Automated Customer Reviews — Project Report

Bruno Pulheze - Ironhack DS & AI - Week 36 Project

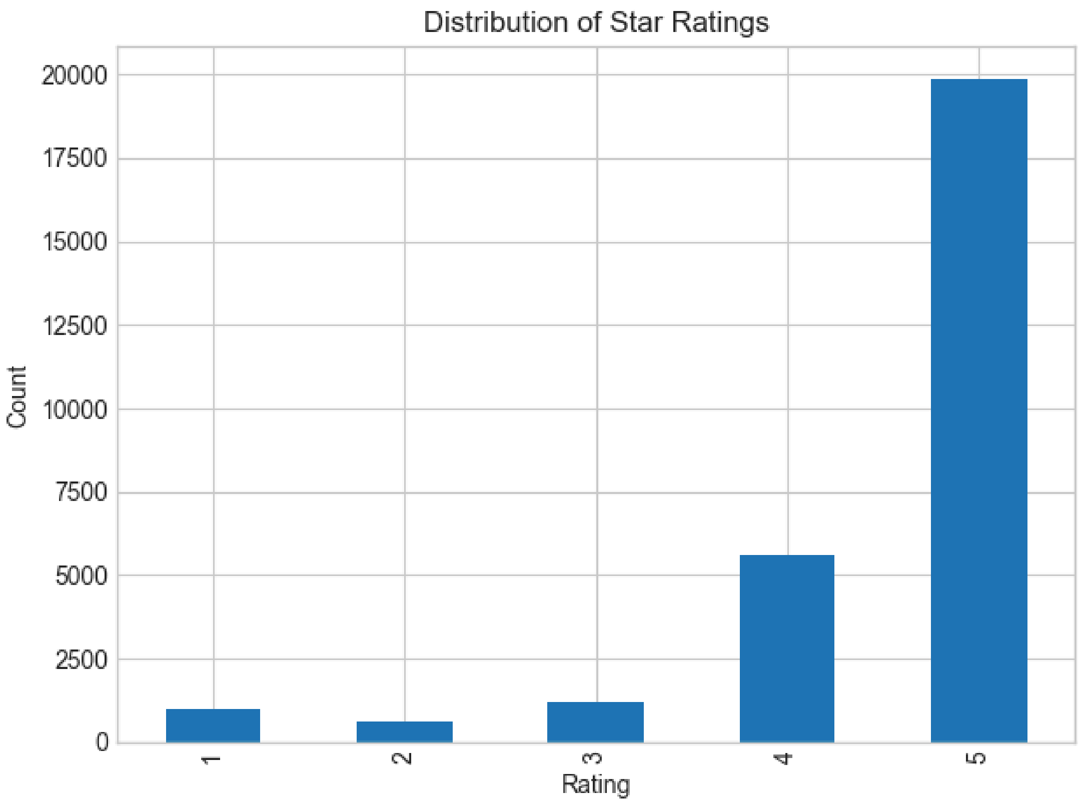
Deliverable: PDF report documenting approach, results, and analysis for the four project notebooks and the TypeScript + FastAPI UI.

Executive Summary

- Objective: build an end-to-end pipeline that preprocesses Amazon product reviews, classifies sentiment, clusters products into meta-categories, and generates recommendation articles.
- Components: four notebooks (`1_data_collection_preprocessing` , `2_review_classification` , `3_product_clustering` , `4_review_summarization`) plus a TypeScript + FastAPI GUI.
- This report documents the approach taken, features used, results obtained, and next steps.

Data & Key Features

- Source: `Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products_May19.csv` .
- Columns kept from raw data: `name` , `primaryCategories` , `reviews.text` , `reviews.rating` , `reviews.title` .
- Derived features:
  - `clean_text` : preprocessed review text (lowercased, HTML removed, stopwords removed, lemmatized).
  - `sentiment` : mapped from star ratings (1–2 → Negative, 3 → Neutral, 4–5 → Positive).
  - `review_length` : word count of `clean_text` .
  - Sentence-transformer embeddings of **product names** (generated in Notebook 3 for clustering).
  - Product-level aggregates (`avg_rating` , `num_reviews` , `positive_pct` , `negative_pct`) computed in Notebook 4 for summarization.



