

Phân tích thảo luận công khai để tìm hiểu sâu về sản phẩm

Khai thác Reddit và Tiki để tìm các vấn đề, tình cảm và xu hướng của sản phẩm

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

[Hình ảnh giữ chỗ: ảnh ghép các đoạn bình luận ẩn danh với chú thích]

Vấn đề & Giá trị

- Sản phẩm nào bị hỏng, tại sao và tần suất—sử dụng các cuộc thảo luận công khai.
- Biến văn bản Reddit và Tiki phi cấu trúc thành các tín hiệu sản phẩm có thể hành động.
- Cung cấp thông tin chi tiết mà các nhóm có thể sử dụng: vấn đề, xu hướng, mức độ nghiêm trọng, ví dụ.

Mục tiêu

- Vấn đề nào? Tần suất? Xu hướng? Mức độ nghiêm trọng?
- Các vấn đề tập trung ở đâu (theo sản phẩm/subreddit/thời gian)?
- Các phương pháp đáng tin cậy như thế nào so với các đường cơ sở?
- Cung cấp một quy trình có thể lặp lại và các phát hiện rõ ràng, được ưu tiên.

Tại sao lại là những nền tảng này?

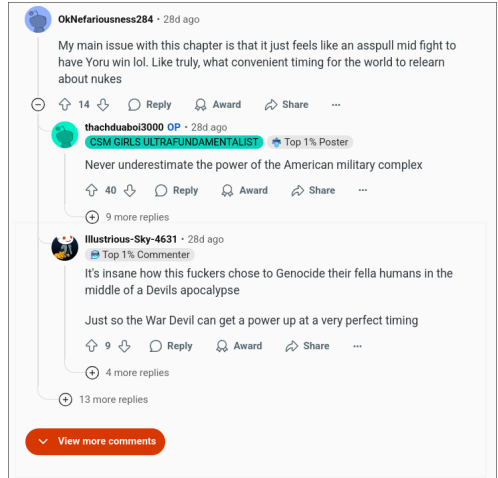
- Reddit: thảo luận theo chuỗi phong phú; API công khai; hỗ trợ NLP mạnh mẽ.
- Tiki (thương mại điện tử): đánh giá có cấu trúc, được xác minh mua hàng; tín hiệu thị trường Việt Nam.
- Bổ sung cho nhau: các chuỗi giàu văn bản so với các bài đánh giá ngắn → phạm vi bao phủ rộng hơn.
- [Hình ảnh giữ chỗ: thẻ ưu/nhược điểm hai cột + bản đồ phạm vi bao phủ]

Thu thập dữ liệu

- **Ban đầu:** PRAW (Python Reddit API Wrapper) để tạo mẫu.
 - **Hạn chế:** Giới hạn 1.000 bài đăng, giới hạn tốc độ.
- **Hiện tại (Kết hợp):**
 - Kho lưu trữ lịch sử Reddit (Academic Torrents) để bỏ qua giới hạn API.
 - Dữ liệu đánh giá Tiki để xác thực chéo nguồn.
 - **Kết quả:** Cửa sổ thời gian rộng hơn, khối lượng lớn hơn.
- [Hình ảnh giữ chỗ: dải băng thời gian PRAW → Kho lưu trữ → Tiki]
- [Hình ảnh giữ chỗ: sơ đồ quy trình với biểu tượng giới hạn tốc độ trên PRAW; khóa bộ lọc thời gian]

Cách PRAW duyệt qua các bình luận

- Reddit trả về cây bình luận với các trình giữ chỗ Xem thêm bình luận.
- Chúng tôi mở rộng một cách lười biếng, duyệt qua và tuần tự hóa—không có “tải xuống” bằng một cú nhấp chuột.
- Ghi lại toàn bộ chiều sâu khi cần thiết; tránh bùng nổ giới hạn tốc độ.



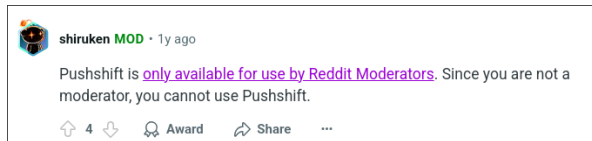
Các phương án thay thế đã được xem xét

Pushshift.io (Pushshift API)

- Mạnh hơn PRAW, có thể lọc bài đăng theo thời gian
- Yêu cầu trạng thái người kiểm duyệt
→ không khả thi

Lưu trữ cá nhân

- Thu thập liên tục trong nhiều tuần
- Không thực tế: phần cứng/thời gian, không thể ghi lại các bài đăng cũ hơn




(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 xuống
- 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

