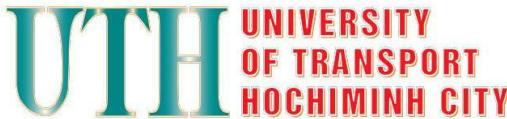


BỘ GIAO THÔNG VẬN TẢI
TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI THÀNH PHỐ HỒ CHÍ MINH



KHAI THÁC DỮ LIỆU BÁO CÁO CUỐI KỲ

PHÂN TÍCH NỘI DUNG TRÒ CHUYỆN TRÍCH XUẤT VĂN ĐỀ PHÔ BIÊN VỀ TÀI SẢN

Giảng Viên Hướng Dẫn: TS Trần Thế Vinh

Họ và Tên Sinh Viên Thực Hiện và MSSV:

Phạm Hoàng Thiện - 051205000064

Ngô Minh Khang - 086250511340

Nguyễn Văn Mạnh - 027205000040

Nguyễn Văn Quang - 038205004237

Lê Huỳnh Cao Dương - 079204051017

Nguyễn Ngọc Anh Thư

Nguyễn Đăng Khoa

Thành phố Hồ Chí Minh, ngày 18 tháng 9 năm 2024

Mục lục

1. Giới thiệu & Phạm vi dự án	3
2. Tóm tắt điều hành	3
3. Nguồn dữ liệu & Thu thập	3
3.1. Lý do chọn Reddit & Tiki	3
3.2. Quy trình Arctic Shift	4
3.3. Ảnh chụp corpus	5
4. Kỹ thuật dữ liệu & tích hợp	5
4.1. Cách PRAW duyệt bình luận	5
5. Phân tích dữ liệu thăm dò (EDA)	6
6. Phân tích sentiment — Kế hoạch ban đầu & quyết định	10
7. Topic modeling & pipeline khía cạnh	11
7.1. Tiền xử lý	11
7.2. Chọn đặc trưng & giảm chiều	11
7.3. Quy trình phân cụm	12
7.4. Trích khía cạnh & kết hợp sentiment	12
7.5. Artefact trọng tâm	12
8. Deliverables & hướng dẫn sử dụng	13
9. Hạn chế & hướng phát triển	13
10. Phụ lục	14
10.1. Tài sản cốt lõi	14
10.2. Ví dụ tóm tắt console (rút gọn)	14

1. Giới thiệu & Phạm vi dự án

PIPELINE : Thu thập Reddit + Tiki → Chuẩn hóa Parquet → EDA → Topic modeling (TF-IDF → SVD → KMeans) + Aspect sentiment

Link github: <https://github.com/khangnm1340/data-mining-nhom-4-reddit-sentiment>

2. Tóm tắt điều hành

- Chúng em khai thác 16 cộng đồng Reddit tập trung vào phần cứng (kèm các đánh giá Tiki bổ sung) để phát hiện vấn đề sản phẩm lặp lại, ưu/nhược điểm và xu hướng chủ đề.
- Công cụ thu thập Arctic Shift vượt giới hạn API, cung cấp dữ liệu nhiều năm được lưu dưới dạng Parquet đồng nhất.
- CLI (`run_pipeline.py`) chuyển tập văn bản đã lọc thành đặc trưng TF-IDF/SVD, cụm KMeans và các artifact sentiment theo khía cạnh.
- Kế hoạch fine-tune transformer đã được lập nhưng tạm hoãn sau phân tích chi phí/lợi ích; VADER đảm nhiệm pipeline sentiment bàn giao.
- Đầu ra gồm joblib/JSON tái sử dụng, dashboard PNG và tư liệu thuyết trình trong `extra/`.

3. Nguồn dữ liệu & Thu thập

3.1. Lý do chọn Reddit & Tiki

- Reddit cung cấp chủ đề giàu văn bản và được cộng đồng kiểm duyệt
- Tiki bổ sung phản hồi mua hàng xác thực từ thị trường Việt Nam.
- Facebook và TikTok bị đặt thấp ưu tiên do thiếu API, nhiều bot và thiên về nội dung ảnh và video(mà dự án này làm NPL).
- Danh sách subreddit bao trùm laptop, điện thoại, âm thanh, nhà thông minh, nhiếp ảnh, lắp ráp PC và công thái học — nắm bắt góc nhìn người chơi và người xử lý sự cố.



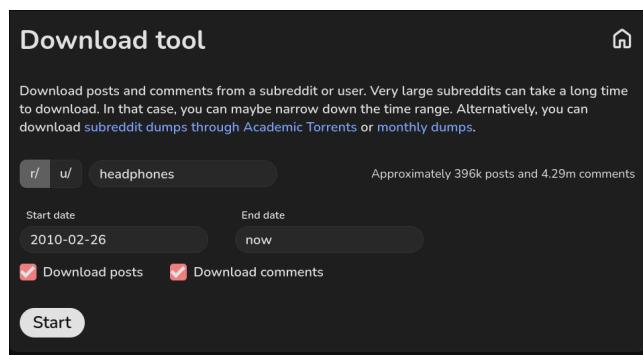
Hình 1: Danh sách subreddit tiêu biểu bao phủ homelab, màn hình, cộng đồng audiophile và hệ sinh thái Apple.

