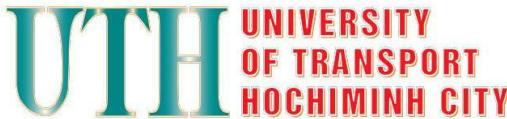


BỘ GIAO THÔNG VẬN TẢI
TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI THÀNH PHỐ HỒ CHÍ MINH



**DATA MINING
BÁO CÁO CUỐI KỲ**

**PHÂN TÍCH NỘI DUNG TRÒ CHUYỆN
TRÍCH XUẤT VĂN ĐỀ PHÔ BIẾN VỀ TÀI SẢN**

Giảng Viên Hướng Dẫn: Trần Thế Vinh

Họ và Tên Sinh Viên Thực Hiện Và MSSV:

Phạm Hoàng Thiện - 05120500064

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.	Introduction & Scope	3
2.	Executive Overview	3
3.	Data Sources & Collection	3
3.1.	Reddit & Tiki rationale	3
3.2.	Arctic Shift workflow	4
3.3.	Corpus snapshot	5
4.	Data Engineering & Integration	5
5.	Exploratory Data Analysis	5
6.	Sentiment Analysis — Initial Plan & Decision	11
7.	Topic Modeling & Aspect Pipeline	11
7.1.	Pre-processing	11
7.2.	Feature Selection & Dimensionality Reduction	11
7.3.	Clustering workflow	12
7.4.	Aspect extraction & sentiment fusion	12
7.5.	Key artifacts	13
8.	Deliverables & Usage	13
9.	Limitations & Future Work	14
10.	Appendix	15
10.1.	Core assets	15
10.2.	Example console summary (abridged)	15

1. Introduction & Scope

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

2. 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/`.

3. Data Sources & Collection

3.1. 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.

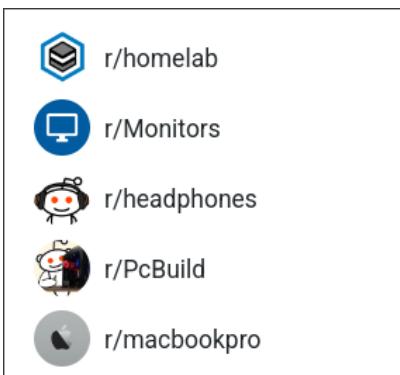


Figure 1: Representative subreddit slate spanning homelab builds, monitors, audiophile hubs, and Apple ecosystems.

product_id	customer_review	rating	text	content	thank_count	CREATE_TS
262790504	Phan Küch	4	Hai King	Máy 2 (Nhi) Vô Khiết 0/1 (Nhi) (Nhi) von kia kia kia (Dung thi 8919) (Dua vay)	0	209705488
179	Lý vín đứt	5	Cực kì hài hước	Cực kì nhẹ nhàng, dễ thương	0	1748314823
180	Nhan Lee	5	Cực kì hài hước	Cực kì nhẹ nhàng, dễ thương	0	1727957609
181	Nhan Ngu	5	Hai King	Cực kì nhẹ nhàng	0	1735555553
182	Kieu Mai Nhung	5	Cực kì hài hước	Hàng rất chảnh hem ảnh rõ nét	0	1717585553
183	Nguyễn Phú Hùng	3	Bình thường	Hình ảnh không rõ nét, Dung lạm dụng	0	1724812759
184	Nguyễn Việt Cường	5	Cực kì hài hước	One. Đẹp mắt, chay nhẹ	0	1719202407
185	Thien Ngan	5	Cực kì hài hước	Cực kì mượt, mượt như lông cát, không cứng cáp	0	1722954563
186	Phạm Hải Ein	5	Cực kì hài hước	Cực kì mượt, mượt như lông cát, không cứng cáp	0	171814794
187	Mie Xuan Thanh	5	Cực kì hài hước	Nhà sản xuất có thể dùng HK shop	0	1741812643
188	Thien Thien	2	Không hài hước	Còn hàng trên dn và bên 2 bên gần như 0.8 km	0	1717212009
189	Phan Khoa	5	Cực kì hài hước	Còn hàng trên dn và bên 2 bên gần như 0.8 km	0	1740297081
190	nguyễn Duy Mạnh	4	Cực kì hài hước	1 mèo em kinh tên với 1 mèo em chửi lung tung trước	0	1702644982
191	nguyễn Nguyễn	4	Hai King	Ok. Đẹp mượt, giao hàng nhanh	0	1711804004
192	Nguyễn Việt Cường	5	Cực kì hài hước	Giao hàng nhanh, đẹp, không có vấn đề	0	1720521228
193	Thu Huu Pham	3	Rất không hài hước	Hình ảnh mờ và ảnh bị biến dạng	0	1722954569
194	Hoang	5	Hai King	Hình ảnh mờ và ảnh bị biến dạng	0	1702550723
195	Inteng Heng	4	Hai King	Đáng 2-3 ngày là OK. Phản hồi tốt, nhưng may không có bảo hành, nếu sự cố gì thì họ bay	0	1702550723
196	Nguyễn Thành Công	3	Bình thường	Bảo hành 1 năm mà 6 tháng camera không được đổi trả	0	1715864223
197	Nguyễn Minh Quốc	5	Cực kì hài hước	Ok	0	1753886023
198	NAK-HOT	5	Cực kì hài hước	Ok	0	1744012387

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

3.2. 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.

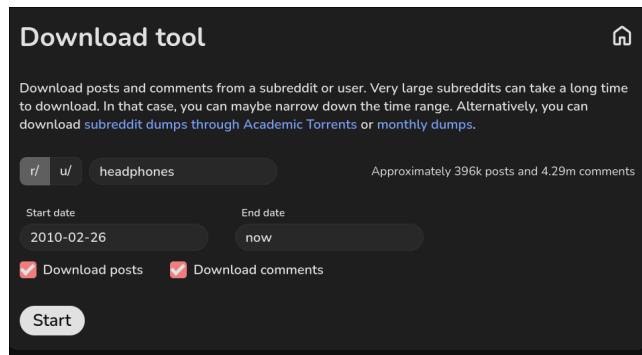


Figure 3: Arctic Shift download tool set to pull the full r/headphones history — bypassing API caps for longitudinal coverage.

