Drashti Shah, Cenrara Widi, Jaejae Zhang
APAN 5205: Applied Analytics Frameworks & Methods II

Amazon Electronics Review Project Report

Launching in 1994, Amazon has become a global e-commerce giant and has surpassed Walmart to become the world's largest retailer. From exclusively focusing on books, Amazon now offers a massive portfolio of products from electronics to fresh produce. According to Statista, the second-largest share of fake product reviews in Amazon worldwide as of December 2018 is the Electronics category, reporting 61% fake reviews which makes exploring this category most interesting. (Statista, 2018) Our team's interest in text mining, passion for the latest electronics products, and the fact that one of us is a member of an incentivized review pumping group currently on Whatsapp led us to choose the Amazon Electronics review dataset (Dataset Source). Overall, we are interested in the following research questions:

- Does the timestamp make any impact on the review score?
- Can we predict the review score based on review text and verified reviews accurately?
- Is it possible to predict the review sentiments using natural language processing?
- Are there any particular trends in the review score ranging from 2016 to 2018 depicting Amazon's efforts in combating the fake reviews?
- Are the reviews for certain products such as earphones, headsets, etc more inclined towards higher review scores and positive sentiments?
- Can we visualize any structure in words in the text reviews using sentiment analysis?
- Do verified reviews display more positive sentiments?

We are using two datasets for this purpose- The first dataset has over 20 million rows of electronics products sold on Amazon dating from 1996 to 2018. The second dataset has the metadata which includes description, price, brand info, and co-purchasing links. It has over 700,000 rows dating from 1996 to 2018. Both these datasets contain information such as ratings, product title, the main category that the product belongs to, text reviews of the users, etc. This makes it suitable for sentiment analysis to understand the social sentiments around the products, and text analytics by transforming unstructured review and title text into a structured format to identify meaningful patterns and use text mining to predict review scores. We narrowed our analysis for the period from 2016-2018 to check the true occurrences of fake reviews in recent years and to make the dataset more manageable for R to handle. This timeframe covers the period up to and after Amazon changed its terms to stop incentivized reviews.

Final Dataset – The table below represents the combined dataset statistics:

Number of Reviews	8,951,484
Number of distinct users	5,112,945
Number of products	443,699
Period of Analysis	Jan 2016 - Dec 2018

Drashti Shah, Cenrara Widi, Jaejae Zhang
APAN 5205: Applied Analytics Frameworks & Methods II

Amazon Electronics Review Project Report

---

1. Data Cleaning

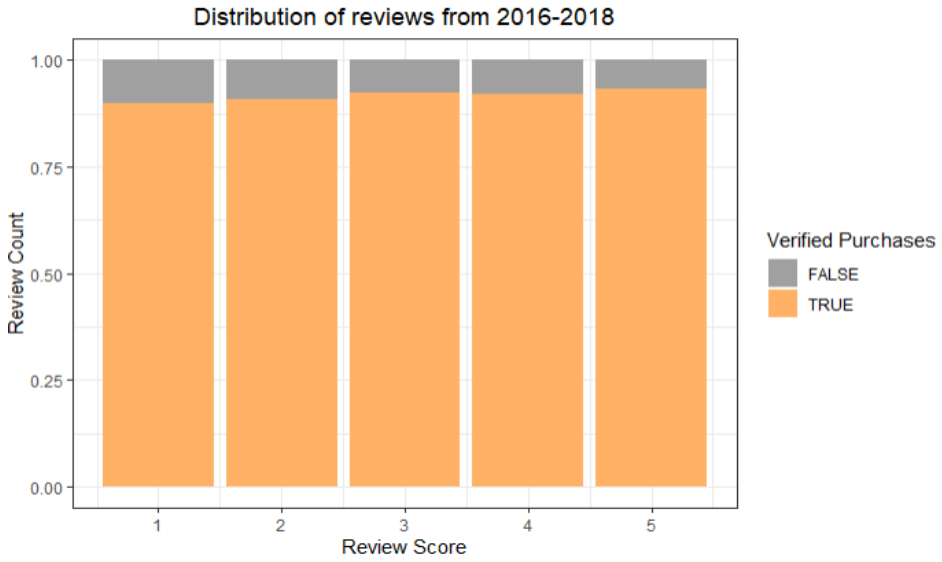
a. From the combined dataset, the records pertaining to categories such as *Books, Collectible Coins, Gift Cards, and Grocery* which is not relevant to the *Electronics* category are removed.

b. The data is cleaned of very few missing values in the title, main category, and review text columns.

c. The text of the reviews, title, and main category is parsed and cleaned from special characters, punctuation, and non-graphical characters (backspace and tab).

