Predicting Amazon Reviews' Helpfulness

By Aren Simmons

Verified Amazon Users

Producing helpful Amazon reviews provides a monetary benefit for users. They have a chance to receive free samples from Amazon if they are verified.

There are a number of factors that may deem a review as helpful and understanding them would give users insight to write reviews that are more helpful for others.

Data Acquisition

The Data comes from a UCSD data repository.

21 Datasets were selected (ones that have ~10M rows or less)- one for each category.

The metadata was also acquired for each category.

All datasets were in separate JSON files which were parsed into a dataframe.

A random sample of products with at least 5 reviews were selected, and only 5 random reviews for each product for 3000 products - then merged with metadata.

Each category had 3000 products, then saved as a csv file. All dataframes were then concatenated.

Data Cleaning

Each review and summary were tokenized, then each token was lemmatized with the part of speech before combining back to its original form.

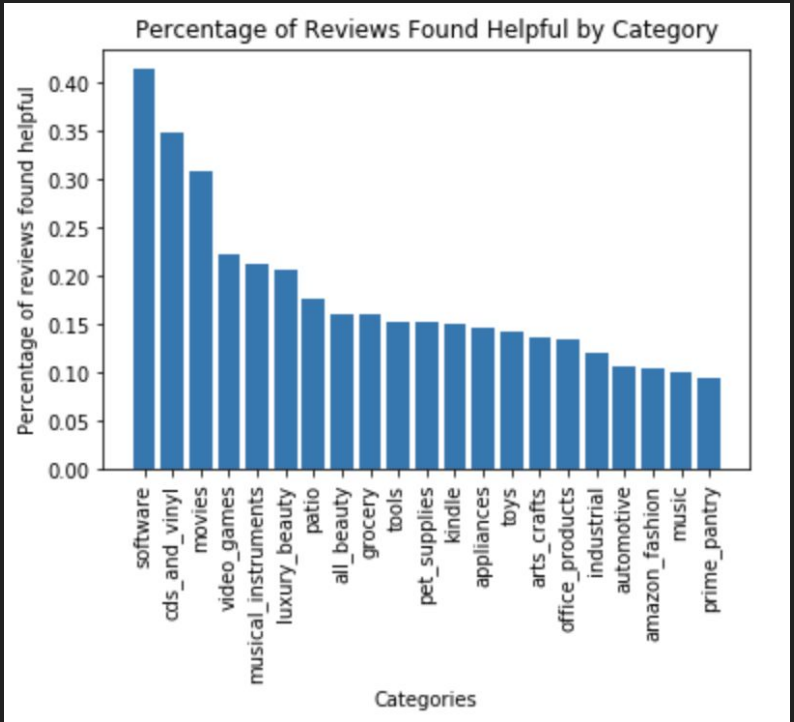
Only digits were extracted from the rank column

Dates were converted to datetime objects.

Reviews and summary lengths were calculated and a separate column where they were grouped into buckets.

Exploratory Data Analysis

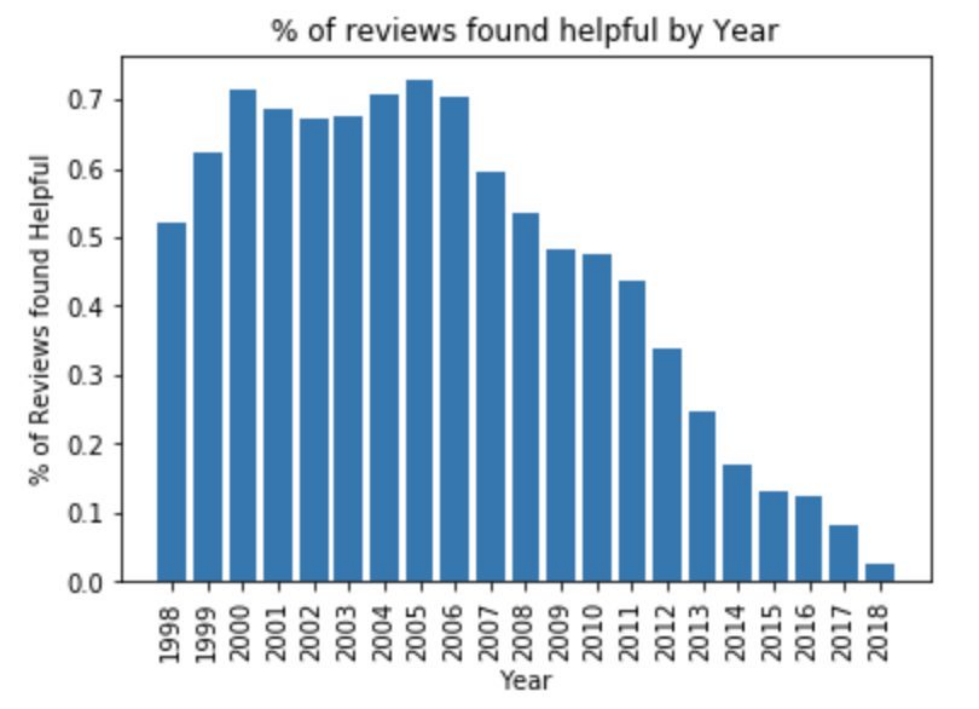
Percentage of Reviews Found Helpful by Category