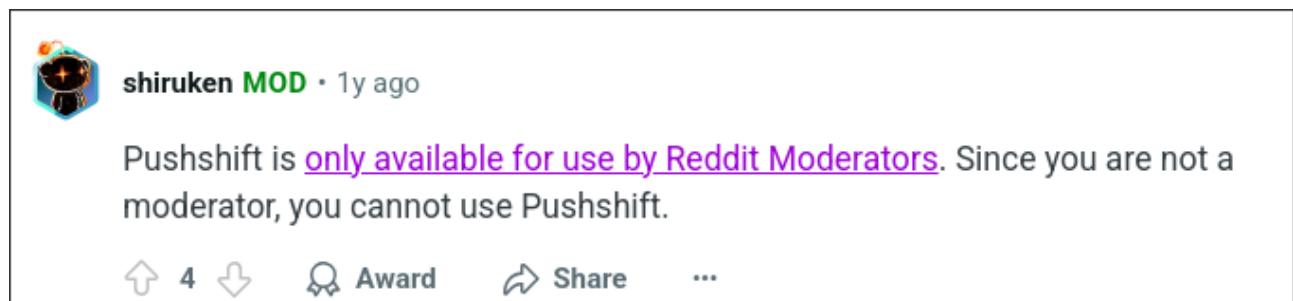


Figure 4: Pushshift access is now moderator-only, reinforcing the need for self-hosted archival strategies.

	name	subreddit	score	link_id	parent_id	body	is_post
789	t3_1la792j	iphone	6			ng to resell this phone and get the new iPhone coming in later this year. Just need opinions	true
790	t3_1m3ywux	iphone	1			ssd cards for recording longer videos, but haven't seen anything about the 16 base model.	true
791	t3_1l5xtrl	laptops	112			I think I messed up while replacing the screen	true
792	t3_1luj9g	laptops	110			air costs in Europe are so high that it's not even worth fixing. Lenovo, this is unacceptable!	true
793	t3_1s7j5e	laptops	118			w old it is, but it's a laptop with a floppy drive. I figured someone here might appreciate this	true
794	t3_1mfhu1q	laptops	247			t gooch on its lower side , but it left this mon matching patch , will it eventually fade away ?	true
795	t3_1mkcmw4	iphone	1			iPhone 15 Pro Makro Foto	true
796	t3_1mkng2m	iphone	1			Shift or True Tone but when I toggle those on and off nothing changes. What is happening?	true
797	t3_1lf7m8l	laptops	0			Help regarding c1g laptop!! 🖥️	true
798	t1_mve6pk1	AppleWatch	35	t3_1l0lvtm	t3_1l0lvtm	Well the picture is actually mine but the idea 😊	false
799	t1_mve6swe	AppleWatch	-7	t3_1l0lvtm	t3_1l0lvtm	you should wrap the watch around the phone horizontally but I like the idea!	false
800	t1_mve70qw	AppleWatch	536	t3_1l0lvtm	t3_1l0lvtm	ap a picture from the watch which is handy for capturing a serial number or model number.	false
801	t1_mve798	AppleWatch	176	t3_1l0lvtm	t3_1l0lvtm	In my days selfie cameras came with a little mirror to be able to see yourself. 😊	false
802	t1_mvea93t	AppleWatch	1	t3_1l0lvtm	t3_1l0lvtm	but why?	false
803	t1_mvewkf1	AppleWatch	9	t3_1l0lvtm	t1_mvea93t	So you can use the much better back camera and also see what you are taking	false
804	t1_mvewb5a3	AppleWatch	17	t3_1l0lvtm	t1_mve70qw	Pretty much the only thing I use it for as well.	false
805	t1_mvewbg7w	AppleWatch	103	t3_1l0lvtm	t3_1l0lvtm	Well if you can get it to work reliably, its always been trouble for me to see once launched.	false
806	t1_mvewe8qa	AppleWatch	20	t3_1l0lvtm	t3_1l0lvtm	I don't see the viewfinder image on my watch, just the controls. Anyone know why?	false
807	t1_mvewefra	AppleWatch	2	t3_1l0lvtm	t3_1l0lvtm	What's the app called?	false
808	t1_mvewerp	AppleWatch	60	t3_1l0lvtm	t3_1l0lvtm	to keep eyes on my truck dash to see where the tire pressure sensors are as I fill the tires.	false
809	t1_mvewf14i	AppleWatch	1	t3_1l0lvtm	t3_1l0lvtm	Or just use the other camera?	false

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

3.3. Corpus snapshot

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

4. Data Engineering & Integration

- Filtering uses case-insensitive alias matching (e.g. “hd600”, “flio 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.

