DS4A Women's Summit | Fall 2020 Capstone Project Report

Predict Potential Usefulness and Positivity of Product Reviews

---

Team 10 | CleaReview

Romane Goldmuntz
Hanh Nguyen
Mansi Parikh
Joyce Yu
Jane Zhang

Table of Contents

1 Introduction

1.1 Business Problem

Both consumers and e-commerce businesses rely on the ubiquity of user-generated product reviews to make better-informed decisions. Reducing the costs of identifying relevant information from vast amounts of product reviews aids potential consumers with their purchase decisions and informs businesses of key product and/or service development areas to focus on based on their importance to their consumers. "Ratings and review content is having greater impact on consumer behavior in the COVID-19 era, providing the validation and social proof necessary to drive sales" (PowerReviews, 2020). Research suggests that consumers are seeking validation for their purchases in such uncertain times more than ever, yet the length of their review periods have shortened with increased stress and/or the greater need to multitask (PowerReviews, 2020). To positively influence consumers' purchase intention in these accelerated and abbreviated consumer journeys with limited information, businesses may identify the most useful reviews in order to increase customer satisfaction via:

- Improving the product/service(s) offered and the overall customer experience based on feedback extracted from quality reviews, which would likely boost consumer confidence in purchases from the brand;
- Reducing friction in the review engagement process of the consumer journey by identifying and presenting high-quality user-generated information about the product/service.

While websites like Amazon allow consumers to vote for reviews found to be helpful, underrated ones that contain relevant information may go unnoticed when buried under top-voted reviews. Furthermore, users may focus on reviews with the highest and lowest ratings, neglecting useful information provided in average-rated reviews or those with a small number of helpful votes. To cope with the inundation of product reviews old and new, businesses may wish to assess and rank reviews by their usefulness and positivity. In particular, this project seeks to understand the properties of product reviews that signal their usefulness, and ultimately, create predictive models that output the helpfulness and positivity of product reviews based on the review content. In turn, product reviews with a low vote count may be brought to stakeholders' attention if they are found to be helpful.

1.2 Business Solution

The project delivers a dashboard that provides five positive reviews and five negative reviews that are deemed the most helpful by consumers to Amazon's vendors. Through the identification of such reviews, the dashboard can be further developed to provide vendors with consumer feedback and other strategic insights (e.g. industry trends, competitor information) that will help them improve the quality of their product/service(s). In turn, the dashboard would improve vendors' experience on the platform and increase their engagement with their consumers.

