BỘ GIAO THÔNG VẬN TẢI
TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI THÀNH PHỐ HỒ CHÍ MINH

DATA MINNG
BÁO CÁO CUỐI KỲ

PHÂN TÍCH NỘI DUNG TRÒ CHUYỆN TRÍCH XUẤT VẤN ĐỀ PHỔ BIẾN VỀ TÀI SẢN

Giảng Viên Hướng Dẫn: **Trần Thế Vinh**

Họ Và Tên Sinh Viên Thực Hiện Và MSSV:
Phạm Hoàng Thiện - 051205000064
Ngô Minh Khang - 086250511340
Nguyễn Văn Mạnh - 027205000040
Nguyễn Văn Quang - 038205004237

Thành phố Hồ Chí Minh, ngày .. tháng .. năm 2024

Table of Contents

1. Analyzing Public Discussions for Product Insights

PIPELINE : Reddit + Tiki collection → Parquet normalization → EDA → Topic modeling (TF-IDF → SVD → KMeans) + Aspect sentiment

1.1. Executive Overview

- We mine 16 hardware-centric Reddit communities (plus complementary Tiki reviews) to surface recurring product issues, pros/cons, and topic trends.
- Arctic Shift harvesting bypasses API caps to deliver multi-year coverage stored as harmonized Parquet.
- A CLI (`run_pipeline.py`) turns filtered corpora into TF-IDF/SVD features, KMeans clusters, and aspect-level sentiment artifacts.
- Planned transformer sentiment fine-tuning was documented but descoped after cost/benefit analysis; VADER backs the delivered pipeline.
- Outputs include reusable joblibs/JSON summaries, PNG dashboards, and presentation assets in `extra/`.

1.2. Data Sources & Collection

1.2.1. Reddit & Tiki rationale

- Reddit supplies rich, text-first, community-moderated threads; Tiki adds verified purchase feedback from the Vietnamese market.
- Facebook and TikTok were deprioritized due to API scarcity, bot noise, and media-first formats.
- Subreddit mix spans laptops, phones, audio, smart home, photography, PC building, and ergonomics—capturing both enthusiast and troubleshooting conversations.

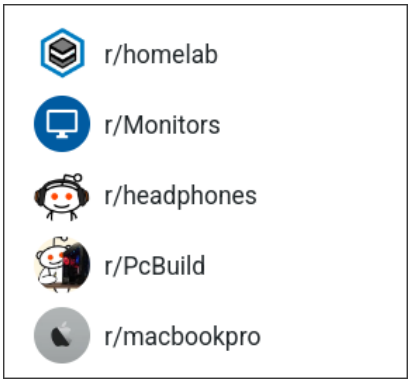

Figure 1: Representative subreddit slate spanning homelab builds, monitors, audiophile hubs, and Apple ecosystems.

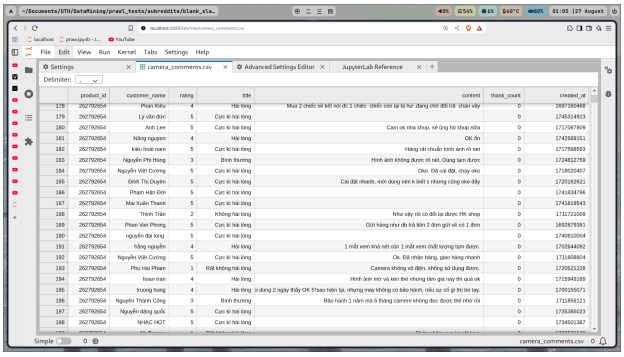

Figure 2: Sample of Tiki camera reviews — structured ratings with Vietnamese free-text complement Reddit narratives.

1.2.2. Arctic Shift workflow

- Historical dumps (Academic Torrents) feed the Arctic Shift collector, eliminating PRAW's 1 000 post ceiling per subreddit.
- Tiki review dumps augment Reddit for cross-source validation when available.
- Data is normalized via Polars/Nushell from 32 JSONL exports to `posts.parquet` and `comments.parquet`, enforcing consistent schemas.

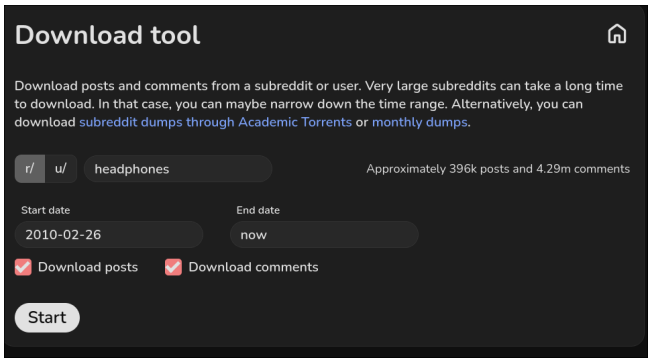

Figure 3: Arctic Shift download tool set to pull the full r/headphones history — bypassing API caps for longitudinal coverage.

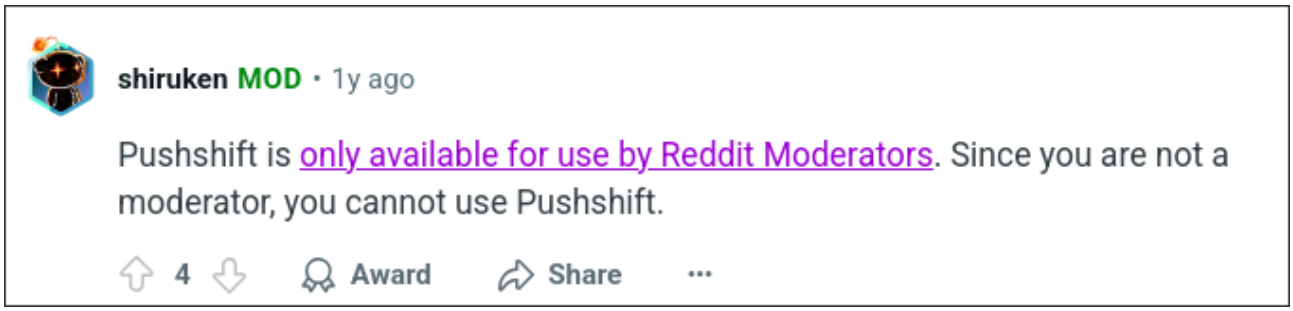

Figure 4: Pushshift access is now moderator-only, reinforcing the need for self-hosted archival strategies.

