Analyzing Public Discussions for Product Insights

Mining Reddit and Tiki for product issues, sentiment, and trends

From noisy threads ⟶ actionable product signals.

[VISION]

Problem & Value

- Which products break, why, and how often—using public discussions.
- Turn unstructured Reddit and Tiki text into actionable product signals.
- Output insights teams can ship on: issues, trends, severity, examples.

Objectives

- Which issues? How frequent? How trending? What severity?
- Where do issues cluster (by product/subreddit/time)?
- How reliable are methods vs baselines?
- Deliver a repeatable pipeline and clear, prioritized findings.

Why These Platforms?

- Reddit: rich, threaded discussions; public API; strong NLP support.
- Tiki (e-commerce): structured, purchase-verified reviews; VN market signal.
- Complementary: text-rich threads vs short reviews ⟶ broader coverage.

Data Source Comparison

Platform	Pros	Cons
Reddit	• Public API (easy access) • Rich text discussions • Strong support for English NLP	• Some noise and off-topic posts • Limited Vietnamese data • API usage restrictions/quotas

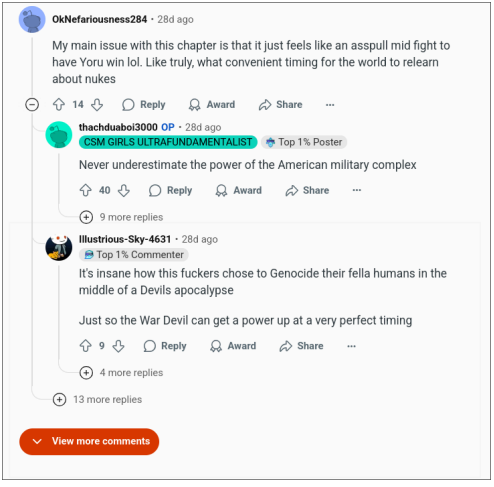
	• Community-driven, topic-focused “subreddits”	
Facebook	• Large user base in Vietnam • Active groups and communities • Rich variety of topics	• Many bots/spam accounts • API restrictions • Difficult to collect clean data • Weak support for English NLP libraries
TikTok	• Very popular among young users • Strong trend/meme insights	• Mostly media (videos, images) • No official API • Lacks group/community structure • Scraping is slow (parse HTML) • Hard to extract relevant textual data

Data Collection

- **Initial:** PRAW (Python Reddit API Wrapper) for prototyping.
  - **Limitation:** 1,000 post cap, rate limits.
- **Current (Hybrid):**
  - Reddit historical archives (Academic Torrents) to bypass API caps.
  - Tiki review dumps for cross-source validation.
  - **Result:** Broader time windows, more volume.

How PRAW Traverses Comments

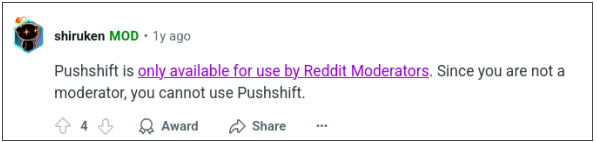
- Reddit returns comment trees with `View more comments` placeholders.
- We expand lazily, traverse, and serialize —no one-click “download”.
- Captures full depth where needed; avoids rate-limit explosions.



Alternatives Considered

Pushshift.io (Pushshift API)

- More powerful than PRAW, can filter posts by time
- Requires moderator status → not feasible



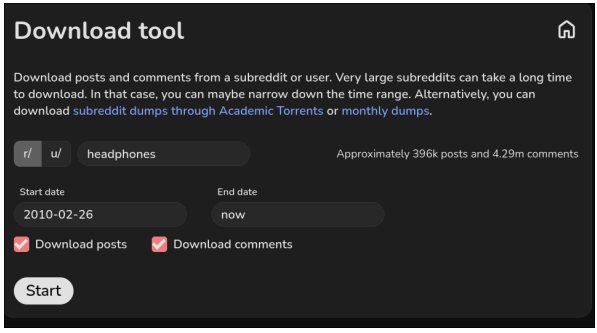
Personal Archive

- Continuous collection over weeks
- Impractical: hardware/time, can't capture older posts

> (Or building one's own archive over a long period of time like the OP mentioned in another comment, that works too – but it does take time. Though they could load it with data from these archives too if they were so inclined.)

