What Makes a Valuable or Fake Review?

Springboard Data Science Career Track
Capstone 2 - Milestone Report 1
by Chantel Clark

The Problem

Business owners often do not have the time to read through every customer review, especially when there are thousands of reviews. The problem with large amounts of customer reviews is that they can be repetitive, and sometimes they are computer generated. Customer reviews are nonetheless vital because they can provide valuable feedback on which the business can use to improve products and services.

Natural Language Processing (NLP) can help businesses to sort and find the informative customer reviews with ease and save a tremendous amount of time. By sifting through the reviews more efficiently, businesses would be able to save time by ignoring the fake reviews, and understand what their customers are truly thinking and feeling. Finding which reviews are similar with a recurring theme (valuable) will provide insight and actionable goals. The aim of this capstone is to explore how NLP can be used to determine which attributes of a review make it valuable or suspicious.

Data

Amazon has an open dataset of over 130 million customer reviews collected between 1995 and 2015, available as URL's at https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt. Each line in the dataset represents one product review. Reviews are grouped by product categories such as apparel, automotive, books, e-books, etc. The columns in the dataset include: 'marketplace' (country code), 'customer_id', 'review_id', 'product_id', 'product_parent' (random identifier for aggregate reviews for the same product), 'product_title', 'product_category', 'star_rating', 'helpful_votes', 'total_votes', 'vine', 'verified_purchase', 'review_headline', 'review_body', 'review_date'.
Complete an exploratory data analysis to answer the following questions:

1. *What is the mean and median number of reviews per customer?*
2. *What do the reviews of a highly active (>500 reviews) reviewer look like?*
3. *What do the reviews that a one review customer look like?*
4. *Do customers who write different amounts of reviews give the same distribution of star ratings?*

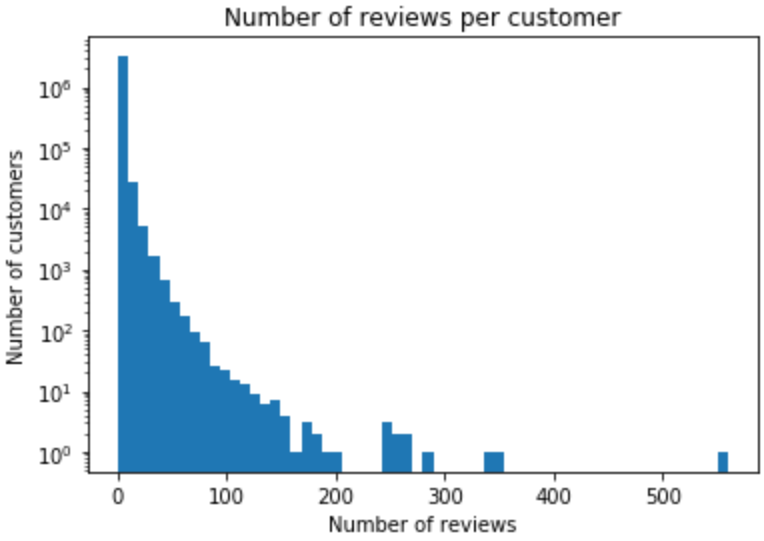
Analysis

To analyze the dataset, an 'apparel' subset with 5,881,873 reviews was used. About 10.1% of the reviews were unverified purchases, meaning that the product was not bought

