

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: T.S 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>

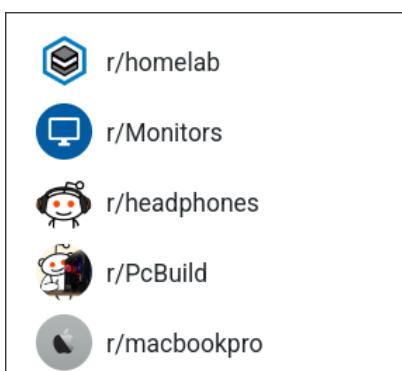
## 2. Tóm tắt điều hành

- 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.



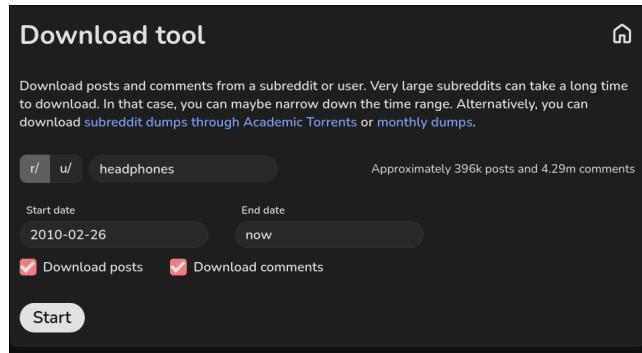
Hình 1: Danh sách subreddit tiêu biểu bao phủ homelab, màn hình, cộng đồng audiophile và hệ sinh thái Apple.

