Thiết Kế và Thu Thập Dataset Đa Lĩnh Vực cho Phân Tích Xu Hướng Giải Trí

MSSV:QE200009- DƯƠNG THỊ MỸ TÂM

MSSV:QE200083- TRẦN VÂN KHÁNH

Tóm Tắt

Báo cáo này trình bày chiến lược thu thập dataset >10.000 bản ghi từ đa lĩnh vực giải trí, tập trung vào phân tích xu hướng không lệch domain. Phương pháp sử dụng ít nền tảng nhưng đa category, với trọng tâm vào YouTube và Spotify.

1. Giới Thiệu

Trong nghiên cứu xu hướng giải trí, dataset đa lĩnh vực (>10.000 bản ghi) là cần thiết để tránh bias. Phương pháp tập trung vào API công khai từ nền tảng như YouTube, Spotify, và TikTok.

2. Mục Tiêu Nghiên Cứu

Thu thập dataset >10.000 bản ghi đa lĩnh vực.

Đảm bảo cân bằng category.

Phân tích xu hướng dựa trên metadata.

3. Phương Pháp Thu Thập Dữ Liệu

3.1. Nguyên Tắc Thiết Kế Dataset

Quy mô: >10.000 bản ghi

Công cụ: Python với API,SQL

Xử lý: Làm sạch và chuẩn hóa metadata.

3.2. Chiến Lược Thu Thập Theo Nền Tảng

Dataset chia thành ba tầng: Core (70%), Phản ánh hành vi (15%), và Mở rộng (15%).

3.2.1. Nền Tảng Core (70% Dữ Liệu)

YouTube Data API:

Lý do: Cung cấp trending video đa category (Music, Comedy, Entertainment, People & Blogs, Gaming, Film & Animation).

Quy mô: ~200 video/ngày/quốc gia; 10 quốc gia × 30 ngày → ~60.000 video (chọn 30.000–50.000).

Cách thu thập: API lấy metadata (video_id, category, views, likes, comments, publish_time).

Spotify Web API:

Lý do: Chuyên âm nhạc (pop, rap, EDM, indie; podcast).

Quy mô: Top 50 + Viral 50/ngày/quốc gia; 20 quốc gia × 30 ngày → ~60.000 track (chọn 30.000–50.000).

Cách thu thập: API lấy metadata (track_id, artist, popularity, tempo, energy, date, country).

3.2.2. Nền Tảng Phản Ánh Hành Vi (15% Dữ Liệu)

TikTok (Hashtag + Sound):

Lý do: Phát sinh trend hài hước/viral (hài hước, challenge, meme, sound).

Quy mô: 200 hashtag/sound, ~50–100 video/hashtag → 10.000+ bản ghi.

Cách thu thập: Crawl metadata (hashtag, content_type, views, growth_rate, date).

Google Trends:

Lý do: Xác nhận xu hướng xã hội (tên bài hát, creator, meme).

Quy mô: Feature phụ.

Cách thu thập: API lấy volume tìm kiếm.

3.2.3. Nền Tảng Mở Rộng (15% Dữ Liệu, Tùy Chọn)

Reddit:

Lý do: Community humor/trends (r/funny, r/videos, r/memes, r/music).

Quy mô: 5 subreddit × 30 post/ngày × 30 ngày → ~4.500 post.

Cách thu thập: API lấy metadata.

X (Twitter):

Lý do: Reaction/meme lifecycle (trending hashtag, tweet volume).

Quy mô: Tích hợp nếu cần.

3.3. Cấu Hình Dataset

Lĩnh Vực	Nền Tảng	Số Bản Ghi Dự Kiến
Video đa lĩnh vực	YouTube	30.000 – 50.000
Âm nhạc	Spotify	30.000 – 50.000
Hài hước / Viral	TikTok	10.000 – 20.000
Xu hướng xã hội	Google Trends	Feature phụ
Tổng		70.000

4. BÀI TOÁN NGHIÊN CỨU (RESEARCH PROBLEM DEFINITION)

3.1. Câu hỏi nghiên cứu tổng quát

Những yếu tố nào (thời điểm đăng tải, mức độ tương tác, tiêu đề, phản hồi tiêu cực) ảnh hưởng đáng kể đến khả năng một nội dung lọt vào danh sách Trending trên YouTube và Spotify?

