Final Report

Yuhan Xu 474154, Wenda Yin 467295, Christy Ren 474884, Rita Shi 474882

1) Data description

The dataset, Amazon Review Data, contains product reviews and metadata from Amazon, including 233.1 million reviews spanning May 1996 - July 2018. Our analysis based on the product reviews data of "All_Beauty" category, which contains 371,345 reviews.

The dataset has 12 columns and 371,345 rows. Here are the columns:

- reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewerName - name of the reviewer
- vote - helpful votes of the review
- style - a disctionary of the product metadata, e.g., "Format" is "Hardcover"
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)
- image - images that users post after they have received the product
- verified –True or False

2) Project objectives

We want to use the data to understand customer opinions and comments about different products to help Amazon better manage its products in "All Beauty" category.

Here are the insights we plan to get from our analysis:

1. What is the sentiment distribution of the "All Beauty" category reviews?
2. What are the most frequent topics when it comes to complaints?
3. What are the most popular topics when it comes to positive feedback?
4. What are the differences in the sentiment between those who frequently leave reviews and those who seldom leave reviews?
5. Does the proportion of positive reviews grow through time or not?
6. What other factors lead to a higher probability of a positive comment?

3) Methodology

Method I：

Sentimental Analysis

Because in the original dataset we do not have labels for these review text, we can only use unsupervised learning methods. That is, the Lexicon-based sentiment analysis method. Lexicon is a collection of words, n-grams with known polarity or emotion associations. Use lexicon-based sentiment analysis, we can analyze words and word patterns known to be associated with different types of sentiment.

To analyze the sentiment of customer reviews, we used VADER Lexicon-Based Sentiment Analysis.

The pros of this methodology:

VADER Lexicon-Based Sentiment Analysis suits unlabeled text. The customer reviews in our dataset don't have predetermined polarities. Therefore, this unsupervised method is suitable for analyzing the sentiment of customer reviews.

Also, this methodology is widely used in analyzing sentiment on social media text since it has been specifically attuned to analyze sentiments expressed in social media. It would be quite suitable for customer reviews post on Amazon.

The cons of this methodology:

First, it is hard to choose the threshold to determine the polarity of the reviews since we do not have labels and do not know the original sentiment.

Moreover, it would be hard for us to measure the accuracy of our classification since we do not know the true sentiment of these reviews.

Method II：

Latent Dirichlet Allocation

LDA plays a very important role in the topic modeling and is often used for text classification. In our analysis, we conduct LDA on the review text to analyze the distribution of words and topics in the documents.

The pros of this methodology:

Being able to identify the underlying subject information in a large document collection or corpus.

Converting the text information into digital information that is easy to model by adopting the method of "bag of words", which treats each document as a word frequency vector.

The cons of this methodology:

LDA adopts the bag-of-words hypothesis in the document, so the position information between words is ignored.

The topic generated by LDA is often occupied by high-frequency words, which leads to the limited role of low-frequency words in practical applications.

4) Results and their discussion

1. **What is the sentiment distribution of the “All Beauty” category reviews and how to choose the threshold for the latter study?**

   Using the Lexicon-based sentimental analysis to determine the VADER score of the review text, we realized that the distribution of VADER score(figure 4.1.1) is skewed to the right, which indicates that there are lots of high concertation of positive words in the reviews. This demonstrates most the consumer who purchased from “All Beauty” category wrote words that is correlated to positive feedbacks based on the threshold 0.5.

We are also interested in the relation between review VADER score and the actual rating of the products. The data contains more than 300,000 rows of data, so it will be hard to put them into the same graph. So, we only use the first 1000 VADER and overall score that the customer rating. At first, we expected the VADER score and overall rating would have a positive linear correlation. We think that people might have more positive reviews when they like the products and give a higher rating. Figure 4.1.8 is the scatter plot that demonstrates the correlation. The result of this figure disagrees with our assumption as it demonstrates that the different ratings will have different VADER scores for review text. We take a close look at that figure and find there are parts of the plot that matches our expectation. First, the positive rating groups (overall >3) have VADER scores that have a high concentration greater than 0.5. This finding is exactly what we expected ---- a higher overall rating will have a high positive VADER score. Secondly, the lowest overall rating (overall=1) has only two reviews with VADER score that exceed 0.5. This shows that the less satisfying consumer will less likely to leave a positive text. Based on the knowledge, we have found some explanations for the difference between our expectations and result. We use the VADER Lexicon-Based Sentiment Analysis, which have intrinsic drawbacks such as not domain-specific which could lead some VADER does not represent the review text. Besides that, there is a lot higher overall score(overall>3) data than the low score data. We think that might have things to do with sample bias. In the A/B testing class, we discuss that people who have positive online shopping experience are more likely to leave a review. We think that will also apply to our study, which explains the high overall score ratings are denser.

After exploring with VADER score alone, we assigned the overall score as our data true outcome. The overall rating higher than 3 points are good that received positive feedback while others are considered as negative feedback. We use a function to determine the VADER Polarity, which means VADER score higher than the set threshold will get positive feedback and the others will get negative scores. As we have set the threshold at 0.5, we get the table 4.1.2 as the outcome of the crosstab. As threshold = 0.5, the accuracy rate is 0.744 = (208383+68163)/371345, the recall rate is 0.738 = 208383/282236 and the precision rate is 0.909 = 208383/229329. The precision level is very high and the accuracy and recall rate are looking great. To view the difference, we increase the threshold to 0.6, which is presented as table 4.1.3. We use the formula to calculate the accuracy is 0.704, the recall rate is 0.673 and the precision rate is 0.916. This comparison shows that while increasing the threshold from 0.5 to 0.6, the level of classifying right decreased, the true positive when predicting positive decreased, but the probability of predict positive and the result positively increased.

This led us to the next step that we want to explore the maximum accuracy, precision and recall rate, which will help us get the threshold that we want to explore in the latter part of our research. We calculate the three rates for the threshold from 0.1-0.9 with an incremental of 0.1. The chart 4.1.4 shows all the result of our calculation. Then we plot these three elements separately. The figure 4.1.5-4.1.7 are the results. We can conclude that there is a negative correlation between the threshold and accuracy, which indicates when we increase the threshold, we are less likely to get correct predictions. The recall rate is also negatively correlated to the threshold. As the total real positive is the same, increase the threshold will decrease the number that we predict as positive, which leads to the decrease of recall rate. However, the precision rate acts the opposite of the others. We can see a clear positive correlation between threshold and precision level. The reason is that the increase of threshold makes the prediction positive number decrease but the number that is true positive decreases less. Based on the graphs, there is no such threshold that has the best rate of all.

To balance the different rates, we decided to use a threshold of 0.5 for our study, which will give us the relatively high precision and the best balance between accuracy and recall rate.

2. What are the most frequent topics when it comes to complaints?

For topic modeling, we apply LDA methods to explore the distribution of topics across documents and the distribution of words across topics.