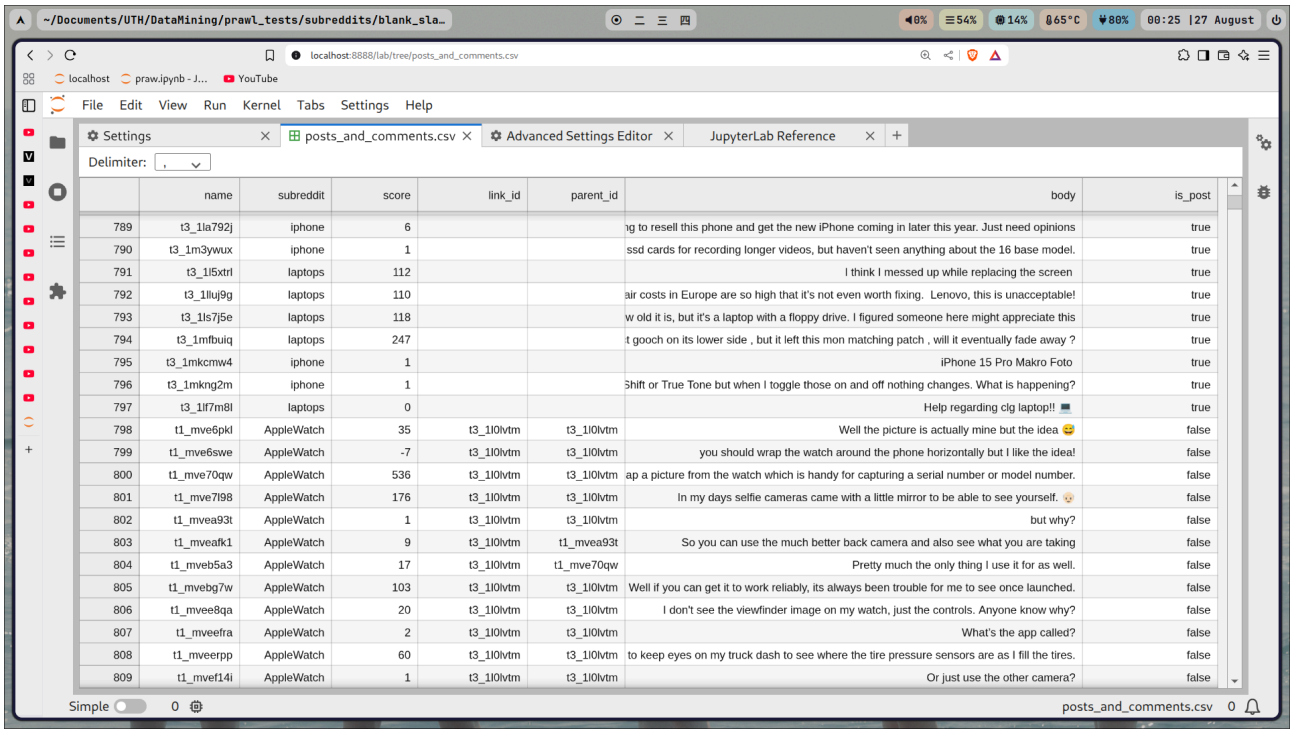

Figure 5: Unified posts-and-comments Parquet preview showing normalized identifiers prior to modeling.

1.2.3. Corpus snapshot

Split	Rows	Columns	Unique subs
Posts	134121	27	16
Comments	1300190	16	16

1.3. Data Engineering & Integration

- Filtering uses case-insensitive alias matching (e.g. “hd600”, “fiio ft1”), with optional subreddit restriction.
- Brand/alias tokens are added to the stopword list to avoid dominating TF-IDF features.
- Post/comment parity is preserved by harmonizing key identifiers (`name`, `subreddit`, `link_id`, `parent_id`) before concatenation.
- `run_pipeline.py` logs reproducibility metadata (command-line, package versions) and writes all intermediate artifacts to the specified output directory.

1.4. Exploratory Data Analysis

- EDA assets reside under `eda/` (PNG + CSV) for quick reuse in slides and dashboards.

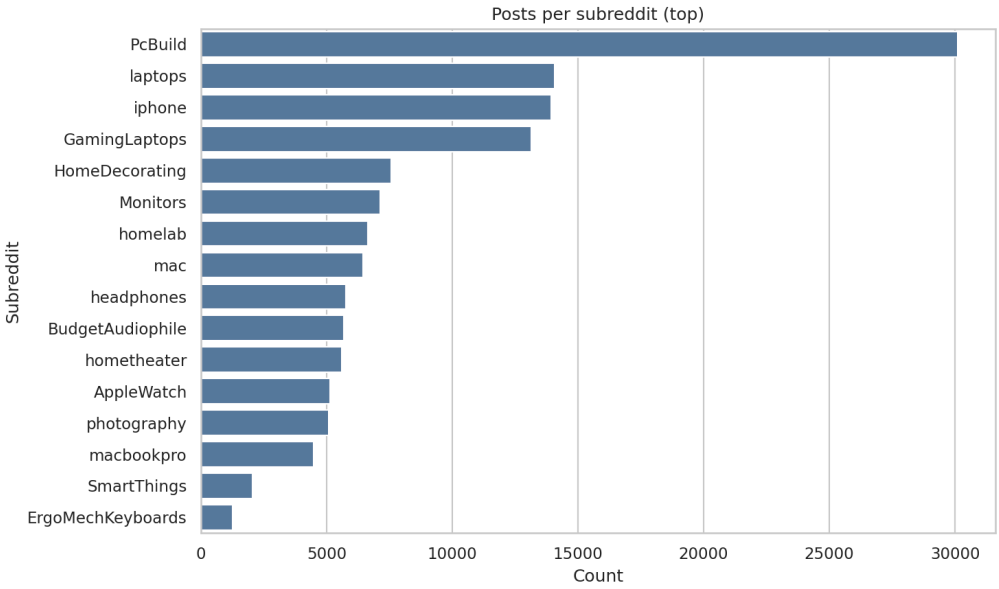

Figure 6: Post counts by subreddit — PcBuild, iPhone, and GamingLaptops dominate volume.