5. Exploratory Data Analysis

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

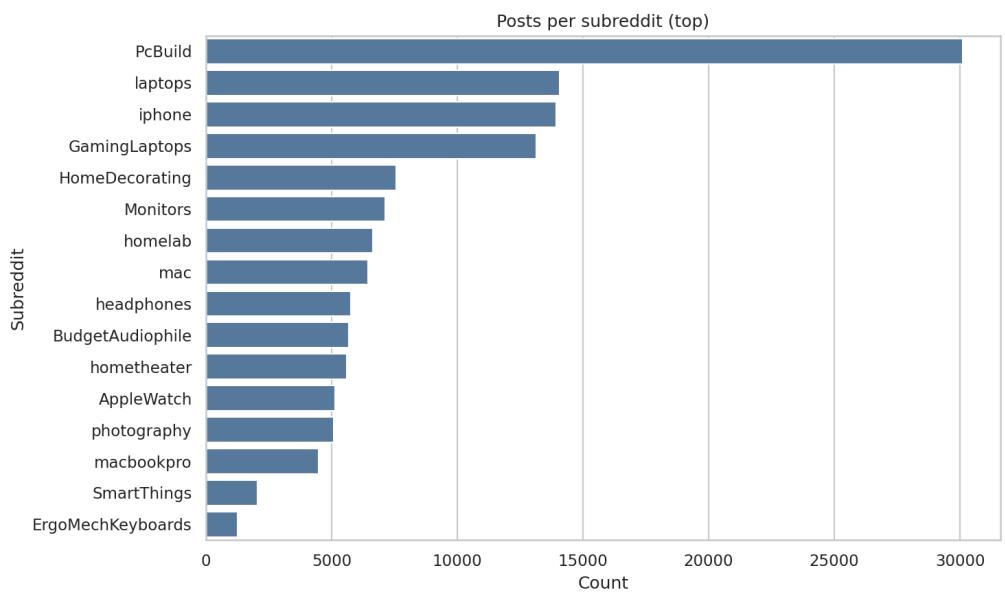


Figure 6: Post counts by subreddit — PcBuild, iPhone, and GamingLaptops dominate volume.

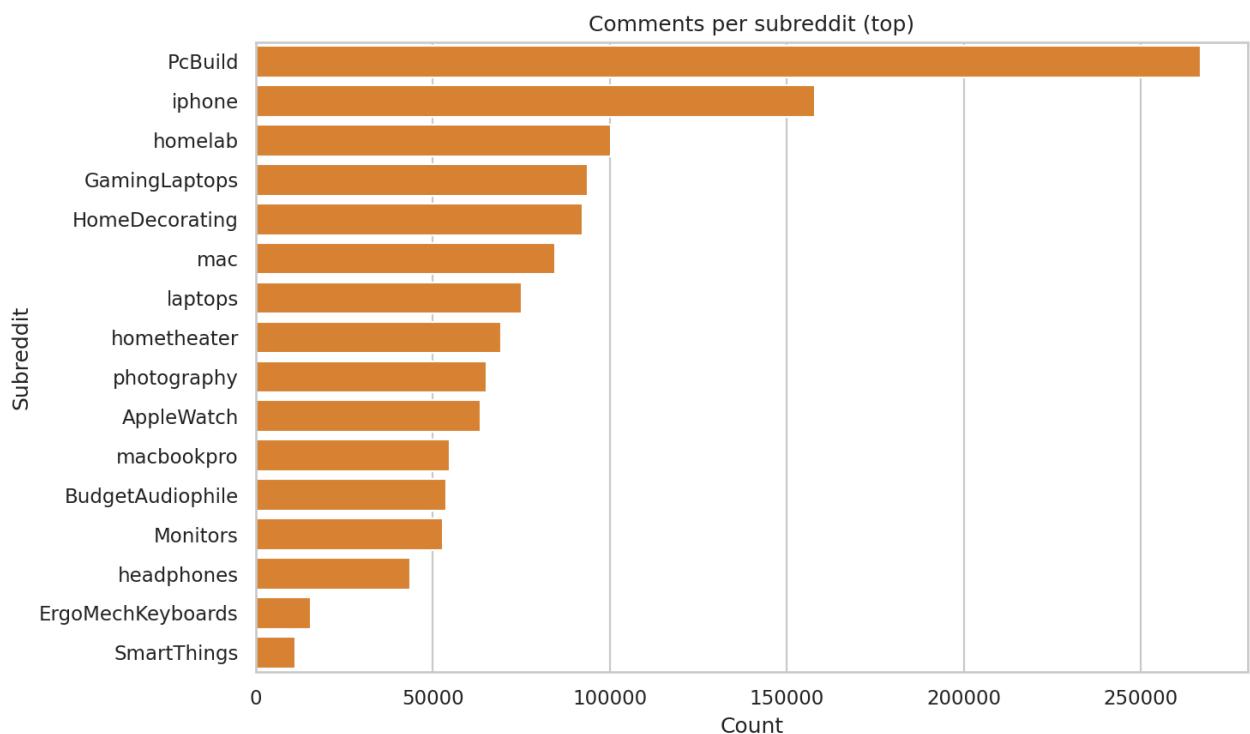


Figure 7: Comment counts: troubleshooting-heavy communities (PcBuild, homelab) lead engagement.

