

# Discovering Product Differences From Consumer Reviews

Yimin Li<sup>1</sup>, Li Liu<sup>2</sup>, and Yuming Zhu<sup>3</sup>

<sup>1</sup>M.A. Student in Computational Social Science, University of Chicago

<sup>2</sup>M.A. Student in Computational Social Science, University of Chicago

<sup>3</sup>M.A. Student in Social Science, University of Chicago

6/5/2020


*Class Project for SOCI 40133 Computational Content Analysis in Spring 2020* <sup>1</sup>

---

<sup>1</sup>We would like to thank Professor James Evans, Bhargav Srinivasa Desikan, Hyunku Kwon, and classmates for helpful comments and suggestions. Replication files are available at [Github](#).

# 1 Introduction

We usually rely on or refers to customers’ reviews when shopping online. Although shopping platforms always recommend us some products by all means, we tend to keep minds clear and find the one we like and we believe in. However, even if we find the best seller with over thousands reviews and over four-star in score, and carefully read some long reviews that seem to include thorough information, we still bypass too much of the power of reviews. Hence, this project aims to provide a prototype of an information revealing system of customers’ reviews.



	Natural Ergonomic Keyboard 4000	Wireless Comfort Desktop 5050	Wireless Desktop 3050	Wireless Desktop 900	Sculpt Ergonomic Desktop
Overview	Ergonomic design, custom palm rest	Ergonomic design, built-in palm rest	Stylish design, built-in palm rest	Modern desktop at a great value	Ergonomic design, cushioned palm rest
Type of design	Natural	Curved	Straight	Straight	Split
Ergonomist Approved	✓	✓			✓
Windows Hotkeys				✓	✓
My Favorite Keys		✓	✓		
Warranty (yrs)	3	3	3	3	3

Figure 1: Product comparison on Amazon

Think about a case when we are going to buy a mechanic keyboard, which is our example in this paper. From Amazon or other shopping platform, we may get as many product detail as we want. Figure 1 is how Amazon compares its products. These information are all producer-based, which means they are provided by producers. However, since most of us are not keyboard experts, we are not able to make decisions immediately (or maybe experts never compare). What we are going to do is to show the same infographic comparison with reviews information. Currently, it is very hard for customers to take in information underlied beneath thousands of reviews. As the reviews are as important a reference for customers to make purchasing decisions, we deem our project valuable.

We are going to briefly introduce the algorithm in section 2. We will going to show

you what will be the input and output, and how the algorithm generally works. In section 3, we will explicitly discuss how our data is going to be in general and the specific sample we use in this paper. Section 4 serves as the main body of this paper, which shows in detail how we apply methods in content analysis to reveal the underlying messages from reviews as a whole. We will demonstrate the main results from our sample in section 5 and discuss our plan for the future work in section 6. Section 7 will conclude the entire paper.

## **2 Algorithm**

Figure 2 shows the design of our algorithm. Firstly, the users could provide several similar products that they are considering buying. Then the algorithm could identify the unique ASIN codes that Amazon uses for these products. Secondly, the algorithm would fetch and sample the associated consumer reviews by the ASIN codes and then pre-processed them. Thirdly, the algorithm would do three parts of analysis in parallel, which are exploratory data analysis, topic modeling, and word embedding. Lastly, the results will be combined to generate a product comparison infographic that quantify the differences of product within reviews.

## **3 Data**

### **3.1 Population**

There are two major parts in the Amazon product review data: consumer reviews and product metadata. The data is scraped by several computer science researchers at the University of California, San Diego and made available for the public to use (Ni, Li, & McAuley, 2019).

The short way to describe the data is it captures all information that consumers would normally see on the product page, such as price, images, descriptions, reviews, and so on.

