Spam Review Detection
Smriti Rastogi¹, Prathiksha. M URS²
Department of Computer Science and Engineering
The National Institute of Engineering, Mysuru, Karnataka, India

Abstract:
Online shopping has become a convenient method for purchasing, be it a car booking or buying simple grocery items. Customers often read online reviews written for a product to decide whether they will purchase the product or not. A high number of positive reviews will increase the sale, whereas negative reviews will deteriorate the sales. Since the customer reviews have so much impact on the business of a product, spam reviews are written to promote or demote products or services. Consequently, manufacturers and dealers are worried about client reviews as these directly affect their organizations. In late years, the spam review issue has increased a lot of consideration from specialists, yet at the same time, there is a need to perform an investigation on true huge scope datasets. In this work, we use spammers’ behavioural features to ascertain the spam score which is then used to distinguish spammers and spam reviews. Trial assessments are led on a true Amazon review data set which breaks down 198,502 reviews and the assessments show that the proposed model has significantly improved the recognition procedure of spam reviews. Specifically, behavioural techniques accomplished 98% precision.

Keywords: Online shopping, Online reviews, positive reviews, negative reviews, promote, demote, spam reviews, spammer behaviour features, spam score, spammers.

1. INTRODUCTION
As the Internet keeps on developing in both size and significance, the amount and effect of online reviews constantly increments. Reviews can impact individuals over an expansive range of enterprises, yet are especially significant in the domain of web-based business, where remarks in regard to items and administrations are regularly the most advantageous, if not by any means the only, route for a purchaser to settle on a choice on whether to get them. In the writing, spam reviews have been sorted into three gatherings, which are essential (1) Untruthful reviews- false opinion on products, (2) Reviews on Brands - where the remarks are just worried about the brand or the vendor of the item and neglect to review the item, and (3) Non-reviews those reviews that contain either inconsequential content or commercials. Detection of type-1 review is the most difficult, as a human appointed authority it is hard to unhesitatingly learn which review is phoney and which is valid. There are no reasonable signs or signals from the content that shows if a particular review is true or not. A data scientist should try to extract certain features from the text when preparing information mining and AI calculations to discover these highlights in the review that will decide whether it is genuine or counterfeit. Data mining and AI methods, principally those for web and content mining, offer an energizing commitment to distinguishing false reviews. It is essential to make reference to that while most existing AI methods are not adequately viable for spam discovery, they have been seen as more dependable than manual recognition as humans do not have the capability to process so many features quantitatively. The essential issue, as is the absence of any distinctive words that can provide a conclusive insight for the order of reviews as genuine or fake. A typical methodology in content mining is to utilize a bag of words approach where the nearness of individual words or little groups of words are utilized as highlights. But a few investigations have discovered that this methodology isn't adequate to prepare a classifier with sufficient execution in review spam. In this manner, extra techniques for feature engineering must be investigated with an end goal to improve review spam detection identification techniques.

2. RELATED WORKS
The first experiment in the field of review spam detection was done by Jindal and Liu[1] in the year 2008, since then a lot of research has been carried out on this subject where earlier methods took the review text into account by calculating the content similarity between two reviews using n-gram based comparison methodology[1] or probability-based language models[2]. As it is a Binary classification problem a lot of features and metadata of reviews was used [3-7]. In such classification methods both linguistic and behavioural features like rating deviation, number of reviews, review length etc are taken into account for building classifiers. Apart from this, researchers also explored unsupervised learning methods with Some positive and unlabelled data. In [8] researchers developed three methods for categorizing reviews as spam, content-based approach gave almost an accuracy of 90 percent. With Support Vector Machine algorithm and bigrams an accuracy of 89.6 percent and combining LIWC software with bigrams, we obtain 89.8 percent accuracy.

In the same research paper, Human observers had maximum accuracy of almost 62 percent which is less than the SVM classifier. Recently researchers have done experiments to find spammer groups (groups of people with a common motive of demoting or promoting a product). In [9] a top-down computing framework, GGSPAM, graph-based technique is used to identify spammer groups. In [10] researchers have exploited frequent itemset mining (FIM) to generate candidate spammer groups.

IJESC, June 2020 26271
III. METHODS

EXISTING SYSTEM
E-commerce websites generally do not flag reviews spam or not spam. Only a few websites like Amazon and Lenskart provide a “Verified Badge” for some reviews. To get the badge websites check if the product is purchased with the same account by the user, but it does not say if the reviews are spam or not. At present, no online shopping websites have an inbuilt classifier to label reviews as spam or not. The customers have to go through all the reviews and decide whether to purchase the product or not so review spam detection is based on the analytical ability of an individual. Manual review spam detection is not efficient and time-consuming.

