

# Evaluation and Prediction of TikTok KOL Effectiveness in the Cosmetics Sector: A Survey and Modeling Study

Ung Hoàng Long

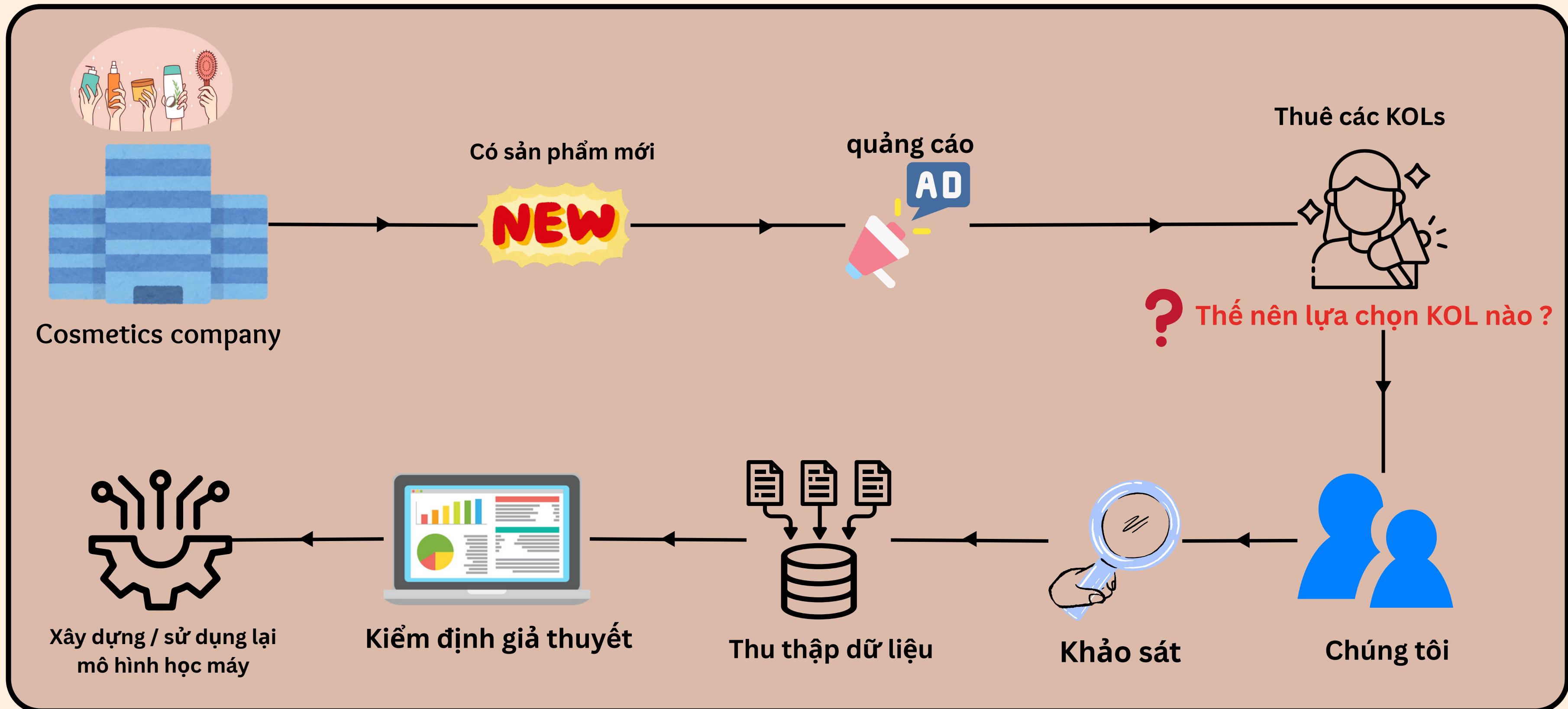


GVHD

Hà Xuân Hoàng

TS. Đỗ Trọng Hợp

# Bối cảnh & Quy trình thực hiện dự án



# Khảo sát

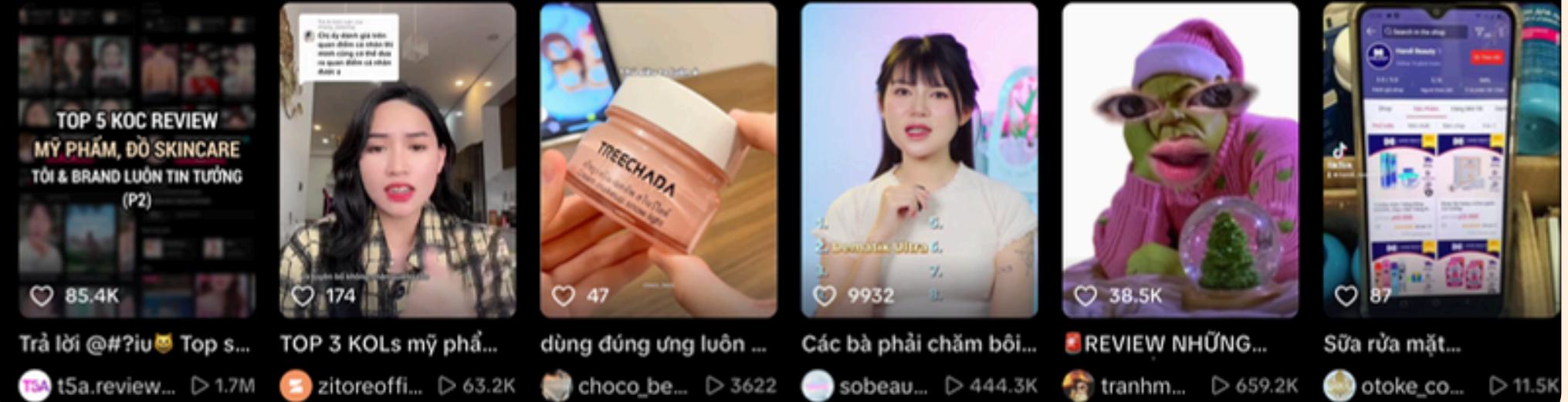
Our update > Brand Marketing > [TOP 10 KOL mỹ phẩm uy tín tại Việt Nam và xu hướng ngành mỹ phẩm](#)