2. Exploratory Analysis

a. Below is **the distribution of review scores** showing almost 90% of purchases across the review rating scores (1-5) are verified (i.e. reviews are from people who didn't receive the product at deeply discounted rates).

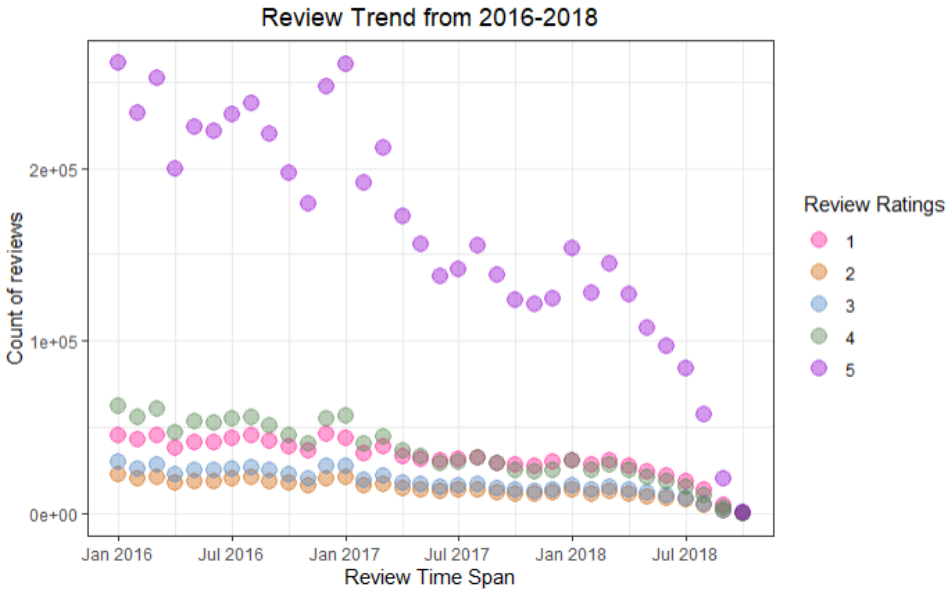


b. **The top three most reviewed main categories** are Computers (3.3 million), All Electronics (1.3 million), and Home Audio and Theater (1.2 million). The review scores in all main categories are dominated by 5 stars followed by 4 stars.

c. *Review Trend:* Despite 90% of reviews being verified, the 5-star reviews have been consistently plentiful, which seems a bit unusual. One might ask if people are generally optimistic, or most possibly there are still fraudulent reviews from purchasers part of Whatsapp, Facebook, and WeChat groups that are artificially inflating the seller's reviews. These accounts might still be getting the verified badge from Amazon. The trend starts to change in early 2017 shortly after Amazon's community guidelines changes to stop compensated product reviews. Additionally, there is a slight increase in reviews in Jan 2016, Jan 2017, and Jan 2018. This could be due to the