PROPOSED SYSTEM

A spammer can be distinguished by investigating its diverse behavioural features. The unlabelled dataset can be utilized with unsupervised learning to recognize the spam reviews. The proposed technique takes unlabelled dataset and produces a yield of a labelled dataset that recognizes spam and non-spam reviews. The procedure begins with the identification and estimation of spammer behavioural features on amazon reviews dataset, then it computes the standardized worth (01) of every feature. Next, the average value is calculated for each review. Each feature is dropped and if the value of average changes by more than 5 percent then that average is considered to be important and given a weight of 2 else 1. Later, Spam score is calculated for each review as:

Spam score = w1x1+w2x2+w3x3……..wnxn(w-weight of a feature, x- normalized value for a feature, n- number of features)

The calculated spam score is compared with Threshold values (0.5,0.55,0.6) to decide if the given review is spam or not. The labelled dataset is then used with supervised classification algorithms to build a model that can test new reviews.

DATASET
The dataset used for this experiment is taken from Amazon website. It has 198,502 reviews of cosmetics in binary format. It contains the following columns-reviewer ID, ASIN (Amazon Standard Identification Number), Reviewer Name, Helpful (number of people who found the Reviews helpful), Review Text, overall (ratings given by the reviewer), Summary and Date. These columns are used to extract new features.

FEATURES
In machine learning, feature extraction is the technique of starting with some initial set of measured columns and building features that are more informative and less redundant. Our dataset has 198,502 reviews and we calculate these features for each of the reviews. (1) Review Similarity- Cosine similarity to find the similarity between two reviews. (2) Maximum number of reviews ratio – It is the ratio of a total number of reviews written by a user divided by the maximum reviews posted in any of the days before the current review date. (3) Review Count – Spammers usually make accounts to write fake reviews for a product they are targeting so the number of reviews is generally less. This feature calculates the count of reviews written by a reviewer. (4) Rating Deviation- In this feature, we check how much a rating given by a particular reviewer differs from the average rating of the product. (5) Ratio of positive reviews-It decides the nature of the reviewer if he is a promoter or not, so we take the ratio of positive reviews by the total reviews. Reviews with rating 4 and 5 are considered to be positive reviews. (6) Ratio of negative reviews- It decides if a reviewer writes negative reviews in general, so we take the ratio of negative reviews by the total reviews. Reviews with rating 1 and 2 are considered negative. The further nature of reviewer is captured by (7) Extreme Ratings. Spammers have the habit of not putting much effort in writing reviews and are very clear with either promoting or demoting the product so they give either extremely low or extremely high ratings. This feature takes the ratio of reviews with extreme ratings by the total number of reviews. (8) Review length -total number of characters in a review and (9) Ratio of capital characters- It is the ratio of the number of capital letters divided by the total characters in a review. (10) Single product reviews Spammers often focus on a single product, this feature calculates if the reviewer has just written reviews for one product. (11) Account activity days- Spammers do not remain active on one commercial website for long. Usually, the time is less than 45 days. This can be calculated as the difference between the times of first and last review. (12) Ratio of First Reviews – Spammers are often the first review writer of many products so that they can affect the business of a product, we calculate the ratio of first reviews by the total reviews.

IV. RESULTS AND DISCUSSION
After labelling the dataset using behavioural features, we used classification algorithms to build a classifier that can successfully detect correct labels for the test dataset. Naive Bayes classifier gave an accuracy score of 0.91. A Decision tree classifier worked better and the accuracy obtained was 0.97. The best results were obtained for Logistic Regression with an
accuracy of 0.98. If we take threshold as 0.5 for spam score the results obtained are as shown in the below figure.

![Accuracy of logistic classifier](image)

**Figure.2. Accuracy score of classifier for threshold 0.5**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Spam Score Threshold</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.5</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>0.55</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.91</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.5</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>0.55</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>0.65</td>
<td>0.97</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.5</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>0.55</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>0.65</td>
<td>0.98</td>
</tr>
</tbody>
</table>

**Figure.3. Accuracy for various classifiers**

V. CONCLUSION

We like to conclude that with the development of our model, spam reviews can be detected with an accuracy of 98 percent. Although a lot of research has been done on this topic with various methods, still e-commerce platforms do not flag reviews as spam or not. The existing system is still manual and based on the intuition of the person reading the review. Since human beings do not have the ability to process numerous factors quantitatively and assess the review, the proposed system takes care of this by generating new features, which are weighted according to the amount of percent change they bring in the average spam score. Our system can be integrated with e-commerce websites to flag such reviews as false. In future, if a dataset with more features like timestamp or location is available, then we can also find spammer groups in a particular area and make the model better.

VI. REFERENCES


