An Efficient Research of Autoranking of Amazon Research using Regression Models

Anshul Rawat, Neetesh Gupta

ABSTRACT--- Before providing services to customers it is very important to know about the requirements as well as the services or products we are providing them contains opinions of other customers in which manner like-it is a positive review from the customer or the negative review. Since, an opinion plays a very important role in purchasing anything. There are some sites running online for the purpose of providing goods to the customers also they focused onto taking the decision over the posting reviews whether it is a positive response or negative. The motive of the work is to analyse the social data or products reviews simultaneously, and then create a model that will automatically create a model for product review. This paper bringing the continuous audits from a web based business website amazon and apply different content mining methods to pre-process the information and afterward apply an AI approach through which results will assess the viability of surveys through an outstanding measure for decency of fit. In this paper a development model with a computational cost model is utilized. The improved cost model with the word handling and positioning is utilized in given research.

Keywords- Artificial Intelligence, Helpfulness Index, Goodness of Fit, Word Count.

INTRODUCTION

The present era is considered to be an era of internet where all tasks are done by the help of internet such as- online shopping, electricity bill payment, book tickets, net banking etc. So, online shopping makes it easier to shop for the desired product or anything else through internet and not by visiting the stores [4-8]. By the help of online shopping one can easily compare the products online in a specific time period while stores may require more time in order to look for the desired products as well as there are some drawbacks also included which are to identify the quality of the products purchased. But now a day's online review of the particular products helped a lot about the products properties and also the satisfaction of the customers. For these reasons Amazon (Shopping website) has created a vote indicator that helps the consumers to rate or to give reviews over that product in order to improve the purchasing rate of other consumers for that product [8-10]. Still Amazon contain a facility for those items which have more reviews because only old and some helpful reviews are showed at the top of all reviews for products. This paper specifically uses Machine Learning algorithms that will attempt to sort the reviews on the basis of their certain features which are suitable for string data and ranking. The proposed algorithm will analyse the review portion which contains enough visibility as well as high number of votes. This paper focuses on building a model that will predict helpful reviews with few or no votes [12-18].

II. LITERATURE REVIEW

Here we can see that many related works that are being done on the reviews of the Amazon products. In the paper [1] author proposed work which uses the binary classification method in order to find out that the given review are helpful or not.

In the paper [2] the author worked on extracting some features from the review texts and then elaborate its performance with different classifiers such as- Naïve Bayes and Support Vector Machines that contains many kernels in this way author achieved the accuracy rate till 72%. Malhar Anjaria used different models for one product that exist in Amazon also observed, a single word may perform better in the form of predictors on the basis of the quality of the products.

Another study done by author in paper [3] uses supervised learning techniques for the generation of the scores for the rank reviews. So that the reviews which were of use were taken from the positive votes and the negative votes received from various consumers this information was then used for the training of the regressor.

One more analysis performed by Minna et al- in this study the author does not try to tune with hyper parameters in order to provide optimization for some of the used classifiers. Here, one fact is noticeable that after recognizing the effects of the regularization the variance of the regressors can reduce the variance in order to provide better performance.

C.S urendhranatha et al suggested that now a days Indian online marketing is becoming more famous and usefull. The speciality of e-commerce market are undergoing drastic change as the technology enabling customers to be more vigilant and logical while making purchase decisions. Online feedbacks and ratings are actually part of marketing strategies of online sellers to encourage the customer to purchase products. Consumers take reference of these

Revised Manuscript Received on August 14, 2019.

Anshul Rawat, M.Tech, Research Scholar, Dept. of CSE, TIT-S Bhopal, Madhya Pradesh, India.(Email: anshul.rawat5@gmail.com)

Neetesh Gupta, Prof. Head of Dept. (CSE), TIT-S Bhopal, Madhya Pradesh, India. (Email: gupta_neetesh81@yahoo.com)

reviews and ratings while they purchase products online. He focuses on understanding the effect of online reviews on purchase decisions of customers and to study the drivers of their involvement in online reviews. Exploratory research design is opted for the purpose of the study and survey was conducted among college students in Hydrabad. [5]

