Final Project Proposal
04/26/2021

Background:

E-Commerce has been one of the most successful field nowadays, especially for fashion. Customers would have confidence in the E-commerce fashion companies if they are more likely to get positive reviews. The feedbacks can range from the quality of the outfits, the price of the clothes, how easily ones can navigate through the website to purchase a piece of apparel, how fast the items got delivered, and how identical the clothes look in-person compared to the advertised images. A Women's Clothing E-Commerce dataset was retrieved from *Kaggle.com* to address the important of reviews. The data set has 23,500 reviews in English with the highest number of reviews being 5-star ratings and lowest being 1-star ratings. The data set consist of the following variables:

Variable	Description	Coding
Clothing ID	Identification of a review from each piece of clothing	Integer categorical
Age	Reviewer's age	Positive integer
Title	Title of the review	Character string
Review Text	Review body	Character string
Rating	Product score granted by customer	Ordinal integer 1 = Worst to 5 = Best
Recommended IND	If the customer would recommend the product	Binary integer 0 = Not recommended 1 = Recommended
Positive Feedback Count	Number of other customers who found this review positive	Positive integer
Division Name	Product high level division	Categorical
Department Name	Product department name	Categorical
Class Name	Product class name	Categorical

Methods:

This analysis will, first, attempt to identify if a review is a *good* or a *bad* review. With the categories (good/bad) of the reviews, it is accessible to characterize the most frequent words that are associated with each type of reviews. In addition, to find the perfect label (good/bad) for each review, an automated terminology extraction process will be applied to the data. This step will allow us to extract the most common vocabularies that are used in positive/negative reviews.

Following the most common terminology in a review, some classification models will also be used to identify the original integer rating of the clothing. The rating variable will be treated as a categorical variable with 1 being the worst rating and 5 being the best rating. The model development process will attempt to generate a few types of model including binomial and multinomial Naive Bayes and SGD Classifier using a split train set of .75 of the whole data set. To ensure a robust model, hyperparameter search will be applied to find the best parameters for the model and error assessment will be constructed to determine the overfitting/underfitting of the

models. A hold-out test set of .25 of the data set will be used to validate the models. The ideal output would be high accuracy and F1 scores.

Deliverables:

Most E-commerce clothing companies desire to know what their customers think about all the items in the shop. The owners can order more items that have 5-star rating and good reviews. In addition, with the rating, owners can understand customers more, in terms of trending, quality, and styles. However, reading all 23,500 feedbacks might be too much for smaller companies, who do not have enough employees. Therefore, a break-down report of all the reviews would help to save the owners and employees' time. In addition, having a few keywords for each item would also help the customers to shop faster and more confidently. For example, if a customer clicks on a pair of pants and see top three most occurred key words from the reviews as "good", "comfortable", and "true to size", they would purchase the items more confidently without reading long paragraph of reviews.