

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?

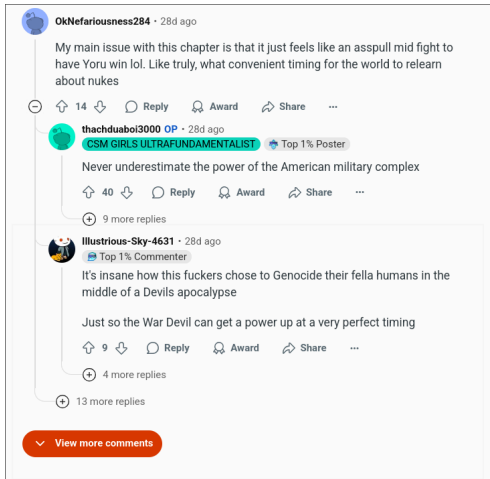
- Reddit: rich, threaded discussions; public API; strong NLP support.
- Tiki (e-commerce): structured, purchase-verified reviews; VN market signal.
- Complementary: text-rich threads vs short reviews → broader coverage.
- [Placeholder image: two-column pros/cons cards + coverage map]

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.
- [Placeholder image: timeline ribbon PRAW → Archives → Tiki]
- [Placeholder image: flowchart with rate-limit icon on PRAW; time-filter lock]

How PRAW Traverses Comments

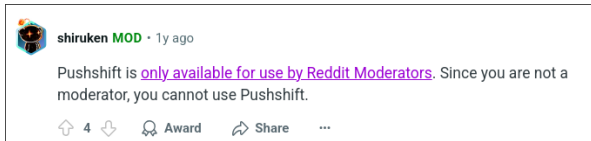
- Reddit returns comment trees with View more comments placeholders.
- We expand lazily, traverse, and serialize —no one-click “download”.
- Captures full depth where needed; avoids rate-limit explosions.



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

Download tool



Download posts and comments from a subreddit or user. Very large subreddits can take a long time to download. In that case, you can maybe narrow down the time range. Alternatively, you can download [subreddit dumps through Academic Torrents](#) or [monthly dumps](#).

r/

u/

headphones

Approximately 396k posts and 4.29m comments

Start date

2010-02-26

End date

now

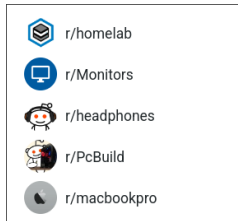
☒ Download posts

☒ Download comments

Start

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 [x!?] via 🐙
) : polars open all_posts_hanni.parquet | polars shape
+-----+
| # | rows | columns |
+-----+
| 0 | 134121 | 45 |
+-----+

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

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 \rightarrow 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.

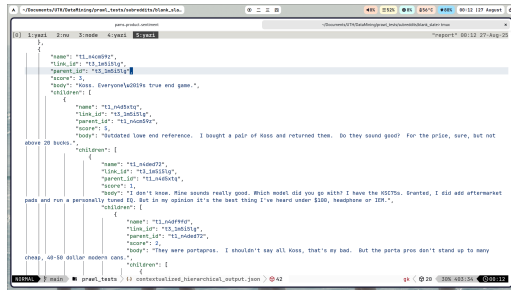
- ▶ Filtered for English-only content for initial analysis. (only 5 thousands comments(out of 1,3 million), but consider keeping since the model is multilingual)
- Top 30 + 20 randomized posts per sub (seeded) to reduce virality bias. (placeholder)

