SENTIMENT ANALYSIS ON AMAZON REVIEWS

BY: THOMAS TAM

TABLE OF CONTENTS

About the Project 01

Dataset 02

Natural Language Processing 03

04 Exploratory Data Analysis

05 Predictive Modeling/Result

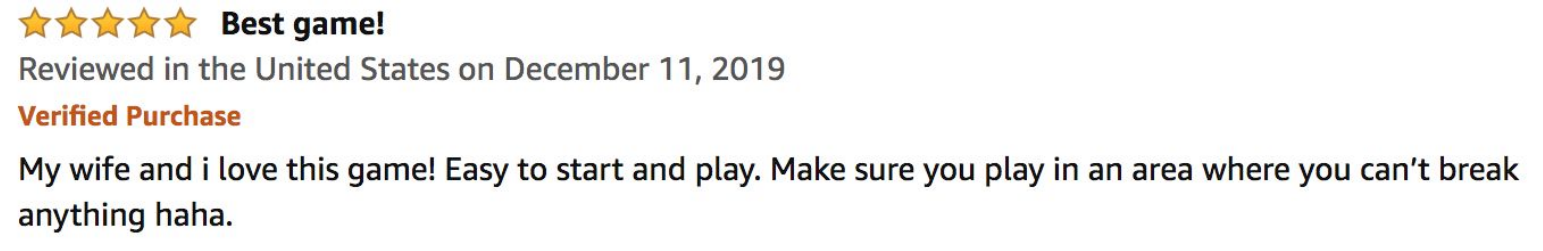
06 Future Work

ABOUT THE PROJECT

The goal of this project is to understand the human language by analyzing text data and create a sentiment analysis model by classifying Amazon reviews using Natural Language Processing. Based on number of stars for each user rating, the model will predict the sentiment for each rating.

BACKGROUND

- Item Focus: Board Games
- Text reviews are used as input values
- Amazon reviews are based on the number of stars between 1 and 5, serves as multi-class labels for classification



★★★★★ **Best game!**

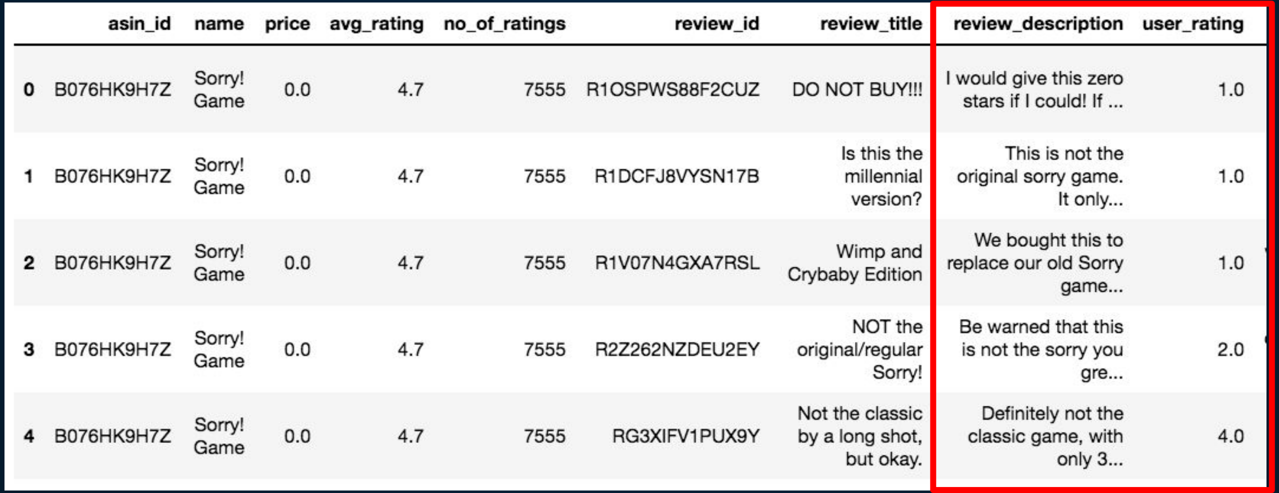
Reviewed in the United States on December 11, 2019

Verified Purchase

My wife and i love this game! Easy to start and play. Make sure you play in an area where you can't break anything haha.

