Recommender Systems for Amazon

Ke Li, Parth Gupta, Senqi Zhang, Vishweshwar Tyagi, Xuanhao Wu
Group 29

Columbia | Engineering
The Fu Foundation School of Engineering and Applied Science

Introduction

- We use the Amazon Review Data (2018) for the video games category that has over 2.5 million reviews and about 84,819 observations in the meta dataset.
- It consists of two datasets, the review dataset that consists of reviews and the meta dataset, which consists of information about the products. These two are linked together by 'asin', which denotes product ID.

review dataset with $2,565,349$ reviews where,

- **reviewerID** - reviewer ID
- **asin** - product ID
- **reviewerName** - reviewer name
- **vote** - no. of votes for the review, indicating its helpfulness
- **style** - dictionary of product attributes
- **reviewText** - review statement
- **overall** - product rating provided by the reviewer
- **summary** - review summary
- **unixReviewTime** - unix time of the review
- **reviewTime** - raw time of the review
- **image** - images posted by reviewer

meta dataset with $84,819$ sample points where,

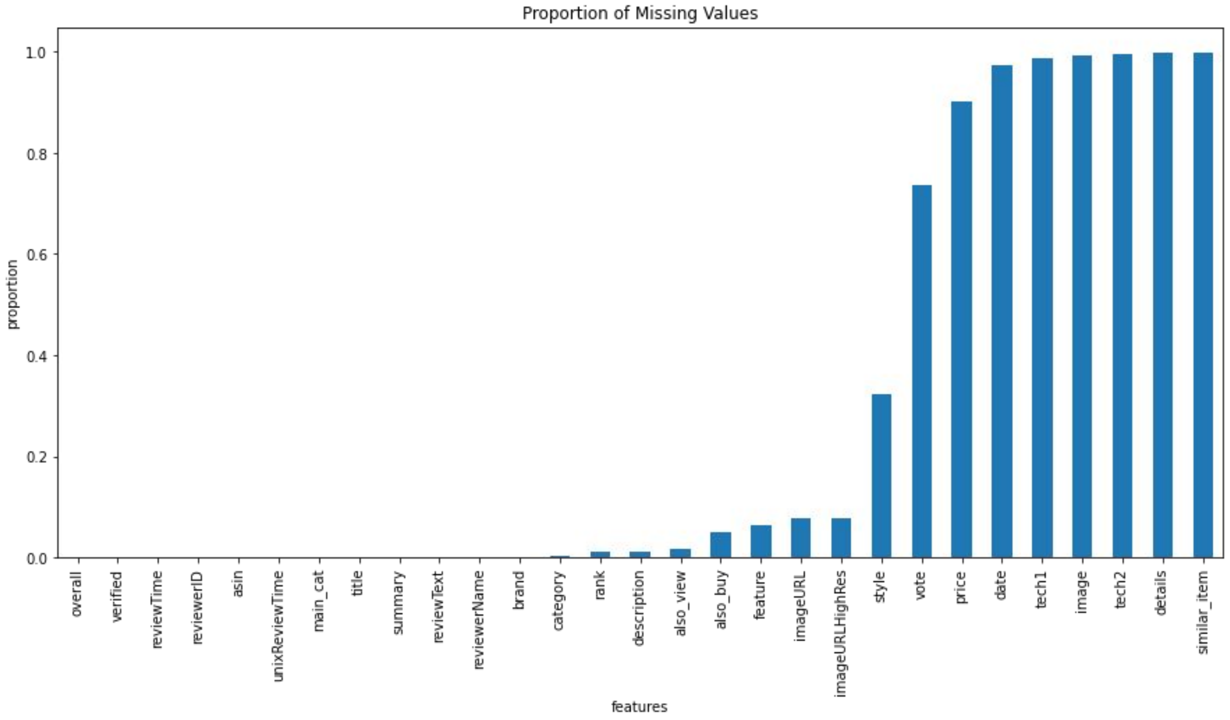
- **asin** - ID of the product
- **title** - name of the product
- **feature** - bullet-point format features of the product
- **description** - description of the product
- **price** - price in US dollars (at time of crawl)
- **imageURL** - url of the product image
- **imageURLHighRes** - url of the high resolution product image
- **related** - related products (also bought, also viewed, bought together, buy after viewing)
- **salesRank** - sales rank information
- **brand** - brand name
- **categories** - list of categories the product belongs to
- **tech1** - the first technical detail table of the product
- **tech2** - the second technical detail table of the product
- **similar** - similar product table

