Sentiment Analysis of Goodreads Reviews

DETERMINING THE PREDICTIVE POWER OF USER REVIEWS

Background

- Goodreads.com
  - Established 2007
  - 65 million members
  - 2 billion books
  - 68 million reviews

  https://www.goodreads.com/about/us
- How accurately can text sentiment analysis of user reviews predict final user score?
- Why? What is the business value of predicting the score?

Overview of Process

- Data Collection & Preparation
- Exploratory Data Analysis
- Model Selection
- Conclusion

Data Collection

- Book data retrieved from Goodreads website
- API allows for limited access to data from site
  - Web Scraping necessary to obtain book data & review text
  - BeautifulSoup module for web scraping

Data Collection

Book Data Points

Data	Description
ID	The Goodreads book ID
Title	Book title
Original title	The Book's original title (e.g. native language title)
Author	Name of book author (first author if multiple)
Published	Publication date
Language	Language of book edition
Avg. Rating	The average user rating based on ratings of 1-5
Rating Count	The number of user ratings
Review Count	The number of user reviews
Genre 1	The #1 book genre based on user votes
Genre 2	The #2 book genre based on user votes
Genre 3	The #3 book genre based on user votes
To read	The count of users who have this book on their "to read" shelf.
Currently reading	The count of users who have this book on their "currently reading shelf.
Favorites	The count of users who have this book on their "favorites" shelves.

- Books in Goodreads DB scanned randomly
- Table lists data points collected for each book scanned
- Features like Genre and Shelves not used in final analysis

Data Collection

Criteria For Inclusion in Analysis

- Book available in English
- Book has at least 40 reviews written
- No duplicates (e.g. books with multiple titles or editions are only included once)

Of 95,000 titles scanned, 7,615 (12.5%) met criteria

Data Collection

Review Data Points

Data point	Description
Review ID	The Goodreads review ID
Book ID	The Book ID associated with this review
Review Date	The date the review was submitted
Rating	The score from 1-5 that the reviewer gave the book. Not all reviews include a rating.
Review Text	The raw review text

- Reviews scraped for each of the 7,615 titles which matched selection criteria
- Between 40 and 300 reviews scraped for each title.
  - 40 was minimum for inclusion
  - 300 was maximum that could be scraped
- Total of 1,366,205 reviews scraped

Data Preparation

Review text processing:

- Remove punctuation (except ‘ and -)
- Identify individual words as separated by spaces
- Reviews with fewer than 30 words excluded

Review Scoring

- Four distinct sentiment lexicons used
- Each review word matched with sentiment lexicons and given a score when matched
- Aggregates of word scores for each review used as features in analysis

Lexicon	Words	Description
AFINN	2,477	Sentiment scores ranging from -5 to +5
Bing Liu	6,787	Polarity scores. 0 for negative, 1 for positive
Harvard Inquirer	3,629	Polarity scores. 0 for negative, 1 for positive
MPQA	6,901	Polarity scores. 0 for negative, 1 for positive