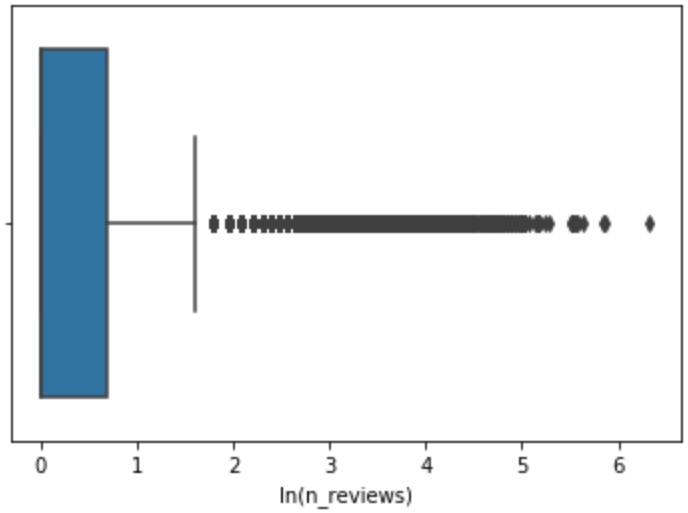
through Amazon. Browsing through the data, it was apparent that some reviews were misclassified in the apparel category. For example, a pineapple corer and beach hammock were contained within the apparel category. In order to get an idea about who the customer reviewers are, the following questions were addressed:

1) *What is the mean and median number of reviews per customer?*

The mean number of reviews that a customer writes is 1.83, while the median number of reviews per customer is 1. Because the average is larger than the median, it is evident that some customers are posting a very large number of reviews which is skewing the dataset to the right (long right tail). The median number of reviews per customer is low, and could be a result of fake accounts where only one review is posted.



There is an extremely large number of customers that post only one review (68.4% of all customers). As an attempt to identify outliers (highly active reviewers), the data was log transformed. The interquartile range (IQR) was computed and multiplied by 1.5 to find the length of the whisker in the box plot.

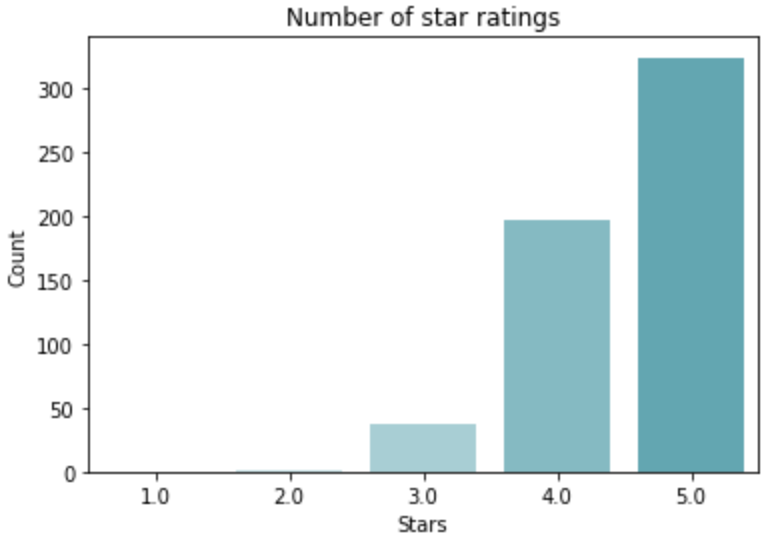


Using this method, customers who wrote more than 5 reviews would be considered as an outlier. In this dataset, there were 130,943 (4.1%) customers who would be classified as an outlier. This obviously does not seem right because on average, a 'real' (not fake) customer could write 6 reviews easily. The large amount of customers that write only one review is still skewing the data. In this case, the multiplier of 1.5 could be increased to 3 (or any desired multiple) to modify the threshold in which to identify customers that could potentially be a fake reviewer.

An alternative approach is to remove the reviewers who only wrote one review. When the 'one hit wonders' (people who only wrote one review) were removed, the threshold for the number of reviews that a customer writes to be considered as an outlier increased to 12. Less than 1% (0.89%) of all customers wrote 12 or more reviews.

