***Spam Ham Classification of Online Reviews***

Apoorva Godasu1, Bhargav Allampally2, Chandana Yenugu3 ,Swetha Nallamalli4

***Abstract***

***In today’s rapidly changing world, most of them are inclined towards trying out new products. In this busy world, customers consider going through the reviews first. For example, on any online website like Amazon or Flipkart, there is a review section that comes along with each product. These reviews are given by other customers who bought the product before. These could be positive, negative or neutral concerns towards the product. A new customer who wants to buy would consider these reviews while deciding about the product. Now the question is, are these reviews veracious in nature? What if these reviews are fake which are given by the product sellers or any third parties associated with it? What if the reviews are posted by the opponents to decline the product sales or ratings? Therefore, it is crucial to filter out the genuine reviews from such fake ones. In this paper we suggest a method to classify the online reviews into spam or ham based on the sentiment of each of them. Yelp dataset is used in this project.***

**INTRODUCTION**

The use of internet and World Wide Web has been increasing enormously day by day, with the ease of accessing the internet from any corner of the world, the rate of online purchases has increased rapidly throughout the years. [1]A survey shows that 209.6 million people from the US are online shoppers by the year 2016 and have surfed and compared the prices and bought products online at least once. If we compare the current situation , there has been a sudden peak in online shopping not just through the years but also due to the unexpected pandemic situation that we are going through. The pandemic itself has brought a situation of staying indoors , this has introduced online shopping to a wider range of population. New customers have started using these online platforms and the existing customers have increased their purchases from the regular curve. With such intense online shopping , online reviews play an important role in the decision making and hence classifying genuine reviews and unreal reviews is vital.

Many retailers would post fake reviews to increase the value of their product via sales, credibility etc. The fact of humongous energy, time and more valuable resources are being invested in writing or generating fake reviews. Fake review spamming conveys how the number of reviews add on value to the product. [2] A research conducted by BrightLocal says that 74% of consumers have read a fake review in 2018, among which the 18-34-year age group was involved on a higher side with 92% stating the same.

Most of the online retailers have a system to handle these fake reviews by tracing out the IP address or by the location. These systems were good at filtering out the fake reviews but classifying the reviews to hold back the reviews by the real customers was a challenge.

**LITERATURE SURVEY**

Previously, ample amount of work and research has been invested to detect fake reviews using sentiment analysis. In Lin et al. the occurrence of duplicate reviews was measured as spam reviews, and with the help of the cosine similarity, the reviews were classified further. Even though cosine similarity works well with spam detection, not all reviews are built the same way. In addition to this, this model is too specific and is not applicable on any dataset other than its training dataset.

[3] Also in another proposal, the characteristics of fake reviews were analyzed. Based on review contents and reviewer behaviors, six-time sensitive features are proposed to highlight the fake reviews. And then, supervised solutions and a threshold-based solution to spot the fake reviews as early as possible were devised.

A substantial volume of training data is very essential, and the gold standard dataset is one such dataset. This contains annotated reviews; this indeed helped a lot of researchers in taking supervised learning approaches which in turn increased the accuracy. On the other side of the coin, this data was generated by a bot and was not likely to be used for spam reviews which collected posts from numerous online websites where the real reviews are posted.

For any classification process, having a substantial volume of training data can demonstrate to be essential. One such dataset is the gold standard dataset which consists of annotated reviews. Many researchers were able to leverage this dataset in their supervised learning approach and push their accuracy. However, the fake reviews in the dataset is often constructed by a bot and are different from the spam reviews that are posted by people on various online websites.

This was considered as a problem of binary classification in Chengai et al.[4] which used a convolutional neural network (CNN) model that captured the product related review features by a linear composition of products and reviews, and then a bagging model that bags the CNN model was introduced with two efficient SVM models. They were successful at the contribution of classification strategy towards the result validity. On the other side, Jindal et al [5] established the importance of feature engineering apart from the classification methods. Usually sentiment analysis was done at word or sentence level, which is a subset of the entire review. When a part of the whole review is analyzed, it is inaccurate to arrive at the results as all sentences could be independent of each other, for instance, pros and cons of a product. Sentiment analysis at this level would result in wrong classification. Also, there was another approach where output of each sentence was compared with the average output by classifying based on the inclination, which was also not accurate enough to calculate the results precisely.

