

Báo cáo về fanpage if.blue

Trường đại học công nghệ - ĐHQGHN

Môn học : Lập trình sử lý dữ liệu với python

Link github : [NguyenPhuongDng/final_project \(github.com\)](https://github.com/NguyenPhuongDng/final_project)

Tên : Nguyễn Phương Đông - 22022593

Báo cáo về fanpage if.blue

▼ Giới thiệu báo cáo

- Vấn đề đặt ra

Trong thời đại số hóa ngày nay, việc sử dụng các nền tảng truyền thông xã hội đã trở thành một phần không thể thiếu trong chiến lược tiếp thị của nhiều doanh nghiệp và tổ chức. Trong tương lai đầy thách thức và cơ hội này, việc đánh giá và hiểu rõ về hiệu suất của các trang fanpage trên Facebook trở nên ngày càng quan trọng.

Trong bối cảnh này, báo cáo nghiên cứu của chúng tôi tập trung vào phân tích chi tiết về một trang fanpage cụ thể trên Facebook, nhằm khám phá các yếu tố quyết định thành công, sự tương tác của cộng đồng, và chiến lược các bài đăng. Trang fanpage này không chỉ là một nơi để chia sẻ thông tin mà còn là một cơ hội để tương tác và xây dựng mối quan hệ với đối tượng của mình.

Chúng tôi sẽ đưa ra cái nhìn sâu sắc vào nội dung được đăng tải, mức độ tương tác, và các chiến lược, bài đăng trên trang fanpage, mục tiêu từ đó làm rõ những vấn đề, điểm mạnh, thách thức, và đề xuất những chiến lược tương lai có thể nâng cao hiệu suất trang fanpage này.

Dưới đây là một hành trình khám phá sự thành công và các thước đo hiệu suất của một trang fanpage trên Facebook, với hy vọng rằng thông tin thu được sẽ mang lại giá trị đối với các doanh nghiệp và nhà tiếp thị trong việc tối ưu hóa chiến lược truyền thông xã hội của họ.

- Nội dung báo cáo

1. Thu thập dữ liệu:
2. Làm sạch và tiền xử lý dữ liệu:
 - (a) Làm sạch dữ liệu
 - (b) Tiền xử lý dữ liệu
3. Phân tích dữ liệu
5. Trực quan hóa dữ liệu (Visualization)
6. Kết luận

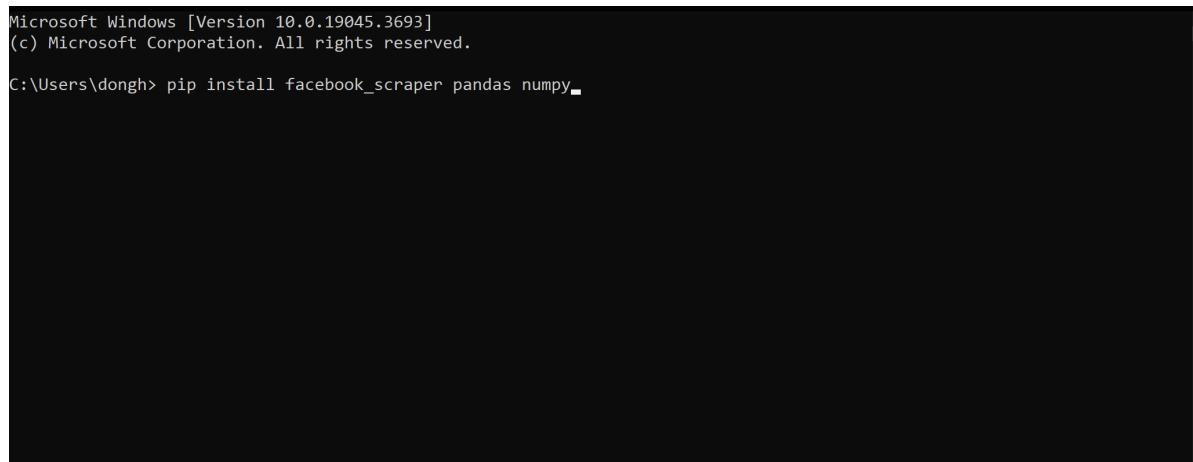
▼ Thu thập dữ liệu

Sau đây chúng ta hãy thu thập dữ liệu từ 1 trang fanpage trên facebook.

