Data Memo - Sentiment Analysis

https://amazon-reviews-2023.github.io/ 1996-2023 amazon reviews with multiple categories

https://github.com/scrapehero-code/amazon-review-api How to create an Amazon review API

An overview of your dataset

1. What does it include?
    - `Toys and Games` is a dataset that contains Amazon review data spanning from 1996-2023 in their toys and games category. This will be used if we are not able to use an API. The dataset includes:
        - **Review Text**: Written feedback from customers.
        - **Star Ratings**: Numerical ratings of products (1–5 scale).
        - **Helpful Votes**: Number of votes indicating how helpful a review is.
        - **Verified Purchase**: Boolean indicating whether the reviewer purchased the product.
        - **Timestamp**: When the review was written.
        - **Product ID (ASIN)**: A unique identifier for each product
2. Where and how will you be obtaining it? Include the link and source. If you plan on creating a web scraper or using an API, describe your plans.
    - Our plan is to attempt to use an Amazon API to pull rating/review data directly.
    - If the API does not workout, we will use the `Toys and Games` Amazon dataset. The dataset is being downloaded from a github repository.
3. About how many observations? How many predictors?

- Observations: The amount of observations is unknown as of this time from the API. `Toys and Games` is a very large dataset, with hundreds of thousands of reviews.
- Predictors:
  - Text (Review content)
  - Rating (Star rating)
  - Verified Purchase
  - Helpful Votes
  - Timestamp
  - Product ID

4. What types of variables will you be working with?
   - There will be multiple types of variables we will be working with including:
     - **String** (str)
     - **Float**
     - **Boolean** (bool)
     - **Integer** (Int)
5. Is there any missing data? About how much? Do you have an idea for how to handle it?

   There is not any missing data in the dataset.

An overview of your research question(s)

1. What variable(s) are you interested in predicting? What question(s) are you interested in answering?

   We are interested in exploring Natural Language Processing programs and developing a better understanding of how they work and applying them to a business-related task.
2. Name your response/outcome variable(s) and briefly describe it/them.
   - Our response will be the sentiment category of reviews based on customer, including:
     - Positive
     - Neutral
     - Negative
3. Which predictors do you think will be especially useful?

- Review Text: Most critical for analyzing sentiment and themes.
- Star Ratings: Useful for aligning reviews with sentiment categories.
- Helpful Votes: Indicates which reviews provide valuable feedback.
- Verified Purchase: Helps distinguish credible reviews from unverified ones.

4. Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

   The goal of our project is to develop a better understanding of Natural Language Processing in Python and create a descriptive program that takes customer reviews as input and outputs the emotional sentiment of the review as Positive, Neutral, or Negative.

Your proposed project timeline

We will start by loading the data from the github data set this week and doing EDA on our reviews to explore the distributions of the extra data beyond the review strings such as star ratings. Then we will begin developing the program to take the review string and process it and then return the emotional sentiment of the review. We are unsure how long this will take, but it is the main focus of our project so we intend to spend the most time on it. We plan to have the program functioning 1-2 weeks before presenting so we can also test our program on fresh data pulled from newly posted Amazon reviews, both to ensure functionality and show we can use an API to pull data as well. We will finish the written portion of the project in the final 2 weeks and then present our project during the final exam time.

Any questions or concerns

1. Are there any problems or difficult aspects of the project you anticipate?
2. Any specific questions you have for me/the instructional team?