When it comes to complaints, we decide to extract 4 topics from reviews text. And Topic 1 is the most frequent one. (33.8% of tokens)

For topic 1, the top-30 most relevant terms are: use; hair; skin; product; work; day; dry really; time; feel; face; like; week ;leave; take; well; apply; long; help; eye; keep; first; difference; cream; start; even; result; look; much; little. According to the content in Topic 1, we can conclude that the Topic 1 is about the **Personal Care**.

Topic 2 is the second frequent one. (25.8% of tokens). The top-30 most relevant terms are: good; use; shave; year; product; razor; buy; work; water; time; blade; last; cut; problem; shaver; teeth; old; new ;month; price; close; much; well; need; long; battery; floss; job; like; first. According to the content in Topic 2, we can conclude that the Topic 2 is about the **Facial Shaving**.

These two topics show the categories which receive negative feedback from consumers. Also, the words inside the topics show what consumers care about when purchasing. Consumers are dissatisfied with the products of Personal Care and Facial shaving. For Personal Care, consumers care about the experience (how they feel when applying the products on hair or skin) and the effect (result and time to get the result). For Facial Shaving, the condition of battery and razor might be the issues.

3. What are the most popular topics when it comes to positive feedback?

When it comes to positive feedback, we decide to extract 4 topics from reviews text. And Topic 1 is the most frequent one. (34.1% of tokens)

For topic 1, the top-30 most relevant terms are: use; skin; product; smell; like; love; feel; scent; time; face; soap; work ;oil; good; really; day; dry; cream; well; leave; time; great; help; great; natural; bottle; wash; look; body; long; little; last. According to the content in Topic 1, we can conclude that Topic 1 is **Personal Skin Care**.

Topic 2 is the second frequent one. (27.9% of tokens). The top-30 most relevant terms are: hair; color; like; look; use; brush; love; good; well; nice; really; perfect; little; great; work; product; nail; wear; size; long; pretty; fit; buy; easy; small; time; beautiful; lip; need; think. According to the content in Topic 2, we can conclude that Topic 2 contains the reviews of the **Beautify**.

The products of Personal Skin Care and Beautify bring consumers a positive experience. Words in these two topics show consumers will comment on the feature of the products, as well as showing their attitudes by the words “like; love; good; help”.

4. What are the differences in the sentiment between those who frequently leave reviews and those who seldom leave reviews?

From the frequency table of customers leaving reviews, we can find that about 77% of the customers only left review once (Table 4.4.1). We consider those customers as the customers who seldom leave reviews. In total, there are 287,784 customers in this category. The rest of the 83,561 customers are the customers who frequently leave reviews.

We use VADER Lexicon-Based Sentiment Analysis to analyze the sentiment. As we mentioned, a drawback of this method and our dataset is that without the true sentiments of customers review, it is hard for us to determine the threshold and to measure the accuracy of our predictions.

To address these issues, we try to use the overall score customers given to the products as the benchmark of the polarity of the reviews. We assumed that a higher overall score indicates a more positive attitude of the customer, and a lower score indicates a more negative attitude of customers.

However, based on our analysis, this assumption is wrong. The overall score and the VADER score resulting from the sentiment analysis do not have any correlation. (Figure 4.4.1) Therefore, we cannot use the overall score as the benchmark to determine the polarity of the reviews. We need to use our own judgment to determine the threshold and the polarity of the customer reviews. Besides, this reveals that some customers may leave negative reviews with a high score or leave positive reviews with a low score, which indicates an inconsistency in customers' opinions. Amazon needs to investigate the reasoning behind the two situations.

Based on our judgment, 0.5 was chosen as the threshold. After splitting the data into two parts based on the frequency of customers leaving reviews, we computed the VADER score of the reviews. From Figure 4.4.2 and Figure 4.4.3, we can see that more reviews had lower VADER score in the second category, that is the reviews of customers who seldom leave reviews.

Among the 83,561 reviews left by the customers who frequently leave reviews, there are 25,438 negative reviews, which accounts for about 30% of the reviews. Among the 287,784 reviews left by the customers who seldom leave reviews, there are 106,044 negative reviews, that is about 40% of the total reviews. One possible explanation for this is that some customers only left reviews when they dislike the products and left negative reviews about them.

5. Does the proportion of positive reviews grow year by year?

With the calculation of the above questions, we now have the VADER score and polarity of each review. We get the number of the year from "reviewTime" and group the data by year in order to learn: 1. The number of reviews over these years. 2. The trend of the proportion of positive reviews among all reviews over the years. 3. The trend of average VADER score changes year by year. The data for 2018 is excluded because the data for this year is not complete. This will influence our calculation of the total number of reviews yearly.

From Figure 4.5.1, we learn that the total number of reviews is constantly growing year by year. However, there is a decrease in 2017. From Figure 4.5.2, we learn that the average VADER score is decreasing overall, from 0.75 in 2000 to 0.42 in 2017. In 2003 and 2007, there are two temporary increases in the average VADER score, but the overall trend is still decreasing. From Figure 4.5.3, we learn that the proportion of positive reviews among all reviews is also decreasing in the All Beauty category. The line trend in this graph follows the same pattern as in Figure 4.5.2.

This should be an alarm to Amazon to find out why people are leaving increasingly more negative comments. Is it because the venders in this category are selling inferior products more than before? Should Amazon pay more effort to supervise on the products sold in this category? Or is it at all because the Amazon platform is giving people easier ways to leave a negative review than before?

6. What other factors lead to a higher probability of a positive comment?

In this part, we use the np.corrcoef(x, y) function to check the Pearson correlation between “image” (0 if not exist 1 otherwise) and “VADER Polarity”(1 of positive and 0 if negative). We get a coefficient of 0.06435181, lower than 0.3, which shows no significant correlation between posting an image and probability of being a positive review.

With additional correlation check between “overall” and “VADER Polarity”, we find a correlation of 0.27840418, which means a correlation between the overall rating and the sentimental polarity of a review. Higher ratings definitely lead to higher more positive polarity. This is consistent with our intuition.

We also check the correlation between “verified” and “VADER Polarity”, and find a -0.14454645 correlation. There is a minor negative correlation between "verified" and "VADER Polarity". This means that if a review is verified, it is more likely a negative review. This can be explained intuitively because a negative review will more likely catch the attention of the Amazon platform and be verified and solved. As customers, we care about the vender’s attitude toward our complaints more than our praise.