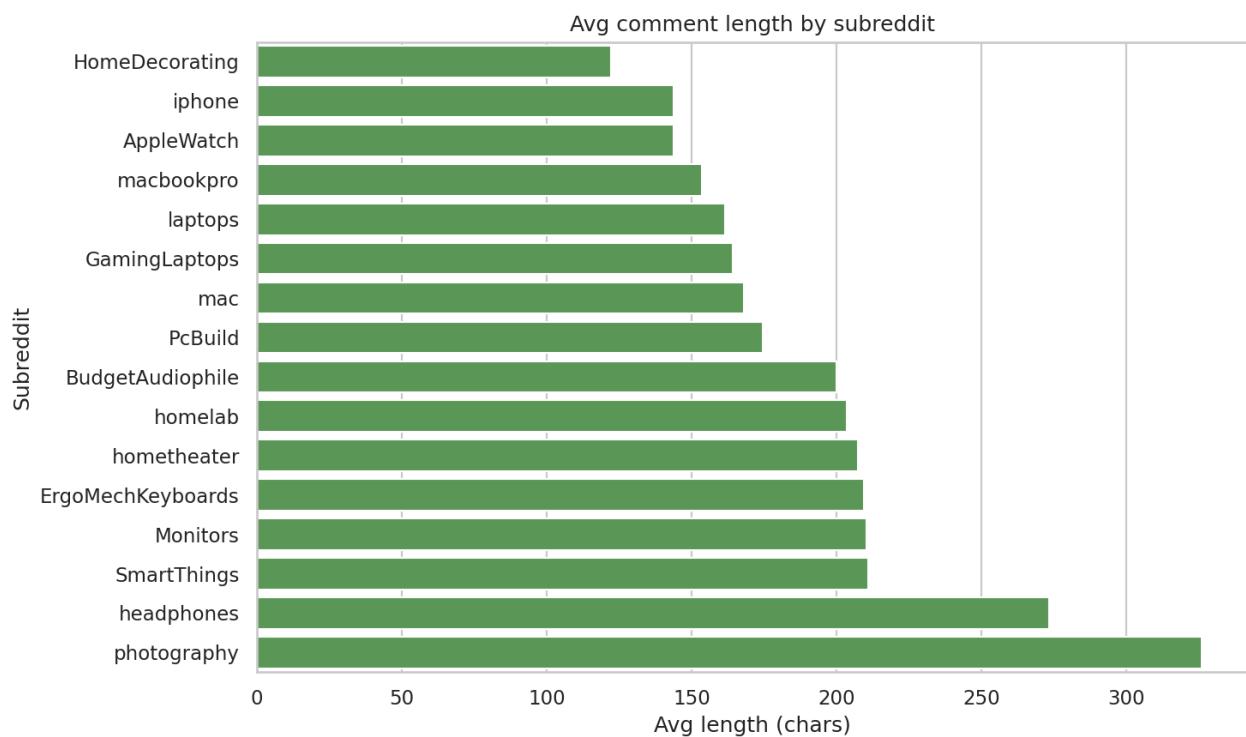


Figure 8: Average comment length highlights deep technical discussions in r/macbookpro and r/AppleWatch.

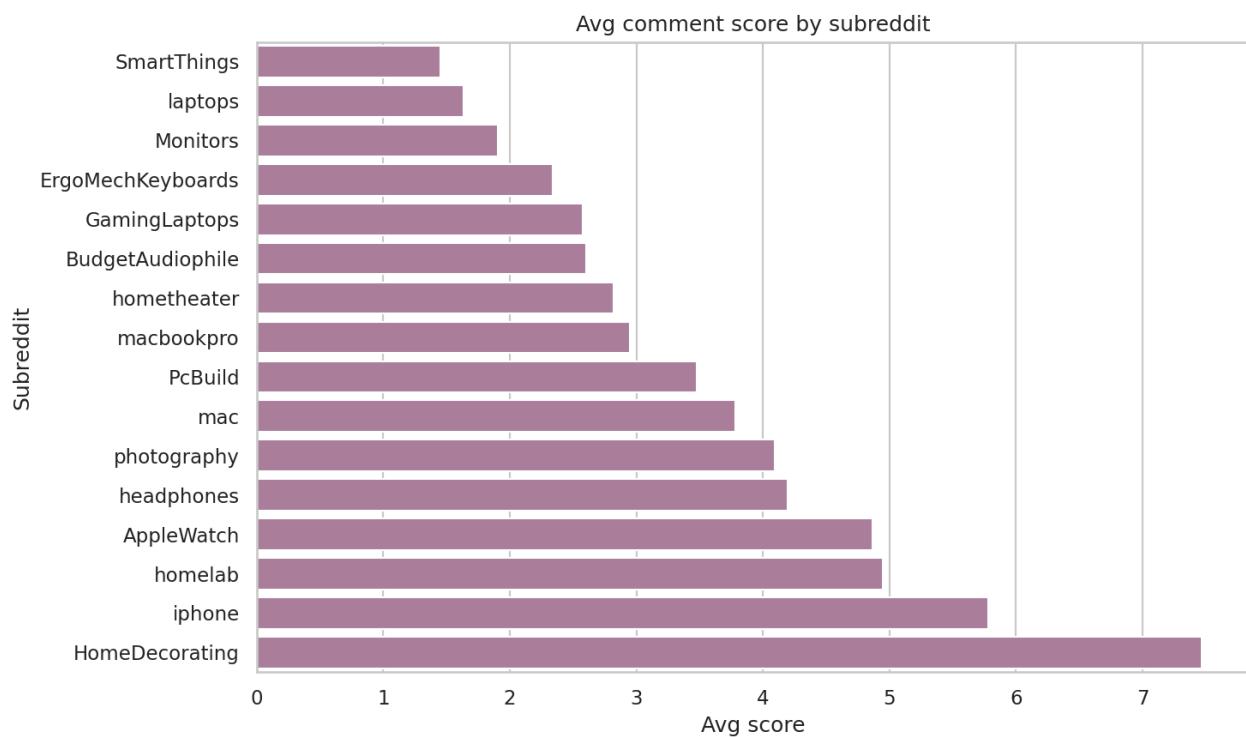


Figure 9: Mean comment scores remain low overall (<2.5), reinforcing the need for textual insight beyond karma.

