Big Data Analytics
BYGB-7990-001

GROUP 5

CASE STUDY - AMAZON BOOK REVIEW

Project Milestone III

CeCe Liu, Charles Bera, Imaobong Undieh, Nisera Vaughan, Sichen Tong, & Yijie Xu

TABLE OF CONTENTS

TABLE OF CONTENTS

01

BUSINESS PROBLEM & RESEARCH QUESTIONS

01

BUSINESS PROBLEM

Background

- Amazon Book Review utilizes the Amazon Books Editors to curate book recommendations
- Reviews and recommendations tend to be biased

Solution

- Build a predictive model to analyze book reviews by utilizing real customer reviews

Benefits of Research

- Unbiased & Diversified Book Recommendations
- Book Pricing
- Inventory Control

01

RESEARCH QUESTIONS

- Are measures of book review sentiments strongly correlated with book ratings measured on a scale of 1-5?
- Is there an association between a book review's sentiment and its helpfulness rating?
- How can book price factor into either an overall positive or negative review sentiment that can ultimately affect its rating?

02

THE DATASET—
AMAZON BOOK REVIEWS

02

AMAZON BOOK REVIEWS

SOURCE

Kaggle

FILE SIZE

2.86GB

TOTAL VALUES

3 Million

DATASET DESCRIPTION

02

VARIABLE	TYPE	DESCRIPTION	EXAMPLE
Id	string	The ID of the book	1882931173
Title	string	Book title	Dramatica for Screenwriters
Price	float	The price of the book	17.46
User_Id	string	ID of user that rated the book	A2KSXSRTMD3ZJ4
profileName	string	Name of user who rated the book	Johnny Appleseed

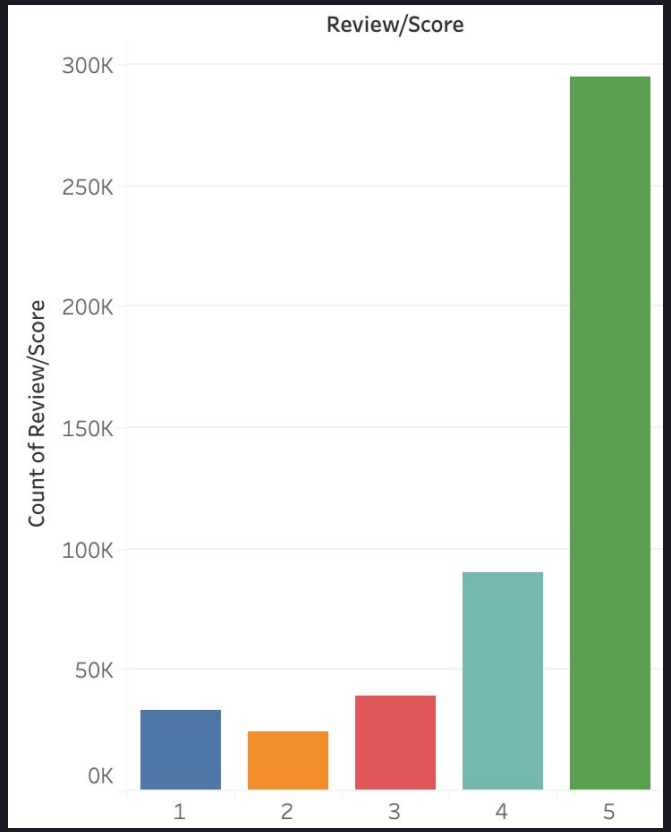
DATASET DESCRIPTION CONT.

02

VARIABLE	TYPE	DESCRIPTION	EXAMPLE
review/ helpfulness	float	Helpfulness rating of the review	5.0
review/score	float	Rating from 0 to 5 for the book	2.0
review/time	float	Time the review was given	1335108095.98
review/ summary	string	Summary of the text review	“Exactly what I needed.”
review/text	string	The full text of the review	“This book answered all my questions about getting into law school in a plain and straightforward way...”