5) Conclusion

1. Our Findings and Recommendations for Amazon

1. There are more positive overall ratings(overall>3) than a negative overall rating. This is not seldom in the online shopping business. Amazon should take a close look at this and figure out if consumers are really satisfying or they just do not want to leave reviews. They can use the review rate (the number of reviews per 100 purchase) that might help them identify the real feedbacks.
2. There is an inconsistency between the customer reviews, their summary, and the overall scores. Customer reviews are important references for the sellers to improve their products and for future customers when choosing the products. The inconsistency and lack of information will have a negative impact on their value. Amazon needs to investigate the reason behind these and tries to make customer comments more consistent and contains more information.
3. The customers who seldom leave reviews are more likely to leave negative reviews. This may be those customers only leave reviews when they have a negative experience. In this circumstance, the customer reviews are biased and do not represent the true voice of all the customers. Another explanation is that those who purchase less are more likely to choose low-quality products. Amazon needs to compare their frequency of leaving negative reviews with their purchase records to find out the reason and try to make the customer reviews more representative.
4. The total number of reviews is increasing year by year. However, the proportion of positive feedbacks is decreasing, and this should remind Amazon to find out the reason for people increasing complaints about All Beauty products. Another doubt is that, do more negative reviews definitely lead to fewer purchases? Intuitively, people tend not to purchase a certain product if he sees a negative review under the description. However, we may confirm this by drawing a link between the purchases amount and polarity of the reviews of a certain product. If we also have the sales data over the years of this All Beauty category, we can confirm that if the sales are negatively influenced by the negative reviews.

2. Possible shortcomings and ways to overcome them.

1. **The limitation of our dataset:** The true sentiments of customer reviews are unknown to us, which is the target. In normal classification problems, we have the outcome variable, so that we can train with the training dataset and predict with the test dataset to check the accuracy. Without sufficient data, we have to choose the VADER threshold based on our judgment, which may be biased.

Specifically, with the 0.5 threshold, we do not get the exact positive and negative reviews split. From the word cloud in Figure 5.2.1, we see an overall positive review pattern. If we classify most of the positive reviews as negative, we may arrive at the wrong conclusion. From the word cloud in Figure 5.2.2, we find that the summary represents people's sentiment better than reviews. Therefore, we did the sentiment analysis with the "summary" column data. However, we find the same problem with the summary data – the accuracy rate has a positive correlation with the threshold, as is shown in Figure 5.2.3. So, we still use our original "histogram" approach.

2. **Lack of purchase data:** We don't have the data about the purchase information for these customers. Therefore, we cannot draw a link between the purchase behavior with the review feedback. If we have the purchase data, we may have a lot of other interesting topics to study. For example, we can find out whether making a positive review lead to more purchase and whether negative reviews lead to less purchase. We can also study if purchase frequency is influenced by the review frequency and sentiment.

6) Appendices

Figure 4.1.1 Histogram of VADER Score Concertation

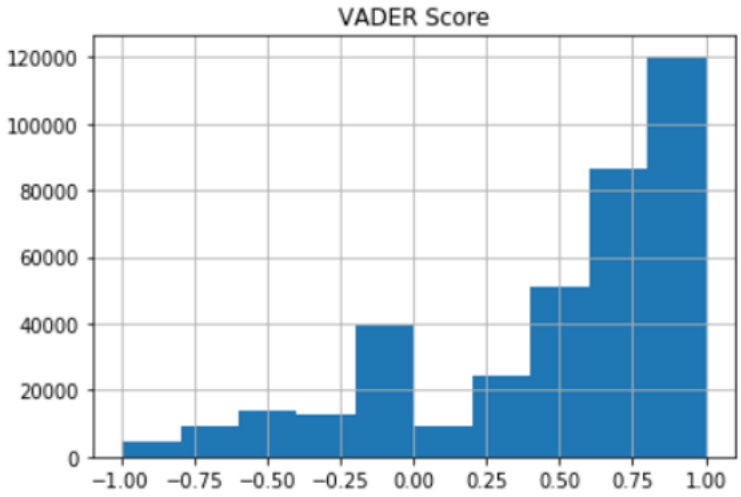

Table 4.1.2 Threshold =0.5 Confusion matrix

Predicted:	negative	positive	All
True:			
negative	68163	20946	89109
postive	73853	208383	282236
All	142016	229329	371345

Table 4.1.3 Threshold =0.6 Confusion Matrix

Predicted:	negative	positive	All
True:			
negative	71868	17241	89109
postive	92423	189813	282236
All	164291	207054	371345

Table 4.1.4 Threshold vs Accuracy, Precision and Recall Rate

Threshold	Accuracy	Precision	Recall
0.1	81.2%	86.9%	88.7%
0.2	81.2%	87.7%	87.7%
0.3	80.7%	88.8%	85.3%
0.4	79.5%	89.9%	82.3%
0.5	74.5%	90.9%	73.8%
0.6	70.5%	91.7%	67.3%
0.7	59.3%	91.7%	51.1%
0.8	51.7%	92.7%	39.5%
0.9	39.0%	93.4%	21.3%