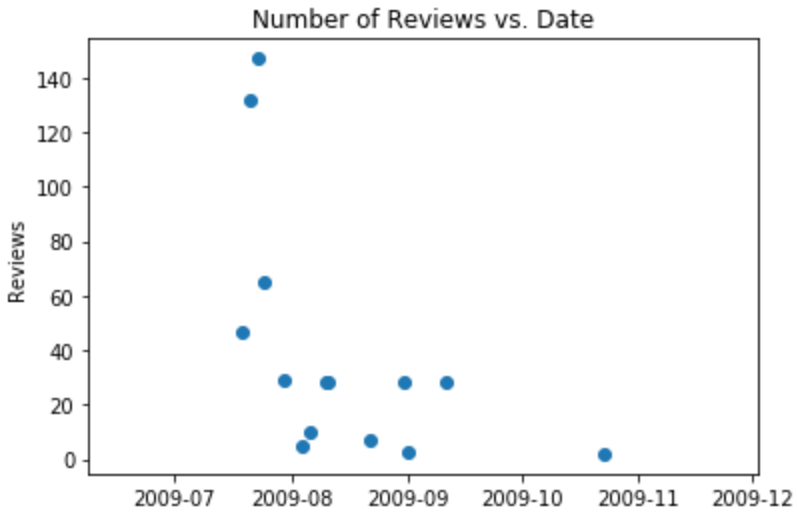
2) *What do the reviews of a highly active (>500 reviews) reviewer look like?*

None of the 559 purchases were verified. The average star rating was 4.5, with just 300 5-star ratings and over 200 4-star ratings. It's actually quite surprising that there were 3-star ratings, and one 2-star rating. Perhaps this is a strategy from getting caught.



There was only one 2-star rating from the reviewer that is clearly a positive review which does not match the 2-star rating and is obviously a promotion for the store:

> *"Bollywood Style Designer Indian Kurti - Unique Indian Kurtis for Women -Summer Dress lucknowi-chikan-- It's the latest in Indian fashion! If you are looking to buy the designer kurtis (tunic) to match with your jeans, trousers or salwars, simply browse this store....."*



The timeline above shows the frequency and number of reviews posted over time. This customer's first review was written on July 19, 2009, and last review on October 23, 2009. The total time that this customer was actively writing reviews was 96 days. On July 21 and 23 of 2009, there were over 100 reviews posted, clearly an anomaly. Number of reviews drastically decreased through October 2009. One way to flag suspicious review activity is to identify a maximum limit of reviews written per day for each reviewer. In this case, it might be reasonable to say that writing over 20 reviews is suspicious. Regardless of what threshold is chosen to identify the number of reviews posted in a day, it can be adjusted in the model.

Many reviews from this particular customer are exactly the same. Index of the review followed by the review body are below:

> *5027866 Stunning Cotton Kurti with gorgeous colored print. This ethnic kurti is master piece of Indian Ethnic Art.*
> *5064403 This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This printed skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.*
> *5064413 This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This crinkle skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.*
> *5064414 This classic cotton boho tier skirt are perfect for all long and short kurtis. A fun & fashionable skirt by Mogul Interior! This printed skirt offers a bohemian style and tie design Skirt is stylish, easy to wear, and is one of this season's hottest looks! Cotton.*

It is quite obvious that these reviews look like fake or unreliable reviews because of the repeated text for different items. Extremely large numbers of similar reviews can be an indicator of unreliable reviews, which could be identified through Natural Language Processing.

3) *What do the reviews that a one review customer look like?*

There were 2,201,632 customers who left only one review (37.4% of all reviews). A random sample of 50 reviews were chosen to inspect the reviews from customers that only posted one review. Most of the reviews in this subset were surprisingly thoughtful and unique, and did not seem to

be computer generated. Even the reviews that were unverified and were not voted as 'helpful' were surprisingly well-written:

> *Fits true to size BUT don't be alarmed when you put them on the hips are a little tight. Trust me that they will loosen to a comfortable fit.*

The distribution of star ratings for the one-review customer were mostly 5-star ratings (54.1%) followed by 4-star ratings (18.9%). While most of these reviews were 5-star ratings, the amount of 1 through 4-star ratings was reasonable, within 1 order of magnitude.

4) *Do customers who write different amounts of reviews give the same distribution of star ratings?*

If customers were grouped by the number of reviews that they have written, are the star distributions different? To compare, customers were segmented into group A if they wrote 1 review, group B for 2-5 reviews, group C for 6-9 reviews, group D for 10-12 reviews, and group E for 13 or more reviews. Because the data is ordinal, and observations (reviews) can come from the same reviewer, a Kruskal Wallis test was used to determine if the star ratings have the same distribution across groups.

