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)?
- 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

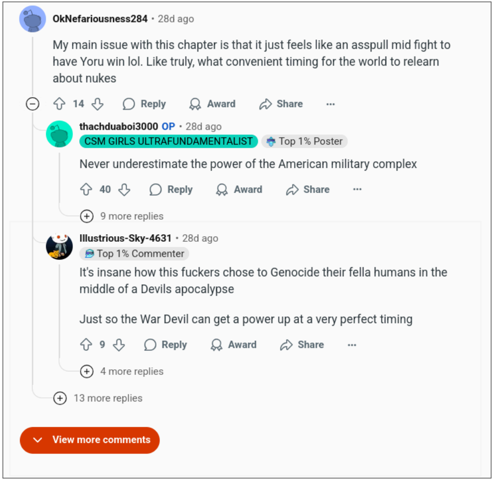
	• Community-driven, topic-focused “subreddits”	
Facebook	• Large user base in Vietnam • Active groups and communities • Rich variety of topics	• Many bots/spam accounts • API restrictions • Difficult to collect clean data • Weak support for English NLP libraries
TikTok	• Very popular among young users • 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.



OkNefariousness284 · 28d ago

My main issue with this chapter is that it just feels like an asspull mid fight to have Yoru win lol. Like truly, what convenient timing for the world to relearn about nukes

thachduaboi3000 OP · 28d ago
CSM GIRLS ULTRAFUNDAMENTALIST
Top 1% Poster

Never underestimate the power of the American military complex

Illustrious-Sky-4631 · 28d ago
Top 1% Commenter

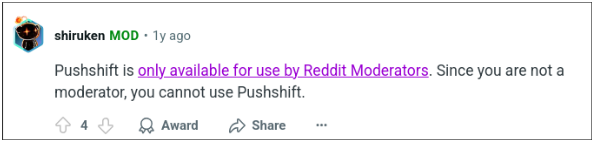
It's insane how this fuckers chose to Genocide their fella humans in the middle of a Devils apocalypse

Just so the War Devil can get a power up at a very perfect timing

Alternatives Considered

Pushshift.io (Pushshift API)

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

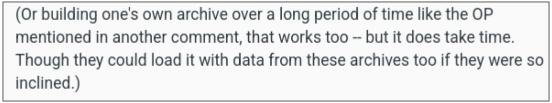


shiruken MOD · 1y ago

Pushshift is only available for use by Reddit Moderators. Since you are not a moderator, you cannot use Pushshift.

Personal Archive

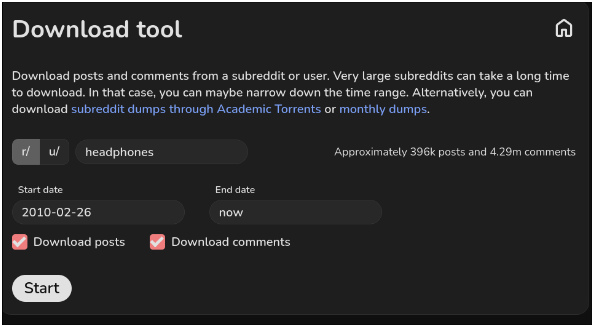
- Continuous collection over weeks
- Impractical: hardware/time, can't capture older posts



(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/ headphones

Approximately 396k posts and 4.29m comments

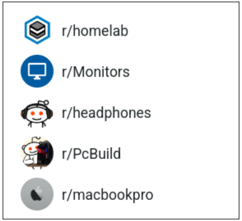
Start date: 2010-02-26

End date: now

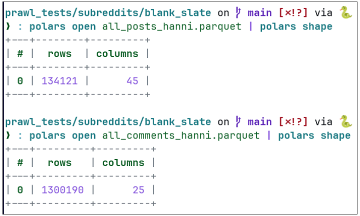
☑ Download posts ☑ Download comments

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.