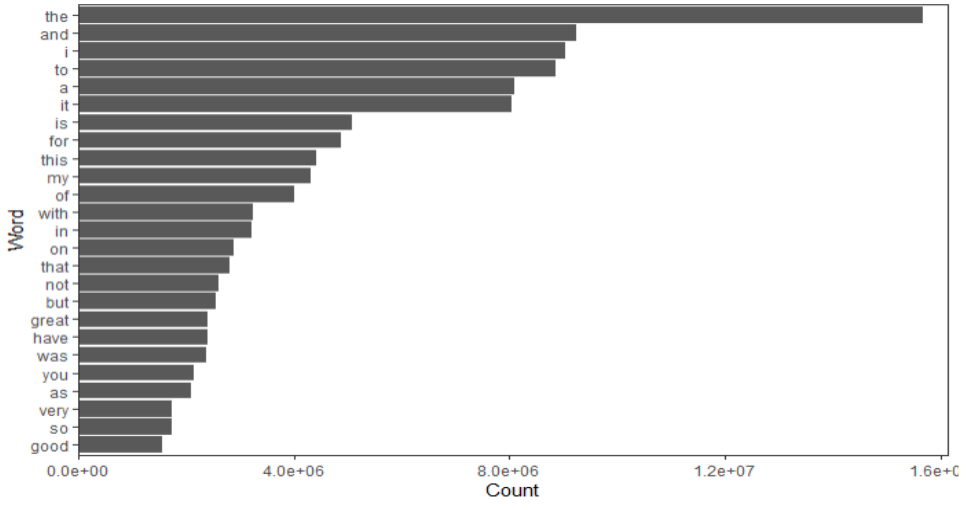
Amazon Electronics Review Project Report

jolly holiday spirit.___



d. *Correlation between Review Length and Review Scores*:
The length of the review does not impact the review score depicted by the weak negative correlation between the number of words/characters in a review and review scores.

e. *Most Frequent Words including stopwords:*
The most frequent words in review text are the unimportant words such as “the”, “and”, “I”, “to”, “a”, etc. which need to be filtered out.



Amazon Electronics Review Project Report

---

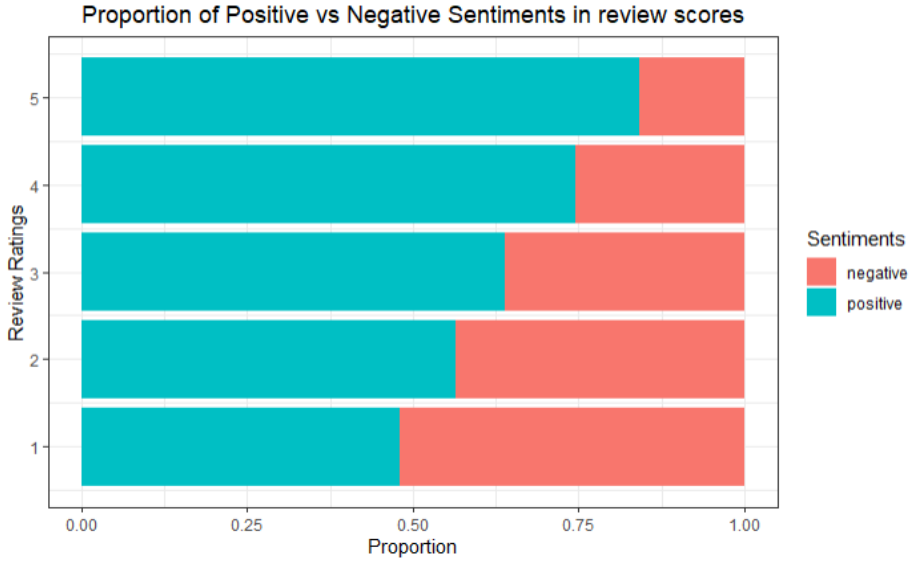
f. *Most Frequent Words excluding stopwords:*
After removing meaningless words, the most frequent 100 words in review text are as below with the top being “product”,” quality”, “sound”, “price” etc.

g. *Comparison Word Cloud:* After splitting the above word cloud of 100 words into positive and negative sentiments, we get a clear picture of what complaints and praises people have regarding the products bought. A few negative terms that stand out are “hard”, “issues”, “cheap”, “bad”, “broke”, etc. A few praises that stand out are “love”, “easy”, “nice”, “perfect”, “recommend”.