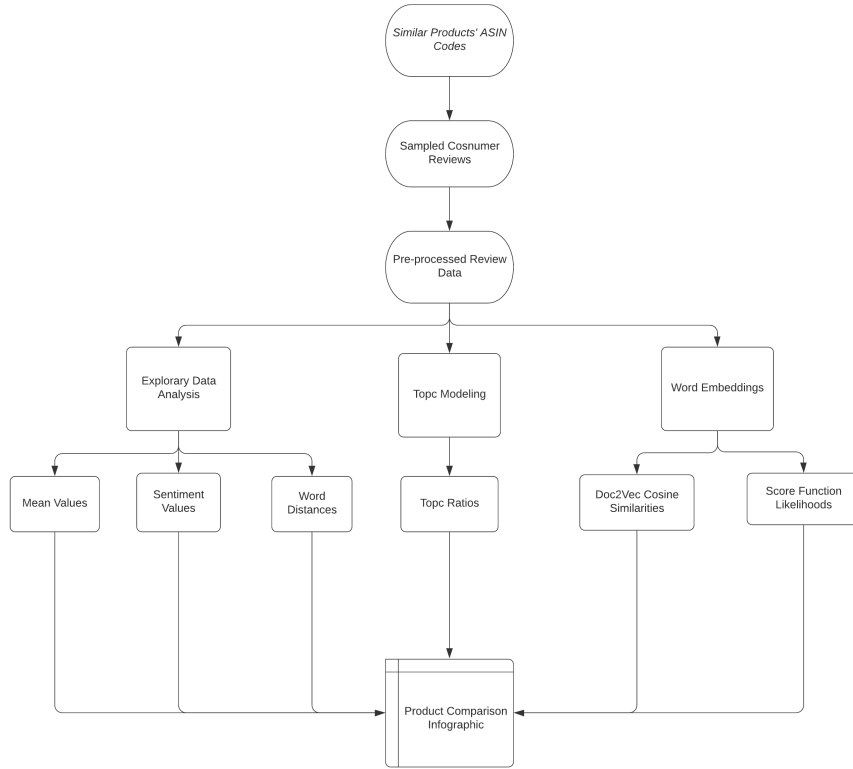


Figure 2: Algorithm Illustration

For each product in the data, the data provides detailed information on its characteristics, such as color, price, package type, descriptions, technical details, similar products, image features, categories information, sales rank, to name a few. For each review associated with the product, it has the reviewer's name, ratings, text, summary, attached image URL, and helpfulness votes, the label for verified purchase, and so on.

Most of the existing studies with this data are in the fields of recommendation systems and natural language processing. The provider of the data Jianmo Ni and Julian McAuley have three published paper in top Computer Science conferences with this data (Ni & McAuley, 2018; Ni, Lipton, Vikram, & McAuley, 2017; Ni et al., 2019). As a result, this project also aims to provide better recommendations for consumers by applying techniques from computational content analysis.

## 3.2 Sample

For this project, we are interested in the scenario when consumers are choosing among similar products for final purchase. As one of the authors is thinking about buying a new keyboard recently, we decide to pull out the relevant data from the keyboards that he is currently considering. Figure 3 shows the screenshots of the 5 keyboards sold on Amazon. In the following sections, we will refer to the respective keyboard by its brand name, such as the Redragon/HP/Azio/Jelly Comb/Microsoft keyboard.

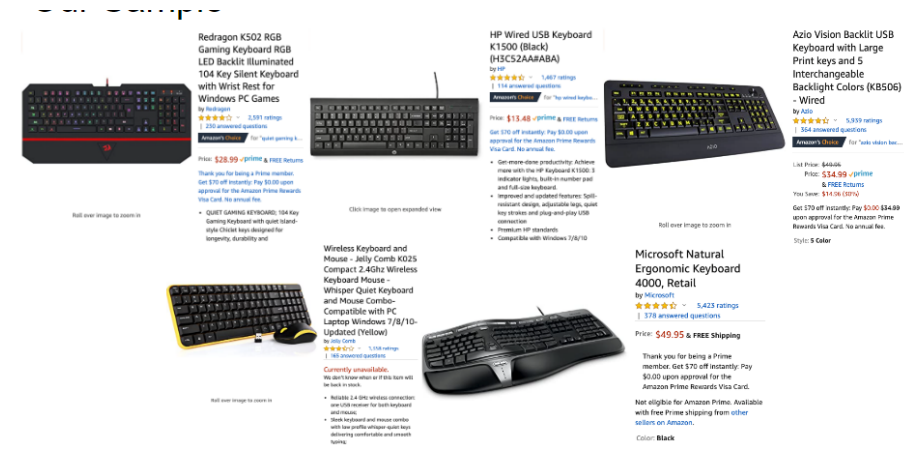


Figure 3: 5 Similar keyboards to Choose From on Amazon

## 4 Methods

### 4.1 Exploratory Data Analysis

Consumers would usually have a overview of the products that they are considering. So in this section, we conduct exploratory data analysis to extract the straightforward differences of the products and reviews.