Feature Selection

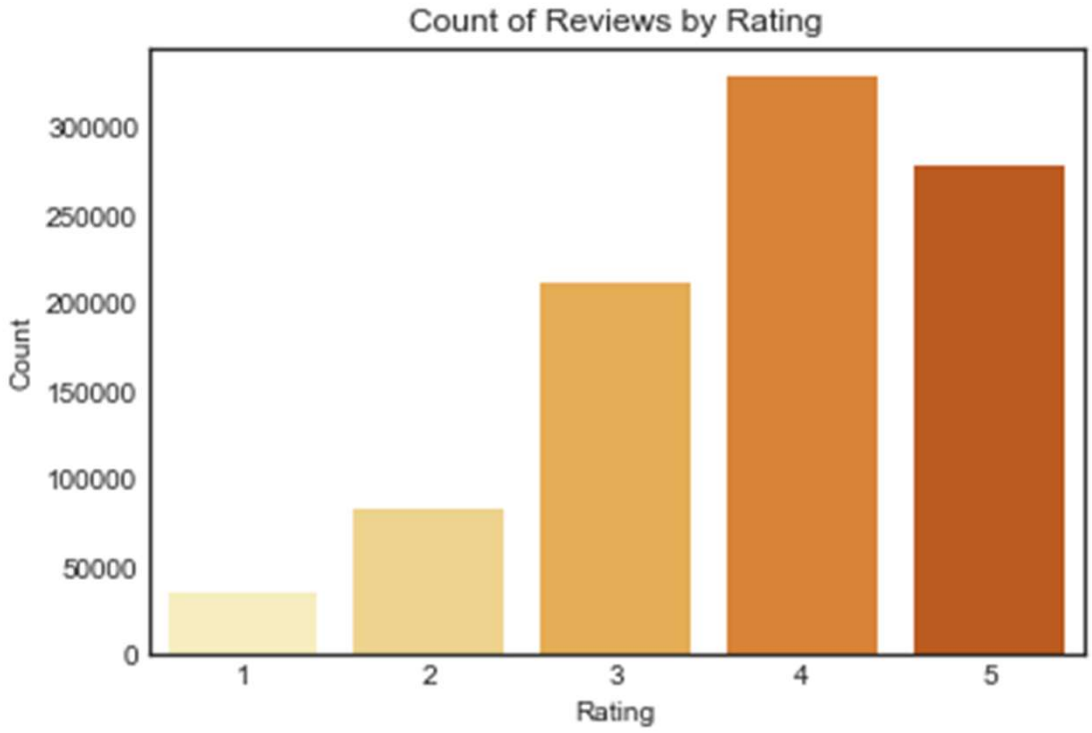
Features	
Review ID	Total MPQA count
Rating	Total Inquirer count
Word Count	Positive AFINN ratio
AFINN Mean	Positive Bing ratio
Bing Mean	Positive MPQA ratio
MPQA Mean	Positive Inquirer ratio
Inquirer Mean	Negative AFINN ratio
AFINN Median	Negative Bing ratio
Bing Median	Negative MPQA ratio
MPQA Median	Negative Inquirer ratio
Inquirer Median	Positive AFINN density
AFINN Sum	Positive Bing density
Bing Sum	Positive MPQA density
MPQA Sum	Positive Inquirer density
Inquirer Sum	Negative AFINN density
Positive AFINN Count	Negative Bing density
Positive Bing Count	Negative MPQA density
Positive MPQA Count	Negative Inquirer density
Positive Inquirer Count	AFINN Words Ratio
Negative AFINN Count	Bing Words Ratio
Negative Bing Count	MPQA Words Ratio
Negative MPQA Count	Inquirer Words Ratio
Negative Inquirer Count	Caps Word Count
Total AFINN count	Exclamation Count
Total Bing count	All Caps Density

- Aggregates of sentiment lexicon scores
- Calculated columns based on aggregates
- Additional features (all caps words, count of exclamation points)
- Mention something about feature choices
  - Note about feature importance
- 50 features total
- Reviews with outlier values (> 3 SD) for count of positive/negative dropped
- Reviews without a rating dropped

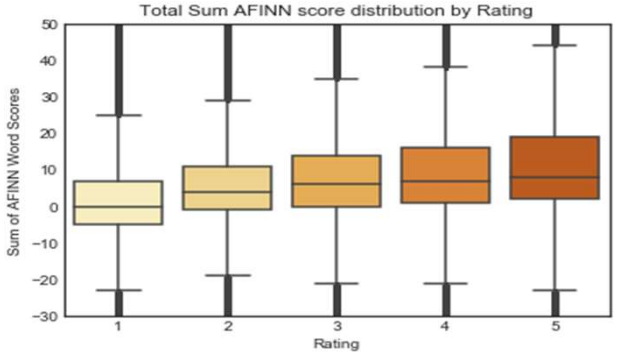
Feature Selection

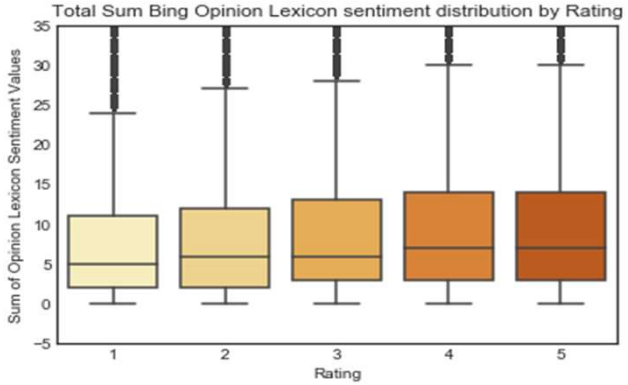
- Top Features – As scored by Random Forest Classifier