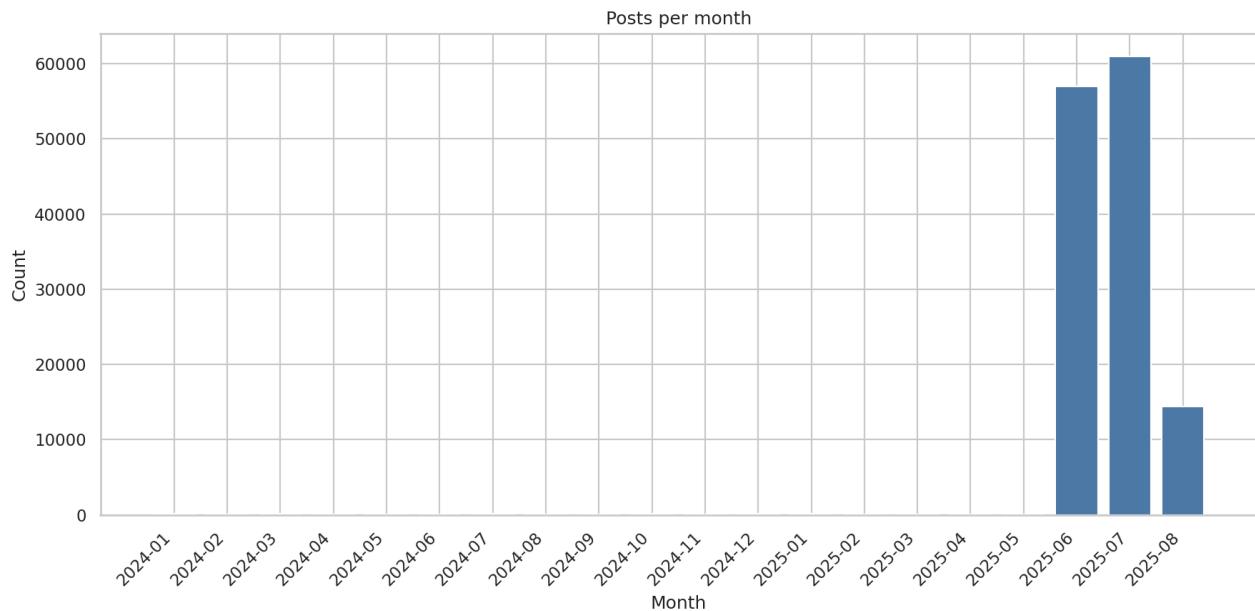


Figure 10: Monthly posting cadence detects seasonal hardware launch spikes.

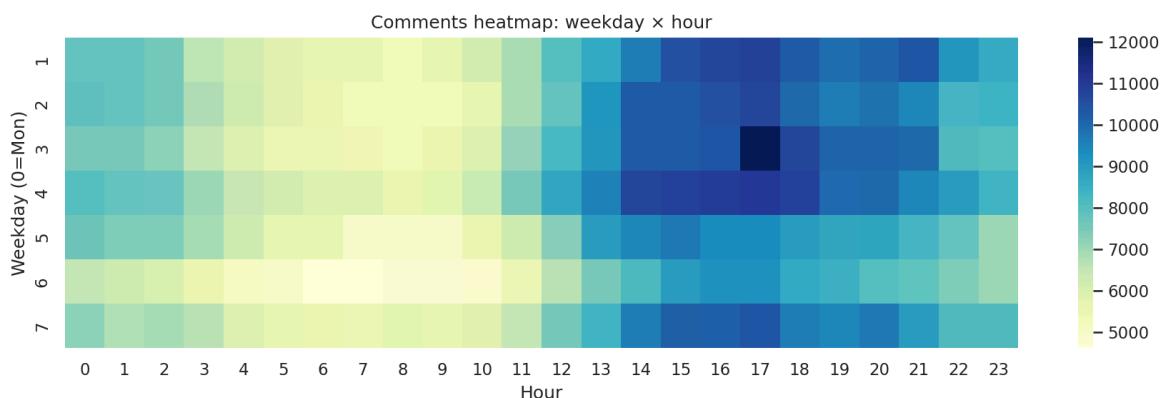


Figure 11: Weekly/hourly heatmap of activity — evenings and weekends drive most conversations.

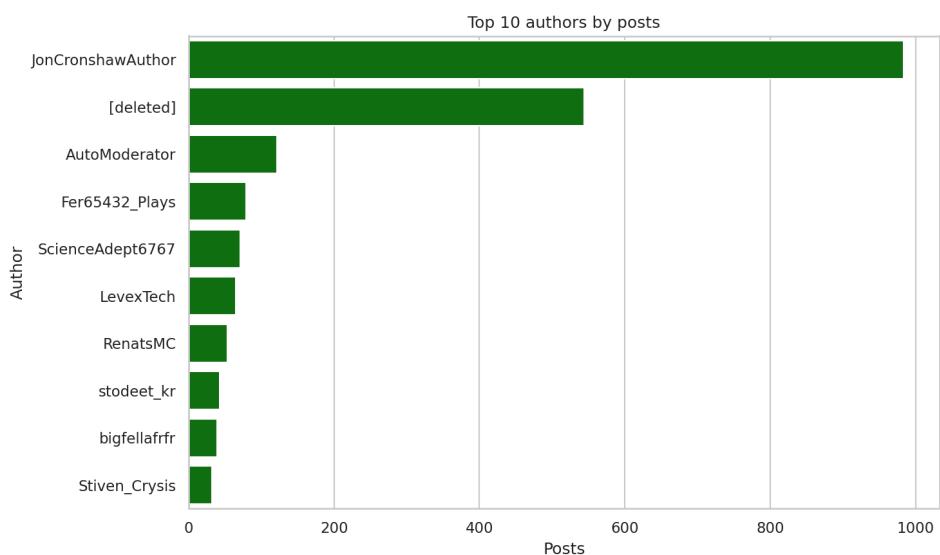


Figure 12: Top post authors — flags power users who shape each hardware narrative.

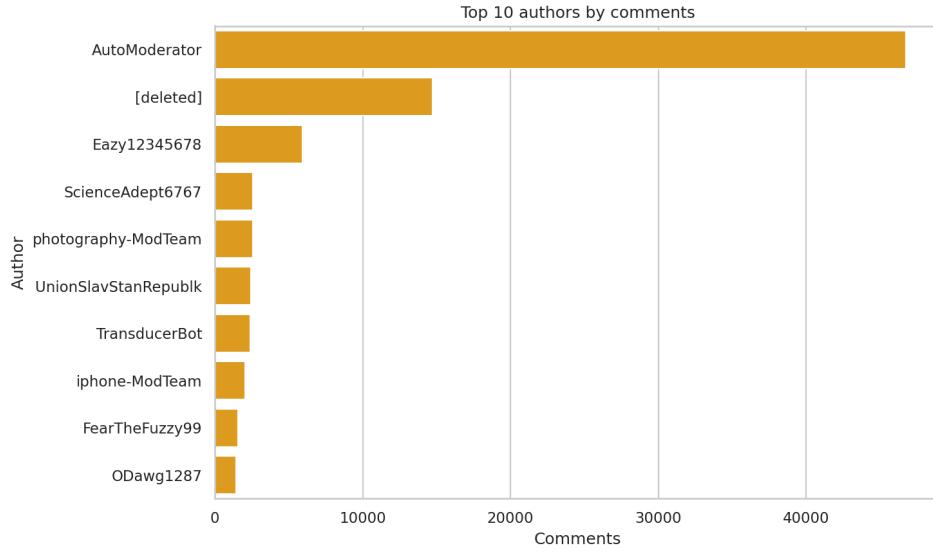


Figure 13: Top comment authors — identifies moderation-heavy and support-oriented contributors.



