Learning category-wise product features from amazon reviews

Yan Qi

Posted on Aug 8, 2019

Overview

As a long-time Amazon Prime member, I rely heavily on the product reviews for my Amazon purchases. I

consider online reviews a very valuable source of information for consumers.

Typically, I would first look at the number of reviews, then the overall numeric rating and its distribution, and read through some text reviews to pick out the customers' likes and dislikes and key features of the product. When reading the review text, unconsciously, I'm also filtering for features that matter to me. Although very useful, digesting and synthesizing qualitative comments this way is rather time consuming and subject to bias, depending on which reviews I decide to read. Wouldn't it be great if we could be presented with a concise and statistically sound summary that captures the essence of all the reviews?

The cognitive burden is aggravated even more when one tries to compare numerous similar products, each with a large number of reviews. As a good comparison shopper, I often find myself evaluating all the options against a set of common features, explicitly or implicitly. For example, when purchasing shoes, one may consider material, style, fit and price. Obviously, different people may place different importance on those features. Reading reviews is a great way to tease out how well the product performs in different aspects. Can machine learning help us systematically, efficiently, and objectively digest the vast number of reviews so that we can match product features with our personal preferences?

These are the problems that I set out to explore in my capstone project, where I used natural language processing techniques to extract meaningful features (which I'd also refer to as topics) common to products within a similar category, based on customer reviews.

R SHINY 412 posts
R VISUALIZATION 353 posts
STUDENT WORKS 1157 posts
WEB SCRAPING 359 posts

Search For

Our Recent Popular Posts

How to Automate Your eCommerce Business with AI Technology in 2020
by Matthew Fritschle
Jan 16, 2020

Predicting NICU Admissions and CCHD
by Paul Lee, Aron Berke, Bee Kim, Bettina Meier and Ira Villar
Jan 7, 2020

Why R is a Must for Data Scientist?
by Aiko Liu, H Zhang and

Got any questions? I'm happy to help.

View Posts by Tags

The output from this project would enable the following use cases:

- A customer would be able to
  - Browse top features learned from reviews across all products in a category and focus further evaluation on only products with features that he/she cares about
  - Look at the most salient features of an individual product and read a few representative reviews for each feature of interest
- A manufacturer or seller would be able to treat top-features extracted from reviews as real-word customer perception of the product. Comparing that against its intended positioning or offerings from competitors and potentially tracking the evolution of perception over time, can inform sales & marketing strategy, as well as product design and customer service.

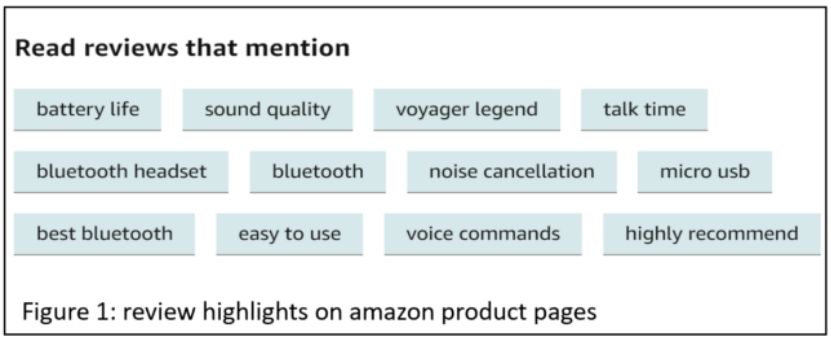


Figure 1: review highlights on amazon product pages

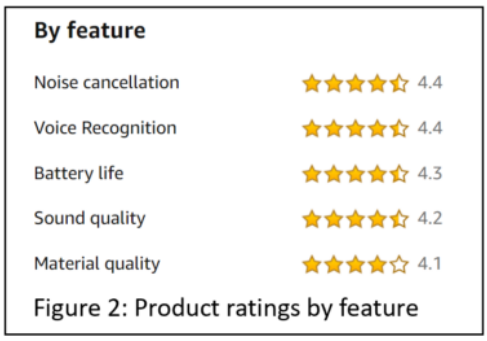


Figure 2: Product ratings by feature

Currently, on each product's page, Amazon provides "review highlights" in the form of a list of phrases summarized from its reviews (figure 1). Most likely, this is based on topic modeling of reviews for one product at a time. Each product will get a different list of unique topics. In my project, a common list of topics is learned from reviews of all products in the categories.

Comparisons across similar products can thus be more easily made along those features.

For some products, a “Rating by features” section is available (Figure 2). It is not clear whether customers were asked to specifically to rate on each feature, or if statistical models were used to obtain the “decomposed” ratings. Although scoring along individual features has not been implemented in this project, I see it as a natural next step in future development. An approach that unifies the overall numeric rating and sentiments on each feature based on the review text would be a very interesting direction to explore. The code for this project can be found on my github repository

Data selection & basic cleaning

For this project, I used the Amazon datasets Dr. Julian McAuley shared on his lab’s website. This dataset contains Amazon product reviews and metadata spanning May 1996 - July 2014. The review data includes numeric ratings, text, helpfulness votes, time of review, etc., and the product meta data includes product category, brand, name, and description. The 5-core dataset where each user and product have at least 5 reviews was utilized.

