Reshape Buying for Your Brand - Take Baby Category as an Example

Part 1 Executive summary

Our first goal is to find the best classifier. It can help baby products producers better understand the general views on their products even from some unlabeled online reviews such as people's comments on Twitter. Obtaining more opinions and feedback from those unlabeled data can help business owners improve their future operational decisions. In addition, we want to understand the general topics in the baby products reviews. Once new ventures know what customers are most concerned with, they can capture more market shares by focusing on improving their competitive advantages in those most concerned areas. In addition, we want to capture customers' sentiments from their reviews to know whether our products can lead to satisfaction or whether we should keep improving our products to nurture loyal customers. We do unsupervised learning to define and predict review sentiments.

Part 2 Data description

Customer ratings have similar percentages in other months. The customer's rating is divided into five grades, among which five stars are the highest score, and 1 star is the lowest score. The standard deviation of the proportion from high to low score is 0.71, 0.65, 0.45, 0.22, and 0.14, indicating that customer scores were less affected by seasonal and time factors.
The proportion of positive feedback is much higher than negative ones. The scores from high to low accounts for 58.11%, 20.56%, 10.73%, 5.73%, and 4.88%. With a score of five being the highest, customers are more likely to give positive feedback than negative ones. Moreover, customers are relatively tolerant and rarely give the lowest score.
The distribution of ratings for the top 10 reviews differs from the overall customer ratings. When we analyze the top ten reviewers with the most comments, we find that 50% of reviewers' proportion of 5 scores is lower than that of the overall ratio. And 70% of them have a proportion of 4 scores higher than the overall proportion. Also, one customer scored 100% of five scores. It's reasonable to assume that some businesses have increased their overall share of 5 scores because they advertise themselves or their customers praise them indiscriminately.

Part 3 Project objectives

All merchants in Amazon hope to increase their sales volume and reputation, which is also our objective in the project. We will help partners grow their sales and positive feedback in three ways.

Customer Analytics. We will build a score prediction model by text classification. By Importing customer reviews to the model, businesses can get the main features that can describe their customers' overall ratings, which will be extremely helpful for those businesses with low store ratings or a high number of reviews on media platforms.
Positive Attributes. We will help you find out the most effective attributes in your specific category by topic modeling. Businesses can adjust their strategies, pricing, and product design based on these keywords.
Sales Growth. Generally speaking, deciding whether to launch a product to boost sales is expensive. By understanding the category's strengths and weaknesses, businesses can save costs and develop appropriate business strategies, which will help them win more sales.

Part 4 Methodology

1. Text Classification.

1.1 Goal

One primary goal of our project is to find the best classifier such that we can use it to understand customers' comments on the baby products from those unlabeled reviews such as reviews from tweets. To help Amazon store owners better capture public comments on their products, we try to improve the forecasting accuracy of the classifiers.

1.2 Classifiers

Since we have split our ratings into three classes: High(with a rating of 5 or 4), Medium(with rating 3), and Low(with rating 2 or 1), our classification is a multiple-class classification problem. We use many general classifiers, including linear discriminant analysis, quadratic discriminant analysis, logistic regression, classification tree, Boosted tree, random forest, support vector machine, and neural networks. As the vectorized text data would be highly non-linear, we decided to use more flexible classifiers such as Trees, SVM, and neural networks. We anticipate that these classifiers will work well since these modern methods are proved to have good performances even on complicated classification problems.

1.3 Process

We randomly selected 10,000 reviews from the whole dataset as the training dataset, consisting of 3,350 High rating reviews, 3,300 Medium rating reviews, and 3,350 Low rating reviews. We also randomly extracted 1,000 reviews from the remaining dataset as the testing dataset, consisting of 335 High rating reviews, 330 Medium rating reviews, and 335 Low rating reviews. Then to implement the random forest, simple tree, naïve Bayesian method, and the support vector machine, we use the Chi-squared statistic as the criterion to select the 500 best features from the training dataset and use these features as the testing features.

After determining the features or the vectorized predictors, we train our model using different classifiers and make predictions on the 1,000 testing reviews correspondingly. We also implemented the Convolutional neural network(Conv1D) and Recurrent neural network to train the model and then compute the model's prediction accuracy on the testing set. The tuning process of neural networks needs a tremendous amount of work. However, our primary focus in this project is not tuning our deep learning architecture. The accuracy rate obtained from the neural networks can serve as a reference to other classifiers, since the complicated deep learning architecture should generally have better prediction accuracy.