Preview of Rating Distribution

02

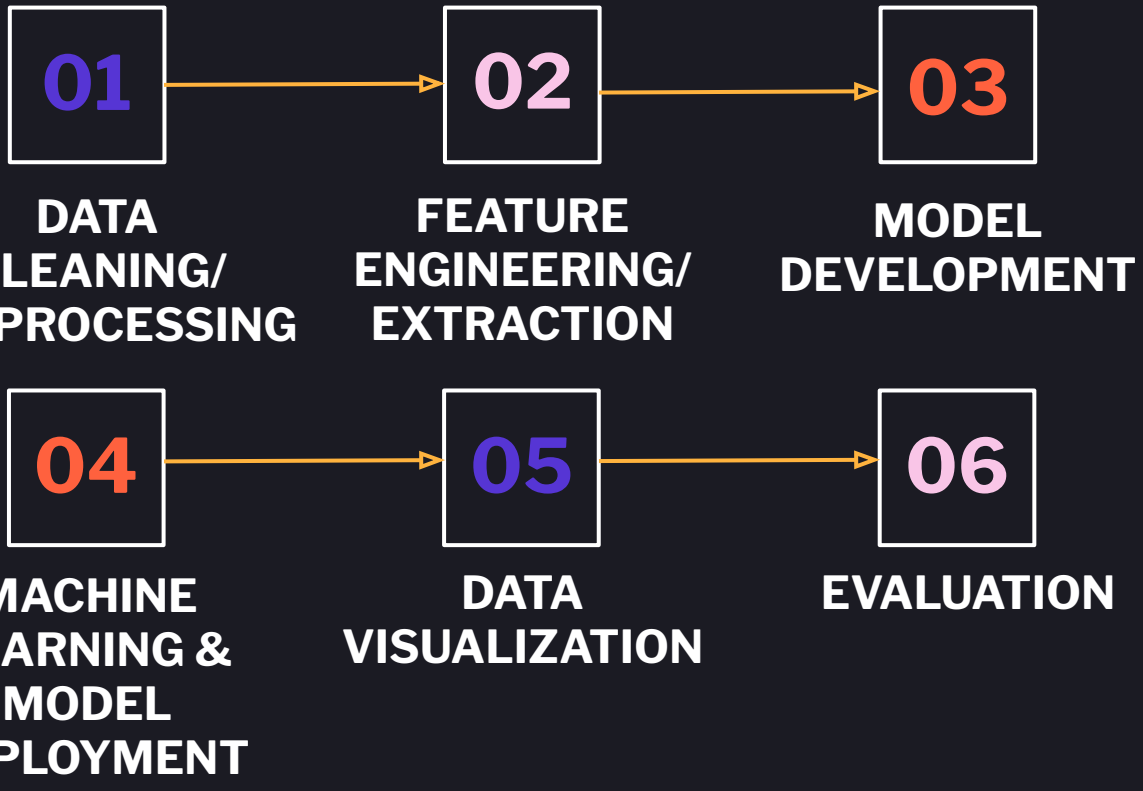


- Histogram describing the distribution of rating score.
- Ratings are heavily skewed to the right

03

SYSTEM DESIGN

03 SYSTEM DESIGN



04

DATA PREPROCESSING

04

DATA PREPROCESSING

Unnecessary Columns and NaN

- ❑ Dropped User_id, profileName, review/time
- ❑ Dropped NaN

Text Cleaning

- ❑ Removed punctuations and non-alphabets
- ❑ Lower case

Fractions Conversions

- ❑ “review/helpfulness” column
- ❑ Fractions into decimals
- ❑ Handling of 0/0

04 DATA PREPROCESSING

Tokenizing

- Split paragraphs and sentences into smaller units

Stemming

- Stems or removes last few characters from a word to get to base form or root

Lemmatization

- Considers the context and converts the word to its meaningful base form or root word