product_id	customer_name	rating	title	content	thank_count	create_at
278	26270504	4	Hai khang	Mua 2 chiếc và hỏi họ có 3 chiếc (mặc dù có 3 chiếc) mà trả cho ta là 2 chiếc. Dùng thử thấy không	0	202405040849
179	26270504	5	Cực khang	Cực khang	0	202405040850
180	26270504	5	Ly khang	Ly khang	0	202405040851
181	26270504	5	Avin Lee	Cực khang	0	202405040852
182	26270504	5	Nắng ragazzi	Hàng khang	0	202405040853
183	26270504	5	Voku hoan nam	Cực khang	0	202405040854
184	26270504	5	Nguyễn Văn Hùng	Bình thường	0	202405040855
185	26270504	5	Nguyễn Văn Hùng	Hàng bình thường	0	202405040856
186	26270504	5	Binh khang	Hàng bình thường	0	202405040857
187	26270504	5	Phạm Hân Hảo	Cực khang	0	202405040858
188	26270504	5	Mae Xuan Khanh	Cực khang	0	202405040859
189	26270504	2	Đỗ Hùng	Nhà này rất đỗi là nhà HKT	0	202405040860
190	26270504	5	Phan Phuoc	Cực khang	0	202405040861
191	26270504	5	Nguyễn Văn Hùng	Gió hàng khang như là đòn bẩy	0	202405040862
192	26270504	5	nguyen da sang	Cực khang	0	202405040863
193	26270504	5	nhang rejaule	Cực khang	0	202405040864
194	26270504	5	Ly khang	Ly khang	0	202405040865
195	26270504	5	Avin Lee	Cực khang	0	202405040866
196	26270504	5	Hoang	Hàng khang	0	202405040867
197	26270504	5	Trang hang	Hàng khang	0	202405040868
198	26270504	5	Thien	Hàng khang	0	202405040869
199	26270504	5	Hoang	Hàng khang	0	202405040870
200	26270504	5	Trang hang	Hàng khang	0	202405040871
201	26270504	5	Thien	Hàng khang	0	202405040872
202	26270504	5	Hoang	Hàng khang	0	202405040873
203	26270504	5	Thien	Hàng khang	0	202405040874
204	26270504	5	Trang hang	Hàng khang	0	202405040875
205	26270504	5	Thien	Hàng khang	0	202405040876
206	26270504	5	Hoang	Hàng khang	0	202405040877
207	26270504	5	Thien	Hàng khang	0	202405040878
208	26270504	5	Trang hang	Hàng khang	0	202405040879
209	26270504	5	Thien	Hàng khang	0	202405040880
210	26270504	5	Hoang	Hàng khang	0	202405040881
211	26270504	5	Thien	Hàng khang	0	202405040882
212	26270504	5	Trang hang	Hàng khang	0	202405040883
213	26270504	5	Thien	Hàng khang	0	202405040884
214	26270504	5	Hoang	Hàng khang	0	202405040885
215	26270504	5	Thien	Hàng khang	0	202405040886
216	26270504	5	Trang hang	Hàng khang	0	202405040887
217	26270504	5	Thien	Hàng khang	0	202405040888
218	26270504	5	Hoang	Hàng khang	0	202405040889
219	26270504	5	Thien	Hàng khang	0	202405040890
220	26270504	5	Trang hang	Hàng khang	0	202405040891
221	26270504	5	Thien	Hàng khang	0	202405040892
222	26270504	5	Hoang	Hàng khang	0	202405040893
223	26270504	5	Thien	Hàng khang	0	202405040894
224	26270504	5	Trang hang	Hàng khang	0	202405040895
225	26270504	5	Thien	Hàng khang	0	202405040896
226	26270504	5	Hoang	Hàng khang	0	202405040897
227	26270504	5	Thien	Hàng khang	0	202405040898
228	26270504	5	Trang hang	Hàng khang	0	202405040899
229	26270504	5	Thien	Hàng khang	0	202405040900
230	26270504	5	Hoang	Hàng khang	0	202405040901
231	26270504	5	Thien	Hàng khang	0	202405040902
232	26270504	5	Trang hang	Hàng khang	0	202405040903
233	26270504	5	Thien	Hàng khang	0	202405040904
234	26270504	5	Hoang	Hàng khang	0	202405040905
235	26270504	5	Thien	Hàng khang	0	202405040906
236	26270504	5	Trang hang	Hàng khang	0	202405040907
237	26270504	5	Thien	Hàng khang	0	202405040908
238	26270504	5	Hoang	Hàng khang	0	202405040909
239	26270504	5	Thien	Hàng khang	0	202405040910
240	26270504	5	Trang hang	Hàng khang	0	202405040911
241	26270504	5	Thien	Hàng khang	0	202405040912
242	26270504	5	Hoang	Hàng khang	0	202405040913
243	26270504	5	Thien	Hàng khang	0	202405040914
244	26270504	5	Trang hang	Hàng khang	0	202405040915
245	26270504	5	Thien	Hàng khang	0	202405040916
246	26270504	5	Hoang	Hàng khang	0	202405040917
247	26270504	5	Thien	Hàng khang	0	202405040918
248	26270504	5	Trang hang	Hàng khang	0	202405040919
249	26270504	5	Thien	Hàng khang	0	202405040920
250	26270504	5	Hoang	Hàng khang	0	202405040921
251	26270504	5	Thien	Hàng khang	0	202405040922
252	26270504	5	Trang hang	Hàng khang	0	202405040923
253	26270504	5	Thien	Hàng khang	0	202405040924
254	26270504	5	Hoang	Hàng khang	0	202405040925
255	26270504	5	Thien	Hàng khang	0	202405040926
256	26270504	5	Trang hang	Hàng khang	0	202405040927
257	26270504	5	Thien	Hàng khang	0	202405040928
258	26270504	5	Hoang	Hàng khang	0	202405040929
259	26270504	5	Thien	Hàng khang	0	202405040930
260	26270504	5	Trang hang	Hàng khang	0	202405040931
261	26270504	5	Thien	Hàng khang	0	202405040932
262	26270504	5	Hoang	Hàng khang	0	202405040933
263	26270504	5	Thien	Hàng khang	0	202405040934
264	26270504	5	Trang hang	Hàng khang	0	202405040935
265	26270504	5	Thien	Hàng khang	0	202405040936
266	26270504	5	Hoang	Hàng khang	0	202405040937
267	26270504	5	Thien	Hàng khang	0	202405040938
268	26270504	5	Trang hang	Hàng khang	0	202405040939
269	26270504	5	Thien	Hàng khang	0	202405040940
270	26270504	5	Hoang	Hàng khang	0	202405040941
271	26270504	5	Thien	Hàng khang	0	202405040942
272	26270504	5	Trang hang	Hàng khang	0	202405040943
273	26270504	5	Thien	Hàng khang	0	202405040944
274	26270504	5	Hoang	Hàng khang	0	202405040945
275	26270504	5	Thien	Hàng khang	0	202405040946
276	26270504	5	Trang hang	Hàng khang	0	202405040947
277	26270504	5	Thien	Hàng khang	0	202405040948
278	26270504	5	Hoang	Hàng khang	0	202405040949
279	26270504	5	Thien	Hàng khang	0	202405040950
280	26270504	5	Trang hang	Hàng khang	0	202405040951
281	26270504	5	Thien	Hàng khang	0	202405040952
282	26270504	5	Hoang	Hàng khang	0	202405040953
283	26270504	5	Thien	Hàng khang	0	202405040954
284	26270504	5	Trang hang	Hàng khang	0	202405040955
285	26270504	5	Thien	Hàng khang	0	202405040956
286	26270504	5	Hoang	Hàng khang	0	202405040957
287	26270504	5	Thien	Hàng khang	0	202405040958
288	26270504	5	Trang hang	Hàng khang	0	202405040959
289	26270504	5	Thien	Hàng khang	0	202405040960
290	26270504	5	Hoang	Hàng khang	0	202405040961
291	26270504	5	Thien	Hàng khang	0	202405040962
292	26270504	5	Trang hang	Hàng khang	0	202405040963
293	26270504	5	Thien	Hàng khang	0	202405040964
294	26270504	5	Hoang	Hàng khang	0	202405040965
295	26270504	5	Thien	Hàng khang	0	202405040966
296	26270504	5	Trang hang	Hàng khang	0	202405040967
297	26270504	5	Thien	Hàng khang	0	202405040968
298	26270504	5	Hoang	Hàng khang	0	202405040969
299	26270504	5	Thien	Hàng khang	0	202405040970
300	26270504	5	Trang hang	Hàng khang	0	202405040971
301	26270504	5	Thien	Hàng khang	0	202405040972
302	26270504	5	Hoang	Hàng khang	0	202405040973
303	26270504	5	Thien	Hàng khang	0	202405040974
304	26270504	5	Trang hang	Hàng khang	0	202405040975
305	26270504	5	Thien	Hàng khang	0	202405040976
306	26270504	5	Hoang	Hàng khang	0	202405040977
307	26270504	5	Thien	Hàng khang	0	202405040978
308	26270504	5	Trang hang	Hàng khang	0	202405040979
309	26270504	5	Thien	Hàng khang	0	202405040980
310	26270504	5	Hoang	Hàng khang	0	202405040981
311	26270504	5	Thien	Hàng khang	0	202405040982
312	26270504	5	Trang hang	Hàng khang	0	202405040983
313	26270504	5	Thien	Hàng khang	0	202405040984
314	26270504	5	Hoang	Hàng khang	0	202405040985
315	26270504	5	Thien	Hàng khang	0	202405040986
316	26270504	5	Trang hang	Hàng khang	0	202405040987
317	26270504	5	Thien	Hàng khang	0	202405040988
318	26270504	5	Hoang	Hàng khang	0	202405040989
319	26270504	5	Thien	Hàng khang	0	202405040990
320	26270504	5	Trang hang	Hàng khang	0	202405040991
321	26270504	5	Thien	Hàng khang	0	202405040992
322	26270504	5	Hoang	Hàng khang	0	202405040993
323	26270504	5	Thien	Hàng khang	0	202405040994
324	26270504	5	Trang hang	Hàng khang	0	202405040995
325	26270504	5	Thien	Hàng khang	0	202405040996
326	26270504	5	Hoang	Hàng khang	0	202405040997
327	26270504	5	Thien	Hàng khang	0	202405040998
328	26270504	5	Trang hang	Hàng khang	0	202405040999
329	26270504	5	Thien	Hàng khang	0	202405041000
330	26270504	5	Hoang	Hàng khang	0	202405041001
331	26270504	5	Thien	Hàng khang	0	202405041002
332	26270504	5	Trang hang	Hàng khang	0	202405041003
333	26270504	5	Thien	Hàng khang	0	202405041004
334	26270504	5	Hoang	Hàng khang	0	202405041005
335	26270504	5	Thien	Hàng khang	0	202405041006
336	26270504	5	Trang hang	Hàng khang	0	202405041007
337	26270504	5	Thien	Hàng khang	0	202405041008
338	26270504	5	Hoang	Hàng khang	0	202405041009
339	26270504	5	Thien	Hàng khang	0	202405041010
340	26270504	5	Trang hang	Hàng khang	0	202405041011
341	26270504	5	Thien	Hàng khang	0	202405041012
342	26270504	5	Hoang	Hàng khang	0	202405041013
343	26270504	5	Thien	Hàng khang	0	202405041014
344	26270504	5	Trang hang	Hàng khang	0	202405041015
345	26270504	5	Thien	Hàng khang	0	202405041016
346	26270504	5	Hoang	Hàng khang	0	202405041017
347	26270504	5	Thien	Hàng khang	0	202405041018
348	26270504	5	Trang hang	Hàng khang	0	202405041019
349	26270504	5	Thien	Hàng khang	0	202405041020
350	26270504	5	Hoang	Hàng khang	0	202405041021
351						