Câu hỏi này được chia nhỏ thành các giả thuyết cụ thể (xem phần 4), và sẽ được kiểm định thông qua dữ liệu thu thập từ APIs của hai nền tảng.

3.2. Biến nghiên cứu (Research Variables)

Để định lượng bài toán, chúng ta định nghĩa các biến nghiên cứu rõ ràng, dựa trên lý thuyết hành vi người dùng và kinh tế nội dung số.

Biến độc lập (Independent Variables – Các yếu tố ảnh hưởng):

Giờ đăng video: Thời điểm trong ngày (e.g., 0-23 giờ).

Ngày trong tuần: Ngày cụ thể (e.g., Monday-Sunday), để kiểm tra xu hướng cuối tuần.

Từ khóa tiêu đề: Phân tích ngôn ngữ, như số lượng từ khóa kích thích (e.g., "SỐC", "BÍ MẬT").

Like, Comment, Dislike: Số lượng tương tác tích cực/tiêu cực.

Tỷ lệ comment tiêu cực: Phân tích sentiment của comments (e.g., sử dụng NLP để tính tỷ lệ negative comments).

Biến phụ thuộc (Dependent Variable – Kết quả cần dự đoán):

Trạng thái Trending (0/1): Nội dung có lọt vào danh sách Trending hay không (dựa trên API data).

Trending Rank: Thứ hạng trong danh sách Trending (nếu có, để phân tích mức độ viral).

Những biến này sẽ được thu thập và xử lý thông qua pipeline dữ liệu, đảm bảo tính khách quan và khả năng kiểm định thống kê.

4. GIẢ THUYẾT NGHIÊN CỨU CHI TIẾT (DETAILED HYPOTHESES)

Dựa trên lý thuyết hành vi người dùng (e.g., từ tâm lý học và kinh tế số), chúng ta đặt ra các giả thuyết có thể kiểm định bằng dữ liệu. Mỗi giả thuyết bao gồm H_0 (null hypothesis – không có ảnh hưởng) và H_1 (alternative hypothesis – có ảnh hưởng), với cơ sở lý thuyết rõ ràng.

Giả thuyết 1 – Thời điểm đăng tải

H_{01}: Thời điểm đăng tải video không ảnh hưởng đến xác suất lọt Top Trending. (Không có sự khác biệt thống kê giữa các khung giờ.)

H_{11}: Video được đăng trong khung giờ nghỉ ngơi (18h–22h) có xác suất lọt Trending cao hơn có ý nghĩa thống kê so với video đăng trong giờ hành chính (8h–17h). (p-value < 0.05 từ t-test hoặc chi-square test.)

Lý do khoa học: Theo lý thuyết về hành vi người dùng (e.g., từ nghiên cứu của Nielsen Norman Group), người dùng có xu hướng tiêu thụ nội dung giải trí nhiều hơn sau giờ làm việc, khi họ thư giãn và có thời gian rảnh. Điều này dẫn đến tỷ lệ tương tác cao hơn, tăng khả năng thuật toán đề xuất đẩy nội dung lên Trending.

Giả thuyết 2 – Ngôn ngữ tiêu đề (Title Linguistics)

H_{02}: Việc sử dụng từ khóa giật gân trong tiêu đề không làm thay đổi đáng kể lượt xem trung bình. (Không có sự khác biệt giữa nhóm có từ khóa kích thích và nhóm không.)

H_{12}: Các video chứa từ khóa mang tính kích thích cảm xúc (SỐC, LỌ, REVIEW, BÍ MẬT, DRAMA) có lượt xem trung bình cao hơn nhóm còn lại. (Sử dụng ANOVA hoặc regression để kiểm định.)

Cơ sở lý thuyết: Hiệu ứng Curiosity Gap (từ tâm lý học, e.g., nghiên cứu của Loewenstein, 1994), trong đó tiêu đề tạo ra sự tò mò, khuyến khích click-through. Điều này được chứng minh trong các nghiên cứu về SEO và viral content trên mạng xã hội.