Notebook 1 — Data Collection & Preprocessing

Approach

- Loaded raw CSV, dropped duplicate rows, and dropped rows with missing `reviews.text` or `reviews.rating`.
- Subset to five columns: `name`, `primaryCategories`, `reviews.text`, `reviews.rating`, `reviews.title`.
- Text preprocessing (`preprocess_text` function applied to `reviews.text`):
  - Lowercasing.
  - HTML tag removal (`re.sub`).
  - Non-alphabetic character removal.
  - NLTK word tokenization.
  - Stopword removal (English stopwords).
  - Lemmatization (`WordNetLemmatizer`).
  - Filtered tokens with length > 2.
  - Result stored in `clean_text` column; rows with empty `clean_text` dropped.
- Sentiment mapping: `reviews.rating` ≤ 2 → Negative, 3 → Neutral, 4–5 → Positive; stored in `sentiment` column.
- Computed `review_length` as word count of `clean_text`.

Results

- Cleaned dataset saved as `data/preprocessed_reviews.csv`.

- Visualizations produced: sentiment class distribution bar chart, star rating distribution bar chart, review length histogram, and word clouds per sentiment class.

Analysis & Notes

- Class imbalance observed (positive class dominant). Class-weighting and oversampling were NOT applied; listed as future work.

Notebook 2 — Review Classification

Approach

- Fine-tuned `distilbert-base-uncased` for 3-class sentiment classification (Negative / Neutral / Positive).
- Label encoding: `{'Negative': 0, 'Neutral': 1, 'Positive': 2}` .
- Data split: `train_test_split` with `test_size=0.2` , `random_state=42` , stratified by label.
- Tokenization: `AutoTokenizer` with `padding='max_length'` , `truncation=True` , `max_length=256` .
- Converted to HuggingFace `Dataset` objects.

Training configuration

- `TrainingArguments` : `num_train_epochs=3` , `per_device_train_batch_size=16` , `per_device_eval_batch_size=32` , `warmup_steps=100` , `weight_decay=0.01` , `eval_strategy='epoch'` , `save_strategy='epoch'` , `load_best_model_at_end=True` , `metric_for_best_model='f1'` .
- No explicit early stopping callback was used; `load_best_model_at_end=True` selects the best checkpoint across epochs.
- No class-weighting or oversampling was applied.