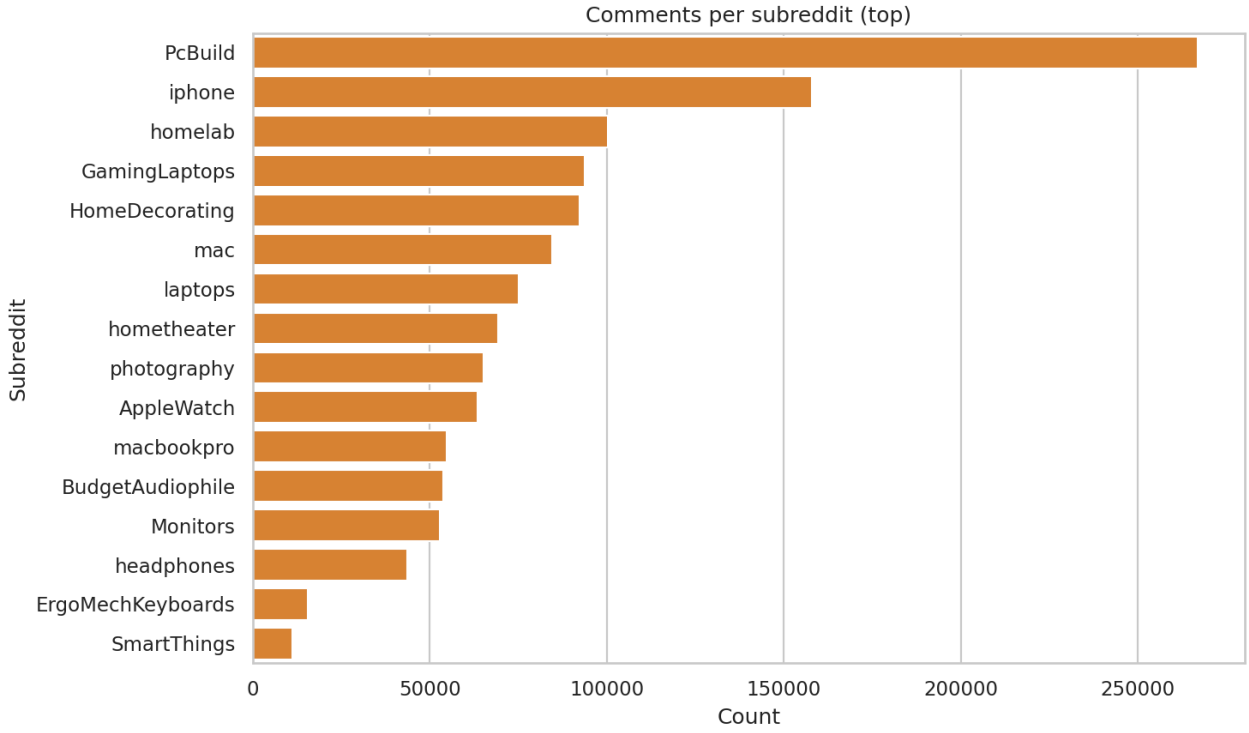

Figure 7: Comment counts: troubleshooting-heavy communities (PcBuild, homelab) lead engagement.

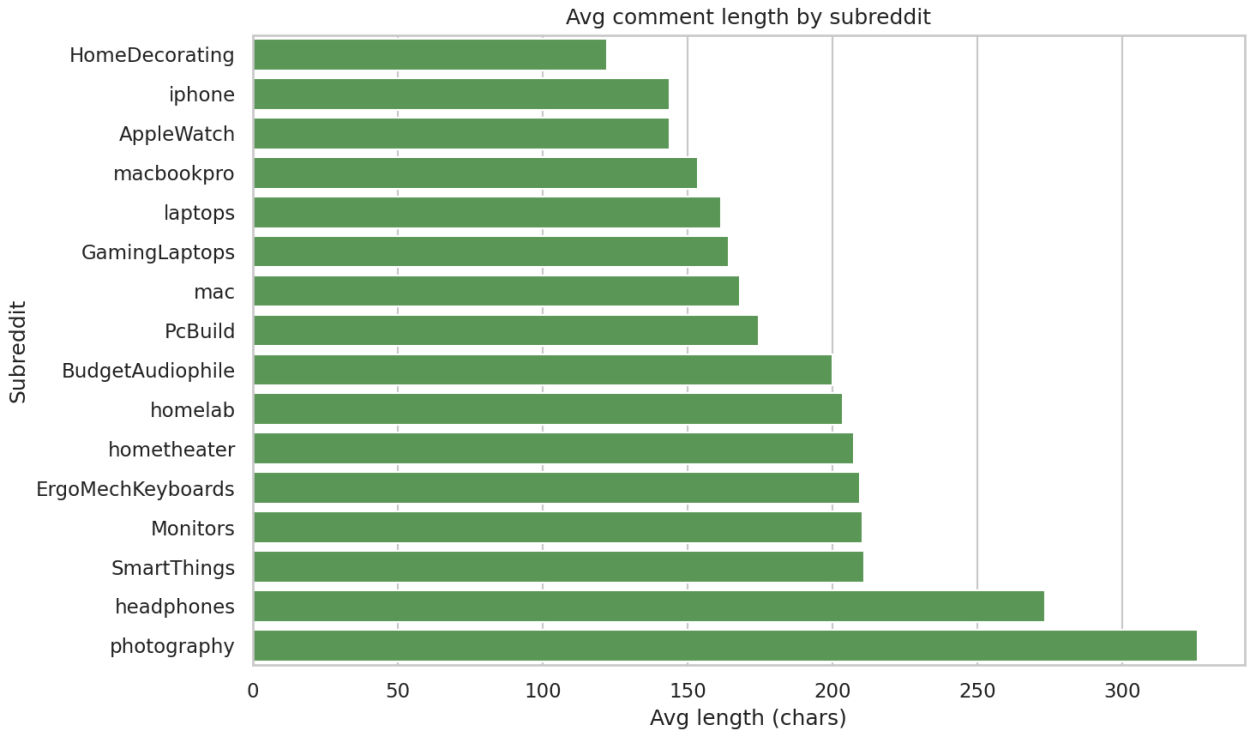

Figure 8: Average comment length highlights deep technical discussions in r/macbookpro and r/AppleWatch.

