Amazon Electronics Product Review Using Text Mining

Iris Deng, Mingze Liu, Zhiqing Sha, Zhuyue Yin, Xiran Zhang

Executive summary

Review systems provide valuable business insights to e-commerces like Amazon. Not only customers, but also business analysts use existing ratings to make decisions. Traditional approaches to analyze this data solely rely on numeric ratings, where many confounding factors, like different individual rating styles, misuse of numeric scores or even fake ratings, exist and make numeric ratings less reliable than expected. As business leaders realize the value of text review in providing reliable, thorough and sometimes inherent attitudes, more and more look at these texts along with numeric ratings for more comprehensive business insights.

Our project aims to identify truly flavored products by analyzing text review records. The data we used was collected by Amazon in 2018, focusing on Amazon's electronic product reviews in the year. This dataset contains more than 300,000 records and various attributes of reviews. We excluded null values and extracted several columns from the dataset to perform analysis. Amazon might use the analysis result to improve their recommendation system by prioritizing marketing positively rated products.

Following data cleaning and normalization, topic modeling and sentiment analysis were performed. We utilized Latent Dirichlet allocation (LDA) to identify two topics in topic modeling, and VADER lexicon, which is an unsupervised learning method in judging sentiments for each product identified by ASIN, to rate review sentiment attached to each product. Results from the two-topic LDA model shows that customers mainly look at two aspects when reviewing electronic products: usage of product and performance/functionality of product. The output from the VADER lexicon-based sentiment analysis listed sentiment scores for each product, and several products stand out with extreme high positive scores.

With what we have found, Amazon could enhance customer shopping experience by refining their recommendation system. For instance, Amazon might rely on the ranking of sentiment scores to build its product display mechanism for the electronic department, in which positively rated products can be displayed first. Additionally, Amazon could consider prioritizing certain products that are excellent in usage or performance, which are two important aspects customers look at the most. Our project outputs will help strengthen the accuracy and effectiveness of Amazon's rating and recommendation system overall.

Data description

The data used in our project is an open source dataset, retrieved on http://jmcauley.ucsd.edu/data/amazon/. The data was collected by Amazon in 2018, in json format. The complete dataset includes product reviews in many years across different product departments. For our project purpose, we only look at electronic products reviews in 2018. This specific dataset contains 377,430 rows of reviews in 2018, and 12 columns of review attributes. These variables include overall (numeric) ratings of the product, votes to the review, review text, verified review or not, review time, review time in unix format, review ID, ASIN, reviewer name, review style, summary, and images attached to the review. Among them, we are mainly interested in the "reviewText" column, which includes text responses from customers on the product. Some of the most frequent words used in these reviews are "one", "camera", and "use".

One problem with the dataset is that there are many null values under the "reviewText" column. Since we are interested in inferring sentiment from text, we filtered these non-applicable records out and performed analysis on the rest.

Project objectives

Certainly, ratings are widely used among E-commerce websites as it is an important parameter for buyers to provide their overall positive, neutral, or negative perceptions of the products. Simultaneously, potential buyers rely heavily on the star ratings to make purchasing decisions. However, five-star reviews do not always represent the best products. Mis-clicks on the star-rating system due to human error, fake ratings maliciously posted by the sellers' competitors, or simply the difference in standards among the reviewers are among one of the few instances where we cannot rely fully on the star rating systems in Amazon. Even though Amazon is trying its best to protect customers from false marketing mechanisms, harmful ratings still survive from the purge. We believe that review in text format contains much more information regarding how customers see a product than the rating in numeric value does. With text mining tools, we hope to solve the issue of the ambiguity that star rating systems created, and fully utilize the personalized review in text format to identify the nuance between an excellent 5-star product and an okay 5-star product.

Our end goal is to provide customers with the best experience in finding positively reviewed electronic products within Amazon and allow manufacturers to understand customers' needs through enhancing product quality and providing better customer satisfaction. The project focuses on the "review Text" column to extract human emotions towards the products, instead of the embedded filter features such as star review (1-5) or sales of the products. By employing text mining tools, we can derive valuable and genuine insights from customer attitudes and sentiments in terms of polarity and intensity on products. Ultimately, we aim to update Amazon's recommendation systems by prioritizing positively scored products on the first page of Amazon Electronics category based on sentiments derived from the review text. This will be detailed in the later section related to Sentiment Analysis.