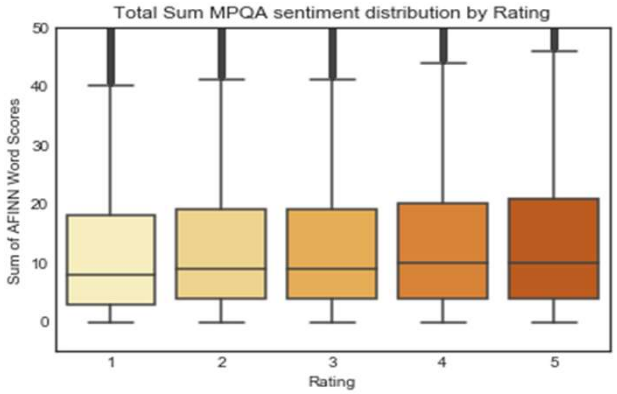
Exploratory Data Analysis – Rating Distribution

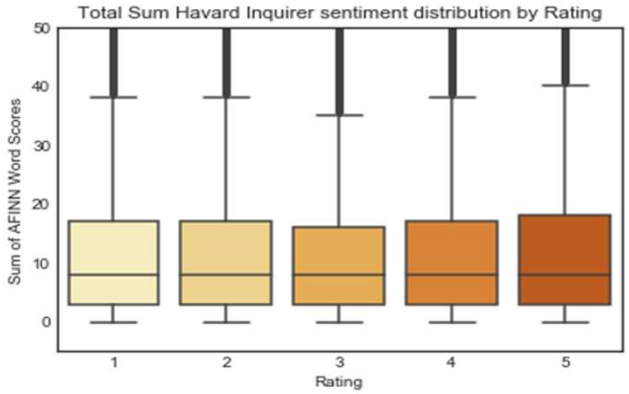


Exploratory Data Analysis – Score Sum Distributions

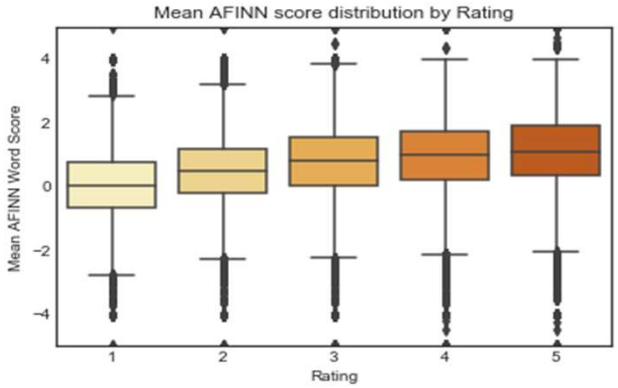


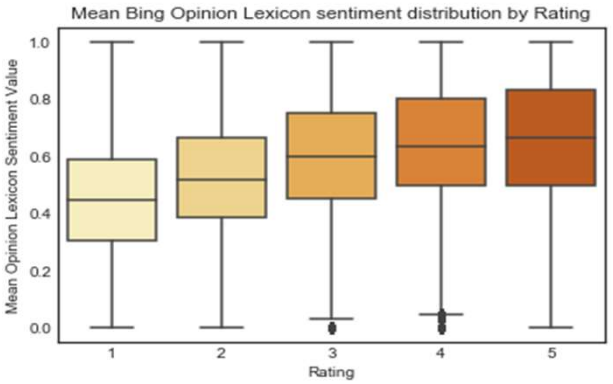


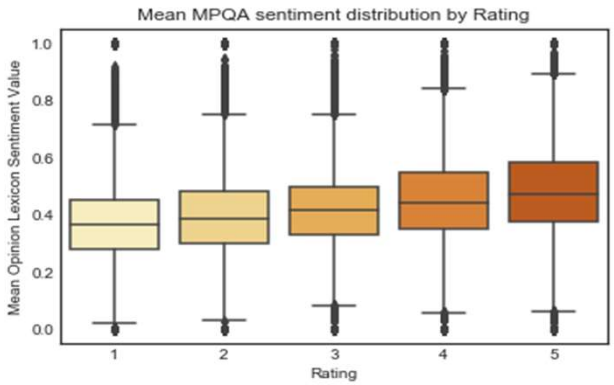


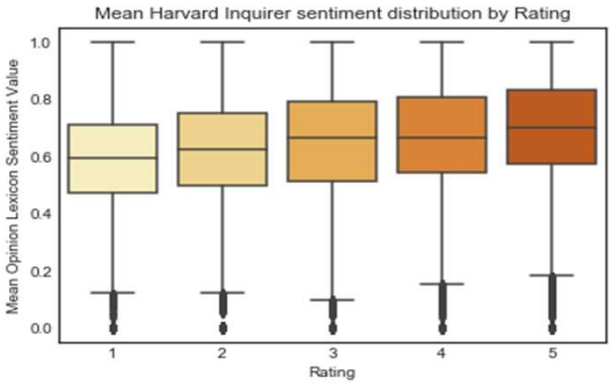


Exploratory Data Analysis – Score Mean Distributions

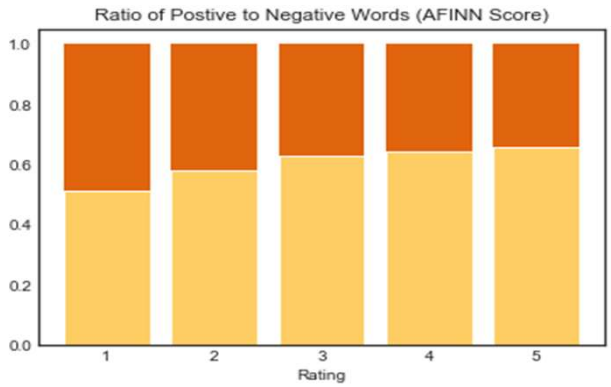


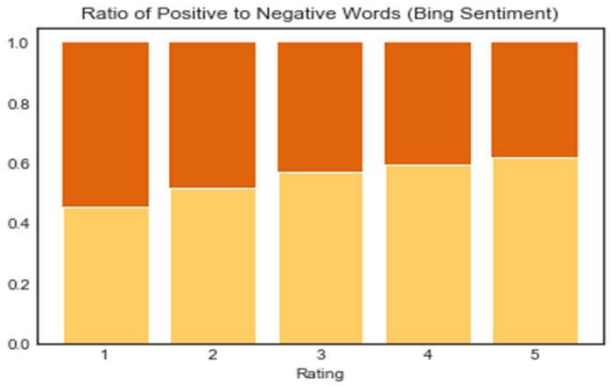


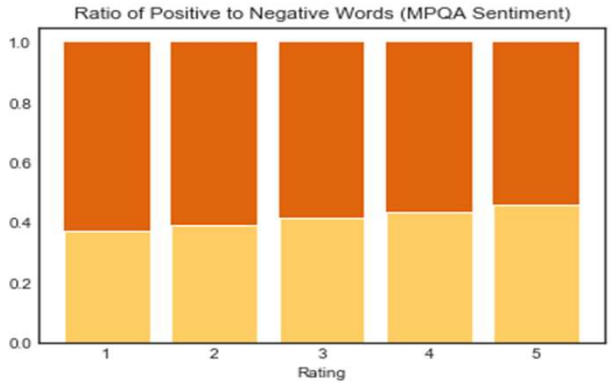


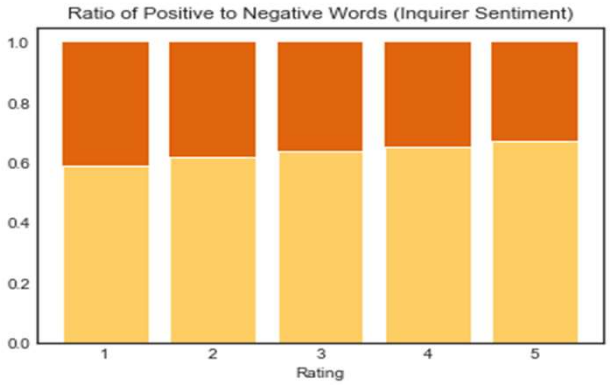


Exploratory Data Analysis – Positive / Negative Ratio









Exploratory Data Analysis - Summary

- Review scores unevenly distributed
  - Heavily skewed toward higher scores
  - Might indicate readers who didn't enjoy book less likely to leave review. Selection Bias
- Examining the distribution of features reveals that there is an upward trend in metrics (e.g. sum of scores, mean of scores, etc.).
  - Higher rated reviews tend to have higher sum, mean, medians, and other aggregates of the word scores
  - This pattern is observed across each lexicon
  - While trend is clear, the data is very spread out in each category, indicative of a wide degree of variance.
- Other metrics such as ratio of positive scored to negative scored words in review also tend to increase as review score increases
  - This pattern is also observed across all four lexicons