product_id	customer_name	rating	title	content	thank_count	created_id
276	Phan Khoa	4	Hai long	Màu 2 (màu) và MÀU NỔ (nhạc) (nhạc) con ta là từ từ đang thử 99% (chưa vậy)	0	2697350468
379	Lý Văn Đức	5	Cực kì hài long	Cám ơn nhạc, sẽ ủng hộ - chúc bạn	0	1748514823
380	Avin Lee	5	Cực kì hài long	Cám ơn nhạc, sẽ ủng hộ - chúc bạn	0	1717694949
381	Nikita Nguyễn	5	Ha long	Chắc chắn sẽ ủng hộ	0	1733055533
382	262792054	5	Cực kì hài long	Hàng chuẩn hình ảnh rõ nét	0	1717585953
383	262792054	5	Bình thường	Hình ảnh không được rõ nét. Dừng làm được	0	1724812759
384	Nguyễn Việt Cường	3	Cực kì hài long	Ông đã có đó, chúc vui	0	1718020407
385	Thien Phuoc	5	Cực kì hài long	Các đk như, thời gian đến nay là 5 tháng cũng chưa đầy	0	1730504943
386	262792054	5	Cực kì hài long	0	1741834796	
387	262792054	5	Cực kì hài long	0	1718105453	
388	262792054	2	Không hài long	Hứa với tôi có đổi lại được 100% shop	0	1717121009
389	Phạm Minh Hùng	5	Cực kì hài long	Ôn tập hàng ngày để bắt đầu 20 bài giờ và sẽ đi	0	1682871981
390	Nguyễn Văn Hùng	5	Cực kì hài long	0	1743305004	
391	262792054	4	Hai long	1 mês sau khi kết hôn 3 tháng xem chất lượng tạm được	0	1702844982
392	262792054	5	Cực kì hài long	Ok. Đã nhận hàng, giao hàng chậm	0	1713084004
393	Phu Hu Pham	1	Rất không hài long	Đang chờ nhận hàng, không rõ ràng	0	1720521228
394	262792054	5	Ha long	nhận định mà nó kèm theo là không rõ ràng	0	1717585953
395	Truong hung	4	Hai long	nhưng thấy 2 ngày OK. Phản hồi tốt, nhưng thấy không có bảo hành, nhà cung cấp bị kêu oai	0	1709550071
396	262792054	3	Bình thường	Bảo hành 2 năm mà ở trong camera không đọc được thẻ nhớ là	0	1713864323
397	Nguyễn Thành Cường	5	Cực kì hài long	0	1733860023	
398	262792054	5	Cực kì hài long	0	1744012887	

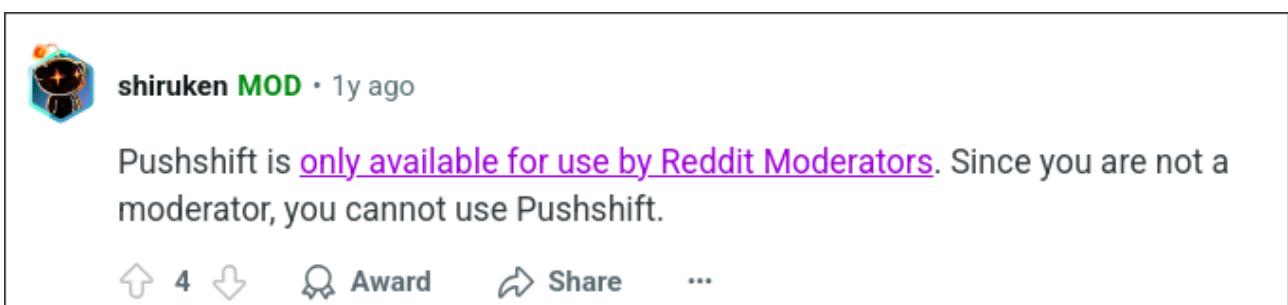
Hình 2: Mẫu đánh giá Tiki cho máy ảnh — chấm điểm có cấu trúc kết hợp với phản hồi tiếng Việt bổ sung cho Reddit.

3.2. Quy trình Arctic Shift

- Các gói dữ liệu lịch sử (Academic Torrents) cấp nguồn cho bộ thu Arctic Shift, loại bỏ giới hạn 1 000 bài/subreddit của PRAW.
- Dump đánh giá Tiki hỗ trợ Reddit để đổi chiếu chéo khi khả dụng.
- Dữ liệu được chuẩn hóa bằng Polars/Nushell từ 32 JSONL sang posts.parquet và comments.parquet, đảm bảo schema thống nhất.



Hình 3: Công cụ Arctic Shift tải toàn bộ lịch sử r/headphones — vượt trán API để có chuỗi thời gian dài hạn.



Hình 4: Pushshift giờ chỉ dành cho moderator, càng củng cố nhu cầu tự lưu trữ.

Hình 5: Xem trước Parquet bài đăng và bình luận hợp nhất, đảm bảo ID thống nhất trước khi mô hình hóa.

3.3. Ảnh chụp corpus

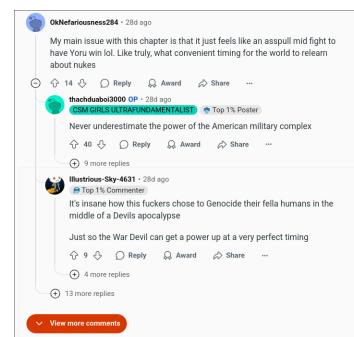
Split	Rows	Columns	Unique subs
Posts	134121	27	16
Comments	1300190	16	16

4. Kỹ thuật dữ liệu & tích hợp

- Lọc sử dụng khớp alias không phân biệt hoa thường (ví dụ “hd600”, “flio ft1”), kèm tùy chọn giới hạn theo subreddit.
- Token thương hiệu/alias được đưa vào danh sách stopword để không chiếm ưu thế trong TF-IDF.
- Giữ cân bằng bài đăng/bình luận bằng cách đồng bộ các khóa (name, subreddit, link_id, parent_id) trước khi gộp.
- run_pipeline.py ghi log metadata tái lập (command-line, phiên bản package) và xuất toàn bộ artifact trung gian vào thư mục chỉ định.