Features used

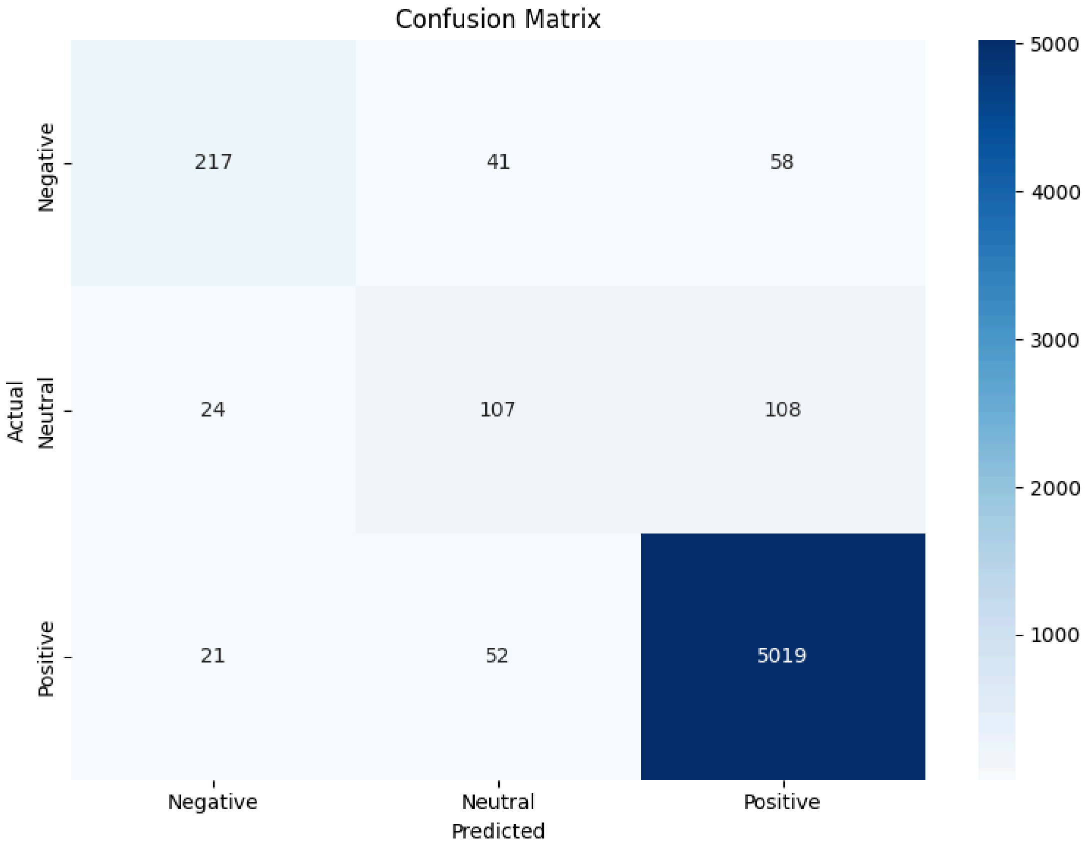
- `clean_text` column (from Notebook 1) as the sole input feature, tokenized for the transformer model.

Results

- Evaluation metrics computed: accuracy, weighted precision, recall, F1-score, and full classification report.
- Confusion matrix plotted as a heatmap.
- Model and tokenizer saved to `models/sentiment_classifier/` .

Class	Precision	Recall	F1
Positive	0.97	0.99	0.98

Class	Precision	Recall	F1
Neutral	0.54	0.45	0.49
Negative	0.83	0.69	0.75



Notebook 3 — Product Clustering

Approach

- Extracted unique product names from `preprocessed_reviews.csv`.
- Generated embeddings for each product name using `SentenceTransformer('all-MiniLM-L6-v2')`.
- Ran elbow method: KMeans for k in range(3, 10) with `n_init=10`, `random_state=42`; computed inertia and silhouette scores.
- Final clustering: `KMeans(n_clusters=5, random_state=42, n_init=10)` on the product name embeddings.
- Mapped cluster labels back to the full dataframe as `df['cluster']`.
- Manually assigned cluster names: `{0: 'Accessories', 1: 'Tablets', 2: 'Smart Assistants', 3: 'E-Readers', 4: 'Batteries and Chargers'}` → stored in `df['meta_category']`.