Data Cleaning and Preprocessing

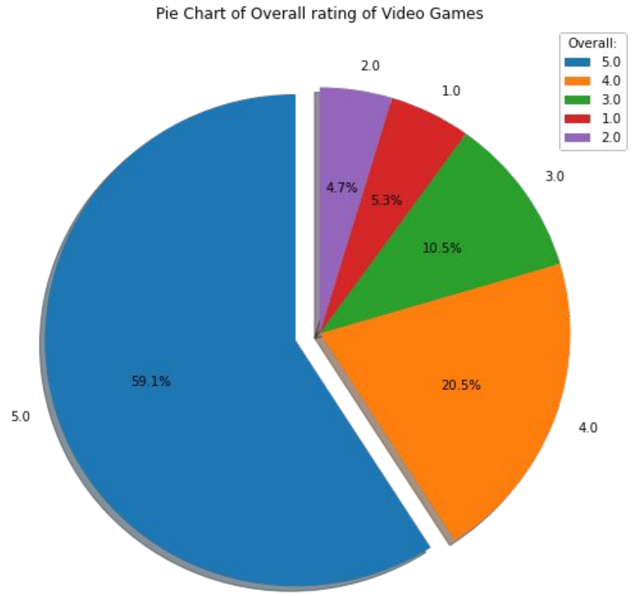
- After initial exploration, we found out that there were duplicates in the meta dataset. So, we removed them.
- Some product IDs in the review dataset were absent from the meta dataset. We got rid of rows from the review dataset containing such product IDs.
- Merged the review and meta dataset using a left join on ‘asin’. Doing so, we lost product IDs present in the meta dataset but absent from the review dataset; no one reviewed these.
- To draw meaningful insights and overcome computational limitations, we subset a 10-core from the merged data. A *k-core* subset ensures that each product is reviewed at least k times as well as each reviewer has provided at least k reviews.
- Doing so reduced the number of reviews from 2,565,349 to 126,703.

Missing Values

- We plotted the missing values of our features.
- All features that have over 40% of missing data will be dropped, except for “vote” in which we will replace the NaN values with 0.
- Features like “imageURL”, “unixReviewTime” will also be dropped as we believe they will not contribute to the model.

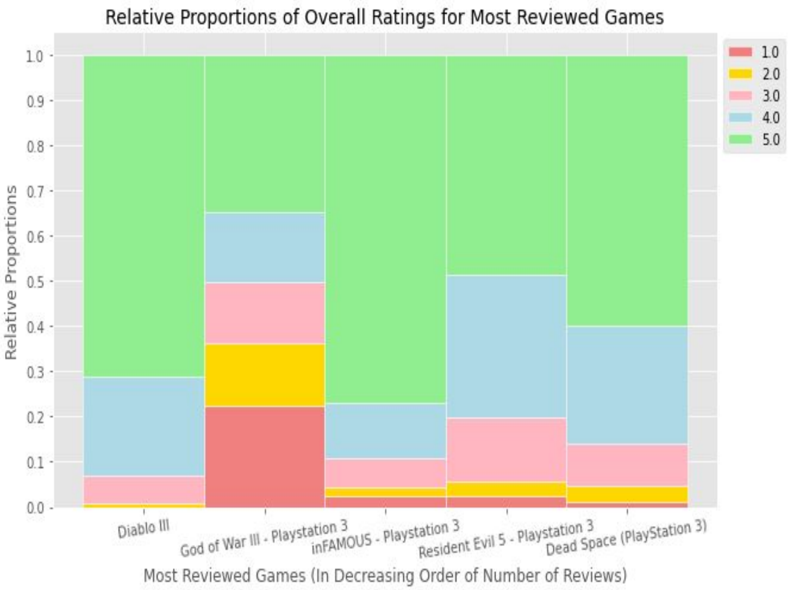


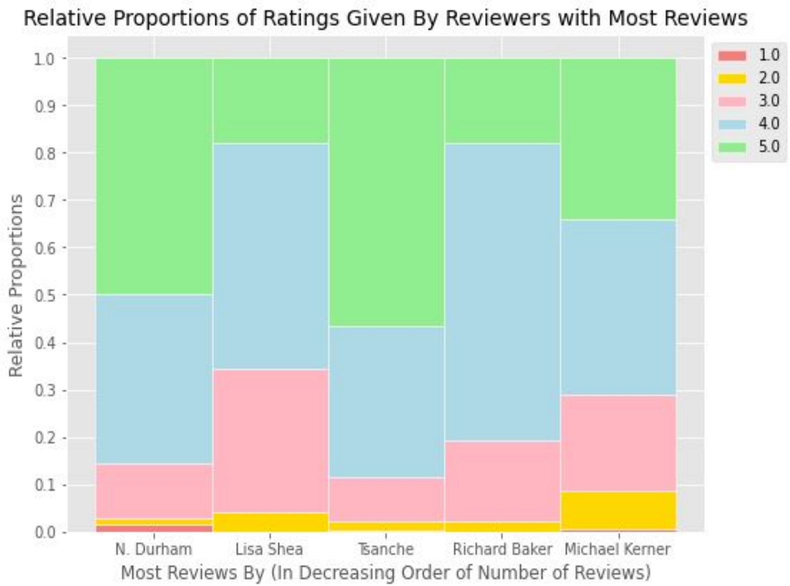
Distribution of Overall Ratings



- We see that the proportion of verified reviews increases with the overall rating.
- We speculate that this might be due to malicious negative reviews from people who did not even buy the product.