Amazon Electronics Review Project Report

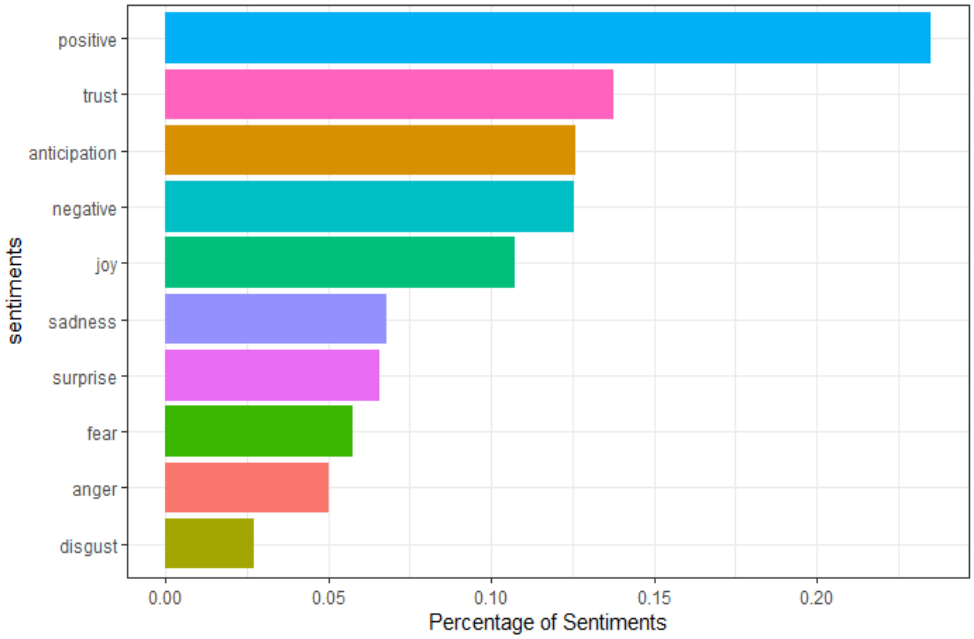
3. Sentiment Analysis

a. **According to Bing lexicon**, the ratio of positive to negative words is higher and the correlation between positive words and review scores is almost 52%. This indicates that almost 50% of reviews with positive words are rated favorably/higher.



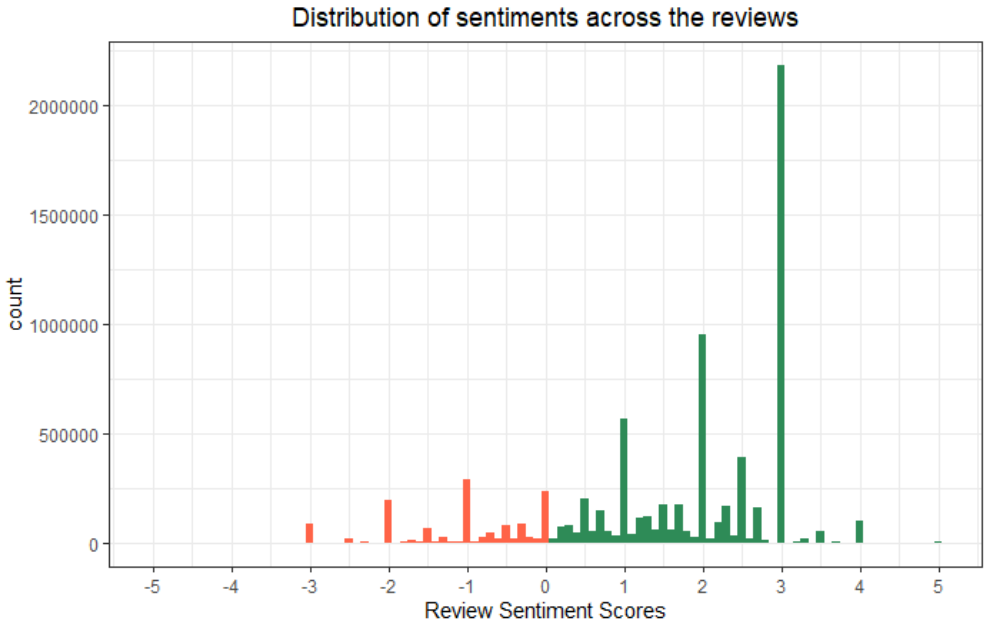
Amazon Electronics Review Project Report

b. **Applying the NRC emotion lexicon**, this bar plot shows a quick and easy comparison of words associated with each emotion in the text. The emotion "positive" has the longest bar and shows that words associated with this emotion constitute just over 20% of all meaningful words in the text. On the other hand, the emotion of "disgust" has the shortest bar and shows that words associated with this negative emotion constitute only 2% of all meaningful words in this text of reviews. There is an interesting split found when looking at the words related to the reviews. 50% of words were associated with positive emotions of "positivity"," trust" and "joy", while 33% of words were associated with negative experiences like "anger", "sadness", "fear", "negativity" and "disgust". It can be interpreted that when a buyer decides to leave a review on a product, there is a higher chance of it being a positive experience with the vendor. Furthermore, the review ratings are not tied to any frequency of emotions expressed noted by the absence of correlation. So, review ratings are mostly random in nature. A potential factor that might be affecting this result is that there still may be a slight impact of the artificial positive reviews with higher review scores but average review text even after the policy change.



