used for this task by an inspiration from distributed representation learning which is progressing well in recent days. By this way user u and products p are modeled with low-dimensional representation vectors vu and dense representation vector VP respectively. By this representation objective functions S(p,u) is defined as inner products between user and products embeddings, i.e.,

$$S(p,u) = v_p^T.v_u \tag{1}$$

Expectation in embedding space is that $v_p^T.v_u > v_p^T.v_{u^l}$ when $u > p^{u^l}$. For learning such embedding's, margin based ranking criterion over the training set T is minimized as follows.

$$l(T) = \sum_{u>p^l \epsilon T} [S(p,u^l) - s(p,u)] \tag{2}$$

$$= \sum_{u>p^{u^l} \epsilon T} [m + v_u^T l.v_p - v_u^T.v_p] \tag{3}$$

2) Learning the Product Embeddings: When a innovative product is launch, users are not able to learn its embedding since no review data exists. Recall that a products p is with a category label c_p and a title explanation T_p . These two types of side information can be used to pertain the products embedding. A label with explanation is a series of word tokens. Assuming that all of this models have the dimension number of L. When the models is learned, can obtain the embedding for words (v_w 2RL), label (v_p^T 2RL) and category labels (v_p^c 2RL). Our product embedding representation is finally a vector concatenation of label embedding and category embedding, i.e., vp = vec(v_p^T;v_p^c), where vec(v_p^T _;v_p^c _) takes two column vectors and returns a concatenated column vector. With v_p 2R2L, we have to set the dimension number for v_u to 2L, i.e., v_u 2R2L.

TABLE I
OPTIMIZATION OF THE NAIVE BAYES & SVM CLASSIFIER

Notations	Description
U	a set of ecommerce users, $u \epsilon U$
P	a set of ecommerce products, $p \epsilon P$
r,s	rating posted by a user with a timestamp s on the product
n_Y,n_N	the number of 'yes' votes and 'no' votes a review received
d	a review d is composed of (u,p,r,s,n_Y, n_N)
c_p,t_p	the category label c_p and title description t_p of a product
L_p	a list of ordered reviews of the product p, $L_p i$-¿d1-¿d2.....-¿d_i.....-¿$d_N p$
$\Delta_L^{(p)}$	the leading gap for product p
$\Delta_T^{(p)}$	the trailing gap for product p
$\Delta_M^{(p)}$	the maximum interval for product p
v_p, v_u	low-dimensional representation vector of product p and user u
v_{tp}, v_{cp}	title embedding and category embedding of product p
S(p,u)	the likelihood that user u becomes an early reviewer of product p

3) Algorithms: In Machine Learning using the sklearn library, we will apply the following templates, Naive Bayes with multinomial model, Support Vector Machines with Stochastic Gradient Descent and Logistic Loss Function, during the trial of Random Forests the calculations were never successful, considering the large number of dimensions. We chose a Multinomial model for Naive Bayes because it is well suited to data (large sparse matrix). A Gaussian Naive Bayes model is not compatible with the sparses matrices, and a Bernoulli Naive Bayes model can only take Boolean input variables. As for the choice of the logistic loss function for the Support Vector Machines model , it is the only function that makes it possible to calculate probabilities on the predicted values and thus to determine the AUC criterion.

We see that the best performing classifier is Naive Bayes on all the text, without the extra features. The addition of additional features does not particularly enhance performance, except for the SVM model, which however remains below the NaiveBayes. In general, the performance of the SVM classifiers is lower than or equal to that of the Naive Bayes classifiers [12]. Finally, to evaluate the performance as shown in Table 6.3, we will draw the OCR curve and evaluate the AUC criterion. To facilitate the readability of the Range, the classifier is eliminated only on the additional variables, since it predicts only 5. It is seen that the contribution of additional features matters only a small gain of AUC to the SVM classifier. Naive Bayes models are still much better in terms of AUC than SVM models[13].

B. Algorithm: Learning algorithm for userembeddings

To study the embedding parameter, as shown in the above Table I with the help of Notations Description, this can purely apply Stochastic Gradient Descent(SGD) for updating user embeddings v_u and product embedding v_p . To handle the cold

TABLE II
ACCURACY OF THE CLASSIFIERS

Classifiers	Accuracy	Precision	Recall)	f1-score
NaiveBayes-Basic	75.2%	75.2%	96.0%	84.3%
NaiveBayes-Optimized	76.1%	76.9%	93.1%	84.3%
SVM-Basic	75.5%	75.2%	96.0%	84.3%
SVM-Optimized	76.5%	78.3%	91.1%	84.2%

start problem, to incorporate the title and category information to relearn the product embeddings v_p . During the learning process, we fix the product embeddings obtained with the labeled doc2vec, and simply optimize the user embeddings. To learn parameter, to implement S.G.D for optimization. The complete optimization process is described in Algorithms. The strategy proposed. At every main iteration of the algorithm, a training triplet hp, $u_u^i 0$, where the incomplete order$u_u^p 0$, is sample from the training set for optimizing the margin based ranking criterion function (T). Imagine that the number of reviews of a products is n the sum number of comparison pair that can be generated is roughly estimated as O (n2). When the number of products is extremely huge, the number of comparison pairs will become enormous.