Figure 4.1.5 Accuracy Rate vs Threshold

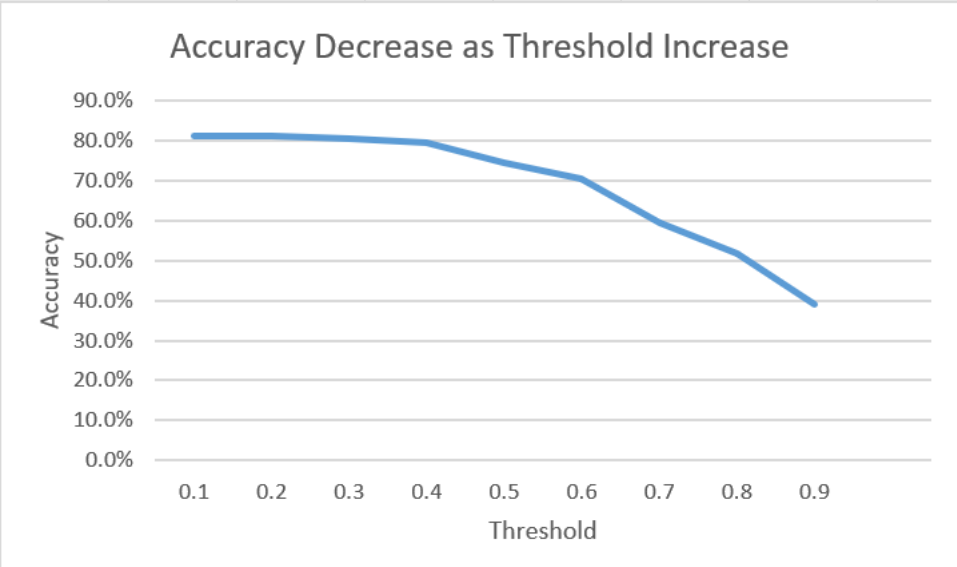

Figure 4.1.6 Recall Rate vs Threshold

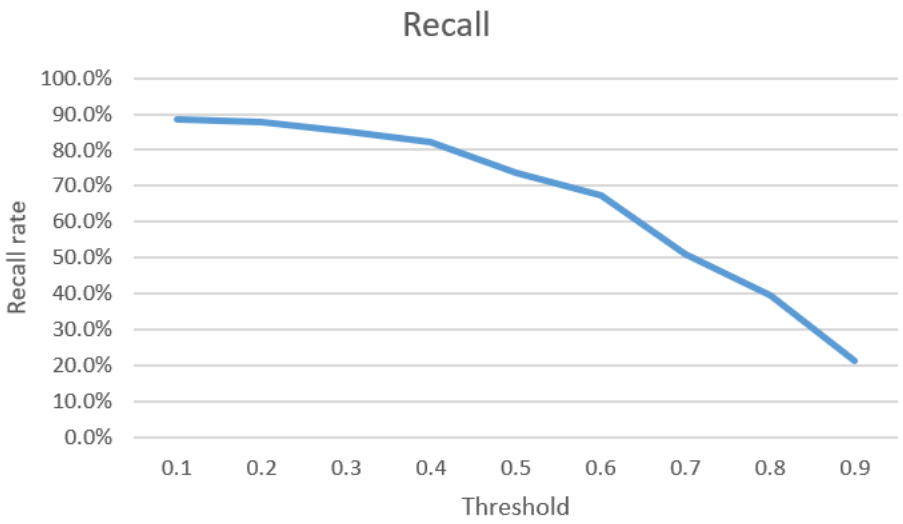

Figure 4.1.7 Precision Rate vs Threshold

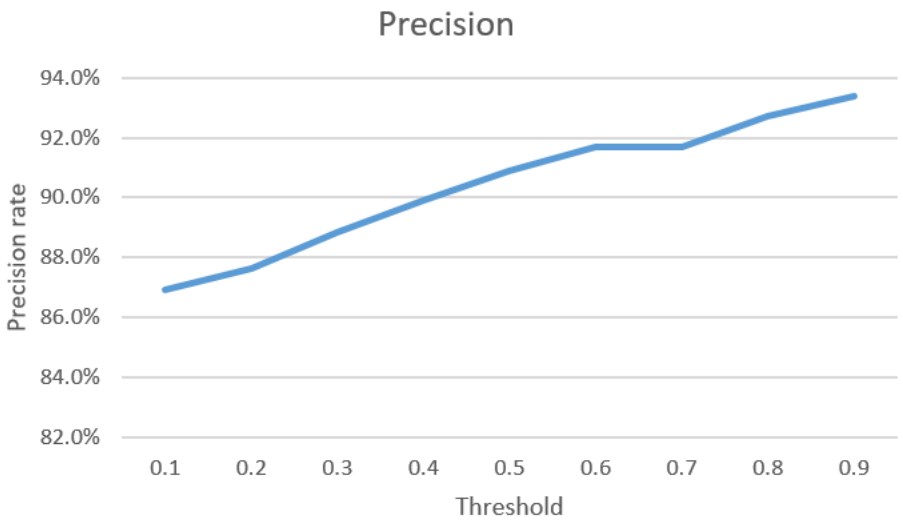

Figure 4.1.8

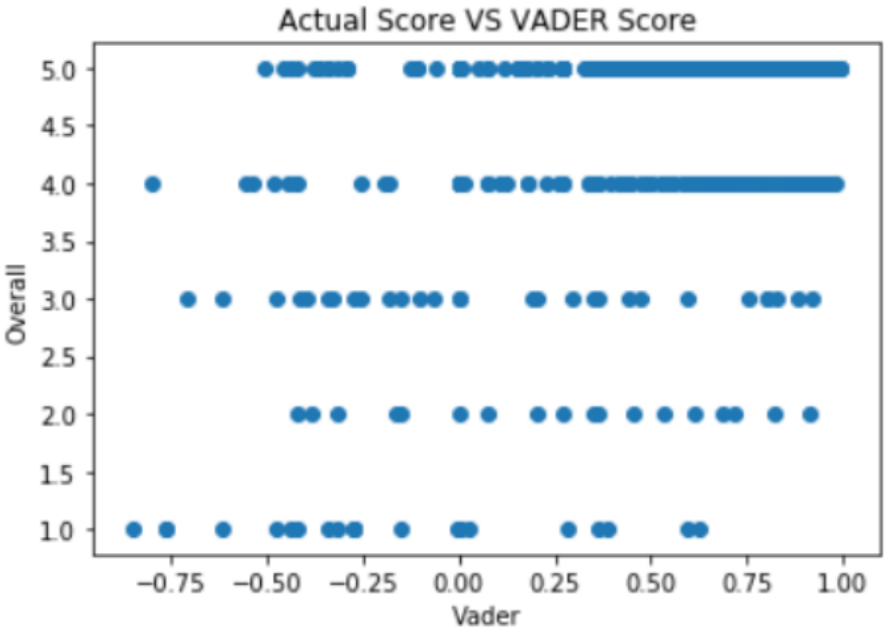

Figure 4.1.9

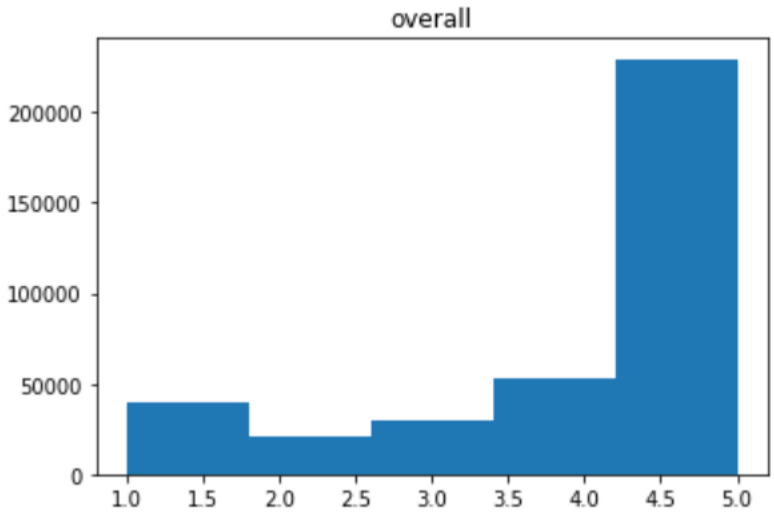

Figure 4.2.1 Topic 1