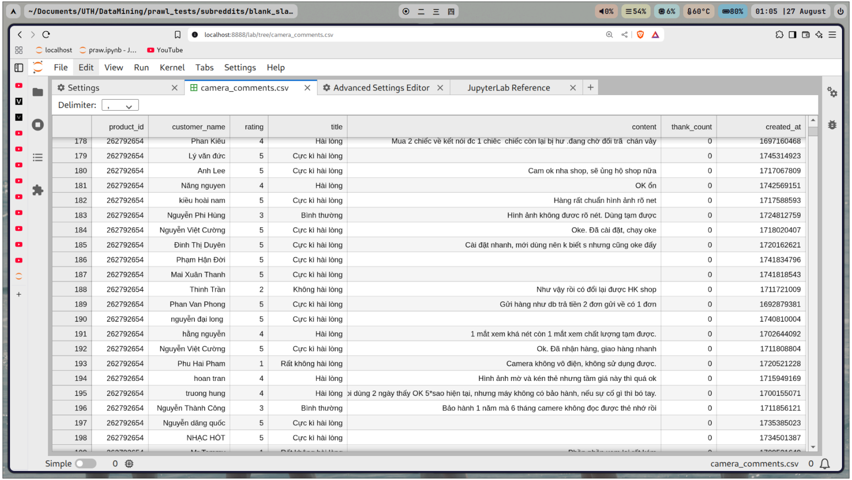
r/homelab
r/Monitors
r/headphones
r/PcBuild
r/macbookpro



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

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

Tiki



Delimiter: ,

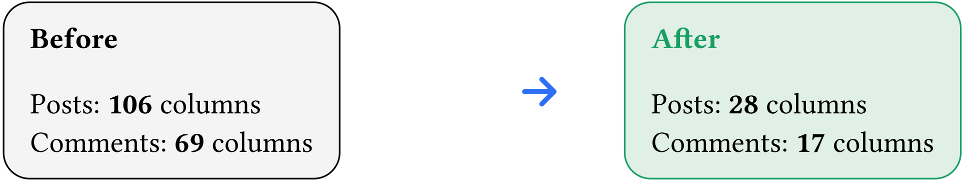
	product_id	customer_name	rating	title	content	thank_count	created_at
178	262792654	Phan Kiêu	4	Hài lòng	Mua 2 chiếc về kết nói đc 1 chiêc chiếc còn lại bị hư .đang chờ đối trã chán vảy	0	1697160468
179	262792654	Lý văn đức	5	Cực kì hài lòng		0	1745314923
180	262792654	Anh Lee	5	Cực kì hài lòng	Cam ok 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	Phạm 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 thẻ nhưng tầm giá này thì quá ok	0	1715949169
195	262792654	truong hung	4	Hài lòng	ɔi dùng 2 ngày thấy OK 5*sao hiện tại, nhưng máy không có bảo hành, nếu sự cố gì thì bó 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 dăng quốc	5	Cực kì hài lòng		0	1735385023
198	262792654	NHẠC HÓT	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

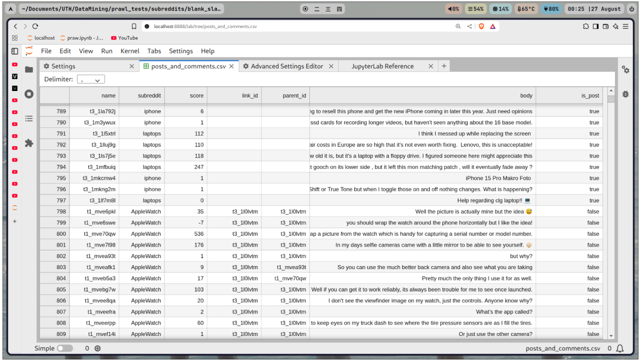
Posts: **106** columns
Comments: **69** columns

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)

