Thomas Screven

Analysis of Amazon Product Reviews

1. Introduction

1.1 Topic

This report is a linguistic analysis on a large dataset of Amazon reviews. The goal was twofold: determine if text classification models can predict the likelihood of a review being useful and identify linguistic differences in reviews across product categories.

Amazon allows users to vote that a product review is helpful. A straightforward technique to optimize the shopping experience is to show reviews with the highest number of helpfulness votes first. However, this heavily favors older reviews as newer reviews will not be seen by as many people. The motivation of the first part of the project was to explore if text classification models can pick out useful fresh reviews to show customers on the front page. The motivation for the second part was to examine part of speech usage to show differences and similarities between the way people write reviews in different product categories.

1.2 Work Performed

I downloaded and processed Amazon review data for 28 different product categories. Every review was labeled as being helpful or unhelpful based on its number of helpfulness votes. For each category, 2 binary classifiers were built: a Naïve Bayes model and a `fastText` model. These models were trained to classify product reviews as helpful or unhelpful. I then ranked test set reviews based on scores produced by each model to evaluate the model's ability to improve the likelihood of presenting useful reviews to a customer.

I also created a frequency distribution table for all part of speech tags in all product categories. Then, I graphed a heatmap of this table showing the difference in POS usage in each category's reviews. I then used a 2d embedding of the frequency distribution vectors to visualize POS usage relationships between categories.

1.3. Utility and Benefit

This work benefits online shoppers by showing useful reviews aiding their buying decision. Similarly online stores, such as Amazon, benefit by providing consumers a smoother experience encouraging repeat shoppers. Furthermore, since classification is based only on review text, differing performance levels across categories and the analysis of part of speech may show further linguistic analysis of review text is warranted.

2. Data

The raw dataset comes from the 2018 version of the Amazon Review Dataset compiled by the McAuley Lab at UC San Diego[1]. Only reviews written by users who had posted at least 5 reviews were included to avoid one-off behavior creating outliers in the data and also allowing us to focus on active users. The dataset consists of 28 JSON files, each for a distinct product category such as "Books" and "Gift_Cards". Reviews without review text were discarded, and for files with more than 100,000 reviews, 100,000 reviews were randomly selected to analyze. Then, duplicate reviews were removed. Finally, all reviews were labeled as being either helpful or unhelpful, where a review is helpful if and only if it has at least one helpfulness vote.

3. Tasks Performed

3.1. Constructing and Evaluating Review Ranking Models

For each product category, I built, trained, and tested a hand coded Naïve Bayes model and a `fastText` model. Both models are binary classifiers that predict the likelihood of a review being helpful or unhelpful. For both, the first 80% of the randomly sorted data was used for training and the rest for testing.

The Naïve Bayes model uses a bag of words technique with logarithmic add-one smoothing, and normalization of the class probabilities. For the `fastText` model, special character strings (such as punctuation) are removed. After that, the model conducts supervised training on the remaining review tokens. In both cases, `nltk` was used for tokenization.

After training, the models are tested on unseen data. The normalized positive (a helpful review) class scores for each review are extracted. After, the test set reviews are ranked based on their helpfulness score output by the model. Then, the top 5% most highly scored reviews are selected, and the precision for this subset is computed. Then, that precision is compared to the prior probability of a review being helpful in the category. This comparison shows the benefit of

[1] Jianmo Ni, Jiacheng Li, Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fined-grained aspects. In *EMNLP*.

using the model to choose fresh reviews to present to customers as compared to choosing reviews randomly.

3.2. Computational Part of Speech Analysis

I performed part of speech analysis on the categories. For each product category, `nltk` tokenization and `nltk` part of speech labeling was used to create a frequency distribution of POS usage in the reviews of each category. The distributions were combined into a matrix with a row for each category and a column for each tag. Each row sums to 1, so each entry in the table represents the fraction of tokens in a category that are the corresponding part of speech.

Using the frequency distribution table, I rendered a heat map using `matplotlib` showing how part of speech distributions in reviews differed between product categories. Then, t-distributed stochastic neighbor embedding (t-SNE) was used to create a 2-D embedding of the per-category part of speech frequency vectors. I created a scatter plot of this embedding to visualize the part of speech use relationships between product categories.