## TOP 10 KOL mỹ phẩm uy tín tại Việt Nam và xu hướng ngành mỹ phẩm

 Ban biên tập Media Lab | BRAND MARKETING | July 08, 2024 8:01 AM

 Media Lab

### Top KOL Review Mỹ Phẩm Đúng ▾



REVU Vietnam

Influencer Marketing Agency and Platform @ REVU Vietnam

Trang chủ > Cộng đồng > Quảng cáo & Truyền thông

## REVU: Danh sách 10 Beauty Influencer các nền tảng

03/03/2024 • ↗ 9,957 • 0

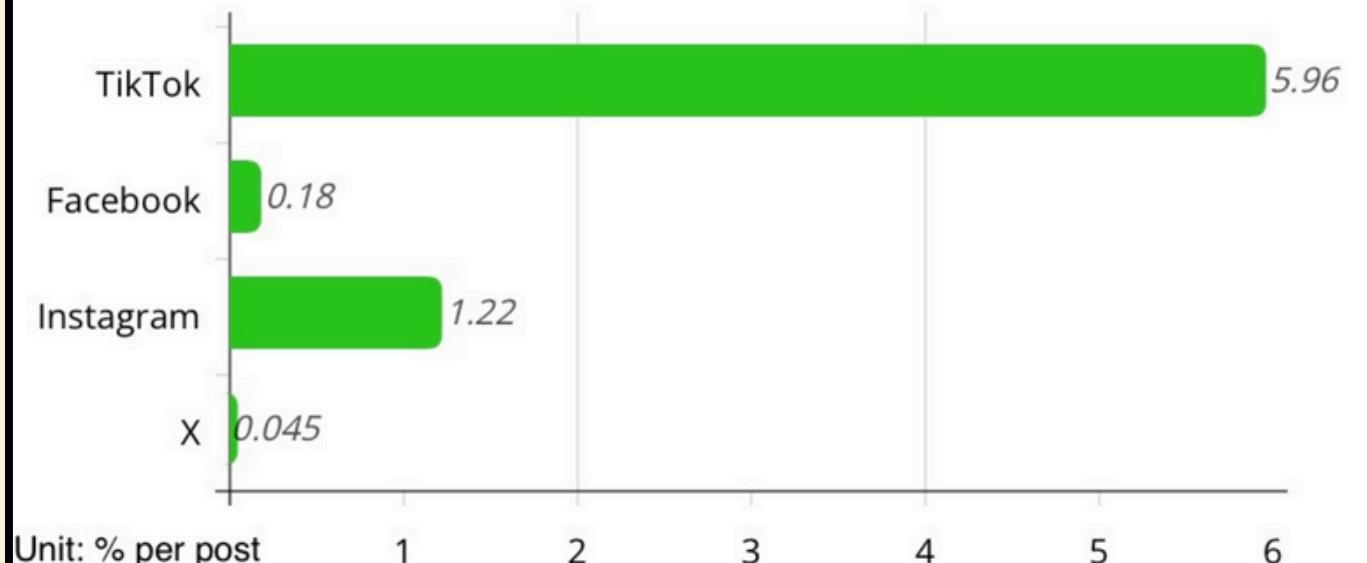
**Tiktok: nền tảng thu thập dữ liệu  
86: KOL chuyên về làm đẹp, pr và review mỹ phẩm**

Vì chúng tôi **không có dữ liệu về doanh thu** của các sản phẩm được các KOL đó PR.

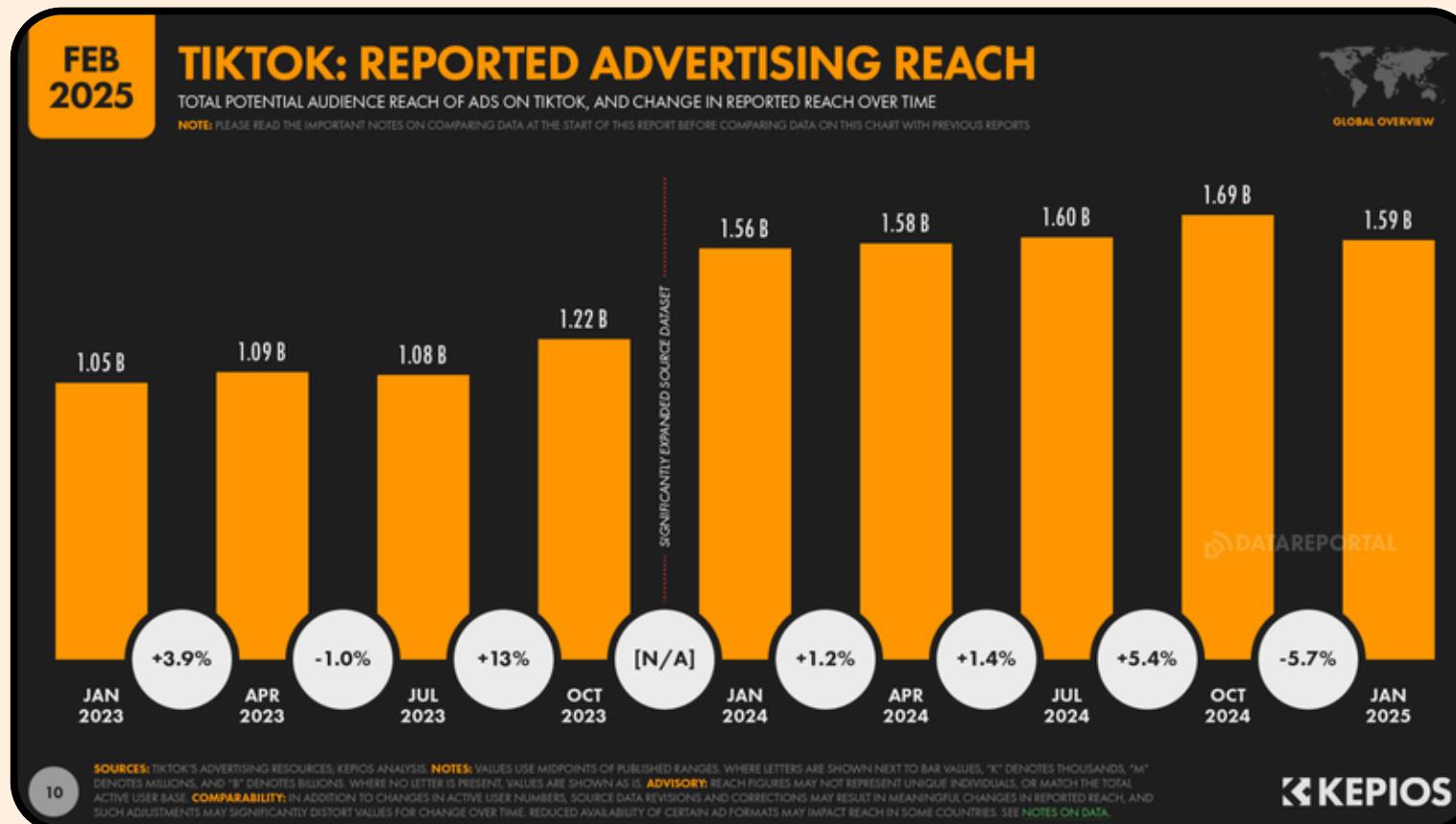
=> Nên chúng tôi chọn **Target là tỷ lệ tương tác trên từng lượt xem trong 3 tháng** của video PR

# Why TikTok ?

## Engagement Rate

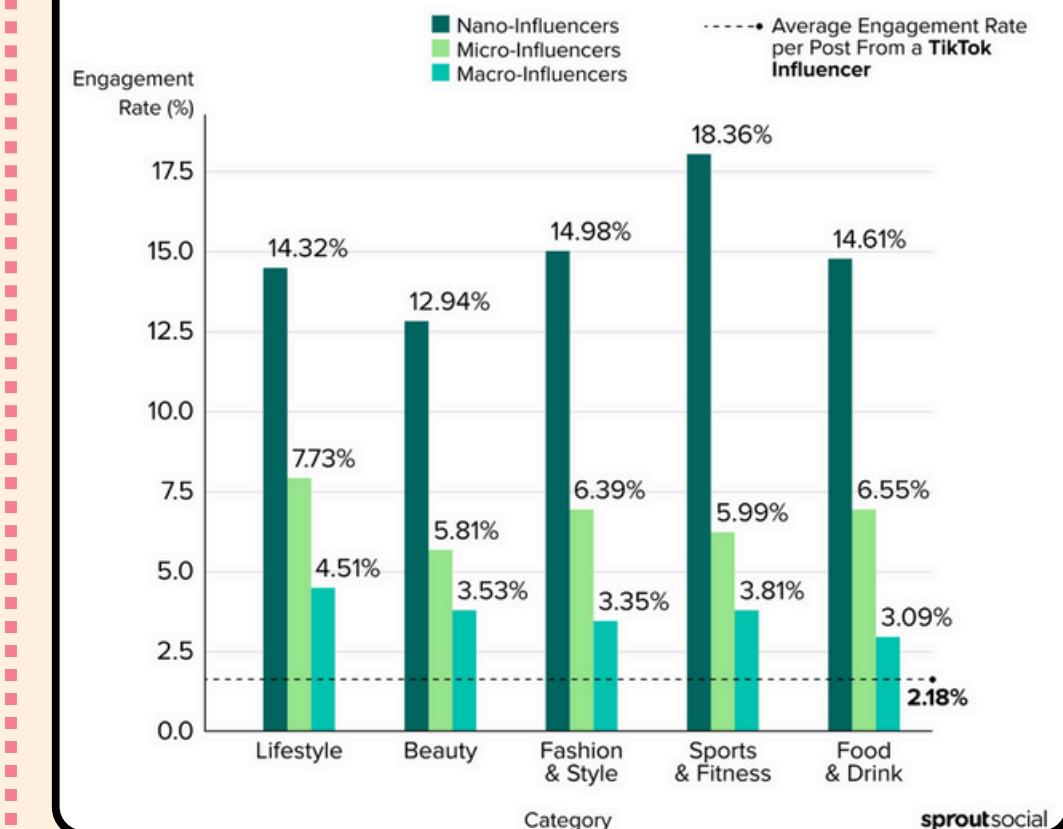


TikTok có mức độ tương tác cao nhất trong các nền tảng mạng xã hội phổ biến – là môi trường lý tưởng để đo lường và phân tích hiệu quả nội dung của KOL trong ngành mỹ phẩm.