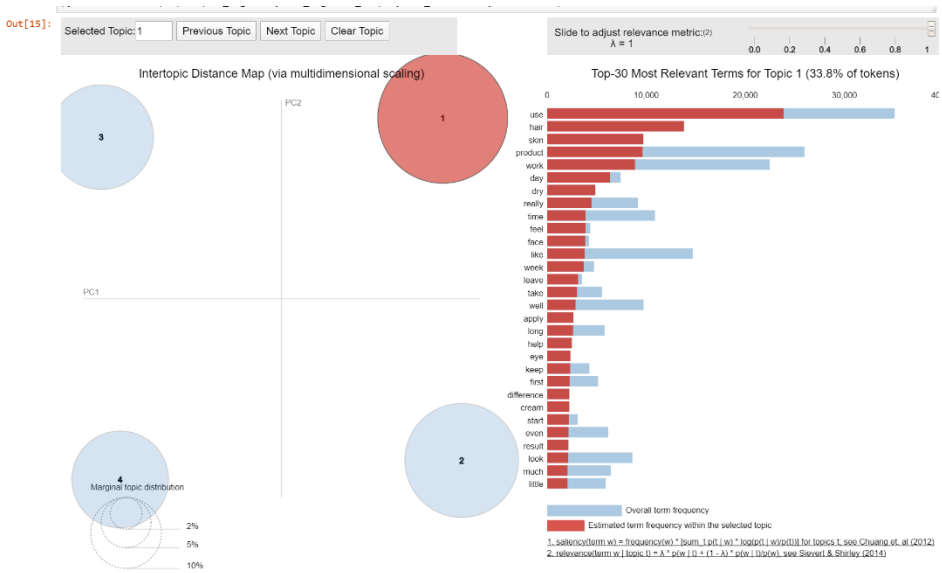

Figure 4.2.2 Topic 2

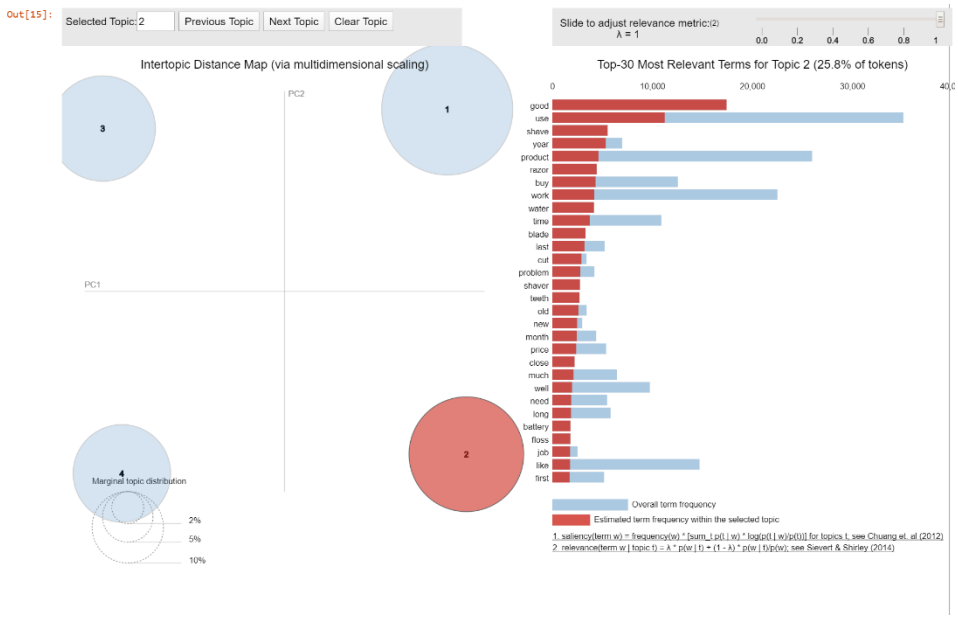

Figure 4.2.3 Topic 3

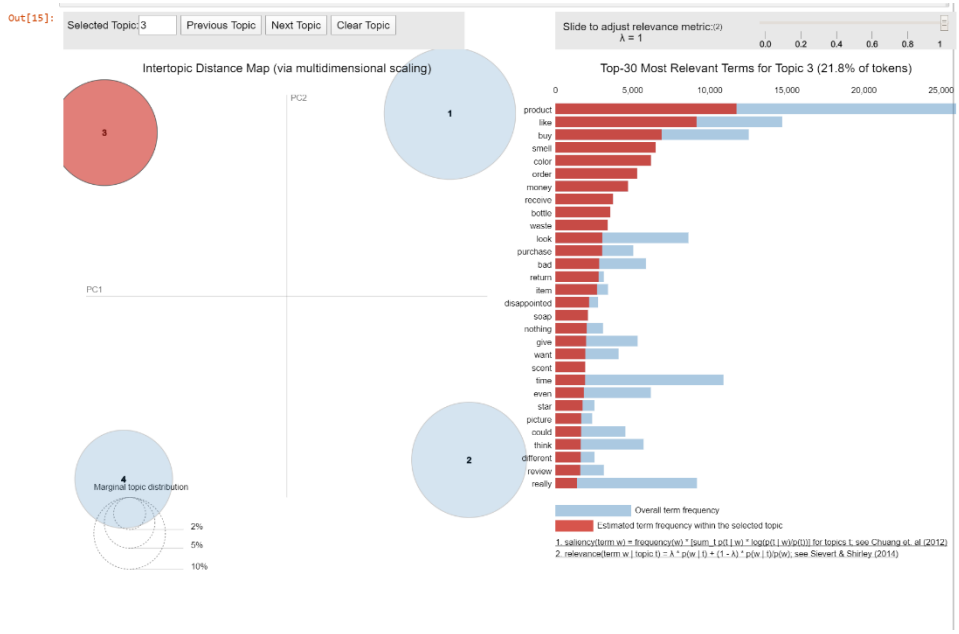

Figure 4.2.4 Topic 4

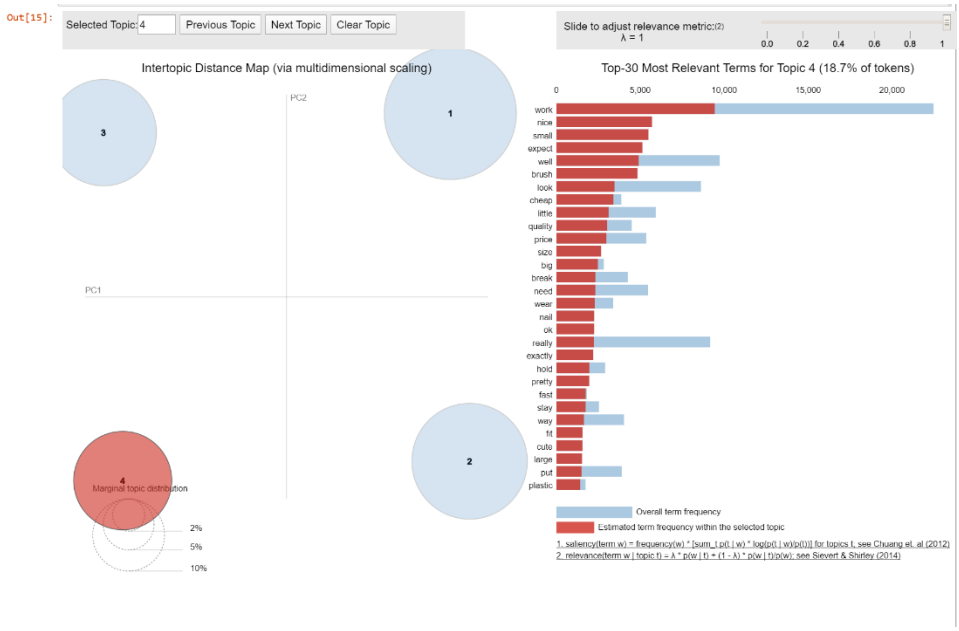

Figure 4.3.1 Topic 1

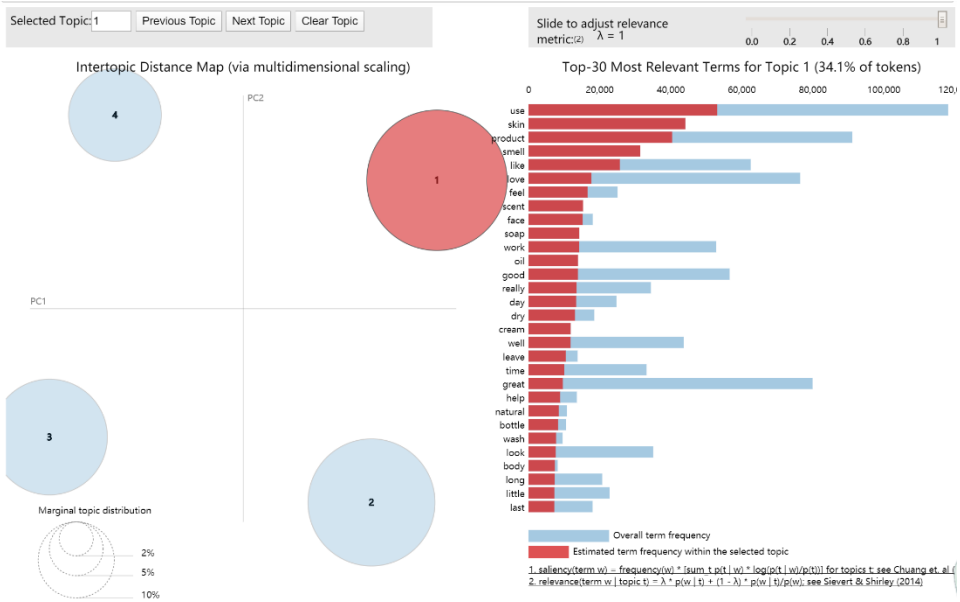

Figure 4.3.2 Topic 2

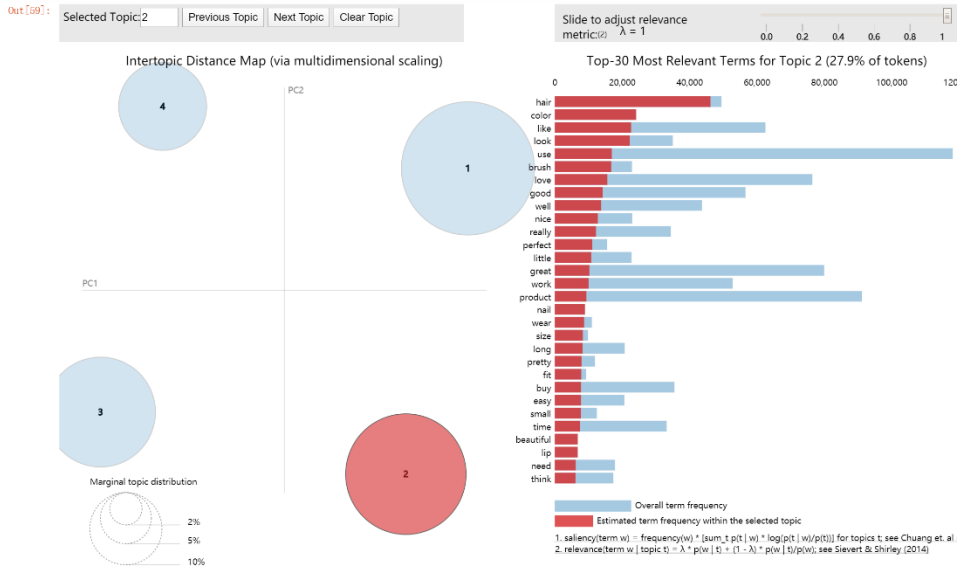

Figure 4.3.3 Topic 3