We see that the categories with the highest share of helpful reviews are software (41.4%), cds (34.7%), movies (30.7%) and video games (22.1%).

It appears that the categories with the highest share of helpful reviews are categories that are digital.

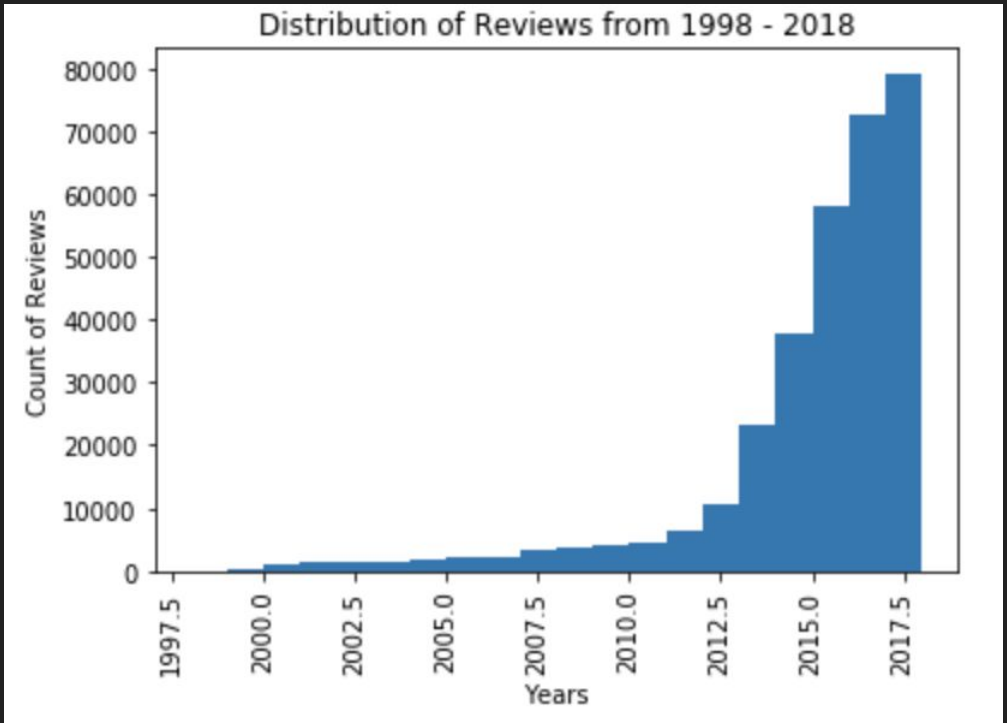
Percentage of Reviews found Helpful by Year



We see a drastic decline in reviews found helpful over the years.

The length of time a review spends on Amazon is a large determining factor on whether or not a review is deemed helpful or not.