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.

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

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

	name	subreddit	score	link_id	parent_id	body	is_post
789	t3_1la792j	iphone	6			ng to resell this phone and get the new iPhone coming in later this year. Just need opinions	true
790	t3_1m3ywu5	iphone	1			ssd cards for recording longer videos, but haven't seen anything about the 16 base model.	true
791	t3_1l5xtr1	laptops	112			I think I messed up while replacing the screen	true
792	t3_1lui9fg	laptops	110			air costs in Europe are so high that it's not even worth fixing. Lenovo, this is unacceptable!	true
793	t3_1ls75je	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_1mfbuq	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 Macro 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_mve6pk1	AppleWatch	35	t3_1l0vtn	t3_1l0vtn	Well the picture is actually mine but the idea 😊	false
799	t1_mve6sws	AppleWatch	-7	t3_1l0vtn	t3_1l0vtn	you should wrap the watch around the phone horizontally but I like the idea!	false
800	t1_mve70qv	AppleWatch	536	t3_1l0vtn	t3_1l0vtn	ap a picture from the watch which is handy for capturing a serial number or model number.	false
801	t1_mve7088	AppleWatch	176	t3_1l0vtn	t3_1l0vtn	In my days selfie cameras came with a little mirror to be able to see yourself. 😊	false
802	t1_mvea93t	AppleWatch	1	t3_1l0vtn	t3_1l0vtn	but why?	false
803	t1_mveaf1	AppleWatch	9	t3_1l0vtn	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_1l0vtn	t1_mve70qw	Pretty much the only thing I use it for as well.	false
805	t1_mvebg7w	AppleWatch	103	t3_1l0vtn	t3_1l0vtn	Well if you can get it to work reliably, it's always been trouble for me to see once launched.	false
806	t1_mveeq8q	AppleWatch	20	t3_1l0vtn	t3_1l0vtn	I don't see the viewfinder image on my watch, just the controls. Anyone know why?	false
807	t1_mveefra	AppleWatch	2	t3_1l0vtn	t3_1l0vtn	What's the app called?	false
808	t1_mverpp	AppleWatch	60	t3_1l0vtn	t3_1l0vtn	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_1l0vtn	t3_1l0vtn	Or just use the other camera?	false

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”, “fio 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 lưới, 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.

OkNefariousness284 · 28d ago  
My main issue with this chapter is that it just feels like an asspull mid fight to have You win lol. Like truly, what convenient timing for the world to relearn about nukes  
14 Reply Award Share ...

thachduaboi3000 OP · 28d ago  
CSM GIRLS ULTRAFUNDAMENTALIST · Top 1% Poster  
Never underestimate the power of the American military complex  
40 Reply Award Share ...