Nội dung thu thập : Thu thập các thông tin về các bài viết của fanpage đó bao gồm reaction, shares, coment, nội dung bài viết,

Để thu thập dữ liệu từ 1 fanpage ta có thể sử dụng thư viện `facebook_scraper` trong python. Đồng thời ta cài đặt luôn 2 thư viện là `pandas` và `numpy` để dễ dàng sử lí dữ liệu. Để có thể sử dụng được thư viện này trước tiên ta cần cài đặt thư viện bằng cmd với câu lệnh

```
pip install facebook_scraper pandas numpy
```



Microsoft Windows [Version 10.0.19045.3693]
(c) Microsoft Corporation. All rights reserved.
C:\Users\dongh> pip install facebook_scraper pandas numpy.

Sau khi cài đặt thư viện ta tiến hành các bước để thu thập dữ liệu

- **B1 : Khai báo thư viện trong python**

```
from facebook_scraper import get_posts  
import pandas as pd
```

```
import numpy as np
```

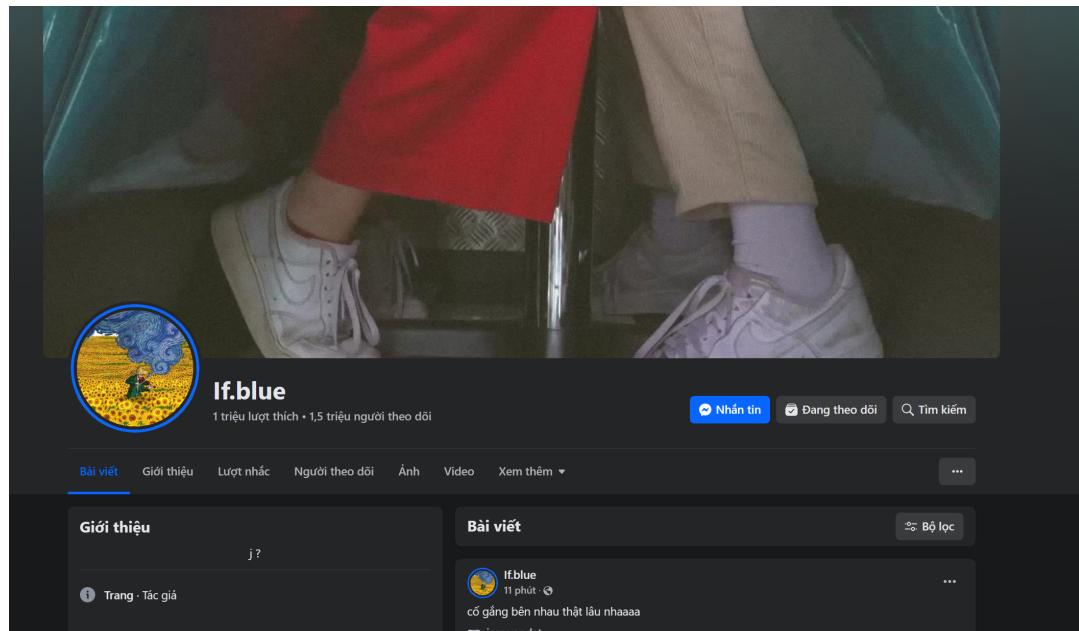
Những dòng này nhập hàm `get_posts` từ thư viện `facebook_scraper`. Chức năng này có thể được sử dụng để xóa các bài đăng từ trang Facebook. Nhập thư viện `pandas` và đặt cho nó bí danh `pd`, đồng thời nhập thư viện `numpy` và đặt cho nó bí danh `np`. Các thư viện này thường được sử dụng để thao tác và phân tích dữ liệu trong Python.

