

# Phân tích thảo luận công khai để khai phá hiểu biết sản phẩm

Khai thác Reddit và Tiki về lỗi sản phẩm, cảm xúc và xu hướng

Từ chuỗi thảo luận ồn ào → tín hiệu sản phẩm có thể hành động.

[TẦM NHÌN]

## Vấn đề & Giá trị

- Sản phẩm nào hỏng, vì sao và tần suất bao nhiêu — dựa trên thảo luận công khai.
- Biến văn bản Reddit và Tiki phi cấu trúc thành tín hiệu sản phẩm hữu ích.
- Xuất ra insight có thể triển khai: lỗi, xu hướng, mức độ nghiêm trọng, ví dụ.

## Mục tiêu

- Lỗi nào? Tần suất? Xu hướng? Mức độ nghiêm trọng?
- Lỗi tập trung ở đâu (theo sản phẩm/subreddit/thời gian)?
- Bàn giao pipeline lặp lại được và phát hiện rõ ràng, có ưu tiên.

## Vì Sao Các Nền Tảng Này?

- Reddit: thảo luận theo chủ đề phong phú; API công khai; hỗ trợ NLP tốt.
- Tiki (TMĐT): đánh giá có cấu trúc, xác thực mua hàng; tín hiệu thị trường VN.
- Bỏ trợ: thread giàu chữ vs đánh giá ngắn → độ phủ rộng hơn.

## So Sánh Nguồn Dữ Liệu

Nền tảng	Ưu điểm	Nhược điểm
Reddit	<ul style="list-style-type: none"><li>• API công khai (dễ truy cập)</li><li>• Thảo luận văn bản phong phú</li><li>• Hỗ trợ NLP tiếng Anh tốt</li><li>• Cộng đồng theo chủ đề “subreddits”</li></ul>	<ul style="list-style-type: none"><li>• Nhiều và lệch chủ đề</li><li>• Dữ liệu tiếng Việt hạn chế</li><li>• Hạn mức/giới hạn sử dụng API</li></ul>

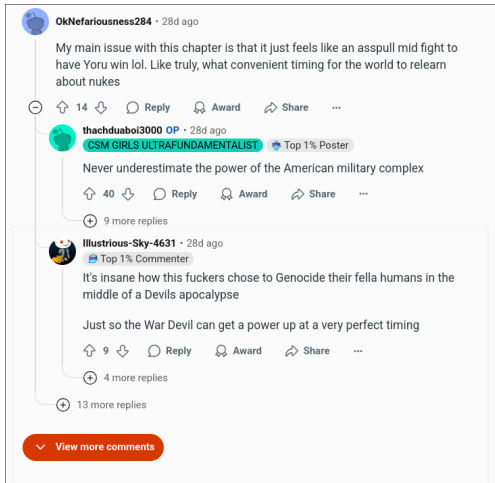
Facebook	<ul style="list-style-type: none"> <li>• Lượng người dùng lớn tại Việt Nam</li> <li>• Nhóm và cộng đồng hoạt động</li> <li>• Chủ đề đa dạng</li> </ul>	<ul style="list-style-type: none"> <li>• Nhiều bot/tài khoản rác</li> <li>• Giới hạn API</li> <li>• Khó thu thập dữ liệu sạch</li> <li>• Hỗ trợ từ thư viện NLP tiếng Anh yếu</li> </ul>
TikTok	<ul style="list-style-type: none"> <li>• Rất phổ biến với người trẻ</li> <li>• Nắm bắt xu hướng/meme tốt</li> </ul>	<ul style="list-style-type: none"> <li>• Chủ yếu nội dung media (video, ảnh)</li> <li>• Không có API chính thức</li> <li>• Thiếu cấu trúc nhóm/cộng đồng</li> <li>• Thu thập chậm (phải parse HTML)</li> <li>• Khó trích xuất văn bản liên quan</li> </ul>

## Thu Thập Dữ Liệu

- Ban đầu: PRAW (Python Reddit API Wrapper) để thử nghiệm nhanh.
  - Hạn chế: giới hạn 1.000 bài, rate limit.
- Hiện tại (lai):
  - Kho dữ liệu lịch sử Reddit (Academic Torrents) để vượt giới hạn API.
  - Dump đánh giá Tiki để đối chiếu chéo nguồn.
  - Kết quả: cửa sổ thời gian rộng hơn, khối lượng lớn hơn.