Singleton reviews came into picture for spam detection. According to Feng et al., 90% of the reviewers who posted only one review turned out to be spam. With an unsupervised model, 86% of accuracy was achieved and is stated as a behavioral model, and when they included Natural Language Processing techniques, the efficiency was further increased. Zhang et.al proposed a to extract the several verbal features in the dataset via Convolutional Neural Network. Additionally, the footprint and the metadata of the reviewer came handy for envisioning the data better.

**PROPOSED METHODOLOGY**

**ABOUT THE DATASET**

In today’s busy life the internet has brought a lot of handy ways in choosing a restaurant or a product. Yelp is one such site where various kinds of people give their reviews. Unlike in the previous days, how people used to ask people, read news and get to know about a place, these days the reviews are one click away and are in abundance. It is not limited to just the people whom we know, it is about the concern and review of a person from any corner of the world. This has provided a user, a quick and easy way to go through the reviews to decide on trying out new restaurants.

Yelp provides reviews of various businesses, and recommendations of best restaurants, entertainment, shopping, services, nightlife, food, etc. Knowing that Yelp has such helpful and useful data that can be used for multiple purposes like analyzing it to fetch a new remedy about it and this may help for further research and development about the user sourcing reviews websites, which in turn can help users to meet their requirements. Thus, yelp delivers a very narrowed though dependable set of data for both open source and academic purpose. We obtained this dataset from the yelp reviews dataset website whose link is provided in the references. Additionally, they also provide a vibrant explanation about the data provided.

This is a huge dataset with 10 gigabytes in size approximately. The reviews dataset alone is 3 gigabytes in size. But reviews of different types of businesses are present in this reviews dataset. Since the goal of our project was to classify online reviews into spam or ham, we used only the reviews. json and business. json file provided out of all the diverse datasets like reviews json, business.json, user.json, check in.json, tips.json and pictures.json. The overall dataset had about 8,021,122 reviews and 209,393 businesses.

**DATA PREPROCESSING**

Firstly, the data files were downloaded and then these files were uploaded into AWS S3 using AWS CLI commands. AWS CLI handles large files of data and hence, we went with AWS CLI to store these large files into S3 by creating a bucket first and then uploaded these files into it.