#### 4.1.1 Average score, votes, verified comments, prices among different products

As mentioned in the dataset section, we first calculated the mean of overall scores, votes and verified comments among five different Amazon keyboard products and plot them

into a bar chart as Figure 4 has shown. Generally speaking, they all have fairly positive reviews, ranging from 3.8 to 4.3; Also, the average votes (“likes”) ranges from 0.4 to 1.1, showing different popularity among different products; Considering the Amazon Review huge dataset, the result seems consistent and can function as an excellent sample of the whole big Amazon Review Dataset.

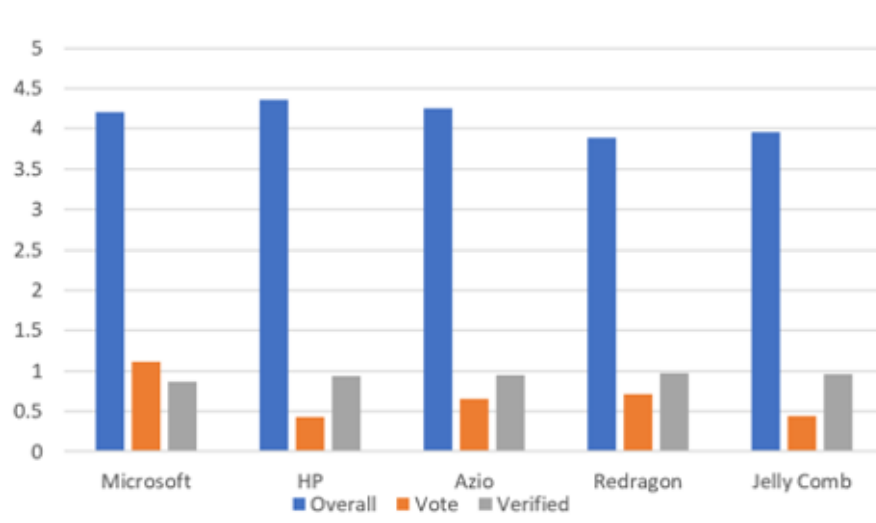


Figure 4: Mean of Overall score, Votes and Verified Purchase Labels Among Different Products

Apart from the high-level discussion, we would also notice some interesting patterns inside our dataset. For example, HP keyboard has the highest average score but with the lowest average popular votes; This basic-level exploratory analysis tells us some basic difference in each review corpus as revealed from the review dataset, paving the way for our future analysis.

#### 4.1.2 Sentimental Analysis

Apart from basic exploratory analysis, we further conducted two-level sentimental analysis to help us draw a more in-depth picture of the sentimental “feelings” of the five keyboard products. We used Python’s TextBlob package to conduct sentimental analysis and the final sentimental score would be a consecutive variable between -1 (negative) and 1(positive). To further test our consistency, we conducted two-stage sentimental analysis

for cross validation. We first calculated the sentimental score of certain products with all its review corpus as the input; Then, we calculated the sentimental score for each review for each product and calculated the mean value of all reviews for certain products. The final sentimental analysis results were reported in Figure 2.

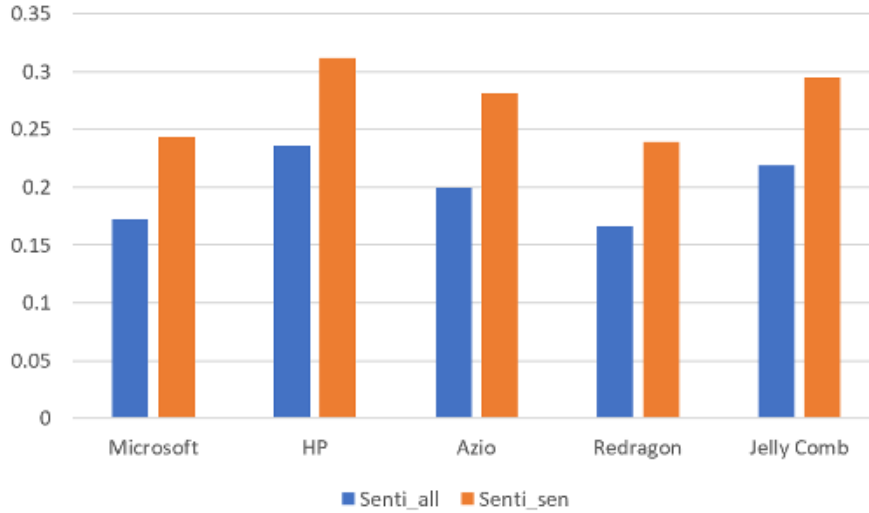


Figure 5: Sentimental Score Among Different Products

The result is super consistent with our prior findings. All five products have a relatively good sentiment score (all above 0.15), meaning the positive character of the Review Corpus. In detail, HP has the highest sentimental score in both methods while Redragon has the lowest, excellently consistent with our prior finding. We would also take the sentimental score into our final model building when giving consumers' recommendation.