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

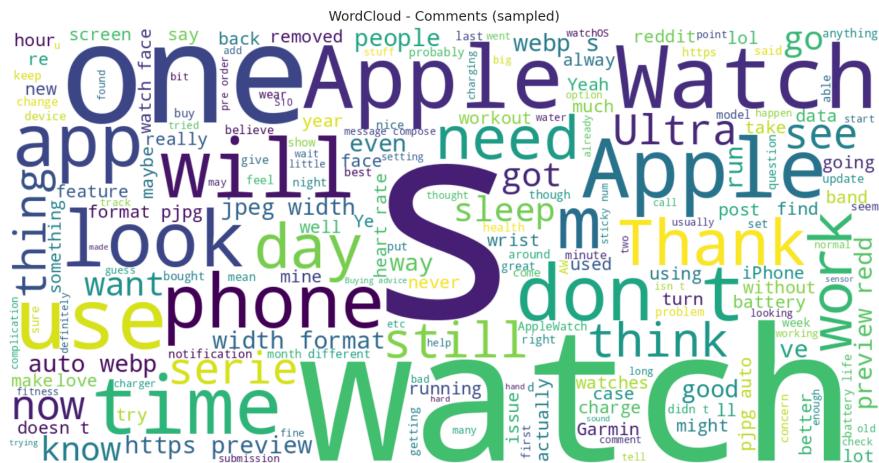


Figure 15: Comment word cloud highlights recurring sentiment cues and component-level jargon.

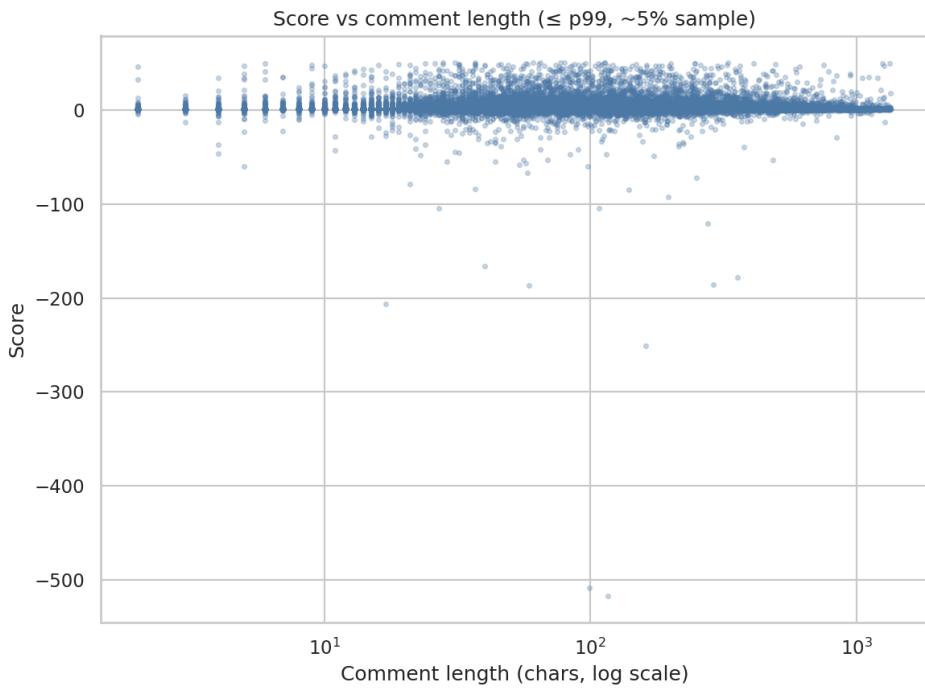


Figure 16: Comment score versus length — longer replies trend toward higher karma among troubleshooting threads.

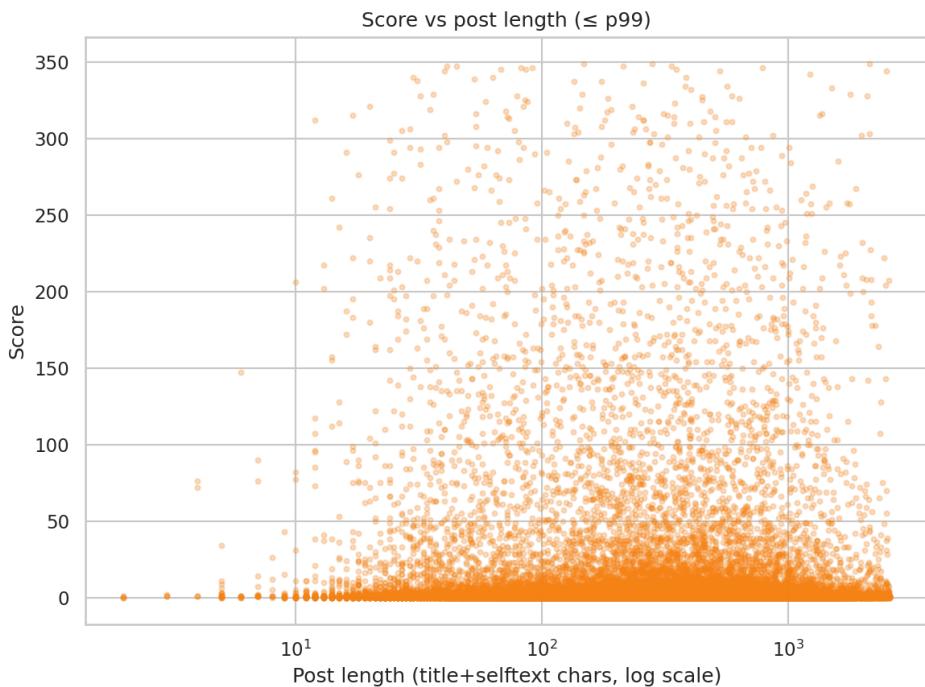


Figure 17: Post score versus selftext length — concise launch rumors and detailed reviews both find audiences.