05

MODEL DEVELOPMENT

05

MODEL DEVELOPMENT -

Vectorizing

- Binary Vectorizer
- Word2Vec Vectorizer
- TD-IF Vectorizer

Splitting the Data

- 70% Training
- 30% Testing

Model Selections

- Multinomial Naive Bayes
- Decision Tree
- Random Forest
- Logistic Regression

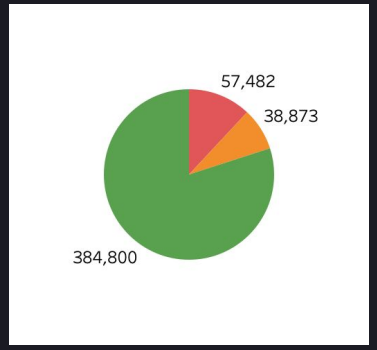
05

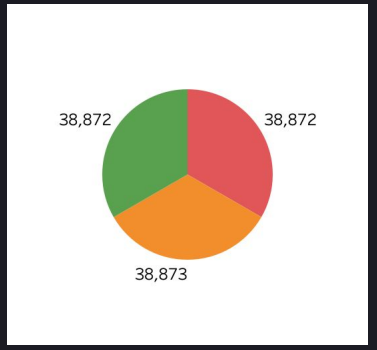
MODEL DEVELOPMENT -

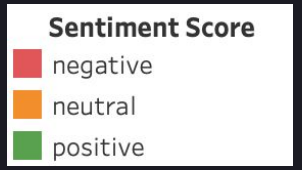
Feature Engineering

- sentiment/score
  - Negative **rating ≤ 2**
  - Neutral **rating = 3**
  - Positive **rating ≥ 4**

Handling Data Bias







- Reduced dimensionality of positive and negative sentiment/score data to reduce bias

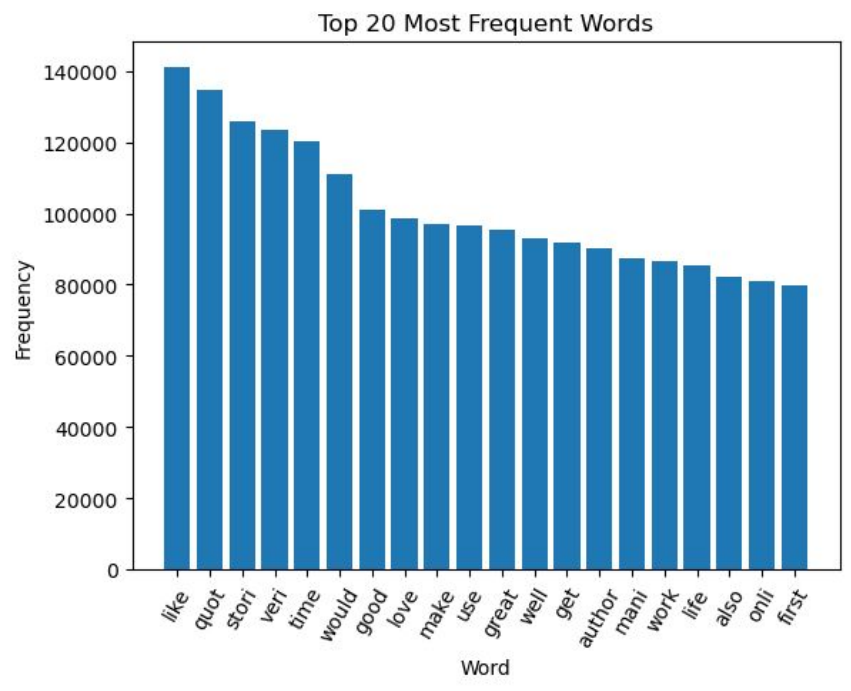
06

RESULTS & MODEL EVALUATION

06

RESULTS & EVALUATION -

Term Frequency



- Term frequency after removing modified stopwords. i.e. word “book” was removed due to nature of our data, meaning it will be everywhere carrying no sentiment.

