Background information (1-2 paragraphs)

Provided by the Dataverse project at Harvard, the Amazon reviews dataset centralizes information for several select consumer goods products. Information includes product rating, number of "helpful" votes, and the review text itself.

Problem statement (1-2 sentences)

Companies frequently want to predict customer satisfaction based on written feedback, but with the large scale of data collection operations that modern-day software facilitates. Thus, it's imperative to design an accurate way to analyze this data en-masse.

Hypothesis (Optional, but could make your write-up more cohesive)

Customer satisfaction can be reliably measured by feedback keywords.

Methods

1. Data cleaning. To ensure the data was in the correct format and did not contain superfluous or outlier information
    a. Involved data visualization
2. Research. To determine the ideal method for estimating the distribution of entries of a categorical variable using qualitative data (text), I ran through quite a few articles and tutorials. I ended up using two approaches to analyze this data.
    a. Linear regression. Under this method, my goal was to estimate the average star number (including fractions) to yield a regression-type analysis. This was my first approach, as I had learned similar techniques involving keywords analysis and linear regression in a data science course.
    b. Text classification model. Under this method, my goal was to predict one of five star categories to yield an ordinal variable analysis. Given that I wanted to predict an ordinal variable using text, this model was particularly suited to the task.
3. Implementation.
    a. Text pre-processing
    b. Model training
    c. Model testing
4. Note: Information about months
    a. More feedback was left in the months around winter and spring, so future research could also explore the relationship between month of review (an approximation for month of purchase) and customer satisfaction. The interest in exploring this question explains its inclusion in several parts of the code.

Results and Discussion

1. Linear regression
    a. Dependent variable: Keywords
    b. The model outputted an usual R^2 value—below -2—which raised a warning flag.

    c. The prediction error (RMSE) was ~2 stars. Given five total stars, this method appears to be a valid way to estimate the range of stars given a particular review. However, it cannot be relied upon to estimate the number of stars used.
2. Text classification
    a. Partially implemented. However, version errors has led to significant delays in debugging.
3. Conclusion: Keywords, without considering other aspects of meaning in a sentence, may still remain a reliable way to estimate customer satisfaction.

Resources

[1] Regression with Text Input Using BERT and Transformers | by La Javaness R&D | Medium
[2] https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12
[3] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UUB774