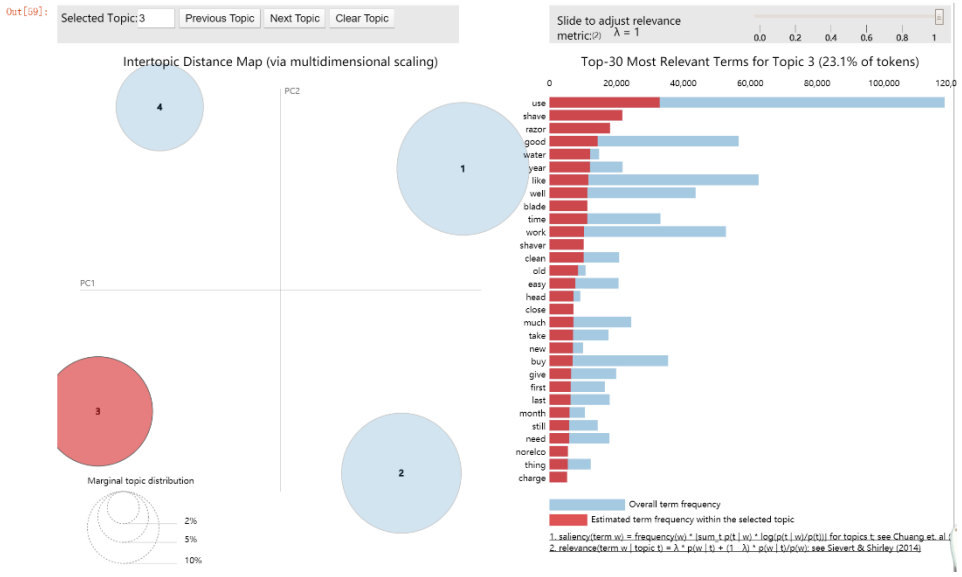

Figure 4.3.4 Topic 4

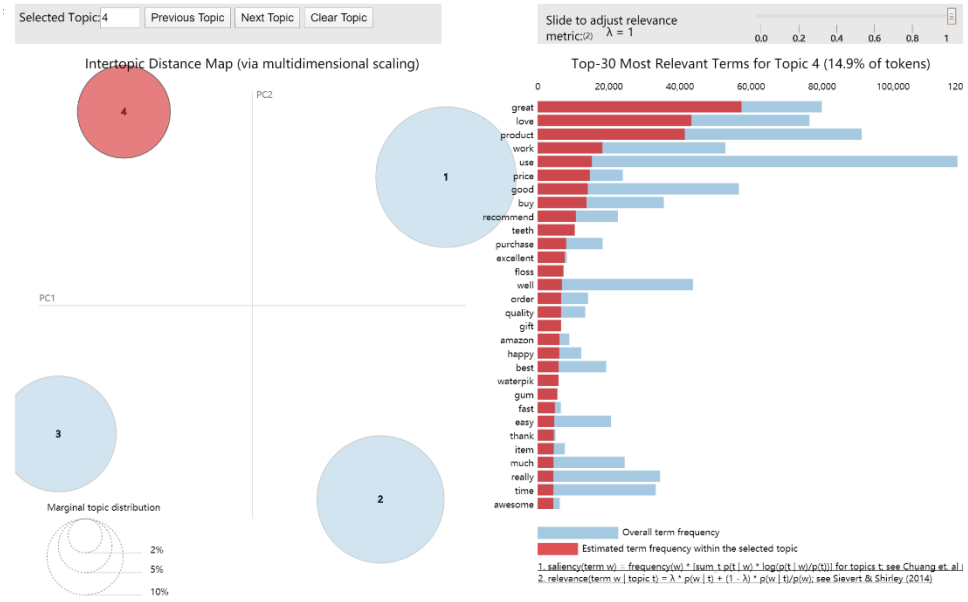

Figure 4.4.1 Correlation between VADER Score and Overall Score

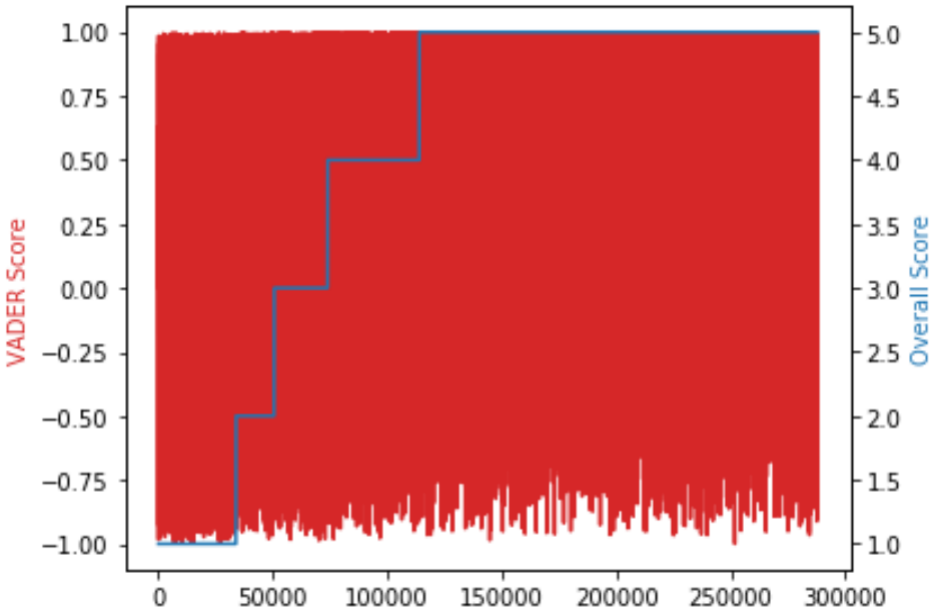

Table 4.4.1 The frequency of customers leaving reviews

The times of review	The number of customers
1	287784
2	30135
3	3702
4	1019
5	774
6	455
7	66
8	35