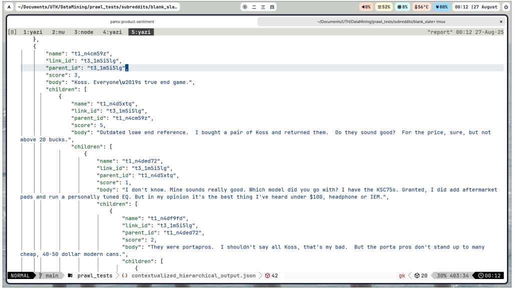


	name	subreddit	score	link_id	parent_id	body	is_post
789	t3_1la792j	iphone	6			ıg to resell this phone and get the new iPhone coming in later this year. Just need opinions	true
790	t3_1m3ywux	iphone	1			ssd cards for recording longer videos, but haven't seen anything about the 16 base model.	true
791	t3_1l5xtrl	laptops	112			I think I messed up while replacing the screen	true
792	t3_1lluj9g	laptops	110			air costs in Europe are so high that it's not even worth fixing. Lenovo, this is unacceptable!	true
793	t3_1ls7j5e	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_1mfbuiq	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 clg laptop!! 💻	true
798	t1_mve6pkl	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_mve7l98	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_mveafk1	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_mveb5a3	AppleWatch	17	t3_1l0lvtm	t1_mve70qw	Pretty much the only thing I use it for as well.	false
805	t1_mvebg7w	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_mvee8qa	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_mveefra	AppleWatch	2	t3_1l0lvtm	t3_1l0lvtm	What's the app called?	false
808	t1_mveerpp	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_mvef14i	AppleWatch	1	t3_1l0lvtm	t3_1l0lvtm	Or just use the other camera?	false

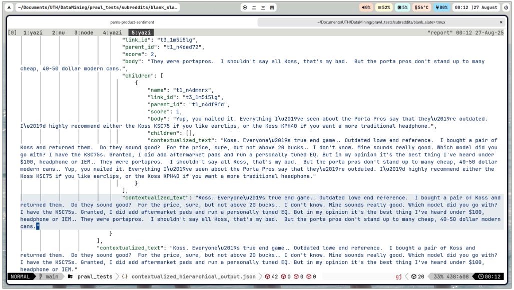
Context Experiment (Per-Comment SA)

- 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.


~/Documents/UTH/DataMining/prawl_tests/subreddits/blank_sla…
◀0%
☰52%
8%
56°C
80%
00:12 |27 August
pams-product-sentiment
~/Documents/UTH/DataMining/prawl_tests/subreddits/blank_slate> tmux

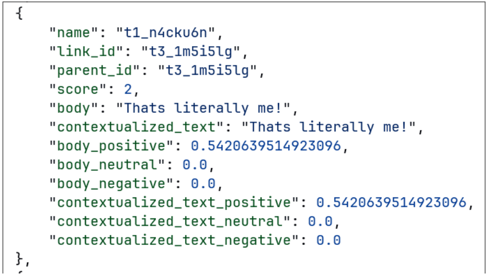
```
[0] 1:yazi  2:nu  3:node  4:yazi  5:yazi                                "report" 00:12 27-Aug-25
        },
        {
            "name": "t1_n4cm59z",
            "link_id": "t3_1m5i5lg",
            "parent_id": "t3_1m5i5lg",
            "score": 3,
            "body": "Koss. Everyone\u2019s true end game.",
            "children": [
                {
                    "name": "t1_n4d5xtq",
                    "link_id": "t3_1m5i5lg",
                    "parent_id": "t1_n4cm59z",
                    "score": 5,
                    "body": "Outdated lowe end reference.  I bought a pair of Koss and returned them.  Do they sound good?  For the price, sure, but not
above 20 bucks.",
                    "children": [
                        {
                            "name": "t1_n4ded72",
                            "link_id": "t3_1m5i5lg",
                            "parent_id": "t1_n4d5xtq",
                            "score": 1,
                            "body": "I don't know. Mine sounds really good. Which model did you go with? I have the KSC75s. Granted, I did add aftermarket
pads and run a personally tuned EQ. But in my opinion it's the best thing I've heard under $100, headphone or IEM.",
                            "children": [
                                {
                                    "name": "t1_n4df9fd",
                                    "link_id": "t3_1m5i5lg",
                                    "parent_id": "t1_n4ded72",
                                    "score": 2,
                                    "body": "They were portapros.  I shouldn't say all Koss, that's my bad.  But the porta pros don't stand up to many
cheap, 40-50 dollar modern cans.",
                                    "children": [
                                        {
NORMAL  main  prawl_tests  contextualized_hierarchical_output.json  42                    gk  20  30% 403:34  00:12
```