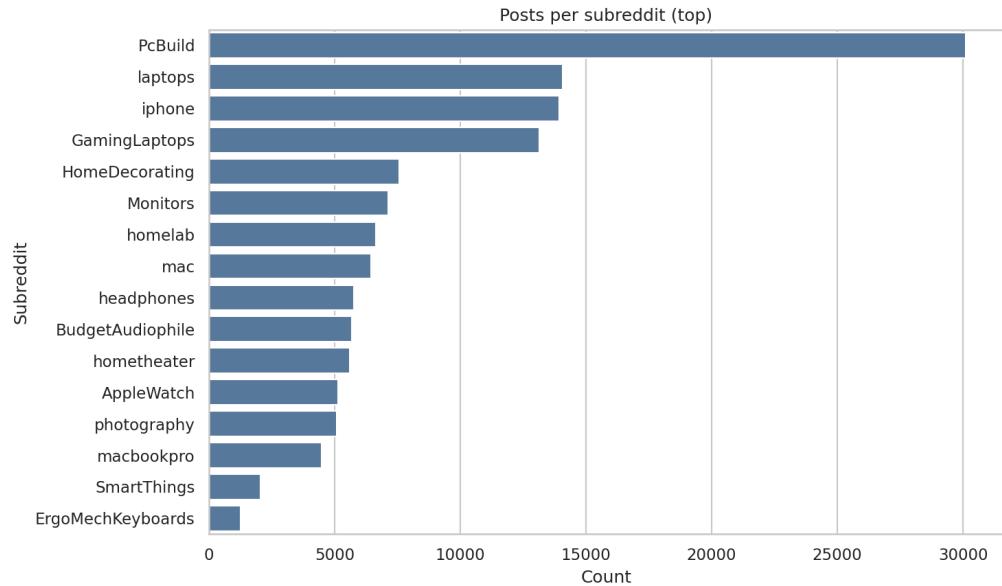
4.1. Cách PRAW duyệt bình luận

- Reddit trả về cây bình luận với nút View more comments.
- Phải mở rộng lối, duyệt và tuần tự hóa thủ công — không có chế độ tải một lần.
- Thu thập đủ mọi tầng sâu khi cần nhưng vẫn tránh vượt rate limit.

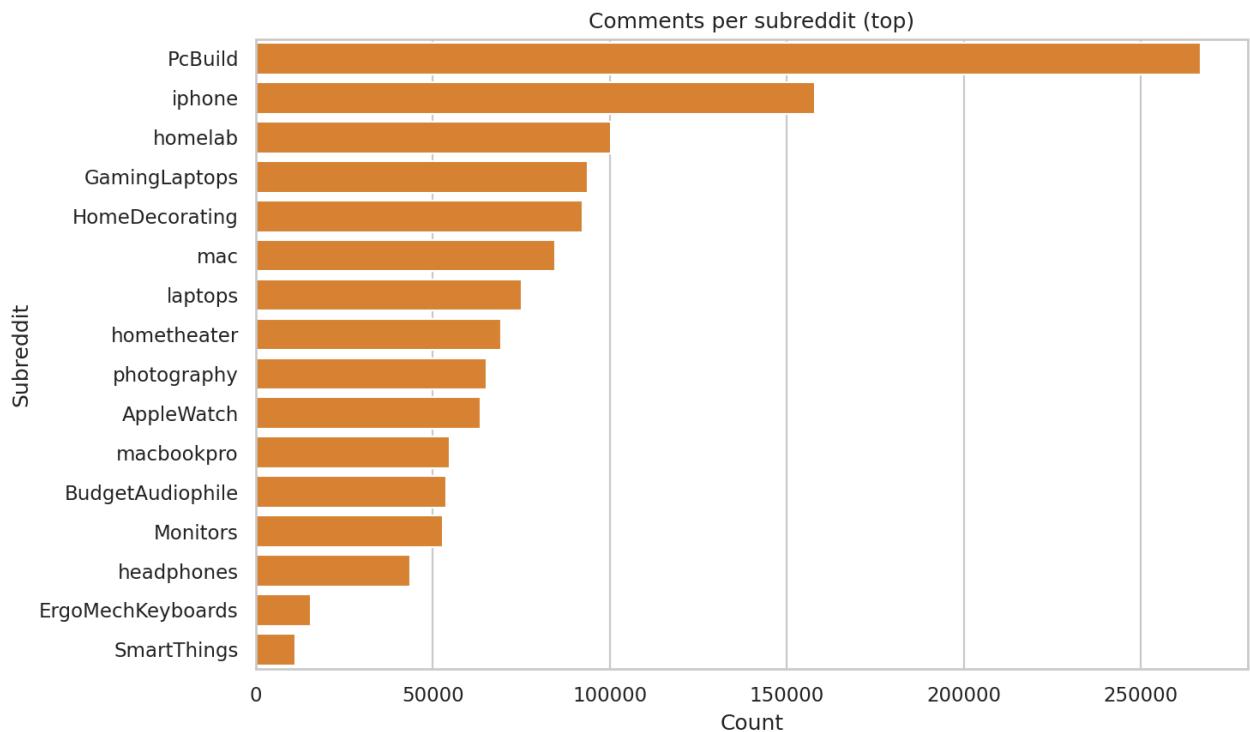


5. Phân tích dữ liệu thăm dò (EDA)

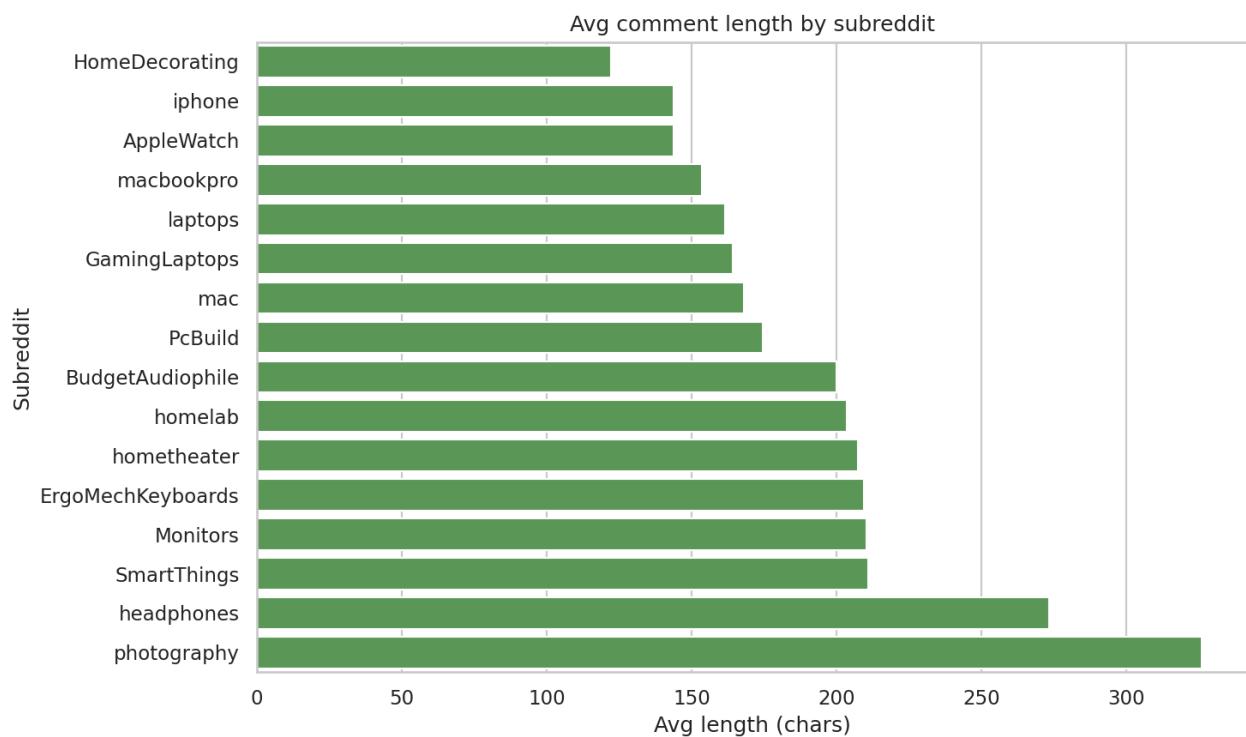
- Tài sản EDA nằm trong eda/ (PNG + CSV) để tái sử dụng nhanh cho slide và dashboard.



