Product Reviews Classifications-A Data Science Approach

Nitin B. Raut, Premkumar M., Kharthikeyan S., Mohan M., Vijayganth V. and Srithar S.

Department of Computer Science and Engineering
KPR Institute of Engineering and Technology, India

cen.nitin@gmail.com, premkumarmurugiah1395@gmail.com, mailkharthik@gmail.com, m.mohansan@gmail.com, kv.vijayaganth@gmail.com, srithar.s@gmail.com

Abstract - Reviewing the product is an important step for e-commerce platforms. Getting reviews from customers and analysis of reviews consumes many resources. As the number of reviews received day by day is increasing very rapidly, reviews should be classified as fake review and Genuine reviews. The total accumulation of reviews and analysis is different from natural language processing problem. Spammers are hired for biased reviewing of products. In this paper we put a novel comparison between purchase list and reviews. We have applied a method for finding duplicate reviews; measure the total numbers of reviews and their mismatch in counts, at the end count dispersion for every product and classification of reviews. We applied data science approach for classification and visualization to get fake reviews. We label the reviews either positive or negative based on comparison between them. A data science approach is applied because for a well-known product the reviews can go up to many thousands.

Keywords - E-Commerce, Data science, Fake review, Genuine review.

1. Introduction

While purchasing anything offline as well as online, we always try to get what others have bought and their experience about product. That acts as an important source of information for us to make a choice. World Wide Web (WWW) has enormous extensions in the E-commerce; many products are sold online through website. The online merchants will always focus on customer satisfaction requirements and the user friendly shopping culture to enable customers to express their views on the product, web is being used by many common users efficiently, many people are writing reviews and publishing them which are very informative for others, as an output there is tremendous growth of product reviews. Many fraudulent cases have been identified [1] where spammers were allowed to flood the reviews. Some products can get many numbers of reviews at specific merchant site. We implement a procedure for analysis and visualizing, the reviews and helping in decision making for purchasing.

2. Literature survey

Various approaches with dictionary methodologies has been proposed in the field. [2] proposed spam detection and opinion mining by using Naïve Bayes. Some researchers use POS tagging where POS tag for each token is stored in database. Establishing mapping between data is very important in data science approach many researchers worked on this area rellim, apriori etc algorithms are developed.

3. Proposed Framework

The flow of execution in our framework is duplicate reviews detection and source mapping, anomaly characterization in count of reviews and its rating distributions. Along with incomplete and incentivized reviews are detected to check the credibility of the review rating. And the overall procedural outcome showed in proper visualization.

Duplication of Review Identification and Source Mapping

Supervised learning model is not feasible in real time, because ground truth is not always labeled. Fake reviews are most of the time negative. Therefore, we have applied an unsupervised approach in labeling the result with the source of reviews. Many researchers published their work by forming combination of words, we have applied same technique. The detection of fake reviews is explained as follows in Figure 1.

Fig. 1: Flow chart for fake review detection

Suppose we have example of review like, “It’s one of the best products” combining words module will generate 3 bi-grams viz. it’s one of the best products. It reduces the burden of tokenization considering each word along with contextual relevance is studied. Thus, the intersection between the sets of words gives similarity. We can use many similarity measures for this. Similarity measures by Jaccard are useful as these are used for data objects.
represented as two groups. Threshold is set to consider duplicate reviews, we set 75% as a threshold, greater than [3] used. However, optimization techniques can be applied to facilitate the large data set with more similarity. Hashing is the best way to minimize the representation of these large data sets [4]. This Hashing values prevents us from computing permutation on each sets of words. Minimum hashing and Cyclic Redundancy Check (CRC-32) is used in our implementation. The inverted index is used [5] as it is widely used strategy in information gain. We build an index-dictionary with all Minimum Hash values of product; therefore, time optimization is done.

Anomaly Characterization in Review count and distribution

Each product gets reviews and ratings progressively for a period with randomly time breaks [8-10], spam review is characterized on the basis of sudden increase in reviews and rating changes in small duration of time, the rating of products may get extreme value because of it the effects of spammer attack is thoroughly studied by [6]. Such spikes are considered as anomalies in seasonal products and non-popular products. Therefore, anomalies detection is important step. The Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD) algorithm is very useful in detection of anomalies with growth and seasonality. Large data sets can be effectively processed by this algorithm. It efficiently finds sudden sharp increases, abnormal picks, and unusual high activities as shown in Figure 2.