- **B2 : Chuẩn bị**

```
FANPAGE_LINK ="if.bluevn"  
COOKIE_PATH = "cookies.txt"  
  
PAGES_NUMBER = 30 # Number of pages to crawl  
post_list = []
```

Ta chuẩn bị 4 biến lần lượt là

1. `FANPAGE_LINK` Biến này lưu trữ đường dẫn của fanpage. Trong trường hợp này, đường dẫn là "if.bluevn".



2. `COOKIE_PATH` Biến này chứa đường dẫn tới tệp cookies. Cookies thường được sử dụng để duy trì trạng thái đăng nhập hoặc các thông tin liên

quan đến phiên làm việc.

Ở đây tôi sử dụng tiện ích trên google để lấy cookie của fanpage đó

Domain	Include subdomains	Path	Secure	Expiry	Name	Value
.google.de	TRUE	/	TRUE	1712648241	CONSENT	PENDING+616
.google.de	TRUE	/	TRUE	1693640241	AEC	ARSKqskX4MPF3ZYlQ...
.google.de	TRUE	/	TRUE	1712216243	SOCS	CAESHAgBEhJnd3Nf...
.google.de	TRUE	/	TRUE	1712274939	__Secure-ENID	10.SE=&K3KzuZpMF...
ogs.google.de	FALSE	/	TRUE	1680680262	OTZ	6929738_52_52_1239...

3. `PAGES_NUMBER` Biến này đặt số lượng trang cần lấy dữ liệu. Trong trường hợp này là 30 trang.
4. `post_list = []` List này là nơi lưu trữ các bài đăng hoặc thông tin thu thập từ trang web hoặc fanpage. Trong trường hợp này, ban đầu nó được khởi tạo rỗng, và được sử dụng để ghi thông tin từ việc crawl trang fanpage

- **B3 : Crawl dữ liệu**

```
for post in get_posts(FANPAGE_LINK,
                      options={"comments": True, "reactions": True,
                      extra_info=True, pages=PAGES_NUMBER, count=1000):
    print(post)
    post_list.append(post)
```

Sau khi crawl dữ liệu ta kiểm tra số lượng post đã thu thập được với hàm `len(post_list)`

ở đây là 300 bài post

```
len(post_list)
```

```
300
```

- **B4 : Lưu dữ liệu**

Để có thể lưu dữ liệu trước tiên ta cần chuyển dữ liệu dạng list sang DataFrame Để làm được điều này ta sử dụng thư viện pandas đã khai báo ở trên

```
df_full_post = pd.DataFrame(post_list)
```

Chuyển `post_list` sang DataFrame và đặt tên là `df_full_post`

Sau đó lưu DataFrame dưới dạng 1 file csv. Để làm được điều này ta cần tạo 1 file csv trước. Có 2 cách để tạo file là tạo thủ công hoặc sử dụng các đoạn code trong python. Sau khi tạo file .csv xong ta tiến hành lưu file:

```
df_full_post.to_csv("Data.csv")
```

Ở đây ta sử dụng file có tên `"Data.csv"` để lưu DataFrame đó.

▼ Làm sạch dữ liệu

Sau khi thu thập dữ liệu, dữ liệu thu thập thường chứa nhiều khuyết thiếu đến từ lỗi của công cụ thu thập hay bản thân bài đăng trên Fanpage (chẳng hạn bài đăng không chứa nội dung, chưa có bình luận, hạn chế quyền truy cập bởi token, . . .). Do vậy chúng ta cần làm sạch dữ liệu đã thu thập được.