1.4 Pros and Cons

Pros: Classifiers are easy to implement and can return reasonable results based on the trained models. The testing accuracy rate can be used as the reference to gauge whether the model has the generalization effects and can help us select the best classifier accordingly. The prediction accuracy is not bad and has practical meaning.

Cons: Some classifiers have closed testing accuracy, and it is not decisive to determine which classifier works the best. In addition, the number of training reviews is still limited, and we may improve the model potency by using more data to train. Our accuracy rate is only about 60%, and it is hard to improve further, which could make business owners skeptical about the prediction results after using our classifiers.

2. Topic Modeling.

2.1 Goal
Our goal for topic modeling is to find some related topics in the Amazon Baby Category reviews. We aim to find the common words in each group representing the baby products and make suggestions from our observation. From the result, we can estimate the high-frequency words with positive feedback due to the higher volume of positive feedback.

2.2 Process
We chose 10,000 reviews at random from the entire dataset as the corpus, including 3,350 High rating reviews, 3,300 Medium rating reviews, and 3,350 Low rating reviews. Later, we used Latent Dirichlet Allocation(LDA) to classify the text into different topics. The end goal of using LDA is to find the optimal distribution representation in each topic. We use the tool above to unique abstract tokens that occur in the reviews by different numbers for the topics. We tested 2-5 topics during the test and discovered that 2-topic is the best outcome for higher coherence, log-likelihood, and perplexity.

2.3 Pros and Cons
Pros: LDA for Topic Modeling can help us organize a larger amount of reviews, including some hidden topics. Moreover, it provides the frequency for each word, both in general and in each topic.
Cons: LDA for Topic Modeling is an unsupervised method, and therefore, the outcome might be a misleading result and not highly relative in the human aspect. Also, some frequency words could appear in all topics.

3. Sentiment Analysis.

3.1 Goals
Our goal for sentiment analysis is to understand customers' actual experiences with our product. Different customers may have different rating standards, but the emotions that they want to convey have something in common. Sentiment analysis can help us accurately capture customers' thoughts and be applied to others' reviews on social media.

3.2 Process
We use the VADER lexicon from the "NLTK" package in Python to analyze and predict the reviewers' sentiments. Also, we compare the predicted sentiments with the actual sentiment, which are substituted by rating scores, to find which sentiment classifier works the best. First, we write a function called "analyze_sentiment_vader_lexicon," within this function, we define the initial thresholds for all three levels. At that time, we assume that "threshold 1 = 0.3", "threshold 2 = 0.7," and the accuracy rate is 45.2%. In order to figure out what factors can help us improve the accuracy rate, we print out the Confusion Matrix(Figure 5). We also compare the advantages and the disadvantages of this method.

3.3 Pros and Cons
Pros: This sentiment analysis is based on the Vader lexicon, which can run really fast compared to other methods. This method is very intuitive and easy to explain, which can help us adjust the parameters.

Cons: It is not good at dealing with sarcasm. The accuracy rate obtained from this method is relatively low. The lexicon is fixed, and it is difficult to change the lexicon specifically for one particular scenario.

Part 5 Results and Discussion

1. Classification,

The classifier model with the best performance is the Random Forest classifier. It has a testing accuracy of 60.51%, and its accuracy on high, medium and low predictions are 69.3%, 62%, and 53.4%, respectively. The Tree method performs the worst, with an overall testing accuracy of only 47.02%, and its classification accuracy for low is as high as 42.69%. The SVM and Naïve Bayesian methods have similar performances, and their prediction accuracies are also very close to that of random forest, which is about 60%. To see whether the accuracy can be improved further, we implemented CNN and RNN with long short-term memory to calculate the testing accuracy by tuning parameters to balance the underfitting and overfitting problem. However, we feel surprised to find that the testing accuracy cannot break through 62% and starts to fall as we train more. Therefore, we kind of believe that the general models can only have 60% accuracy on unobserved data.

In terms of individual class prediction accuracy, the prediction of models on the high rating is the best, followed by medium rating, and a low rating is the worst. We found that the model seems to confuse high-rated and low-rated reviews easily, with about 23% of high-rated information classified as low-rated and 23% of low-rated information classified as high-rated. We find that much low-rated information may use a large number of positive words to express sarcastic meaning or praise before criticism. The trained model cannot accurately distinguish the low-rated testing reviews from the high-rated ones.