Another question we attempt to answer is "what features or characteristics of the electronic products do customers care the most?" Particularly within the electronics category, we are interested in knowing whether customers value appearance or functionality more. Questions like these can be helpful to the

product manufacturers as well as Amazon in analyzing the contributing factors and logics behind the positive or negative reviews. The Topic Modeling section below discussed the key topics and attributes related to the electronic products in large.

Methodology

Topic Modeling

To do topic modeling, first we set up all the review text into a dataframe since only review text contains sentences with topics. To do the normalization, we also need to wipe out all the floats in our dataframe. Then we create the Bag-of-Words representation of the data. We limited the number of features to 1000 most frequent features to compute the topic model faster and more simply. After that we fit the topic model by LDA. We decided to set the number of topics to 2 and 4 since we were not sure how many topics our dataset has. Because our dataset only contains the reviews of electronic devices, so we assume there won't be too many topics, 2 and 4 would be good choices. To show the final result, we used visualization and drew a Intertopic Distance Map via multidimensional scaling. We found if we set 2 topics, we can conclude the 2 topics we get. However, it's very hard for us to conclude 4 topics if we set the number of topics to 4 since topic 2 and 3 are very similar with each other, and there are a few common words among all 4 topics. As a result, we decided to set the number of topics to 2.

In our topic modeling we mainly used Latent Dirichlet allocation (LDA). LDA is a generative model, but in text mining, it introduces a way to attach topical content to text documents. LDA assumes the dataset contains multiple topics and is good at working with long text such as essays and books. It can also evolve as users process new documents with the same model. Document is separated into multiple distinct topics. Another advantage of the LDA technique is that one does not have to know in advance what the topics will look like, it will determine the topics for users automatically.

However, there are still some shortcomings in LDA. The results we get from LDA are not deterministic, which means users might get different results each time for the same dataset. LDA also suffers from "order effects" when it was adopted on different topics. For example, different topics are generated if the order of training data is shuffled. Such order effects introduce a systematic error for any study. This error can relate to misleading results; specifically, inaccurate topic descriptions and a reduction in the efficacy of text mining classification results.

Sentiment Analysis

The focus of our analysis of Amazon reviews is the sentiment analysis with unsupervised machine learning. The CSV file in which the data is contained divides the data into several columns, and similar to topic modeling, the sentiment analysis uses the ASIN and review text columns.

The review column contains unstructured text, and we use performed normalization on these review texts for more convenient analysis.

The objective is to update Amazon's priority of listing products based on the sentiment that customers have on each electronic product. Hence, the first step we should consider is to put all the reviews for one product together and link the combined review to its respective ASIN. In this way, we could potentially come up with the polarity and intensity of the opinion customers had towards one product by analyzing all its reviews.

The best way to achieve this is to incorporate a dictionary with keys and values as previously mentioned, the correctly established link between reviews and the product to which they belong is extremely important. We used a simple for loop to loop through the dataset and assign ASINs to keys while combining all the reviews for the same ASIN into one list of strings and assigning it to the key.

The second thing to consider is to define a function that could produce a comparable output among product reviews. We chose to use the Vader Lexicon-Based sentiment analysis for its advantage in simple comparability. The Vader Lexicon-Based function should have a binary polarity, positive or negative, and an intensity indication, the Vader score, as its output after feeding it normalized unstructured text as strings.

After defining the Vader Lexicon-Based sentiment analysis function, we need to loop through our predefined dictionary of ASINs and reviews to get polarity and Vader score for all products. The results are contained in a separate PDF file of over 10,000 pages for viewing.

We believe that there are several advantages to our method of computing sentiment analysis on product reviews, and one of them is its comparability. The Vader Lexicon-Based sentiment analysis could output a numerical number that indicates intensity which lots of other methods, such as the SGDClassifier of a supervised machine learning method lack. Another advantage of Vader is that it does not require any training data, making it efficient while processing large datasets like ours. Finally, it has a great 'understanding' of text emotions as well as urban slang, making it powerful to analyze social media comments and product reviews, which generally includes a large number of everyday informal text.