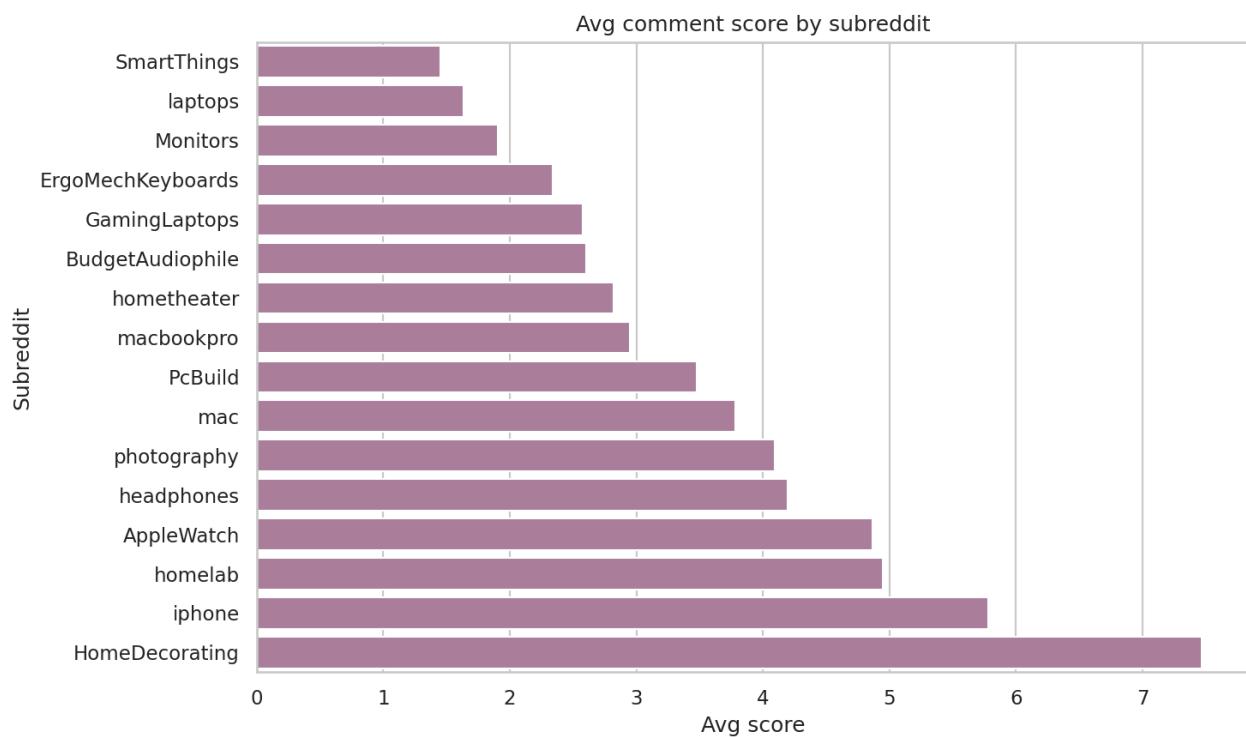
Hình 6: Số bài theo subreddit — PcBuild, iPhone và GamingLaptops chiếm sản lượng lớn.



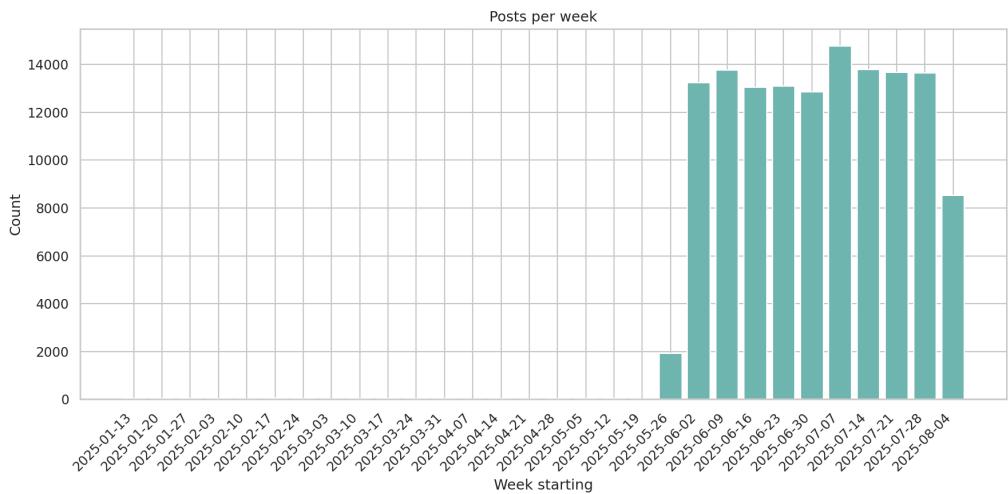
Hình 7: Số bình luận: cộng đồng xử lý sự cố (PcBuild, homelab) dẫn đầu về tương tác.