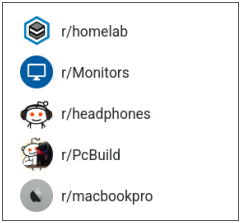
Academic Torrents (Arctic Shift)

- Downloadable historical Reddit datasets
- Good for history + scale



Subreddits Chosen

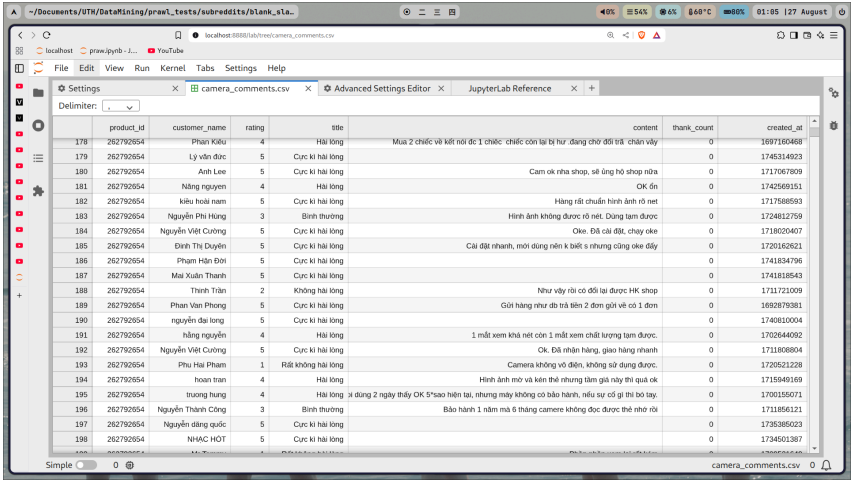
- A diverse mix of tech and lifestyle communities:
- r/macbookpro, r/GamingLaptops
- r/iphone, r/AppleWatch, r/Monitors, r/headphones, r/homelab, r/photography
- ...and several others covering home, audio, and PC building.
- total : `134121` posts, `1300190` comments, 2025-06-01 -> 2025-07-31.




```
prawl_tests/subreddits/blank_slate on  main [×!?] via 🐍
❯ : polars open all_posts_hanni.parquet | polars shape
+---+--------+---------+
| # |  rows  | columns |
+---+--------+---------+
| 0 | 134121 |      45 |
+---+--------+---------+

prawl_tests/subreddits/blank_slate on  main [×!?] via 🐍
❯ : polars open all_comments_hanni.parquet | polars shape
+---+---------+---------+
| # |  rows   | columns |
+---+---------+---------+
| 0 | 1300190 |      25 |
+---+---------+---------+
```