While Vader Lexicon-Based sentiment analysis has attractive advantages, it also has drawbacks. One of which is the problem of irony could be entirely misunderstood by Vader and provide an opposite polarity and intensity. However, we believe that irony does not often occur in Amazon product reviews and hence this drawback should not impact our results severely. Another drawback is that Vader does not understand non-English comments or reviews. We also don't believe that's a major issue for our analysis given that this dataset is based on Amazon.com instead of other regions such as Amazon.cn. Finally, outside of Vader analysis, the Pandas package very slow and inefficient in processing large dataset like ours, and one alternative that could be chosen is MODIN, which has a way faster processing speed and a large Pandas API coverage; one could simply consider MODIN a more enhanced version of Pandas for processing large datasets.

Results and Discussion

Topic modeling

As we can see in our Intertopic Distance Map, topic 1 mainly contains the name of devices and some verbs that appear when using them, so we can conclude them into names of devices and instruction verbs. In topic 2, most of the words are adjectives often used to evaluate items, so we conclude them into adjectives. Topic 1 shows that when evaluating an electronic device, people tend to describe their experience in using rather than other factors like appearance. Electronic devices customers of Amazon focus more on machines' performance. The high frequency of 'issue' and 'problem' indicates most of the customers mentioned their experience to describe their dissatisfaction. In topic 2, most of the high frequency adjectives are positive adjectives, others are all neutral adjectives, which means most of the review text had a positive evaluation of products. 'Great' and 'good' are the 2 words with the highest frequency.

Sentiment Analysis

As previously mentioned, the result is included in a separate PDF file for viewing (see Appendices for a screenshot of it).

The result from the Vader Lexicon-based sentiment analysis is clear and visually easy to understand. For each ASIN, there is a binary polarity, positive and negative, and a sentiment score between -1 and 1. There are a number of products that have a positive polarity with a 1.0 positive intensity, such as product B00009R8XD, B00009XVCZ and B0000BVYT3. Although the Vader scores are printed in 2 decimal places, we keep 4 decimal places for comparing purposes (as displayed in the brackets under the Vader score), drastically decreasing the number of the same score for intensity.

We would like to put the highest positively scored product as the very first product customers see in the electronics department page of Amazon, and as there are several products with positive 1.0 even with 4 decimal places, we would put the one with more sales in front the ones with less sales. The same logic applies to the rest of the products with the same polarity and intensity.

It seems that the most negatively intense product is unique. It is product B0001654K4 with a negative intensity score of -0.97. This product will be placed at the bottom of the very last page of the electronics department.

Generally, we believe the results of our sentiment analysis prove a good guide for Amazon on how to sort and prioritize their product list for the electronics department.