# PRAW duyệt cây bình luận như thế nào

- Reddit trả về cây bình luận với chỗ trống  
View more comments.
- Mở rộng lười, duyệt và tuần tự hóa — không có “tải xuống” một cú bấm.
- Bắt toàn bộ độ sâu khi cần; tránh bùng nổ rate limit.



# Phương Án Thay Thế Đã Xem Xét

## Pushshift.io (Pushshift API)

- Mạnh hơn PRAW, có thể lọc bài theo thời gian
- Yêu cầu quyền moderator  
→ không khả thi

## Lưu Trữ Cá Nhân

- Thu thập liên tục trong nhiều tuần
- Bất khả thi: phần cứng/thời gian, không lấy được bài cũ



shiruken MOD • 1y ago

Pushshift is [only available for use by Reddit Moderators](#). Since you are not a moderator, you cannot use Pushshift.



4



Award



Share




(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)

- Bộ dữ liệu Reddit lịch sử có thể tải về
- Tốt cho lịch sử + quy mô

### 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

End date

2010-02-26

now

☒ Download posts

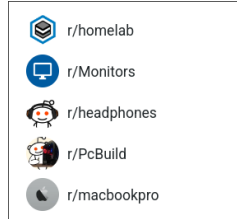
☒ Download comments

Start



# Subreddit Đã Chọn

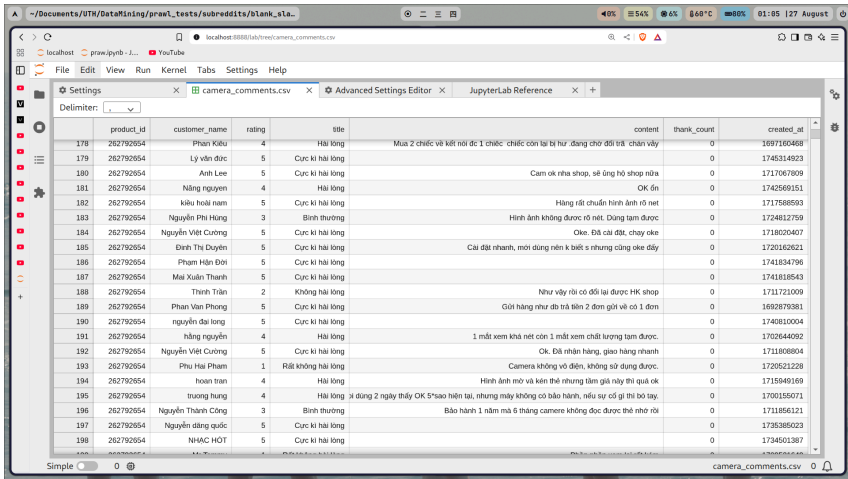
- Tổ hợp đa dạng các cộng đồng công nghệ và đời sống:
- r/macbookpro, r/GamingLaptops
- r/iphone, r/AppleWatch, r/Monitors, r/headphones, r/homelab, r/photography
- ...và vài subreddit khác về gia đình, âm thanh, và lắp ráp PC.
- tổng: 134121 bài viết, 1300190 bình luận, 2025-06-01 → 2025-07-31.



```
prawl_tests/subreddits/blank_slate on 1/ main [x!?] via 🐙
> : polars open all_posts_hanni.parquet | polars shape
+-----+-----+
| # | rows | columns |
+-----+-----+
| 0 | 134121 | 45 |
+-----+-----+

prawl_tests/subreddits/blank_slate on 1/ main [x!?] 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 dc 1 chiếc chiếc còn lại bị hư .đang chờ đổi trả .chân vầy	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 nguyên	4	Hài lòng	OK ến	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	Ok. 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 ok đấy	0	1720162621
186	262792654	Phạm Hân Đới	5	Cực kì hài lòng		0	1741834796
187	262792654	Mai Xuân Thanh	5	Cực kì hài lòng		0	1741818543
188	262792654	Thịnh Trần	2	Không hài lòng	Như vậy rồi có đổi lại được HK shop	0	1711721009
189	262792654	Phan Văn Phong	5	Cực kì hài lòng	Gửi hàng như đã 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 nguyên	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	1711808084
193	262792654	Phu Hai Phạm	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	xi đúng 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 camera không đọc được thẻ nhớ rồi	0	1711856121
197	262792654	Nguyễn dâng quốc	5	Cực kì hài lòng		0	1735385023
198	262792654	NHAC HÓT	5	Cực kì hài lòng		0	1734501387

## EDA Nổi Bật

- [Hình minh họa: chuỗi thời gian số bình luận/tuần theo subreddit]
- [Hình minh họa: phân bố độ dài (boxplot/histogram)]
- [Hình minh họa: sản phẩm được nhắc nhiều (cột) qua từ khóa/NER]
- Nhận xét (placeholder): đỉnh sau ra mắt X; r/homelab dài gấp 2×.

# Quy Trình Tiền Xử Lý

Chuyển 32 JSONL → 2 Parquet (Nushell + Polars); giữ tối đa các cột có ý nghĩa.

## Trước

Bài viết: **106** cột  
Bình luận: **69** cột



## Sau

Bài viết: **28** cột  
Bình luận: **17** cột

- Làm sạch:
  - Loại bỏ URL.
