MIT 805 Assignment 1 - Part 1

Armand de Wet u 16005326

23 August 2024

1 Technical Aspects of Amazon reviews 2023 dataset

I decided to make use of the Amazon Review 2023 dataset [1] as it presented large multi modal data that can contribute to business insights and value. I specifically focus on the electronics category, thus any value generated may be to the benefit of electronic online retailers selling there product on Amazon.

Each category in the complete dataset is split into 2 sub-categories namely user review (numerical ratings, textual rating, helpfulness votes etc.) and item metadata (descriptions, price, links to different raw images etc.).

This data is available in JSONL (JSON Lines) format, which has the same structure as JSON data but uses a line by line format for easier streaming and processing of individual objects [2]. This allows for easy access and manipulation of the data. The electronics category makes up about 27.2 GB of space.

There are multiple versions of this dataset, but I make use of the latest version that was collected in 2023. (up-to-date as of September 2023 and contains interactions from May 1996.)

2 Data Overview

2.1 Data Fields

2.1.1 User Reviews

Field	Type	Explanation
rating	float	Rating of the product (from 1.0 to 5.0).
title	str	Title of the user review.
text	str	Text body of the user review.
images	list	Images that users post after they have received the product. Each image has different sizes (small, medium, large), represented by the small_image_url, medium_image_url, and large_image_url respectively.
asin	str	ID of the product.
parent_asin	str	Parent ID of the product. Note: Products with different colors, styles, sizes usually belong to the same parent ID. The "asin" in previous Amazon datasets is actually parent ID. **Please use parent ID to find product meta.**
user_id	str	ID of the reviewer
timestamp	int	Time of the review (unix time)
verified_purchase	bool	User purchase verification
helpful_vote	int	Helpful votes of the review

Figure 1: Description of user review data fields

2.1.2 Item Metadata

Field	Type	Explanation
main_category	str	Main category (i.e., domain) of the product.
title	str	Name of the product.
average_rating	float	Rating of the product shown on the product page.
rating_number	int	Number of ratings in the product.
features	list	Bullet-point format features of the product.
description	list	Description of the product.
price	float	Price in US dollars (at time of crawling).
images	list	Images of the product. Each image has different sizes (thumb, large, hi_res). The "variant" field shows the position of image.
videos	list	Videos of the product including title and url.
store	str	Store name of the product.
categories	list	Hierarchical categories of the product.
details	dict	Product details, including materials, brand, sizes, etc.
parent_asin	str	Parent ID of the product.
bought_together	list	Recommended bundles from the websites.

Figure 2: Description of item metadata data fields

2.2 Data collection

This data was collected by McAuley Lab by crawling and scraping data from Amazon, specifically to use in there article "Bridging Language and Items for Retrieval and Recommendation" [1]. In this article they present pre-trained sentence embedding models for recommendation purposes.

3 Expected Relationships and Correlation

- I suspect items with more reviews to have higher average ratings.
- Different brands of the same type of product have different number of reviews, average rating and successful purchase verification.
- Longer reviews have on average more helpful votes.
- The average rating of a product has an influence on the number of verified purchases.

- The average rating length of a product has an influence on the number of verified purchases.
- The number of ratings of a product has an influence on the number of verified purchases.
- Expect to see cyclical nature (e.g. seasonal or monthly) of the number of product reviews of certain items.
- Product price can correlate with the number of reviews a product has and the number of verified purchases of said product.
- Product price and average rating can correlate.
- Certain brands have higher item pricing on average.

4 Data Description (The V's)

The 5 most well known features of Big Data are the following: velocity, volume, value, variety and veracity [3]. I will thus proceed by describing the Amazon 2023 reviews data at the hand of these 5 V's.

4.1 Velocity

The data is not real-time/streaming data as it was collected through web crawling and scraping and then stored for analysis. There doesn't seem to be a set frequency to the data collection as the previous versions were collected in 2013, 2014 and 2018 respectively; with the latest version being collected in 2023.

4.2 Volume

The entire dataset is very large as it is made up of 34 different categories, with the electronics category already making up 27.2GB. Thus even this subsection of the data constitutes "Big Data".

4.3 Value

This dataset can contribute to immense value for electronic online retailers selling their product through Amazon, as it can be used to identify hidden patterns, trends, relationships and correlations as mentioned in section 3. This data can also be fed into ML models to, for example, predict seasonal product performance and also, since image data is included in this set, different metrics can be predicted based on image alone e.g. average expected rating, number of reviewed sales etc.

4.4 Variety

This dataset contains a rich variety of data types. It contains structured data such as customer item ratings. It also contains unstructured data in the form of text based customer reviews and product descriptions. It is also possible to incorporate product images from the product URLs provided, thus making up visual unstructured data. This set also contains semi-structured data in the form of product metadata.

4.5 Veracity

The quality and accuracy of this dataset is high as it was collected directly from the Amazon website. One issue in the structure of the data is that the sub-fields of the description field in the item metadata are not homogeneous, hence grouping the data by those sub-fields or extracting data from the sub-fields may prove difficult.

References

[1] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, and J. McAuley, "Bridging language and items for retrieval and recommendation," *arXiv preprint arXiv:2403.03952*, 2024.

[2] S. Mudadla. "Difference between ordinary json and json lines?" (2023), [Online]. Available: `https://medium.com/@sujathamudadla1213/difference-between-ordinary-json-and-json-lines-fc746f93d75e#:~:text=In%20summary%2C%20the%20key%20difference,and%20processing%20of%20individual%20objects.` (visited on 2024).

[3] S. Robinson and A. S. Gillis. "5v's of big data." (2023), [Online]. Available: `https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data#:~:text=The%205%20V's%20of%20big%20data%20%2D%2D%20velocity%2C%20volume%2C%20value,innate%20characteristics%20of%20big%20data.` (visited on 2024).