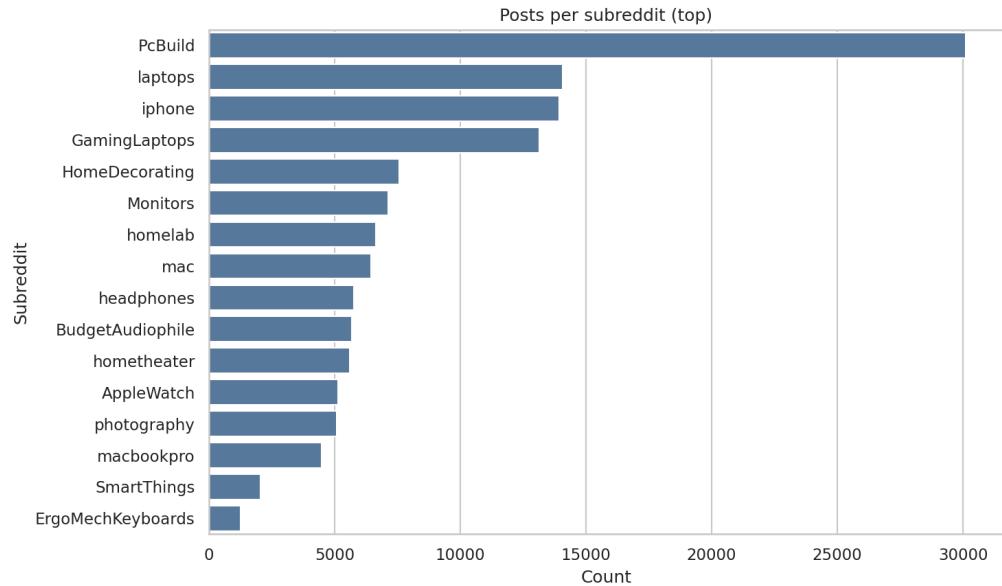
Illustrious-Sky-4631 · 28d ago  
Top 1% Commenter  
It's insane how this fuckers chose to Genocide their fella humans in the middle of a Devils apocalypse  
Just so the War Devil can get a power up at a very perfect timing  
9 Reply Award Share ...

+ 4 more replies  
+ 13 more replies

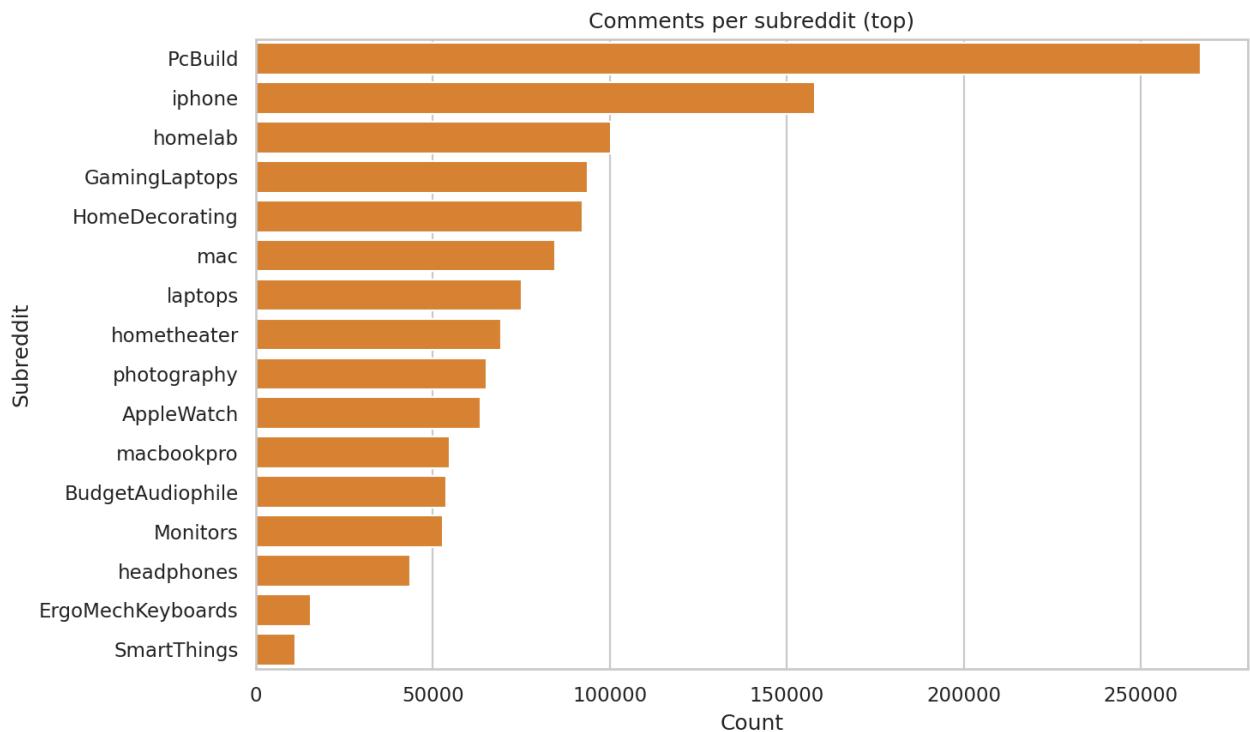
[View more comments](#)