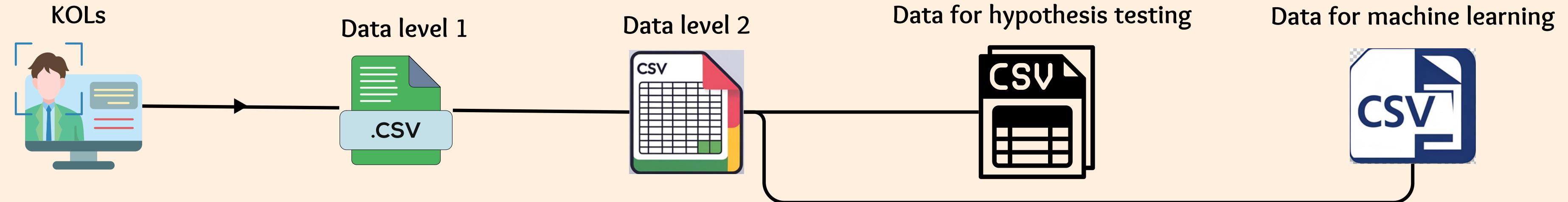
TikTok sở hữu quy mô tiếp cận quảng cáo toàn cầu lớn và đang tăng trưởng ổn định – tạo điều kiện thuận lợi cho các chiến dịch PR sản phẩm mỹ phẩm tiếp cận người tiêu dùng tiềm năng.

## TikTok Influencer Engagement Rates by Category and Influencer Size (2024)



Tỷ lệ tương tác cao của KOL trên TikTok – đặc biệt là nhóm nano & micro – là cơ sở chọn nền tảng này để phân tích hiệu quả PR sản phẩm làm đẹp.

# Tổng quan về dữ liệu



## Data level 1:

- Thời gian: 1/1/2024 - 31/5/2025
- Số lượng KOL: **86** người ( 1 người ~ 350 bài đăng )
- mỗi sample là 1 bài đăng

## Data level 2:

- Số lượng KOL: **75** người
- Dạng **window sliding**
- Nhóm theo tháng ( 6 tháng input, 1 tháng buffer, 3 tháng output )
- Mỗi sample đại diện cho mỗi KOL trong ( 1 KOL có thể được biểu diễn thành 2 sample vì tính chất window sliding )

## Data for hypothesis testing:

- **Input:** 8/2024- 1/2025
- **Output:** 3/2025 - 5/2025

## Data for machine learning:

- **Tương tự như Data level 2**
- **Được chia train, dev, test cẩn thận** ( **dataleakage problem** )

## Data level 1

Chúng tôi thu thập thông tin của  APIFY  
tất cả bài đăng của 86 KOL  
từ 1/1/2024 - 31/5/2025

- Thông tin cá nhân về KOL (nickname, tổng tương tác, tổng số follower)
- Các chỉ số tương tác (tim, share, comment, save collection, số lượt xem )
- Thông tin về bài đăng ( caption, hashtag, ngày đăng, music, độ dài video, chất lượng video, video có phải pr hay không,...)

### Thuộc tính

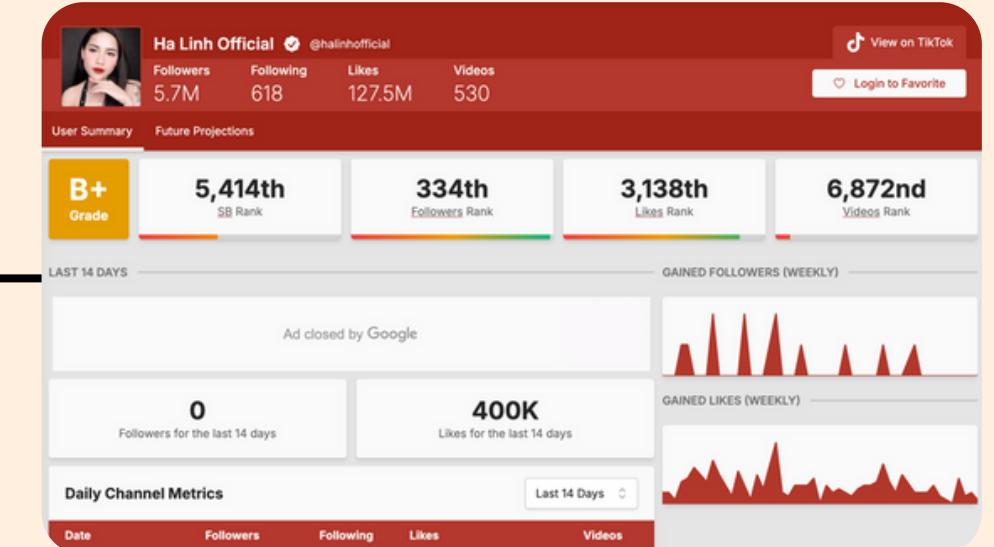
- ‘text’: caption bài đăng
- ‘createTime’: Thời gian đăng bài
- ‘isAD’: Đánh dấu bài đăng PR
- ‘musicName’: Tên bài hát
- ‘musicId’: Id nhạc của Tiktok
- ‘duration’: Thời lượng bài đăng
- ‘diggCount’: Số tim
- ‘shareCount’: Số share
- ‘playCount’: Số lượt xem
- ‘collectCount’: Số lưu về collection
- ‘commentCount’: Số comment
- ‘hashtags’: List hashtag được sử dụng
- ‘isSlideshow’: Bài đăng dạng slide
- ‘isSponsored’: Bài đăng được tài trợ
- ‘ttSeller’: KOL được Tiktok hỗ trợ
- ‘commerceUser’: Người dùng thương mại

# Data level 1

- Số lượt follower thay đổi mỗi ngày
- Tỷ lệ giới tính của followers
- Tỷ lệ follower là người Việt Nam

