Capstone #1: NLP with Amazon Reviews

Problem statement: Which words distinguish low versus high ratings in customer reviews of Amazon Alexa products? From a linguistic perspective, can we capture key differences in positive versus negative reviews? Finally, can we use a supervised machine learning approach to classify new, unseen reviews as positive or negative?

This is a useful question to ask, as automated processing of reviews as positive versus negative, from a binary perspective, can help a company tailor its customer service attention to those characterized as negative (i.e. for those that might warrant a follow-up). Further, by looking at the words most often used in a positive versus a negative review, these words can help to inform a business strategy as to which aspects of a product warrant improvement versus which aspects of a product can be boasted as being of a higher quality.

Such a problem would be of interest to companies who receive feedback and reviews from customers in a written format.

The data: The data were obtained from Kaggle (link to data here), publicly available at no cost.

To clean the data, first I omitted rows that contained no text review (despite the fact that the dataset actually only contained observations containing a text review). Then, I explored the data using .info() and .describe() methods, before grouping the data into low versus high reviews. While I also looked at the data, as a whole, because the data had such high positive skew, it was informative to weed out the positive reviews in order to actually get a sense of what the small number of negative reviews were saying. To facilitate looking at only the most helpful of words, I used stopwords to filter out proper nouns (i.e. ‘Alexa’ and ‘Amazon’), which were the most frequent but least helpful when considering things like sentiment and product components correlating with positive or negative ratings. Finally, to get a better qualitative sense of the data, I used a variety of wordclouds to look at the data as a whole and to look at specific subsets of the data.

Findings: Positive reviews, generally, were identified through the use of highly emotive words (i.e. ‘great’, ‘love’), while negative reviews more frequently included words about concepts that the product is related to (i.e. ‘sound’, ‘time’). A Naive Bayes approach was used for the classification step, and after tuning some of the model parameters, the performance of the model was improved over the model using default parameter settings.