Conclusion

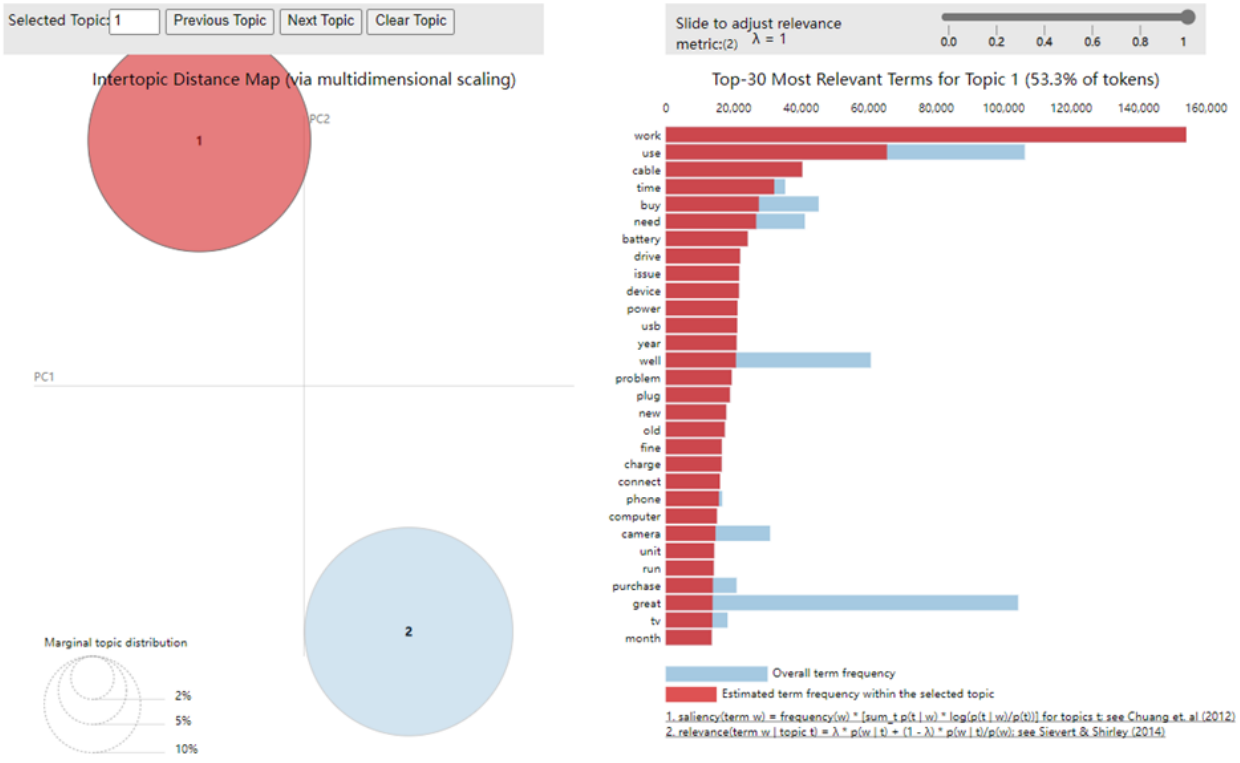
In comparison to solely relying on Amazon's existing star ratings, our project greatly controls for the factors, such as human errors and fake star ratings, that may affect the reliability of Amazon's recommendation systems. It enhances the trustworthiness of the review system by prioritizing the best products to the potential buyers based on reviews. Our project concludes that customers value more about the functionality and usability rather than the appearance of the electronics products based on Amazon review text data in 2018. Manufacturers of the products may leverage the keywords mentioned most frequently to update its product design, features, and potentially predicting the future trends. Although we cannot identify clear topics among the electronics category, it would be interesting to conduct further studies for the betterment of this system. One of which is to perform topic modeling based on brands, or on more specific categories of electronics such as phones, televisions, etc. Our dataset contains limited information about the product name as it only provides product id as an identification. We would have drawn more valuable insights from the results if specific product name and sub-category are provided.