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

6. 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.

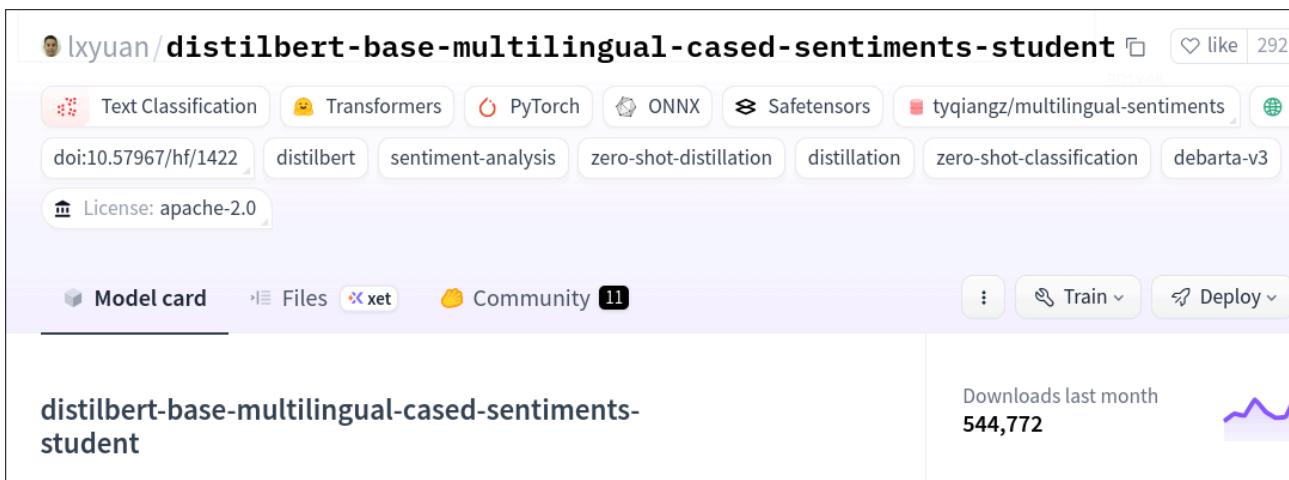


Figure 18: Deferred fine-tuning target `lxyuan/distilbert-base-multilingual` — parked due to GPU and labeling constraints.

7. 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.

7.1. 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.

7.2. Feature Selection & Dimensionality Reduction

- TF-IDF (`ngram_range = 1-2`) with adaptive `min_df` (3 if $N \geq 300$ 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.

7.3. 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.

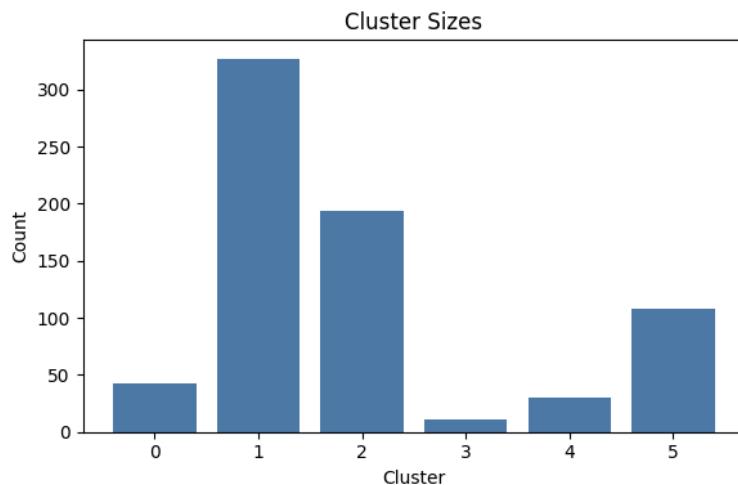


Figure 19: Cluster size distribution (Flio FT1 example) after SVD-reduced KMeans.

7.4. 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.