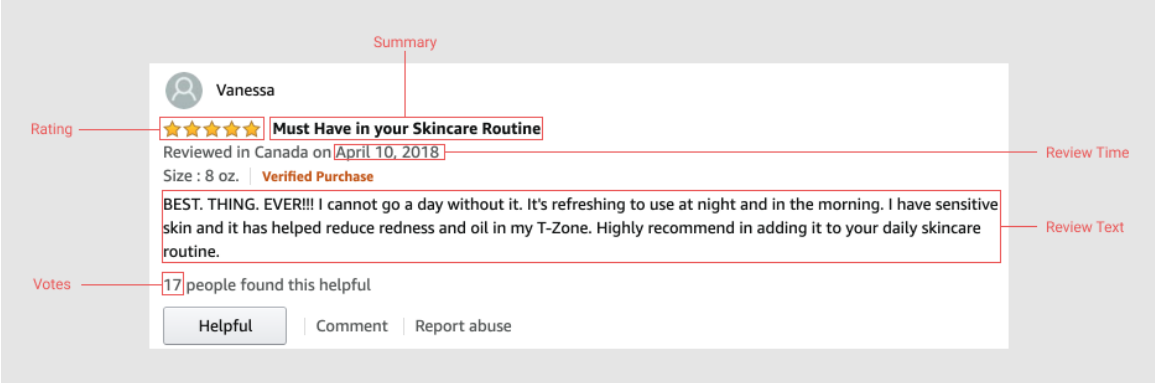


Figure 1.2.1: Sample Amazon review

In order to determine the top reviews to be displayed on the dashboard, reviews are to be rated on two different aspects: helpfulness and sentiment. Both are binary outcomes that will be determined by existing attributes in the review data (see Figure 1.2.1), particularly vote count to determine helpfulness and the star rating and NLP methods to determine the sentiment of a post. A relevant set of independent variables is to be used to determine both outcomes, namely important discussions of product features for helpfulness and the inclusion of words that would indicate positive or negative sentiment for the sentiment model. While the outcomes are binary, the developed model would output a probability, which can be used to rank all reviews against both outcomes and find the most helpful positive and negative reviews for each product that a brand owns.

2 Data Analysis and Computation

2.1 Datasets, Data Wrangling and Cleaning

We have chosen to focus on Amazon reviews datasets that are related to the beauty industry:

- All Beauty reviews (UCSD) - 370K reviews that represent 33K products
- Luxury Beauty reviews (UCSD) - 575K reviews that represent 12K products

All Beauty reviews and Luxury Beauty reviews both contained 2 subdatasets, the reviews and the product metadata respectively. The schema of such datasets is as follows:

Field	Type	Description
overall	FLOAT	Rating of the product
verified	BOOL	Whether it is verified that the user bought the product.
reviewTime	STRING	The timestamp for the review (format: yyyy-mm-dd).
reviewerID	STRING	ID of the reviewer
asin	STRING	ID of the product
reviewText	STRING	Text of the review
summary	STRING	The summary of the review
vote	FLOAT	Number of people who found the review helpful
style	DICT	A dictionary of the product metadata

Table 2.1.1: `reviews` dataset schema

Field	Type	Description
title	STRING	The name of the product
brand	STRING	The name of the brand.
rank	FLOAT	The rank of the product in the Beauty category.
asin	STRING	The ID of the product
description	LIST	The description of the product.
also_view	LIST	Related products that other customers viewed.
also_buy	LIST	Related products that other customers bought.
price	FLOAT	The price of the product in USD
similar_item	LIST	Similar product table.

Table 2.1.2: `metadata` dataset schema

Substantial filtering was implemented to identify the reviews that are suitable as model inputs. For example, all recent reviews published at the time that the data was pulled are to be eliminated from analysis given that they may have had little time to accumulate adequate votes on the platform to be deemed useful. Such filtering was conducted with the use of `reviewTime` in the reviews datasets. Additionally, since at least 5 positive and 5 negative reviews are shown for each product, the analysis focused on the products that have a substantial number of reviews.

The operations performed to clean both datasets are the following:

- Modify the data types in the Luxury Beauty datasets to be consistent with those from the All Beauty datasets in order to combine them for further cleaning;
- Remove reviews with NaN votes (see Exploratory Data Analysis, Figure 2.2.1);
    - 13.98% and 18.04% of the All Beauty and Luxury Beauty reviews datasets respectively contained non-zero vote values, totalling a sufficient amount of 155,588 reviews.
- Apply feature scaling to the vote counts in both shopping categories (separately) to account for very different shopper universes;
- Remove reviews published within 7 days of the most recently published in each respective dataset to ensure all reviews had at least one week to potentially be identified as helpful by users;
- Merge (inner join) data frames (`reviews` and `metadata`) on the product IDs;
- Order entries by votes and eliminate duplicates with fewer votes;
    - 5.18% of the combined datasets were duplicate reviews that were removed.
- For the models, remove unverified reviews since they are deemed less credible than verified ones.