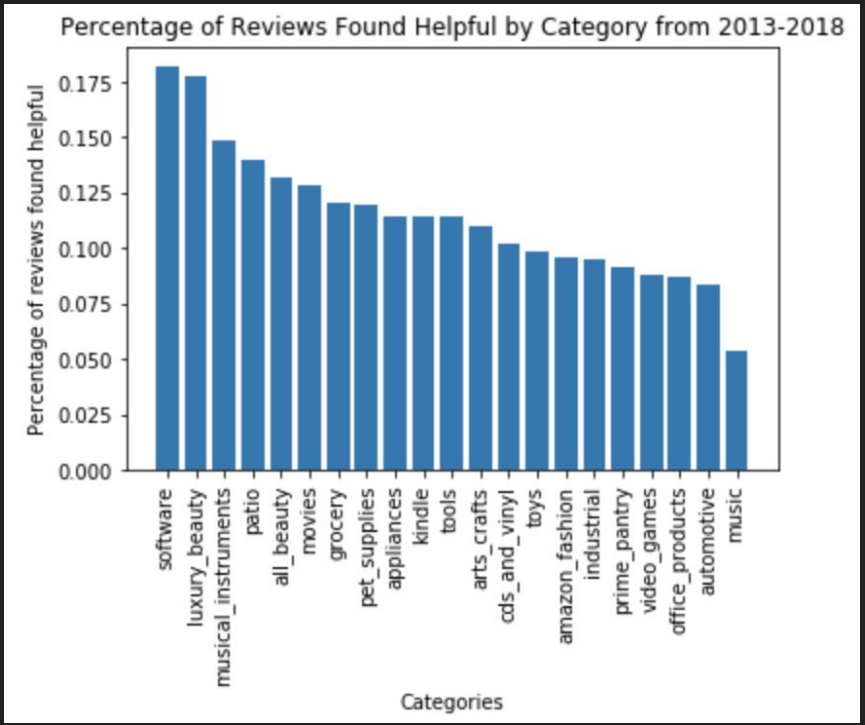
Distribution of Reviews by Year



The distribution of reviews are skewed to the left. We see that a majority of the reviews from the dataset are recent.

Roughly 79% of reviews are from 2013 to 2018.

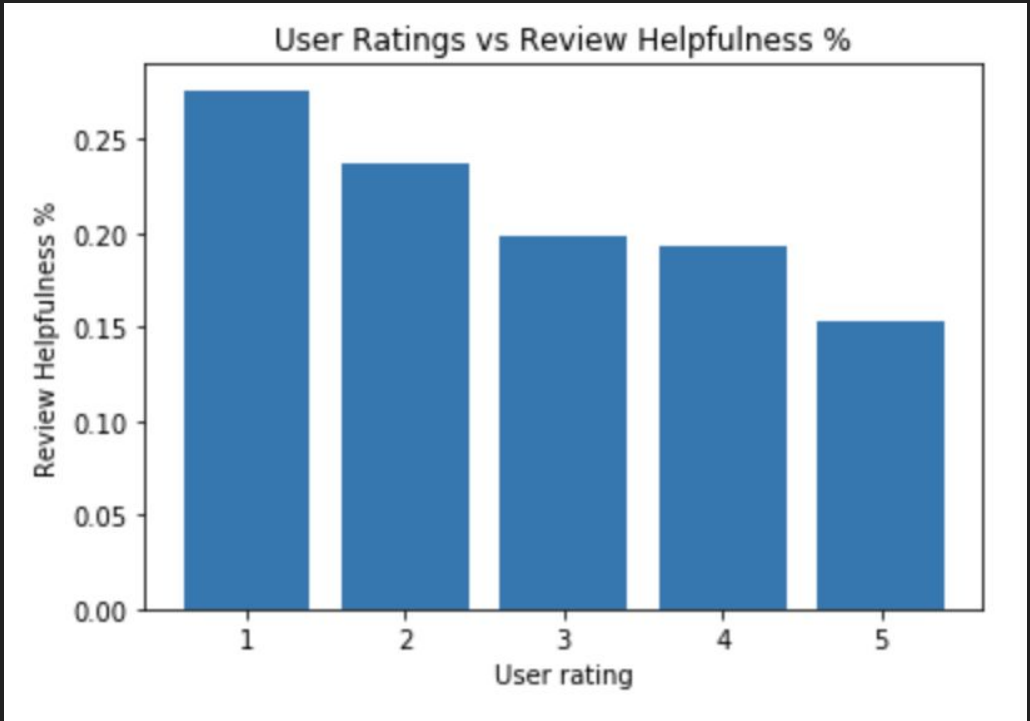
Percentage of Recent Reviews Found Helpful by Category



Without older reviews software is still on top, but only with 18% of reviews found helpful. We see a dramatic decrease in reviews found helpful now that older reviews have been queried out.

My first intuition was that digital categories had a higher share of helpful reviews, however upon removal of earlier years, we see that is not the case. Ultimately the length of time that a review has been posted has the highest impact of determining a reviews' helpfulness.