The Amazon datasets are classified into broad categories, such as “clothing, shoes and Jewelry”, “electronics,” “books” etc. The meta data includes the hierarchical categories that a product belongs to. For the business use I envisioned, I wanted to focus on a sub-category that’s fairly specific but still includes a large number of products and reviews where users can benefit from machine learning. Then I remembered the time when a friend was shopping for a bluetooth

headset on Amazon and complained that the number of choices was overwhelming, and it was too laborious to read through so many reviews. With these considerations in mind, I picked “Bluetooth Headsets,” a subcategory within “Cell Phones & Accessories,” which includes 10284 reviews for 422 products from 2004 to 2014.

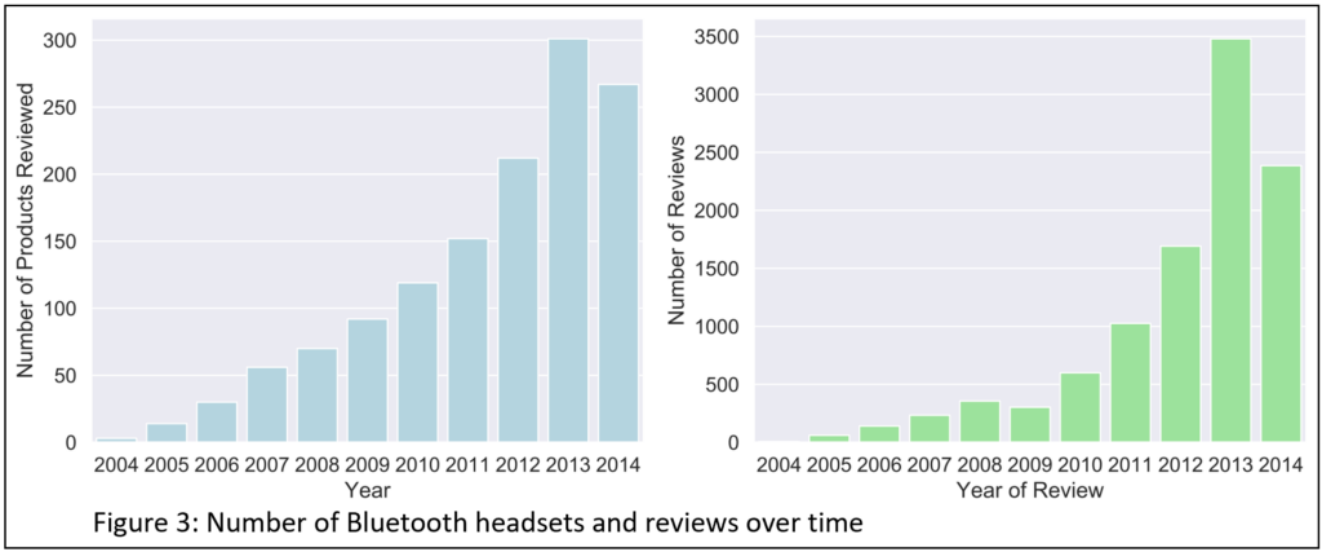


Figure 3: Number of Bluetooth headsets and reviews over time

From here on, some simple cleaning steps were applied, including removing html tags, dropping a few records with missing review text or product name, filtering out products that are not truly Bluetooth or headsets based on keywords in product names. Missing values in the original “brand” column were filled by the brand in the “title” (product name) column.

Review text processing

The first step in natural language processing is always to transform text into a format that the computer can understand, i.e. numbers. The transformation depends on the type of modeling intended. As I was planning to do topic modeling on the reviews, I treated each review as one document and the reviews for the 390 Bluetooth headset as the collection. I then used the Python packages NLTK, spaCy, and regex-based custom

functions to process each review following the steps in Figure 4.

There were quite a lot of HTML entity numbers and tags in the review text, such as "'" and "<a href='xyz.html'>". These can cause problems for subsequent steps when punctuations were removed and nonsensical words could form. Applying the html parser from BeautifulSoup package got rid of the HTML entity numbers and tags. Expanding contractions means mapping "aren't" to "are not" and "it's" to "it is" etc while removing punctuation means getting rid of symbols such as !, &, %. Since our aim is to extract topics from text, we are only interested in words and phrases, hence the symbols can be safely removed. However, punctuation may be useful for some NLP tasks. For example, double exclamation marks could mean more intense emotions in sentiment analysis.

