Module 5 Technique Practice Report

ALY6040 – Data Mining, Northeastern University

Professor Justin Grosz

14 May 2022

Submitted by,

Sharon Appoline Rosary (rosary.s@northeastern.edu)

Abstract

The agenda of this assignment is to analyze Amazon fine food reviews data that is available on Kaggle.[1] We will be using mining techniques to bring out analysis of review texts using concepts of Natural Language Processing. The goal of the assignment is to find a theme across reviews and discover hidden topics in the review data.

Data cleaning

The data consists of 568,454 fine food reviews with 10 features namely Product-ID, Review Score, Summary of the review, Complete text entry of the review, and time of the review. The data was sampled for random 300,000 rows to align with the local system processing power and analysis. The Score column consisted of values ranging from 1 to 5 – 1 being lowest score and 5 being the highest score. There were around 22,529 reviews with neutral entry. These reviews were no way helping our goal to bring out themes across reviews. Hence, these rows were removed. Post removing these rows the dataset consisted of 277,471 reviews. Since we have eliminated reviews with score 3, can categorize the review score in to positive (4 & 5) and negative (1 & 2) reviews.

Doing that we have 234,109 datapoints with positive reviews and 43,189 negative reviews. The data has 2 columns Helpfulness numerator and Helpfulness denominator, using these two columns, a new column called sentiment was created. The value was entered as positive if the ratio of Helpfulness numerator and Helpfulness denominator is greater than 0.8. The review columns were mostly in Camel case format. It was first converted into all lower case for standardization purposes using regular expression. A new column called Summary_clean was created using Summary feature, by removing special characters and symbols. Post this point Summary_clean feature was used to analyze the review data. In the next steps, the cleaned data is used to perform exploration and visualization.

Exploratory Data Analysis

In the exploratory part of the analysis of text analysis, like numerical data we cannot build a correlation plot between variables. Applying the concept of Count Vectorizer I visualized a word cloud that is clustered based on positive and negative review score. Count vectorizer transform a given text into a vector based on the frequency (count) of each word that occurs in the entire text. As we can see in the figure-1, words like tasty, cracker, yummy, good are the part of reviews that have a high review score. The above method was applied to the summary column of the data frame. The same was applied to the negative sentiment or low scores to see the frequently appearing words. As we can see from the figure2 we can see that words that are usually used in the negative context such as salty, horrible, unusable are the most frequent ones that are used in the customer food review. Post this, I analyzed the most frequently appearing bigrams. Bigrams which are two words coming together in a sentence. As we can see that most customers who thought that the product is of high quality and gave a high score have commented frequently as 'great product'. Same procedure was also applied for analyzing trigrams. The results appeared to be same with most of the comments to be love, best dog food, and great product.

