**ALY 6040: Data Mining Applications**

Technique Practice 5

College of professional studies, Northeastern

Rohit Meena

Prof. Justin Grosz

May 15, 2022

Abstract:

In this report, I will use the Amazon Food Review data set and conduct an analysis on it using the NLP method, as well as discuss the issues and problems with the data, as well as how I cleaned the data. With this data, I can also discuss interpretation and recommendations.

Report

There are 10 columns and 568454 rows in the Amazon food review data set. It's difficult to convey how people feel about the food because we have a large amount of data and can't read all of the customer reviews. We can use the NLP method to solve this issue. Natural language processing is a branch of computer science specifically, a branch of artificial intelligence concerning the ability of computers to understand text and spoken words in the same way that humans can. As a result, I can do text analysis using the Topic molding method of NLP.

Before I can do the text analysis, I need to clean the data, However, we must first perform feature extraction. So, first, we count the number of stop words using feature extraction. The number of punctuation characters, hashtag characters, numerical characters, and uppercase words. After that, we're finished with features that can only be obtained prior to data cleaning. We're all set to clean up the data. In any machine learning model, but especially in NLP, data cleaning is critical. The dataset is frequently a cluster of words that the computer does not understand without the cleaning process. The first step in the preprocessing process is to lowercase the reviews. This eliminates the need for multiple copies of the same words. If we ignore the transformation, "dog" and "dog" will be treated as different words when calculating the word count. The following method is Punctuation should be removed because it introduces noise into the data. For the time being, NLP methods do not have a useful way of analyzing punctuation. As a result, they were excluded from the text data. These characters were removed in this step: [! "# $ percent & '()\*+,-./: => [] |] , I can also remove stop words, URLs, Emojis, and Emoticons, among other things; you can see how I removed all of these in the code.

It's time to visualize the data after it's been cleaned. We can see the score distribution first. The code can be found in the file, as well as the graph in the appendix. As you can see in the bar chart, 5 out of 5 is the highest score, implying that approximately 250000 customers have given the food a 5 out of 5 rating. The lowest rating is 2 out of 5; approximately 20000 to 25000 customers have given the food a 2 out of 5 rating. A positive review, in my opinion, is when we receive a score of 3, 4, 5, out of 5, and a negative review is when we receive a score of 1, 2, out of 5. Now I converted the score values of 1 and 2 to 0 and 3, 4, and 5 to 1. Now, using a bar chart (as shown in the appendix), I compare the 0 and 1 to see which one has the highest value. It shows that approximately 300,000 customers give positive reviews and approximately 75,000 customers give negative reviews. I also made a bar chart, which you can see in the appendix, to show the number of reviews by time between 1999 and 2012; as the chart shows, the number of reviews increases over time. Now, as you can see in the appendix, I created a world cloud. A word cloud is a simple but effective visual representation object for text processing that displays the most frequently used words in larger, bolder letters and with different colors. Where you can see the more frequent word.

**Appendix**

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

***Fig 1 Distribution of score Fig 2 Distribution of Good Reviews ( 1 is good and 0 is bad )***

A picture containing graphical user interface

Description automatically generated

***Fig 3 Number of Reviews by times***

Text

Description automatically generated with low confidence

***Fig 4 World Cloud***

**Reference**

* Gokce, Enes. “Topic Modeling with NLP on Amazon Reviews: An Application of Latent Dirichlet Allocation (LDA).” *Medium*, towardsdatascience.com, 16 May 2022, <https://towardsdatascience.com/topic-modeling-with-nlp-on-amazon-reviews-an-application-of-latent-dirichlet-allocation-lda-ae42a4c8b369>.
* EnesGokceDS. “GitHub - EnesGokceDS/Amazon\_Reviews\_NLP\_Capstone\_Project: In This Repository, You Will Find All Process of NLP from the Scratch.” *GitHub*, github.com, [https://github.com/EnesGokceDS/Amazon\_Reviews\_NLP\_Capstone\_Project. Accessed 20 May 2022](file:////Users/rohitmeena/Downloads/EnesGokceDS.%20“GitHub%20-%20EnesGokceDS/Amazon_Reviews_NLP_Capstone_Project:%20In%20This%20Repository,%20You%20Will%20Find%20All%20Process%20of%20NLP%20from%20the%20Scratch.”%20GitHub,%20github.com,%20https:/github.com/EnesGokceDS/Amazon_Reviews_NLP_Capstone_Project.%20Accessed%2020%20May%202022.).