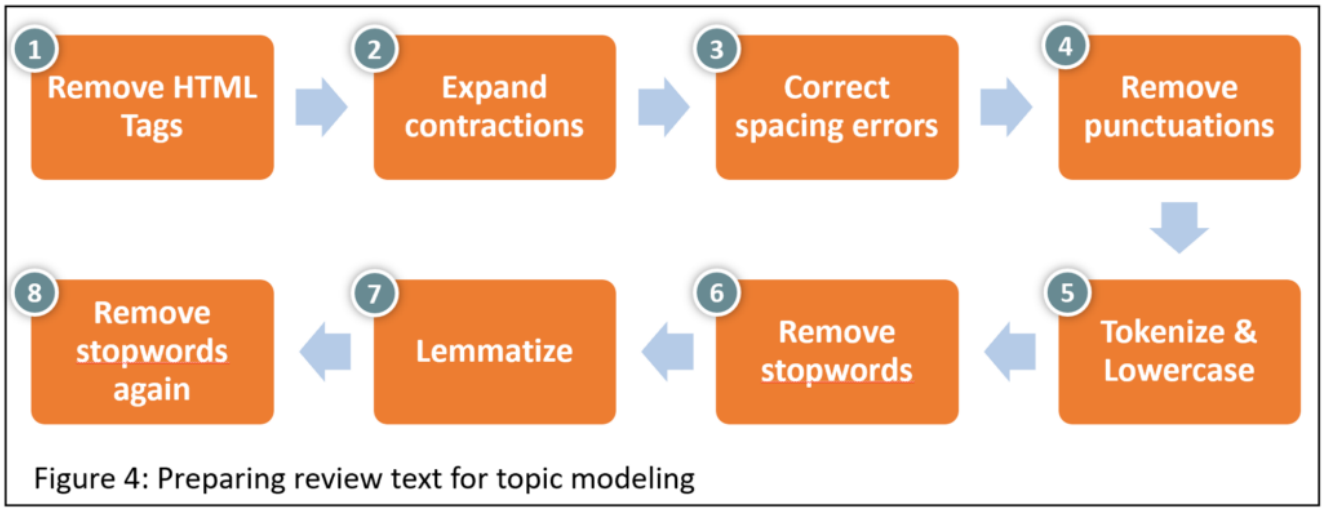


Figure 4: Preparing review text for topic modeling

Next each review is split into individual words (tokens) and converted to lowercase. To remove stopwords, it is often necessary to extend the list of English stopwords from NLTK with frequently occurring but trivial words that are specific to the dataset at hand. In this case, "headset", "Bluetooth", "headphone", "headphones" would appear in almost every review but have little

value as part of a topic. Those words were added to the stopwords list and removed.

Lemmatization refers to the process of converting various forms of a word to its canonical form, also known as lemma. For example, the lemma for “teaches, teaching, taught” is “teach,” the lemma for “mice” is “mouse,” and the lemma for “best” is “good” (with spaCy lemmatizer). I chose the spaCy lemmatizer over others as lemma depends on the POS (part-of-speech) of the word, and the spaCy lemmatizer includes a step to determine the POS before assigning the corresponding lemma.

As lemmatization produced some additional stopwords, another round of stopwords removal was applied. Numbers were also removed, as they were not likely to be part of an interesting topic. Apart from the more standard text processing above, some customized cleaning was done to handle the special “noise” in this dataset such as missing space in “This is great.Recommend to anyone” and patterns such as “wordA…wordB…wordC” and “wordA/wordB”.

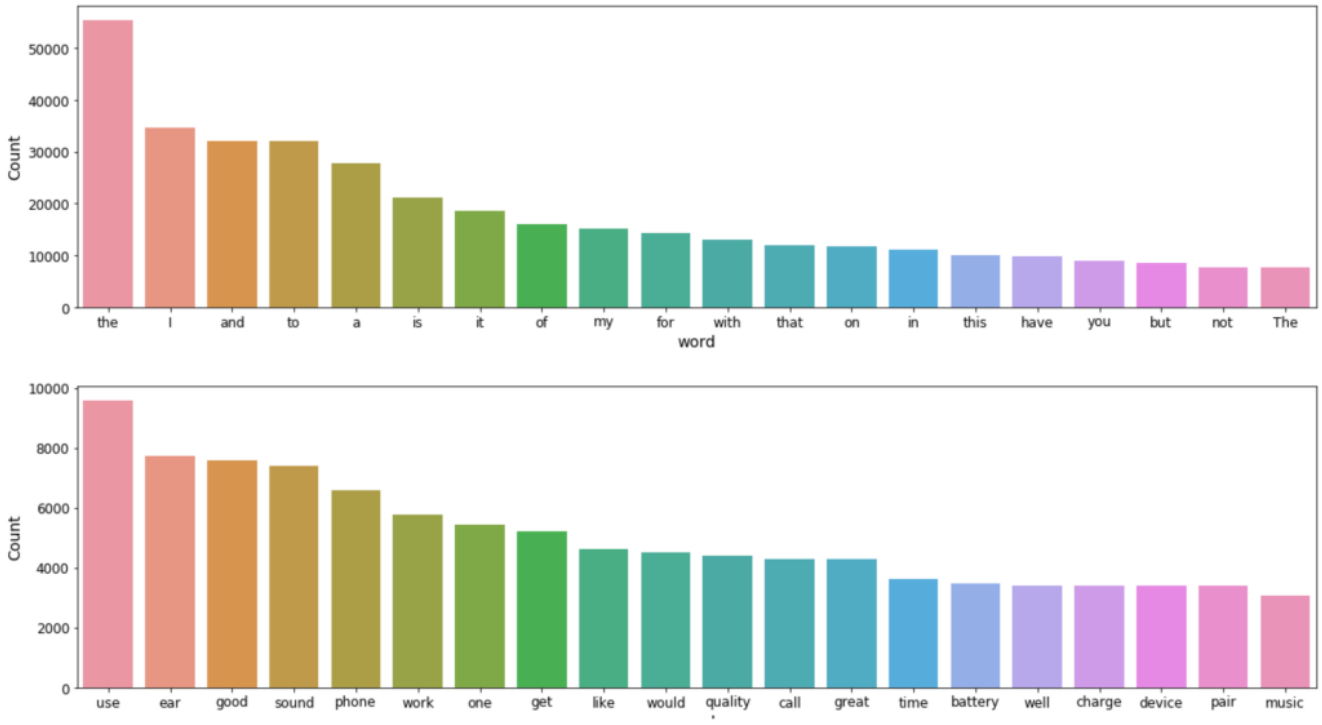


Figure 5: Top 25 most frequent words in all review text before and after processing

Figure 5 shows the top 25 most frequent words from all the review text before and after the text processing steps. The most frequent words in the cleaned text provides much more insights about Bluetooth headsets than those from the original reviews.