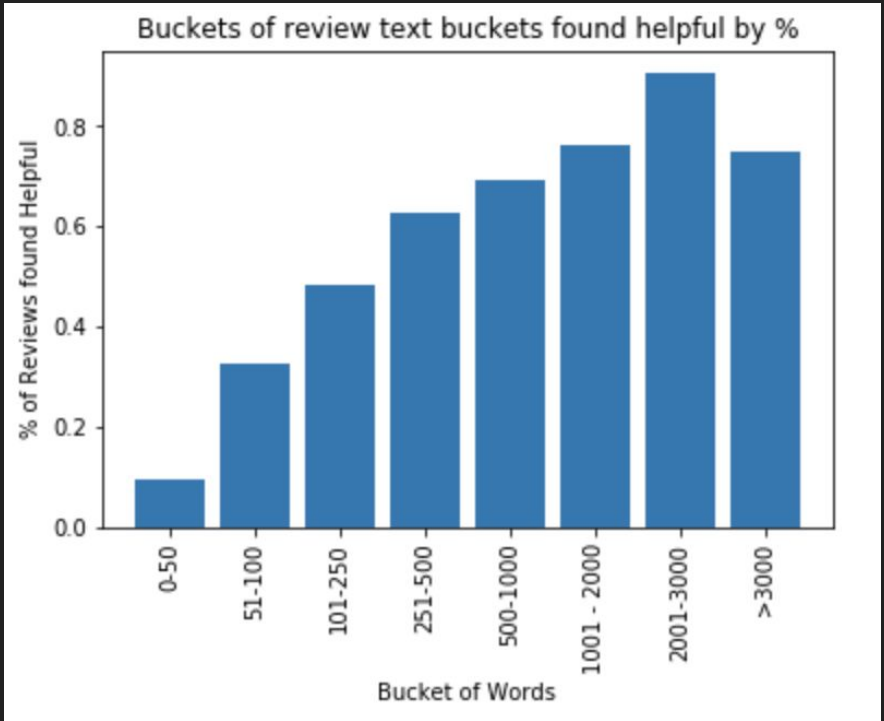
User Ratings vs Helpfulness %



We see that reviews that give a product higher ratings are found to be less helpful.

It is likely that reviews that are more critical of a product are found to be more helpful than a review that praises the product.

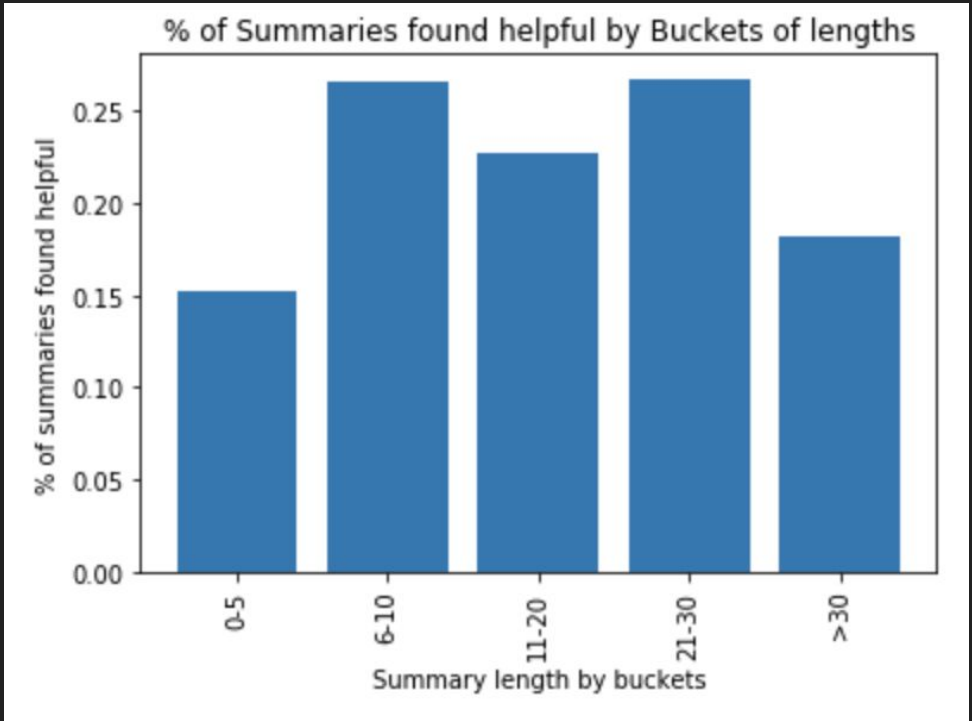
Buckets of Review Length by Helpfulness %



We see that longer reviews are found to be more helpful than short reviews.