#### 4.1.3 Word Clouds

We also applied the word clouds to count the frequency of words appeared in the Review Corpus and visualize it by using the word clouds method. As shown in the word cloud, it quite makes sense that keyboards and keys appear much frequent inside the corpus since we are analyzing the keyboard review data. What's more, we also discovered some other interesting patterns from the word clouds graph. First, it shows depicts a positive sentimental incline which is consistent with our prior findings. For example, the frequency of "good" words (i.e. "great", "good") is much greater than the frequency of negative





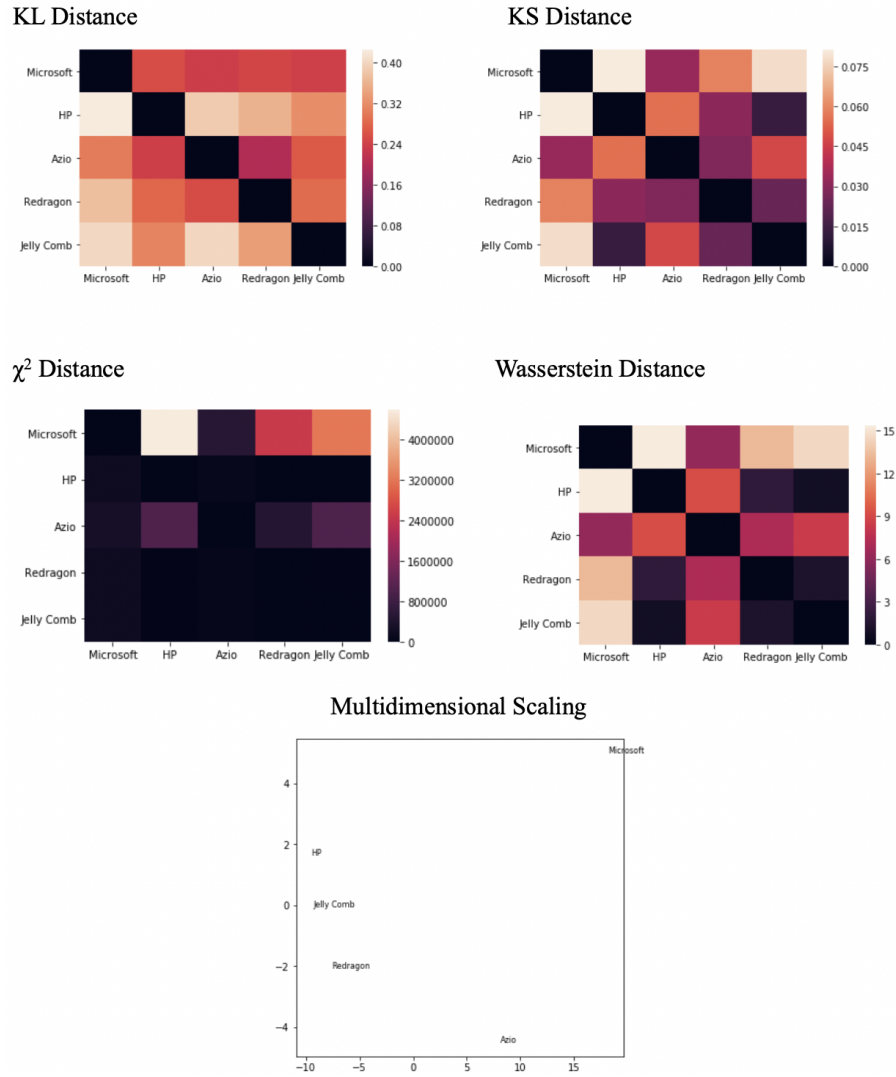


Figure 7: Distance and Multidimensional Scaling

The excellent classification is further examined in the topic modeling section, which we would also examine which products might share similar topics, helping us to give potential decisions to the customers. Also, this distance calculation has excellent implications in real-world data-based decision making, making it possible to classify several products into several category and can recommend customers with the products with the products that is closed to it in terms of distance, providing abundant similar products to the consumers.