4. Results

Below are tables reporting the precision of the reviews with the 5% most highly ranked helpfulness scores. The tables are compilations of the statistics for all 28 Naïve Bayes models and `fastText` models, one for each product category.

Naïve Bayes

	P@5%	Helpfulness Prior Probability	Difference between P@5% and Prior Probability	Ratio of P@5% vs. Prior Probability
Mean Accuracy with 95% CI	56.8% (50.5% - 63.1%)	14.4% (12.0% - 16.9%)	42.4% (37.4% - 47.4%)	4.31 (3.73 - 4.90)
Maximum	96.7%	34.0%	87.0%	10.0
Minimum	20.7%	4.63%	13.8%	2.41

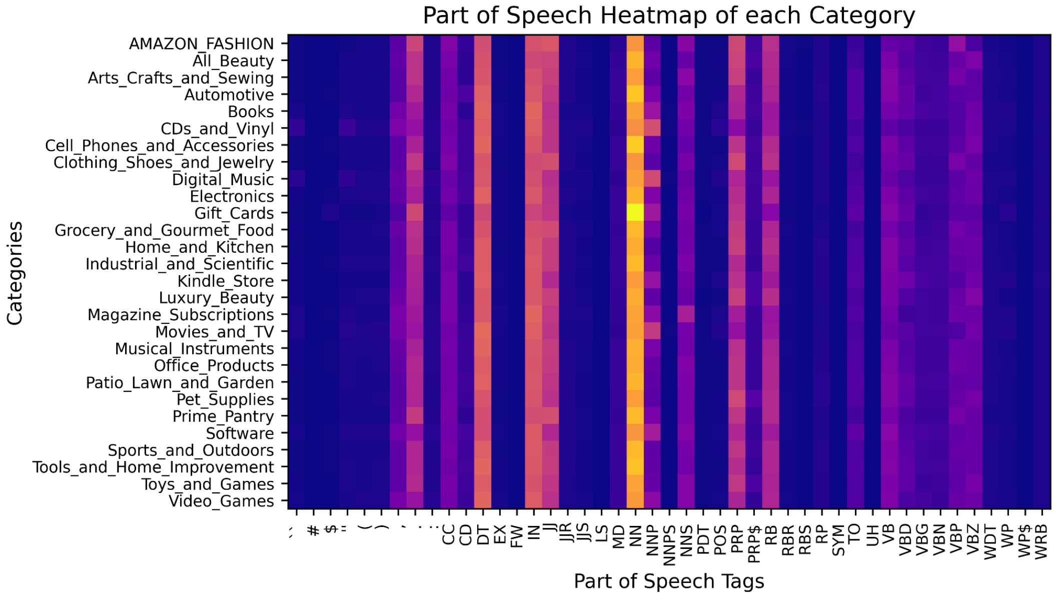
FastText [2]

	P@5%	Helpfulness Prior Probability	Difference between P@5% and Prior Probability	Ratio of P@5% vs. Prior Probability
Mean Accuracy with 95% CI	50.4% (43.9% - 56.8%)	14.4% (12.0% - 16.9%)	35.9% (30.3% - 41.5%)	3.91 (3.15 - 4.66)
Maximum	100%	34.0%	90.4%	10.4
Minimum	10.3%	4.63%	3.47%	1.51