$H_0: \ median_A = median_B = median_C = median_D = median_E$
$H_1:$ Two or more of the distributions do not have the same median

The Kruskal Wallis test was run for groups C, D, and E, and returned a test statistic of 1298.9, and a p-value of 8.706e-283. The p-value is very close to zero, therefore the null hypothesis can be rejected and we conclude that the medians of groups C, D, and E are different. This answers the question at hand - customers from different groups give different distributions of star ratings.

To further analyze the differences between two groups, a Mann-Whitney U test was used for each pair of groups.

$H_0 : \ median_A = median_B$
$H_1 : \ median_A \neq median_B$

Groups	Statistic	P-value
(C, D)	6.505e+10	4.009e-23
(C, E)	1.459e+11	1.299e-283
(D, E)	5.144e+10	6.551e-66

Rest of results to be determined… < code still running > . From the table above, p-values are all near zero, therefore we reject the null hypothesis for each test. The medians and distributions for groups C, D, and E are significantly different from the other groups.

Natural Language Processing

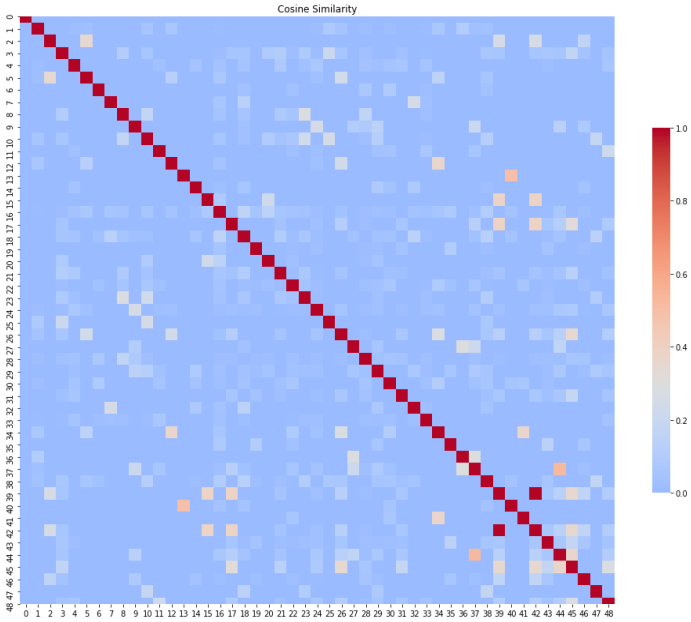
785 reviews without text were removed for NLP analysis. The remaining reviews were tokenized to break up the review by unique words and to build a vocabulary for the entire corpus (collection of reviews). Then stemming was implemented to account for the varying forms of a word with the same meaning. For example, ‘bad’ and ‘badly’ will both be changed to the stem word ‘bad’. Each review was vectorized with TF-IDF (term frequency - inverse document frequency), and stop words that are not descriptive such as ‘a’, ‘the’, ‘she’ or ‘he’ were removed with the ‘english’ dictionary provided by tfidfVectorizer. TF-IDF returns a vector with weights between 0 and 1 based on word frequency in all reviews, where words that are very common in other reviews receive a lower weight than the words that are used infrequently. Besides the TF-IDF weights, the vectorized reviews contain many zeros for words that do not appear in the review but are in the corpus vocabulary from other reviews.

To evaluate the similarity among the reviews, cosine similarity was computed with the following formula:

$$similarity \ = \ cos\theta \ = \ \frac{u \bullet v}{||u||\ ||v||}$$

Cosine similarity also ranges between 0 and 1, where a score of 0 represents very dissimilar reviews, and a score of 1 represents very similar reviews.

The cosine similarity matrix for sample data containing 49 reviews is below (full data set is still running).


Cosine Similarity
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
1.0
0.8
0.6
0.4
0.2
0.0