~/Documents/UTH/DataMining/prawl_tests/subreddits/blank_sla…
◀0%
☰52%
5%
56°C
80%
00:12 |27 August
pams-product-sentiment
~/Documents/UTH/DataMining/prawl_tests/subreddits/blank_slate> tmux

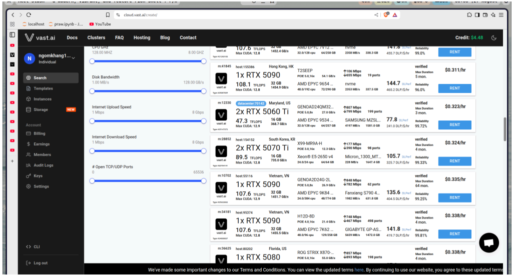
```
[0] 1:yazi  2:nu  3:node  4:yazi  5:yazi                                "report" 00:12 27-Aug-25
                                    "link_id": "t3_1m5i5lg",
                                    "parent_id": "t1_n4ded72",
                                    "score": 2,
                                    "body": "They were portapros.  I shouldn't say all Koss, that's my bad.  But the porta pros don't stand up to many
cheap, 40-50 dollar modern cans.",
                                    "children": [
                                        {
                                            "name": "t1_n4dmnrx",
                                            "link_id": "t3_1m5i5lg",
                                            "parent_id": "t1_n4df9fd",
                                            "score": 1,
                                            "body": "Yup, you nailed it. Everything I\u2019ve seen about the Porta Pros say that they\u2019re outdated.
I\u2019d highly recommend either the Koss KSC75 if you like earclips, or the Koss KPH40 if you want a more traditional headphone.",
                                            "children": [],
                                            "contextualized_text": "Koss. Everyone\u2019s true end game.. Outdated lowe end reference.  I bought a pair of
Koss and returned them.  Do they sound good?  For the price, sure, but not above 20 bucks.. I don't know. Mine sounds really good. Which model did you
go with? I have the KSC75s. Granted, I did add aftermarket pads and run a personally tuned EQ. But in my opinion it's the best thing I've heard under
$100, headphone or IEM.. They were portapros.  I shouldn't say all Koss, that's my bad.  But the porta pros don't stand up to many cheap, 40-50 dollar
modern cans.. Yup, you nailed it. Everything I\u2019ve seen about the Porta Pros say that they\u2019re outdated. I\u2019d highly recommend either the
Koss KSC75 if you like earclips, or the Koss KPH40 if you want a more traditional headphone."
                                        }
                                    ],
                                    "contextualized_text": "Koss. Everyone\u2019s true end game.. Outdated lowe end reference.  I bought a pair of Koss and
returned them.  Do they sound good?  For the price, sure, but not above 20 bucks.. I don't know. Mine sounds really good. Which model did you go with?
I have the KSC75s. Granted, I did add aftermarket pads and run a personally tuned EQ. But in my opinion it's the best thing I've heard under $100,
headphone or IEM.. They were portapros.  I shouldn't say all Koss, that's my bad.  But the porta pros don't stand up to many cheap, 40-50 dollar modern
cans."
                                }
                            ],
                            "contextualized_text": "Koss. Everyone\u2019s true end game.. Outdated lowe end reference.  I bought a pair of Koss and
returned them.  Do they sound good?  For the price, sure, but not above 20 bucks.. I don't know. Mine sounds really good. Which model did you go with?
I have the KSC75s. Granted, I did add aftermarket pads and run a personally tuned EQ. But in my opinion it's the best thing I've heard under $100,
headphone or IEM."
NORMAL  main  prawl_tests  contextualized_hierarchical_output.json  42  0  0  0          gj  20  33% 438:608  00:12
```