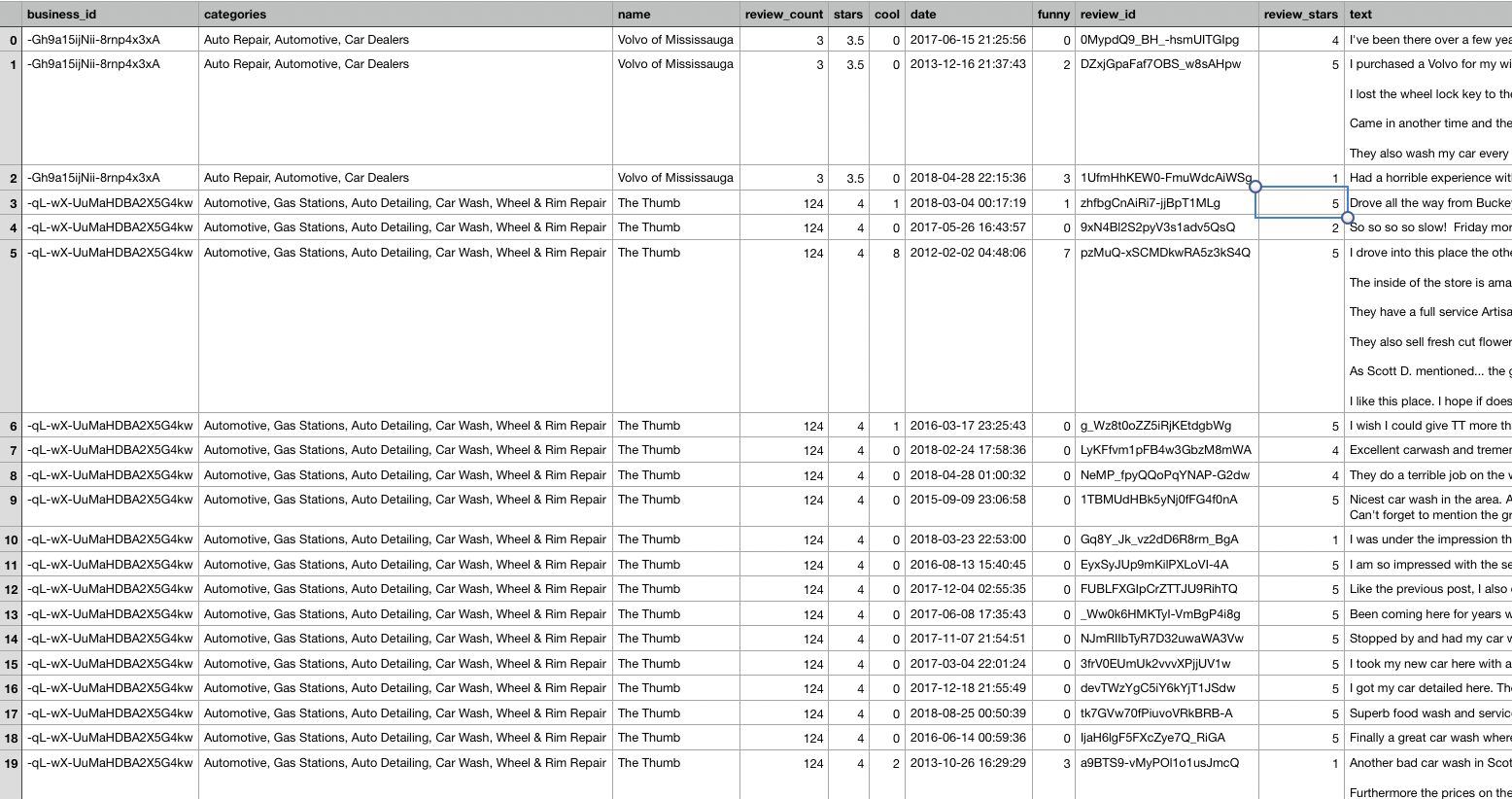
Using spark both the files, reviews.json and business.json are read and were then converted into dataframes. There were multiple columns in business.json file like business\_id, name, address, city, state, postal\_code, latitude and longitude, review\_count, starts, is\_open, attributes and categories and hours. But not all were needed, so we took only business\_id and is\_open fields to check if the respective business was still running or not. Because, reviews on running businesses made more sense than the ones which were not up and running. Later, to fine-tune the categories, we used the explode function to convert the vector of categories/types in the field called category into a separate column. We were able to extract 10 categories out of 1290 listed.

As a next step, we have selected the automotive category and combined the reviews of automotive from reviews.json along with the user\_id, review\_id, review\_stars , useful, funny, cool votes etc. Also, dropped all the unwanted columns from the dataset and converted the final json file into csv. This conversion was made to ease the upcoming work we had going further. The converted csv file was saved into S3 bucket.

**PREPARING THE DATASET TO FIT IN THE MODEL**

The preprocessed data is stored in AWS S3 in CSV format filtering out all the white spaces such that it would not cause any hurdle while reading and converting back to the dataframe.

Below is the snapshot of the csv file obtained after pre processing



For further work to progress, we chose Jupyter notebook on EMR cluster. The data was read into the data frame via sparkcontext and spark session. Then all the columns that were no longer necessary were dropped out. Then considering the business\_id, review\_date and the entire review, a new and unique id was generated for each review we had in place, making a path for these features to be dataset ready. There were numerous types of voting given by the customers, for example funny, cool, useful etc. We clubbed all these into a single column and added the total votes to a review. This processed data is provided to the algorithm we were using to classify the online reviews as spam or ham.

**PROPOSED SOLUTION/MODEL**

As per outlier analysis in statistics, an outlier is a data point which varies significantly from other data points present in the dataset. Based on this outlier analysis, our algorithm was built in such a way that we classified a review to be spam if it showed any traits of being an outlier. But there was no strong foundation for this from the dataset we had, and therefore we chose to use clustering to classify the dataset. But there was a need for appropriate features to fit into the model, to form clusters and hence to determine the outliers from there. Therefore, the features were built in such a way that they can be fed into a clustering algorithm and the desired output was obtained.