## 5. Phân tích dữ liệu thăm dò (EDA)

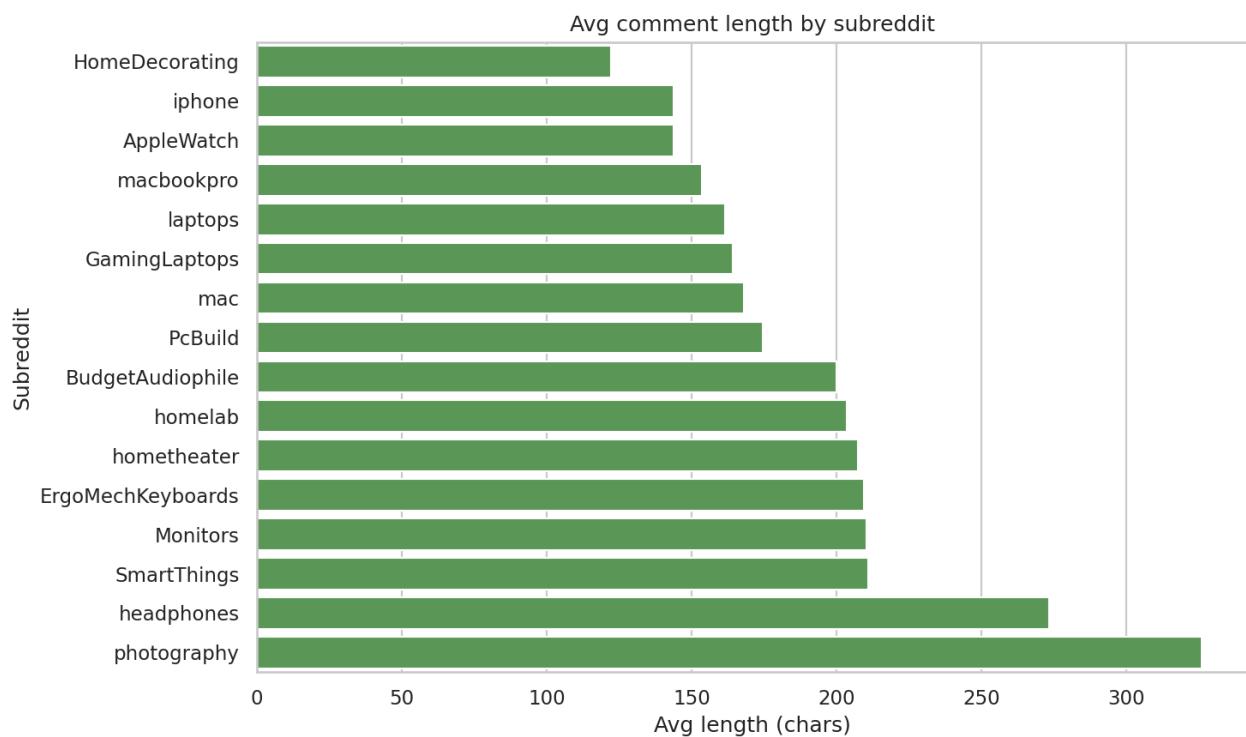
- Tài sản EDA nằm trong eda/ (PNG + CSV) để tái sử dụng nhanh cho slide và dashboard.



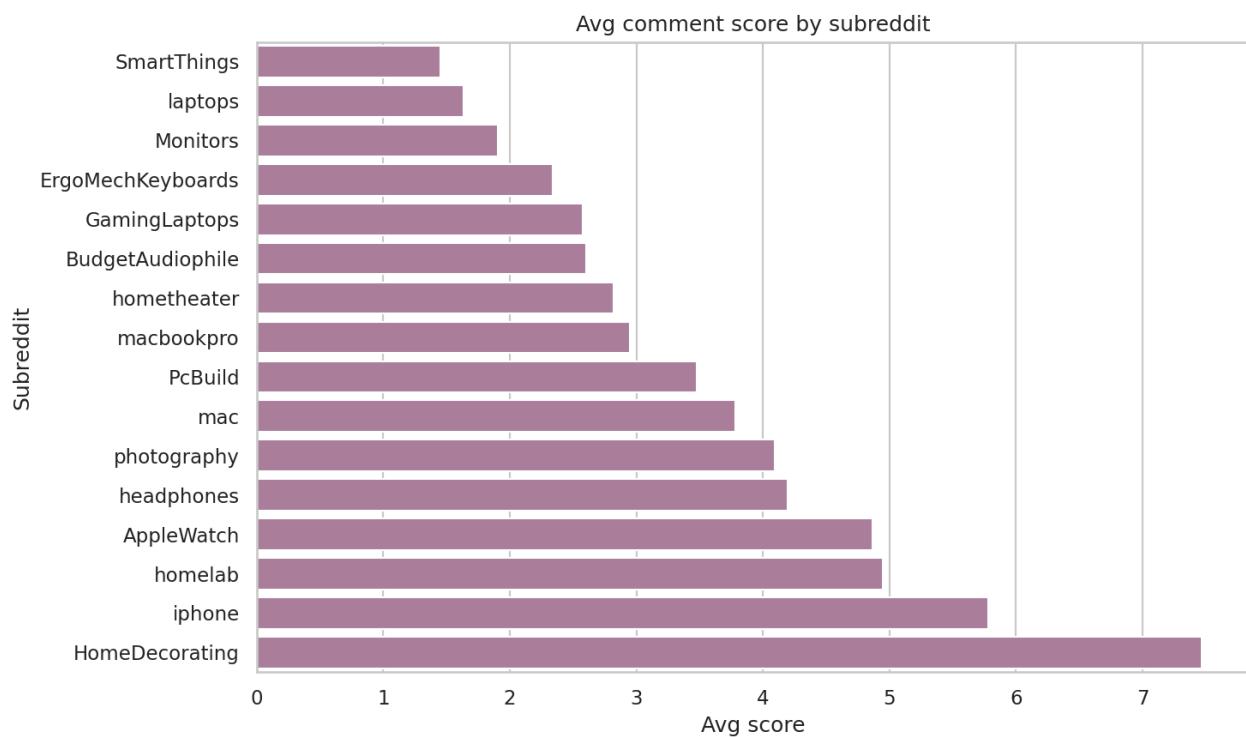
Hình 6: Số bài theo subreddit — PcBuild, iPhone và GamingLaptops chiếm sản lượng lớn.