2.2 Exploratory Data Analysis

Preliminary EDA showed that there is a lot of imbalance within the combined datasets, especially among the selected outcome variables. Fortunately, this is not a sparse matrix, and there are large volumes of data along with an adequate supply of predictors to use to train and test the predictive models.

	reviewText
2	Great hand lotion
3	This is the best for the severely dry skin on my hands
4	The best non- oily hand cream ever. It heals overnight.
5	Ive used this lotion for many years. I try others occasionally and always come back to Gardners. Please don't change a thing.
6	Works great for dry hands.
7	The best hand cream ever.
8	LOVE THIS SCENT!! But Crabtree and Evelyn make so many. Washes off easily too!!
9	Its a great moisturizer especially for gardners
10	This hand cream is the best! Have been using it for years. Keeps my hands soft all day
11	I am a healthcare care professional that suffers from dry hands. This hand lotion has changed my life. I keep it on the desk nearby. The hand therapy instantly improves the discomfort from dry hands.

Figure 2.2.1. Sample reviews with NaN vote values

As mentioned in Section 2.1, product reviews with NaN vote values were excluded from analysis. It is unclear whether or not NaN vote values indicate zero votes on a given product review. As shown in Figure 2.2.1, there exists product reviews with NaN vote values that contain arguably useful comments. Indeed, one possible assumption is that the NaNs account for the 0-vote reviews. However, after having a look at these reviews, some were very similar to high-voting-score reviews in the dataset. As a result, such reviews were excluded as this would introduce unnecessary assumptions into the analysis given the conflicting conclusions that can be drawn from this subset of reviews.

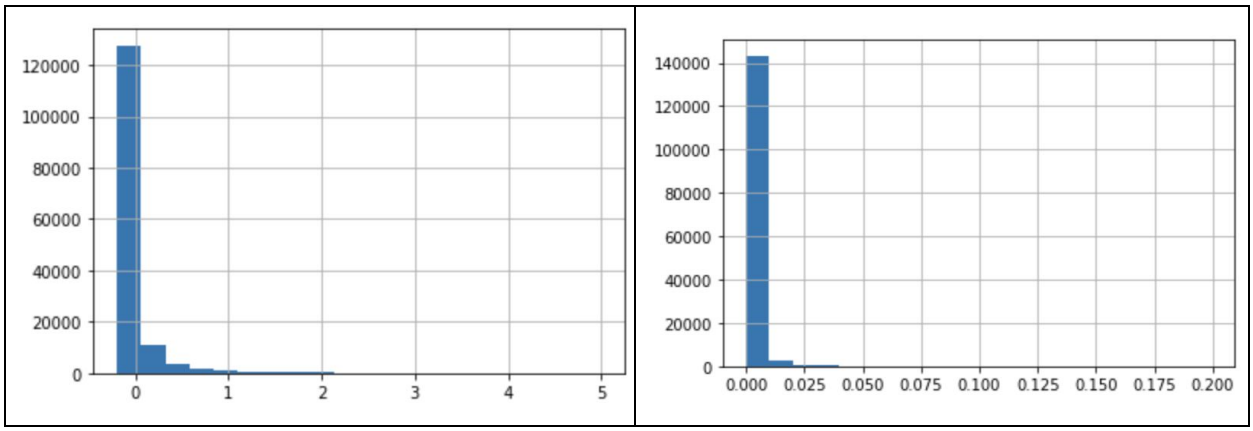


Figure 2.2.2. Feature scaling graphs of reviews by number of votes

As shown in Figure 2.2.2, feature scaling (left: standardization, right: normalization) was applied to the vote count for the reviews for both the All Beauty and Luxury Beauty categories before they were combined, and then the results were plotted for the combined data set. Unfortunately, both methods revealed that most of the reviews are concentrated in the region of low engagement (i.e., most reviews get very little vote activity) and so it would be difficult to choose a cutoff value for the binary casting of the vote count variable into a boolean helpfulness outcome. This makes us want to consider other ways to define helpfulness for a review, whether it's using other variables to do this so that the distribution is widened and/or possibly transforming the outcome variable in order to reduce the skewness. However, we still used the results of the normalized scale to label reviews as this was the most practical option.