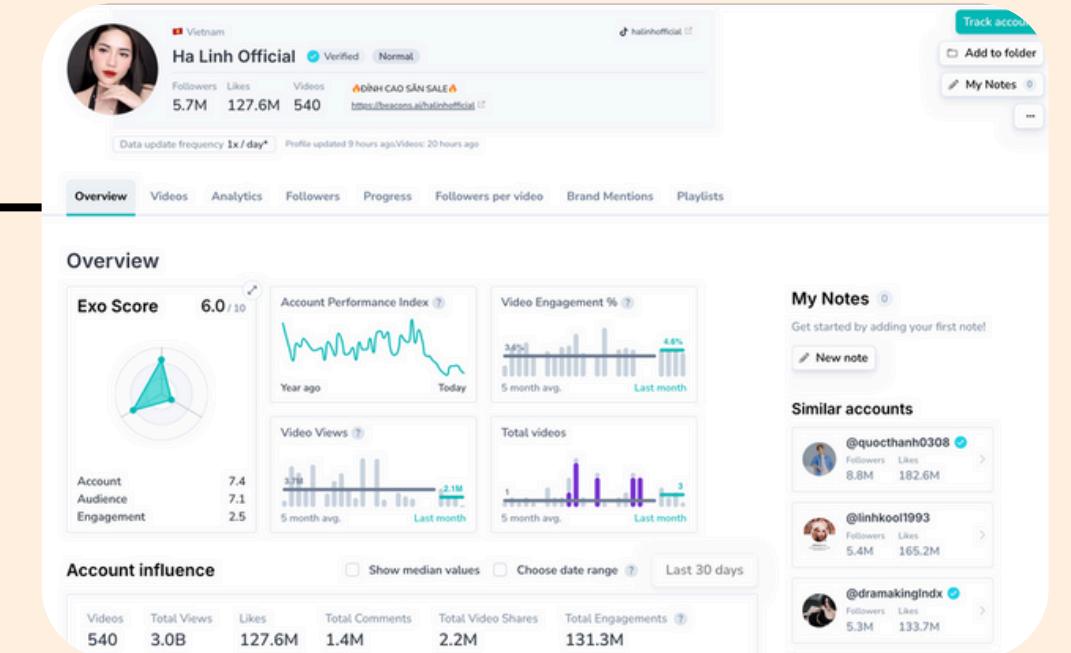


SOCIALBLADE



- Số chữ trong caption
- Số icon trong caption
- Số hashtags
- Khung giờ đăng bài
- Bài đăng vào cuối tuần

exolyt



- ‘is\_trending’: Bài đăng có caption trending

Gemini 2.0 Flash

RPM: 30

TPM: 1,000,000

RPD: 200

Knowledge cutoff: August 2024

## Data level 2

Mỗi sample đại diện cho mỗi khung thời gian cho KOL.

Vì các sample lúc này không còn là mỗi bài đăng nữa nên phải tạo thêm các thuộc tính đại diện cho KOL đó trong khoảng thời gian cố định

**Input: 6 tháng**

- Từ các thuộc tính ở Data level 1:

- Tổng các thuộc tính thô
- Tỷ lệ từ các thuộc tính thô
- Tổng số bài PR có lượng tương tác đột biến ( outlier ) trong khung thời gian cố định
- Tổng bài hát độc nhất được sử dụng

Bỏ 1 tháng buffer -----

**Output: 3 tháng**

- Tỷ lệ tương tác / mỗi lượt xem:

$$\frac{\text{Digg Count} + \text{Share Count} + \text{Collect Count} + \text{Comment Count}}{\text{Play Count}}$$

## Khung thời gian

| Số sample cho mỗi KOL | Input           | Output           |
|-----------------------|-----------------|------------------|
| 1                     | 1-6/2024        | 8-10/2024        |
| 2                     | 2-7/2024        | 9-11/2024        |
| 3                     | 3-8/2024        | 10-12/2024       |
| 4                     | 4-9/2024        | 11/2024 - 1/2025 |
| 5                     | 5-10/2024       | 12/2024 - 2/2025 |
| 6                     | 6-11/2024       | 1-3/2025         |
| 7                     | 7-12/2024       | 2-4/2025         |
| 8                     | 8/2024 - 1/2025 | 3-5/2025         |

# Data for hypothesis testing

## Đối với doanh nghiệp:

- Khám phá các xu hướng mới trong hành vi hoặc đặc điểm của KOL theo thời gian gần đây.
- Hiểu rõ các yếu tố ảnh hưởng đến hiệu quả (target) của từng KOL.
- Phân nhóm KOL dựa trên các thuộc tính đặc trưng, từ đó xác định các nhóm nổi bật.
- Sàng lọc trước các KOL tiềm năng, giúp tối ưu chi phí và hiệu quả trước khi đưa vào mô hình dự đoán.

## Đối với mô hình dự đoán:

- Hoạt động như một bước lọc dữ liệu đầu vào, chỉ giữ lại những KOL có tín hiệu tích cực, phù hợp với xu hướng hiện tại.
- Giảm nhiễu và tăng độ chính xác cho quá trình huấn luyện mô hình.
- Là cơ sở để ra quyết định chiến lược:
  - Nên tiếp tục cập nhật mô hình bằng mini-batch online learning
  - Hay cần retrain mô hình hoàn toàn mới nếu xuất hiện xu hướng/thuộc tính mới.

## Khung thời gian

| Input           | Output   |
|-----------------|----------|
| 8/2024 - 1/2025 | 3-5/2025 |

Mỗi sample là 1 KOL.  
Tổng KOL: 61

## Giới thiệu về các phương pháp kiểm định

### 1. T - test

- Là một phương pháp thống kê tham số (parametric test) dùng để so sánh trung bình hai nhóm, nhằm xác định xem có sự khác biệt về mặt thống kê giữa hai nhóm hay không
- Có 3 loại t-test chính :
  1. Independent t-test
  2. Paired t-test
  3. One-sample t-test
- Giả định của t - test :
  - Dữ liệu phân phối chuẩn (gần chuẩn).
  - Dữ liệu liên tục và độc lập.
  - Phương sai của hai nhóm bằng nhau.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

### 2. ANOVA

- Là một phương pháp thống kê tham số (parametric test) dùng để so sánh trung bình của ba nhóm độc lập trở lên.
- Nguyên lý hoạt động :

Phân tích sự khác nhau giữa các nhóm bằng cách so sánh phương sai giữa các nhóm.

Sử dụng F thống kê để đánh giá
- Giả định của ANOVA:
  - Dữ liệu phân phối chuẩn (gần chuẩn).
  - Dữ liệu liên tục và độc lập.
  - Phương sai của hai nhóm bằng nhau.

$$F = \frac{MS_{between}}{MS_{within}}$$

- MSbetween : Phương sai giữa các nhóm
- MSwithin : Phương sai trong mỗi nhóm

# Hypothesis testing

Nếu như bộ dữ liệu không thỏa mãn các giả định thì sao ?

## 1. Mann-Whitney U - t - test

- Là một kiểm định phi tham số.
- Sử dụng các tham số (mean) và phương sai (var) để mô hình hóa.
- Vượt trội hơn t - test khi phân phối không chuẩn.
- Mann-Whitney U so sánh phân phối (thường là trung vị) của hai nhóm

## 2. Kruskal-Wallis - ANOVA

- Là một kiểm định phi tham số.
- Là một phiên bản mở rộng của Mann-Whitney U.
- Gộp tất cả dữ liệu từ các nhóm và xếp hạng (rank) chúng từ nhỏ đến lớn.
- Sử dụng thống kê H để đánh giá xem các nhóm có khác biệt về phân phối hay không.

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

Muốn biết được mức độ khác nhau giữa 2 nhóm thì phải làm sao ?

## Effect Size

- Effect size là một thước đo định lượng mức độ khác biệt giữa hai nhóm.
- Dùng Cohen's d để đo lường

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}$$

Điễn giải (theo Cohen, 1988):

- $d \sim 0.1$  : Hiệu ứng nhỏ
- $d \sim 0.3$  : Hiệu ứng trung bình.
- $d \sim 0.5$  : Hiệu ứng lớn

- Khi đã biết mức độ khác nhau giữa hai nhóm. Việc xác định xem nhóm nào lớn hơn nhóm nào bằng cách :

- Dựa trung bình của hai nhóm :

Group1.mean() & Group2.mean()

## Hypothesis testing

- Giả thuyết này kiểm tra xem KOL đánh dấu tài khoản của mình là bán hàng trên tiktok có mang lại tỉ lệ tương tác tại 3 tháng output khác biệt so với tài khoản không được đánh dấu hay không ?
- **H0 :** Trung bình tỉ lệ tương tác 3 tháng sau tại hai nhóm True và False không có sự khác biệt.
- **H1 :** Trung bình tỉ lệ tương tác 3 tháng sau tại hai nhóm True và False có sự khác biệt.