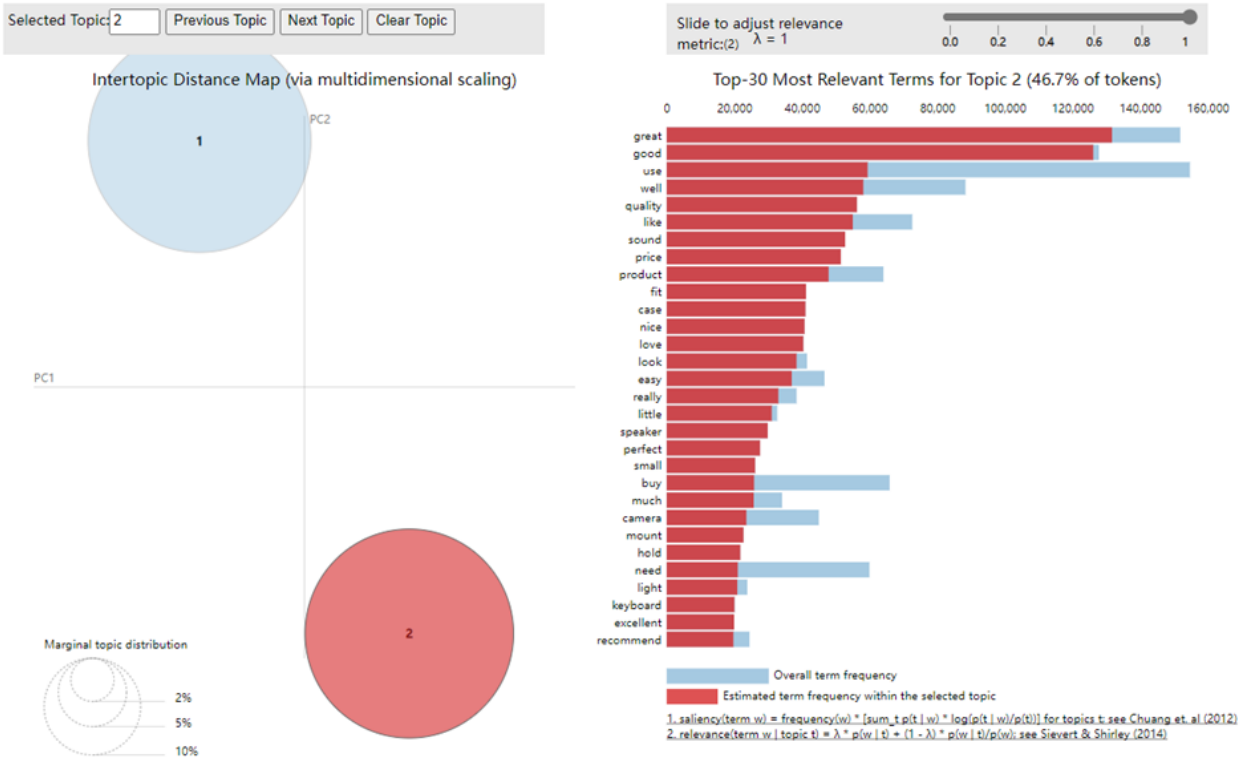
Using topic modeling and sentiment analysis, we found that positive sentiments are more prevalent than negative sentiments, though there is one caveat in drawing this conclusion due to selection bias - not all buyers leave reviews. We assume that often those with strong opinions towards the products, either good or bad, leave text reviews.

In conclusion, Amazon, manufacturers, as well as potential buyers can all benefit from this large load of data by performing text mining on Amazon review data. It adds another assurance to the existing star rating system by analyzing customer sentiments and extracting key attributes of the products.

Appendices


```
Product:

B00V7CAKKK
VADER Polarity (Binary): positive
VADER Score: 0.62
('positive', 0.6249)

Product:

B00V7B5B7I
VADER Polarity (Binary): positive
VADER Score: 0.62
('positive', 0.6249)

Product:

B00V7CBH6G
VADER Polarity (Binary): positive
VADER Score: 0.86
('positive', 0.8555)

Product:

B00V7UL6OG
VADER Polarity (Binary): positive
VADER Score: 0.68
('positive', 0.6808)

Product:

B00V7V02YK
VADER Polarity (Binary): positive
VADER Score: 0.98
('positive', 0.9756)

Product:
```


Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
Marginal topic distribution
2%
5%
10%
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Top-30 Most Relevant Terms for Topic 1 (53.3% of tokens)
0 20,000 40,000 60,000 80,000 100,000 120,000 140,000 160,000
work
use
cable
time
buy
need
battery
drive
issue
device
power
usb
year
well
problem
plug
new
old
fine
charge
connect
phone
computer
camera
unit
run
purchase
great
tv
month
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
Marginal topic distribution
2%
5%
10%
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Top-30 Most Relevant Terms for Topic 2 (46.7% of tokens)
0 20,000 40,000 60,000 80,000 100,000 120,000 140,000 160,000
great
good
use
well
quality
like
sound
price
product
fit
case
nice
love
look
easy
really
little
speaker
perfect
small
buy
much
camera
mount
hold
need
light
keyboard
excellent
recommend
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

References

DeLancey, J. (2020, May 29). *Pros and cons of NLTK sentiment analysis with vader*. CodeProject. Retrieved December 17, 2021, from https://www.codeproject.com/Articles/5269447/Pros-and-Cons-of-NLTK-Sentiment-Analysis-with-VADE

Evaluating KNN, LDA and QDA classification for embedded online feature fusion - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/OVERVIEW-OF-PROS-AND-CONS-OF-KNN-LDA-AND-QDA_tbl1_224375624

Source of text data: http://jmcauley.ucsd.edu/data/amazon/