Giả thuyết 3 – Tương tác tiêu cực

H_{03}: Tỷ lệ dislike và comment tiêu cực không ảnh hưởng đến khả năng viral. (Không có mối quan hệ giữa tỷ lệ tiêu cực và trạng thái Trending.)

H_{13}: Tỷ lệ tương tác tiêu cực cao làm giảm khả năng xuất hiện trong Trending, dù tổng lượt xem lớn. (Correlation analysis hoặc logistic regression để kiểm định.)

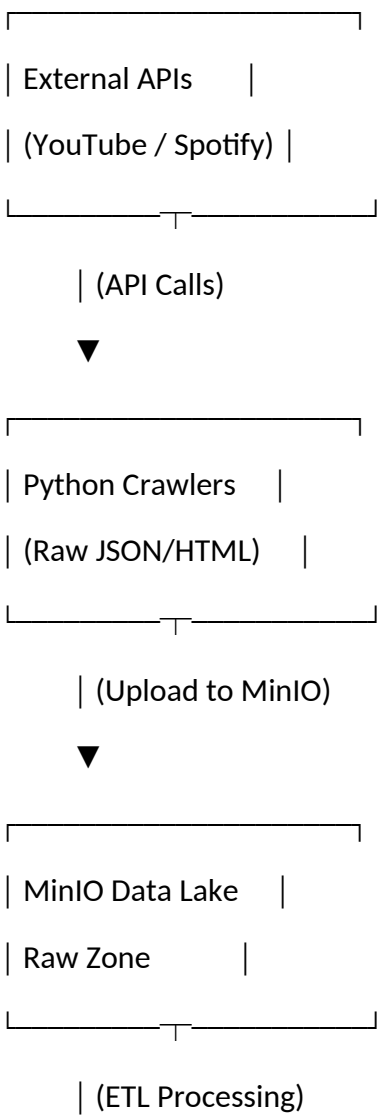
Ý nghĩa: Phân biệt giữa “viral tích cực” (e.g., nội dung hài hước) và “viral tiêu cực” (e.g., drama hoặc scandal), dựa trên lý thuyết về social proof và thuật toán đề xuất (e.g., YouTube giảm xếp hạng nội dung có nhiều dislike để tránh trải nghiệm xấu cho người dùng).

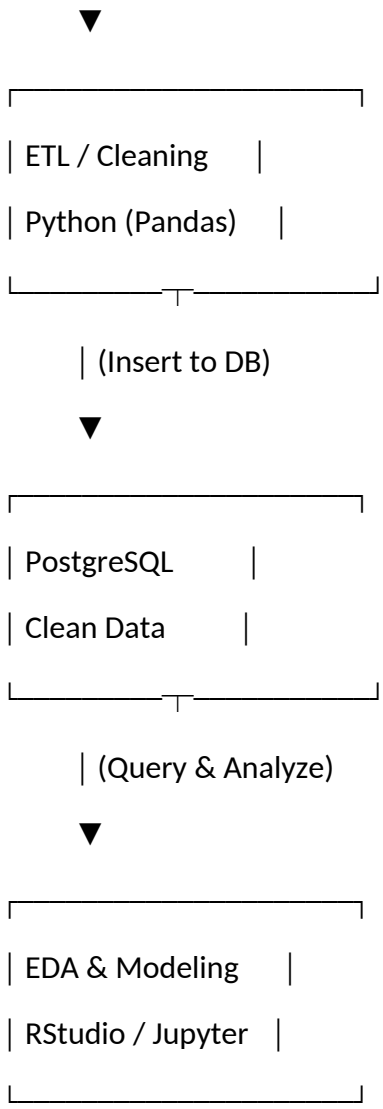
Những giả thuyết này sẽ được kiểm định trong Report 2 và 3 thông qua EDA và mô hình thống kê.

5. KIẾN TRÚC HỆ THỐNG (ENTERPRISE-LIKE ARCHITECTURE)

5.1. Tổng quan kiến trúc

Kiến trúc được thiết kế theo mô hình Enterprise Data Pipeline, với các layer rõ ràng để đảm bảo scalability và maintainability:





Kiến trúc này mô phỏng hệ thống doanh nghiệp, với Data Lake để lưu trữ thô và Database cho dữ liệu sạch, hỗ trợ cả batch và real-time processing trong tương lai.