- **True group :** Tỉ lệ tương tác tại 3 tháng output ứng với tài khoản là tài khoản bán hàng ('commerceUser' = True)
- **False group :** Tỉ lệ tương tác tại 3 tháng output ứng với tài khoản là tài khoản bán hàng ('commerceUser' = False)

### Kết luận :

- Phân phối của hai nhóm không chuẩn.
- Phương sai của hai nhóm không đồng đều.  
→ Sử dụng Mann-Whitney U.
- **p\_value = 0.4427** → Không có sự khác biệt phân phối giữa hai nhóm True và False
- Việc chỉ gán nhãn tài khoản "bán hàng" không mang lại tỉ lệ tương tác ở 3 tháng sau khác biệt so với việc không gán nhãn.

## Hypothesis testing

- Giả thuyết này kiểm tra xem KOL được Tiktok đánh dấu tài khoản của mình là bán hàng có mang lại tỉ lệ tương tác tại 3 tháng output khác biệt so với tài khoản không được đánh dấu hay không ?
- **Ho :** Trung bình tỉ lệ tương tác 3 tháng sau tại hai nhóm True và False không có sự khác biệt.
- **H1 :** Trung bình tỉ lệ tương tác 3 tháng sau tại hai nhóm True và False có sự khác biệt.

- **True group :** Tỉ lệ tương tác tại 3 tháng output ứng với tài khoản là tài khoản bán hàng ('ttSeller' = True)
- **False group :** Tỉ lệ tương tác tại 3 tháng output ứng với tài khoản là tài khoản bán hàng ('ttSeller' = False)

### Kết luận :

- Phân phối của hai nhóm không chuẩn.
- Phương sai của hai nhóm không đồng đều.  
→ Sử dụng Mann-Whitney U.
- **p\_value = 0.7147** → Không có sự khác biệt phân phối giữa hai nhóm True và False.
- Việc được Tiktok gán nhãn tài khoản “bán hàng” không đồng nghĩa với việc tạo ra sự khác biệt về tỉ lệ tương tác với tài khoản không được gán nhãn.
- Tự gán nhãn hoặc được Tiktok gán nhãn tài khoản “bán hàng” không được người dùng tin tưởng cao, thương hiệu cá nhân, uy tín, độ nổi tiếng có thể mang lại sự tin tưởng cao hơn thay vì một dấu tích “tài khoản bán hàng”.
- Các doanh nghiệp không thể dựa vào 'ttSeller' hay 'commerceUser' để chọn lựa KOL.

# Hypothesis testing

- Giả thuyết : Giả thuyết này nhằm kiểm tra xem mức độ tương tác ở 3 tháng sau("EngagementRateOnPRPost\_target") có bị ảnh hưởng bởi ngưỡng tỷ lệ tương tác 6 tháng đầu vào ("pr\_engagement\_rate\_input") hay không ?
- Ho : Trung bình của 'EngagementRateOnPRPost\_target' giữa group\_low (nhóm có 'pr\_engagement\_rate\_input' < 0.02) và group\_high (nhóm có 'pr\_engagement\_rate\_input' >= 0.02) không khác biệt có ý nghĩa thống kê.
- H1 : Trung bình của 'EngagementRateOnPRPost\_target' giữa hai nhóm khác biệt có ý nghĩa thống kê.

- **High group** : Tỉ lệ tương tác tại 3 tháng sau ('EngagementRateOnPRPost\_target') ứng với tỉ lệ tương tác tại 6 tháng ban đầu ('pr\_engagement\_input') >= 0.02.
- **Low group** : Tỉ lệ tương tác tại 3 tháng sau ('EngagementRateOnPRPost\_target') ứng với tỉ lệ tương tác tại 6 tháng ban đầu ('pr\_engagement\_input') < 0.02.

## Kết luận :

- Phân phối của hai nhóm không chuẩn
- Phương sai của hai nhóm không đồng đều.  
→ Sử dụng Mann-Whitney U.
- **p\_value = 0.0119** → Có sự khác biệt giữa high group và low group.
- Hiệu ứng trung bình (**R = 0.323**) :
  - Việc tỉ lệ tương tác cao ở quá khứ sẽ mang lại tỉ lệ tương tác cao ở tương lai. Giả thuyết cũng chỉ ra được rằng các KOL có tỉ lệ tương tác cao thường ổn định (lượng Share, Comment, Digg, ... ) không thay đổi nhiều qua các tháng).
  - Các công ty nên lựa chọn các KOL có tỉ lệ tương tác trong quá khứ tốt, như vậy thì tỉ lệ tương tác cho các bài PR booking sau này sẽ có tỉ lệ tương tác cao hơn.

## Hypothesis testing

- Giả thuyết này kiểm tra xem liệu ở mức ngưỡng cụ thể của tổng số bài đăng là `isSlideshow_nonpr` có ảnh hưởng đến tỉ lệ tương tác ở 3 tháng sau hay không ?
- $H_0$  : Trung bình của 'EngagementRateOnPRPost\_target' giữa `group_true` (nhóm có '`isSlideshow_nonpr`' < 10) và `group_false` (nhóm có '`isSlideshow_nonpr`' >= 10) không khác biệt có ý nghĩa thống kê.
- $H_1$  : Trung bình của 'EngagementRateOnPRPost\_target' giữa hai nhóm khác biệt có ý nghĩa thống kê.

- **High group** : Tỉ lệ tương tác tại 3 tháng sau ('EngagementRateOnPRPost\_target') ứng với số lượng bài đăng không phải là PR dạng slideShow “`isSlideshow_nonpr`” >= 10.
- **Low group** : Tỉ lệ tương tác tại 3 tháng sau ('EngagementRateOnPRPost\_target') ứng với số lượng bài đăng không phải là PR dạng slideShow “`isSlideshow_nonpr`” < 10.

### Kết luận :

- Phân phối hai nhóm không chuẩn.
- Phương sai hai nhóm không đồng đều.  
→ Sử dụng Mann-Whitney U.
- **p\_value = 0.0275** → Có sự khác biệt giữa high group và low group.
- Hiệu ứng nhỏ (**R = 0.27**) :
  - Mặc dù có sự khác biệt về mặt thống kê, tổng số bài đăng là SlideShow và không phải bài PR không mang lại sự khác biệt đáng kể cho tỉ lệ tương tác trên bài đăng PR cho cả hai nhóm (thấp hoặc cao).

# Hypothesis testing

- Giả thuyết này kiểm tra xem liệu ở mức ngưỡng cụ thể của tổng số bài đăng là PR có ảnh hưởng đến lượng bài đăng có tỉ lệ tương tác thuộc nhóm vượt trội hay không?
- $H_0$  : Trung bình của lượng bài đăng PR có tỉ lệ tương tác thuộc nhóm vượt trội giữa group\_low (nhóm có 'total\_posts\_pr'  $\leq 100$ ) và group\_high (nhóm có 'total\_posts\_pr'  $> 100$ ) không khác biệt có ý nghĩa thống kê.
- $H_1$  : Trung bình của lượng bài đăng PR có tỉ lệ tương tác thuộc nhóm vượt trội giữa hai nhóm khác biệt có ý nghĩa thống kê.

- **High group** : Số lượng bài đăng ngoại lai (“e\_outlier\_pr”) ứng với tổng số bài đăng PR “total\_posts\_pr”  $> 100$ .
- **Low group** : Số lượng bài đăng ngoại lai (“e\_outlier\_pr”) ứng với tổng số bài đăng PR “total\_posts\_pr”  $\leq 100$ .

