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 them, as they already cover a wide range of topics for customer review analysis. 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.

Data Preprocessing The data cleaning and transformation process involved several key steps to manage the dataset's size and ensure it was suitable for analysis. Initially, the dataset contained 100,000 samples per category, which amounted to 2.92 GB and was too large to fit in memory for plotting. This necessitated reducing the dataset size to make it manageable. First, we considered sampling 10,000 samples per category, which would reduce the dataset size to 0.33 GB. However, this would have resulted in an inference time of 8 days due to the large number of items (339,880) and the slow processing speed of 2 items per second. Subsequently, we explored sampling 1,000 samples per category, which would further reduce the dataset size to 0.03 GB. However, this would still require 19 hours for inference, which was deemed too long for practical use. Finally, we decided to sample only 100 samples per category, resulting in a dataset size of less than 0.00 GB. This smaller dataset contained 3,399 items, which could be processed within a reasonable timeframe of approximately 2 hours at a rate of 2 items per second. This approach allowed for effective model inference and analysis while maintaining a manageable dataset size and was small enough to push to git.

The data transformation process involved (1) loading the data from HuggingFace and intermediate caching on multiple levels to avoid redundant bandwidth usage, (2) sampling the data to reduce the dataset size using a fixed seed for reproducibility, (3) preprocessing the data by dropping unnecessary columns, converting timestamps to datetime objects, removing rows with missing values in critical columns, and cleaning text fields by stripping HTML tags and whitespace, and filtering out empty strings. Finally, the cleaned data was prepared for inference tasks and the subsequent analysis.

Missing Values Missing values in the dataset are addressed through a combination of data preprocessing techniques. The primary method used is listwise deletion, where rows with missing values in specific columns are removed. This is evident in the preprocess function, where the `dropna` method is applied to remove rows with missing values in the "text", "title", and "rating" columns. This approach ensures that only complete cases are included in subsequent analyses, which can help maintain the integrity of the dataset but may lead to a loss of data if many entries are incomplete. Additionally, the code handles missing values by cleaning text data to ensure that any remaining entries are valid. For instance, HTML tags are stripped from the "text" and "title" fields, and whitespace is removed. This step ensures that any non-informative or empty text entries are filtered out, thereby improving data quality.

The choice of listwise deletion is appropriate when missing data is assumed to be missing completely at random (MCAR), as it can prevent bias in analysis if this assumption holds true. However, if the data is not MCAR, this method might introduce bias and reduce statistical power. Alternative methods such as imputation (mean, median, mode) or more sophisticated techniques like multiple imputation could be considered if preserving all available data is critical and if the missingness pattern allows for such methods.

[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.

Potential Biases Firstly, there is a significant sampling bias due to the nature of the data collection process, which relies on reviews from Amazon. This platform-specific focus can lead to a lack of representativeness, as it excludes opinions from other e-commerce platforms and may reflect the purchasing habits and preferences unique to Amazon users. Additionally, many reviews are from users who have only left a single review, which suggests a reporting bias where occasional reviewers might not provide as balanced or informed feedback as more frequent reviewers.

The data processing phase introduces further biases. For instance, reviews are truncated to 512 tokens for analysis, potentially omitting important context or nuances found in longer reviews. Moreover, the language detection and sentiment analysis models used are primarily trained on English data, which could introduce language bias if reviews in other languages are not accurately processed or interpreted.

The models employed for sentiment analysis and aspect extraction also carry intrinsic biases. These models are trained on specific datasets that may not fully capture the diversity of consumer opinions or product types found on Amazon. For example, sentiment analysis models might oversimplify complex sentiments into basic categories like positive or negative, thus failing to capture more nuanced consumer feedback.

Additionally, there is a notable presence of fake, paid, or incentivized reviews on Amazon, which can significantly skew the perceived quality and satisfaction levels of products. These practices compromise the reliability of reviews and can mislead consumers by presenting an inaccurate picture of product quality. The combination of these biases suggests that while Amazon reviews can provide valuable insights, they should be interpreted with caution and supplemented with more objective sources when possible.