Features used

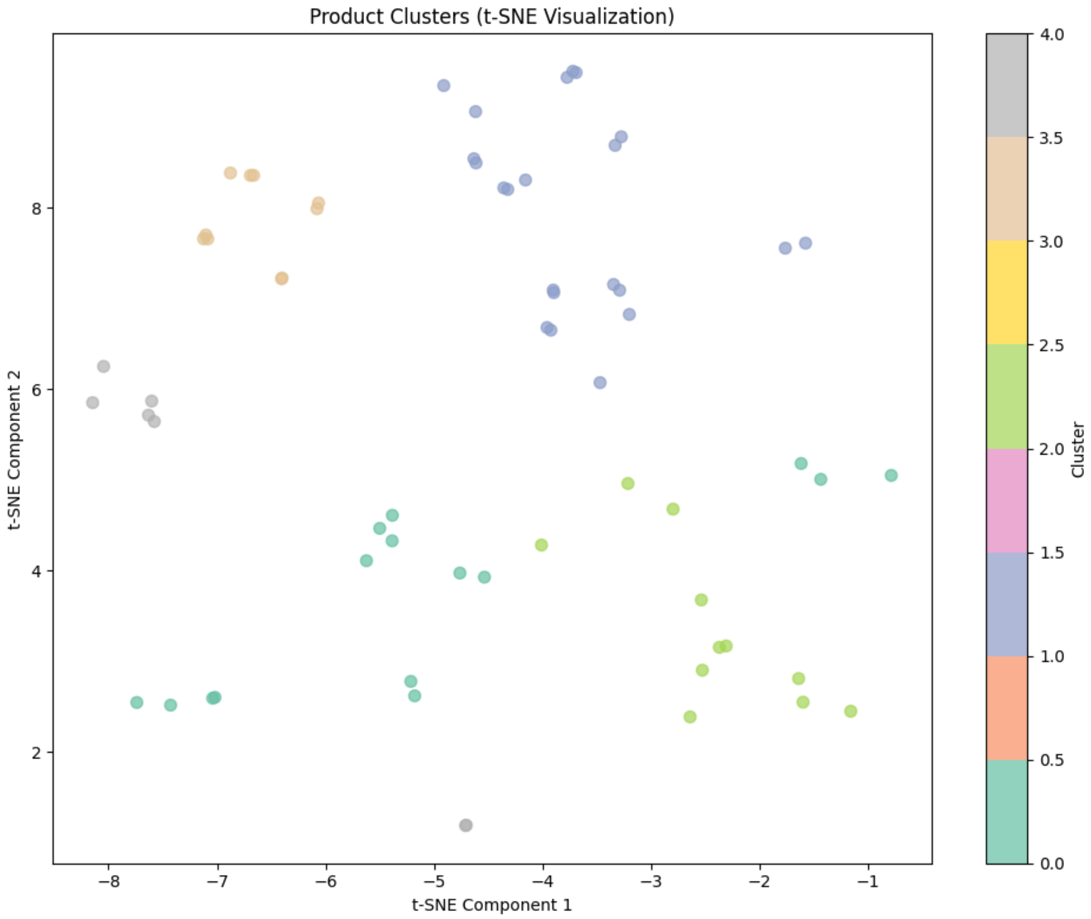
- Sentence-transformer embeddings of **product names** (not review text).

Results

- k=5 selected based on elbow and silhouette analysis.
- Visualizations: dual subplot (inertia + silhouette vs k), and a t-SNE 2D scatter plot colored by cluster label.
- Output saved as `data/clustered_reviews.csv`.

Analysis

- Clusters correspond to distinct product categories (tablets, e-readers, smart assistants, accessories, batteries/chargers).
- The accessories cluster was the most heterogeneous with many peripherals, likely due to more generic product names.
- The `meta_category` column feeds into Notebook 4 for per-category article generation.



Notebook 4 — Review Summarization

Approach

- Goal: generate a recommendation article for each meta-category, highlighting top products and main complaints.
- Loaded `data/clustered_reviews.csv` and aggregated per product within each category: computed `avg_rating`, `num_reviews`, `positive_pct`,

`negative_pct`.
- `get_category_insights()`: for each category, sorted products by `avg_rating` and selected the top 3 and worst product.
- `get_complaints()`: extracted up to 5 negative-sentiment reviews for the worst product.
- `build_prompt()`: constructed a text prompt containing top products, sample reviews (truncated to 200 chars), worst product, and its negative reviews.
- Summarization model: **OpenAI GPT-3.5-turbo** via `openai.chat.completions.create()` with `max_tokens=500`, `temperature=0.7`.
- API key loaded via `dotenv`.
- Note: `transformers` (`pipeline`, `AutoTokenizer`, `AutoModelForSeq2SeqLM`) were imported but **never used**; all summarization was performed exclusively via the OpenAI API.