## Kết luận:

- Phân phối của hai nhóm không chuẩn.
- Phương sai giữa hai nhóm không đồng đều.  
→ Sử dụng Mann-Whitney U.
- **p\_value = 0.0141** → Có sự khác biệt giữa hai nhóm high group và low group.
- Hiệu ứng trung bình (**R = 0.36**):
  - Việc đăng nhiều bài PR ( $> 100$  bài) có thể tác động lên số lượng bài PR có tỉ lệ tương tác thuộc nhóm vượt trội.
  - Các KOL nên thúc đẩy số lượng bài PR làm tăng số lượng bài PR lên trending có thể thu hút nhiều lượt tương tác của khách hàng.
  - Các công ty cũng nên chọn lựa các KOL có số lượng đăng bài cao để tăng khả năng bài đăng PR công ty trở nên viral.

## Hypothesis testing

- Giả thuyết này kiểm tra xem liệu ở mức ngưỡng cụ thể của tổng số caption của bài đăng PR mang nội dung trending có ảnh hưởng đến lượng follower mới hay không?
- $H_0$ : Trung bình của `follower_change_in_input_window` trong nhóm `high_trending` bằng trung bình của `follower_change_in_input_window` trong nhóm `low_trending`.
- $H_1$ : Trung bình của `follower_change_in_input_window` trong nhóm `high_trending` khác với trung bình của `follower_change_in_input_window` trong nhóm `low_trending`.

- **High trending**: Tổng lượt thay đổi follower (“`follower_change_in_input_window`”) ứng với tổng số lượng caption của bài PR có nội dung trending “`is_trending_pr`”  $\geq 20$ .
- **Low trending**: Tổng lượt thay đổi follower (“`follower_change_in_input_window`”) ứng với tổng số lượng caption của bài PR có nội dung trending “`is_trending_pr`”  $< 20$ .

### Kết luận:

- Phân phối của hai nhóm không chuẩn.
- Phương sai giữa hai nhóm không đồng đều.  
→ Sử dụng Mann-Whitney U.
- **p\_value = 0.0395** → Có sự khác biệt giữa hai nhóm High trending và Low trending.
- Hiệu ứng nhỏ (**R = 0.26**):
  - Mang dù có sự khác biệt về mặt thống kê, nhưng caption thuộc nhóm trending không có tác động đáng kể đến việc tăng/giảm follower dù là nhiều hay ít.
  - Vì không mang lại một kết quả thật sự rõ rệt, KOL không cần quá chú trọng việc đưa nội dung liên quan đến trending mà hãy tập trung vào các yếu tố khác (caption ngắn gọn, không caption...).

# Hypothesis testing

- Giả thuyết này nhằm kiểm tra xem liệu tổng số bài đăng PR được đăng vào cuối tuần có ảnh hưởng đến tổng số lượng bài mang tỉ lệ tương tác vượt trội hay không ?
- Ho : Trung bình tổng số bài đăng vượt trội cả ba nhóm (low, medium, high) là bằng nhau.
- H1 : Có ít nhất một nhóm có trung bình tổng số bài đăng vượt trội khác với các nhóm còn lại.

- **Low group** : Tổng số bài đăng vượt trội (“e\_outlier\_pr”) ứng với ‘is\_weekend\_pr’ = 0.
- **Medium group** : Tổng số bài đăng vượt trội (“e\_outlier\_pr”) ứng với  $0 < \text{'is\_weekend\_pr'} \leq 15$ .
- **High group** : Tổng số bài đăng vượt trội ('e\_outlier\_pr') ứng với ‘is\_weekend\_pr’ > 15.

## Kết luận:

- Phân phối của ba nhóm không chuẩn.
- Phương sai của ba nhóm không đồng đều.  
→ Sử dụng Kruskal-Wallis.
- **p\_value = 0.0028** → Có ít nhất một nhóm có phân phối khác với các nhóm còn lại.
- Kiểm định hậu nghiệm (post-hoc). Có sự khác biệt thống kê ở các cặp nhóm sau:
  - Phân phối nhóm low khác với nhóm high: **p\_value = 0.039 < 0.05**.
- Hiệu ứng lớn (**R = 0.79**):
  - Có nghĩa là tổng số lượng đăng bài pr vào cuối tuần lớn ( $> 15$ ) mang lại tổng số lượng bài đăng vượt trội cao hơn đáng kể so với việc không đăng bài đăng nào vào cuối tuần.
  - Bởi vì cuối tuần là thời gian người dùng có thời gian sử dụng nền tảng lớn nhất, vậy nên các KOL nên tích cực đăng tải bài post PR vào cuối tuần để có lượng tương tác cao hơn.

## Data for machine learning & Models used

- Có format giống Dataset level 2

- Test Set: Khung thời gian cuối cùng của mỗi KOL ( **96 samples** )
- Dev Test: Khung thời gian cuối cùng sau khi chia test ( **61 samples** )
- Train Test: Còn lại ( **69 samples** )