Comparison of Each Classifier Model						
	Naive bayesian	**Tree method**	**Random Forest**	**SVM**	**CNN**	**RNN**
Accuracy	0.5771	0.4702	0.6051	0.5963	0.6032	0.5523

2. Topic Modeling

According to the scores of each model, 2-Topic is the best. The performance gets lower when the number of topics increases.

Comparison of Each Number of Topic				
	2-Topic	**3-Topic**	**4-Topic**	**5-Topic**
Coherence of the model	-1.58	-1.631	-1.758	-1.79323
Coherence by topic (higher values are better)	[-1.63 -1.52]	[-1.495 -1.915 -1.482]	[-1.6708 -1.7848 -1.5817 -1.9948]	[-1.83277 -1.77149 -1.82501 -1.85964 -1.67724]
Log-Likelihood (higher values are better)	-3335755.82	-3347485.53	-3354507.54	-3372711.93
Perplexity (lower values are better)	1581.77	1623.28	1648.65	1716.27

However, when we look into the clusters. We could see that the topics are mostly product categories and some coherent words regarding the specific sectors. Therefore, we can gain insight into different products from the cluster. Combining the scores and observed outcomes, we choose the 3-topic model for further evaluation.

From the 3-topic model, we can define three groups:

1. Baby stroller/seat/bed/car - most people are concerned about the comfort, the fitness for the size, convenience to take, and the safety from strap and crib. (Figure 1)
2. Baby bottle/diaper/cup/nipple - buyers care about cleanness, leaking problem, status after wash, and convenience to hold the baby. (Figure 2)
3. Baby toys - most reviewers rate for appearance, color, and whether the baby likes it or not. (Figure 3)

Also, from the general analysis, we can analyze the most popular products for reviewers. Bottle, seat, and diaper are the top 3 products (Figure 4) which might be the most purchased items on Amazon. As a result, there is a more outstanding total available market for those popular products, which means the new vendor can sell these products for their first stage to the Amazon platform. These results could help us understand the critical factors for each product and help our clients improve their products' rate and review from an emphasis on these features.

3. Sentiment Analysis

As (Figure 5) shows, it is not good at predicting low rating scores. Lots of low rating scores are predicted as median or high rating scores. As a result, we adjust the threshold as follows: "threshold 1 = 0.5", "threshold 2 = 0.8," and the accuracy rate is about 46.4%, which improves our accuracy rate by about 1.2%. Next, we print out the newest Confusion Matrix(Figure 6)

From (Figure 6), we notice that we cannot significantly improve our model's accuracy rate even though we improve the threshold. The reasons to explain the confusion matrix are that there are many reviews on our dataset that do not follow the primary trend evaluation of good and bad. In other words, there is lots of sarcasm. For example, when the model tells us this review should be given a High score while actually, it gets a low score, there are about 60 reviews like that.

Part 6 Conclusion

Recommendation and Application:

For new clients: The new vendors can utilize topic modeling results to observe the rather popular products for their strategy to launch the new brands on Amazon. For instance, bottle, seat, and diaper are the top 3 products the vendors can launch in the first stage. After all, we suggest ventures on Amazon can give out anonymous surveys and collect more comments on our products from new customers such that we can use the best classifier to see the overall feedback. Suppose the classifier tells us most anonymous reviews are labeled with terrible consuming experiences. In that case, we should be alert about this and try to discern what is going wrong in our business. Thus, the classifier can be used very easily as long as we can set review collecting sections and allow new clients to make some voice.

For current clients: We can provide insights regularly to help existing clients monitor customer feedback. To combine the frequency and the importance of each word, the clients can optimize its products and strategies on Amazon. For example, we can email our current clients regularly and then use the classifiers to label their comments to understand whether our current customers got satisfied recently and then decide whether we can do more work to make them become our loyal clients. We can also use the sentiment classifier to understand the current customers' sentiment drift such that we can do some remedies to avoid them from escaping. In addition, once we know some customers have become more interested in our products, we can keep in touch with them more often and make them loyal by providing better service to them.

Shortcomings and Ways to overcome:

The main shortcoming of our classification is that the overall prediction accuracy cannot be improved further, such that owners may feel reluctant to use our model. Prediction accuracy problem is challenging to overcome, especially when we try to classify the sequence data with semantic linkage.