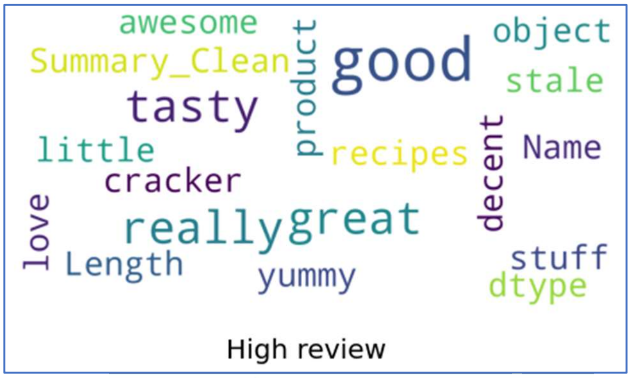


Figure 1: Word cloud of high review score

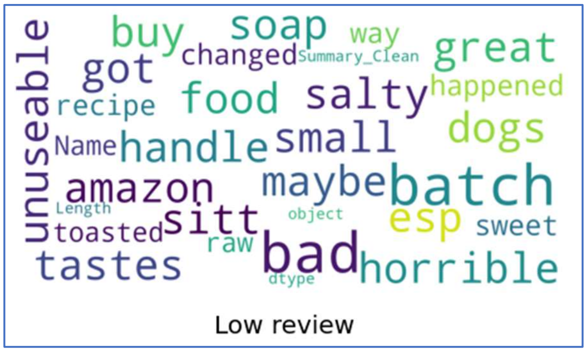


Figure 2: Word cloud of low review score

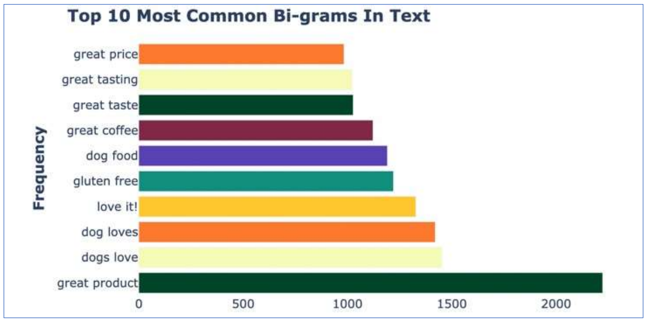
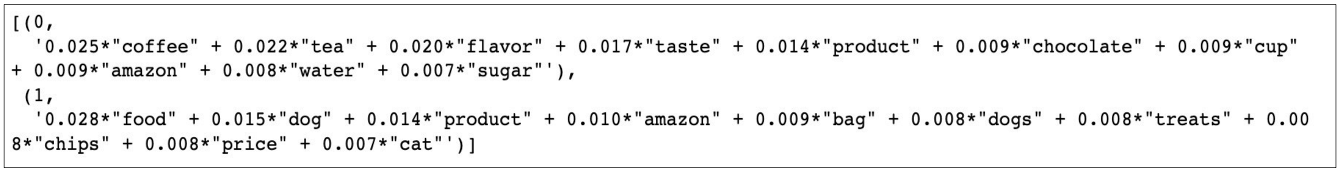


Figure 3: Most frequent bi-grams

Text Analysis

After exploring the data, in this step Latent Dirichlet Allocation (LDA) was applied. Latent Dirichlet Allocation (LDA) is used to classify text in a document to a certain topic. The topics that we want to extract from the data are also “hidden topics” and it is yet to be discovered. In the first step, Documentterm matrix was created. Text data is represented in the form of a matrix. The rows of the matrix represent the sentences from the data which needs to be analyzed and the columns of the matrix represent the word. The document term matrix is used for topic modelling. In the second we must specify the number of topics the algorithm has to pick up. The goal of Topic modeling to find a theme across reviews and Latent Dirichlet Allocation (LDA) is one of the techniques of topic modelling. It’s an unsupervised classification method. When we have a large text and don’t know where to start, then we can apply topic modeling and can see what kind of groups are emerging. LDA can be done in 3 methods using All the text data, only nouns from the text, and Only nouns and adjectives. In this assignment I have attempted to perform LDA with nouns only. I created two topic models with the 20 most relevant words that were most relevant.

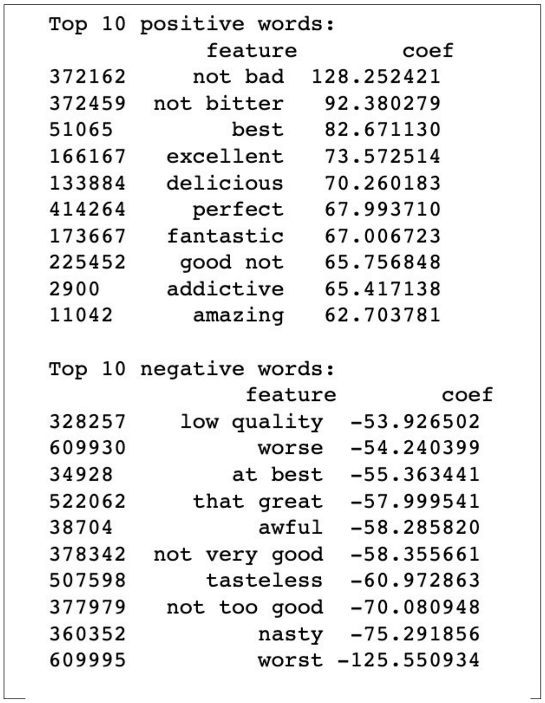


```
[(0,
  '0.025*"coffee" + 0.022*"tea" + 0.020*"flavor" + 0.017*"taste" + 0.014*"product" + 0.009*"chocolate" + 0.009*"cup"
+ 0.009*"amazon" + 0.008*"water" + 0.007*"sugar"'),
 (1,
  '0.028*"food" + 0.015*"dog" + 0.014*"product" + 0.010*"amazon" + 0.009*"bag" + 0.008*"dogs" + 0.008*"treats" + 0.00
8*"chips" + 0.008*"price" + 0.007*"cat"')]
```