Amazon Electronics Review Project Report

c. **According to the Afinn emotion lexicon,** more than 2 million reviews have a sentiment score of +3. The average sentiment throughout the reviews is trending positive.



4. Predictive Analysis

a. *Data Preparation:*

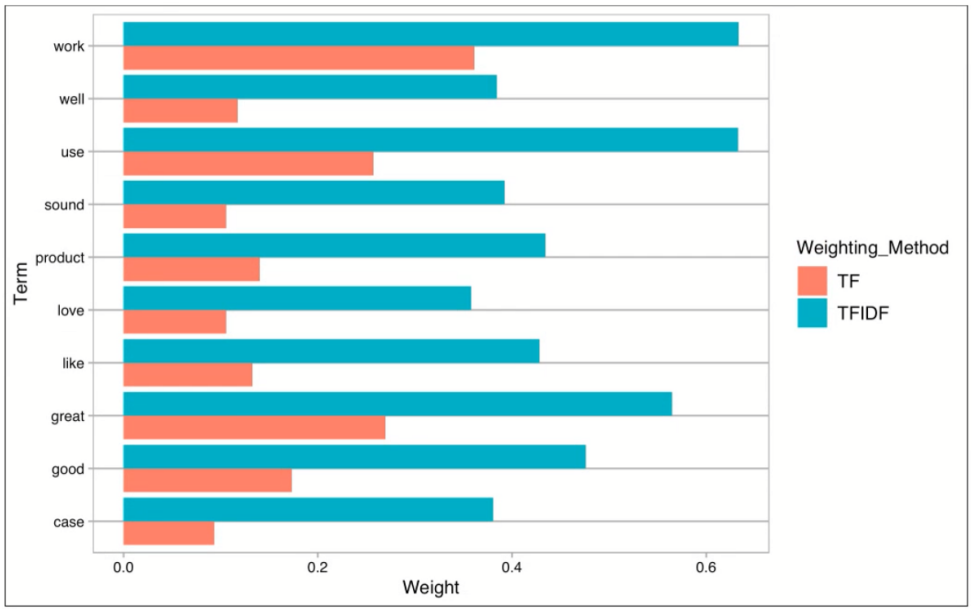
We use the bag of words approach in which a corpus is created for each review and the reviews within the corpus are cleaned again by removing punctuation, stopwords, whitespace, and converting the text to lowercase.

b. *Tokenization:*

Once the corpus is clean, we begin the tokenizing process where we extract individual words and generate a document term matrix derived from term frequencies. The document term matrix contains a very large number of tokens that cause more variables than observations. So, we remove infrequently occurring words that appear in fewer than 5% of the reviews. We only retain all terms that appear in 5% or more reviews. Furthermore, we also create a matrix of the inverse document frequency of the word across a set of documents to be utilized further in the analysis.

