5.0

SPRINT 2

BOOK RATINGS

SABRINA RUIZ

OVERVIEW & PROBLEM STATEMENT

1. many kids do not like reading

2. explore text data analysis (NLP fun!)

“Can we use customer book review comments to predict the book’s rating score?”

“Can we use customer book review comments to return n book recommendations?”

IMPACT

1. Time Savings and Improved Reading Experience:
2. Enhanced book recommendations?
3. Universal Design and Educational Application

PROPOSED VISION

1. EDA

- researched null data
- duplicates
- 2 data sets > 1
- re-addressed categorical data

2. NLP

- custom tokenizer
  - remove common words
- sentiment classifying (unbalanced data)
- vectorization - 2 versions
- how many features?

3. MODEL

- predictor
  - grid search
- recommender
  - classification

“COMBINED DATA”

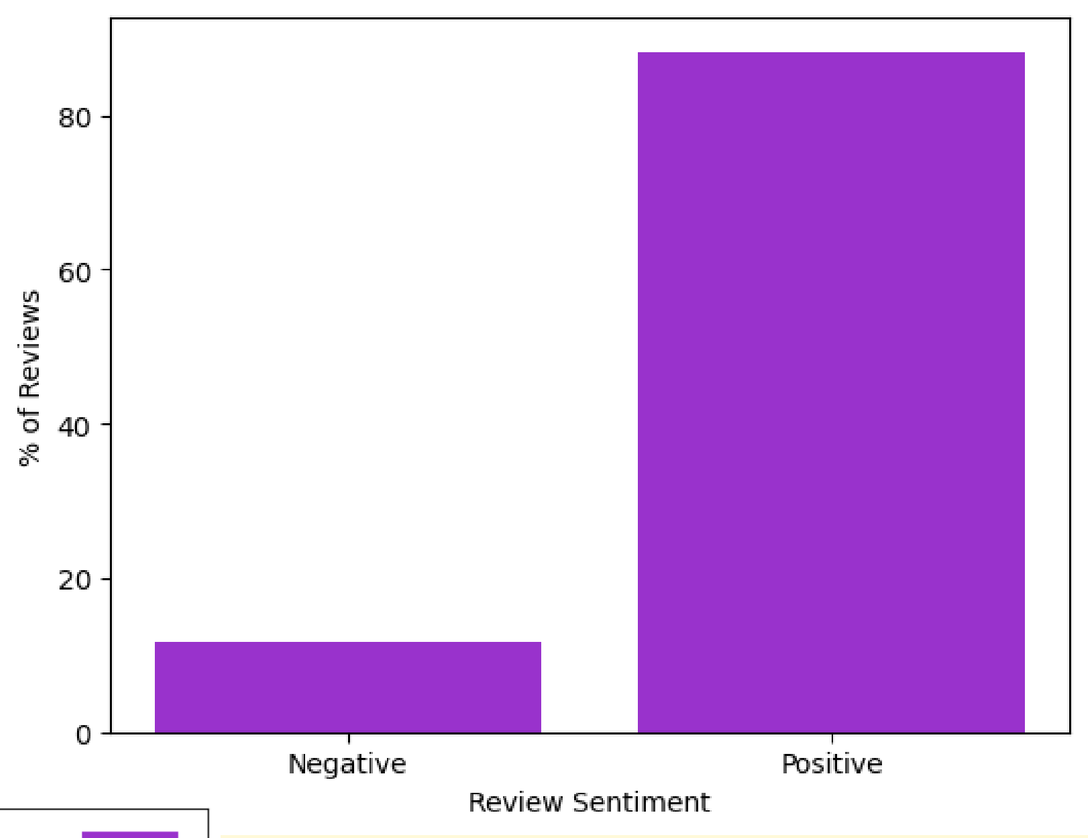
- 3 mil. reviews × 10 features
- reviewer id , review text, book info (author, title, year, genre)
- target feature = ‘review’ score

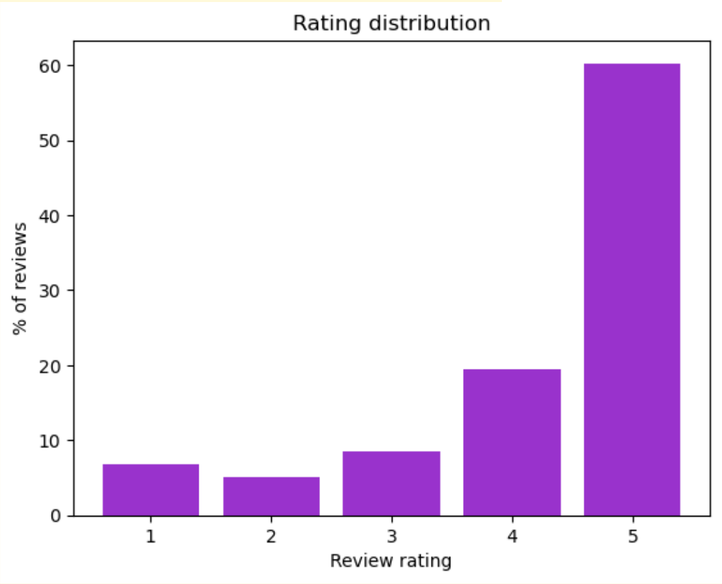
initial logreg model w/out text features:

65% accuracy

initial logreg model w/ text features:

0% accuracy





- **Sentiment Distribution**
  - 88% positive
  - 12% negative (my oversight!)

NEXT STEPS

1. ADDRESS DATA INBALANCE

perhaps downsize reviews where score > 3

positive/ negative cut off change (2.5? 2.7?)

2. GET TO MODELING

decide between a “recommendation” system or with the rating predictor as before

- Unsupervised -cluster --> recommender
- simple text similarity

3. SENTIMENT CLASSIFYING TWEAKS

custom tokenizer - fix stemming, common words, n-grams