Data Science Project Proposal

Greg Pollock

The main question I’m trying to answer is if it’s possible to predict a positive or negative customer review given the text of the review. Specifically, can I create an accurate classification model that correctly differentiates between 1, 2, 3, 4, and 5 star reviews based on the text of the review.

I’m getting my data from datasets available on Kaggle. I’m taking a dataset from amazon food reviews and a dataset from Yelp reviews.

https://www.kaggle.com/snap/amazon-fine-food-reviews?select=Reviews.csv

https://www.kaggle.com/omkarsabnis/yelp-reviews-dataset

My target will be the star rating. I’m considering simplifying the target to be a positive / negative review which would mean grouping 1-2 star reviews and 4-5 star reviews and excluding neutral 3 star reviews. This might allow for greater accuracy, but I will decide once I am able to take a look at the data and the models.

The only feature I will have is the review text.

I originally wanted to access the full yelp academic dataset, which is a 9GB json file, but my little Macbook Air couldn’t handle it, so I opted for a smaller yelp dataset available on Kaggle. The amazon dataset has 568,454 reviews and the yelp dataset I’m using has 10,000 reviews, so in total 578,454 reviews will be seen by my models.

This data will help me answer my question because the data includes real product or service reviews and how they scored the product or service, so by looking at the words used by real people I should be able to tell over time the common words or phrases indicating positive or negative reviews and perhaps even the magnitude of positivity or negativity.