A majority of reviews with 250 words or more are found helpful (62% and higher).

Summary Bucket Length by Helpfulness %



Summary Bucket Length seems to have a sweet spot between 6 and 30 words that are found more helpful.

Helpful and Unhelpful Words

Predicting helpful and unhelpful word probabilities

I used a Naive Bayes classifier to predict Helpful and Unhelpful word probabilities and ranked them from highest to lowest.

Word probabilities were split up by review words and summary words.

I repeated this for each category for all years and for recent years (post 2013).

Helpful Review And Summary Words

Helpful words	P(Helpful | word)
scherzo	0.89
sonata	0.88
seventh	0.88
brahms	0.87
bartok	0.86
richter	0.86
quartets	0.84
frontpage	0.83
schubert	0.83
poser	0.81
Unhelpful words	**P(Helpful | word)**
manifold	0.02
historia	0.02
drumstick	0.02
carburetor	0.02
goodie	0.02
reeds	0.02
libro	0.02
peavey	0.02
producto	0.01
sheath	0.01

Helpful words	P(Helpful | word)
acquaint	0.85
baritone	0.85
challenged	0.85
96	0.80
collision	0.71
crop	0.67
coffee	0.66
crunch	0.66
dances	0.65
crazy	0.64
Unhelpful words	**P(Helpful | word)**
confirm	0.04
cooked	0.04
categorize	0.04
abilities	0.04
creation	0.04
cakes	0.04
bedtime	0.03
ballad	0.03
compression	0.03
babe	0.03

We see that words that are listed in languages other than english (libro, producto) have low probabilities of being considered helpful. The helpful words are typically terms or names that are associated with the specific product (scherzo and sonata are latin words used in classical music).

Helpful Review Word Probabilities

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	op	0.938433	chant	0.018605	music
1	adagio	0.833471	whenever	0.018605	music
2	di	0.812117	sea	0.018605	music
3	ii	0.803728	hearing	0.017530	music
4	walter	0.803728	blessing	0.017199	music
5	andante	0.784481	remix	0.015990	music
6	verdi	0.761050	cassette	0.014019	music
7	concerto	0.761050	lol	0.013206	music
8	cobham	0.761050	oldie	0.013206	music
9	liszt	0.747313	relate	0.012481	music

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	thule	0.856139	opening	0.020454	automotive
1	lbs	0.789856	chain	0.019200	automotive
2	dump	0.652696	yes	0.019200	automotive
3	charcoal	0.632719	miles	0.019200	automotive
4	buffer	0.632719	thread	0.018629	automotive
5	65	0.610303	toyota	0.017750	automotive
6	tv	0.610303	described	0.016873	automotive
7	fasten	0.610303	2000	0.014696	automotive
8	crv	0.610303	amaze	0.014037	automotive
9	weigh	0.610303	explorer	0.013435	automotive

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	iron	0.700845	exactly	0.018365	appliances
1	warn	0.670747	shipment	0.016693	appliances
2	cooker	0.670747	perfect	0.016621	appliances
3	capacity	0.670747	took	0.016036	appliances
4	alert	0.647074	received	0.015429	appliances
5	cascade	0.647074	saved	0.015429	appliances
6	claim	0.647074	promised	0.015429	appliances
7	soda	0.647074	fits	0.015236	appliances
8	queen	0.633914	worked	0.012809	appliances
9	utensil	0.619734	described	0.008331	appliances

There is a pattern in the types of words that appear in all categories for helpful words with higher probabilities.

Words that describe a family member are common for Unhelpful words across categories (nephew, grandson, granddaughter). These are likely reviews for products meant as a gift.

Helpful words are those that pertain to the specifics of each category.