## 4.2 How do reviews say differently?

### 4.2.1 Topic Modeling

Different reviews talk about different topics. As a popular tool for extracting latent topics, the topic model is a type of unsupervised machine learning method that groups the words in the reviews into several latent topic groups (Alghamdi & Alfalqi, 2015). To find these latent topics, the topic model calculates the statistical correlations and groups the words that used together more frequently (Gentzkow, Kelly, & Taddy, 2019). In particular, as the reviews for different products often contains several topics, the model will assign different topic ratios to each product.

With the help of genism python package, we will examine what properties each product has to the perspectives of the customers as well as the relationship and correlation among these products.

### 4.2.2 Latent Topics

We use the topic coherence score to determine the optimal topic number. As shown in Figure 8, the coherenc score reaches a local optimum with 5 topics. Hence, we set the topic number to be 5 and used the genism package to conduct topic extraction. The topic extracted from the genism package is shown in Figure 9.

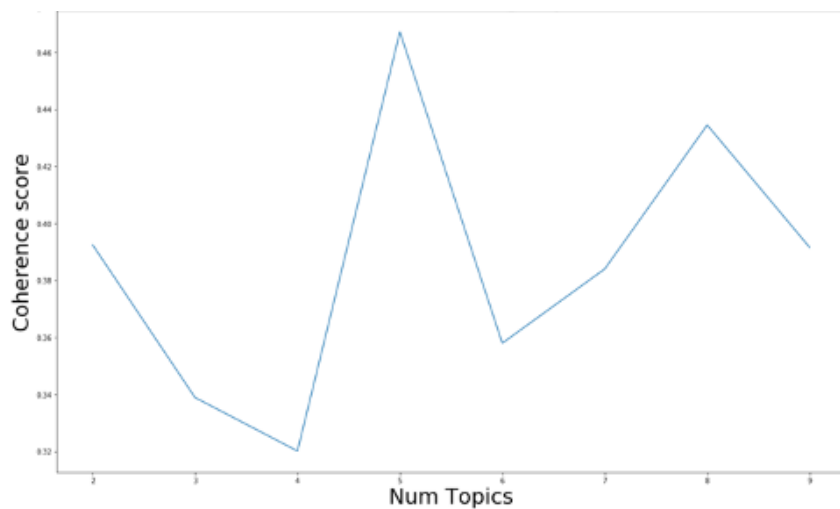


Figure 8: Determining Optimal Topic Number – Topic Coherence Score

Figure 9 reports topic covered in each model, ranking from the highest weight to the lowest weight in each model. Actually, since all five products come from keyboard category in Amazon, it is not surprising at all to see the huge homogeneity among each other: All models listed keyboards/keys as the most important words followed by use and great, remaining consistent with our findings in the EDA section. Also, Model 0 performs slightly different from others, which include the word “typing/types” for two times, highlighting its functional use.

So how could these models help consumers data-based decision making? In the later visualization section, we would see that even though the difference is so tiny, we could also observe the difference performance among different products in the visualization section, hence this method would be an excellent method to be explicated into other studies to further help consumers to extract the topic from products that are so inter-correlated with each other.

	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4
0	keyboard	keyboard	keyboard	keyboard	keyboard
1	keys	keys	keys	keys	keys
2	use	use	key	great	like
3	like	key	great	key	great
4	key	like	use	good	key
5	keyboards	great	like	use	use
6	typing	good	work	keyboards	good
7	type	work	keyboards	like	work
8	work	love	feel	love	m
9	love	space	microsoft	typing	keyboards

Figure 9: Topics Extracted from the Genism Package

### 4.2.3 Results and Visualization

After determining the optimal number of topics and topic listing above, we further conducted topic modeling to further see the intersection and topic modeling results for different products on different topics. The plot and heatmap are shown in Figure 10.