06

RESULTS & EVALUATION -

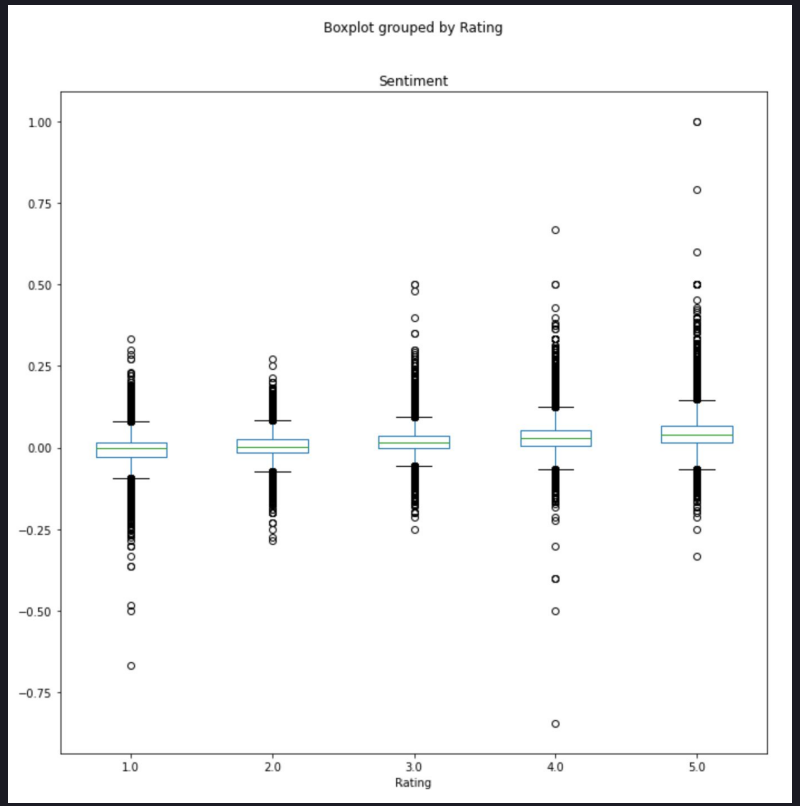
Logistic Regression Model - Rating and Reviews

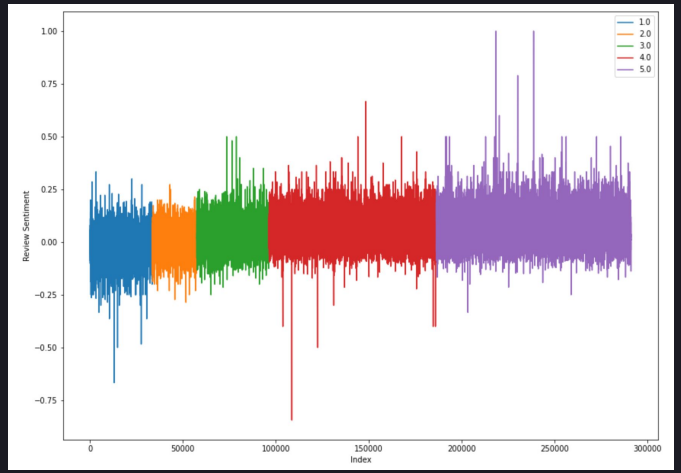
Ratings	5.0	4.0	3.0	2.0	1.0
Precision	0.793	0.495	0.479	0.376	0.714
Recall	0.868	0.080	0.070	0.048	0.340
Accuracy	0.780	0.814	0.918	0.948	0.945

- **Observation**: Model performance plummets as we go to rating lower than 5.0
- **Future Improvement**: Take skewness into account and train the model based on area under curve
- **Specifics**: Tune the model with gridsearch to find the optimal hyper parameter combination