Topic modeling with LDA

The purpose of topic modeling is to extract "topics" from a collection of documents. Applications include document classification and summarization. On our Bluetooth headsets review data, the goal is to discover a list of meaningful, non-overlapping, exhaustive "topics" that best reflect the product features and aspects that customers commented on.

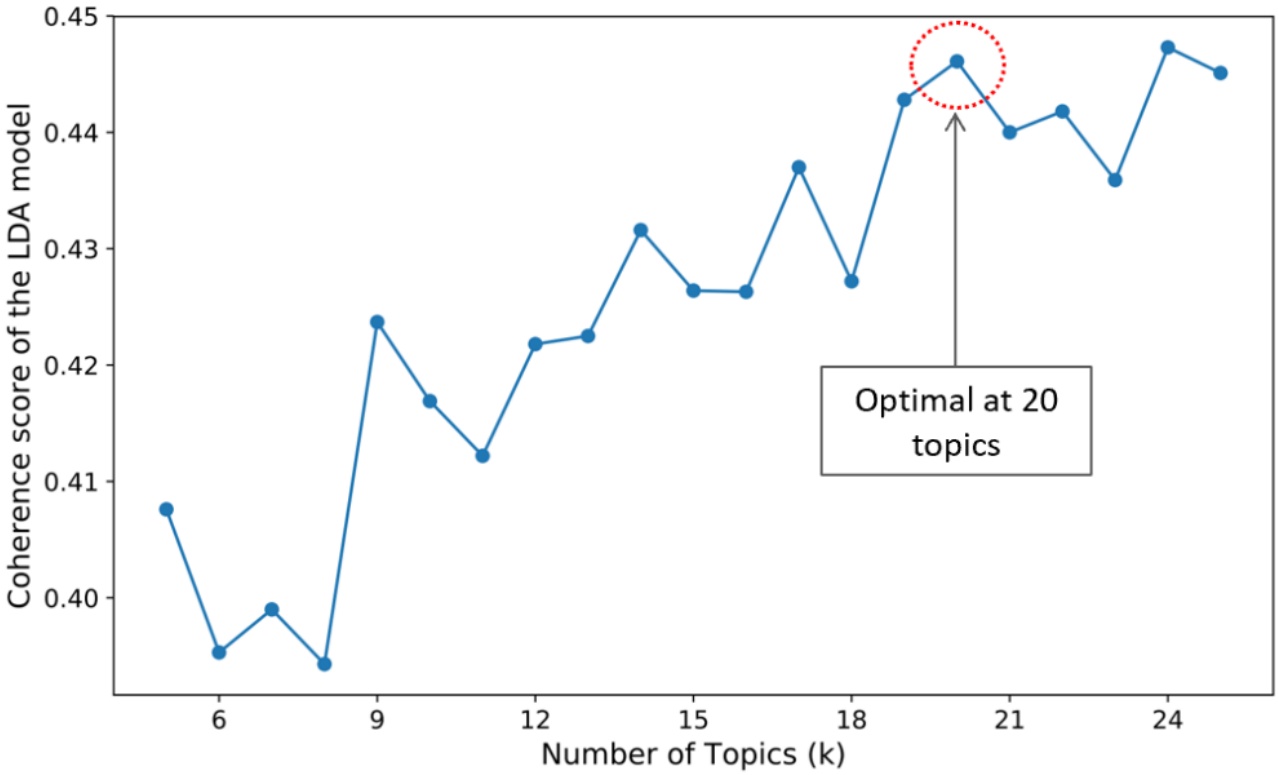
Latent Dirichlet Allocation is a generative probabilistic model for a collection of documents (corpus). Each document is generated from a random mixture of latent topics and each word in a document is generated from multinomial distributions conditioned on topics. The Goal of LDA topic modeling is then to infer the

unobserved “topics” and the associated probability distributions from the observed data. In the model output, each topic is represented by a weighted list of words, and each document is assigned a weighted list of topics, where the weights represent multinomial probabilities. For each topic, weights for all words add to 1. And for each document, weights for all topics add to 1. This probabilistic model makes sense for review data, where each review could address several topics with different amount of emphasis, and different reviewer may talk about the same topic using slightly different wording.

In my implementation, I used Gensim (python package) for preparing the corpus and Mallet (Java package) with Gensim wrapper for LDA modeling. I experimented with both Gensim and Mallet for LDA modeling and ultimately decided to go with Mallet for this dataset. Gensim LDA has a lot of nice features such as online learning and multicore mode and constant memory requirement, which makes it well suited for large-scale applications. However, as some articles have pointed out, Mallet often outperforms Gensim in terms of the quality of the topics learned. Indeed, in my own experiments on two amazon review datasets (Bluetooth headsets, Laptops), Mallet LDA with default setting gave better or similar results than Gensim LDA with optimized parameters. Here the “Goodness of model” was assessed by the topic coherence score in Gensim, a mutual-information based quantitative measure of how coherent a list of words are for describing a common topic. Mallet LDA produced topics with higher coherence scores than those from Gensim LDA, and they were also more interpretable upon human inspection. This better performance is likely due to the fact that Mallet uses an optimized version of collapsed

Gibbs Sampling, which is more precise than the variational Bayesian optimization used by Gensim.