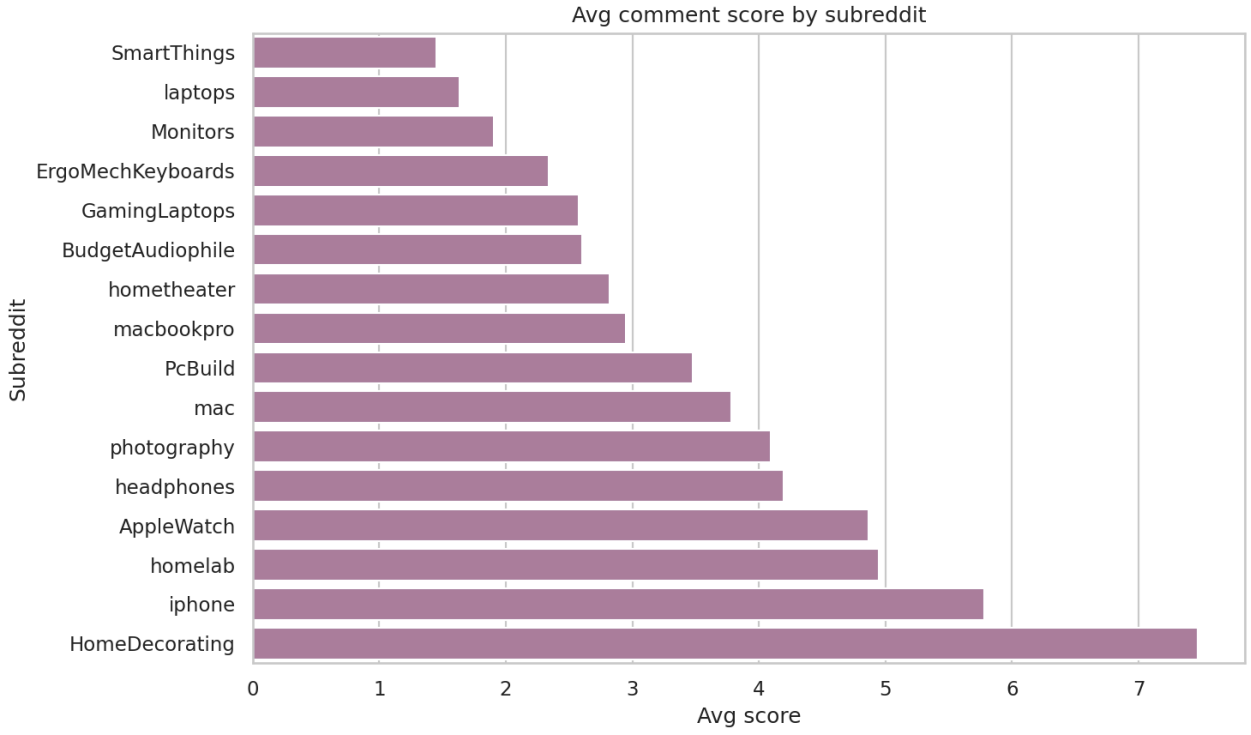

Figure 9: Mean comment scores remain low overall (≤2.5), reinforcing the need for textual insight beyond karma.

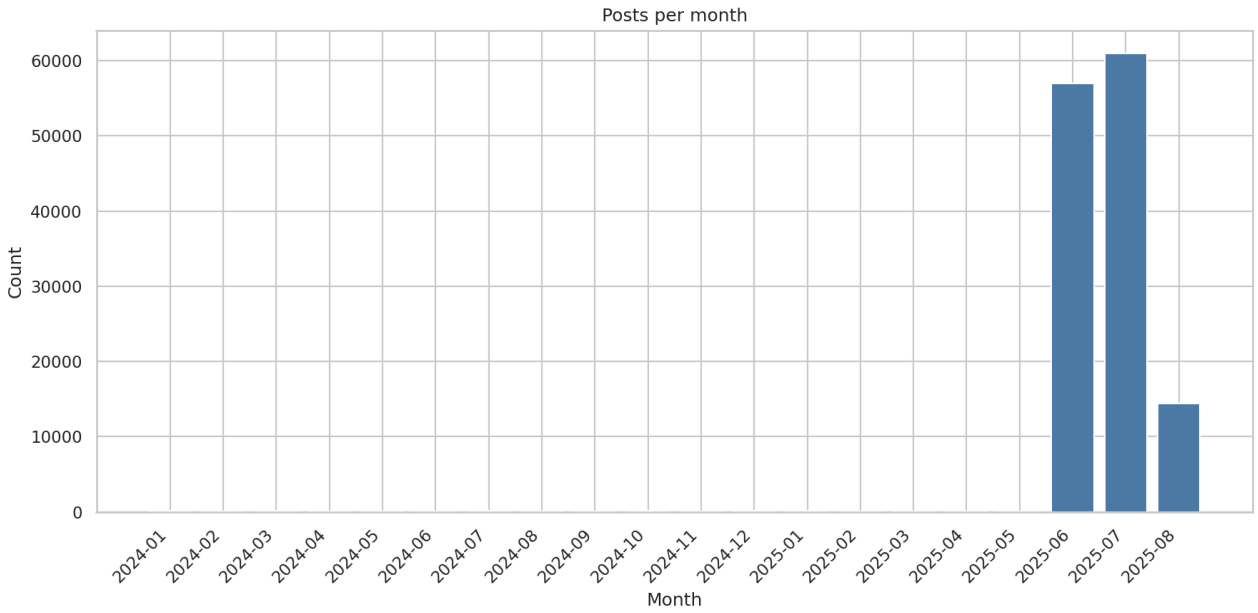

Figure 10: Monthly posting cadence detects seasonal hardware launch spikes.