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
},
```



CPU GHz 128.00 MHZ – 8.00 GHZ

Disk Bandwidth 1.00 MB/s – 128.00 GB/s

Internet Upload Speed 1 Mbps – 8 Gbps

Internet Download Speed 1 Mbps – 8 Gbps

Open TCP/UDP Ports 0 – 65536

107.6 TFLOPS Max CUDA: 12.8 | 1452.4 GB/s | 32.0/128 cpu 64/258 GB | nvme 2358 MB/s 338.3 GB | 455.7 DLP/$/hr | Reliability 99.0%

m:41845 host:155386 Hong Kong, HK | 1x RTX 5090 | 108.1 TFLOPS Max CUDA: 12.8 | 32 GB 1454.9 GB/s | T2SEEP PCIE 5.0,16x 54.1 GB/s | AMD EPYC 9654 ... 48.0/192 cpu 72/290 GB | ↑106 Mbps ↓655 Mbps 19 ports | nvme 2353 MB/s 337.5 GB | 144.7 DLPerf 465.2 DLP/$/hr | verified Max Duration 5 mon. | Reliability 96.0% | $0.311/hr

m:12330 datacenter:70142 Maryland, US | 2x RTX 5060 Ti | 47.3 TFLOPS Max CUDA: 12.9 | 16 GB 368.7 GB/s | GENOAD24QM32... PCIE 5.0,8x 27.0 GB/s | AMD EPYC 9534 ... 32.0/128 cpu 64/257 GB | ↑807 Mbps ↓799 Mbps 199 ports | SAMSUNG MZSL... 4197 MB/s 1581.0 GB | 77.8 DLPerf 241.0 DLP/$/hr | verified Max Duration 3 mon. | Reliability 99.72% | $0.323/hr

m:28852 host:154152 South Korea, KR | 2x RTX 5070 Ti | 89.5 TFLOPS Max CUDA: 12.8 | 16 GB 735.0 GB/s | X99-MR9A-H PCIE 3.0,16x 12.3 GB/s | Xeon® E5-2650 v4 24.0/24 cpu 64/64 GB | ↑177 Mbps ↓390 Mbps 98 ports | Micron_1300_MT... 228 MB/s 1647.4 GB | 105.7 DLPerf 325.7 DLP/$/hr | verified Max Duration 4 mon. | Reliability 99.33% | $0.324/hr

m:10702 host:55116 Vietnam, VN | 1x RTX 5090 | 107.6 TFLOPS Max CUDA: 12.9 | 32 GB 1451.7 GB/s | GENOA2D24G-2L PCIE 5.0,8x 26.9 GB/s | AMD EPYC 9K84 ... 24.0/384 cpu 48/774 GB | ↑848 Mbps ↓782 Mbps 62 ports | Fanxiang S790 4... 1982 MB/s 631.8 GB | 135.6 DLPerf 404.5 DLP/$/hr | verified Max Duration 64 mon. | Reliability 99.25% | $0.335/hr

m:34181 host:95374 Vietnam, VN | 1x RTX 5090 | 107.6 TFLOPS Max CUDA: 12.8 | 32 GB 1455.5 GB/s | H12D-8D PCIE 4.0,16x 21.4 GB/s | AMD EPYC 7K62 ... 48.0/96 cpu 129/258 GB | ↑748 Mbps ↓467 Mbps 498 ports | GIGABYTE GP-AS... 5659 MB/s 1472.0 GB | 141.8 DLPerf 419.7 DLP/$/hr | verified Max Duration 4 mon. | Reliability 99.81% | $0.338/hr

m:36625 host:80202 Florida, US | 1x RTX 5080 | ROG STRIX X870-... PCIE 5.0,16x 55.0 GB/s | ↑857 Mbps ↓893 Mbps 198 ports | verified Max Duration 4 mon. | $0.338/hr

We've made some important changes to our Terms and Conditions. You can view the updated terms here. By continuing to use our website, you agree to these updated terms

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