- Conclusion – Clear visible trend between rating value and overall sentiment score, but wide variance may mean the association is weak.

Classification Model

Five Category Classification – Summary of Results

Classifier	Score	Avg. Precision	Avg. Recall
K-Nearest Neighbors	0.27	0.30	0.29
Decision Trees	0.27	0.26	0.37
Naïve Bayes	0.28	0.36	0.28
MLP	0.37	0.37	0.38
Random Forest	0.38	0.38	0.38

- Random Grid cross-validation used to optimize hyperparameters
- MLP and Random Forest classifiers resulted in best accuracy.
- Use of scalars to standardize features did not produce meaningful improvements in the results.

Classification Model

Two Category Classification – Summary of Results

Rating	Score	Precision – 0	Recall – 0	Precision – 1	Recall – 1
5	0.69	0.72	0.91	0.45	0.17
4	0.64	0.65	1.00	0.50	0.00
3	0.77	0.77	1.00	0.00	0.00
2	0.91	0.91	1.00	0.00	0.00
1	0.96	0.96	1.00	0.00	0.00

- Converted classification into five separate two-category classification problems to see if model could predict individual ratings.
- Used MLP classification model
- Scores and precision are high, but only because model guessed 0 for nearly all results.
  - Notice Recall for scores 2-5 are zero, indicating that the model did not predict that review was in given score category a single time.

Classification Model

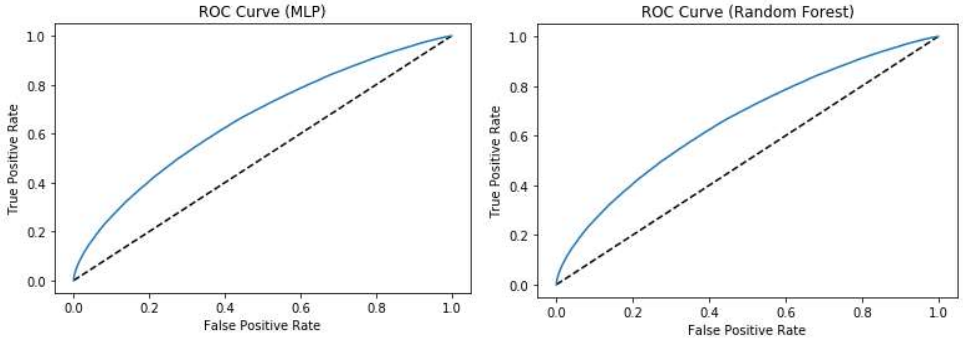
Two Category Classification – Summary of Results

Classifier	Score	Precision – Neg.	Recall – Neg.	Precision – Pos.	Recall – Pos.
MLP	0.66	0.55	0.22	0.68	0.90
Random Forest	0.62	0.66	0.78	0.53	0.37

- Reformatted classification into two categories –
  - Positive (Rating >= 4)
  - Negative (Rating <= 3)
- Used MLP and Random Forest which had best accuracy in 5-category classification
- Reducing classification to two categories significantly improved model accuracy, though still low

Classification Model

Two Category Classification – ROC Curves



- ROC Curves visualize model accuracy
- Curve close to 45 degree line indicate low accuracy

Conclusion

- Sentiment Analysis can be used to provide a view into the attitude of the reader, but falls short of being able to accurately predict the final score.
- Trends exist showing a relationship between higher rating and higher sentiment score values, however this association is not strong enough to produce an accurate model.
- MLP and Random Forest classifiers produced the most accurate predictions
- Additions that may potentially improve model accuracy
  - Additional lexicons
  - Analysis of text beyond positive or negative word sentiment, e.g. emotional connotations associated with words
  - Inclusion of additional review features not related to sentiment