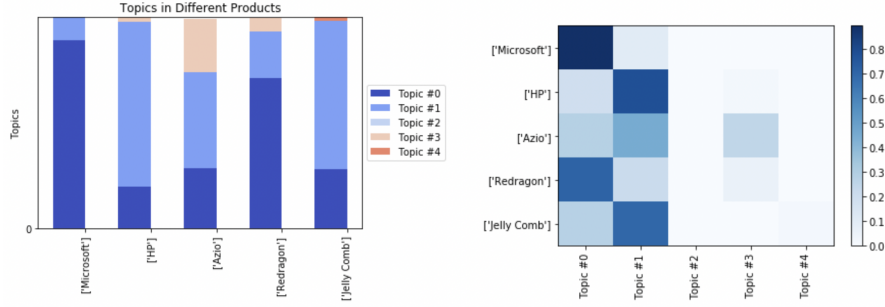


Figure 10: Topics Extracted from the Genism Package

As is shown in the two graphs above, there is a huge divergence for five products on five topics. It shows that almost no products come from Topic 2 while all of them have a huge portion covered in Topic 0 and 1, with some products referred to Topic 3. This result is quite surprising and insightful that they can extract the topics so well even though the difference between topics is not so prevalent, providing excellent implications for our future work in topic extraction.

We also see the results of the topic extraction matched with the EDA results, the two outliers in the distance classification has the highest topic incline in Topic 0 and 1 while the other three seems not so strong preference in topic modelling. This not only verifies the exploratory results found in the previous section, but also shows us excellent real-world implications and functions that it can achieve.

### 4.2.4 Conclusion and Discussion

To conclude, topic modeling provides an excellent perspective for us to conduct topic extraction and depict the different topic meaning behind it. Although the difference for the topic generated from unsupervised machine learning is not so prevalent since our

	Topic 1 (Functionality)	Topic 2 (Brand)
Product A	0.2	0.6
Product B	0.5	0.3

Table 1: Final Output of the Topic Extraction (Illustration)

dataset consists of reviews from highly seminar products, we could also see an excellent topic extraction result, having two strong preference towards a certain topic.

The result is consistent with the distance and multidimensional graph in the EDA section, verifying the excellent prediction of both methods, providing an excellent approach for assisting data-driven decision for customers. Our estimated final output is shown in Table 1, which we might quantitatively tell the difference between product A and B in terms of different topics and it can be achieved currently via both distance calculation and topic extraction, although the difference of topic is not so prevalent.

We intend to extrapolate this model built here to more products in the Amazon Review Data. Since the results have been prevalent even with such tiny differences, we believe that we could achieve great success in extrapolating it to different products with less homogeneity where we could perform excellent topic extraction there.

### 4.3 How do reviews mean differently?

One limitation in topic model is we only consider what words appear but ignore their relative locations in the sequences. Thus, two paragraphs generated by a bag of same words could have same latent topics but different semantic meanings.

This difference is important from consumers’ perspectives. When reading reviews, consumers not only look at what reviews say but also what reviews mean. As a result, we use Doc2vec model (Le & Mikolov, 2014) to examine how different products’ reviews relate to each other within the semantic space.

The heatmap 11 shows how the product reviews are similar with each other by calculating the pairwise cosine similarity. For example, for the Microsoft keyboard, its reviews are most similar to the reviews of Jelly Comb’s keyboard (0.98) and least similar to the

reviews of Azio’s keyboard (0.83).

Consumers can use this results to help them evaluate how products are different by the semantic meanings of reviews. As a result, if a consumer is considering buying the Microsoft keyboard and would like to consider one more option from the 4 other keyboards, the Jelly Comb keyboard should be more likely to be recommended by the algorithm.