Đầu tiên ta cần khai báo thư viện. Ở đây ta sử dụng thư viện `pandas` và `numpy` để đọc và làm sạch dữ liệu

```
import pandas as pd  
import numpy as np
```

Sau đó ta tiến hành đọc và làm sạch dữ liệu. Để làm sạch dữ liệu chúng ta cần đọc dữ liệu từ file `"Data.csv"` ta đã lưu trước đó.

```
df_full_post = pd.read_csv("Data.csv")
```

Đồng thời đặt DataFrame là `df_full_post` được lưu dưới dạng bảng.

```
df_full_post.shape  
  
(300, 53)
```

Ban đầu bảng `df_full_post` bao gồm có 300 hàng và 53 cột.

Ta tiến hành làm sạch dữ liệu:

1. Làm sạch với cột

- Xóa các cột có giá trị các hàng không có giá trị (đều là NULL):

```
df_full_post.dropna(axis=1, how='all', inplace=True)
```

- Xóa các cột có giá trị các hàng là trùng nhau :

```
df_full_post = df_full_post.T.drop_duplicates().T
```

- Xóa các cột không cần thiết:

```
del df_full_post['text']
del df_full_post['images']
del df_full_post['image_ids']
del df_full_post['images_description']
del df_full_post['fetched_time']
del df_full_post['images_lowquality']
```

Các cột không cần thiết bao gồm : `text` , `images` , `image_ids` , `images_description` , `fetched_time` , `images_lowquality` .Những cột này thường có các giá trị trùng với các cột khác hoặc các giá trị không dùng đến trong quá trình phân tích.

2. Làm sạch với hàng

- Xóa các hàng có giá trị NULL ở các cột quan trọng ảnh hưởng đến quá trình phân tích :**

```
df_full_post.dropna(subset=['reactions'], inplace=True)
df_full_post.dropna(subset=['reaction_count'], inplace=True)
```

```
df_full_post.dropna(subset=['shares'], inplace=True)
df_full_post.dropna(subset=['comments'], inplace=True)
df_full_post.dropna(subset=['comments_full'], inplace=True)
df_full_post.dropna(subset=['time'], inplace=True)
```

Nếu các hàng không tồn tại giá trị tại các cột trên ta tiến hành xóa hàng đó. Thật may mắn là sau khi đoạn code trên chạy thì số lượng các hàng không thay đổi

Các cột cần kiểm tra là : `reactions`, `reaction_count`, `shares`, `comments`,
`comments_full`, `time`

Sau khi làm sạch dữ liệu thì dữ liệu còn lại 300 hàng và 29 cột

▼ Tiền xử lí dữ liệu

Sau khi làm sạch dữ liệu. Chúng ta sẽ tiến hành lọc các dữ liệu cần thiết cho quá trình phân tích dữ liệu lúc sau

- **Xác định khoảng thời gian để phân tích dữ liệu**

```
df_full_post = df_full_post[(df_full_post['time'].dt.year == 2023) & (df_full_post['time'].dt.month.isin([8, 9, 10, 11]))]
```

Ở đây ta chọn các bài post trong khoảng thời gian từ tháng 8/2023 đến giữa tháng 11 năm 2023 để phân tích dữ liệu, để có cái nhìn trực quan nhất về fanpage trong giai đoạn vài tháng gần đây.

Đồng thời ta xóa các bài post được đăng vào ngày 19/11/2023. Vì đây là ngày mà ta thu thập dữ liệu. Vì thu thập ở giữa ngày nên dữ liệu về ngày này có thể không đầy đủ và đảm bảo.

```
row_to_drop = df_full_post[df_full_post["date_only"] == "2023-11-19"]
df_full_post.drop(row_to_drop, inplace=True)
```

- Thêm trường thông tin `"date_only"` và `"hour_only"`
 - `"date_only"`

```
df_full_post["time"] = pd.to_datetime(df_full_post['time'])
df_full_post["date_only"] = df_full_post["time"].dt.date
df_full_post["hour_only"] = df_full_post["time"].dt.hour
```

```
df_full_post['date_only'] = df_full_post['time'].dt.strftime('%Y-%m-%d')
```