Features used

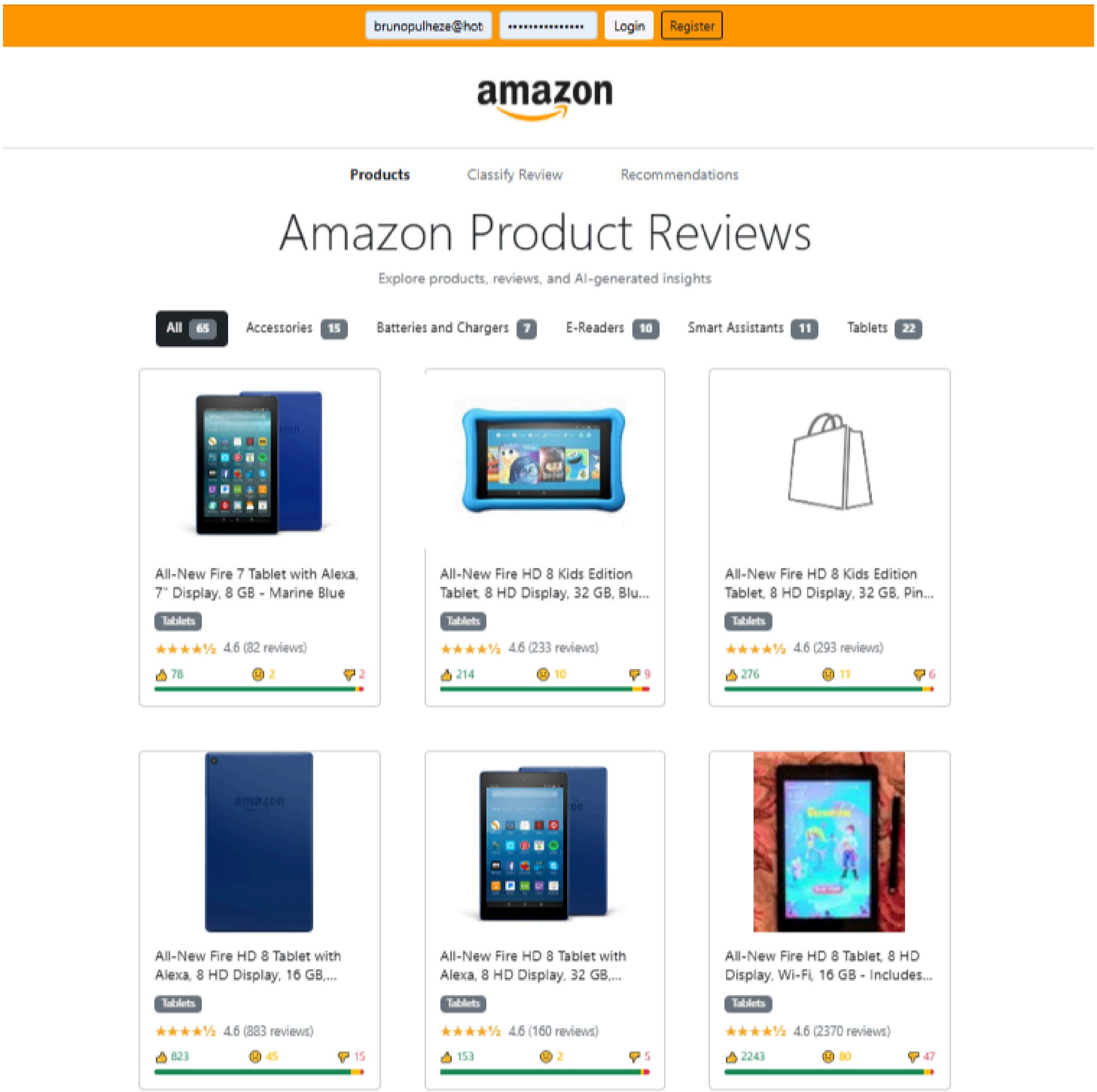
- `meta_category`, `name`, `reviews.rating`, `sentiment`, `clean_text` from `clustered_reviews.csv`.

Results

- Generated one recommendation article per meta-category.
- Articles saved to `data/recommendation_articles.txt` (separated by `=` dividers).
- No automated evaluation metrics (ROUGE, BERTScore, etc.) were computed.

TypeScript + FastAPI UI

- Frontend: TypeScript (Create React App) + Bootstrap providing product browsing, review lists, and presentation of generated articles.
- Backend: FastAPI exposing endpoints used by the frontend: `/products`, `/products/{id}`, `/categories`, `/categories/{name}/article`, and `/classify`.
- Integration notes: backend serves preprocessed CSVs from `data/`; the `POST /classify` endpoint uses a local fine-tuned model if present or a rule-based fallback.