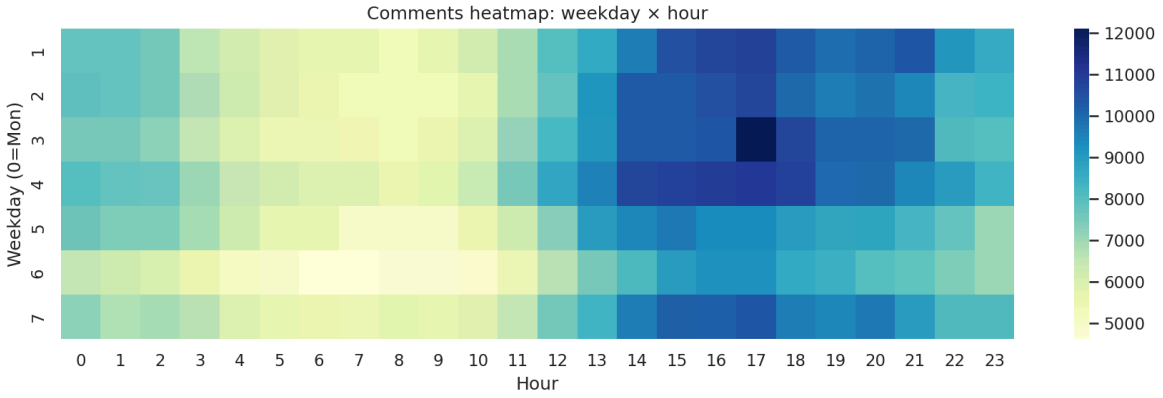

Figure 11: Weekly/hourly heatmap of activity — evenings and weekends drive most conversations.

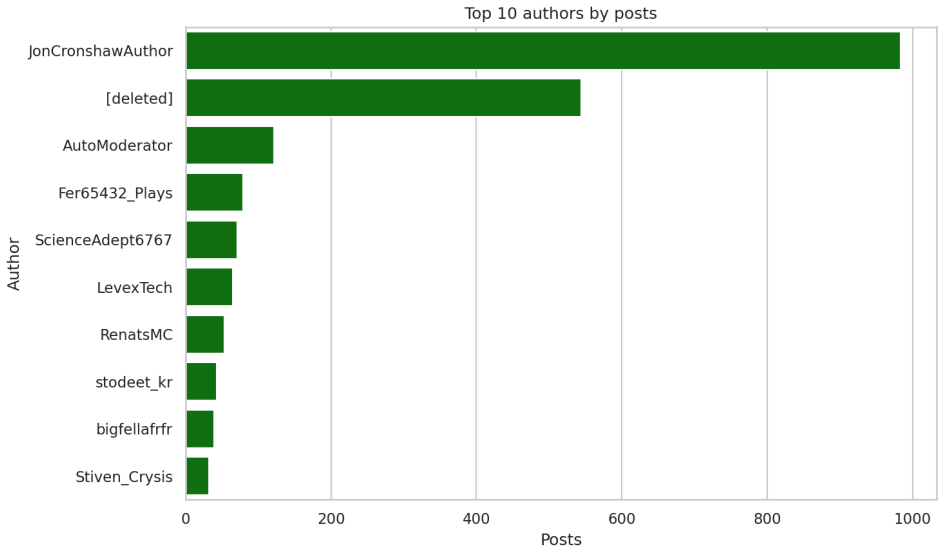

Figure 12: Top post authors — flags power users who shape each hardware narrative.

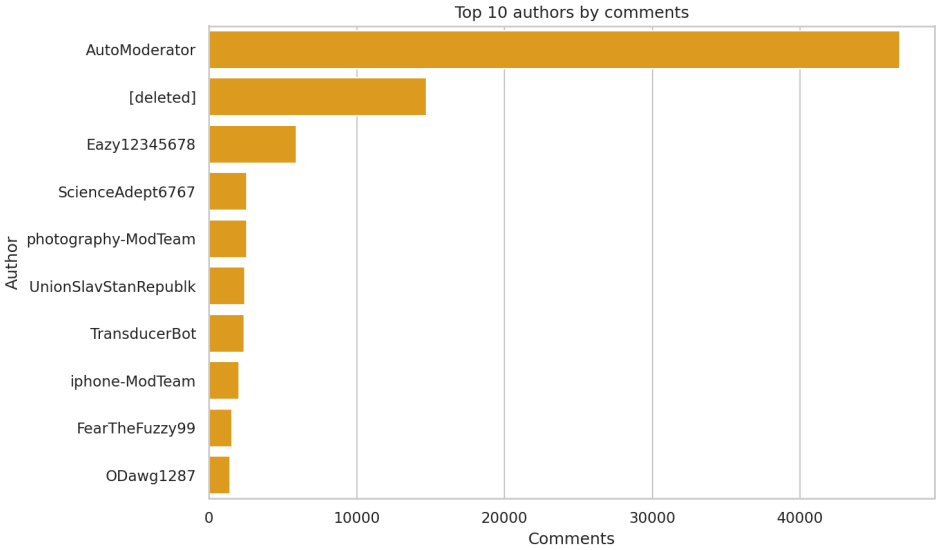

Figure 13: Top comment authors — identifies moderation-heavy and support-oriented contributors.

Figure 14: Title word cloud surfaces dominant product classes and troubleshooting topics.

Figure 15: Comment word cloud highlights recurring sentiment cues and component-level jargon.