  - Lọc chỉ tiếng Anh cho phân tích ban đầu. (chỉ khoảng 5 nghìn bình luận trong 1,3 triệu, nhưng cân nhắc giữ vì mô hình đa ngôn ngữ)
- Top 30 + 20 bài ngẫu nhiên mỗi sub (có seed) để giảm thiên lệch lan truyền. (placeholder)

~/Documents/ETH/DataMining/psml\_tests/subreddits/51ank\_sla... 40% 54% 65°C 88% 00:25 127 August

localhost prml.py:16 A... YouTube

File Edit View Run Kernel Tabs Settings Help

Settings posts\_and\_comments.csv Advanced Settings Editor JupyterLab Reference

Delimiter: , v

	name	subreddit	score	link_id	parent_id	body	is_post
789	c3_1ia75Q	iphone	6			g to reveal this phone and get the new iPhone coming in later this year. Just need options	true
790	c3_1mcyauX	iphone	1			sed cards for recording longer videos, but haven't seen anything about the 16 base model.	true
791	c3_1U5a9t	laptops	112			I think I messed up while replacing the screen	true
792	c3_1hu8fy	laptops	110			air costs in Europe are so high that it's not even worth fixing. Lenovo, this is unacceptable!	true
793	c3_1s7j5e	laptops	118			ve old it is, but it's a laptop with a floppy drive. I figured someone here might appreciate this	true
794	c3_1mf8uq	laptops	247			I gooch on its lower side, but it left this moon matching patch, will it eventually fade away?	true
795	c3_1mko9eA	iphone	1			iPhone 15 Pro Max Pro Foto	true
796	c3_1m9g2m	iphone	1			Shift or True Tone but when I toggle those on and off nothing changes. What is happening?	true
797	c3_1U7m6l	laptops	0			Help regarding cly laptop!!	true
798	t1_mve6d4	AppleWatch	85	c3_1J0vtn	c3_1J0vtn	Well the picture is actually mine but the idea	false
799	t1_mve65oe	AppleWatch	-7	c3_1J0vtn	c3_1J0vtn	you should wrap the watch around the phone horizontally but I like the idea!	false
800	t1_mve70q9	AppleWatch	536	c3_1J0vtn	c3_1J0vtn	ap a picture from the watch which is handy for capturing a serial number or model number.	false
801	t1_mve7568	AppleWatch	176	c3_1J0vtn	c3_1J0vtn	In my days selfie cameras came with a little mirror to be able to see yourself.	false
802	t1_mve693z	AppleWatch	1	c3_1J0vtn	c3_1J0vtn	but why?	false
803	t1_mve68k1	AppleWatch	9	c3_1J0vtn	t1_mve683z	So you can use the much better back camera and also see what you are taking	false
804	t1_mve66a3	AppleWatch	17	c3_1J0vtn	t1_mve70q9	Pretty much the only thing I use it for as well.	false
805	t1_mve6b7w	AppleWatch	103	c3_1J0vtn	c3_1J0vtn	Well if you can get it to work reliably, its always been trouble for me to see once launched.	false
806	t1_mve6tq9	AppleWatch	20	c3_1J0vtn	c3_1J0vtn	I don't see the viewfinder image on my watch, just the controls. Anyone know why?	false
807	t1_mve6rta	AppleWatch	2	c3_1J0vtn	c3_1J0vtn	What's the app called?	false
808	t1_mve6erp	AppleWatch	60	c3_1J0vtn	c3_1J0vtn	to keep eyes on my truck dash to see where the tire pressure sensors are as I fill the tires.	false
809	t1_mve714l	AppleWatch	1	c3_1J0vtn	c3_1J0vtn	Or just use the other camera?	false

Simple 0 posts\_and\_comments.csv

# Thử Nghiệm Ngữ Cảnh (SA Theo Bình Luận)

- So sánh phân loại có ngữ cảnh cha vs theo-bình-luận trên tập nhãn nhỏ.
- Kết quả: theo bình luận tránh cắt 512 token và giảm nhiễu ngữ cảnh.

contextualized\_hierarchical\_output.json: minh họa bình luận kèm ngữ cảnh cây của nó.