Identification of Incentivized Reviews

Many sellers offer discounts and some other products as a gift, for more promising reviews. These are termed as incentivized reviews; some reviews are biased in nature. Natural language toolkit–Wordnet labels the semantic relationship between the words also it is universally accepted, the dictionary of similar meaning of both single word and doubly pairing words is created. The regular expression is used to find the synonyms as shown in Figure 3. The time duration for incentives is also recorded by algorithm developed by [7] we are implementing data science for large data sets as many media houses, marketing gurus and human psychology comes in picture.

```
Fig. 2: Anomaly detection
```

```
Distribution of reviews and ratings
Storage Time
S-H-ESD Module
Anomaly detection
```

```
Fig. 3: Identification of incentivized reviews
```

Identification of incentivized reviews

Many sellers offer discounts and some other products as a gift, for more promising reviews. These are termed as incentivized reviews; some reviews are biased in nature. Natural language toolkit–Wordnet labels the semantic relationship between the words also it is universally accepted, the dictionary of similar meaning of both single word and doubly pairing words is created. The regular expression is used to find the synonyms as shown in Figure 3. The time duration for incentives is also recorded by algorithm developed by [7] we are implementing data science for large data sets as many media houses, marketing gurus and human psychology comes in picture.

```
Fig. 3: Identification of incentivized reviews
```

Visualization

A statistical approach is used to maintain the score of each product. Depending on score we can visualize the result as shown in Table 1. Each product category is assigned an average scoring scale, the list of the products along with reviews recorded. Duplication in reviews, spam and incentivized reviews are stored, anomalies in reviews are also calculated as discussed above. After this each product’s scores are compared for suitability against the average score scale category, average ratio of duplicate reviews can be estimated for product category. A colored circle scheme is used to visualize the overall result.

```
Table 1: Visualization scheme
```

```
| Score | Color Code | Description     |
|-------|------------|-----------------|
| 0     |            | Non Alarming    |
| 1-2   |            | Moderate Alarming|
| >=3   |            | Extremely dangerous|
```

Efficiency

Our approach is very simple and can be scalable with any data. it is cloud based application approach, therefore very storage efficient and robust. The admin can easily notice the result and find out reviews classification. The admin can delete the fake reviews.
4. Conclusion
We conclude with a method for detection of spam reviews. We applied duplicate review identification, anomaly characterization in review count and its distribution, incentivized review detection and rating distribution, these serve as platform to get fraudulent reviews to get effective visualization, we have proposed this algorithm for big data sets with large number of product categories which will give more efficient and robust system.

References
[1]. D. Meyer, Fake reviews prompt belkin apology. Cnet News (2009).
[2]. C. Miller. Company settles case of reviews it faked. New York Times (2009).
[3]. D. Streitfeld, for $2 a star, an online retailer gets 5-star product reviews. New York Times 26 (2012).
[4]. N. Jindal, and B. Liu, Review spam detection. In Proc of the 16th International Conference On World Wide Web, ACM, pp.1189-1190, (2007).
[5]. A.Z. Broder, On the resemblance and containment of documents. In Compression and Complexity of Sequences 1997. Proc., IEEE, pp. 21–29, (1997).
[6]. S-H-ESD. Anomaly detection with r code access: https://github.com/twitter/Anomaly Detection.
[7]. A.Prabakarn, Min Chen, Product Review Credibility Analysis, IEEE Workshop on Comp., Networking, and Comm. (CNC) 2019.
[8]. Kumar, S. P., Sumithra, M. G., & Saranya, N. (2019). Artificial bee colony-based fuzzy c means (ABC-FCM) segmentation algorithm and dimensionality reduction for leaf disease detection in bioinformatics. The Journal of Supercomputing, 75(12), 8293-8311.
[9]. Karthick, A., Manokar Athikesavan, M., Pasupathi, M. K., Manoj Kumar, N., Chopra, S. S., & Ghosh, A. (2020). Investigation of inorganic phase change material for a semi-transparent photovoltaic (STPV) module. Energies, 13(14), 3582.
[10].Suriya, M., Chandran, V., & Sumithra, M. G. (2019). Enhanced deep convolutional neural network for malarial parasite classification. International Journal of Computers and Applications, 1-10.