Trường thông tin `"date_only"` trích xuất phần ngày và định dạng nó là 'tháng-ngày-năm' của cột `"time"`. Điều này có thể thực sự hữu ích để phân tích và trực quan hóa các hoạt động của fanpage theo ngày

- `"hour_only"`

```
df_full_post["time"] = pd.to_datetime(df_full_post['time'])
```

```
df_full_post['hour_only'] = df_full_post['time'].dt.strftime('%H:%M:%S')
```

Trường thông tin `"hour_only"` trích xuất phần giờ và định dạng nó là 'giờ - phút' của cột `"time"`. Điều này có thể thực sự hữu ích để phân tích và trực quan hóa các hoạt động của fanpage theo các giờ trong ngày.

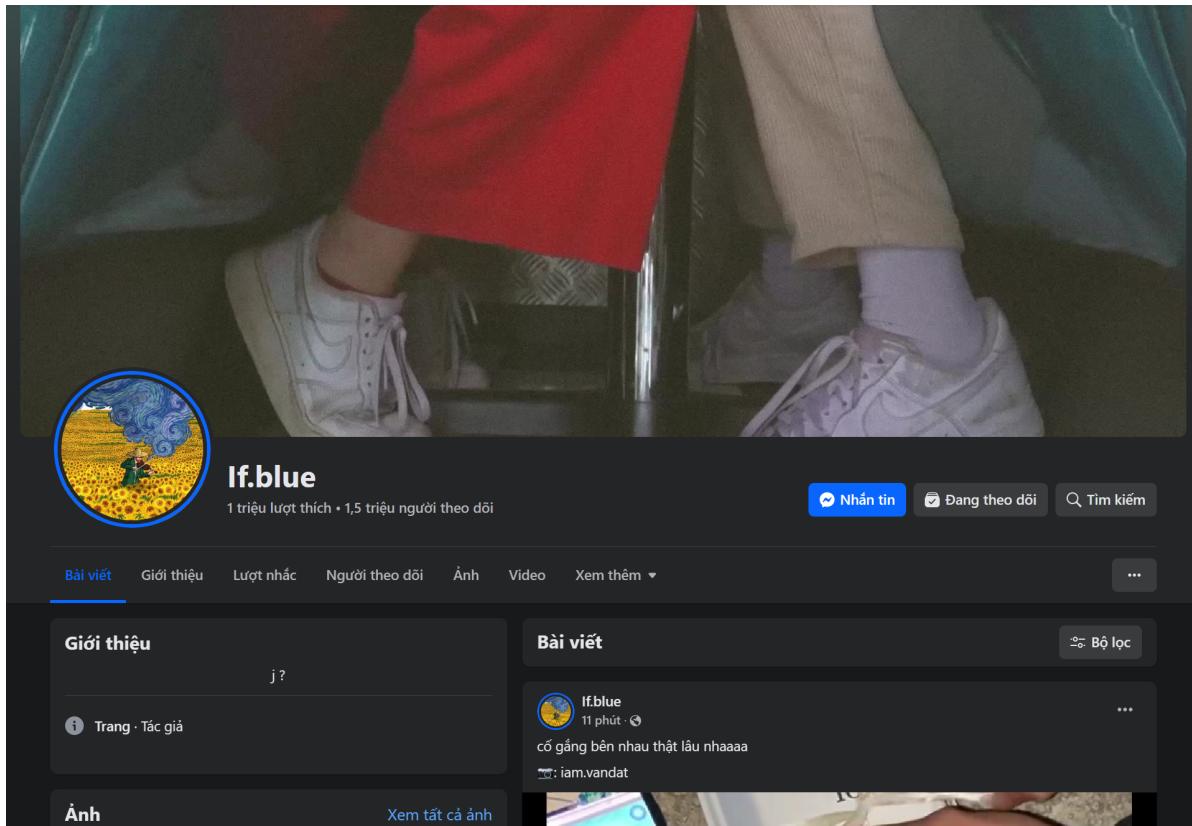
Sau khi xử lý, lọc dữ liệu để chuẩn bị cho quá trình phân tích. Ta tiến hành lưu Dữ liệu mới được làm sạch và xử lý vào 1 file khác để thuận tiện cho quá trình phân tích sau này.

```
df_full_post.to_csv("Data_clean.csv")
```