DATASET

- Data scraped on October 13, 2020
- # Unique Products: 228 items
- # Unique Reviews: 5,539 reviews
- predict user_rating (ground truth labels) based on review_description



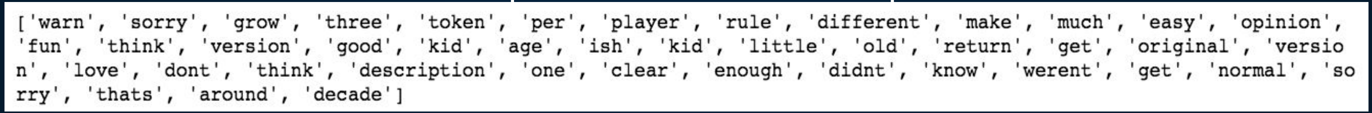
	asin_id	name	price	avg_rating	no_of_ratings	review_id	review_title	review_description	user_rating
0	B076HK9H7Z	Sorry! Game	0.0	4.7	7555	R1OSPWS88F2CUZ	DO NOT BUY!!!	I would give this zero stars if I could! If ...	1.0
1	B076HK9H7Z	Sorry! Game	0.0	4.7	7555	R1DCFJ8VYSN17B	Is this the millennial version?	This is not the original sorry game. It only...	1.0
2	B076HK9H7Z	Sorry! Game	0.0	4.7	7555	R1V07N4GXA7RSL	Wimp and Crybaby Edition	We bought this to replace our old Sorry game...	1.0
3	B076HK9H7Z	Sorry! Game	0.0	4.7	7555	R2Z262NZDEU2EY	NOT the original/regular Sorry!	Be warned that this is not the sorry you gre...	2.0
4	B076HK9H7Z	Sorry! Game	0.0	4.7	7555	RG3XIFV1PUX9Y	Not the classic by a long shot, but okay.	Definitely not the classic game, with only 3...	4.0

NATURAL LANGUAGE PROCESSING

NLP involves capturing and understanding the underlying meaning behind each review.

Preprocessing includes removing punctuation, emojis, numbers, stopwords, lemmatization, etc.