Challenges & Lessons Learned This project involved a range of data science tools and techniques. From a technical perspective, we used a reproducible `virtualenv` environment, a `venv` to `docker` and `conda` transpiler written in a makefile, to ensure both portability and reproducibility through the compilation of a requirements text file. Additionally, we used the `datasets` library to load and preprocess the dataset. This is standard practice. Using these technologies we learned how to handle the challenges of working on data in teams, and managing large datasets with very limited computational resources.

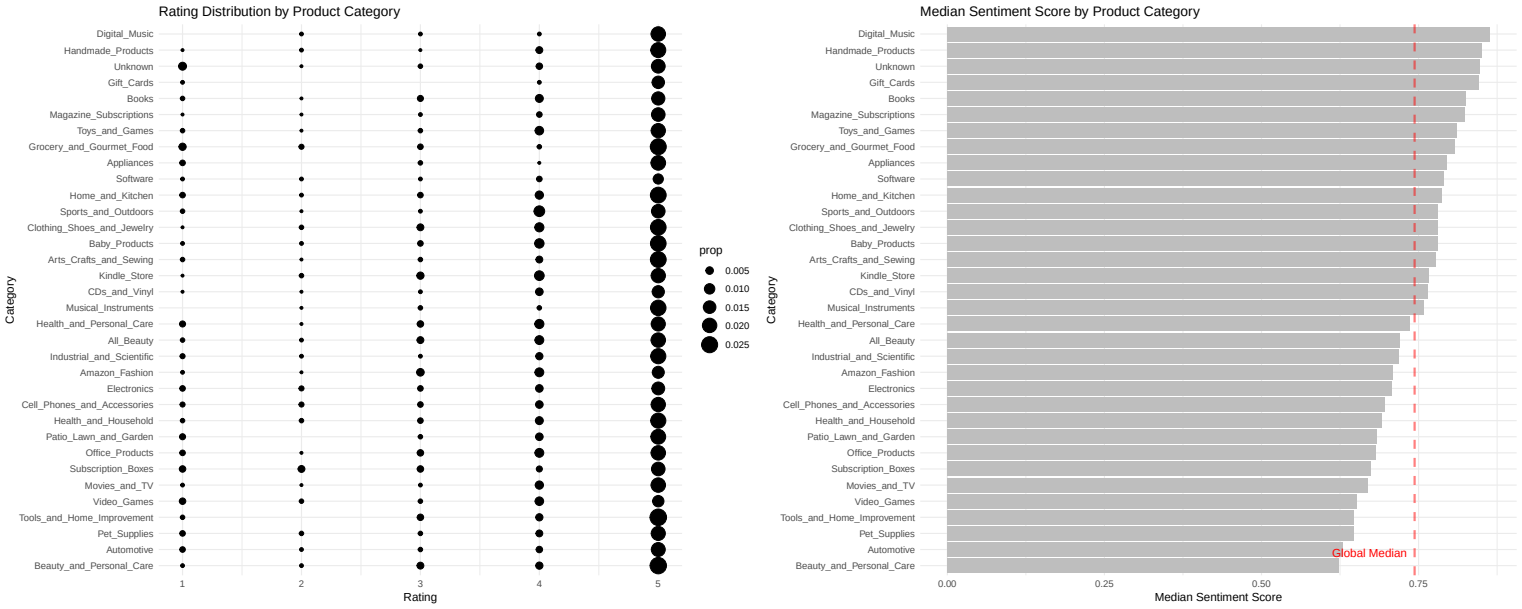
From a theoretical perspective We also explored aspect extraction using keyword extraction and aspect-based sentiment analysis. The project provided hands-on experience with real-world data analysis, including data preprocessing, model training and evaluation. The team also gained insights into the challenges 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 exclusively worked in-person and collaborated through pair-programming sessions.

Results

After some initial exploratory data analysis and visualization, we were able to answer the research questions visually and quantitatively. The results are as follows.