Amazon Product Reviews

Explore products, reviews, and AI-generated insights

All-New Fire 7 Tablet with Alexa, 7" Display, 8 GB - Marine Blue

Tablets

★★★★½ 4.6 (82 reviews)

👍 78 😐 2 👎 2

All-New Fire HD 8 Kids Edition Tablet, 8 HD Display, 32 GB, Blu...

Tablets

★★★★½ 4.6 (233 reviews)

👍 214 😐 10 👎 9

All-New Fire HD 8 Kids Edition Tablet, 8 HD Display, 32 GB, Pin...

Tablets

★★★★½ 4.6 (293 reviews)

👍 276 😐 11 👎 6

All-New Fire HD 8 Tablet with Alexa, 8 HD Display, 16 GB,...

Tablets

★★★★½ 4.6 (883 reviews)

👍 823 😐 45 👎 15

All-New Fire HD 8 Tablet with Alexa, 8 HD Display, 32 GB,...

Tablets

★★★★½ 4.6 (160 reviews)

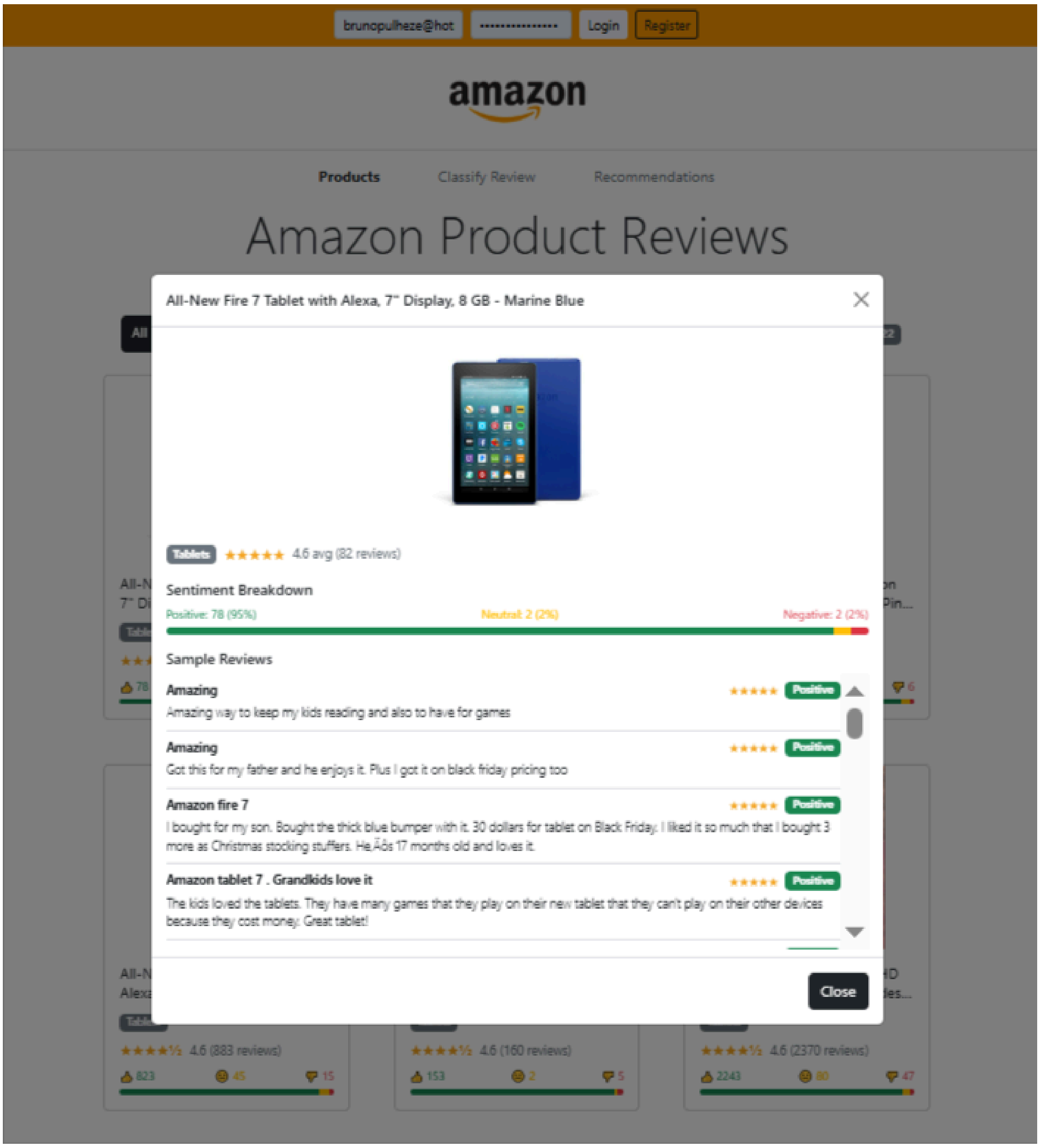
👍 153 😐 2 👎 5

All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Includes...

Tablets

★★★★½ 4.6 (2370 reviews)

👍 2243 😐 80 👎 47



Amazon Product Reviews

All-New Fire 7 Tablet with Alexa, 7" Display, 8 GB - Marine Blue

Tablets ★★★★★ 4.6 avg (82 reviews)

Sentiment Breakdown

Positive: 78 (95%) Neutral: 2 (2%) Negative: 2 (2%)

Sample Reviews

Amazing ★★★★★ Positive

Amazing way to keep my kids reading and also to have for games

Amazing ★★★★★ Positive

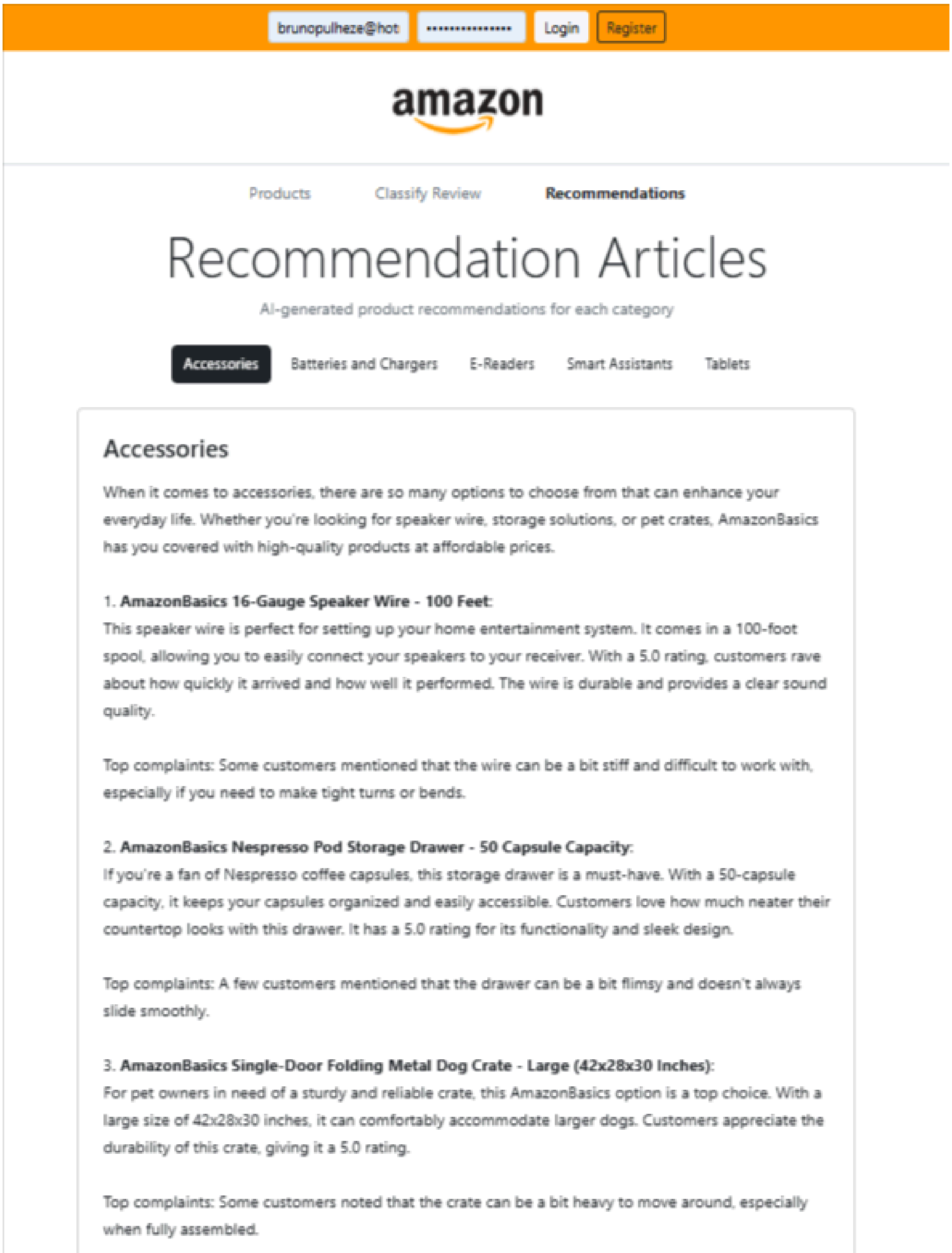
Got this for my father and he enjoys it. Plus I got it on black friday pricing too