Zan Mo et al observed that, in order to understand the value of online reviews on custmer purchase aspect, more than 300 Taobao shops' online reviews are gathered. According to S-O-R model (Stimulus-Organism-Response Model), this paper focues the influence on custumer purchase behaviour according to online reviews of experience goods from a new perspective of custmer learning.[6]

III. PROBLEM DEFINITION

There are some problems arise in the existing work which are as follows-

1. *Unrelated data:*
   this was the major issue when large no. of reviews that are not related to particular product is available on the shopping websites.
2. *Data utilization:*
   because of lack presence of tools which are required to monitor large data was also an issue.
3. *Increasing content:*
   if the no. of websites will increase then the content management will also difficult task.
4. In Shopping website such as Amazon, most of the reviews that exist in Amazon have either very few or zero votes. Especially for popular product that have a high number of reviews, it's difficult for newly written reviews that are helpful to be read by people, since only the old and helpful reviews show up on the top of the review list for those product.

IV. PROPOSED WORK

Optimized Rank Cost Model (ORCM):

In order to work with the extension work, a proposed algorithm using optimized ranking cost model using an advance mathematical equation and co-relation has been presented.

Algorithm outline steps:

- There is a need for both predictive helpfulness index between the current reviews and new arriving reviews.
- *n* Reviews, all of them are of same size and can therefore be interchanged with each other and thus similar community redundancy by checking similar review.
- t_{hi} ... transp. intensity, i.e. cost between existing reviews and incoming Reviews i
- d_{ij} ... distance between *i* input and Reviews currently available.

Also need Jaccard distance computed from earlier approach Jdij (Jaccard distance).

Decision Variables

If T*list*→Review listing *all values*
and *i*→ input *first value which is giving as new review*
j →comparison review *value to incoming value*

Transportation intensity Thij = Intensity = Working with transportation intensity of each Review traversal. Computing word frequency in review itself and finding it uniqueness strength.

Jdij = j**accard Index = (the number in both sets) / (the number in either set) * 100**

The same formula in notation is:
X is weight value of input review and Y is the weight review of compared review.

J(X,Y) = |X∩Y| / |X∪Y|

X1, x2 is weight min value and max value weight of input incoming review and Y1, y2 is the weight min, max value of review being compared.

Transportation cost per unit transported from one input to i *to* process individual Reviews *j*

$$x = \text{Jdij} + \frac{\sqrt{((x2-x1)*(x2-x1) + (y2-y1)*(y2-y1))Thij}}{xy}$$

Total transportation cost from one entry to all other entry in data group-

$$x = \sum_{Jd=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{th=1}^{n} Jd_{ij}t_{ij}d_{ij}(x_{ij}y_{ij})$$

This approach helps in minimizing the total cost of processing individual entity. Next, will explain three models: linear regression, ridge regression and support vector machines for regression**.**

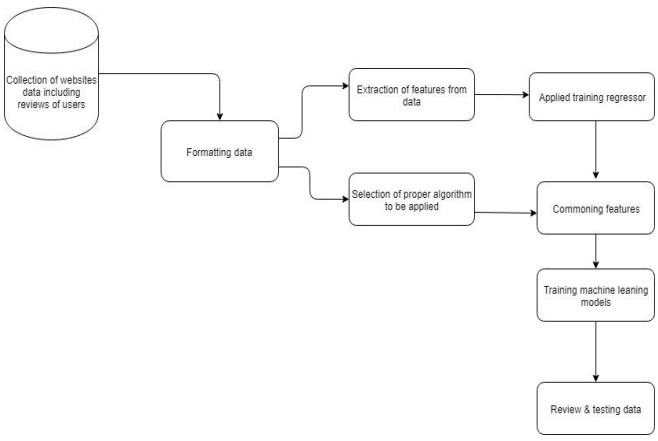

Figure 1: Work flow of the system.

In the above figure 1 the flow of the proposed algorithm has been shown in order to generate the meaningful outputs.

Step 1: it includes the collection of the data from the Shopping websites that contains reviews which will further use for generating outputs as desired.

Step 2: in this step the filtrations of the data as per the need will be done.

Step 3: in this step the meaningful features from the users reviews will be extracted.

Step 4: choose the best appropriate algorithm.

Steps 5: in this step the training regressor will be applied on the extracted data.