Ta lưu dữ liệu dưới dạng file.csv có tên là `Data_clean` theo các bước tương tự như lưu file `Data.csv` trước đó.

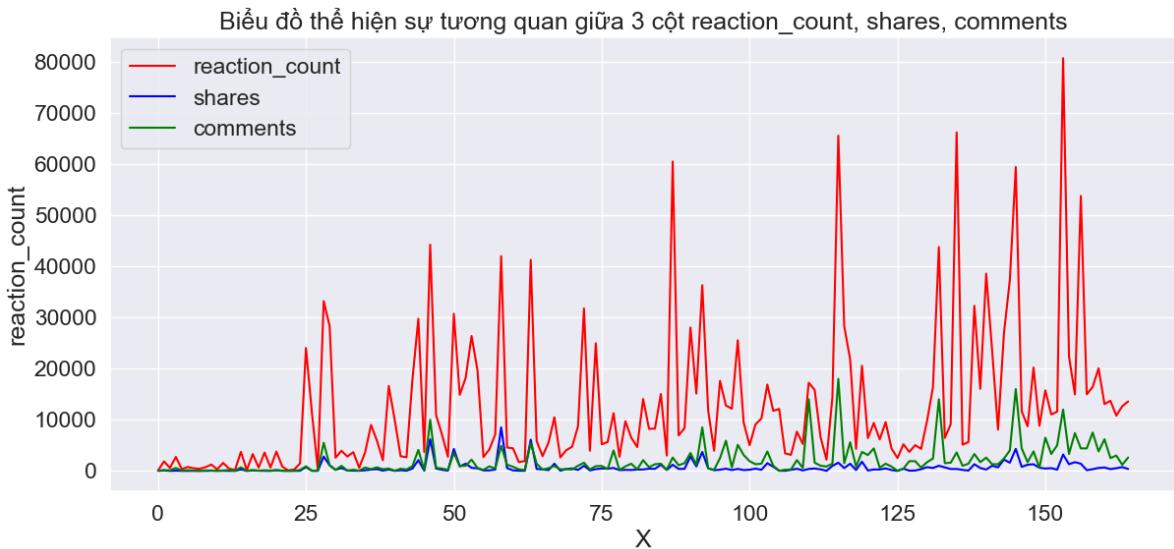
▼ Phân tích dữ liệu

If.blue là 1 fanpage có lượng người theo dõi khá lớn với 1 triệu lượt thích và 1,5 triệu lượt theo dõi. Các bài post của fanpage này chủ yếu về đề tài giải trí. Mang lại tiếng cười sau 1 ngày làm việc mệt mỏi.