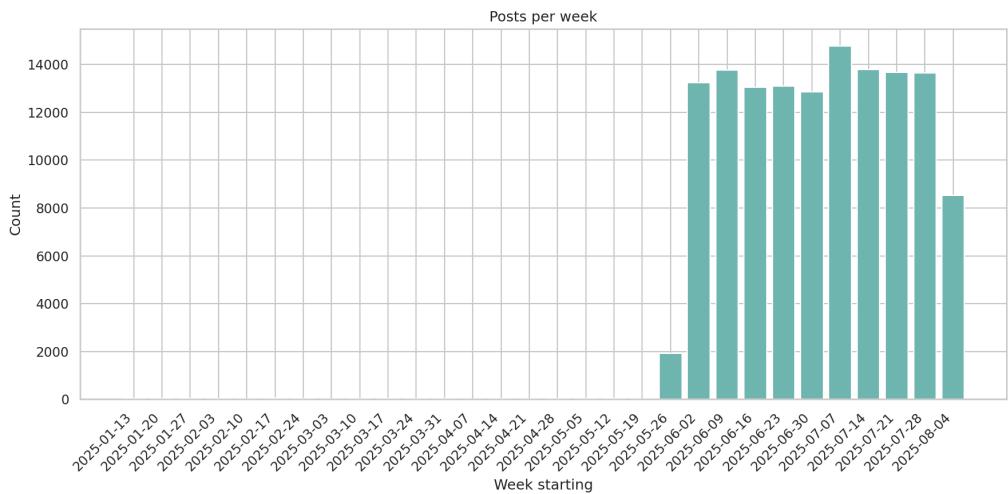
Hình 7: Số bình luận: cộng đồng xử lý sự cố (PcBuild, homelab) dẫn đầu về tương tác.



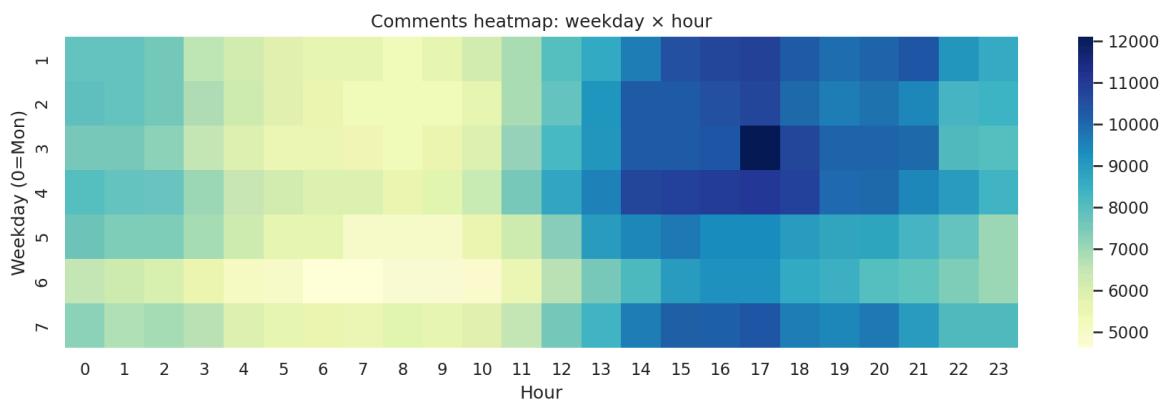
Hình 8: Độ dài bình luận trung bình cho thấy thảo luận kỹ thuật sâu ở r/macbookpro và r/AppleWatch.



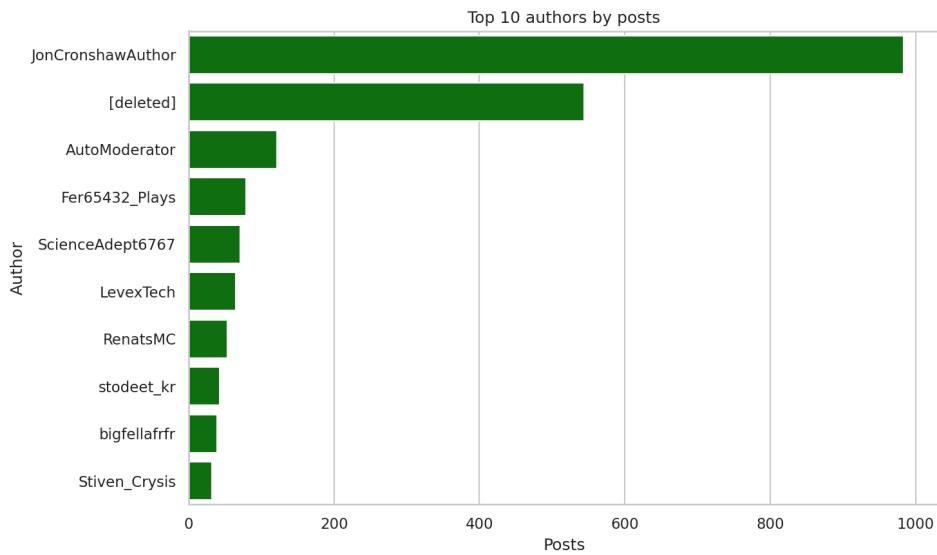
Hình 9: Điểm bình luận trung bình khá thấp (<2.5), nhấn mạnh nhu cầu insight văn bản ngoài karma.