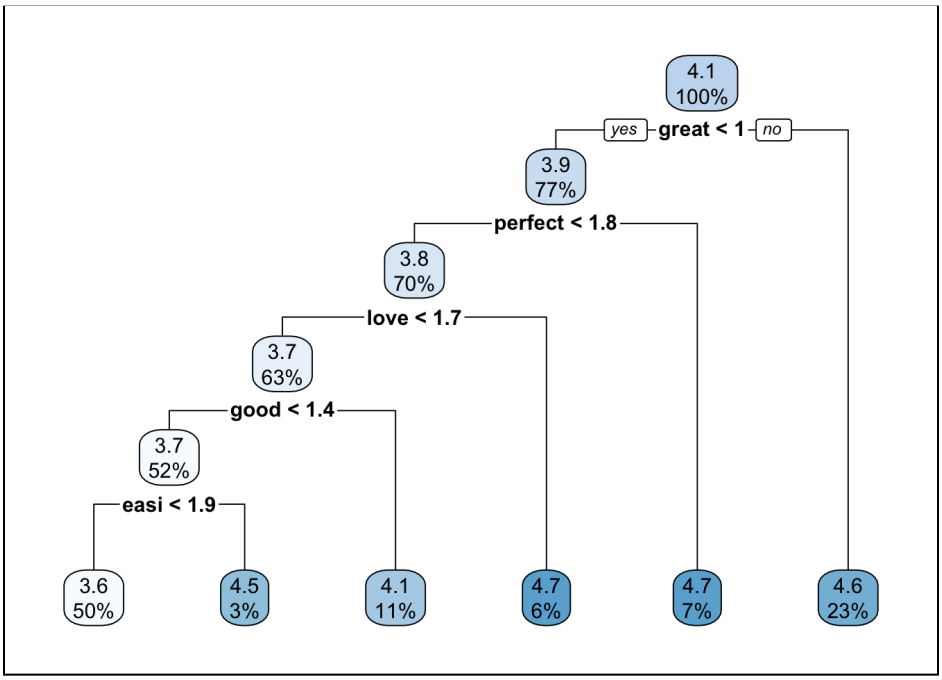
The following horizontal bar chart contrasts the weights of term frequency (TF) and term frequency-inverse document frequency (TF-IDF) for the top 10 terms. The term “sound” occurs approximately 0.1 times in a document and it is less common in the given document corpus as it has a higher TF-IDF weight. Another thing to note is that the term “work” has the highest frequency in a document. However, its occurrence is rare in the document set.

Amazon Electronics Review Project Report



c. *Predictive Models Based on Terms:*

We incorporate the CART model and regression model to predict the review rating. The following CART model result shows that reviews that contain the terms “great”, “perfect”, “love”, “good”, and “easy” are generally rated higher than those that do not contain said terms. Customers tend to rate lower when the reviews contain fewer of these words.



Amazon Electronics Review Project Report

On the other hand, using a regression model as seen in the figure below, the most frequently occurring term, which is the term "work", is predictive of the review score. The model seems to include a good number of positive words, and all the words are statistically significant to predict review score.

	reviewScore		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.92	3.91 – 3.92	**<0.001**
need	0.06	0.06 – 0.06	**<0.001**
work	-0.12	-0.12 – -0.11	**<0.001**
easi	0.10	0.10 – 0.10	**<0.001**
product	-0.04	-0.04 – -0.04	**<0.001**
perfect	0.20	0.20 – 0.20	**<0.001**
great	0.27	0.27 – 0.27	**<0.001**
buy	-0.08	-0.08 – -0.07	**<0.001**
one	-0.04	-0.04 – -0.03	**<0.001**
well	0.09	0.09 – 0.09	**<0.001**
get	-0.07	-0.07 – -0.07	**<0.001**
recommend	0.04	0.03 – 0.04	**<0.001**
time	-0.06	-0.06 – -0.06	**<0.001**
dont	-0.08	-0.08 – -0.08	**<0.001**
love	0.16	0.16 – 0.16	**<0.001**
like	0.02	0.01 – 0.02	**<0.001**
look	-0.02	-0.02 – -0.02	**<0.001**
purchas	-0.03	-0.03 – -0.03	**<0.001**
nice	0.07	0.07 – 0.07	**<0.001**
sound	-0.03	-0.03 – -0.03	**<0.001**
bought	-0.04	-0.04 – -0.04	**<0.001**
use	-0.01	-0.01 – -0.01	**<0.001**
can	0.03	0.03 – 0.03	**<0.001**
just	-0.03	-0.03 – -0.03	**<0.001**
price	0.05	0.05 – 0.05	**<0.001**
will	-0.04	-0.04 – -0.04	**<0.001**
good	0.08	0.08 – 0.08	**<0.001**
realli	-0.01	-0.01 – -0.00	**<0.001**
case	-0.01	-0.01 – -0.01	**<0.001**
littl	0.02	0.02 – 0.02	**<0.001**
fit	-0.05	-0.05 – -0.04	**<0.001**
qualiti	-0.01	-0.01 – -0.01	**<0.001**
Observations	6266882		
R^2 / R^2 adjusted	0.146 / 0.146		

d. *Prediction Error:*

Below is a table comparing RMSE for each model we have run so far. The regression model seems to be performing better than the CART model.