9	24
10	16
11	10
12	7
13	3
14	2
21	2
23	1
18	1
20	1
27	1

Figure 4.4.2 The VADER Score of the sentiment of customer who frequently leave reviews

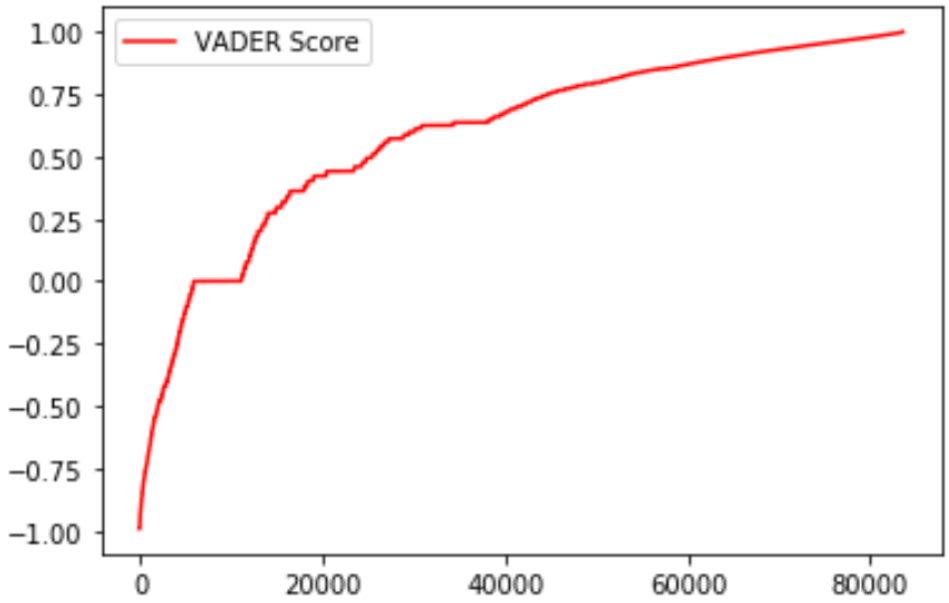

Figure 4.4.3 The VADER Score of the sentiment of customer who seldom leave reviews

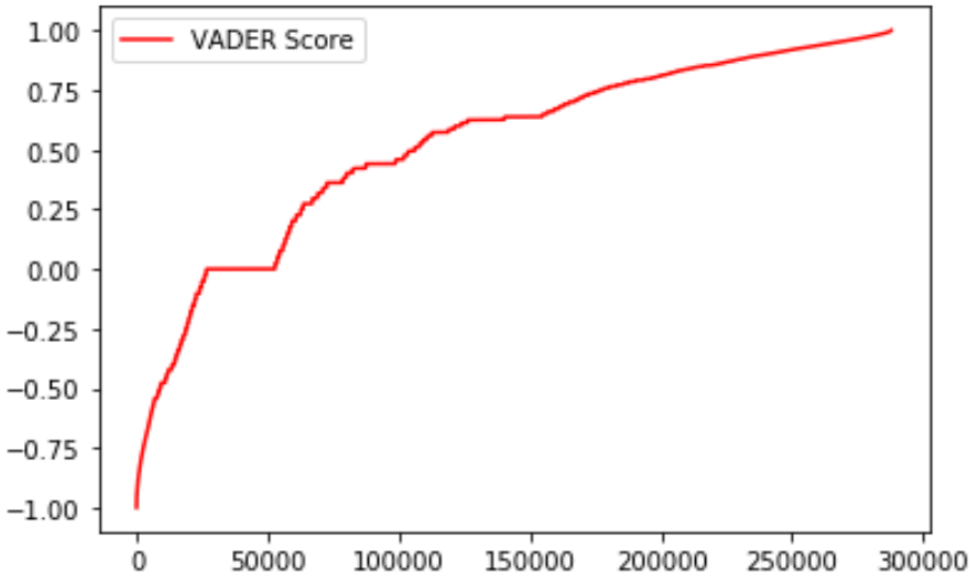

Figure 4.5.1 The Number of Reviews Over the Years (exclude 2018)