06 RESULTS & EVALUATION -

Review Sentiment Analysis

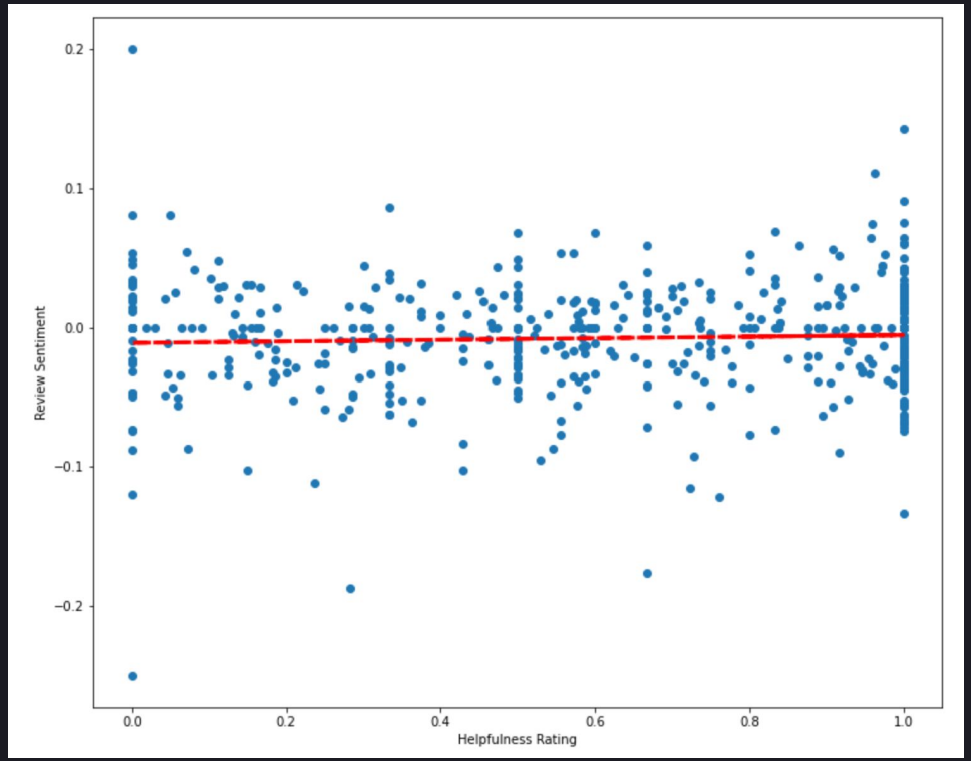




- **Frequency Distribution**: Rows were taken out to account for skewness since 5.0 was a more frequently occurring rating. Left as most occurring rating to simulate a real-world rating response.
- **Review Sentiment**: Calculated as a value between 1 and -1 using positive and negative dictionaries in NLTK, 0 representing neutral reviews.
- **Interpretation**: 4.0 reviews show that customers are still willing to give really high ratings despite having negative things to say.

06 RESULTS & EVALUATION -

Review Sentiment and Helpfulness Rating

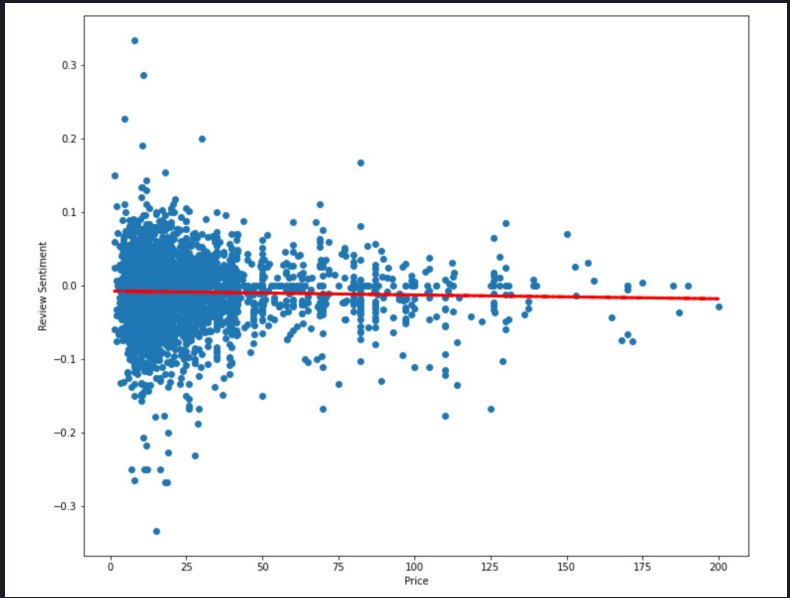


- **Observation**: There is no strong correlation between the review sentiment and the helpfulness rating.
- **Future Improvement**: If sentiment dictionaries could be updated to include negative/positive words associated with book reading specifically. Helpfulness rating could be better defined.
- **Specifics**: Before normalizing rating, eliminate helpfulness ratings that were difficult to infer meaning from like “0/0”.