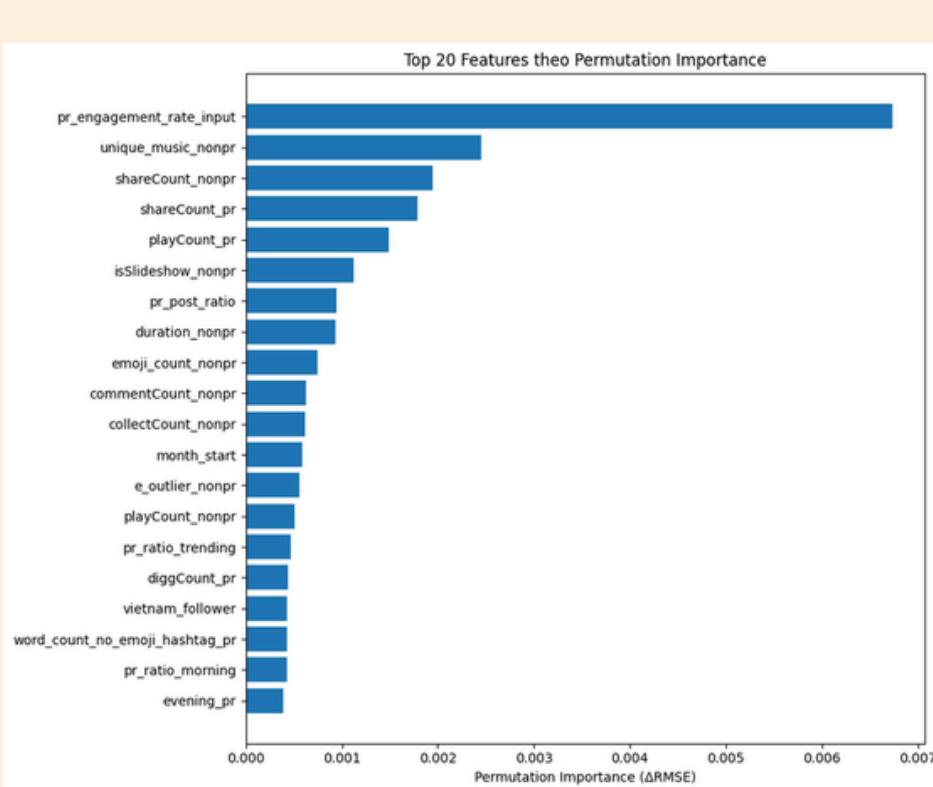
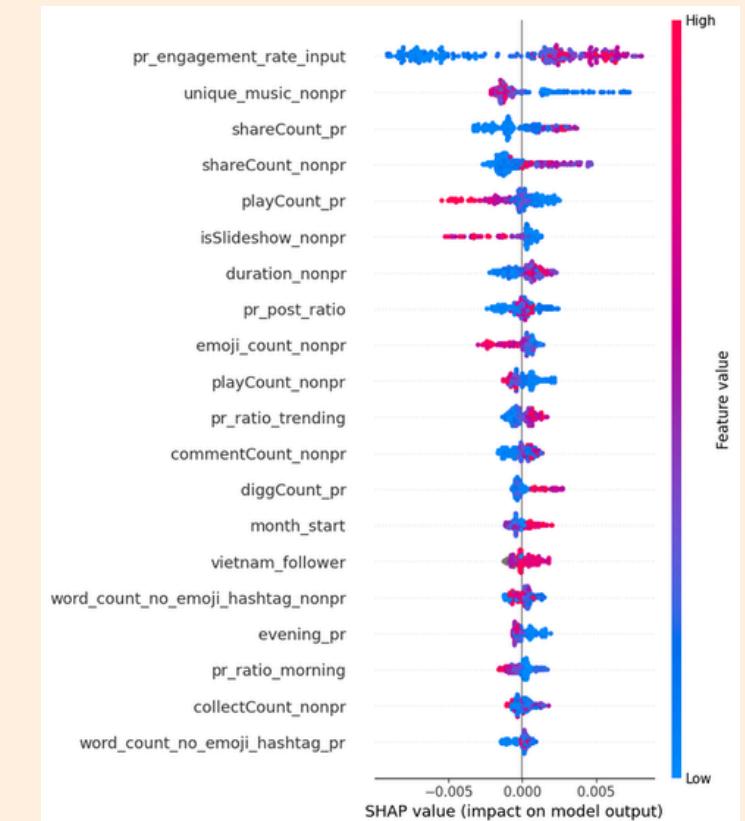
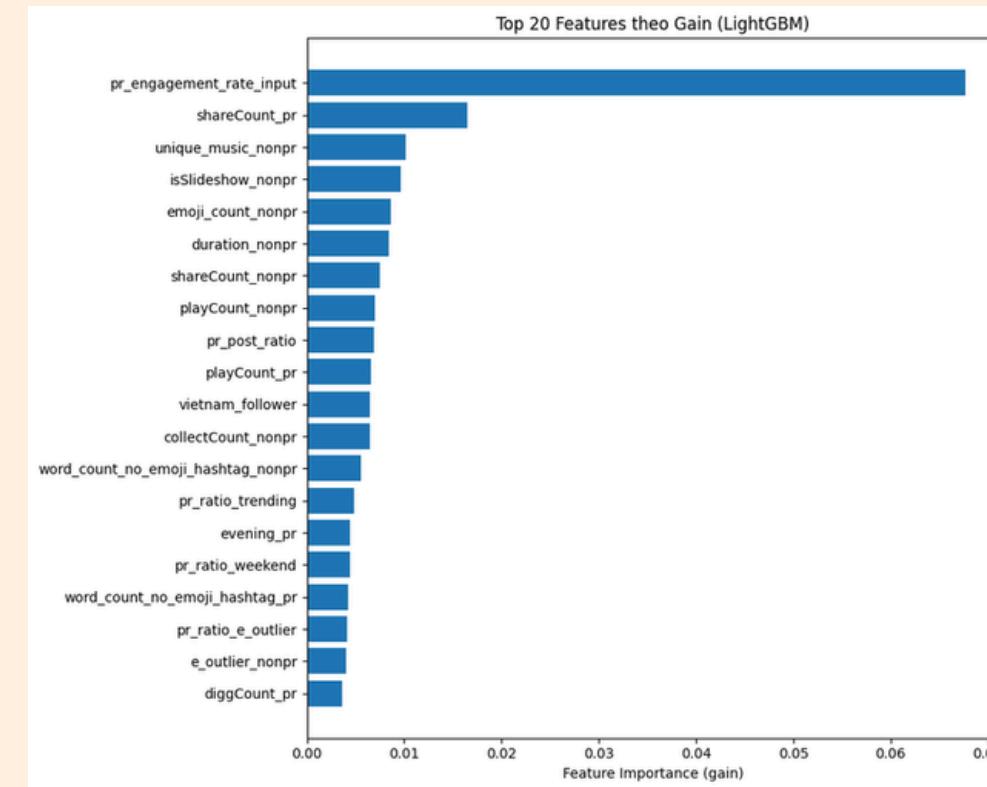
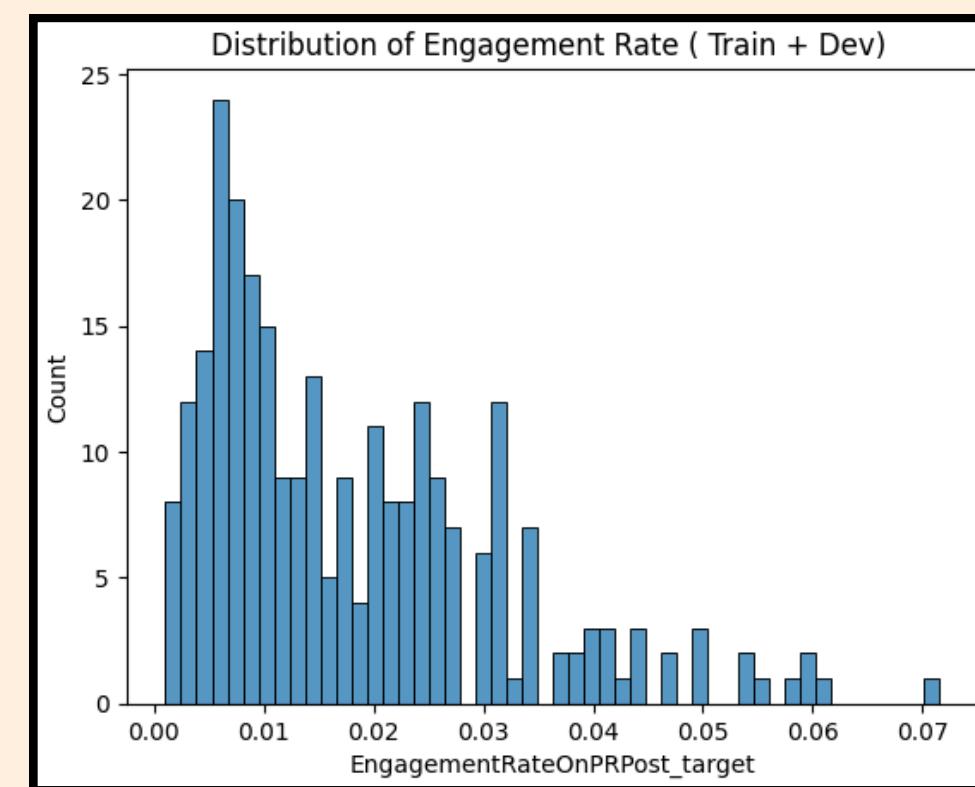
### Quy tắc chia:

- Đảm bảo khoảng thời gian của input và output trong **train test** của một KOL **không trùng** với khoảng thời gian output trong **test set**
- 

### Models

- **Linear Model:** Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression
- **Tree based Model:** RandomForestRegressor, ExtraTreesRegressor, HistGradientBoostingRegressor, XGBoost, LightGBM, CatBoost