```

A /Documents/NTT/DataMining/prawl_tests/vabrdists/vLink_vla. 48% 52% 6% 85% 80:12 127 August
para-prod: context
[0] 1:razi 2:ru 3:node 4:razi 5:razi "report" 80:12 27-Aug-29
{
  "name": "t1_n4c0992",
  "link_id": "t3_in5151g",
  "parent_id": "t3_in5151g",
  "score": 3,
  "body": "Koss. Everyone's2019e true and game.",
  "children": [
    {
      "name": "t1_n4d5xtq",
      "link_id": "t3_in5151g",
      "parent_id": "t3_in5151g",
      "score": 5,
      "body": "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.",
      "children": [
        {
          "name": "t1_n4d672",
          "link_id": "t3_in5151g",
          "parent_id": "t3_in5151g",
          "score": 1,
          "body": "I don't know. Mine sounds really good. Which model did you go with? I have the K5175s. 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_n4d9f9d",
              "link_id": "t3_in5151g",
              "parent_id": "t3_in5151g",
              "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_n4e099d",
                  "link_id": "t3_in5151g",
                  "parent_id": "t3_in5151g",
                  "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."
                }
              ]
            }
          ]
        }
      ]
    }
  ]
}

```

```

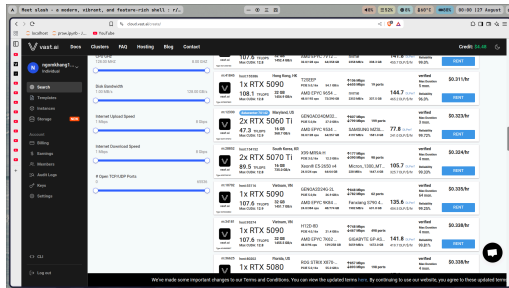
A /Documents/NTT/DataMining/prawl_tests/vabrdists/vLink_vla. 48% 52% 6% 85% 80:12 127 August
para-prod: context
[0] 1:razi 2:ru 3:node 4:razi 5:razi "report" 80:12 27-Aug-29
{
  "name": "t1_n4c0992",
  "link_id": "t3_in5151g",
  "parent_id": "t3_in5151g",
  "score": 3,
  "body": "Koss. Everyone's2019e true and game.",
  "children": [
    {
      "name": "t1_n4d5xtq",
      "link_id": "t3_in5151g",
      "parent_id": "t3_in5151g",
      "score": 5,
      "body": "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.. I don't know. Mine sounds really good. Which model did you go with? I have the K5175s. 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.. Yap, you nailed it. Everything I've2019e seen about the Porta Pros say that they're2019e outdated. I've2019e highly recommend either the Koss K5C75s if you like earclips, or the Koss KPM40 if you want a more traditional headphone.",
      "children": [