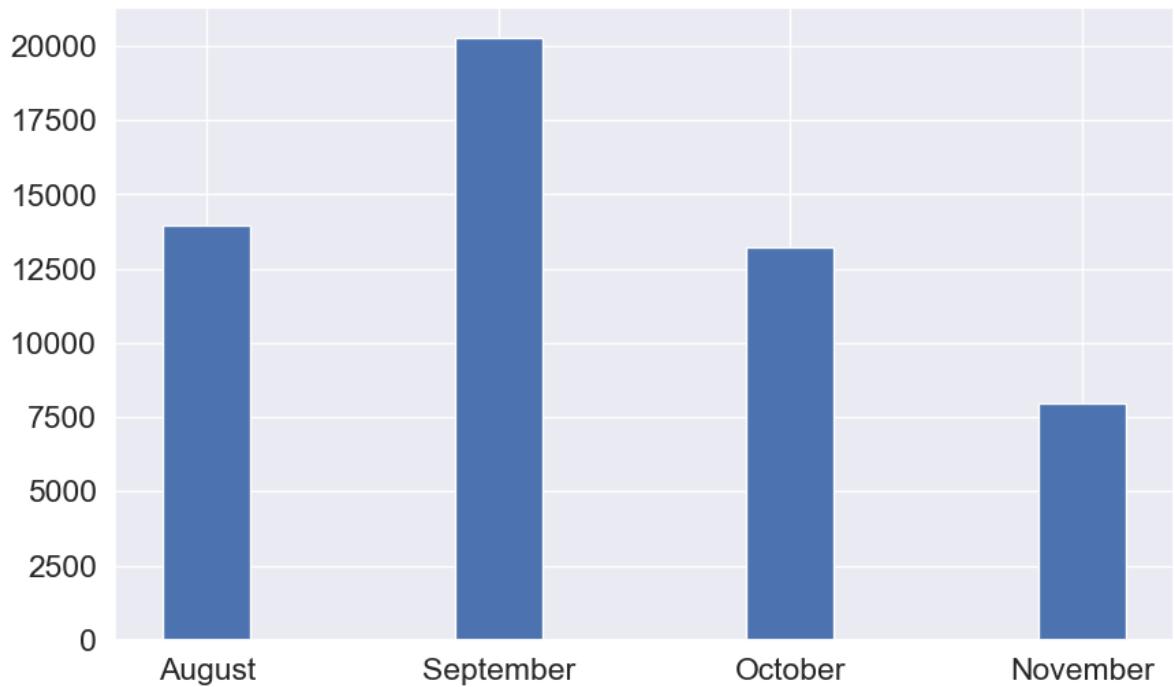
- **Lượt tương tác của fanpage**

Fanpage này có một lượng tương tác khá lớn .Trong báo cáo này chúng ta hãy xem xét hoạt động của fanpage trong khoảng thời gian từ tháng 8 đến nửa tháng 11 năm 2023. Dưới đây là biểu đồ thể hiện sự tương tác của các bài post.



Theo biểu đồ trên các bài post được sắp xếp theo thời gian giảm dần theo thứ tự từ 0 đến 162. Có thể thấy lượt tương tác của các bài post trong các giai đoạn gần đây đang có dấu hiệu giảm dần. Tuy cũng có những bài có lượt tương tác khá cao nhưng xét về tổng thể thì có dấu hiệu giảm. Dưới đây là biểu đồ cho thấy sự giảm tương tác qua các tháng gần đây:

Lượt tương tác trung bình qua các tháng



Ta có thể thấy rằng tương tác của bài viết, số lượt shares và số lượt comments có mối quan hệ tương đối tương đồng theo thời gian. Khi có sự tăng lên hoặc giảm xuống trong tương tác của bài viết, số lượt shares và số lượt comments cũng thay đổi theo cùng một hướng. Điều này cho thấy mức độ quan tâm và phản hồi của người dùng đối với các bài viết trên fanpage đều có xu hướng thay đổi đồng thời và cũng cho thấy người dùng thích reaction hơn là việc shares và comment khi mà lượt reaction luôn cao nhất sau đó đến comment rồi mới đến shares.

Trong khoảng thời gian trên .Về những cái nhất của các bài viết :

- Bài viết có lượt reaction cao nhất là bài thứ 151
('<https://facebook.com/if.bluevn/posts/625462799741047>') với 80748 lượt reaction được đăng vào ngày 09-09-2023 lúc 16:37:18.

	post_text	reaction_count	comments	shares	time	date_only	hour_only	video	image
151	ƯỚC ĐƯỢC NGỒI CHUNG VỚI CRUSH  : tnhy.zz	80748	12000	3200	2023-09-09 16:37:18	2023-09-09	16-37	https://scontent.fhph3-1.fna.fbcdn.net/v/t42.1...	NaN