	CART	**Regression**
RMSE	1.329206	1.307354

5. Topic Modeling

We include topic modeling in our analysis to uncover some hidden or latent variables (topics) that shape the meaning of our document and corpus. For topic modeling, the assumption is that each document consists of a mixture of topics and each topic consists of a collection of words. With Latent Dirichlet Allocation (LDA), it focuses on the document-topic and word-topic distributions, lending itself to better generalization. Meanwhile, Latent Semantic Analysis (LSA) decomposes our document term matrix

Amazon Electronics Review Project Report

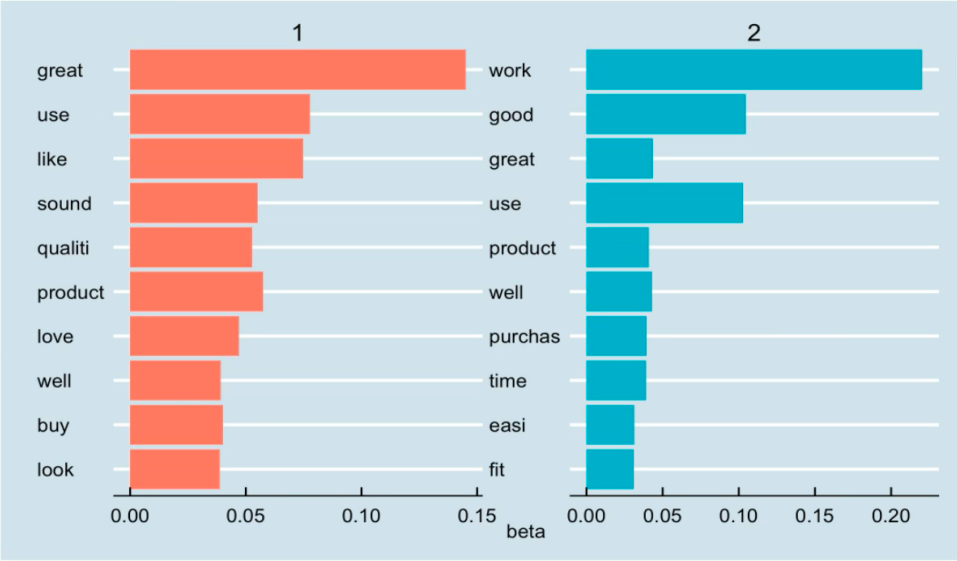
(TF-IDF) into a separate document-topic matrix and a term-topic matrix to find a few latent topics that capture the relationship among the words and documents.

a. *Removing Documents with Zeros:*

We utilize the corpus created from predictive modeling to carry on. When creating the document term matrix, we use term frequency matrix instead of term frequency-inverse document frequency matrix because the LDA function only takes integer values. Topic modeling is a selective model as it can only work with non-zero documents. As a result, 989,255 documents from the document term matrix are removed.

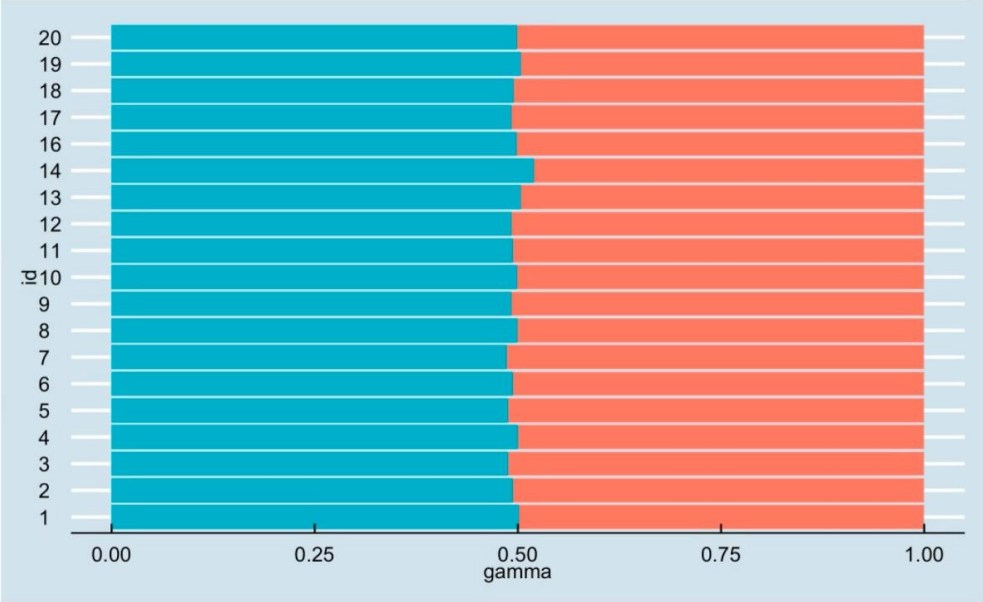
b. *Term-Topic Probabilities/Document-Topic Probabilities:*