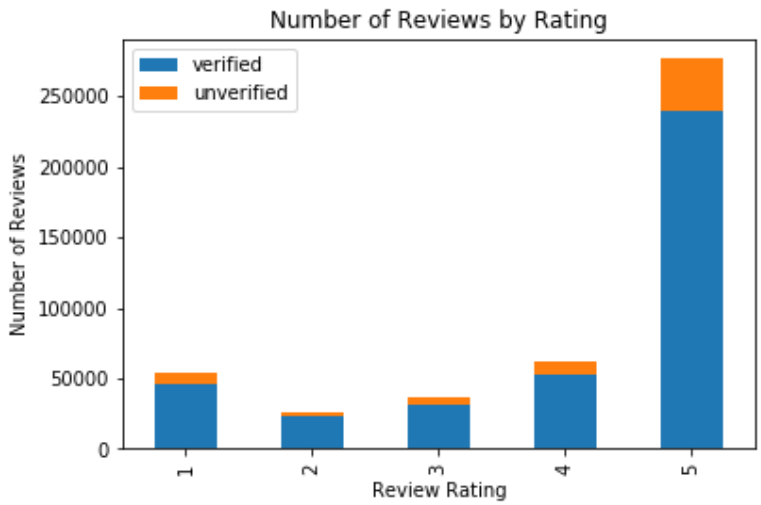


Figure 2.2.3: Stacked bar plot of filtered review count by rating and buyer verification status

As seen in Figure 2.2.3, a large amount of the filtered reviews contain a 5-star rating. Those top reviews aside, most buyers have rated their purchases as one or four stars. The prevalence of such reviews with extreme ratings are then more likely to be viewed by potential buyers as "extreme values are often more salient than more moderate values" (Filieria, Raguseo, & Vitari, 2018). Furthermore, most reviews that have a non-zero number of votes are published by buyers that made verified purchases. This may suggest that the verification status of a review contributes to its helpfulness by reinforcing the perceived trustworthiness of the reviewer.