Figure 6: Optimizing number of topics in LDA model



We can use the topic coherence score as the metric for optimizing the number of topics in an LDA model. On the Bluetooth headset data, we obtained the best model with 20 topics.

Topic interpretation and labeling

The topics returned by LDA are lists of words. Some human inspections are still needed to turn those lists into more interpretable "topics." When the quality of the topic model is good, the word list itself can be very telling. Can you guess from the lists in Figure 6 what those topics are?

Topic 7: 0.340*"ear" + 0.100*"fit" + 0.048*"piece" + 0.041*"small" + 0.029*"stay" + 0.028*"size" + 0.028*"bud" + 0.025*"fall" + 0.024*"hook" + 0.012*"loop"

Topic 16: 0.120*"call" + 0.090*"voice" + 0.070*"phone" + 0.037*"feature" + 0.029*"answer" + 0.024*"button" + 0.021*"app" + 0.021*"command" + 0.014*"number" + 0.012*"turn"

Topic 8: 0.227*"great" + 0.100*"love" + 0.062*"work" + 0.057*"recommend" + 0.052*"easy" + 0.030*"comfortable" + 0.025*"highly" + 0.024*"purchase" + 0.021*"blue" + 0.020*"awesome"

Figure 7: Topics output from LDA model as weighted list of words

To be more precise and get a bit more color, I reviewed the topic word lists, as well as the review texts with the highest probabilities being generated from each topic, in order to label the topics with a short name and a detailed description (Table 1).

Topic Number	Short Name	Meaning
0	**buttons**	various buttons e.g. power and volume control
2	**fit_head**	Headset's fit on the head e.g. with glasses
7	**fit_ear**	fit on into/ear, due to e.g. earbud depth, cover, ear hook shape, ear wires
6	**phone_connection**	connection between headset and phone, signal range, reconnection
13	**charging**	everything related to charging the headset: cables, charging speed, via car-charger
19	**battery**	Quality of battery e.g. talk time on full charge, battery indicator, deterioration over time
3	**noise_cancellation**	whether background noise can be heard, especially in cars, whether has noise cancellation feature
14	**audio_performance**	experience (not just referring to sound) listening to music, audiobooks, video
9	**sound_quality**	sound quality, sometimes with detailed evaluation, bass
16	**voice_command**	quality of voice command functionality/app for making and taking phone calls
4	**misc_issues_problems**	miscellaneous problems and issues (negative)
17	**customer_experience**	Customer service experience related to shipping, return, replacement
5	**Cheap**	cheaply priced (positive: good quality for the price; negative: breaks easy don't buy)
8	**great_overall**	highly satisfied overall, especially good quality for the price
18	**active_lifestyle**	Usability in an active/sports setting, such as during running and gym workout
10	**Plantronics**	(primarily) high praises for the Plantronics brand
12	**Jabra**	for the Jabra brand
1		
11	Low quality topics: no obvious theme identified	
15		
Legend:	Individual feature or function	Holistic assessments / Specific brands

Table 1: Labelled and interpreted topics

Out of the 20 topics, I could easily interpret 17 of them. Out of these 17 topics, two identifies the brands Plantronics and Jabra (blue rows in Table 1). "Plantronics" topic mostly captured high praises from loyalists who have owned many products from this brand.

The remaining 15 topics cover diverse aspects highly relevant for Bluetooth headsets. Ten topics are about individual product features (green rows in Table 1), such as the quality of voice command functionality, sound quality, and noise cancellation capability. The model also correctly separated fit of the headset on the head vs. fit in/on the ear into two topics. The other 5 topics (yellow rows in Table 1) were not referring to just one product feature, but more holistic assessments. For example, topic 4 is about “miscellaneous issues and problems” and the actual review often contained such language as “issues, problem”. Topic 17 was not about the product itself, but the overall customer service experience especially related to shipping speed, ease of exchange and returns, through the manufacturer or amazon. Topic 18 highlights product usage in an active setting such as during “running or gym workout”. Topic 8 captures an expression of high overall satisfaction.

Those 15 topics thus succinctly summarizes the aspects that customers consider important for evaluating Bluetooth headsets based on their reviews for all products in this category.

Drawing insights from topic distribution in reviews and products

Now that we have a set of well-labeled high-quality topics learned from the data, there is a lot we can do to derive insights on the products. We start by looking at what topics are present in each review, then calculate the extent to which each topic is present in a product’s review.

From the trained LDA model, we can obtain a matrix of document-topic probabilities X where each row

corresponds to a document and each column corresponds to a topic. Element *Xik* represents the probability that review *i* is generated from topic *k*. Each row contains the probabilities that a particular review is generated from each of the 20 topics and those probabilities sum to 1. Each column contains the probabilities of every review being generated from a particular topic.