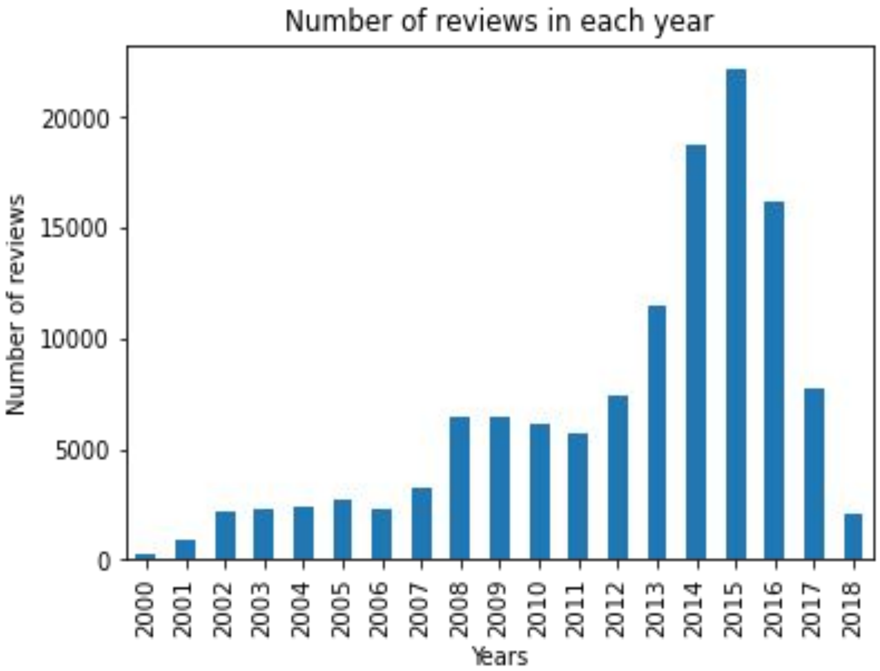
Most Reviewed Games and Most Frequent Reviewers





- Except for God of War 3, which has mixed reactions with roughly equal proportions for each rating point, most popular video games were positively reviewed with the most common rating as 5.0.
- Among those who review the most, we see that they generally give a rating of 3 or more. This tells us that these reviewers are more likely to leave positive feedback.

Number of Reviews in Each Year



The review ranges from 2000-2018. We see an increase in the reviews. The peak appears in 2015 and follows a decrease. We suspect that this is caused by the rise of PC and mobile games like PUBG and Fortnite. More and more players are shifting from playing on the console to PC and mobile.

Sentiment Analysis of Review Text

5 random reviews with the highest positive sentiment polarity:

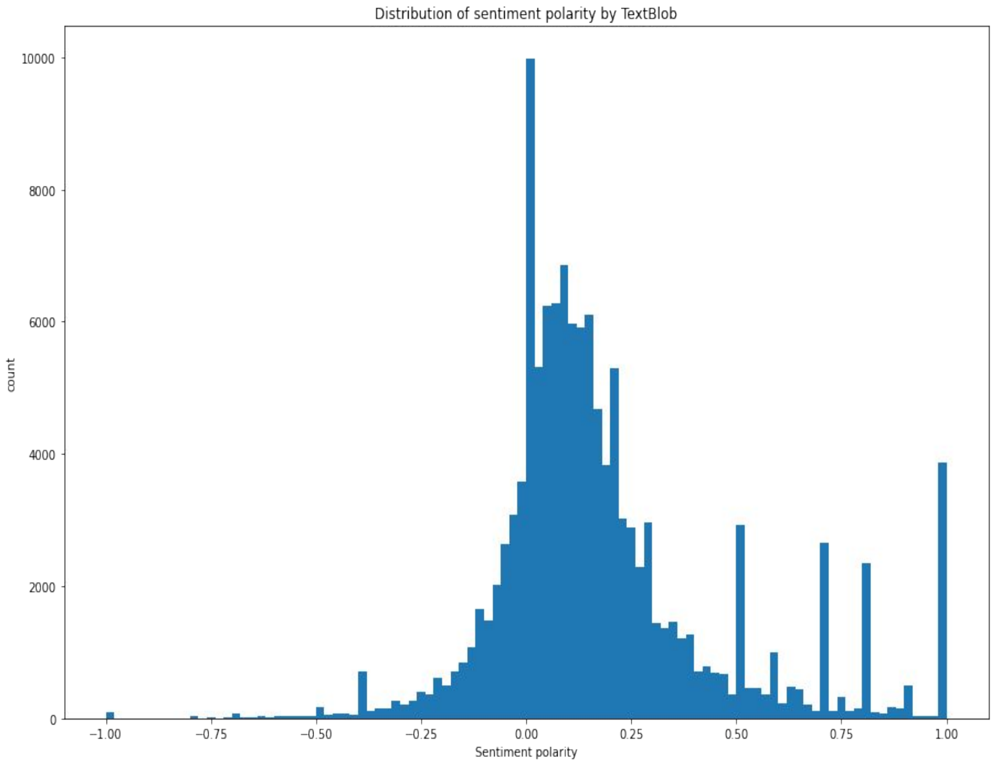
- Best experience on planet earth.
- Excellent recommended
- Awesome! Thank you!!!!
- Excellent.
- Excellent