Figure 4: Topic modelling with two topics

The above figure-4 shows the outputs of topic modelling with parameter of number topics being 2. We can see that the 10 words relevant in each topic 0 & 1. So, there are 20 words which are relevant and most frequent in the topics. Based on the first topic 0, the words suggests that the topic is more about consumer packed goods such as water, chocolate, cup, and sugar. Looking at the topic model 1, based on the relevant words repeated the topic is mostly on cat and dog treats.

To predict the summary text of the food review, logistic regression model was applied on the summary feature. The column was transformed using Count vectorizer for the train data to fit the logistic regression model. As we can see in the figure 5, these are the coefficients for a high review score in the food review summary. These top 10 words were ranked based on the value of their respective coefficient. If the review contains words like worst, nasty, low quality it is highly likely that the rating was either 0 or 1 and has negative feedback. The model obtained a weighted average accuracy of **94.90%,** where it was successful in predicting **97%** of the times as a negative review and **83%** of the times as a positive review.



```
Top 10 positive words:
            feature        coef
372162      not bad  128.252421
372459   not bitter   92.380279
51065          best   82.671130
166167    excellent   73.572514
133884    delicious   70.260183
414264      perfect   67.993710
173667    fantastic   67.006723
225452     good not   65.756848
2900      addictive   65.417138
11042       amazing   62.703781

Top 10 negative words:
                 feature        coef
328257       low quality  -53.926502
609930             worse  -54.240399
34928            at best  -55.363441
522062        that great  -57.999541
38704              awful  -58.285820
378342     not very good  -58.355661
507598         tasteless  -60.972863
377979      not too good  -70.080948
360352             nasty  -75.291856
609995             worst -125.550934
```

Figure 5: Words with coefficient weightage

Conclusion

Pursuing this assignment, we saw the steps and procedure followed in the Latent Dirichlet Allocation (LDA) a technique to analyze text data. We also obtained the most frequently appearing unigrams, bigrams and trigrams in the food review by customers on Amazon. Based on the statistical results, Amazon as a Business to Consumer services perspective can suggests and predict the likeliness of preferences of customers and give it as feedback to the respective producer. This would fuel in helping business to make necessary changes in their marketing and strategy planning. Using the results of the text analysis, it is beneficial for Amazon to understand user preferences, user choices and help in profiling individual customer preferences. This will help Amazon in better understanding of what products to recommend to what customers.

References

1. Amazon Fine Food Reviews. (2017, May 1). Kaggle. https://www.kaggle.com/datasets/snap/amazon-fine-foodreviews/code?datasetId=18&sortBy=voteCount
2. Gokce, E. (2021, December 15). Topic Modeling with NLP on Amazon Reviews - Towards Data Science. Medium. https://towardsdatascience.com/topic-modeling-with-nlp-on-amazon-reviewsan-application-of-latent-dirichlet-allocation-lda-ae42a4c8b369
3. S. (2021, August 26). Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-andlatent-dirichlet-allocation-lda-using-gensim-and-sklearn/

Appendix

The assignment was completed using Python on Jupyter notebook for data cleaning, data exploration & modelling. I have attached the files on canvas & names of the files attached are *Module-5Assignment.ipynb*.