Helpful Summary Word Probabilities

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	performances	0.735191	use	0.038149	music
1	important	0.689542	happy	0.035696	music
2	power	0.689542	want	0.033540	music
3	read	0.689542	awesome	0.032749	music
4	beyond	0.624876	artist	0.031629	music
5	search	0.624876	song	0.029442	music
6	type	0.624876	four	0.015617	music
7	under	0.624876	three	0.014402	music
8	sad	0.624876	stars	0.007227	music
9	loves	0.624876	five	0.006528	music

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	stainless	0.751827	fits	0.036487	appliances
1	avoid	0.751827	fit	0.033618	appliances
2	food	0.707907	arrive	0.032564	appliances
3	reliability	0.707907	replace	0.031712	appliances
4	beautiful	0.707907	original	0.029404	appliances
5	blue	0.707907	item	0.026802	appliances
6	stop	0.707907	four	0.024482	appliances
7	steel	0.707907	three	0.020125	appliances
8	cool	0.679551	stars	0.016704	appliances
9	garbage	0.645097	five	0.016694	appliances

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	compact	0.692397	way	0.030315	automotive
1	scanner	0.628005	job	0.028766	automotive
2	inside	0.628005	installation	0.028766	automotive
3	mind	0.628005	tool	0.027367	automotive
4	mine	0.628005	bad	0.022910	automotive
5	give	0.584517	fine	0.020417	automotive
6	weak	0.529516	them	0.019035	automotive
7	repair	0.529516	stars	0.016598	automotive
8	base	0.529516	five	0.015468	automotive
9	lol	0.529516	four	0.009092	automotive

For summaries I found in every category that the most common unhelpful words are numbers and stars (i.e. “Five Stars”).

Helpful Review Word Probabilities (recent)

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	regardless	0.225628	stuff	0.000238	prime_pantry
1	hemp	0.172268	some	0.000235	prime_pantry
2	sesame	0.142732	flavor	0.000208	prime_pantry
3	gram	0.142732	taste	0.000203	prime_pantry
4	consume	0.099908	good	0.000186	prime_pantry
5	nutrient	0.097654	them	0.000182	prime_pantry
6	hardly	0.088525	these	0.000168	prime_pantry
7	seed	0.086875	try	0.000159	prime_pantry
8	wave	0.076850	love	0.000151	prime_pantry
9	fatty	0.076850	great	0.000043	prime_pantry

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	aftershot	0.513592	easy	0.001616	software
1	32	0.363130	works	0.001528	software
2	plugin	0.232875	hard	0.001335	software
3	color	0.197672	great	0.001271	software
4	black	0.191968	love	0.001226	software
5	transcription	0.174357	found	0.001226	software
6	spot	0.174357	dvd	0.001146	software
7	written	0.174357	computer	0.000858	software
8	proficient	0.174357	game	0.000837	software
9	lightroom	0.174357	learn	0.000621	software

	Helpful_words	Helpful_prob	Unhelpful_words	Unhelpful_prob	category
0	3rd	0.138617	hair	0.000408	luxury_beauty
1	warning	0.138617	bottle	0.000348	luxury_beauty
2	physical	0.118246	cream	0.000313	luxury_beauty
3	blackhead	0.107694	product	0.000302	luxury_beauty
4	coola	0.091386	love	0.000297	luxury_beauty
5	burnt	0.074470	good	0.000291	luxury_beauty
6	opt	0.074470	from	0.000259	luxury_beauty
7	mainly	0.074470	scent	0.000235	luxury_beauty
8	newborn	0.074470	great	0.000182	luxury_beauty
9	chipping	0.074470	smell	0.000134	luxury_beauty

Words that are subjective to the user have are unhelpful. Words like “great”, “love” and “perfect” are all unhelpful words while the helpful words are objective.

Applying Machine Learning

Data used for modeling

The inputs I used for modeling were the reviews, summaries and categories.

Reviews and Summaries were vectorized separately while categories were encoded.

The same train test split was used for all models.

Models and Evaluation

Models used: Naive Bayes, Random Forest, SGDC, Logistic Regression

All models were going to be used to predict helpfulness based on only reviews, only summaries, and all of the factors (reviews, summaries and categorical data).

F1 score was used to evaluate model performance.

A list of discrimination thresholds were used to optimize F1 Scores.

These models were going to be trained on all categories.