Amazon fire 7 ★★★★★ Positive

I bought for my son. Bought the thick blue bumper with it. 30 dollars for tablet on Black Friday. I liked it so much that I bought 3 more as Christmas stocking stuffers. He‚Äôs 17 months old and loves it.

Amazon tablet 7 . Grandkids love it ★★★★★ Positive

The kids loved the tablets. They have many games that they play on their new tablet that they can't play on their other devices because they cost money. Great tablet!



Recommendation Articles

AI-generated product recommendations for each category

Accessories

When it comes to accessories, there are so many options to choose from that can enhance your everyday life. Whether you're looking for speaker wire, storage solutions, or pet crates, AmazonBasics has you covered with high-quality products at affordable prices.

1. **AmazonBasics 16-Gauge Speaker Wire - 100 Feet**:
This speaker wire is perfect for setting up your home entertainment system. It comes in a 100-foot spool, allowing you to easily connect your speakers to your receiver. With a 5.0 rating, customers rave about how quickly it arrived and how well it performed. The wire is durable and provides a clear sound quality.

Top complaints: Some customers mentioned that the wire can be a bit stiff and difficult to work with, especially if you need to make tight turns or bends.

2. **AmazonBasics Nespresso Pod Storage Drawer - 50 Capsule Capacity**:
If you're a fan of Nespresso coffee capsules, this storage drawer is a must-have. With a 50-capsule capacity, it keeps your capsules organized and easily accessible. Customers love how much neater their countertop looks with this drawer. It has a 5.0 rating for its functionality and sleek design.

Top complaints: A few customers mentioned that the drawer can be a bit flimsy and doesn't always slide smoothly.

3. **AmazonBasics Single-Door Folding Metal Dog Crate - Large (42x28x30 Inches)**:
For pet owners in need of a sturdy and reliable crate, this AmazonBasics option is a top choice. With a large size of 42x28x30 inches, it can comfortably accommodate larger dogs. Customers appreciate the durability of this crate, giving it a 5.0 rating.

Top complaints: Some customers noted that the crate can be a bit heavy to move around, especially when fully assembled.

Appendix & Next Steps

- Future work:
  - Address class imbalance (class-weighting or oversampling for the classifier).
  - Add automated evaluation metrics (ROUGE / BERTScore) for the generated articles.
  - Explore clustering on review-text embeddings in addition to product-name embeddings.
  - Deploy frontend (Vercel / GitHub Pages) and backend (cloud hosting).