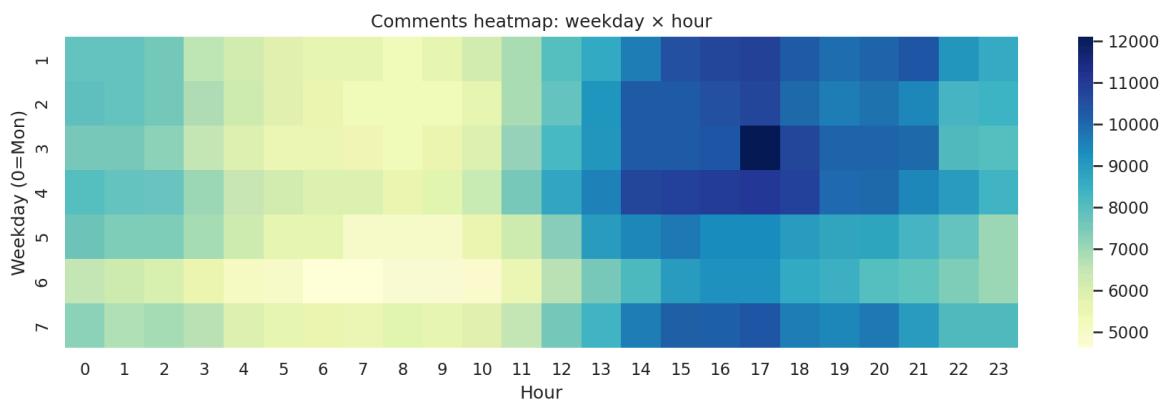
Hình 8: Độ dài bình luận trung bình cho thấy thảo luận kỹ thuật sâu ở r/macbookpro và r/AppleWatch.



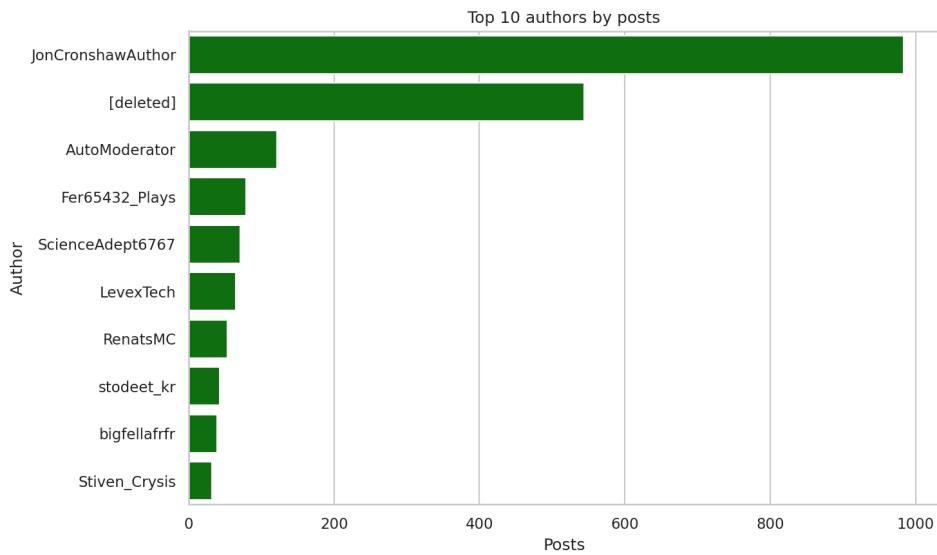
Hình 9: Điểm bình luận trung bình khá thấp (<2.5), nhấn mạnh nhu cầu insight văn bản ngoài karma.



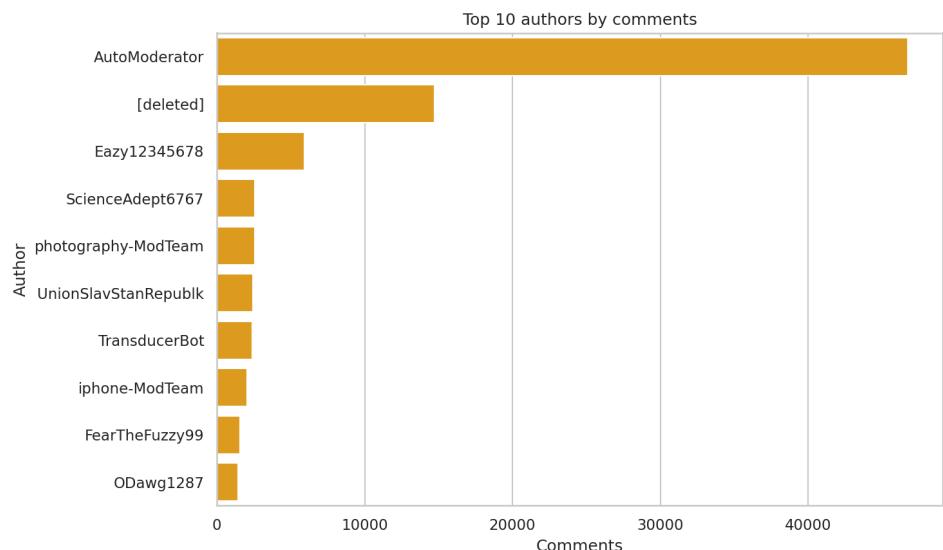
Hình 10: Sản lượng bài đăng theo tuần phát hiện các đợt ra mắt phần cứng theo mùa.



Hình 11: Heatmap theo ngày/giờ — buổi tối và cuối tuần thúc đẩy bàn luận.



