Capstone 2: Sentiment Prediction from Amazon product review data

Business Problem:

Our client would like to improve their internal recommendation system on their ecommerce site. Using customer text data about our products, we will build, evaluate and compare models to estimate the probability that a given text review can be classified as “positive”, “neutral”, or “negative”.

While some of this data is accompanied by some sort of numeric data (stars, like, etc), often it is not. Our goal is to build a text classifier using Amazon product review data which can be used to analyze customer sentiment which does not have accompanying numeric data.

Stakeholders:

Our fictional client and their technical team.

Data:

The dataset is quite large, in total over 233 million reviews. We are going to select 3 subcategories to analyze:
Electronics
Clothing and Jewelry
Home and Kitchen

https://nijianmo.github.io/amazon/index.html

Data Science Methods:

We will model our business problem as a supervised classification problem. Using the star number accompanying our text reviews, we will label reviews as “positive”, “neutral”, or “negative”.

All the models built will be evaluated and compared according to appropriate performance metrics that align with the business problems. Additionally, we will conduct interpretability analyses to try to find connections between extracted features and the probabilities associated with the classes being considered.

Possible Constraints:
Our main assumption is that the number of stars represents the sentiment expressed by the text of a customer review. The customer’s experience is a combination of things such as the delivery

experience (condition upon arrival, length of delivery), their prior perception of the company and the actual product itself. There is also the issue of possible fake reviews which may have noisier text than reviews from actual buyers.

Desired Outcomes:

With the developed models, our business can now use text reviews which do not have accompanying numeric data to decide which new products to recommend to customers. This will reduce returns and increase customer satisfaction by providing better recommendations.

Unfortunately, sentiment analysis can never reach perfect predictions. Even two human reviewers may disagree on whether a text is positive or negative.