Based on the results from the three methods, we have a 60% accuracy by Random Forest, which is a

decent outcome because some words are duplicated in both positive and negative feedback and also some neutral words that are hard to define. If we need further improvement for the classification, we can add on Checkpoint and Early Stopping to prevent overfitting. Combining with professional insights, some weights of the words can be adjusted to build a more accurate model.

For topic modeling, if the brands can use focus groups, the outcomes can avoid misleading problems. In the future, we can also derive the advanced topic modeling from the original outcome, which means we can figure out the sub-topic in different kinds of products with more organized features to make the analysis more related to each topic.

Lastly, we can improve our results by using special lexicon**s for sentiment analysis**. For example, we can use the results from text modeling as our lexicon, which I believe can improve the sentiment analysis result.

Part 7 Appendices

Figure 1(Topic Modeling)

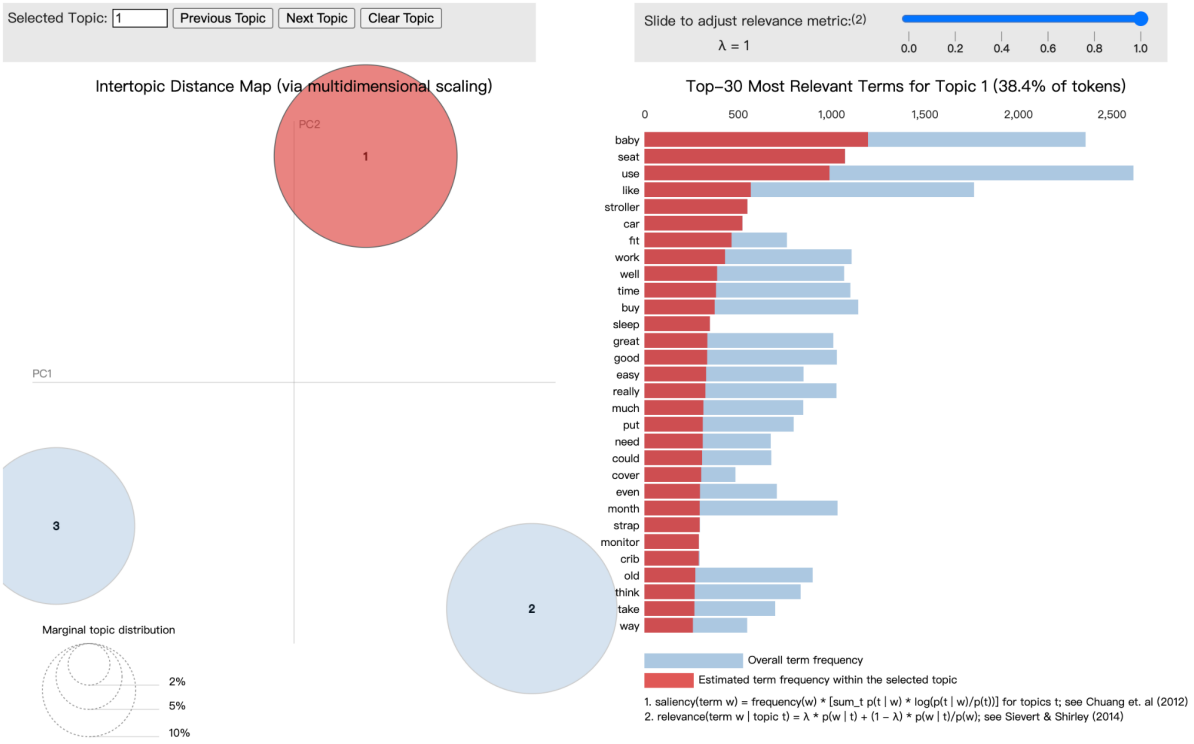

Figure 2(Topic Modeling)