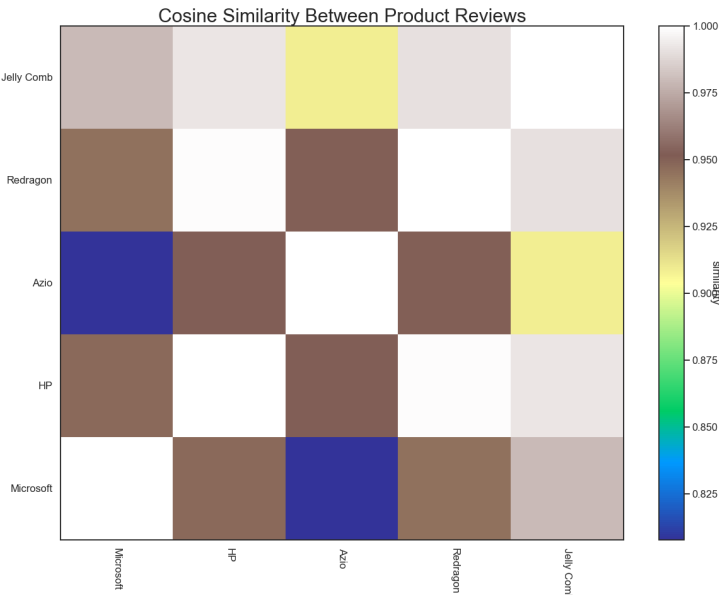


Figure 11: Product Description Provided by the Keyboard Seller

**4.4 Are product descriptions trustworthy?**

When consumers compare similar products online, they will likely read the product descriptions provided by the sellers. These short descriptions help highlight the unique features and advantages of the products. However, consumers will not just be convinced by what the seller says. Instead, they will go reading the reviews to verify these points. If the reviews say the same thing, consumers are more likely to trust the seller and purchase the product. On the other hand, if the reviews don’t say the similar thing (or even opposite side), consumers would probably be suspicious of the seller and less likely to buy the product.

To evaluate how trustworthy is the sellers' descriptions, we measure the likelihoods that the descriptions would be generated by the reviewers' mouths. If the likelihood is high, the reviews align with the descriptions. If the likelihood is low, the descriptions might be somewhat different from the reviewers' opinions.

We use the score function based on word-embedding model. Developed by Matt Taddy, the score function calculates the likelihood that a provided text could be generated by a word-embedding model trained by the reviews (Taddy, 2015). It works by summing the inner product between each pair of the text's word vectors.



Figure 12: Product Description Provided by the Keyboard Seller

After applying the score function to each of the five products, we find the likelihoods are indeed different, indicating they could have different level of trustworthiness. In order to make the results more readable, we calculate the percentage difference of trustworthiness from the baseline result (Microsoft Natural Ergonomic Keyboard 4000). In 13, we find 3 other keyboards' descriptions (keyboards by Redragon, Jelly Comb, and Azio) are less trustworthy, while the HP keyboard's description has a higher likelihood (3% increase) and slightly more trustworthy than the one from Microsoft keyboard.

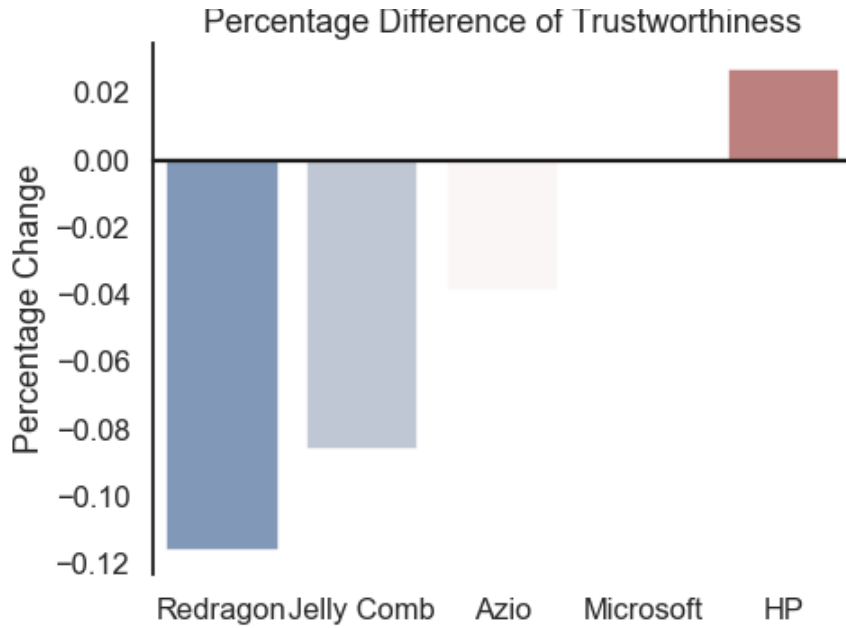


Figure 13: Percentage Difference of Trustworthiness

## 5 Results

Combining the result from the methods (Counting Words, Sentiment Analysis, Topic Modeling, Doc2Vec, and Score Function), Figure 14 provides a infographic that could help consumers make easier decisions by quantifying the product differences from consumer reviews. For example, if the consumer is interested in picking up a keyboard that is both popular and with a trustworthy seller, the Azio keyboard would be the best option as it satisfies these criteria.

Also, this infographic is the prototype of an algorithmic product that we are building. To solve consumers' decision problem, we plan to build a web app where users could type in the names of products that they are comparing. The web app could generate a infographic automatically that helps consumers accelerate decisions without spending too much time in reading reviews.