Using textblob, a library that returns the sentiment score, we performed sentiment analysis on the body of the review. The results were then stored back into a dataframe. Also, there were columns which showcased the review rating i.e., review\_stars, using this column we were able to find the average of review\_stars and rating for a given business\_id. As a next step the difference between the values of average rating and the sentiment score was taken to see how much the values were deviating from the average. Based on this deviation, it was easy to classify the reviews by tagging them as outliers. For instance, a restaurant was not that great according to 98% of the reviewers i.e., the average has been on the lower side and if there were a couple of reviews where the sentiment score was on a pretty higher side like the highest value, then these were considered as outliers and could be given by the sellers itself, so as to increase their ratings or to at least uplift their business. Therefore, these parameters have become the features for the clustering model i.e., distances from their average rating and sentiment scores along with the votes for each review.

Now comes the question of finding the outliers, for this it is crucial to know the behavior of the review within each cluster. The Gaussian mixture model came into the picture in this case, which is a soft k-means model. Maximum likelihood is the key for cluster prediction and gives the probability of how much it belongs to that cluster. Hence, the outliers were distinguished based on maximum likelihood which was rooted within the model. To classify a review to be genuine or say non-spam/ham, the review had to belong to a cluster with a minimum probability of 90%, which defines good intra cluster distance. As a result, this review is not considered as an outlier. Thus, all such reviews with more than 90% probability were marked as ham/ non-spam reviews. This has been the engine of our algorithm.

This was performed on various clusters and the results have been noted. The maximum iterations were set to 100, which was another crucial parameter. This was not enough to state how many clusters have been giving the good outliers. Hence, the next challenge was finding the accurate number of clusters say k. To figure this out, two other methods popped into the picture. The first one is elbow curve method which choose k value as the value at the elbow of the curve. Even though this suffices the need, we wanted to check the outliers based on the cluster formation. Another precise way was introduced by silhouette distance, whose score was based on the inter and intra cluster distance which helps in concluding on k value. The score is calculated based on Euclidean squared distance and cosine distance. Various values of k were found using the cosine distance and calculated the silhouette score. Based on this, we concluded if a review was a spam or ham.

**EXPERIMENTS AND RESULTS**

***A. Results***

The below are the results obtained from the above model. We found the silhouette score for k values between 2 to 8 and got results of various clusters on different k values. We got a good silhouette score for the k=2 number of clusters and all the remaining k values retrieved bad scores as considered and compared. Thus, we found that the best number of clusters required will be 2 and labelled a review as a non-spam if it contributed to if it belonged more than 90% to a cluster and as spam if it contributed to less than 90% to any of the two clusters else.

|  |  |
| --- | --- |
| **Silhouette\_distance** | **k-value(Number of Clusters)** |
| **0.11950472014657175** | **2** |
| **0.11950472014657175** | **2** |
| **-0.0634518694278572** | **3** |
| **-0.10135028043040387** | **4** |
| **-0.15671248130914417** | **5** |
| **-0.0795122407621355** | **6** |
| **-0.05629476779391.** | **7** |
| **-0.08737859156300563** | **8** |

**CONCLUSION AND FUTURE WORKS**

We can conclude that this project helped a lot in learning more about this big data field and our model was also able to classify the reviews as spam or non-spam. Our dataset is highly unbalanced as it has 20 non spam reviews for one spam review. This might be the reason we got such a low number of clusters.we can actually work on larger amounts of datasets which are balanced to get better classification. we can also try adding more constraints to this model by determining whether a user is fake and classifying their reviews as spam reviews, cause while analysing the data we found that there are many users who still post reviews for the businesses which are no longer open,So all such users can be determined as fake.These additional constraints can be used to improve the model in delivering much better results.

**References:**

1. <https://www.statista.com/statistics/183755/number-of-us-internet-shoppers-since-2009/>
2. <https://www.brightlocal.com/learn/fake-reviews-are-a-real-problem-8-statistics-that-show-why/>
3. <https://www.computer.org/csdl/proceedings-article/asonam/2014/06921594/12OmNCfAPGK>
4. https://www.researchgate.net/publication/305893213\_Exploiting\_Product\_Related\_Review\_Features\_for\_Fake\_Review\_Detection
5. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, pp. 309–319 (2013).