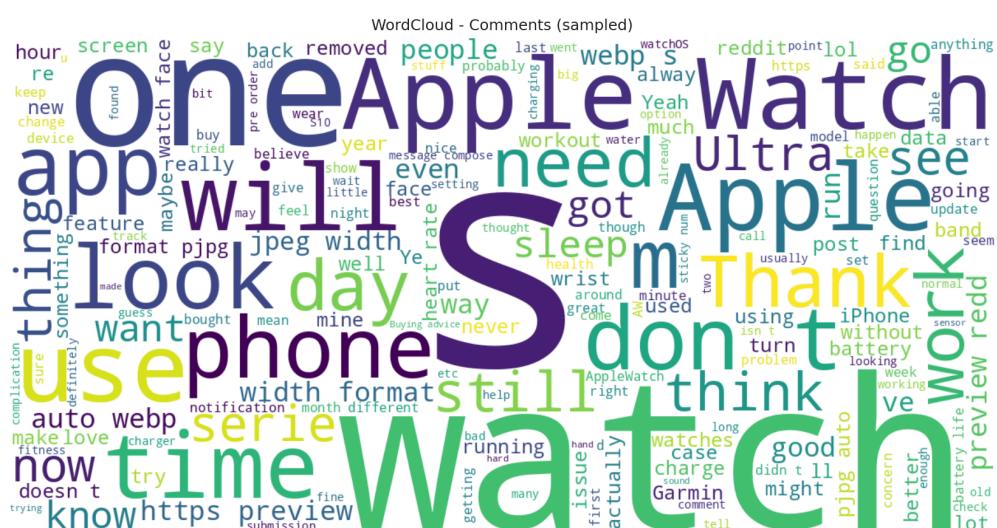
Hình 12: Tác giả bài đăng hàng đầu — chỉ ra phần lớn những account đăng nhiều nhất là bot và những account bị xóa sẽ quy chung về [deleted].



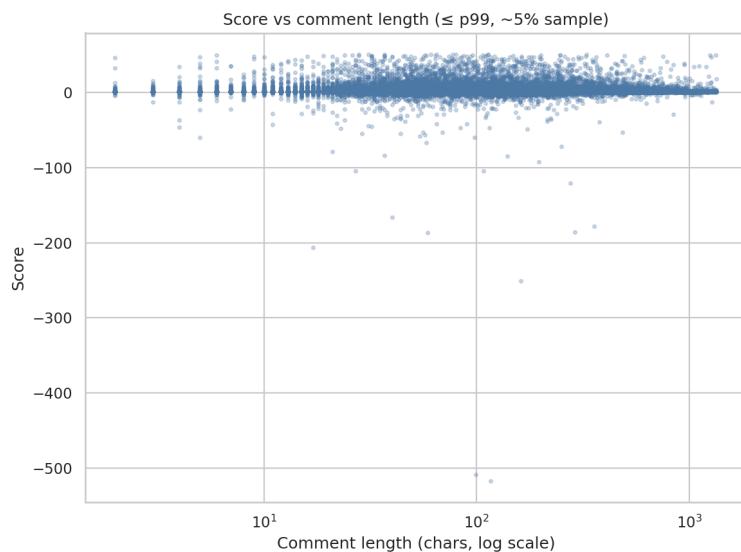
Hình 13: Tác giả bình luận hàng đầu — nhận diện người thiên về hỗ trợ và điều phối.



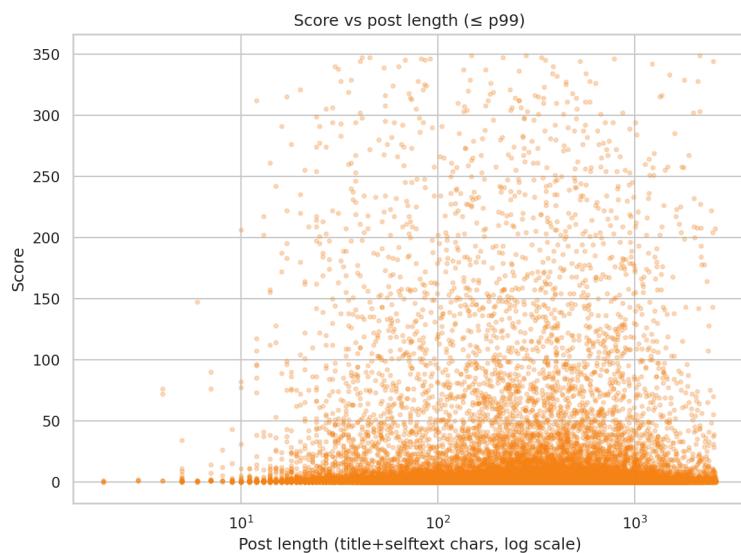
Hình 14: Word cloud tiêu đề nêu bật nhóm sản phẩm chủ đạo và chủ đề xử lý sự cố.



Hình 15: Word cloud bình luận cho thấy các cụm từ cảm xúc lặp lại và thuật ngữ linh kiên.



Hình 16: Biểu đồ score so với độ dài bình luận — phản hồi dài hơn có xu hướng nhận karma cao trong chủ đề hỗ trợ.



Hình 17: Biểu đồ score so với độ dài bài đăng — tin đồn ngắn gọn và bài review chi tiết đều thu hút người xem.

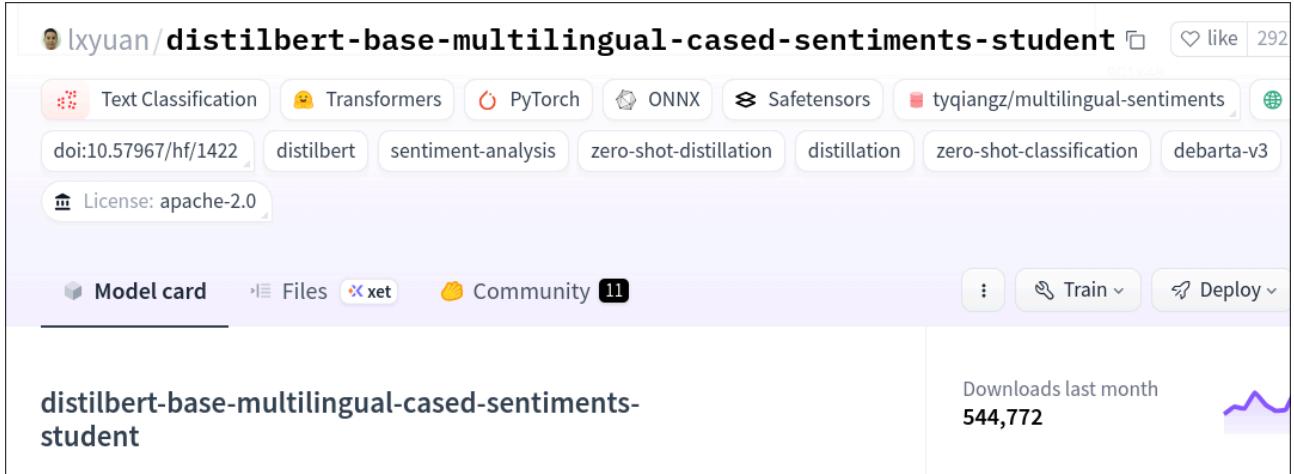
- Các hình bổ sung: word cloud, scatter plot score-theo-độ-dài, bảng xếp hạng tác giả và tập CSV cho phân tích sâu hơn (eda/*.csv).