VI. RESULTS

The first step of dataset exploration data is cleaned by removing stop words, eliminating the patterns and Punctuation removal etc. and transformed into a data frame. Figure 4 shows the accuracy of the classifier Naïve Bayes and SVM classifier to see the best performing classifiers.

Table II shows the Performance Evaluation, which drawn in OCR curve and evaluate the AUC criterion. To facilitate the readability of graph for Sentimental analysis for Classifiers.

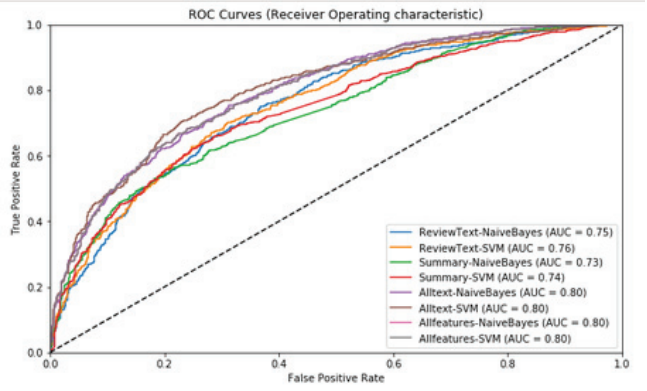

Fig. 4. Results of Early reviewer Prediction

VII. CONCLUSION

Studies were carried out on characterizing the early reviewer's and also predicting two real-world online review datasets. The analysis gives strong strength for all the theoretical conclusions from sociologies and economics. It's found that highest average rating scores are assigned by the early reviewers and they do also post most useful reviews.

In this present task, review contents are not considered. As part of future work have to try to include the reviews contents into these early reviewer prediction models. Due to difficulty in extracting applicable information from reviewer information, could not study the communication channels and social networks structures in diffusion of innovations partly. Many sources of data like Flixster is studied from which social networks are extracted and some analysis is performed. At present target is on analysis and predictions of early reviewers but there is an addressable problem of improvising the products marketing with identified early reviewers. As a future work there can be investigation with actual e-commerce cases in association with e-commerce companies in the upcoming days.

REFERENCES

[1] K. PRIYANGA and R. R. YOGESWARI, "Characterizing and predicting early reviewers for effective product marketing on e-commerce websites."

[2] B. K. Shah, A. K. Jaiswal, A. Shroff, A. Dixit, O. N. Kushwaha, and N. K. Shah, "Sentiments detection for amazon product review," *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6, 2021.

[3] M. V. Sagvekar and P. Sharma, "Study on product opinion analysis for customer satisfaction on e-commerce websites," 2021.

[4] B. Baiju *et al.*, "Describing and foreseeing early commentators for successful item showcasing on internet business sites," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 10, pp. 803–811, 2021.

[5] K. Dorthi and V. Prashanth, "A study on effective product marketing on e-commerce based on early reviews," in *2021 International Conference on Intelligent Technologies (CONIT)*. IEEE, 2021, pp. 1–6.

[6] M. Rajesh *et al.*, "Study on product opinion analysis for customer satisfaction on e-commerce websites," *Recent Trends in Intensive Computing*, vol. 39, p. 284, 2021.

[7] S. C. NUNE, "Characterizing and predicting early reviewers for effective product marketing on ecommerce websites," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 14, pp. 2870–2877, 2021.

[8] T. Bai, W. X. Zhao, Y. He, J.-Y. Nie, and J.-R. Wen, "Characterizing and predicting early reviewers for effective product marketing on e-commerce websites," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2271–2284, 2018.

[9] B. SUMATHI and M. TILAK, "Characterizing and predicting early reviewers for effective product marketing on e-commerce websites."

[10] Z. Liu, "Yelp review rating prediction: Machine learning and deep learning models," *ArXiv*, vol. abs/2012.06690, 2020.

[11] M. Hawlader, A. Ghosh, Z. K. Raad, W. A. Chowdhury, M. S. H. Shehan, and F. B. Ashraf, "Amazon product reviews: Sentiment analysis using supervised learning algorithms," *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, pp. 1–6, 2021.

[12] A. D, "Sentiment classification on tweets for event detection via clustering and filtering framework," *International Journal amp; Magazine of Engineering, Technology, Management and Research*, vol. 5, no. 5, p. 56–62, 2018.

[13] H. H. Kumar, Y. Gowramma, S. Manjula, D. Anil, and N. Smitha, "Comparison of various ml and dl models for emotion recognition using twitter," in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE, 2021, pp. 1332–1337.