It turns out that the probabilities in all 20 columns follow a similar distribution. For each topic, more than 80% of all reviews have a probability that is less than 0.1. Based on these observations and inspection of review text for various topics with different probabilities, we arrived at a cutoff of 0.09. With that we can turn the probability matrix X into an indicator matrix Y where *Yik* is 1 only if review *i* contains topic *k*.

So how many topics does a review contain according to our LDA model? Well, that depends on the length of the review. Most reviews are short. More than 50% of all reviews have fewer than 40 tokens each. There are also, however, some extremely detailed reviews. About 7% of reviews are more than 200 tokens long. Some techie users are very serious about their Bluetooth headsets and create head-to-toe meticulous reviews.

Figure 8: Relationship between number of topics and length of review

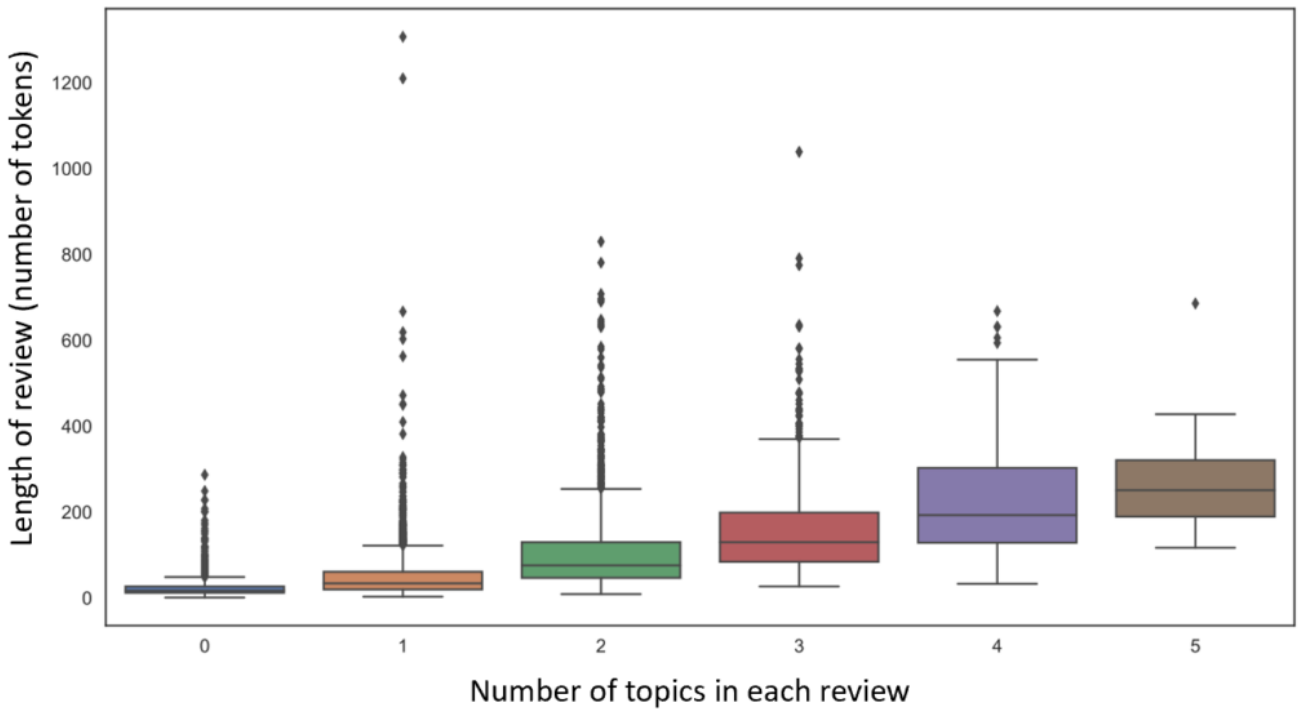


Figure 8 shows that our model does not capture any topic when the review is extremely short and simply does not contain enough occurrences of the keywords that comprise a topic. Longer reviews do tend to contain more topics, with a maximum of 5 on this dataset. But longer reviews are not always better. Since for each review, the probabilities for all topics are forced to add to 1, topics have to compete with each other to stand out. Only the few strongest ones win. This explains why only 1 topic was detected in each of the two longest reviews. When one read through these two reviews, they apparently just ramble on and on, seeming to be talking about everything, though nothing really stands out. Even though the reviews definitely touched upon several topics, these topics inevitably were buried in too much noise. Conversely, the reviews with 5 topics are on average 200-300 words long, with focused discussion on each topic. To the human eye, those are “well written” reviews, conveying clear messages, providing