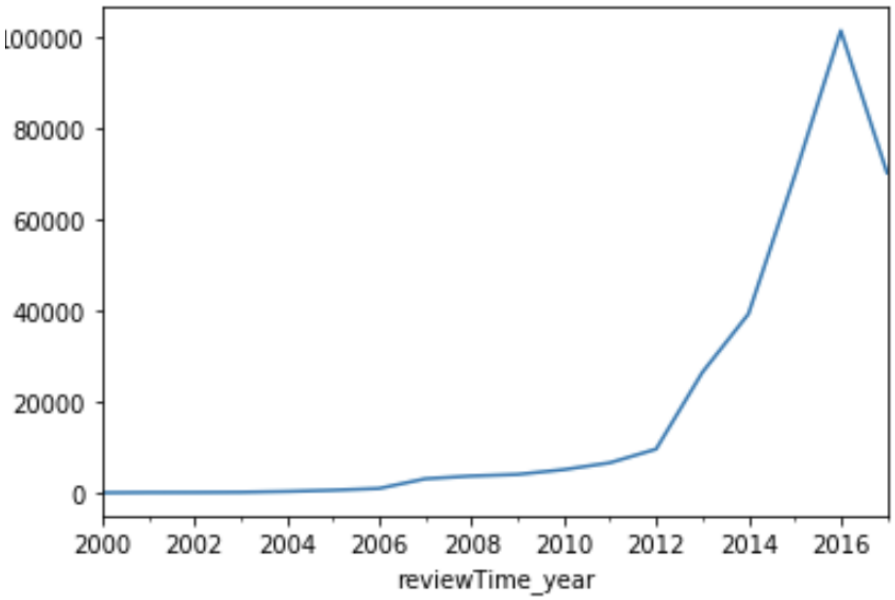

Figure 4.5.2 The Average VADER Score Over the Years (exclude 2018)

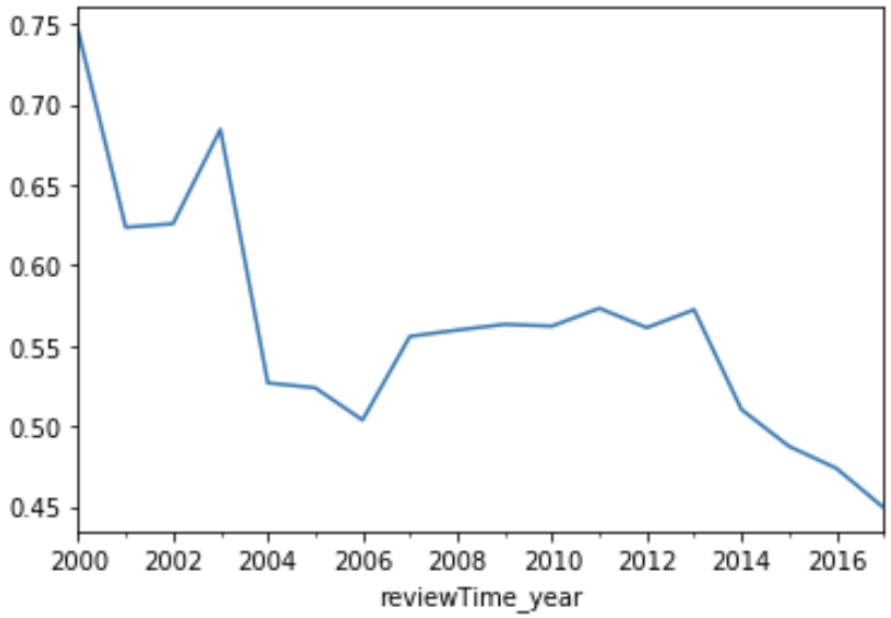

Figure 4.5.3 The Number of Positive Reviews Proportion Over the Years (exclude 2018)

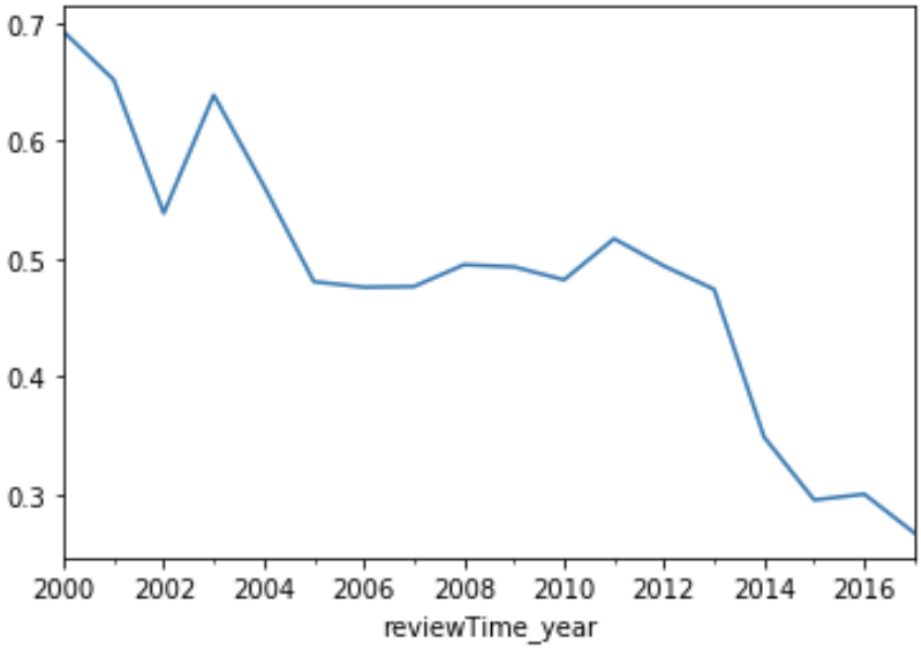

Figure 5.2.1 Word Cloud for "Review"

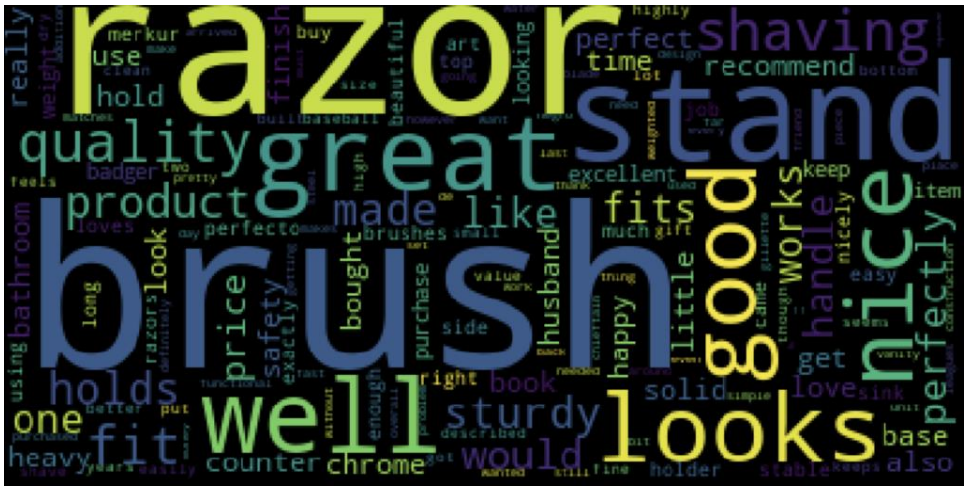

Figure 5.2.2 Word Cloud for "Summary"

Figure 5.2.3 Accuracy Rate vs Threshold