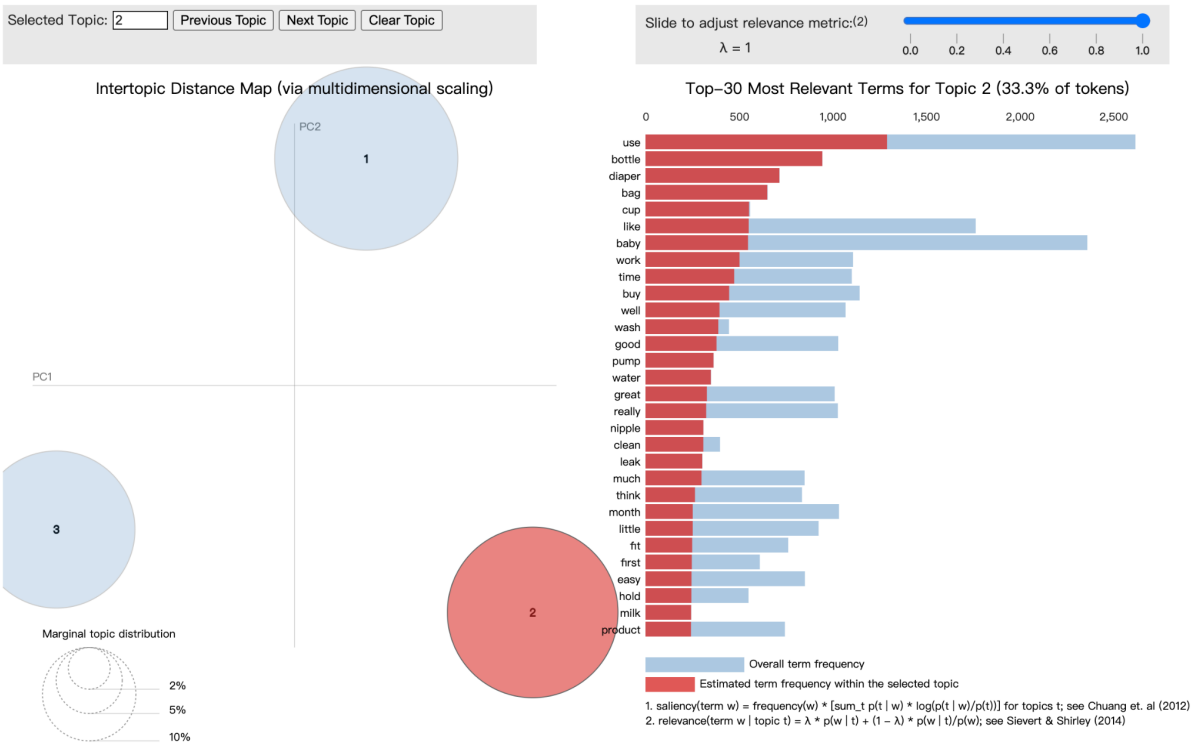

Figure 3(Topic Modeling)