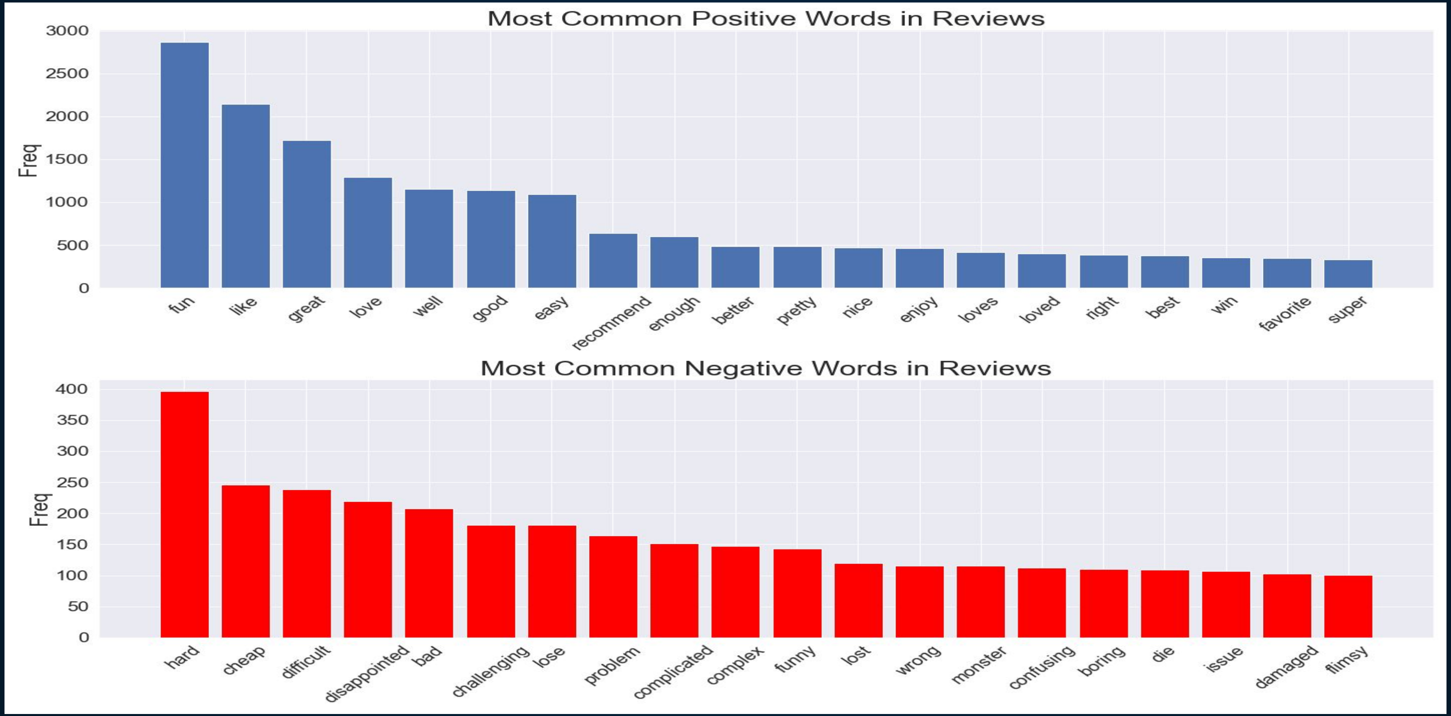
Tokenization example collection of processed words



```
['warn', 'sorry', 'grow', 'three', 'token', 'per', 'player', 'rule', 'different', 'make', 'much', 'easy', 'opinion',
'fun', 'think', 'version', 'good', 'kid', 'age', 'ish', 'kid', 'little', 'old', 'return', 'get', 'original', 'versio
n', 'love', 'dont', 'think', 'description', 'one', 'clear', 'enough', 'didnt', 'know', 'werent', 'get', 'normal', 'so
rry', 'thats', 'around', 'decade']
```

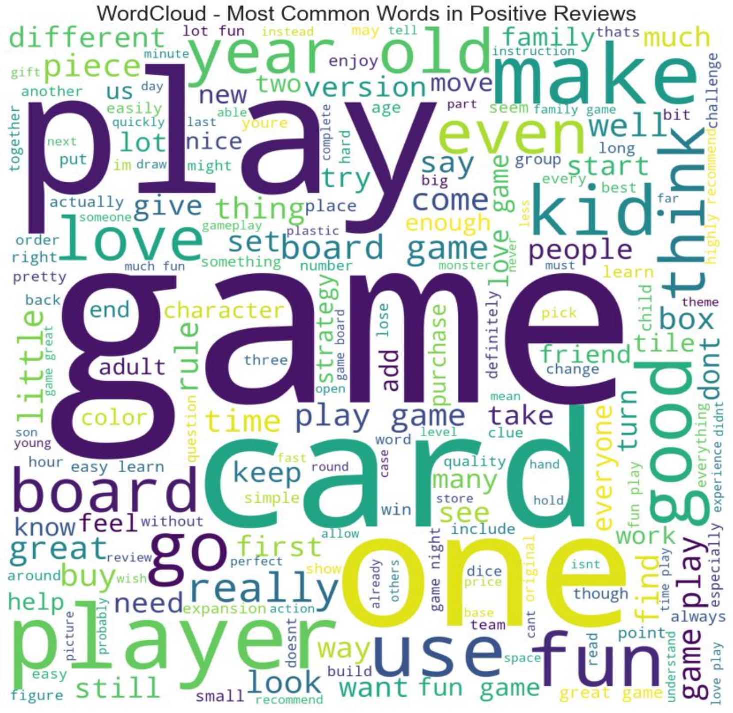
Word Vectorization extracts information from the text and converts it into numbers.

