Report: Amazon Reviews'23

Code: `github.com/sueszli/amazeballs/`

Motivation Understanding online reviews isn't just about interpreting customer opinions; it's a window into how people perceive and interact with products across diverse categories. The Amazon Dataset'23 offers a treasure trove of insights, allowing us to explore patterns in sentiment, subjectivity and the elements that matter most in consumer decision-making. By digging into this data, we aim to uncover the subtle relationships between what customers say and how they rate products, shedding light on the dynamics of trust, satisfaction and expectation in the digital marketplace.

Beyond the findings, this report highlights the (data science) process of turning raw, unstructured data into actionable knowledge. Through techniques like sentiment analysis, topic modeling and classification, we're not just addressing key questions about product reviews – we're also demonstrating the iterative, hands-on nature of data science itself.

Process The process which we followed, is more formally known as CRISP-DM (Cross-Industry Standard Process for Data Mining). It begins with (1) business understanding, where we refine the research questions in consultation with a supervisor for our project, define variables and metrics and build hypotheses while being mindful of biases. Next, we move to (2) data understanding, where we sample and preprocess the data, ensuring privacy and assess the accuracy, biases and reliability of the measurements. In (3) data preparation, we clean the data by checking for missing values, outliers and inconsistencies, calculating descriptive statistics and transforming the data as needed. If the data is insufficient to answer the research questions, we may combine columns, look for additional datasets, or modify the questions. In (4) modeling, we calculate correlations and build models to explore the relationships between variables. During (5) evaluation, we plot the data, identify patterns and anomalies, visualize the findings and check predictions to assess if the models answer the original questions. Finally, (6) deployment involves using the results to make decisions or share insights with stakeholders.

Methodology

Research Questions First we define the research questions that we aim to answer. Our team has selected task 21 from the list provided by the course team and did not further modify it. The research questions are as follows:

- RQ1: Are reviews for some categories of product on Amazon overall more positive than for other categories?
- RQ2: Are reviews more subjective for some classes of products than for others?
- RQ3: Which aspects of different classes of products are the most important in the reviews?
- RQ4: Can one predict the star rating from the review text?

The first research question is a comparison of sentiment across categories, the second is a comparison of subjectivity across categories, the third is a topic modeling task, commonly referred to as aspect-based sentiment analysis and the fourth is a classification task to predict the star rating from the review text.

Dataset Selection

- https://amazon-reviews-2023.github.io/
- https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023 - easier to download from huggingface - Larger Dataset: We collected 571.54M reviews, 245.2% larger than the last version; - Newer Interactions: Current interactions range from May. 1996 to Sep. 2023; - Richer Metadata: More descriptive features in item metadata; - Fine-grained Timestamp: Interaction timestamp at the second or finer level; - Cleaner Processing: Cleaner item metadata than previous versions; - Standard Splitting: Standard data splits to encourage RecSys benchmarking.

- sample because too large - 100,000 samples per category (2.92 GB): doesn't fit in memory for plotting - 10,000 samples per category (0.33 GB): inference would take 8 days (339880 items with 2it/s) - 1,000 samples per category (0.03 GB): inference would take 19 hours (33994 items with 2it/s) - 100 samples per category (<0.00 GB): inference would take 2 hours (3399 items with 2it/s) — this is what we used
- lots of languages, so models had to be multilingual - some of them were, others weren't

Findings