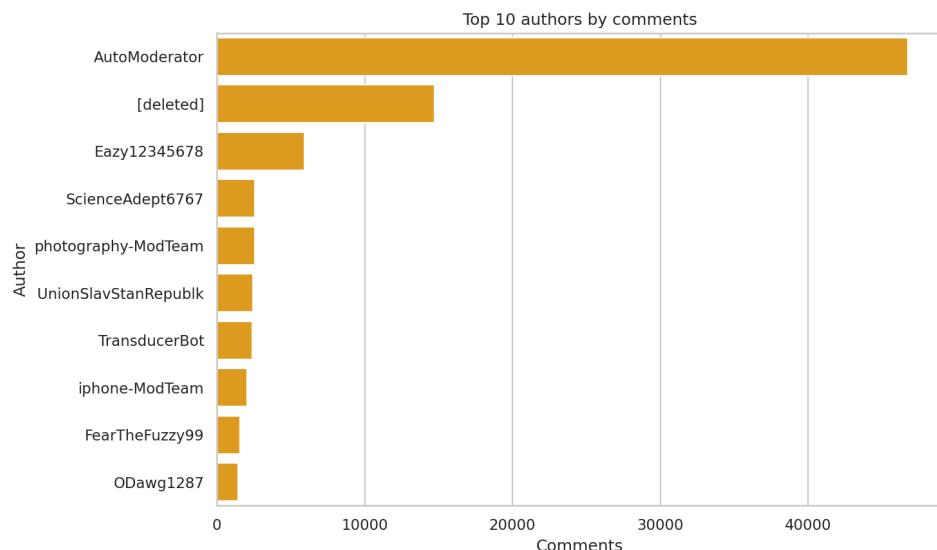
Hình 10: Sản lượng bài đăng theo tuần phát hiện các đợt ra mắt phần cứng theo mùa.



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



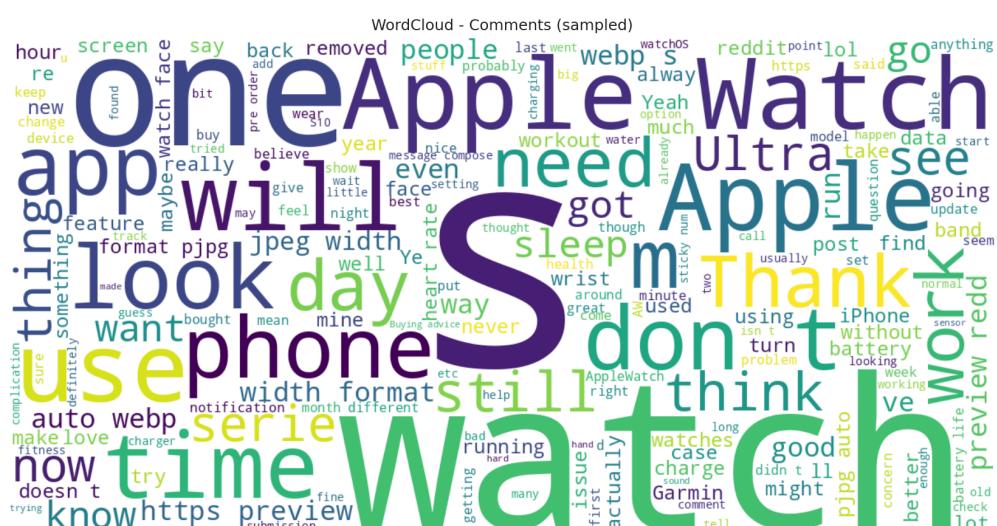
Hình 12: Tác giả bài đăng hàng đầu — chỉ ra phần lớn những account đăng nhiều nhất là bot và những account bị xóa sẽ quy chung về [deleted].



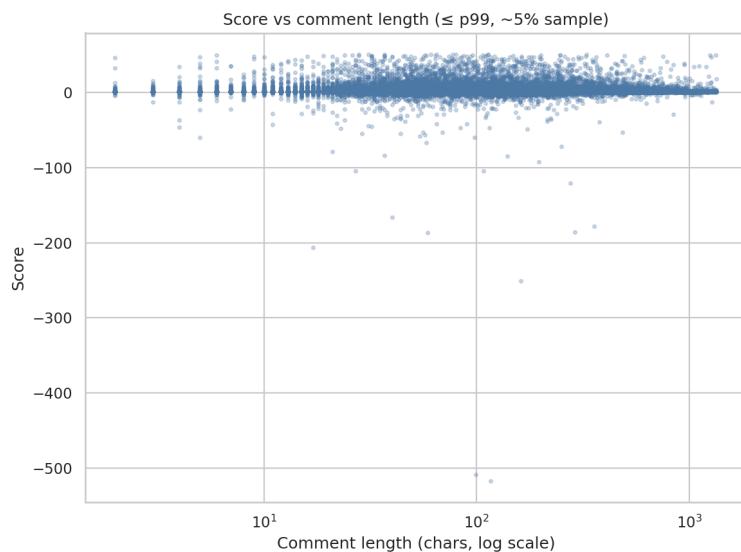
Hình 13: Tác giả bình luận hàng đầu — nhận diện người thiên về hỗ trợ và điều phổi.



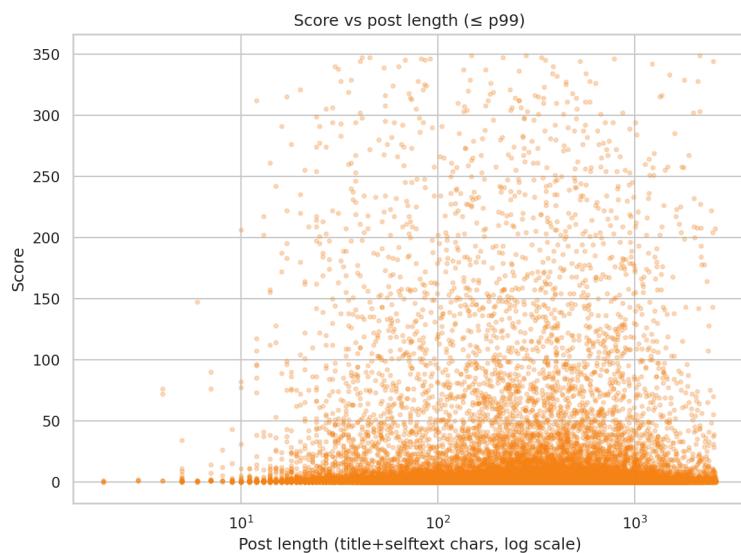
Hình 14: Word cloud tiêu đề nêu bật nhóm sản phẩm chủ đạo và chủ đề xử lý sự cố.