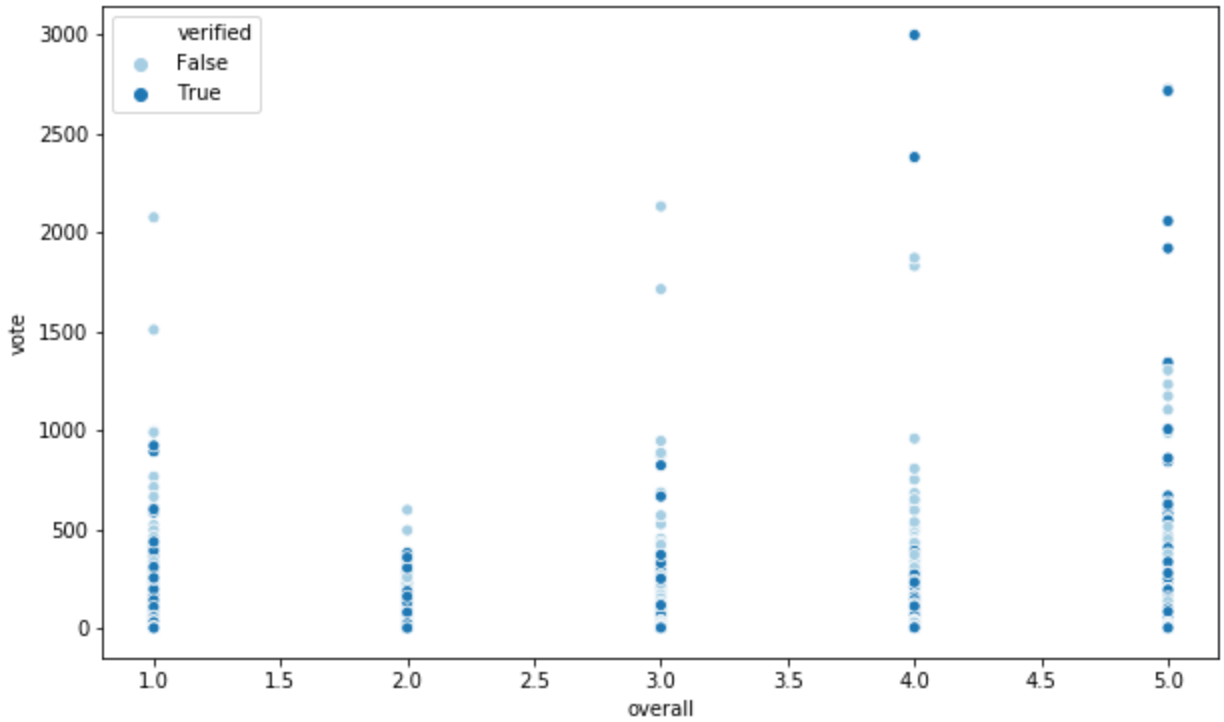


Figure 2.2.4: Scatter plot of filtered reviews by rating ('overall') and number of votes received ('vote')

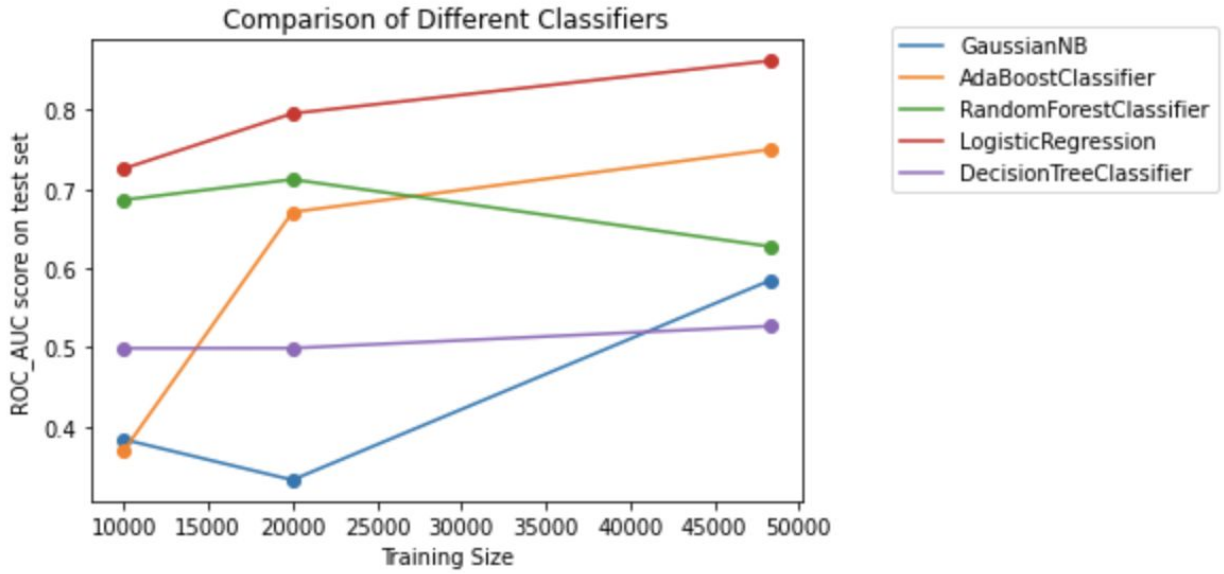
Figure 2.2.4 shows that most verified reviews receive up to 500 votes. With consideration towards the quantity of reviews by rating (shown in both Figure 2.2.3 and 2.2.4), the reviews at the extremes of the rating scale (1-5 star reviews) have received the most votes (out of total votes given to all reviews in the analyzed product categories). However, the height of the distributions for each rating in Figure 2.2.4 suggest that reviews with intermediate star ratings can be helpful despite the aforementioned user gravitation towards those with extreme ratings.