EDA - UNIGRAMS FOR COMMON WORDS

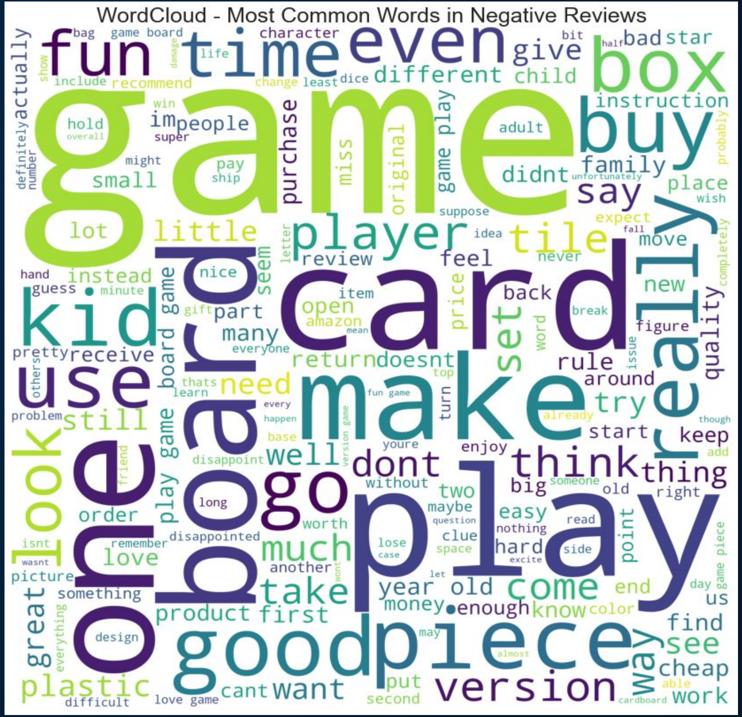


EDA - WORDCLOUD UNIGRAM

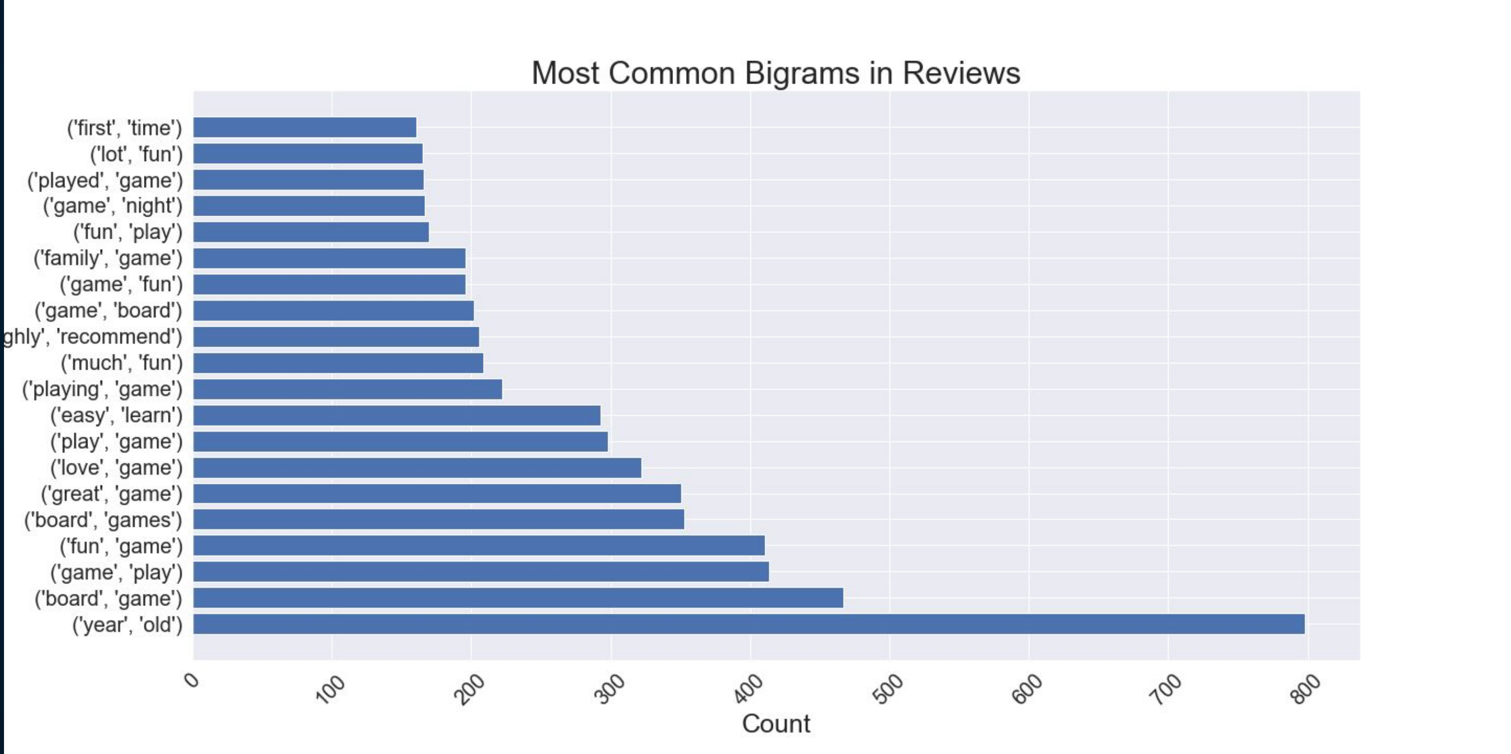
Most Common Words in Positive Reviews



Most Common Words in Negative Reviews

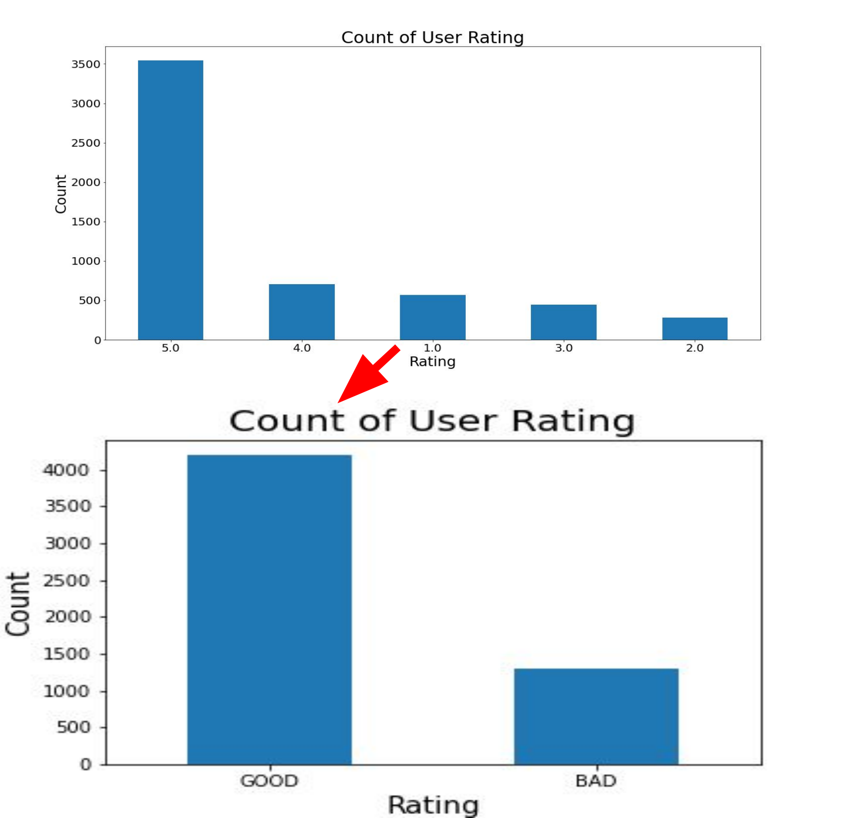


EDA - BIGRAMS



Most Common Bigrams in Reviews
('first', 'time')
('lot', 'fun')
('played', 'game')
('game', 'night')
('fun', 'play')
('family', 'game')
('game', 'fun')
('game', 'board')
ghly', 'recommend')
('much', 'fun')
('playing', 'game')
('easy', 'learn')
('play', 'game')
('love', 'game')
('great', 'game')
('board', 'games')
('fun', 'game')
('game', 'play')
('board', 'game')
('year', 'old')
0
100
200
300
400
500
600
700
800
Count