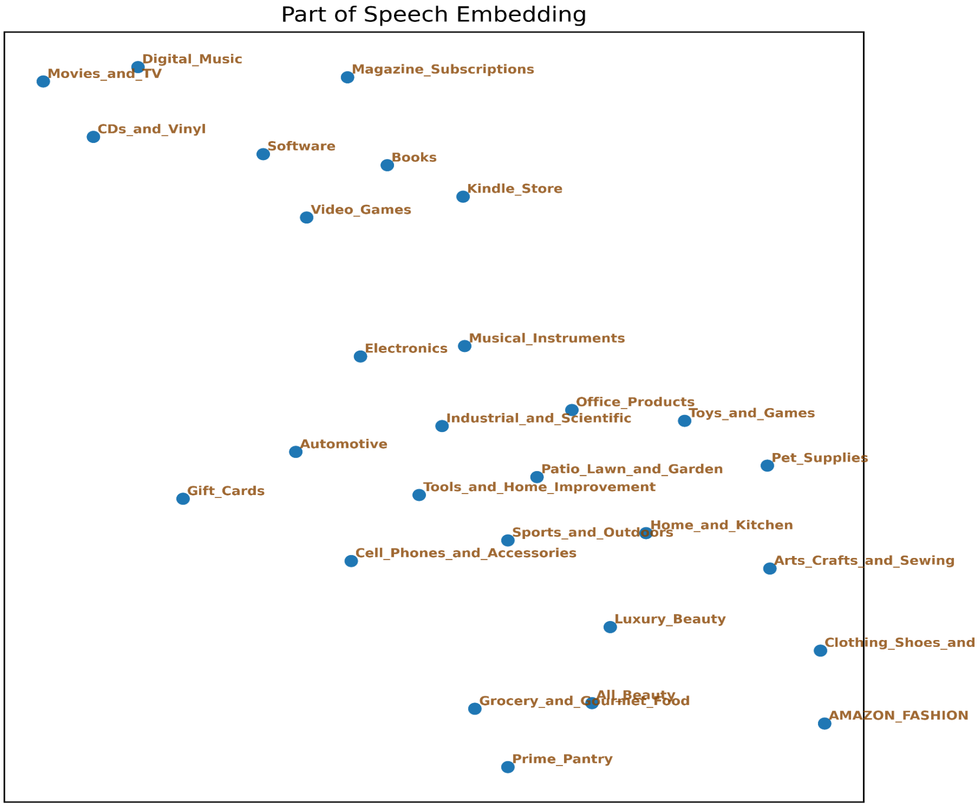
Both models significantly outperformed random chance. Even in the worst case, the Precision@5% metric is much higher than the helpfulness prior probability. Naïve Bayes performs somewhat better than `fastText`. The 5% precision threshold was chosen based on the following hypothetical situation: Amazon has a pool of 100 fresh reviews for a given product, and they wish to include 5 on the first product review page. The reviews selected by the models correspond to the Precision@5% metric. In that hypothetical situation, the models on average chose around 4 times as many helpful reviews than choosing randomly would. That is a substantial improvement.

However, while performance was at least good in every category, it varied between categories, implying there are related language differences. I analyzed these differences by calculating POS frequency distributions for each of the 28 categories and visualizing them in a heat map.

[2] Slight randomness in fastText model performance but the difference between trials is not significant.



As one may expect, there are several POS tags which dominate in usage, such as singular nouns (NN), adjectives (JJ), and determiners (DT). However, we do see the relative frequency of each part of speech differs across product categories. To inquire further, I plotted a t-SNE 2-D embedding of these POS frequencies.



This obtained a fascinating result: product categories which are similar in function also are similar in POS usage in their product reviews. Similarity in POS usage across categories is reflected in how short the distance is between scatterplot points. Below are several examples of clustering but there are many more combinations in the embedding:

1. Bottom right: luxury beauty, clothing/shoes, Amazon fashion, and all beauty.
2. Center: patio/lawn/garden, tools/home improvement, and home/kitchen.
3. Top left: digital music, movies/tv, and CDs/vinyl.
4. Top center: magazine subscriptions, books, and Kindle store.

We clearly see there are distinguishable patterns in the language in reviews related to the type of product. Some combination of a product type's nature and the population likely to write reviews for a product type influences how people use language to write reviews.

5. Continuing, Improving, and Extending

It would be interesting to implement ML frameworks which use deep learning. Those richer frameworks could incorporate other features such as the rating given by the reviewer and their history. Another area to examine is if other high level linguistic characteristics also follow the part of speech trends uncovered in this report. Finally, it would be compelling to explore if linguistic analysis could be effectively used in a recommender system to suggest products to users based on the linguistic properties of their own reviews.