We split the document term matrix into two clusters of classification using the LDA function to measure the term topic probabilities within each topic. We construct a table using the beta from the topics and select the top 10 terms to visualize their probabilities.



As the graph above demonstrates, each topic includes multiple words that it is most strongly associated with. However, our objective is to access the model based on the document level. Therefore, we combine the documents with gamma from topics and create the document-topic probability table. Thus, we can combine the topics with the review scores from original data. The following graph is the document probabilities for the first 20 documents.

Amazon Electronics Review Project Report



c. *Topic/LDA Model:*

The finalized data from the previous section is split into 70% in train and 30% in the test. The CART model produces an RMSE score of 1.349.

d. *LSA Model:*

The Latent Semantic Analysis (LSA) model is an attempt to compare the RMSE score from the LDA/Topic model. Unlike the LDA model, which is incompatible with TF-IDF, the LSA model accepts both TF and TF-IDF matrices. Therefore, we select the TF-IDF matrix when applying the LSA function because it accounts for the significance of each word in the document instead of using raw counts. With the same seed and data split, the LSA model generated an RMSE score of 1.35882.

e. *LDA vs. LSA Model:*

Below is a table comparing RMSE for each topic model:

	LDA	**LSA**
RMSE	1.350	1.359

While the LDA model generated a better RMSE score, the nature of the model is incompatible with documents that contain zeros. One-ninth of documents are dropped in the process of the LDA technique.

6. Conclusion and Future Outlook

Overall, there is an alarming number of 5-star reviews that may indicate signs of the abuse of the Verified Purchase system. A lot of purchases might be getting the verified badge without the necessary due diligence from Amazon. This may be the case because refunds from sellers are routing through Paypal, allowing sellers to go unnoticed by Amazon. The declining 5-star reviews after the end of 2016 suggests success in combating fake reviews to some extent by Amazon. More than 80% of reviews did not have a helpfulness vote so, we ignored that measure but in the future if the same is treated as a binary classification measure by casting the reviews that received any votes as "helpful" and the reviews that did not receive votes as "unhelpful" then it could help discover hints in review text that relate to helpfulness.

The correlation between the length of review and review score is negligent, suggesting that there is no link between review length and score. This means the artificial reviews don't have any noticeable patterns or trends in the length of the review. The analyzed data set only takes into consideration the timeframe of 2016-2018 due to capacity and processing constraints in R. Should a longer time frame be analyzed, there is a possibility of uncovering patterns like keywords in the fake reviews. All of this emphasizes that Amazon needs to allocate more resources to deal with this unwieldy problem. For future works, it would be interesting to combine social network modeling techniques and deep learning to understand the drivers for helpfulness ratings and review scores.

As for predicting the review score, the reviews that contain more positive words like "great", "perfect", "love", "good", "easy" are predicted to receive higher ratings. We have also incorporated two of the most popular topic modeling techniques, LDA and LSA. We are able to extract human-interpretable topics from a corpus, where each topic is characterized by the terms they are most strongly associated with, and see how they are distributed in each topic. However, with LSA, we are not able to see what the topics are. Nevertheless, we are able to compare the RMSE for both techniques. Comparing all RMSEs from all the models we have run, regression produces the lowest RMSE.