2.3 Statistical Analysis and Machine Learning

2.3.1 Helpfulness Predictive Model

We determined the features of a review that deem it most helpful by users, which we assessed by vote counts. While we expected that the most helpful reviews contain information about cost, durability, and other key features of a product, we did not view each of these different categories separately or even take into account a post's level of detail (which is a function of how long it is), punctuation, capitalization, and any other related characteristics. Instead, we made each word in the reviews its own independent attribute which would be used as the predictor variables in determining if a post is helpful or not. Unfortunately, helpfulness itself is still considered a rather subjective quality of a post, but if several people hold the same belief (as seen by the vote count), we can safely assume that the post is indeed helpful so the class itself does have some level of objective truth.

After trying out various classification models, we saw that logistic regression performed the best in terms of ROC so we pursued that to the finish. Not only was the ROC of this classifier better than the rest, but also the average ROC from 100 simulated trials was 0.86. Once we refined the model, we applied the results to an 'out-of-phase data set,' or one that was meant for only the dashboard.



2.3.2 Sentiment Analysis Model

The goal of the sentiment analysis model is to predict whether an ambiguous review (2 or 3-star review) is positive or negative.

To achieve this goal, the Linear SVM algorithm was selected, based on its known performance on text data for this specific task (Reddy, 2018).

The data was preprocessed by removing stop words defined by SpaCy small english model, to which the word 'not' has been to removed, because of its possible contribution and significance to negative reviews. For example, in the sentence 'I did not like this product', removing the 'not' would change the negative review into a positive one. Other preprocessing included the removal of ASCII characters, the lemmatization of the words within a review, and the application of tf-idf. Bigram were considered beside unigrams to ensure that significant word associations would not be missed (e.g. 'not like' considered as well as 'not' and 'like' separately).

The model was trained and tested only on reviews that were considered negative and positive with high certainty (1-star reviews for negative reviews and 4- and 5-star reviews for positive reviews) before being applied on the ambiguous reviews that had been set apart before the training of the model.

The results reach an accuracy of 94.7% (see confusion matrix and classification report below). It was observed that the precision, recall and f-1 score are pretty high for both classes, but that the number of False Positives was also high.

True value/predicted value	negative	positive
negative	1946	422
positive	264	10360

Confusion Matrix: Baseline Model

	precision	recall	f1-score	support
0.0	0.88	0.82	0.85	2368
1.0	0.96	0.98	0.97	10624
accuracy			0.95	12992

Classification Report: Baseline Model

After applying the model to the ambiguous reviews, the problem seems to be confirmed. The issue might be due to the imbalanced dataset and the dominance of positive reviews. To counter that, the Linear SVM model was augmented with an oversampling method called SMOTE, which artificially generates samples for the minority class - negative reviews in this case.

The results are a slight decrease in accuracy (92.9%) and a significant decrease in the precision for the negative class (88% to 77%). The reason behind such a drop is a significant increase in the number of False Negatives. As the number of False Positives has decreased, adding SMOTE to the model seems to have inverted the bias in the baseline model.

True value/predicted value	negative	positive
negative	2080	288
positive	630	9994

Confusion Matrix: SMOTE model

	precision	recall	f1-score	support
0.0	0.77	0.88	0.82	2368
1.0	0.97	0.94	0.96	10624
accuracy			0.93	12992

Classification Report: SMOTE model

The literature seems to suggest that combining a SMOTE approach and a biased SVM model could increase performance (Wang, 2008). The method was therefore applied on the project data but led to worse results than when SMOTE was added alone.

True value/predicted value	negative	positive
negative	2086	282
positive	665	9959

Confusion Matrix: SMOTE and Biased SVM Model

	precision	recall	f1-score	support
0.0	0.76	0.87	0.82	2368
1.0	0.97	0.94	0.96	10624
accuracy			0.93	12992

Classification Report: SMOTE and Biased SVM Model

For this project, the decision was made to go with the first model. The reason behind that decision is that in the worst case scenario, a negative review being considered highly helpful is going to be labelled positive and be under the Positive Reviews tab of the dashboard. This will only give more information to

the vendor, who would likely identify the review as negative and misclassified, about what is strongly disliked by their customers and give them a chance to fix it before it drives new potential customers away.