6. Phân tích sentiment — Kế hoạch ban đầu & quyết định

- Quy trình dự kiến: fine-tune lxyuan/distilbert-base-multilingual-cased-sentiments-student trên bộ nhãn phân tầng, hỗ trợ bởi Gemini (mục tiêu $\geq 5\ 000$ annotate cho hiệu chỉnh và đánh giá).
- Trở ngại: chi phí thuê GPU, tốc độ gán nhãn bị giới hạn và lợi ích biên

thấp so với việc tận dụng topic insight không giám sát.

- Cách triển khai: VADER (rule-based) tính polarity câu trong giai đoạn tổng hợp khía cạnh, giữ lại kế hoạch nâng cấp cho tương lai.



Hình 18: Mục tiêu fine-tune lxyuan/distilbert-base-multilingual bị tạm hoãn do hạn chế GPU và nhãn.

7. Topic modeling & pipeline khía cạnh

CLI vận hành vòng lặp xác định — làm sạch → TF-IDF → SVD → KMeans → sentiment theo khía cạnh — cân bằng tự động hóa cho lặp nhanh với các flag rõ ràng cho người dùng muốn tùy chỉnh.

7.1. Tiền xử lý

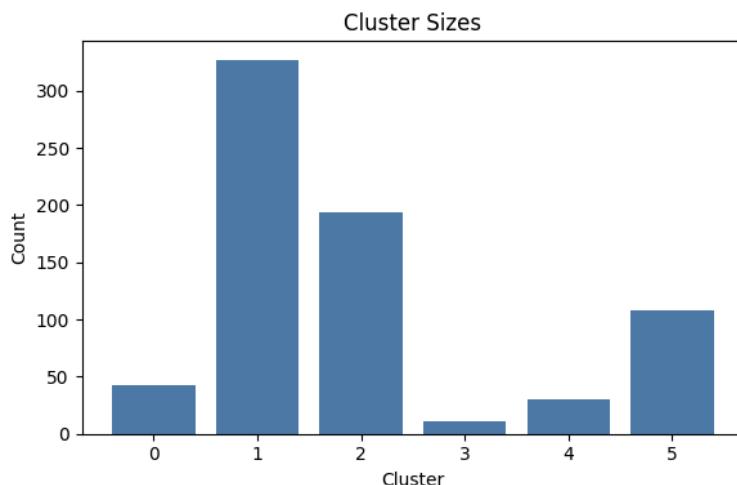
- Chuẩn hóa Unicode (NFC), loại URL/khoi code, chuyển chữ thường, loại ký tự không phải chữ cái nhưng giữ dấu nháy.
- Lọc token giữ từ dài hơn hai ký tự và loại stopword mở rộng (bao gồm alias thương hiệu).
- Bộ lọc alias/subreddit cho phép tạo corpus tập trung theo từng sản phẩm.

7.2. Chọn đặc trưng & giảm chiều

- TF-IDF (`ngram_range = 1-2`) với `min_df` thích ứng (3 nếu $N \geq 300$, ngược lại 2) và `max_df = 0.95`; vocabulary giới hạn ở `min(2000, 3N)` khi chạy auto (`run_pipeline.py:545`).
- TruncatedSVD giới hạn thành phần ở `min(round(0.25N), 200)` với đáy 50, ghi lại explained variance để minh bạch (`run_pipeline.py:560`).
- Bộ bảo vệ giới hạn SVD ở `min(V - 1, N - 1)` và mở override qua `--min-df`, `--max-feat`, `--svd`, giúp corpus nhỏ vẫn an toàn mà không mất kiểm soát chuyên gia.
- Cơ chế bảo vệ vocabulary rỗng tự động thử lại TF-IDF với `min_df = 1`, tránh việc chạy alias góc cạnh bị thất bại.

7.3. Quy trình phân cụm

- KMeans quét `k_min..k_max` (mặc định 3–8) dựa vào silhouette; `method.json` ghi lại phạm vi tìm kiếm, seed và score tốt nhất để phân tích biết vì sao chọn `k` đó.
- Chẩn đoán cụm kết hợp trung bình TF-IDF (`top_terms_by_cluster`) với bài đại diện gần centroid (`representatives_by_cluster`) để tạo chủ đề dễ diễn giải và trích dẫn.
- Tùy chọn `--save-plots` sinh bar chart kích thước cụm nhằm QA nhanh trước khi chia sẻ.



Hình 19: Phân bố kích thước cụm (ví dụ Fiio FT1) sau khi giảm chiều bằng SVD và phân cụm KMeans.

7.4. Trích khía cạnh & kết hợp sentiment

- Thành viên subreddit tự động chọn bộ seed khía cạnh (pin, nhiệt, độ thoái mái ...); tùy chọn TF-IDF expansion bổ sung thuật ngữ đặc thù corpus.
- Ghép mẫu ở cấp câu; polarity VADER tạo bảng mentions, sentiment trung bình và danh sách trích dẫn đã lọc.
- `assignments.jsonl` gắn nhãn cụm cho từng tài liệu cùng profile sentiment khía cạnh, trong khi `aspect_summary*.json` loại khía cạnh nhiễu dưới ngưỡng tần suất thích ứng (`max(5, round(0.01N))`).
- Phân tích viên có thể mở rộng coverage với `--expand-seeds`, thêm top term TF-IDF (trừ alias thương hiệu) vào từ điển khía cạnh cho cách diễn đạt tự nhiên.
- Bộ đếm trích dẫn chỉ giữ câu có compound polarity vượt ± 0.05 và giới hạn tối đa năm ví dụ mỗi khía cạnh, cân bằng tín hiệu và ngắn gọn.