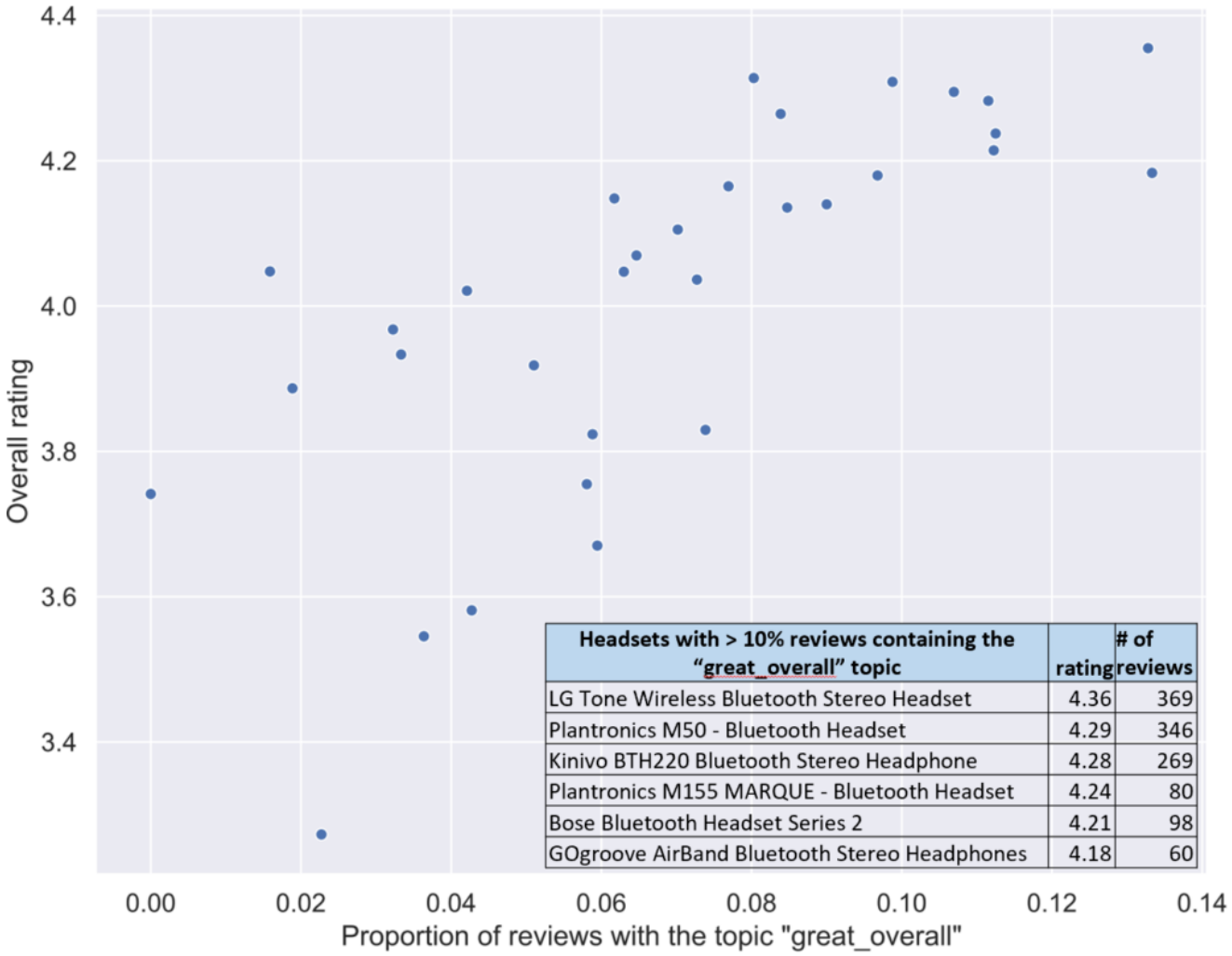
just the right amount of details, but not so much to induce cognitive fatigue.

The presence/absence of topics in reviews can be aggregated at the product level to create a matrix **Z** where rows correspond to products and columns correspond to topics and element *Zjk* represent the proportion of reviews for product *j* that contain topic *k*. This "product-topic proportion" matrix **Z** would allow us to ask many interesting business questions. We demonstrate a few below.

Case 1 - Topics that predicts numeric rating

Let's start with a simple topic "great_overall". Typically, reviews with this topic express a high level of overall satisfaction, with strong positive phrases like "awesome" or "highly recommend". The scatterplot in Figure 10 shows the products with a higher % of reviews containing the topic "great_overall" tend to have a higher overall numeric rating. The 6 products with > 10% of reviewing containing "great_overall" are among the top 10 most highly rated Bluetooth headsets. This feature can be a good predictor for the product's overall high performance. Similarly, the % of reviews mentioning topic "misc_issues_problems" correlates negatively with a product's numeric rating.

Figure 9: Product rating vs. % of product reviews with topic “great_overall”