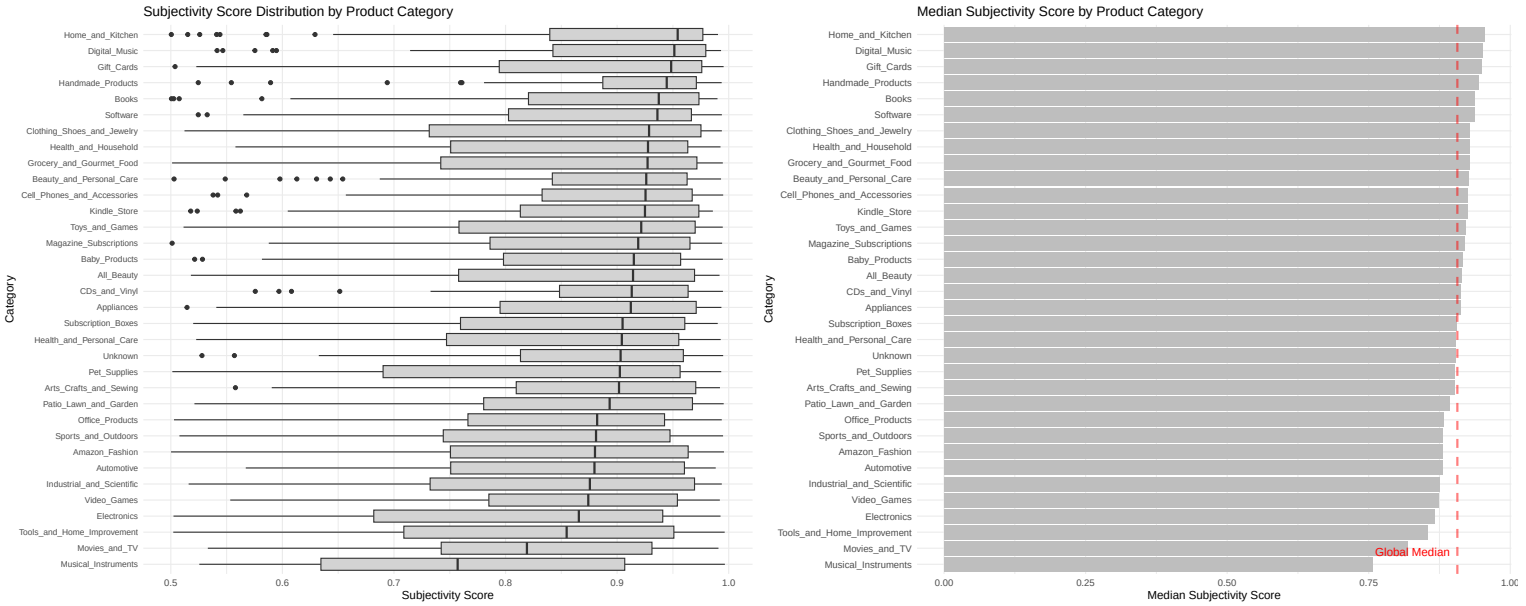
RQ1: Sentiment Analysis



First, we analyzed the results of our sentiment analysis model and additionally compared it with the distribution of ratings per category. For the reviews we used a scatter plot to visualize the distribution of ratings per category and for the median sentiment we used a horizontal bar plot. Against our expectations, the sentiment scores did not correlate with the ratings. Specifically, categories with significantly fewer 5-star reviews, such as “Software” still had a beyond global median sentiment score of around 0.8. This suggests that the sentiment scores are not directly related to the ratings, and that the sentiment analysis model may be capturing a different aspect of the reviews than the star ratings.