# Data preprocessing & Feature Selection



Right skewed: log1p transform

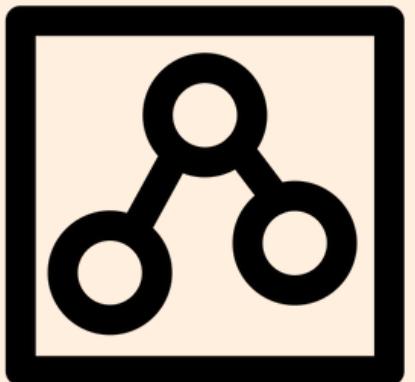
FEATURE IMPORTANCE

SHAP VALUE

PERMUTATION IMPORTANCE

Target

Tree-based Models



StandardScaler

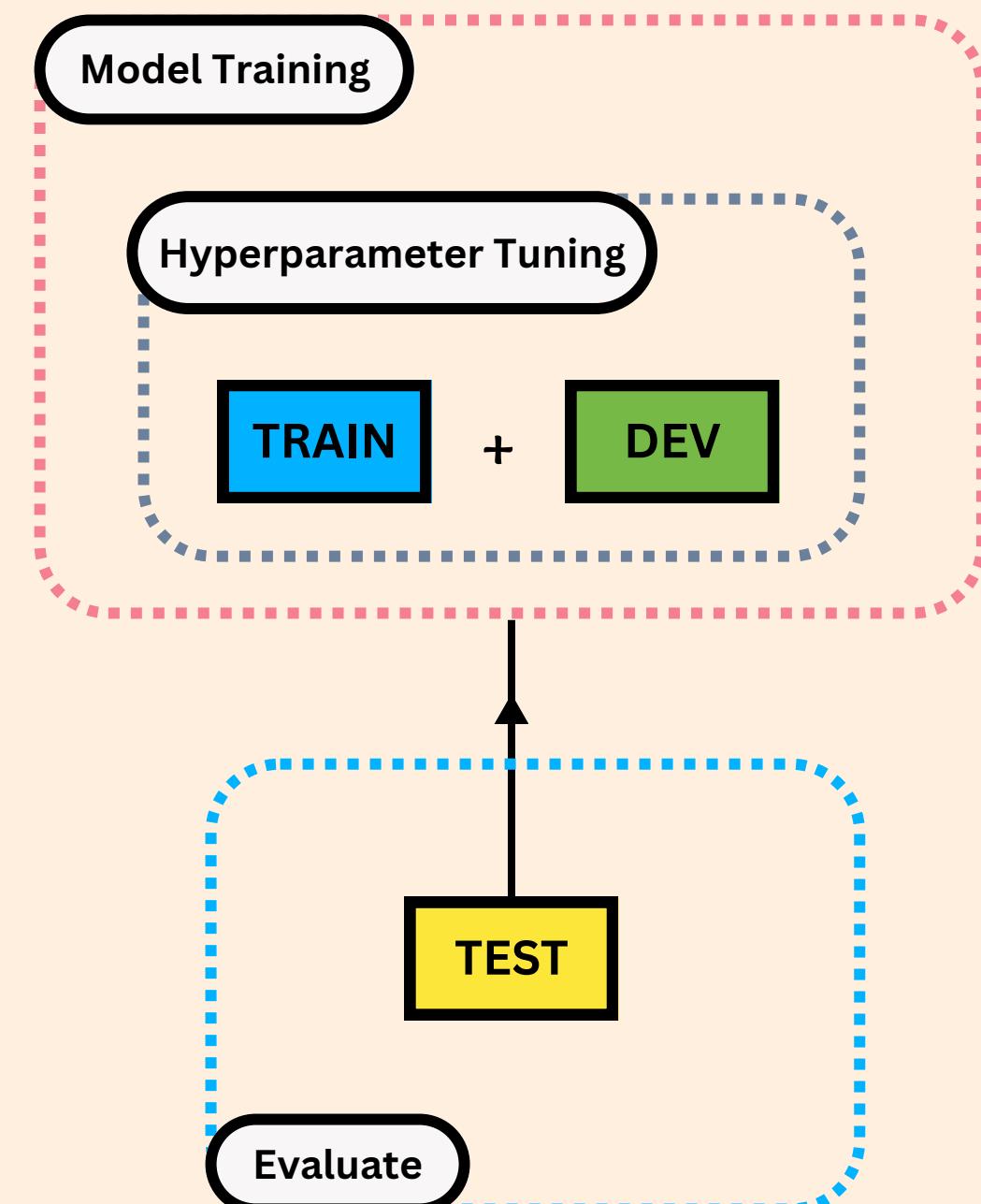
+

PCA: giữ lại ít nhất 95% phương sai gốc

Linear Models

# Results

| Models                           | MAE    | MAE / Mean Target | R^2      |
|----------------------------------|--------|-------------------|----------|
| Linear Regression                | 0.0179 | 71.39%            | - 0.1060 |
| Ridge Regression                 | 0.0173 | 68.79%            | - 0.0312 |
| Support Vector Regression        | 0.0250 | 99.77%            | - 0.0295 |
| Random Forest                    | 0.0156 | 62.37%            | - 0.0445 |
| Extra Trees Regressor            | 0.0154 | 61.23%            | - 0.0218 |
| Hist Gradient Boosting Regressor | 0.0162 | 64.49%            | - 0.0422 |
| XGBoost                          | 0.0167 | 66.78%            | - 0.0251 |
| LightGBM                         | 0.0164 | 65.46%            | - 0.0504 |
| CatBoost                         | 0.0153 | 61.07%            | - 0.0258 |



## Discussion

### Hiệu suất hiện tại của các mô hình còn thấp:

- Tất cả mô hình đều có  $R^2$  âm → nghĩa là dự đoán tệ hơn cả việc đoán trung bình.
- MAE (Mean Absolute Error) tuy nhỏ (dao động ~0.015–0.025), nhưng khi so với giá trị trung bình của target, thì MAE chiếm tới 61–99% giá trị trung bình target → độ sai lệch tương đối cao.
- Extra Trees Regressor và CatBoost là 2 models tốt nhất hiện tại, có thể cân nhắc trong tương lai

### Nguyên nhân mô hình hoạt động chưa tốt:

#### 1. Dữ liệu quá ít:

- Chỉ 267 mẫu train và 69 mẫu test → không đủ để mô hình học được quy luật sâu hơn.
- Chưa có độ phủ (đa lĩnh vực)
- Bài toán mang tính chất time series (sliding windows), nên không thể sử dụng cross-validation truyền thống để tăng hiệu quả đánh giá.

#### 2. Target lệch phải (right-skewed):

- Dù đã log1p transform, nhưng vẫn có thể gây khó khăn trong việc mô hình hóa nếu mô hình không đủ mạnh hoặc feature chưa nắm rõ đặc trưng.

#### 3. Thiếu thông tin đặc trưng mạnh (strong features):

- Một số đặc trưng đầu vào có thể không đủ sức phân biệt các KOL → cần bổ sung thêm đặc trưng

## Limitation & Future Work

### Limitation

- Dữ liệu còn ít, chưa đa lĩnh vực và thuộc tính còn giới hạn vì ngân sách hạn chế trong quá trình thu thập dữ liệu
- Chưa tận dụng được hết các tài nguyên như video, âm thanh mang tính xu hướng.
- Các thuộc tính được tạo thêm nhờ sử dụng LLM phiên bản khá cũ, chỉ ổn ở mức độ thử nghiệm
- Chưa thể hợp tác với các tổ chức để lấy được các dữ liệu độc quyền như doanh thu, số lượt click,..
- Không thể đánh giá bằng Cross-Validation

### Future work

- Xin tài trợ từ phía công ty, giúp cải thiện về việc thu thập dữ liệu
- Sử dụng các SOTA LLMs để tạo ra các thuộc tính mới và kiểm tra độ đồng thuận, giúp thuộc tính có sự tin tưởng hơn
- Mở rộng phạm vi dữ liệu sang các lĩnh vực khác → Không chỉ tập trung vào mỹ phẩm mà còn thử nghiệm với thời trang, thực phẩm, sức khoẻ,...
- Kết hợp nội dung đa phương tiện → Trích xuất đặc trưng từ video, âm thanh, hình ảnh để mô hình hiểu rõ hơn về nội dung quảng bá.
- Áp dụng mô hình học liên tục (online learning) → Cho phép hệ thống cập nhật theo thời gian thực khi có dữ liệu KOL mới.
- Cải thiện phương pháp đánh giá mô hình theo thời gian → Thiết kế lại quy trình đánh giá phù hợp hơn với cấu trúc dữ liệu dạng cửa sổ thời gian (time-window).

# Xin cảm ơn