	name	subreddit	score	link_id	parent_id	body	is_post
789	t3_1ia7502	iphone	6			to resell this phone and get the new iPhone coming in later this year. Just need opinions.	true
790	t3_1m2y9ux	iphone	1			sad cards for recording longer videos, but haven't seen anything about the 16 base model.	true
791	t3_1id5d1	laptops	112			I think I messed up while replacing the screen	true
792	t3_1lu9p9	laptops	110			air costs in Europe are so high that it's not even worth being. Lenovo, this is unacceptable!	true
793	t3_1bc7j6n	laptops	118			is old it is, but it's a laptop with a floppy drive. I figured someone here might appreciate this	true
794	t3_1orh4uk	laptops	247			I gooch on its lower side, but I left this mon-matching patch, will it eventually fade away?	true
795	t3_1nkorre4	iphone	1			iPhone 15 Pro Maxro Foto	true
796	t3_1m9uz2n	iphone	1			Shift or True Tone but when I toggle those on and off nothing changes. What is happening?	true
797	t3_1l7m8l	laptops	0			Help regarding cly laptop!!	true
798	t3_1mve6p4	AppleWatch	35	t3_130vtn	t3_130vtn	Well the picture is actually mine but the idea	false
799	t3_1mve60w	AppleWatch	-7	t3_130vtn	t3_130vtn	you should wrap the watch around the phone horizontally but I like the idea!	false
800	t3_1mve70q9	AppleWatch	536	t3_130vtn	t3_130vtn	ap a picture from the watch which is handy for capturing a serial number or model number	false
801	t3_1mve7188	AppleWatch	176	t3_130vtn	t3_130vtn	in my days selfie cameras came with a little mirror to be able to see yourself.	false
802	t3_1mvea93t	AppleWatch	1	t3_130vtn	t3_130vtn	but why?	false
803	t3_1mvea8d1	AppleWatch	9	t3_130vtn	t3_1mvea93t	So you can use the much better back camera and also see what you are taking	false
804	t3_1mvea5a3	AppleWatch	17	t3_130vtn	t3_1mve70q9	Proty much the only thing I use it for as well.	false
805	t3_1mveb7he	AppleWatch	103	t3_130vtn	t3_130vtn	Well if you can get it to work reliably, its always been trouble for me to see once launched.	false
806	t3_1mveb9qs	AppleWatch	20	t3_130vtn	t3_130vtn	I don't see the viewfinder image on my watch, just the controls. Anyone know why?	false
807	t3_1mvebtra	AppleWatch	2	t3_130vtn	t3_130vtn	What's the app called?	false
808	t3_1mveevp	AppleWatch	60	t3_130vtn	t3_130vtn	to keep eyes on my truck dash to see where the tire pressure sensors are as I fill the tires.	false
809	t3_1mvef34i	AppleWatch	1	t3_130vtn	t3_130vtn	Or just use the other camera?	false

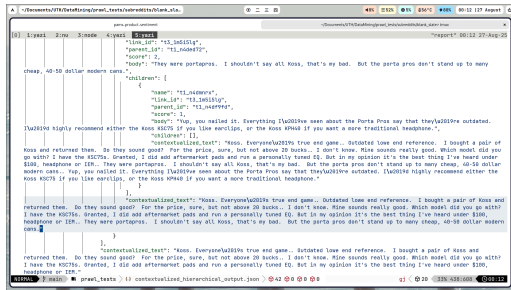
Context Experiment (Per-Comment SA)

- 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.



The screenshot shows a code editor with a JSON tree structure. The root node is a comment with the text "Koss. Everyone's true end game..". It has a score of 5 and a body that says "Outdated low end reference. I bought a pair of Koss and returned them. Do they sound good? For the price, sure, but not above 20 bucks..". The comment has three children: a link to a Koss website, a link to a Koss website, and a link to a Koss website. The editor shows the JSON structure for the comment and its children, including the text of the comment and the links to the Koss website.

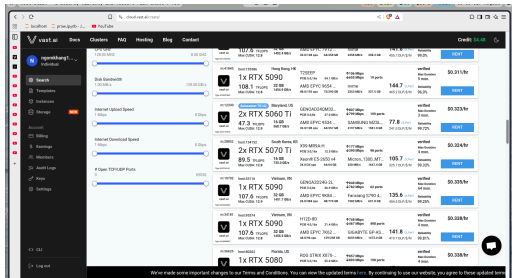


The screenshot shows a code editor with a JSON tree structure, similar to the one in the previous image. The root node is a comment with the text "Koss. Everyone's true end game..". It has a score of 5 and a body that says "Outdated low end reference. I bought a pair of Koss and returned them. Do they sound good? For the price, sure, but not above 20 bucks..". The comment has three children: a link to a Koss website, a link to a Koss website, and a link to a Koss website. The editor shows the JSON structure for the comment and its children, including the text of the comment and the links to the Koss website. A specific field, "contextualized_text", is highlighted in the JSON structure, showing the full context of the comment and its children.

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