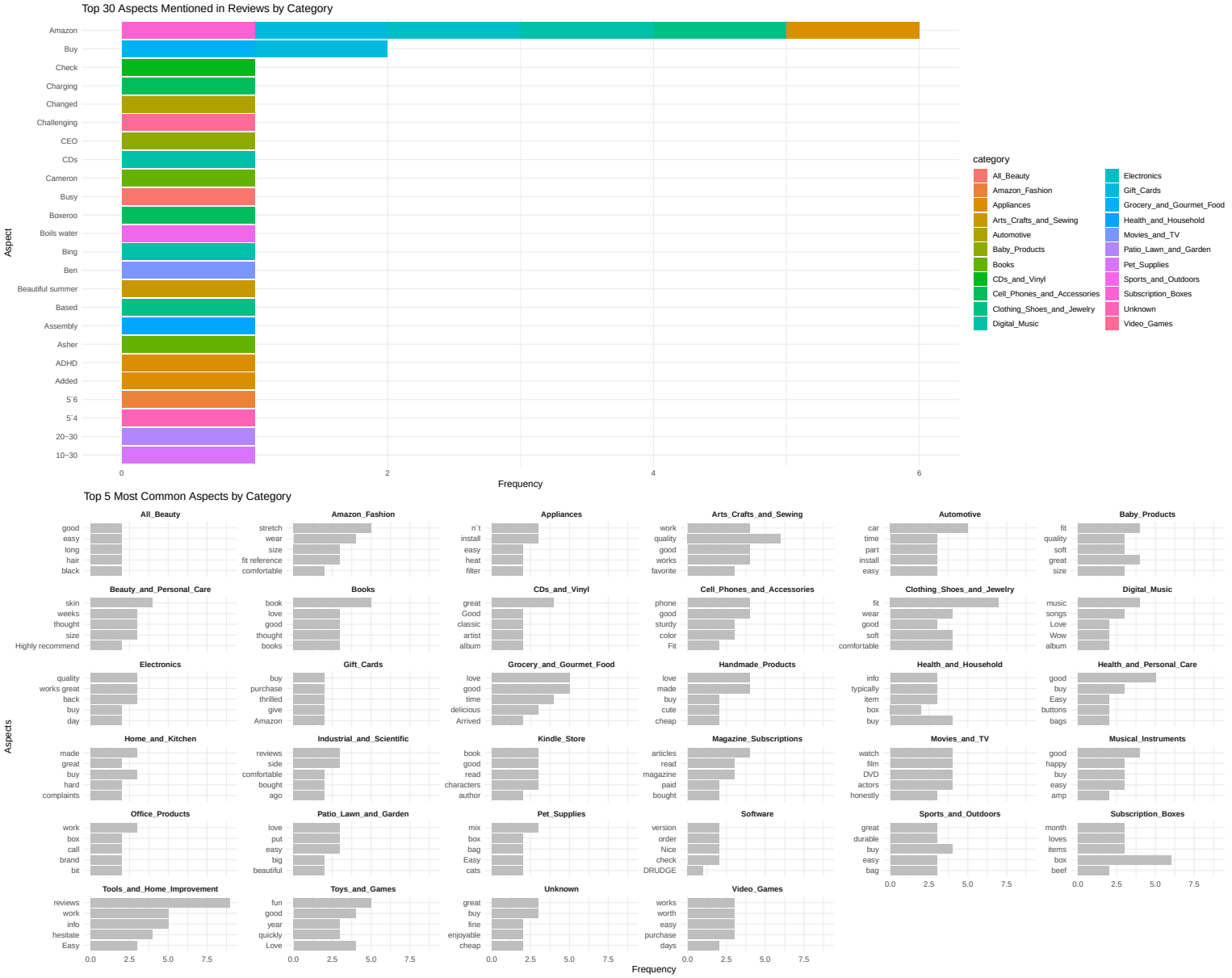
The sentiment scores were generally slightly skewed towards the positive side, with a total median sentiment score of 0.74. The sentiment scores were also quite consistent across categories, and only dropped below the global average on around half of the categories. The lowest median sentiment score was observed for the “Beauty and Personal Care” category at around 0.62 and the highest for the “Digital Music” category at around 0.85.

RQ2: Subjectivity Analysis



Next, to inspect the subjectivity of each category, we used a boxplot to visualize the distribution of subjectivity scores per category and a horizontal bar plot to show the median subjectivity score. These two plots are essentially identical, given that a box-plot already marks the median, but for the sake of comparibility with the previous plot we decided to use a bar plot as well. We can observe a view highly objective outliers, particularly in the "Home and Kitchen" and "Digital Music" categories despite having the highest median subjectivity scores. The least median subjectivity scores were counterintuitively assigned to "Musical Instruments", "Movies and TV" and "Tools and Home Improvement". This counterintuitive distribution of subjectivity scores suggests that (1) either the model may not be accurately capturing the subjectivity of the reviews or that (2) the reviewers use a different language or style in their reviews independent of their emotional attachment to the product.