5 random reviews with the most neutral sentiment(zero) polarity:

- Gave as Gift.
- Kids
- Bought it for my Grandson. He loves the Destiny games.
- it works for what i need
- easier to get it here then in store will buy again

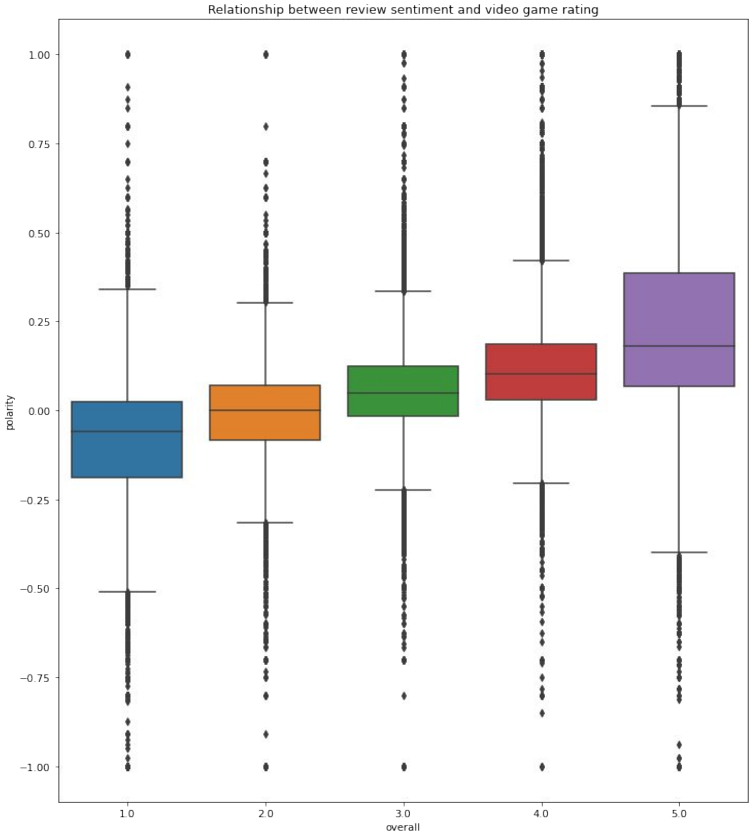
5 random reviews with the most negative sentiment(-1) polarity:

- Just awful. I suggest a ps4 and SFV
- Horrible
- Boring
- Worst of the series....
- Boring



Relationship between Review Sentiment (polarity) and Video Game Rating

This is the distribution of sentiment scores for each rating score. We see a clear increase in sentiment score as the rating increases. However, we do see outliers that does not make much sense. For example, 1.0 ratings can have very positive reviews and 5.0 ratings can have very negative reviews. We will further explore those outliers and drop them if necessary.



COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Techniques

- We will compare classical recommender system techniques like Content based filtering and Collaborative filtering.
  - Content-based filtering - Track the user's action such as the products bought or reviewed by the user to create a user profile, and compare with product categories to make recommendations.
  - Collaborative filtering - Track the user's preference and compare with other users who have similar tastes to predict what the user will also like.
- We will try to explore the effects of metadata (reviews and summary) in content-based filtering. We will do this by implementing a review-based recommender system that extract informative text data from user-generated reviews as criterias used in the recommender systems to enhance accuracy.
- Since text data is the core of a review, we will apply sentiment analysis and topic modeling on them.
  - Sentiment Analysis - Determine the sentiment scores for each reviews and classify them as "Extremely negative", "Negative", "Netural", "Positive", "Extremely positive".
  - Topic Modeling - Determine topics within the reviews. We will classify each review into a certain topic using keywords.