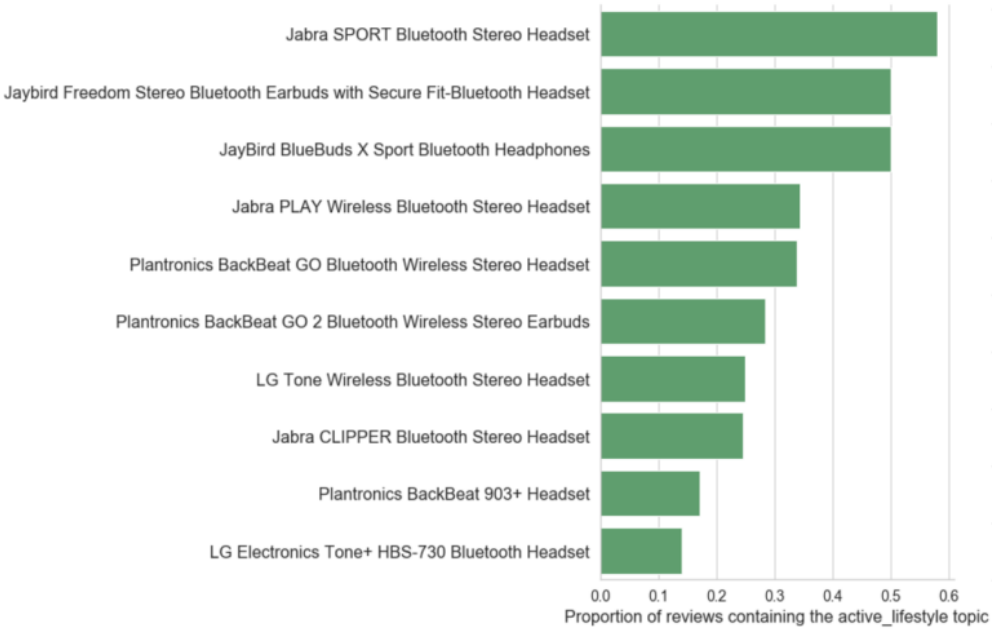
Headsets with > 10% reviews containing the “great_overall” topic	rating	# of reviews
LG Tone Wireless Bluetooth Stereo Headset	4.36	369
Plantronics M50 - Bluetooth Headset	4.29	346
Kinivo BTH220 Bluetooth Stereo Headphone	4.28	269
Plantronics M155 MARQUE - Bluetooth Headset	4.24	80
Bose Bluetooth Headset Series 2	4.21	98
GOgroove AirBand Bluetooth Stereo Headphones	4.18	60

Case 2 – Headsets most suitable for an active lifestyle

As another example, a customer might be looking for Bluetooth headsets that he can use while running or working out in the gym. We would look at all products with a minimum number of reviews (say 50), and check out the top 10 with the highest product-topic proportions for the topic “active_lifestyle”. Two of the top products from Jabra and Jaybird indeed have SPORT in their names and are designed for this setting. By looking at the overall rating, two LG products are the only ones with > 4 stars. They also have larger number of reviews than the other 8 products. The customer may decide to only consider the headsets in this list with a

rating > 3.7 and further evaluate based on price and other features.

Figure 10: Products most mentioned for usage in "active_lifestyle"



Overall Rating	total # of reviews
3.27	88
3.97	62
3.92	98
3.74	58
3.82	68
3.75	155
4.36	369
3.89	53
3.58	117
4.18	434

Case 3 – Perception of most reviewed products

In the last case, we will take a look at how the most reviewed Bluetooth headsets are perceived by users, reflected by the top topics that are most associated with those products. LG and Plantronics are two renowned brands in the Bluetooth headset space. Among the 5 most reviewed (based on the number of reviews) products, 2 are from LG and 2 are from Plantronics. If we take our product-topic proportion matrix **Z** and use a cutoff of 0.08, we get 4 or 5 top mentioned topics for each product (Figure 11). The topics for products from the same brand are fairly concordant while we can see clear differences between the topics across brands. While customers of LG often talked about its use in an active setting, the phone connection capability, and audio performance, Plantronics headphones seem to stand out with the noise cancellation and voice command functionalities.

Figure 11: Top features for most reviewed LG and Plantronics headsets

LG Electronics Tone+ HBS-730	LG Tone Wireless Bluetooth Stereo Headset	Plantronics M50	Plantronics Voyager Legend
4.18 stars (434 reviews)	4.36 stars (426 reviews)	4.29 stars (346 reviews)	4.07 stars (137 reviews)
phone_connection (19%) *active_lifestyle (14%)* *great_overall (10%)* *audio_performance (9%)* *customer_service (8%)*	*active_lifestyle (25%)* *great_overall (13%)* *phone_connection (10%)* *audio_performance (9%)*	*noise_cancellation (12%)* *great_overall (11%)* *fit_ear (11%)* *voice_command (9%)* *battery (9%)*	*voice_command (21%)* *charging (18%)* *customer_service (12%)* *noise_cancellation (12%)*

Conclusions and future development

The business cases explored here demonstrate how topic modeling on customer review data is useful for both customers and e-commerce companies. Other potential uses include product segmentation, rating predictions, and predicting topics on unseen reviews. A really neat expansion of this work is to calculate customer satisfaction along each of the identified topics (features) via sentiment analysis. This would allow us to then see not only which features are most talked about for each product but also whether the customers thought of that aspect of the product negatively or positively. Given more time, it will also be easier to visualize the results through a web-app powered by Flask or R shiny where users can explore different business cases to mine insights.

About Author

Yan Qi

Yan is an experienced business analytics professional with well-balanced skills in quantitative analysis, strategic thinking, and communications. She received a Ph.D. in biomedical engineering from the Johns Hopkins University. During graduate studies, Yan innovated statistical inference and machine...

View all posts by Yan Qi >

Related Articles

STUDENT WORKS

Latent Dirichlet Allocation and Topic Modelling: A Link Between The Quantitative and Qualitative?

by Kyle D. Weber

Nov 4, 2019

Introduction/Motivation: When discussing the application of machine learning techniques and the use of data science in industrial settings,...

Continue Reading

MACHINE LEARNING

Optimize Conversion Rates with Sentiment Analytics

by Ted Dogan

Oct 20, 2019

I decided to apply sentiment analytics to understand how to go about improving customer targeting optimizing conversions. In...

Continue Reading

Leave a Comment

You must be logged in to post a comment.
No comments found.