2010-02-26

End date

now

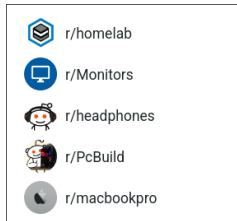
☒ Download posts

☒ Download comments

Start

Các subreddit đã chọn

- Một sự kết hợp đa dạng của các cộng đồng công nghệ và lối sống:
- r/macbookpro, r/GamingLaptops
- r/iphone, r/AppleWatch, r/Monitors, r/headphones, r/homelab, r/photography
- ...và một số cộng đồng khác về nhà cửa, âm thanh và xây dựng PC.
- tổng cộng : 134121 bài đăng, 1300190 bình luận, 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 |
+-----+
```

Điểm nổi bật của EDA

- [Hình ảnh giữ chỗ: chuỗi thời gian của bình luận/tuần theo subreddit]
- [Hình ảnh giữ chỗ: phân phối độ dài (biểu đồ hộp hoặc biểu đồ tần suất)]
- [Hình ảnh giữ chỗ: các đề cập sản phẩm hàng đầu (thanh) qua từ khóa/NER]
- Điểm chính (giữ chỗ): số lượng tăng đột biến sau khi ra mắt X; bình luận trên r/homelab dài gấp 2 lần.

Quy trình tiền xử lý

[dữ liệu trước và sau]

- **Nạp dữ liệu:** Nạp JSONL vào các lược đồ Parquet đã nhập bằng Polars & Nushell.
- **Làm sạch:**
 - Xóa URL, loại bỏ đánh dấu, chuẩn hóa khoảng trắng.
 - Lọc nội dung chỉ bằng tiếng Anh để phân tích ban đầu.
- **Lựa chọn mô hình hóa:** Lập mô hình từng bình luận riêng lẻ sau khi các thử nghiệm cho thấy ngữ cảnh gốc làm ô nhiễm tín hiệu tình cảm.
- Ngẫu nhiên hóa 100 bài đăng/sub (có hạt giống) để giảm sai lệch do lan truyền. (giữ chỗ)
- [Hình ảnh giữ chỗ: đoạn văn bản trước/sau (đã xóa URL/biểu tượng cảm xúc)]
- [Hình ảnh giữ chỗ: biểu đồ tròn ngôn ngữ (giữ lại so với loại bỏ)]
- [Hình ảnh giữ chỗ: sơ đồ căn chỉnh lược đồ JSONL → Parquet đã nhập]

Thử nghiệm ngữ cảnh (SA mỗi bình luận)

- So sánh phân loại theo ngữ cảnh gốc và mỗi bình luận trên một tập dữ liệu nhỏ được gán nhãn.
- Kết quả: mỗi bình luận tránh được việc cắt ngắn 512 mã thông báo và ô nhiễm ngữ cảnh.
- [Hình ảnh giữ chỗ: biểu đồ cột cắt bỏ (F1 có và không có ngữ cảnh gốc)]

Phân tích tình cảm (SA)

- **Phương pháp:** Các máy biến áp được đào tạo trước so với các đường cơ sở đơn giản hơn (VADER, TextBlob).
- **Mô hình:** lxyuan/distilbert-base-multilingual-cased-sentiments-student
 - **Tại sao:** Hỗ trợ đa ngôn ngữ tốt, nhẹ, hiệu suất không cần đào tạo mạnh mẽ.
- **Thực thi:** Thuê GPU trên vast.ai để suy luận quy mô lớn.
- **Bước tiếp theo:** Tinh chỉnh trên một tập dữ liệu được gắn nhãn dành riêng cho miền.
- [Hình ảnh giữ chỗ: thanh so sánh mô hình (VADER/TextBlob/DistilBERT) trên tập được gắn nhãn]
- [Hình ảnh giữ chỗ: thanh thông lượng/chi phí CPU so với GPU trên vast.ai]
- [Hình ảnh giữ chỗ: ma trận nhầm lẫn hoặc đường cong hiệu chuẩn]
- [Bảng điều khiển giữ chỗ: 2–3 ví dụ định tính với các nhãn được dự đoán]

Lập mô hình chủ đề (Đã lên kế hoạch)

- **Mục tiêu:** Khám phá các chủ đề và vấn đề chính trên mỗi subreddit.
- **Phương pháp:** Đánh giá BERTopic so với các phương pháp truyền thống (LDA/NMF).
- **Quy trình:** Tiền xử lý với các từ dừng miễn, bổ sung.
- **Đầu ra:** Các chủ đề hàng đầu, bình luận đại diện và đường xu hướng.
- Mục tiêu: ngưỡng c_v mạch lạc và các chủ đề ổn định trên các mẫu con. (giữ chỗ)
- [Hình ảnh giữ chỗ: bản đồ khoảng cách giữa các chủ đề (mô hình BERTopic)]
- [Hình ảnh giữ chỗ: bảng các từ hàng đầu cho 2 chủ đề mẫu]

Công cụ CLI (Đã lên kế hoạch)

- Một quy trình để phân tích có thể lặp lại:
- ingest: Dữ liệu thô sang Parquet.
- clean: Lọc và chuẩn hóa dữ liệu.
- sentiment: Chạy phân tích tình cảm hàng loạt.
- topics: Đào tạo và áp dụng các mô hình chủ đề.
- report: Tổng hợp kết quả và xuất.

Rủi ro & Giảm thiểu

- **Sai lệch:** Lấy mẫu ngẫu nhiên trên nhiều subreddit đa dạng.
- **Trôi dạt theo thời gian:** Bao gồm các lát thời gian để so sánh các nhóm.
- **Khả năng tái tạo:** Môi trường được ghim, chạy theo cấu hình và các hạt giống được lưu trữ.
- **Đạo đức:** Sử dụng dữ liệu công khai, tuân theo Điều khoản dịch vụ của nền tảng, tổng hợp kết quả.

Các bước tiếp theo

- Hoàn thiện trực quan hóa EDA.
- Chạy phân tích tình cảm ở quy mô lớn.
- Thử nghiệm và chọn một phương pháp lập mô hình chủ đề.
- Xây dựng và trình diễn quy trình làm việc CLI cốt lõi.
- Bắt đầu gắn nhãn để tinh chỉnh mô hình tình cảm.

Hỏi & Đáp