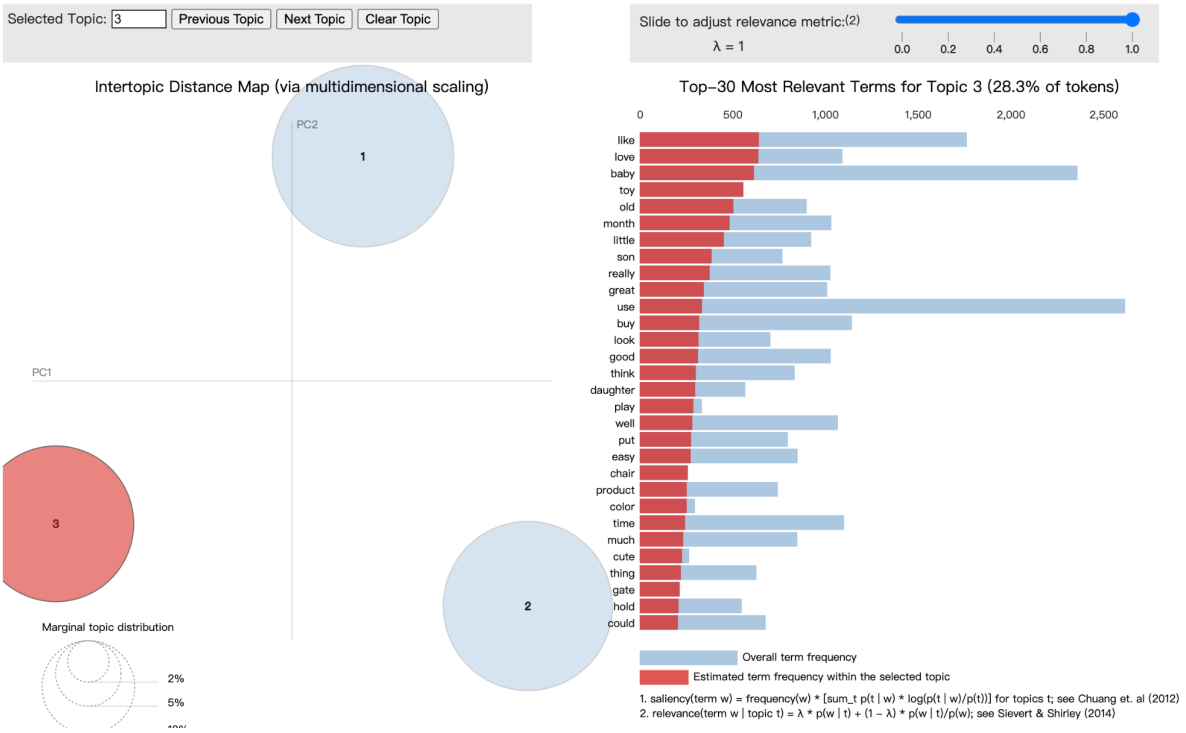

Figure 4(Topic Modeling)

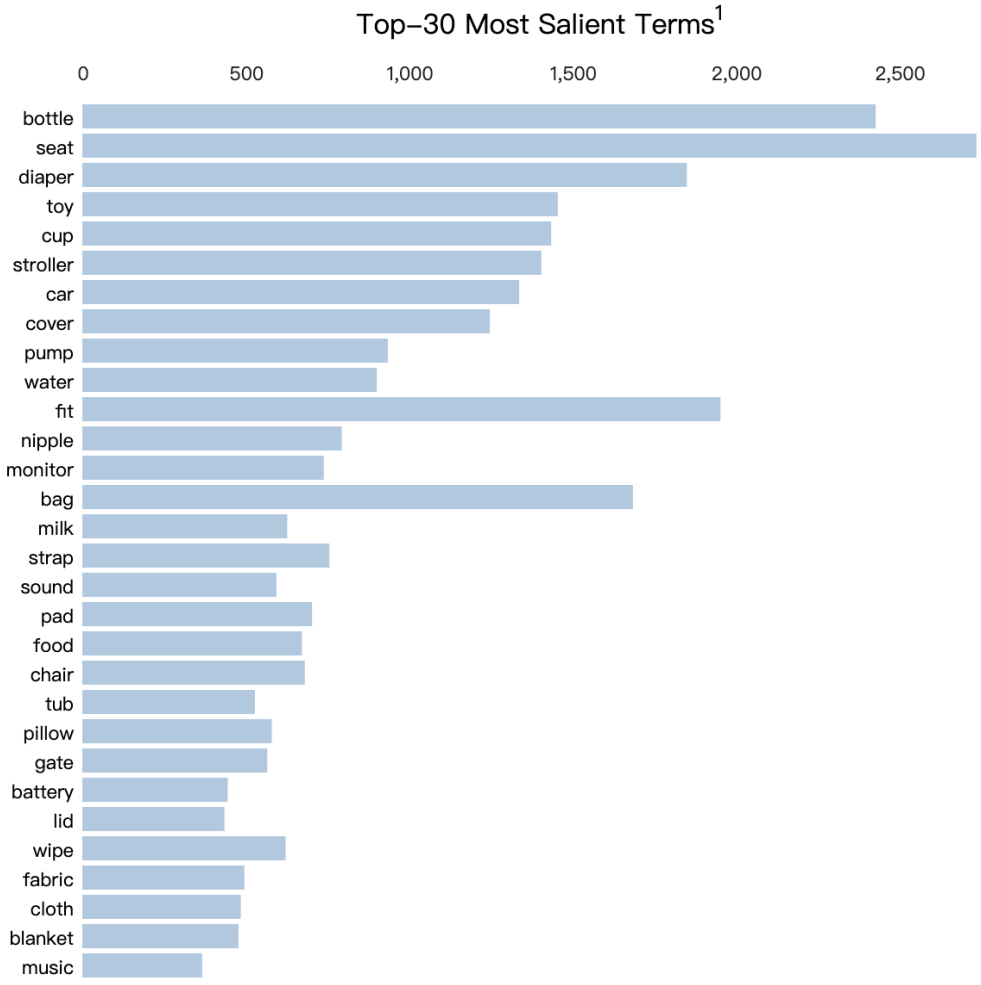

Figure 5(Sentiment Analysis)

Confusion matrix:

	High	Medium	Low
High	271	23	41
Medium	145	121	69
Low	226	57	70

Figure 6(Sentiment Analysis)

Confusion matrix:

	High	Medium	Low
High	236	39	60
Medium	114	146	75
Low	182	78	93

Part 8 Reference

Source: Github - Amazon Review Data