06

RESULTS & EVALUATION -

Review Sentiment and Price



Real Customer Reviews

"This edition broke off mid-sentence in the last page or two. Don't order it until the problem has been fixed."

- a 1.0 rating on 'The Life of Jesus Critically Examined' priced at $205

"Ham radio overview should have been the title. This book is all bones and no meat and would be a waste of money to even a beginning ham. I would highly recommend Ham Radio for Dummies which in my opinion is the best book out there for beginners and as a reference."

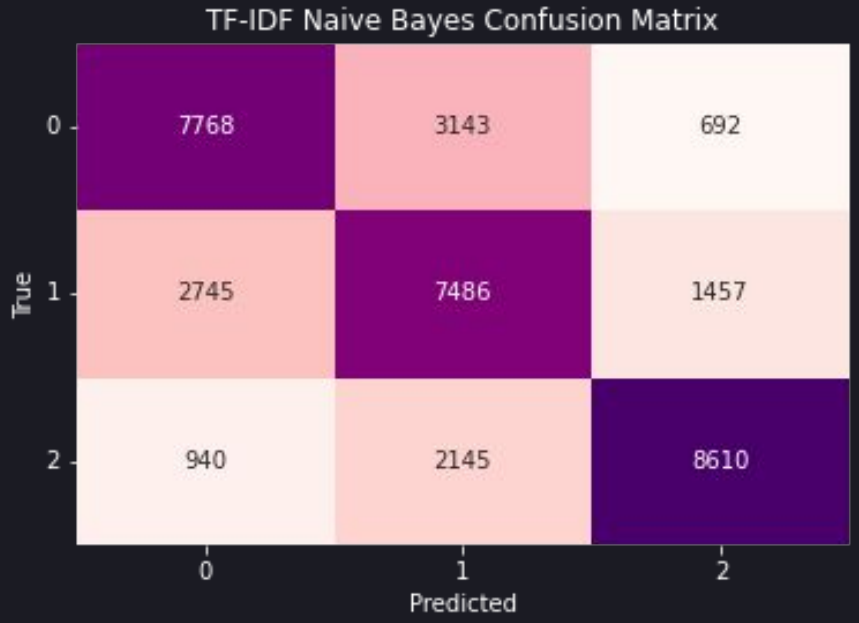
- a 1.0 rating on 'Ham Radio: Simplified' priced at $8.95

- **Observation**: No strong correlation although a wide range of sentiment towards cheaper books. This could be the case for a number of reasons.
- **Future Improvement**: Sentiment analysis did not pick up some meaning since it can't read context.
- **Specifics**: The second review example, although clearly negative to a reader, was given a sentiment score of 0.019231 which is positive.

06

MODEL EVALUATION -

Best Model Performer - Customer Reviews and Ratings



	Precision	Recall	Support	F1-Score
Negative	0.69	0.66	0.67	11754
Neutral	0.58	0.65	0.62	11671
Positive	0.80	0.74	0.22	11561

Accuracy			0.68	34986
Macro Avg.	0.69	0.68	0.68	34986
Weighted Avg.	0.69	0.68	0.68	34986

07

SCOPE & CONCLUSIONS

07

SCOPE & CONCLUSIONS

Review Score Summary and Rating

- There is a positive correlation between the review sentiment score and its rating from only 1 to 3 and 5, but no correlation in rating of 4. Reviews associated with ratings of 4 were amongst the most negative.

Sentiment Analysis: Helpfulness Rating & Price

- There is no positive or negative association between the review sentiment score and the helpfulness rating.
- There is a weak correlation between the sentiment and the price as well, although higher priced books see primarily neutral reactions.

Ideal Model - TF-IDF MNB

- Using TF-IDF with MNB allowed for the best results.
- With an accuracy of about 70%, and balanced Precision, Recall, Support, F1 scores, this model had the best performance.

08

REFERENCES & QUESTIONS

08

REFERENCES

- https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews
- https://www.amazon.com/amazonbookreview/

THANKS!

QUESTIONS?

CREDITS: This presentation template was created by **Slidesgo**, and it includes icons by **Flaticon**, infographics & images by **Freepik**