        {
          "name": "t1_n4d672",
          "link_id": "t3_in5151g",
          "parent_id": "t3_in5151g",
          "score": 1,
          "body": "I don't know. Mine sounds really good. Which model did you go with? I have the K5C75s. 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.. Yap, you nailed it. Everything I've2019e seen about the Porta Pros say that they're2019e outdated. I've2019e highly recommend either the Koss K5C75s if you like earclips, or the Koss KPM40 if you want a more traditional headphone."
        }
      ]
    }
  ]
}

```

# Phân Tích Cảm Xúc (SA)

- Cách tiếp cận: Transformer tiền huấn luyện thay vì baseline đơn giản (VADER, TextBlob).
- Mô hình: lxyuan/distilbert-base-multilingual-cased-sentiments-student
  - Vì sao: hỗ trợ đa ngôn ngữ tốt, nhẹ, hiệu năng zero-shot mạnh.
- Thực thi: thuê GPU trên vast.ai để suy luận quy mô lớn.
- Bước tiếp theo: fine-tune trên tập nhãn theo miền.

```
{  
  "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  
}
```



## Mô Hình Chủ Đề (Dự Kiến)

- Mục tiêu: khám phá chủ đề và lỗi chính theo subreddit.
- Phương pháp: đánh giá BERTopic so với LDA/NMF truyền thống.
- Quy trình: tiền xử lý với stopword theo miền, lemmatization.
- Đầu ra: chủ đề top, bình luận tiêu biểu và đường xu hướng.
- Mục tiêu: ngưỡng coherence  $c_v$  và chủ đề ổn định qua mẫu con. (placeholder)
- [Hình: bản đồ khoảng cách chủ đề (BERTopic mock)]
- [Hình: bảng top từ cho 2 chủ đề mẫu]



## Công Cụ CLI (Dự Kiến)

- Pipeline cho phân tích lặp lại:
- ingest: từ dữ liệu thô sang Parquet.
- clean: lọc và chuẩn hóa dữ liệu.
- sentiment: chạy phân tích cảm xúc hàng loạt.
- topics: huấn luyện và áp dụng mô hình chủ đề.
- report: tổng hợp kết quả và xuất.

## **Rủi Ro & Giảm Thiểu**

- Thiên lệch: lấy mẫu ngẫu nhiên trên nhiều subreddit đa dạng.
- Trôi thời gian: bao gồm các lát thời gian để so sánh cohort.
- Tái lập: ghim môi trường, chạy theo cấu hình, lưu seed.
- Đạo đức: dùng dữ liệu công khai, tuân thủ ToS, kết quả đã tổng hợp.

## Bước Tiếp Theo

- Hoàn thiện trực quan hóa EDA.
- Chạy phân tích cảm xúc ở quy mô lớn.
- Thử nghiệm và chọn phương pháp mô hình chủ đề.
- Xây và demo workflow CLI cốt lõi.
- Bắt đầu gán nhãn để fine-tune mô hình cảm xúc.

## Hỏi & Đáp