 Comparing five keyboards					
Features	Microsoft Natural Ergonomic Keyboard 4000	Azio Vision Backlit USB Keyboard	HP Wired USB Keyboard K1500	Redragon K502 Gaming Keyboard	Jelly Comb K025 Wireless Keyboard and Mouse
Rating	4.2	4.4	4.2	3.9	4.0
Popularity	✓	✓			
Sentiment	✓	✓			
Topic - Functionality	✓				✓
Topic - Brand		✓		✓	
Semantic Distance	/	95%	83%	94%	98%
Trustworthiness of Seller	/	3%	-4%	-11%	-8%

Figure 14: Product Comparison Infographic

## 6 Future Work

This project is motivated to solve our real problem in online shopping experience. We also realize our current result is just a start of comparing product differences from consumer reviews. The followings are some limitations in the current work and possible ideas for future work.

### 6.1 Inputs from Consumers

Imagine how you made the final decision from similar products in your last Amazon purchase. The decision is a function of not only product features but also consumers' intrinsic preferences. For example, some consumers prefer the ones with highest ratings while some others prefer the ones with most unique feature.

Our current infographic is static, meaning it cannot interact with consumers directly. As a result, in our future web app, users could provide weights for their intrinsic prefer-

ences and the app will recommend the best product based on different weights.

## **6.2 Speed-Accuracy Trade off**

In this project, we sampled 300 reviews from each of the product reviews subset. There are two reasons for doing this. Firstly, it enables us to train the models faster. Secondly, online consumers also read reviews randomly instead of reading all of them.

However, this might also decrease the accuracy of the model results. From our experience, reviews data are highly skewed as only a few of them are very well-written and receive many "Helpful" votes. They are usually listed on top of the page so that consumers could see them. While for sampling, there is no guarantee to be included.

To solve this issue, we think we should probably use weighted sampling instead of random sampling. The "Helpful" votes, which is highly correlated with the likelihood of being read, could be used as the weights for sampling the reviews.

## **6.3 Sensitivity of Results**

Since we choose the products that are similar in nature, it is likely that there are no significant differences in the results. As a result, we still need to do more work to understand whether a difference, such as in likelihoods from score function, is detected purely by chance or caused by the different underlying trustworthiness levels.

We also notice that the product descriptions vary in length, implying we probably should not compare the likelihoods from score functions directly.

## **6.4 Review Amount**

Products, especially keyboards, evolve rapidly. It is likely that a consumer would consider a relative new product with few or no reviews. In this case, we cannot compare that new product with the older product from reviews.

## 7 Conclusion

Online shopping enables consumers to have more choices. With the abundant information presented to them, consumers need to spend a lot of time on comparing the similar products. Our project tries to solve this decision problem by designing an algorithmic infographic that could compare similar products by their differences in consumer reviews. In particular, we quantify these differences by exploratory data analysis, topic modeling, and word embedding. We are also excited to bring this project to the next level, such as building a web app that consumers could get personalized recommendations based on their inputs and preferences.

## References

- Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, 6(1). Retrieved 2020-05-01, from <http://thesai.org/Publications/ViewPaper?Volume=6&Issue=1&Code=ijacsa&SerialNo=21> doi: 10.14569/IJACSA.2015.060121
- Gentzkow, M., Kelly, B. T., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, forthcoming.
- Le, Q. V., & Mikolov, T. (2014, May). Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*. Retrieved 2020-06-05, from <http://arxiv.org/abs/1405.4053> (arXiv: 1405.4053)
- Ni, J., Li, J., & McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 188–197).
- Ni, J., Lipton, Z. C., Vikram, S., & McAuley, J. (2017, November). Estimating Reactions and Recommending Products with Generative Models of Reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 783–791). Taipei, Taiwan: Asian Federation of Natural Language Processing. Retrieved 2019-10-30, from <https://www.aclweb.org/anthology/I17-1079>
- Ni, J., & McAuley, J. (2018, July). Personalized Review Generation By Expanding Phrases and Attending on Aspect-Aware Representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 706–711). Melbourne, Australia: Association for Computational Linguistics. Retrieved 2019-10-30, from <https://www.aclweb.org/anthology/P18-2112> doi: 10.18653/v1/P18-2112
- Taddy, M. (2015, July). Document Classification by Inversion of Distributed Language Representations. *arXiv:1504.07295 [cs, stat]*. Retrieved 2020-06-05, from

<http://arxiv.org/abs/1504.07295> (arXiv: 1504.07295)