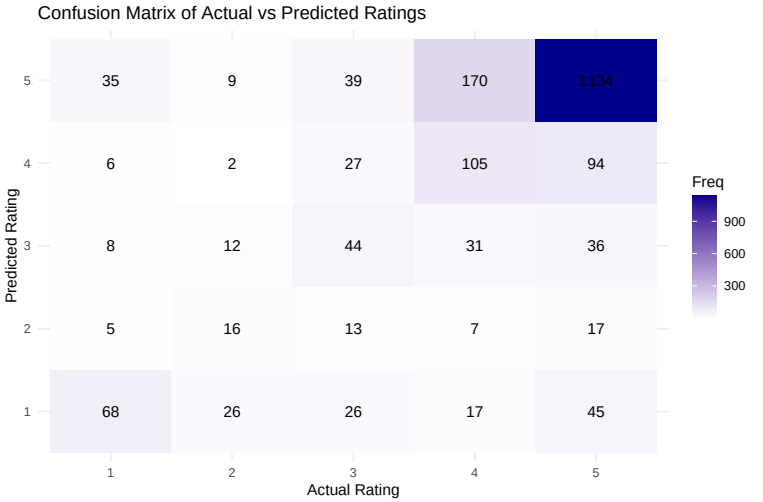
RQ3: Aspect Extraction

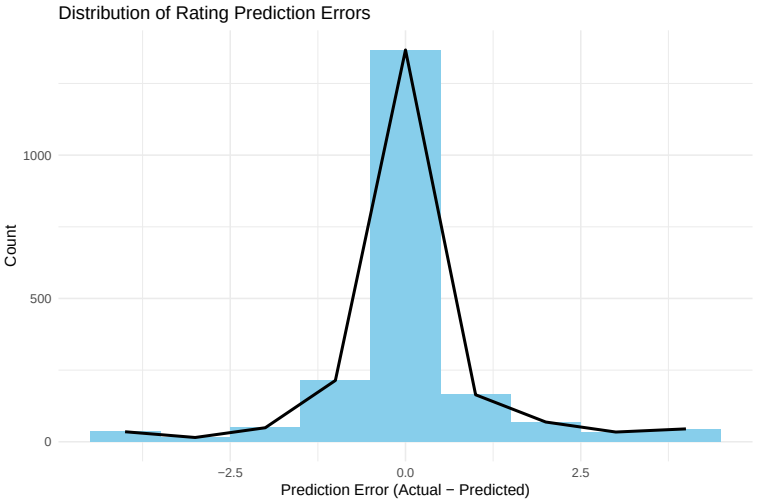


The keyword extraction algorithm did not perform as well as expected, but it was able to extract some of the most common aspects mentioned in the

reviews. The top 30 aspects mentioned in reviews by category are visualized in the first horizontal bar-plot and each total sum is further decomposed into the occurences per category. The second plot shows the top 5 most common aspects mentioned in reviews by category. The most common aspects mentioned in reviews across all categories were “Amazon”, “Buy” and “Check”. The top 5 aspects per individual category were more diverse but as expected, related to the domain of the category. For example, the “Magazine Subscriptions” category had aspects like “articles”, “read”, “magazine”, “paid”, “bought” which all relate to the content of the magazine. Overall the aspect extraction task was not as successful and insightful as we had hoped, but it was nonetheless an interesting technical challenge.

RQ4: Star Rating Prediction





Actual	Predicted	Count
Positive	Positive	16
Positive	Negative	26
Negative	Positive	5
Negative	Negative	68

Finally, we evaluated the performance of our star rating prediction model. The confusion matrix shows the distribution of actual vs predicted ratings, with the diagonal representing correct predictions. The model achieved an accuracy of 0.68, with most errors occurring in the misclassification of 4-star and 1-star reviews. The distribution of rating prediction errors shows a slight skew towards underestimating the rating, with more negative prediction errors than positive ones. This suggests that the model may be more conservative in its predictions, tending to assign lower ratings than the actual reviews. However these metrics have to be taken with a grain of salt as we don’t have any insights into the training process of the model and it might as well be overfitting.

In conclusion, we successfully answered the research questions using a combination of NLP techniques and data analysis. The results provided valuable insights into review sentiment, subjectivity, aspect extraction and star rating prediction, highlighting the complexities and nuances of consumer feedback on Amazon.