Step 6: in this step some features are now sort out on the basis of their related features.

Step 7: apply suitable machine learning algorithm.

Step 8: now, in the final step the review of the data will be done.

V. METHODOLOGY AND WORK DESCRIPTION

A. Dataset

This shows the usage of set of Flipcart reviews written within the date range of June 2008 - August 2012, in particular for the mobile phones and accessories category. Each review within the set contains information about the reviewer ID, product ID, review up votes and down votes, the review text itself, the rating which reviewer gave to the perticular product, the title of the review, and the time the review was written. The original dataset contained 1.25 million reviews. In order to able to extract enough number of trustworthy features and reflect the objective of this study better, filtered the review data with the following criteria:

(1) each review needs to have more than 12 votes (thereby having enough visibility), and

(2) the review should exist in popular products (with more than 16 reviews). By using these criteria, will reduced the data set from 1.25 million down to about 30,000 reviews.

Then randomly selected 80% of the data as the training set, and the other 20% as the test set.

B. Outcome Variable

Borrowing the idea from the study done by Kim et al [3], will define the following quantity to provide a measure of the sense of helpfulness a review provides given the i-th review:

$$Y^{(i)} = \frac{upvotes^{(i)}}{upvotes^{(i)} + downvotes^{(i)}}$$

Originally, it was a thought of defining the outcome variable to be the difference between upvotes and downvotes of a particular review. This definition, however, doesn't differentiate the case between a review that has 104 up votes and 99 down votes versus a review that has 5 up votes and 0 down votes, although one would think that the first review could have been less helpful due to the review having down voted by people. To visualized the data to get a better understanding of it. The histogram of the review counts grouped by the outcome variable indicates skewness in the data. Approximately 68% of these views have helpfulness scores between0.7 to 1.

C. Features

It used the following features extracted from each review text, classified into broad categories.

Textual Features

• Text Length - The count of characters in the review text, including punctuation and spaces.

• Character Count - The number of alphabetical characters in the review text

• Word Count - The number of words in the review text.

• Unique Word Count - The number of unique words in the review text.

• Sentence Count - The number of sentences in the review text.

Metadata Features

By using the number of stars that a reviewer has given a product in his/her review as part of a feature set.

Bag of Words

Also used the bag of words model to generate additional features for each review. The bag of words model uses a large- dimensional, sparse vector to count word occurrences in a review text with respect to some vocabulary of words. More precisely, let x (i) ∈ R n be the sparse vector containing word occurrences in the i^{th} review, where n is the size of a vocabulary. If review i contains c occurrences of the j^{th} word in a vocabulary, then $x_j^{(i)}=c$.

VI. TOOLS & RESULT ANALYSIS

Tools

- JetBrainsPyCharm Community Edition 2017.2.1 x64
- JetBrainsWebStorm 2017.3 x64
- Jupyter Notebook

Libraries

- Pandas
  - Used for the formation of dataset, as required for the machine learning model
- NumPy
  - Used for the creation of compressed arrays
- Scikit Learn
  - Used for implementing the inbuilt machine learning models and utilizing them according to the needs of the project
- MatPlotLib
  - Used for the visualizing and plotting the datasets
- ReactJs
  - Used for the designing the user interface

RESULT ANALYSIS

Regressor	Training Accuracy	Testing Accuracy
Linear Regression	**19.76 %**	**19.62%**
Ridge Regression	**19.85 %**	**19.72%**
Support Vector Regression	**84.99 %**	**32.48%**
ORC Model	**89.46 %**	**48.90%**

Table 1: Comparison analysis.

In the above table 1 the comparison has been shown in between the proposed and existing algorithm.