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



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.

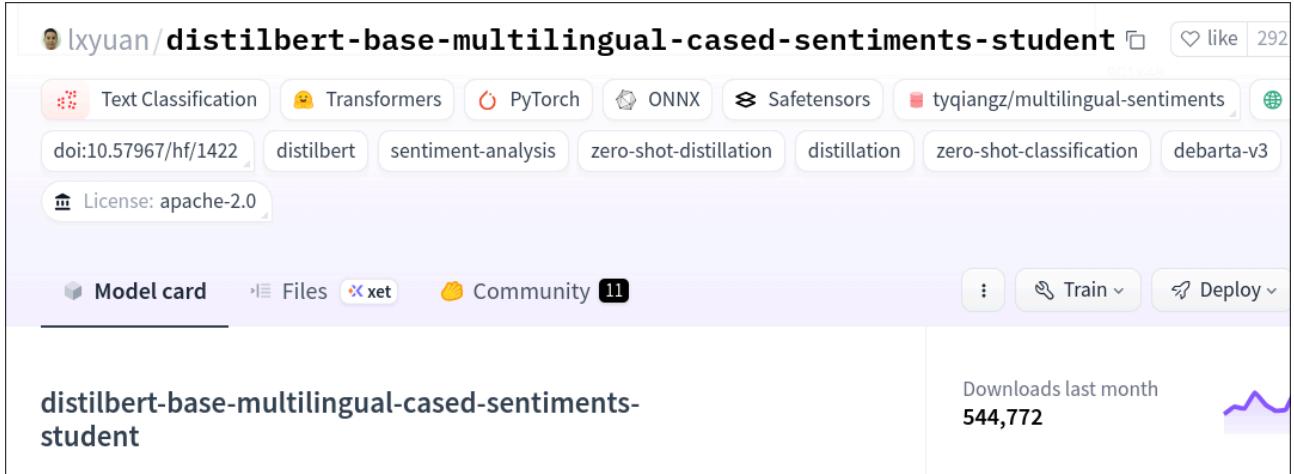
- Các hình bổ sung: word cloud, scatter plot score-theo-độ-dài, bảng xếp hạng tác giả và tập CSV cho phân tích sâu hơn (eda/\*.csv).

## 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

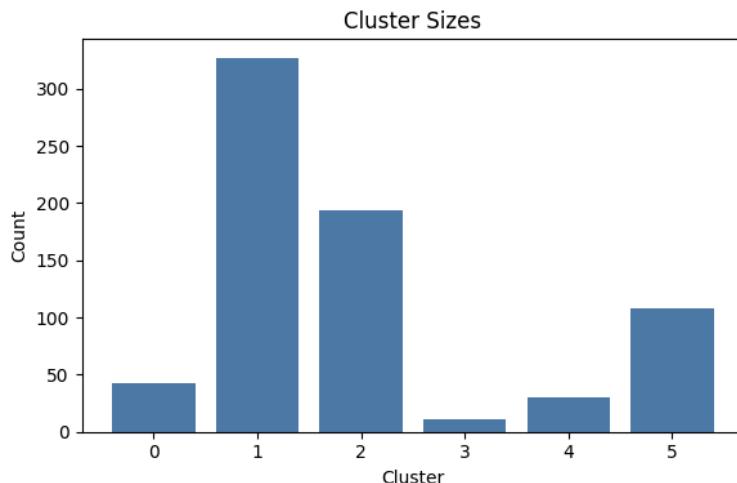
- Chuẩn hóa Unicode (NFC), loại URL/khoi code, chuyển chữ thường, loại ký tự không phải chữ cái nhưng giữ dấu nháy.
- Lọc token giữ từ dài hơn hai ký tự và loại stopword mở rộng (bao gồm alias thương hiệu).
- Bộ lọc alias/subreddit cho phép tạo corpus tập trung theo từng sản phẩm.

### 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 thoả 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  $\pm 0.05$  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

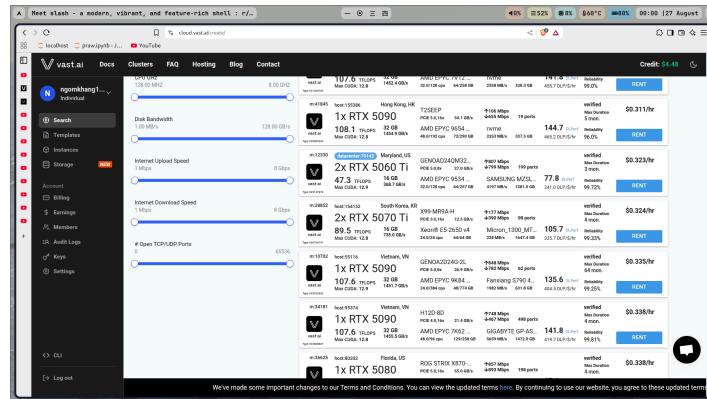
- Điểm vào CLI: `run_pipeline.py --data <parquet> --product <name> --aliases <comma-separated> --out <dir> [options]`.
- Tham số cấu hình điều khiển vector hóa (`--min-df`, `--max-df`, `--max-feat`, `--ngram-*`), giảm chiều (`--svd`), phân cụm (`--method`, `--k-min`, `--k-max`) và hành vi khía cạnh (`--aspect-category`, `--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.