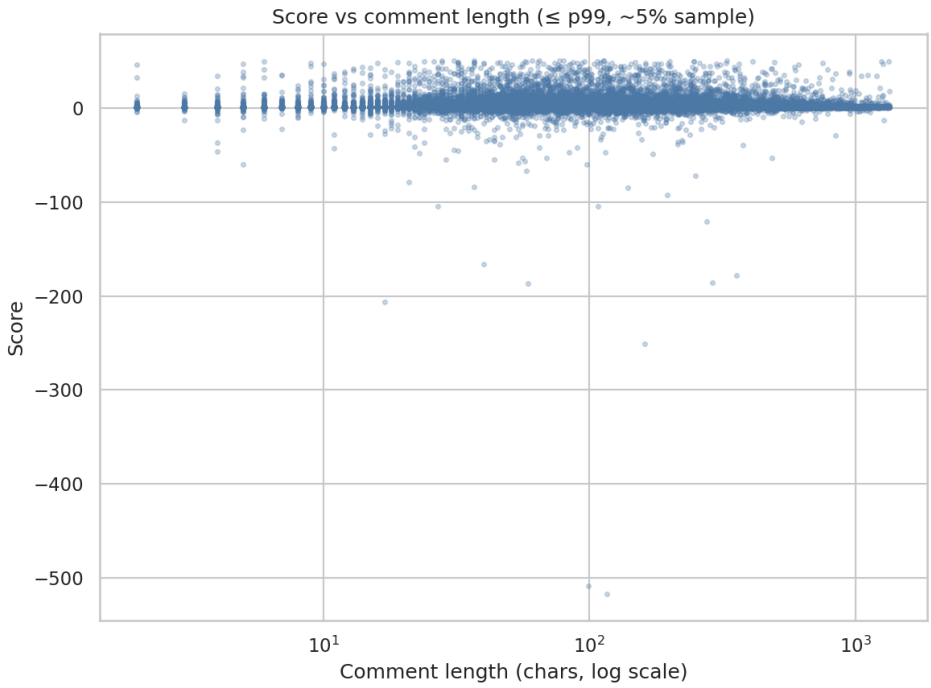

Figure 16: Comment score versus length — longer replies trend toward higher karma among troubleshooting threads.

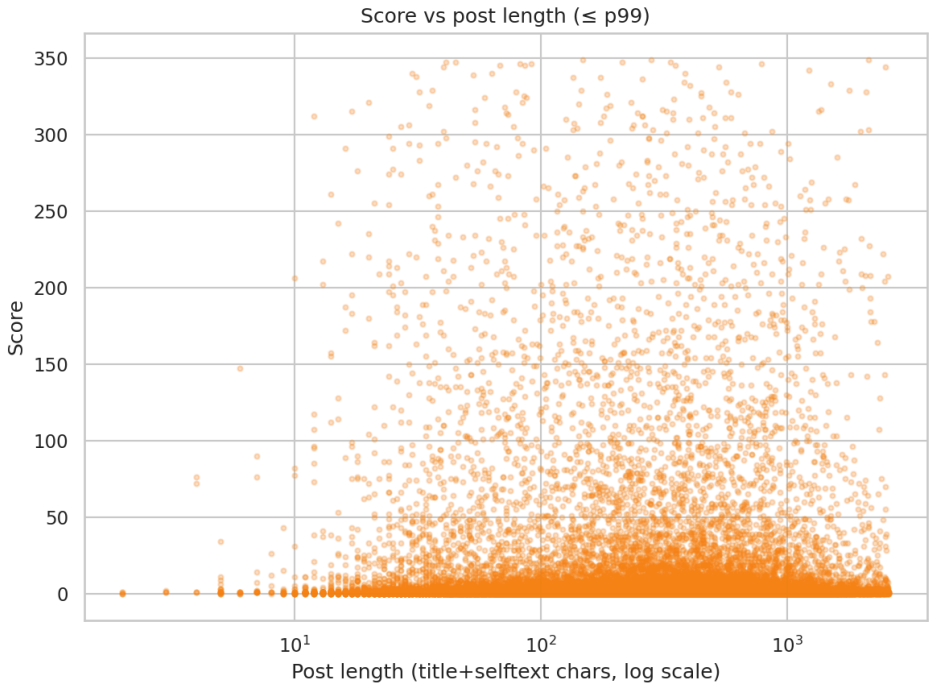

Figure 17: Post score versus selftext length — concise launch rumors and detailed reviews both find audiences.

- Additional visuals: word clouds, score-versus-length scatterplots, author leaderboards, and CSV exports for deeper analysis (`eda/*.csv`).

1.5. Sentiment Analysis — Initial Plan & Decision

- Intended flow: fine-tune `lxyuan/distilbert-base-multilingual-cased-sentiments-student` on a stratified, Gemini-assisted label set (target ≥5 k annotations for calibration and evaluation).
- Obstacles: GPU rental costs, rate-limited labeling throughput, and limited marginal benefit vs. leveraging unsupervised topic insights.
- Delivered approach: VADER (rule-based) for sentence polarity during aspect aggregation, retaining the documented plan for future enhancement.

Figure 18: Deferred fine-tuning target `lxyuan/distilbert-base-multilingual` — parked due to GPU and labeling constraints.

1.6. Topic Modeling & Aspect Pipeline

The CLI executes a deterministic loop—cleaning → TF-IDF → SVD → KMeans → aspect sentiment—balancing automation for rapid iteration with explicit flags for power users who need to override defaults.

1.6.1. Pre-processing

- Normalize Unicode (NFC), strip URLs/code fences, lowercase, remove non-alphabetic characters while preserving apostrophes.
- Token filtering retains words longer than two characters and excludes expanded stopwords (including brand aliases).
- Aliases/subreddit filters yield a focused corpus per product run.

1.6.2. Feature Selection & Dimensionality Reduction

- TF-IDF (`ngram_range = 1–2`) with adaptive `min_df` (3 if N ≥ 300 else 2) and `max_df = 0.95`; vocabulary capped at `min(2000, 3N)` when auto mode is used (`run_pipeline.py:545`).
- TruncatedSVD clamps components to `min(round(0.25N), 200)` with a floor of 50, logging explained variance for transparency (`run_pipeline.py:560`).
- Guardrails clamp SVD to `min(V - 1, N - 1)` and expose overrides via

`--min-df`, `--max-feat`, and `--svd`, giving small corpora a safe path without losing expert control.

- Empty vocabulary protection automatically retries TF-IDF with `min_df = 1`, so edge-case alias runs still generate features instead of failing fast.

