**Module 5 Technique Practice**

Sharon Appoline Rosary

College of Professional Studies, Northeastern University

ALY6040 : Data Mining Applications

Prof. Justin Grosz

06/26/2022

**Introduction**

The objective of this assignment is to analyze Amazon fine food reviews available on Kaggle. We will employ mining techniques to analyze review texts using concepts of Natural Language Processing. The goal of this assignment is to identify a common theme across reviews and determine hidden patterns within the data. This analysis can help in predicting the preferences of customers and provide feedback to the respective supplier. Additionally, it can help Amazon understand customer preferences better and recommend products accordingly.

**Data Cleanup**

The dataset consists of 568,454 records with 10 features namely Product-ID, User Id, Helpfulness numerator & denominator, Review Score, Summary of the review, Complete text review, and time of the review. While inspecting the score column, the values ranged from 1 to 5. The lowest score was 1 and the highest score was 5. The reviews with scores 4 and 5 were categorized into ‘good review’ (443,777 reviews) whereas the reviews with scores 1 and 2 were considered as ‘bad review’(82,037 reviews). There were around 426,40 reviews with neutral score i.e., 3. Since neutral reviews do not assist in highlighting themes across reviews, they have been removed. This amount of review records can be hard to handle by computers. So, considering the number of records, a sample of data containing 10,000 was considered for the analysis.

The dataset also contains 2 variables Helpfulness numerator and Helpfulness denominator. Helpfulness numerator denotes the number of people who found the review helpful to them and Helpfulness denominator indicates the number of users whether they found the review useful or not. Helpfulness numerator should always be less than Helpfulness denominator. Two new variables %helpful and %upvote were introduced to analyze if the review scores were useful for customers. For analysis, the upvote variable was categorized into bins to show the number of customers who utilized the scores before purchasing. The text review column had comments mostly in Camel case format. So, it was converted to lower case. Then, it was checked for special characters and symbols. Stop words such as ‘like', 'im', 'know', 'just', 'dont', etc. were added to default English stop words which will be used while creating document-term matrix. Thus, the cleaned data can be used for further analysis.

**Exploratory Data Analysis**

Chart, table, treemap chart

Description automatically generatedAs part of exploratory data analysis, I wanted to visualize how many users find review scores helpful. Based on the helpfulness ratio calculated, the bins were categorized, and it was compared against the scores. From the heatmap, we can infer that reviews are skewed towards positive, and more than half of the reviews are with 0 votes. However, most of the users agree with score 5 reviews.

Figure1 : How many users find review scores useful

A picture containing text, bottle

Description automatically generatedThen, I visualized word clouds that were clustered based on positive and negative review score. A logistic regression model was fit to check the correlation of words with the good(scores with 4 & 5) and bad review scores(scores with 1 & 2). The top 20 positive words that were highly correlated to the good review score were populated in the word cloud. As we can see in the figure 2, words like perfect, delicious, yummy, great, excellent, etc. are the part of reviews that have a high review score

Figure2: Word cloud with positive sentiments

Logo

Description automatically generatedThe second word cloud contained words that were in reviews with bad review score. The 20 words that had least correlation with the review score(negative sentiment) were populated in the word cloud as shown in figure 3. Words such as poor, worse, undrinkable, tasteless, stale, bland, etc. were used in reviews that had a bad review score.

Figure3: Word cloud of negative sentiments

Text

Description automatically generatedText

Description automatically generatedThe coefficients for a high review score in the food review summary are given in this figure . These top 5 words were ranked based on the value of their respective coefficient. If the review contains words like great, delicious, best, perfect, or excellent, it is highly likely that the rating was either 4 or 5 and has positive feedback. Whereas, if the review contains words like awful, disappointed, terrible, or worst, it is highly likely that the rating was either 1 or 2 and has negative feedback. The model obtained an accuracy of 93.26%, where it was successful in predicting 83% of the times as a positive review.

Figure5: Top 5 negative sentiments

Figure4: Top 5 positive sentiments

**Text Analysis**

After data exploration, text analysis was performed using Latent Dirichlet Allocation (LDA) modeling. During LDA Topic modeling, we create many different topic groups. we will obtain different numbers of groups. Then, we examine and compare topic modelings, and decide which topic model makes more sense, most meaningful, have the clearest distinction within the model. Then, the group (model) that makes the most sense will be chosen among all topic groups.

The first step was to create a Document-term matrix. Text data is represented in the form of a matrix. The rows of the matrix represent the sentences in the data and the columns represent the words in the data. Then, we will choose the number of topics the algorithm has to pick up. LDA technique is mainly employed to find a theme or hidden topics across large texts. There are many ways to create topic modelings. i.e., we can create topic modeling by using entire text data, only nouns from the text, only nouns and adjectives, or a combination of nouns, adjectives, verbs, etc. In this analysis, I have chosen to perform LDA by using nouns from the reviews text. I performed LDA with 2,3 and 4 topic groups. These results will be compared and a decision on which topic model makes more sense will be arrived finally. After creating the three LDA models, the model with two topics was the most meaningful and relevant.

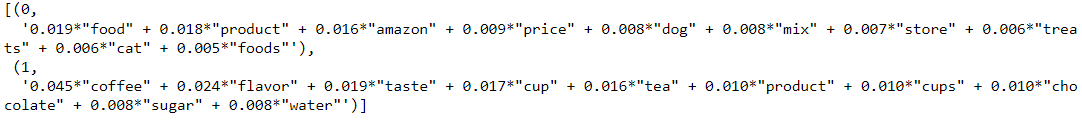


Figure6: Final Topic modelling with two topics

The above figure 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. Based on the first topic 0, the words suggests that the topic is more about cat and dog treats. Whereas in topic model 1, it contains words related to drinking such as water, tea, cup, water, and taste, etc. By considering all the steps of topic analysis with the LDA method, it can be concluded that good Amazon Reviews can be categorized into two main topics: beverages and pet items.

**Conclusion**

In this assignment, Latent Dirichlet Allocation (LDA) technique was employed to analyze Amazon fine food reviews. We obtained the frequently appearing positive sentiments and negative sentiments in the reviews. The negative sentiments in the reviews can be focused by Amazon and changes can be made to the products accordingly. This can enhance the likelihood of positive feedbacks. Based on the results obtained, the reviews were mostly focused on beverages and pet items. The outcomes of the text analysis can enable Amazon to better understand customer choices, likings, and assist in identifying the preferences of specific customers. As a result, Amazon will have a better grasp of which products to suggest to which 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/>