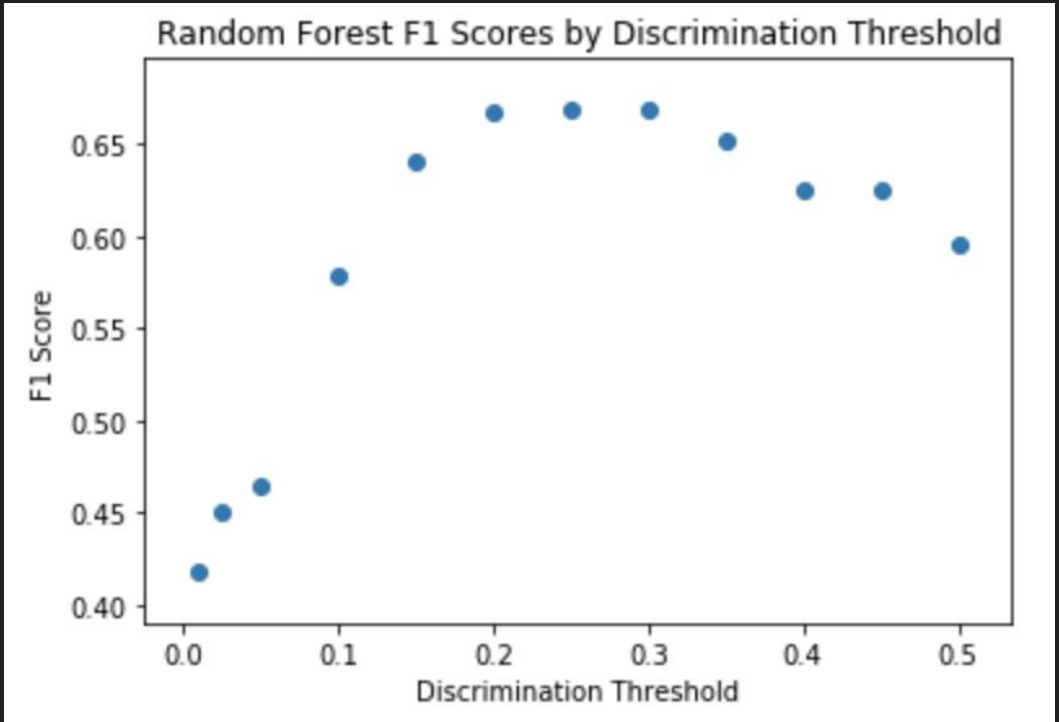
After identifying which model performed the best, this model would be used to train on each specific category then evaluating overall performance.

Results

	Review_Scores	Summary_Scores	Total_scores
NB	0.649979	0.612599	0.451159
RF	0.668417	0.590886	0.651031
SGDC	0.531218	0.563519	0.531218
Log_Reg	0.523929	0.565899	0.523929

We see that a random forest model that was trained on just the reviews outperformed all other models with an f1 score of .668.

Random Forest Discrimination Threshold



The model performs the best when the discrimination threshold is set at 0.3

Category Specific Vs Generalized Models

	category	category_specific	generic
0	office_products	0.599596	0.617719
1	toys	0.591861	0.639055
2	amazon_fashion	0.527179	0.584936
3	video_games	0.685761	0.692331
4	pet_supplies	0.597508	0.626997
5	all_beauty	0.584817	0.601474
6	automotive	0.590505	0.612326
7	tools	0.607845	0.626624
8	movies	0.712731	0.693148
9	music	0.661414	0.666858
10	grocery	0.590846	0.603313
11	kindle	0.516871	0.563855
12	software	0.677868	0.688860
13	musical_instruments	0.660593	0.661102
14	luxury_beauty	0.605751	0.598231
15	appliances	0.712780	0.693046
16	arts_crafts	0.703600	0.708822
17	industrial	0.603840	0.628368
18	prime_pantry	0.569972	0.593354
19	patio	0.639369	0.628803
20	cds_and_vinyl	0.725741	0.732843

The random forest model that was trained on all categories outperformed the category specific model in 17 out of 21 categories.

The categories that the category specific model outperformed the generalized model were movies, luxury beauty, appliances and patio.

The category specific model however outperformed the generalized model when all of the predictions were concatenated with an f1 score of .676 vs .668.

Closing Remarks

- Reviews that reference the product with terms that pertain to the category specifically are found more helpful.
- Words with a bad probability of being helpful are those that pertain to a family member. These are reviews for products that were purchased for someone else.
- Words that are subjective and share the reviewers sentiment are much more likely to be found unhelpful than words that are objective which describe the product.
- Words denoting a number and “stars” showed up as bad predictors for helpfulness in summaries.
- Reviews that are more critical of the product are also found more helpful than those that praise the product.
- Time has a dramatic effect on the helpfulness of reviews.
- Helpful reviews and summaries are longer than unhelpful reviews and summaries.
- A random forest model only using reviews as an input yields the highest f1 score.

Fin.