Tiki



	product_id	customer_name	rating	title	content	thank_count	created_at
178	262792654	Phan Kiều	4	Hài lòng	Mua 2 chiếc về kết nối đc 1 chiếc chiếc còn lại bị hư đang chờ đổi trả chán vậy	0	1697160468
179	262792654	Lý văn đức	5	Cực kì hài lòng		0	1745314923
180	262792654	Anh Lee	5	Cực kì hài lòng	Cam ok nha shop, sẽ ủng hộ shop nữa	0	1717067809
181	262792654	Năng nguyen	4	Hài lòng	OK ổn	0	1742569151
182	262792654	kiều hoài nam	5	Cực kì hài lòng	Hàng rất chuẩn hình ảnh rõ net	0	1717588593
183	262792654	Nguyễn Phi Hùng	3	Bình thường	Hình ảnh không được rõ nét. Dùng tạm được	0	1724812759
184	262792654	Nguyễn Việt Cường	5	Cực kì hài lòng	Oke. Đã cài đặt, chạy oke	0	1718020407
185	262792654	Đinh Thị Duyên	5	Cực kì hài lòng	Cài đặt nhanh, mới dùng nên k biết s nhưng cũng oke đấy	0	1720162621
186	262792654	Pham Hận Đời	5	Cực kì hài lòng		0	1741834796
187	262792654	Mai Xuân Thanh	5	Cực kì hài lòng		0	1741818543
188	262792654	Thinh Trần	2	Không hài lòng	Như vậy rồi có đổi lại được HK shop	0	1711721009
189	262792654	Phan Van Phong	5	Cực kì hài lòng	Gửi hàng như db trả tiền 2 đơn gửi về có 1 đơn	0	1692879381
190	262792654	nguyễn đại long	5	Cực kì hài lòng		0	1740810004
191	262792654	hằng nguyễn	4	Hài lòng	1 mắt xem khá nét còn 1 mắt xem chất lượng tạm được.	0	1702644092
192	262792654	Nguyễn Việt Cường	5	Cực kì hài lòng	Ok. Đã nhận hàng, giao hàng nhanh	0	1711808804
193	262792654	Phu Hai Pham	1	Rất không hài lòng	Camera không vô điện, không sử dụng được.	0	1720521228
194	262792654	hoan tran	4	Hài lòng	Hình ảnh mờ và kén thế nhưng tầm giá này thì quá ok	0	1715949169
195	262792654	truong hung	4	Hài lòng	ại dùng 2 ngày thấy OK 5*sao hiện tại, nhưng máy không có bảo hành, nếu sự cố gì thì bỏ tay.	0	1700155071
196	262792654	Nguyễn Thành Công	3	Bình thường	Bảo hành 1 năm mà 6 tháng camere không đọc được thẻ nhớ rồi	0	1711856121
197	262792654	Nguyễn đăng quốc	5	Cực kì hài lòng		0	1735385023
198	262792654	NHẠC HỘT	5	Cực kì hài lòng		0	1734501387

EDA Highlights

- [Placeholder image: time series of comments/week by subreddit]
- [Placeholder image: length distribution (boxplot or histogram)]
- [Placeholder image: top product mentions (bar) via keyword/NER]
- Takeaways (placeholder): spikes follow launch X; r/homelab comments 2× longer.

Preprocessing Pipeline

Convert 32 JSONL → 2 Parquet (Nushell + Polars); keeping as many meaningful columns as possible.

Before

Posts: **106** columns
Comments: **69** columns

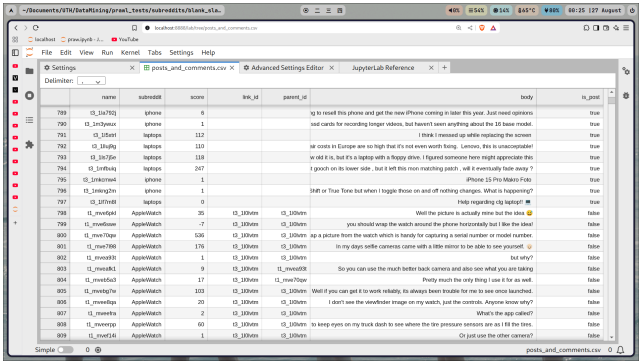
→

After

Posts: **28** columns
Comments: **17** columns

- **Cleaning:**
  - Removed URLs.