2.4 Dashboard Application

The interactive dashboard was implemented with Python (Dash), HTML, and CSS. It contains three primary components, which are updated upon product selection:

- *Summary Info*: the average review rating and the total number of reviews available for the selected product;
- *Top Reviews*: the five most helpful positive and negative reviews for the selected product, categorized based on the sentiment classified by the sentiment analysis model, and ranked by the helpfulness predictive model;
- *Keyword Insights*: word clouds to present the keywords found in positive and negative reviews for the selected product.

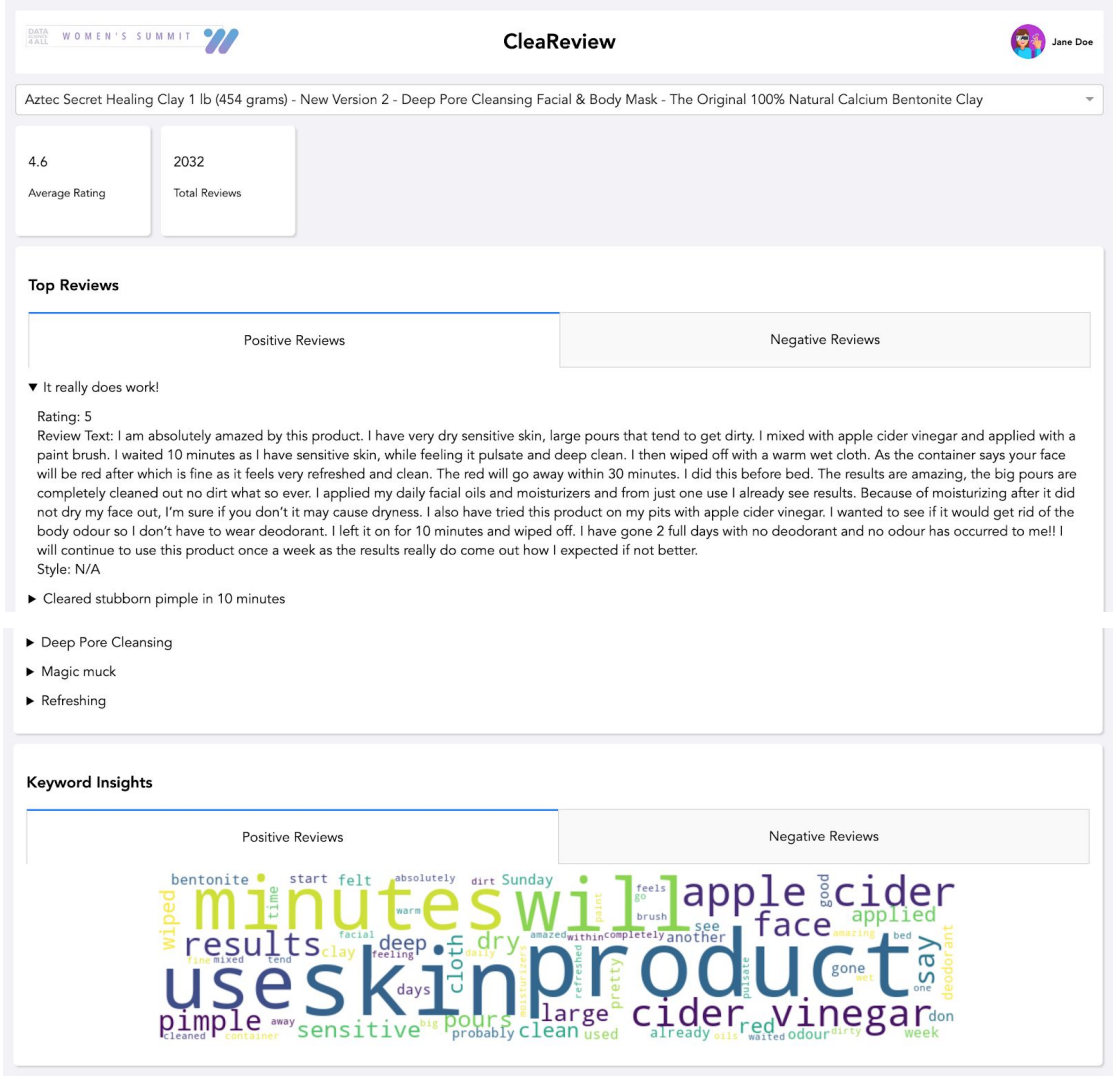


Figure 2.4.1: Screenshot of dashboard for sample product

3 Conclusions

After consulting other work done in this area, we found that there were several limitations to our project that we couldn't necessarily control. One example is that we were not working with the best possible data set and that limited us from being able to construct, with confidence, a proper outcome variable. It was due to relative numbers of upvotes that we did this, but many other researchers also had downvotes included in their datasets and they could make more confident decisions as to the class of each review. Another issue was that not all the reviews should be considered equal. Many have had a very long tenure on the platform and so have had the chance to accumulate many votes, whereas recently published posts that weren't necessarily removed from our dataset, did not have a comparable chance to accumulate votes. Had the reviews been comparable, we would have been able to rely on upvotes alone.

In any case, we discovered that it is difficult to determine the helpfulness of a post and its sentiment! The English language is varied with a single word having many meanings depending on the phrasing. This makes it hard to isolate words and work with features independently. However, this only makes the need more important to incorporate important phrasing and slang into models to account for different interpretations.

4 Future Work

There are several areas of improvement in this project that may be considered, especially in regard to the further iterations of the dashboard, as well as the analytics that happens behind the scenes which fuels the dashboard.

4.1 Dashboard Extensions