PREDICTIVE MODEL - PREP

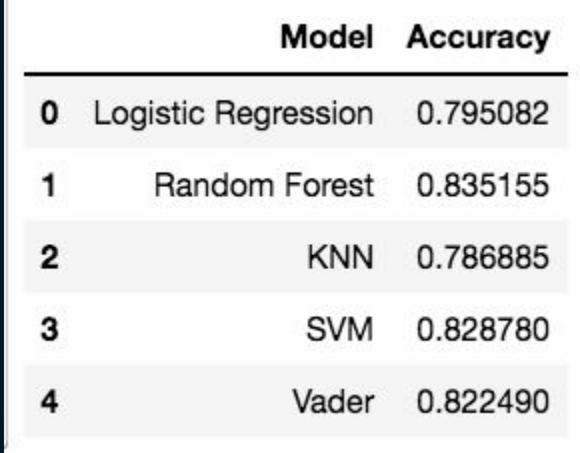


<u>Binary Classification</u>
Convert multiclass of 5 different ratings to binary classes

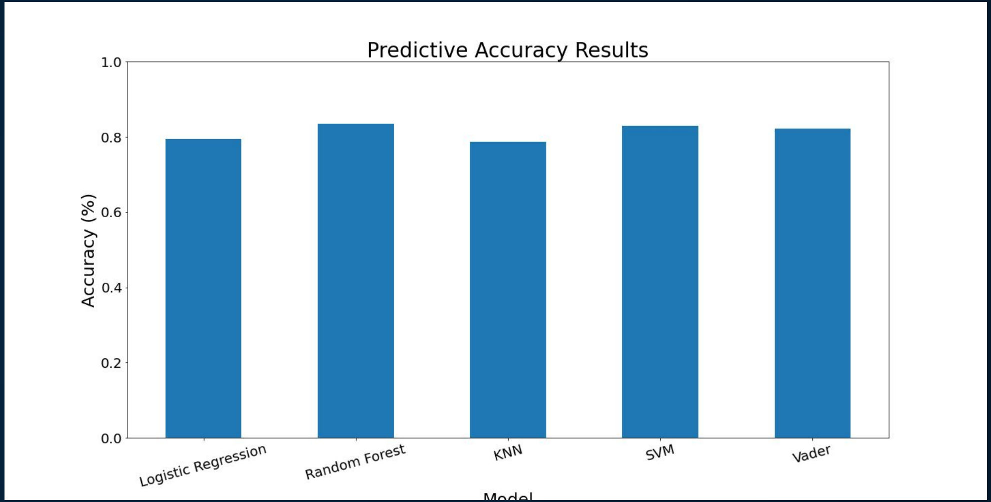
4-5 ratings: “GOOD”
1-3 ratings: “BAD”

<u>Class Imbalance Problem</u>
Balance class weights in algorithms

PREDICTIVE MODEL



	Model	Accuracy
0	Logistic Regression	0.795082
1	Random Forest	0.835155
2	KNN	0.786885
3	SVM	0.828780
4	Vader	0.822490



Metric: Accuracy to determine how many number of reviews the model predicts correctly

Rule-Based Algorithm: Vader
Machine Learning Algorithms: Logistic Regression, Random Forest, KNN, SVM

Best Predictive Algorithm: Random Forest with accuracy of 0.835

FUTURE WORK

- Increase the number of items and reviews to improve prediction results through Amazon API
- Model does not currently handle misspelled words, sarcasm and irony
- Improve to multi-class classification model

THANK YOU