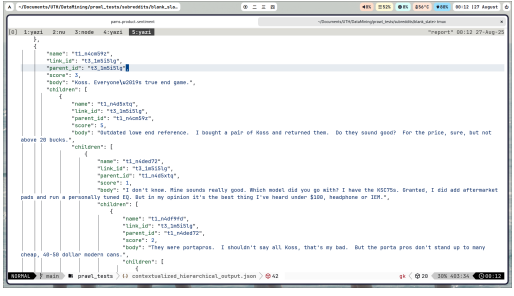
- Filtered for English-only content for initial analysis. (only 5 thousands comments(out of 1,3 million), but consider keeping since the model is multilingual)
- Top 30 + 20 randomized posts per sub (seeded) to reduce virality bias. (placeholder)

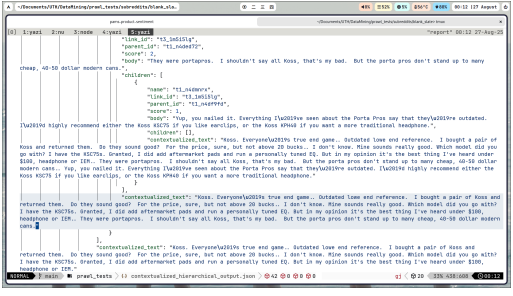


Context Experiment (Per-Comment SA)

- Parent-context vs per-comment classification compared on small labeled set.
- Result: per-comment avoids 512-token truncation and context pollution.

contextualized_hierarchical_output.json : shows what a comment with the context of it's comment tree looks like.

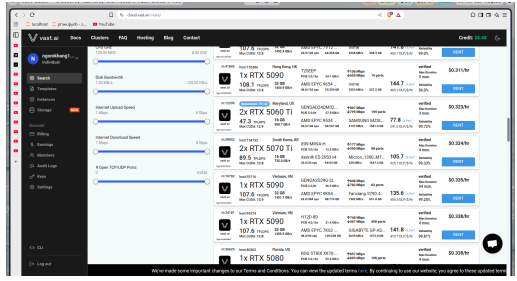




Sentiment Analysis (SA)

- **Approach:** Pretrained transformers over simpler baselines (VADER, TextBlob).
- **Model:** `lxyuan/distilbert-base-multilingual-cased-sentiments-student`
  - **Why:** Good multilingual support, lightweight, strong zero-shot performance.
- **Execution:** Rented GPU on `vast.ai` for large-scale inference.
- **Next Step:** Fine-tune on a domain-specific labeled dataset.


```
{
    "name": "t1_n4cku6n",
    "link_id": "t3_1m5i5lg",
    "parent_id": "t3_1m5i5lg",
    "score": 2,
    "body": "Thats literally me!",
    "contextualized_text": "Thats literally me!",
    "body_positive": 0.5420639514923096,
    "body_neutral": 0.0,
    "body_negative": 0.0,
    "contextualized_text_positive": 0.5420639514923096,
    "contextualized_text_neutral": 0.0,
    "contextualized_text_negative": 0.0
},
```

Topic Modeling (Planned)

- **Goal:** Discover key themes and issues per subreddit.
- **Method:** Evaluate BERTopic vs. traditional methods (LDA/NMF).
- **Process:** Preprocess with domain stopwords, lemmatization.
- **Output:** Top topics, representative comments, and trend lines.
- Target: coherence c_v threshold and stable topics across subsamples. (placeholder)
- [Placeholder image: intertopic distance map (BERTopic mock)]
- [Placeholder image: top-words table for 2 sample topics]

CLI Tool (Planned)

- A pipeline for repeatable analysis:
- `ingest`: Raw data to Parquet.
- `clean`: Filter and normalize data.
- `sentiment`: Run batch sentiment analysis.
- `topics`: Train and apply topic models.
- `report`: Aggregate results and export.

Risks & Mitigations

- **Bias:** Sampled randomly across multiple, diverse subreddits.
- **Time Drift:** Included time slices to compare cohorts.
- **Reproducibility:** Pinned environments, config-driven runs, and stored seeds.
- **Ethics:** Used public data, followed platform ToS, aggregated results.

Next Steps

- Finalize EDA visualizations.
- Run sentiment analysis at scale.
- Pilot and select a topic modeling approach.
- Build and demo the core CLI workflow.
- Begin labeling for fine-tuning sentiment model.

Q & A