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

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.

For the sentiment analysis task (RQ1) we used a pre-trained and distilled version of a multi-lingual Bidirectional Encoder Representations from Transformers (BERT) model to classify the sentiment of the reviews into positive, negative or neutral classes in addtion to a sentiment score. We were able to notice that lanauges other than English were also present in the dataset by using the `langdetect` library, however, not all following models were multi-lingual and we thus had to tolerate some errors, especially in the aspect extraction task.

Subjectivity (RQ2) was again determined using a BERT model. But this time it was tuned on the "Wiki Neutrality Corpus dataset" which indirectly adopts Wikipedia's NPOV policy as the definition for "neutrality" and "subjectivity". The NPOV policy may not fully reflect an end users assumed or intended meaning of subjectivity because ironically enough, the policy itself is subjective. However, it is a good starting point for a model to learn what is considered neutral and what is not and suitable for our small scale project.

The aspect extraction task (RQ3) was done using an inaccurate but highly efficient keyword extraction algorithm YAKE! which is based on the TextRank algorithm. Due to compute limitations, we were not able to use a more accurate models like SetFitABSA [1] or `pyabsa` for aspect extraction. However we did implement them in case the reader is interested in running them on their own machine.

Finally, for the star rating prediction task (RQ4) we used a pre-trained BERT model fine-tuned on an older and exclusively English version of the Amazon Reviews dataset, reaching an accuracy of 0.8. This model was able to predict the star rating of a review with a high degree of accuracy.

Dataset Selection To answer these questions, we chose the Amazon Reviews'23 dataset which is the standard dataset for the Amazon product reviews in the RecSys and NLP communities. This dataset is a collection of 571.54M reviews, 245.2% larger than the last version, with interactions ranging from May 1996 to September 2023. It includes richer metadata, fine-grained timestamps and cleaner processing, making it an ideal choice for our analysis. Most importantly it is easily accessible through the Hugging Face Datasets library, which simplifies the data loading and preprocessing steps.

- 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, it was small enough to push to git (2 MB)
- lots of languages, so models had to be multilingual - some of them were, others weren't

Potential Biases Due to our sampling strategy, we might introduce bias by selecting a limited number of reviews per category, which might not be representative of the entire dataset. Additionally, the pre-trained sentiment analysis and aspect extraction models are trained on specific data, potentially leading to errors in non-English reviews or niche categories. We tried to mitigate the latter by filtering aspects with high confidence scores and using multilingual models, but challenges remain in handling these cases.

Data Science Tools & Techniques Learned This project involved a range of data science tools and techniques. From a technical perspective, we used Python libraries like Pandas for data manipulation, Hugging Face Transformers for sentiment analysis and text classification, and sampling techniques for efficient handling of large datasets. We also explored aspect extraction using keyword extraction and aspect-based sentiment analysis. The project provided hands-on experience with real-world data analysis, including data preprocessing, model training, and evaluation. The team also gained insights into the challenges of working with large datasets, multilingual reviews, and the importance of sampling and model selection in data analysis.

Team Work Division Tasks were divided based on expertise: data preprocessing, sentiment analysis, aspect extraction and sampling were handled by different team members. Each person focused on the components that best suited their skills. We mostly worked in person and collaborated exclusively through pair programming sessions.

Findings

[1] Jayakody, D., Isuranda, K., Malkith, A. V. A., De Silva, N., Ponnamperuma, S. R., Sandamali, G. G. N., & Sudheera, K. L. K. (2024, August). Aspect-based Sentiment Analysis Techniques: A Comparative Study. In 2024 Moratuwa Engineering Research Conference (MERCon) (pp. 205-210). IEEE.