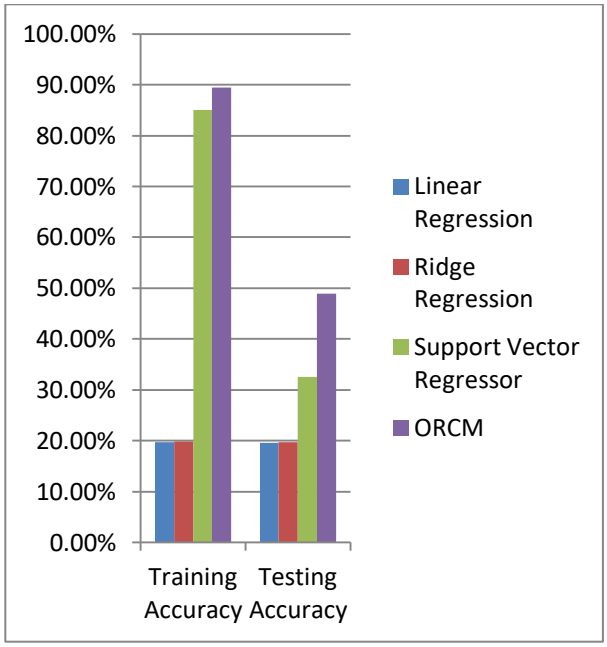

Figure 2: Graphical analysis.

The above is the graphical analysis of the algorithms used.

The given research discussed about the existing regression model and their technique. A proposed linear regression formulae is implemented and proposed by the given research.

Result obtained from the implementation of proposed cost function shows the efficiency in terms of training and testing accuracy.

VII. CONCLUSION & FUTURE SCOPE

The results shows that it is possible to automatically evaluate judgments. Analyzing the problem as regression and extract properties that represent the data. By using different regression methods and compared their properties. The obtained results found that optimized rank cost model has achieved the good results. Due to the complexity of the upcoming issues, a believe is that there is plenty of scope to improve upcoming work by adapting the three important designs: developing a bit complex function, increasing size and achieving victory. The text features here used were very basic forms, all about long-term properties and readability. Other useful features in natural language, such as sending the analysis. Bag-of-words can also be improved by using the terms Frequency Inverted Document Frequency (tf-idf) or the Google-developed library word2vec [9], which continues with a continuous or continuous N-Gram model. Another feature may be the location of evaluator. The future work will include the higher chances of making more helpful reviews. The feature space is comparatively high dimensional due to which it is necessary to have a big data set in order to get accurate results. Also aware of the fact that some of the features are highly inter related. To focus on these problems, feature selection methods or principal component analysis can be applied. Atlast, skewness has a huge effect on the efficiency of models. Especially, optimized rank cost model are known to be susceptible to skewness of data, as studies have indicated [10].To sort this issue, simple techniques such as up sampling, down sampling, or stratified sampling of the data can be employed, while the study mentioned also provided another entire methodology to mitigate skewness. wqe

VIII. REFERENCES

1. J. Rodak, M. Xiao, and S. Longoria, Predicting supportiveness appraisals of amazon product reviews, 2014, Accessed: 2015-11-08.
2. S. Bolter, Predicting item audit support utilizing AI and particular characterization models, 2013. [Online]. Accessible.
3. S.- M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, Automatically evaluating survey support, in Proceedings of the 2006 Conference on observational strategies in normal language processing, 2006, pp. 423–430.
4. J. McAuley, R. Pandey, and J. Leskovec, Inferring systems of substitutable and reciprocal items,. New York, NY, USA: ACM, 2015, pp. 785– 794. [Online].
5. Impact of Online Consumer Reviews on Consumer purchase Decision in Bangalore. International journal of Allied Practice, research and Review. Vol. IV, Issue III, March 2017, p.n.01-07.
6. Effect of Online Reviews on Consumer purchase Behaviour. Journal of Service Science and Management, 8, (2015), 419-424.
7. Shruti Kohli, Himani Singal, “Data Analysis with R” in 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing.
8. Arun Jalanila, Nirmal Subramanian, “Comparing SAS® Text Miner, Python, R” in 2016 IEEE International Conference on Healthcare Informatics.

9. Anjali Ganesh Jivani , A Comparative Study of Stemming Algorithms, International Journal of Computer, Technology and Application, Volume 2, ISSN:2229-6093.

10. V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60–76, 2009.

11. W. He, "Examining students online interaction in a live video streaming environment using data mining and text mining," Computers in Human Behavior, vol. 29, no. 1, pp. 90–102, 2013.

12. R. Agrawal and M. Batra, "A detailed study on text mining techniques," International Journal of Soft Computing and Engineering (IJSCE) ISSN, pp. 2231–2307, 2013.