1.6.3. Clustering workflow

- KMeans sweep across `k_min..k_max` (default 3–8) guided by silhouette score; `method.json` captures the search bounds, seeds, and best score so downstream analysts know why a `k` was picked.
- Cluster diagnostics combine TF-IDF means (`top_terms_by_cluster`) with centroid-nearest exemplars (`representatives_by_cluster`) to surface interpretable themes and quote-ready posts.
- Optional `--save-plots` renders bar charts of cluster sizes for quick QA before sharing results.

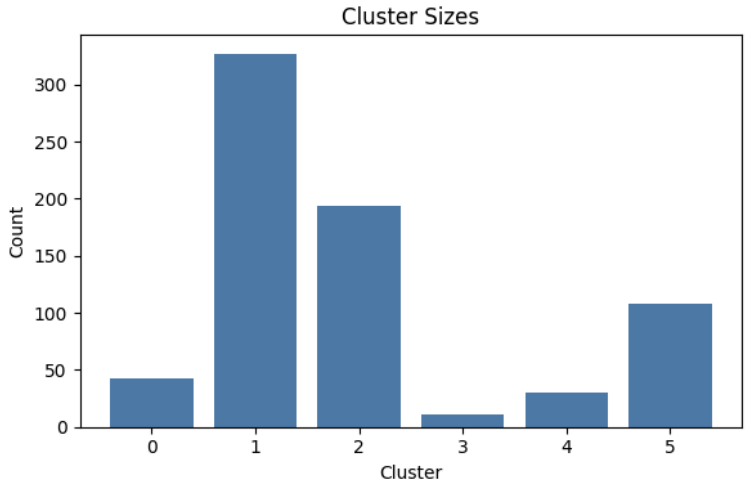

Figure 19: Cluster size distribution (Fiio FT1 example) after SVD-reduced KMeans.

1.6.4. Aspect extraction & sentiment fusion

- Subreddit membership auto-selects aspect seed dictionaries (battery, thermals, comfort, etc.); optional TF-IDF expansion adds corpus-specific terms.
- Sentence-level pattern matching assigns hits; VADER polarity populates mentions, average sentiment, and curated quote lists.
- `assignments.jsonl` marries each document's cluster label with its aspect sentiment profile, while `aspect_summary*.json` filters out noisy aspects below an adaptive frequency threshold (`max(5, round(0.01N))`).
- Analysts can deepen coverage with `--expand-seeds`, which injects top TF-IDF terms (minus brand aliases) into the aspect dictionaries for more organic phrasing.
- Quote buffers only keep sentences past ±0.05 compound polarity and cap the stored examples at five per aspect, balancing signal and brevity.

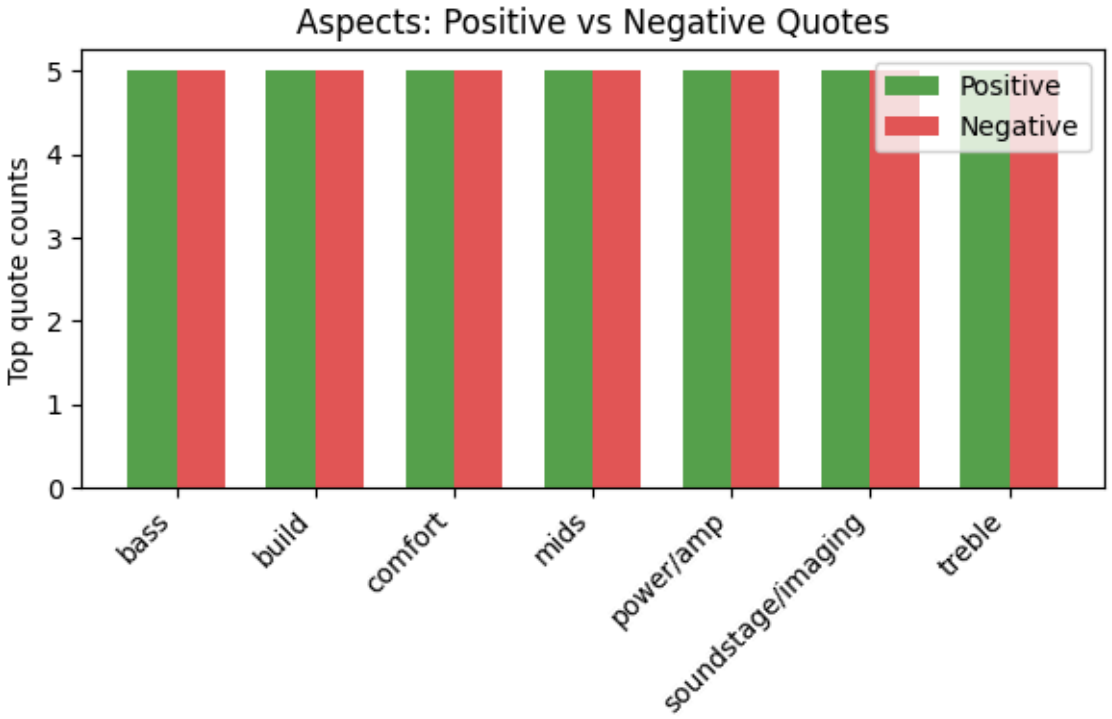

Figure 20: Aspect sentiment summary (Fiio FT1 run): positive vs. negative quote counts.

1.6.5. Key artifacts

- `tfidf_vectorizer.joblib`, `svd_model.joblib`, `svd_explained_variance.json` — reproducible feature pipeline.
- `kmeans_clusters.json` — cluster sizes, TF-IDF top terms, and centroid-nearest representative posts in one package; `method.json` logs the selected hyper-parameters.
- `assignments.jsonl`, `aspect_summary.json`, `aspect_summary_by_subreddit.json` — document-level clusters with aspect sentiment, filtered by the adaptive frequency floor.
- Optional PNGs: `cluster_sizes.png`, `aspects_pos_neg.png` for quick sanity checks.