The dashboard delivered through this project serves as a MVP for the end product that we envisioned would be used by all vendors on an online marketplace platform like Amazon. As static datasets from UCSD were used for the project to ensure data integrity, there was no immediate need to store, maintain, or update the data used via an online database. However, in order for real vendors to use the dashboard as intended, an online database should be created to retrieve the most up-to-date product review data for a given vendor. Furthermore, account authentication features should be implemented such that each vendor views the data for its own brands and products, as opposed to viewing all available product review data as shown on the current version of the dashboard.

In terms of functionality, additional features may be introduced to the dashboard based on the insights that may be extracted from the product review data. The following are potential features that may be added to the dashboard:

- Industry insights (e.g. keyword word clouds) based on all reviews in a given product category;
- Competitor insights (e.g. keyword word clouds) based on all reviews for a given product type.
    - This would require identifying new data sources to be used to classify the specific product type, as such data is not available in the UCSD datasets. However, the `similar_item` column in the `metadata` dataset may be used to group products together.
- Product improvement recommendations may be inferred from the delta between vendor insights, industry insights, and competitor insights.

4.2 Model Improvements

There are several model improvements that can be implemented in future iterations:

- Find better ways to construct the target variable, namely relying on variables other than just the count of positive votes, which was all that was available;
- Use the summary of the review rather than solely the longer review description for determining helpfulness and sentiment;
- Ensure that the model is built in a statistically sound fashion by inputting a more balanced data set into the models

References

Filieria, R., Raguseo, E., & Vitari, C. (2018). When are extreme ratings more helpful? Empirical evidence on the moderating effects of review characteristics and product type. *Computers in Human Behavior, 88*, 134-142.

Wang, H.-Y. (2008). Combination approach of SMOTE and biased-SVM for imbalanced datasets. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 228-231. Retrieved from https://ieeexplore.ieee.org/document/4633794

PowerReviews. (2020, May 5). *PowerReviews Market Trends Snapshot – April 2020*. Retrieved from PowerReviews: https://www.powerreviews.com/insights/impact-consumer-ratings-covid-19/

Reddy, V. (2018, November 12). *Sentiment Analysis using SVM*. Retrieved from Medium: https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1