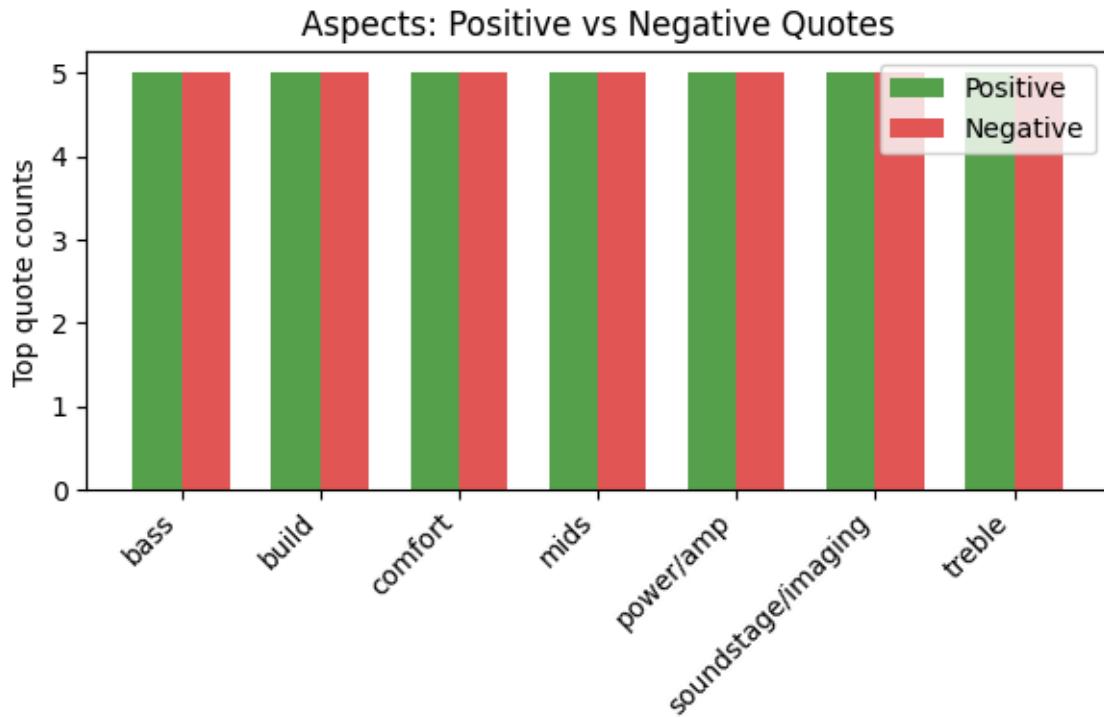


Figure 20: Aspect sentiment summary (Fiio FT1 run): positive vs. negative quote counts.

7.5. 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.

8. 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.

9. 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.

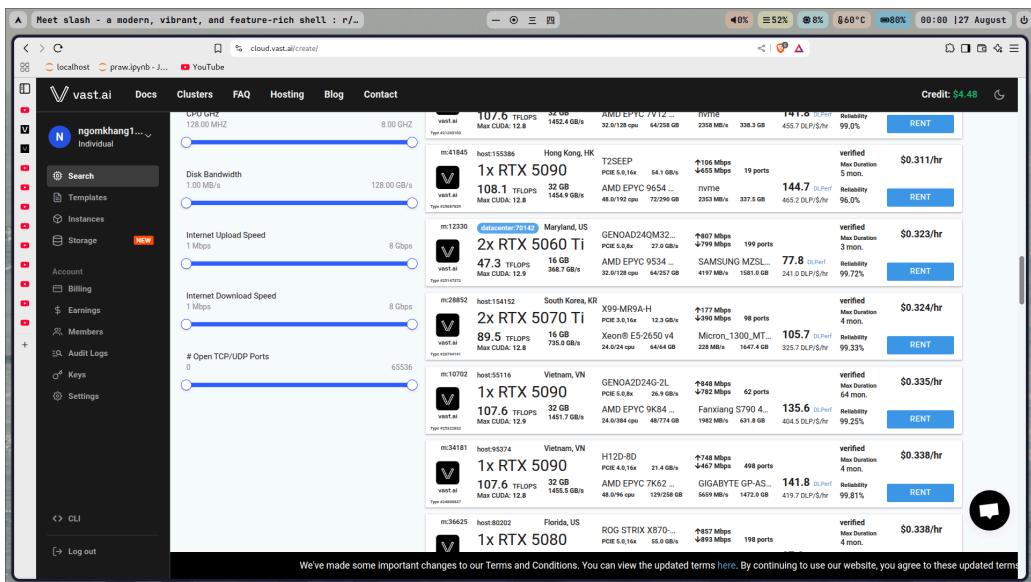


Figure 21: Vast.ai GPU marketplace snapshot — projected fine-tuning spend that led to deferring transformer training.

```

~Documents/UTH/DataMining/pams-product-sentiment> tmux
          33% 28% 11% 857°C 76% 07:06 | 27 August
          ~/Documents/UTH/DataMining/pams-product-sentiment> tmux
          ~/Documents/UTH/DataMining/pams-product-sentiment" 07:06 27-Aug-25
          07:06:26 [11/289]

[0] 1:yazi 2:node 3:[tmux]
pams-product-sentiment

5833 | Les exemples ne sont pas forcément les bons.
| Est ce que la température vous indique si vous avez besoin d'une veste ou non ?
| Est ce que la qualité ou la puissance des brûleurs d'une cuisinière plutôt que le nombre, indique la qualité de la cuisinière ?
| Peut-être pas seulement mais ce sont des facteurs importants.
| Est ce que la puissance du moteur indique la qualité d'une voiture ?
| Peut-être pas uniquement mais c'est un indicateur extrêmement important des capacités et donc des performances possibles de la voiture.
| Quels autres éléments doit-on prendre en compte aussi important ou plus que les mp ?
| C'est une question pas une contradiction, je voudrais justement comprendre.
| Merci

5834 | Wybaczcie reaktywowanie wątku, ale dopiero zaczynam z fotografią analogową i jestem zupełnie zielona w temacie 😊 Rozumiem, że dyskusja dotyczy się
| niewyłącznych filmów "luzem", a czy aparat z klaszą w środku może przejść przez kontrolę? Czy klasza również ulegnie zniszczeniu?
| ↑ This OP ↑

5835 | Edit or retouch?

5836 | Genau die selbe Erfahrung hab ich auch gemacht. Ohne Hinweis darauf einfach das Geld abgezogen. Hab 50 Prozent angeboten bekommen. Richtige
| Abzocke für ein Dienst den ich nie genutzt habe. War nirgends ein Hinweis darauf das es ein Abo mit automatischer Verlängerung ist also
| Aufpassen!!!!
| Yo también lo estoy probado, durante un tiempo tenía Lightroom para edición profesional de fotografía de Stock, pero al dejar el Stock pasé a
| Luminar porque pensé que para mis fotos personales me vendría mejor y editarla más rápido.

5838 | La realidad ha sido que todo lo que avanza con los
| ajustes pre establecidos y de IA que incorpora lo atraso con el renderizado pasando de foto a foto, en general el programa es muuuuy lento en un
| modo de trabajo
```

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

10. Appendix

10.1. 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.

10.2. 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.