1.7. Deliverables & Usage

- CLI entry point: `run_pipeline.py --data <parquet> --product <name> --aliases <comma-separated> --out <dir> [options]`.
- Configuration knobs control vectorization (`--min-df`, `--max-df`, `--max-feat`, `--ngram-*`), dimensionality (`--svd`), clustering (`--method`, `--k-min`, `--k-max`), and aspect behavior (`--aspect-category`, `--expand-seeds`, `--min-aspect-freq`).
- Artifacts for Fiio FT1, Sennheiser HD600, and Sony WF-1000XM4 are archived under `extra/artifacts_ft1`, `extra/artifacts_hd600`, and `extra/artifacts_m4`.

1.8. Limitations & Future Work

- Sentiment remains rule-based; executing the documented fine-tuning plan would handle sarcasm and domain-specific jargon better.
- The NMF/LDA branch is scaffolded but not productized—adding it would support overlapping topic mixtures.
- Broader Tiki integration and multilingual expansion would capture non-English sentiment more faithfully.
- Additional evaluation (topic coherence, stability under resampling, human-in-the-loop validation) should accompany the next iteration.

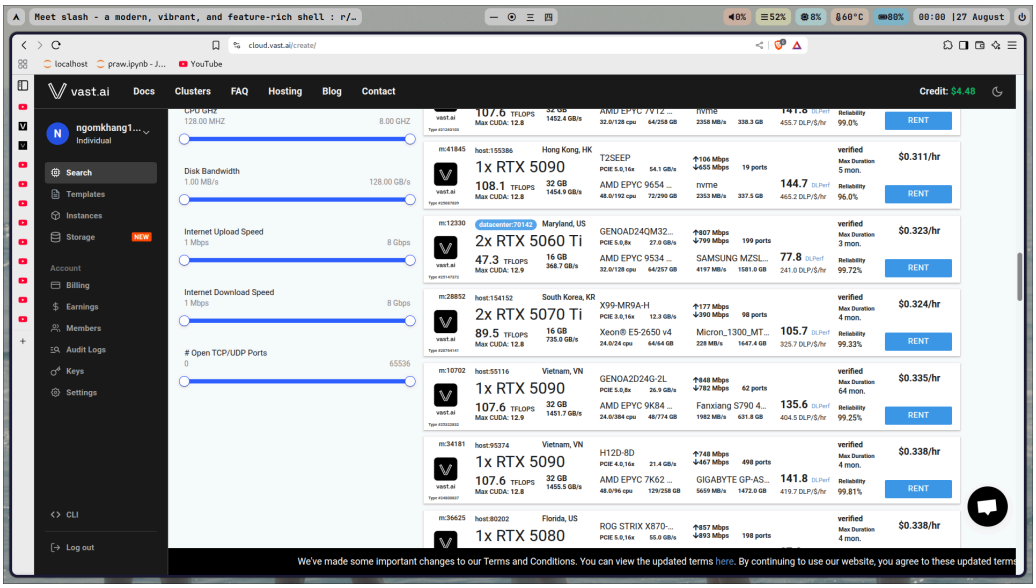

Figure 21: Vast.ai GPU marketplace snapshot — projected fine-tuning spend that led to deferring transformer training.

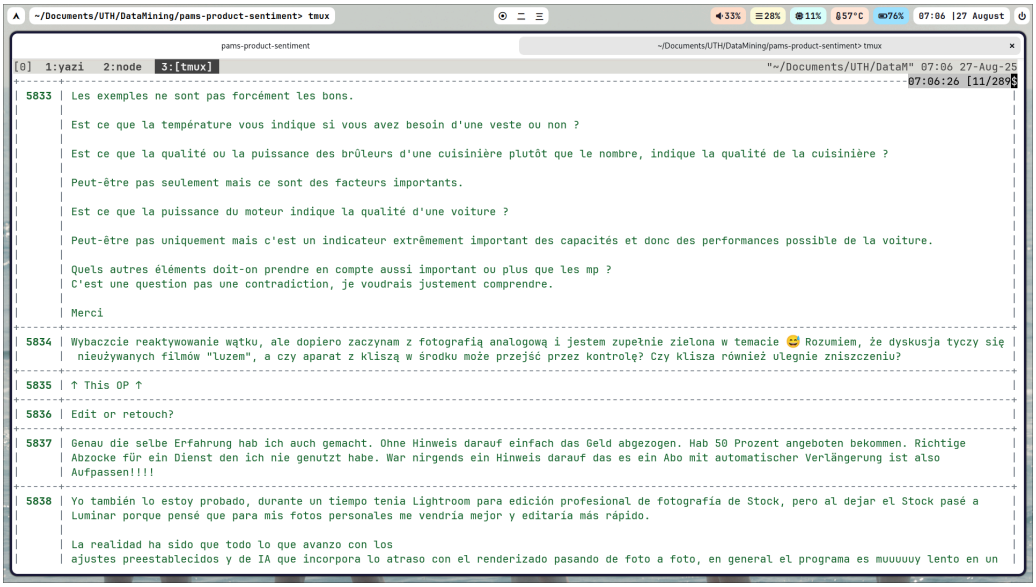

Figure 22: Multilingual Reddit excerpts highlight the need for Vietnamese and European-language handling in the next release.

1.9. Appendix

1.9.1. Core assets

- `plan.md` — detailed architecture, alternatives, evaluation heuristics.
- `eda/` — exploratory plots and CSV summaries.
- `run_pipeline.py` (root & `extra/`) — configurable topic + aspect

sentiment CLI.

- `extra/bai_thuyet_trinh.typ` — Typst slide deck presented in class.

1.9.2. Example console summary (abridged)

```
N=742, V=1625, min_df=3, max_df=0.95, max_features=2000
SVD components=150, cumulative explained variance=0.72
Chosen K=5 with silhouette=0.41
Aspect battery POS: “Battery life has been excellent…” (+0.68)
Aspect battery NEG: “Battery drains fast when streaming…” (-0.52)
```

End of Report

Next steps (optional):

1. Run typst compile final_report.typ final_report.pdf.
2. Re-run run_pipeline.py for any new products and swap image paths if you generate additional artifacts.
3. If you revisit sentiment fine-tuning, add the new evaluation results to the “Limitations & Future Work” section.