7.5. Artefact trọng tâm

- `tfidf_vectorizer.joblib`, `svd_model.joblib`,
- `svd_explained_variance.json` — pipeline đặc trưng có thể tái sử dụng.
- `kmeans_clusters.json` — kích thước cụm, top TF-IDF và bài đại diện gần centroid gói trong một tệp; `method.json` ghi lại siêu tham số đã chọn.
- `assignments.jsonl`, `aspect_summary.json`,

`aspect_summary_by_subreddit.json` — cụm cấp tài liệu kèm sentiment khía cạnh, lọc bằng ngữ cảnh tần suất thích ứng.

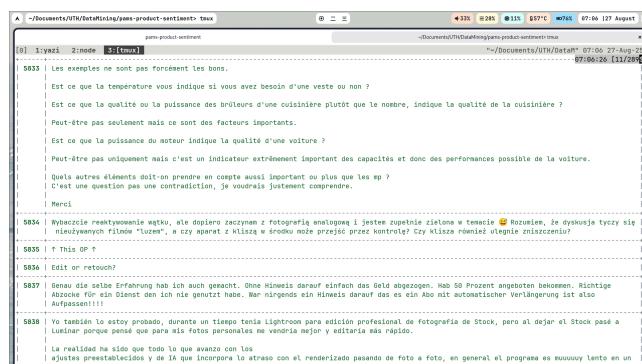
- PNG tùy chọn: `cluster_sizes.png`, `aspects_pos_neg.png` để kiểm tra nhanh.

8. Deliverables & hướng dẫn sử dụng

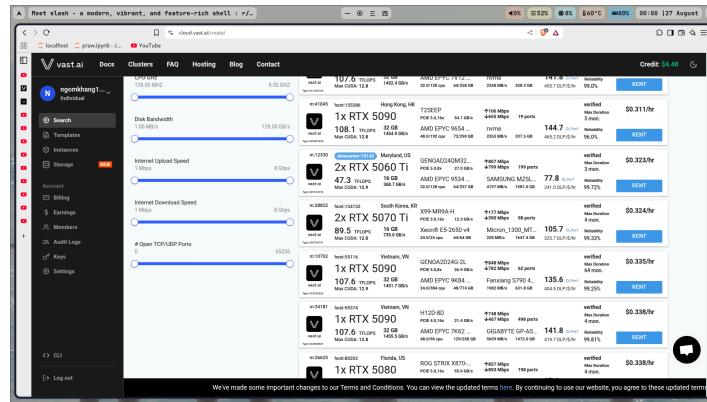
- Điểm vào CLI: `run_pipeline.py --data <parquet> --product <name> --aliases <comma-separated> --out <dir> [options]`.
- Tham số cấu hình điều khiển vector hóa (`--min-df`, `--max-df`, `--max-feat`, `--ngram-*`), giảm chiều (`--svd`), phân cụm (`--method`, `--k-min`, `--k-max`) và hành vi khía cạnh (`--aspect-category`, `--expand-seeds`, `--min-aspect-freq`).
- artifact cho FiiO FT1, Sennheiser HD600 và Sony WF-1000XM4 được lưu tại `extra/artifacts_ft1`, `extra/artifacts_hd600` và `extra/artifacts_m4`.

9. Hạn chế & hướng phát triển

- Sentiment vẫn dựa rule-based; thực thi kế hoạch fine-tune sẽ xử lý châm biếm và thuật ngữ đặc thù tốt hơn.
- Nhánh NMF/LDA mới dựng khung, chưa sản phẩm hóa — bổ sung sẽ cho phép chủ đề chồng lấn.
- Mở rộng tích hợp Tiki và hỗ trợ đa ngôn ngữ tốt hơn sẽ phản ánh sentiment phi tiếng Anh chính xác hơn.
- Cần bổ sung đánh giá (topic coherence, độ ổn định khi lấy mẫu lại, kiểm chứng có con người) cho vòng lặp kế tiếp.



Hình 20: Ví dụ Reddit đa ngôn ngữ cho thấy nhu cầu xử lý tiếng Việt và châu Âu ở bản phát hành tiếp theo.



Hình 21: Ảnh chụp marketplace GPU Vast.ai — dự toán chi phí fine-tune khiến nhóm hoãn huấn luyện transformer.

10. Phụ lục

10.1. Tài sản cốt lõi

- `plan.md` — kiến trúc chi tiết, phương án thay thế, heuristic đánh giá.
- `eda/` — biểu đồ khám phá và tổng hợp CSV.
- `run_pipeline.py` (thư mục gốc & `extra/`) — CLI chủ đề + sentiment khía cạnh tùy chỉnh.
- `extra/bai_thuyet_trinh.typ` — slide Typst trình bày trên lớp.

10.2. Ví dụ tóm tắt console (rút gọn)

$N=742$, $V=1625$, $\text{min_df}=3$, $\text{max_df}=0.95$, $\text{max_features}=2000$

Số thành phần SVD=150, cumulative explained variance=0.72

Chọn K=5 với silhouette=0.41

Aspect battery POS: “Battery life has been excellent...” (+0.68)

Aspect battery NEG: “Battery drains fast when streaming...” (-0.52)

KẾT THÚC BÁO CÁO

Trong khuôn khổ môn Kỹ thuật Khai thác Dữ liệu, nhóm đã xây dựng một pipeline hiện đại để gom, làm sạch và phân tích thảo luận cộng đồng xoay quanh sản phẩm công nghệ. Chuỗi công việc này thể hiện khả năng kết nối nhiều nguồn dữ liệu, triển khai quy trình máy học có thể tái lập và rút ra insight thực tiễn cho doanh nghiệp. Chúng em tin rằng những cải tiến tương lai về sentiment và đa ngôn ngữ sẽ giúp pipeline trở thành nền tảng phân tích sản phẩm toàn diện hơn. Xin chân thành cảm ơn thầy đã hỗ trợ, góp ý suốt môn học vừa qua.