- Bài viết có lượt shares cao nhất là bài thứ 56 :
('<https://facebook.com/if.bluevn/posts/655065576780769>') với 8500 lượt shares được đăng vào ngày 03-11-2023 lúc 02:00:51.

	post_text	reaction_count	comments	shares	time	date_only	hour_only	video	image
56	DETHUONG QUA TR LUNNNN  : bctuti05 So disgus...	42018	4900	8500	2023-11-03 02:00:51	2023-11-03	02-00	https://scontent.fhph3-1.fna.fbcdn.net/v/t42.1...	NaN

- Bài viết có lượt comment cao nhất là bài thứ 115 :
("'<https://facebook.com/if.bluevn/posts/639590194994974>'") với 18000 lượt comment được đăng vào ngày 05-10-2023 lúc 22:13:34.

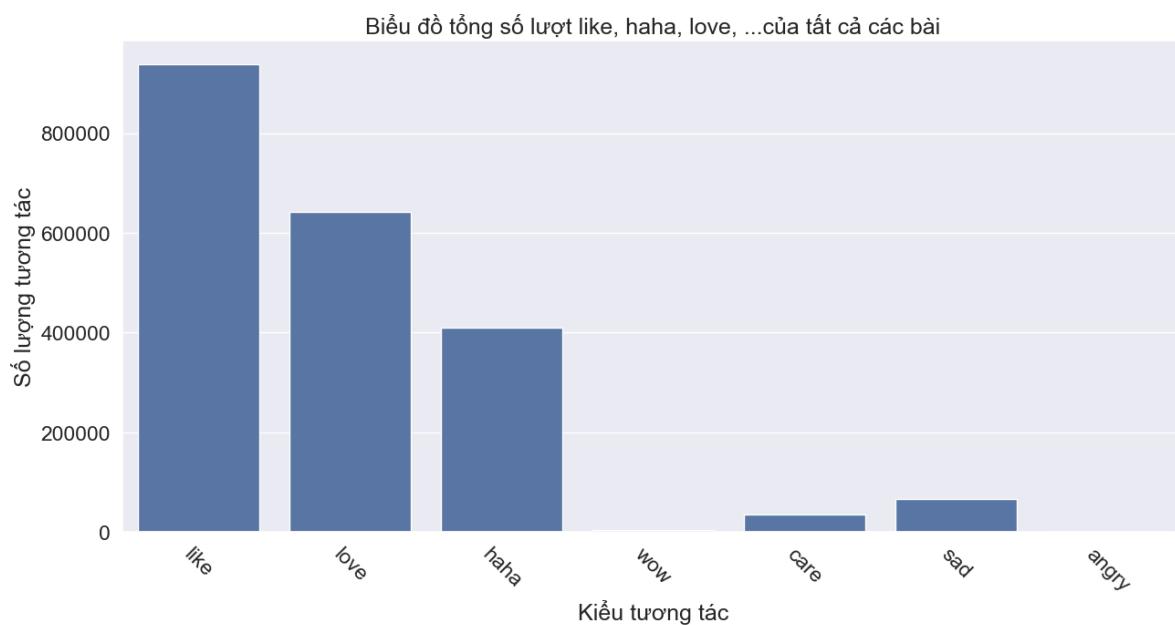
	post_text	reaction_count	comments	shares	time	date_only	hour_only	video	image
115	BỊ GHIỀN CÁI NÀY RỒI MÂY NÌ  : vomaxema8	65599	18000	1600	2023-10-05 22:13:34	2023-10-05	22-13	https://scontent.fhph3-1.fna.fbcdn.net/v/t42.1...	NaN

Điểm chung của 3 bài viết này đều được đăng vào khoảng chiều tối và đều có loại bài đăng đính kèm video.

Và dưới đây là dữ liệu thống kê về tổng các kiểu tương tác mà người dùng hay sử dụng:

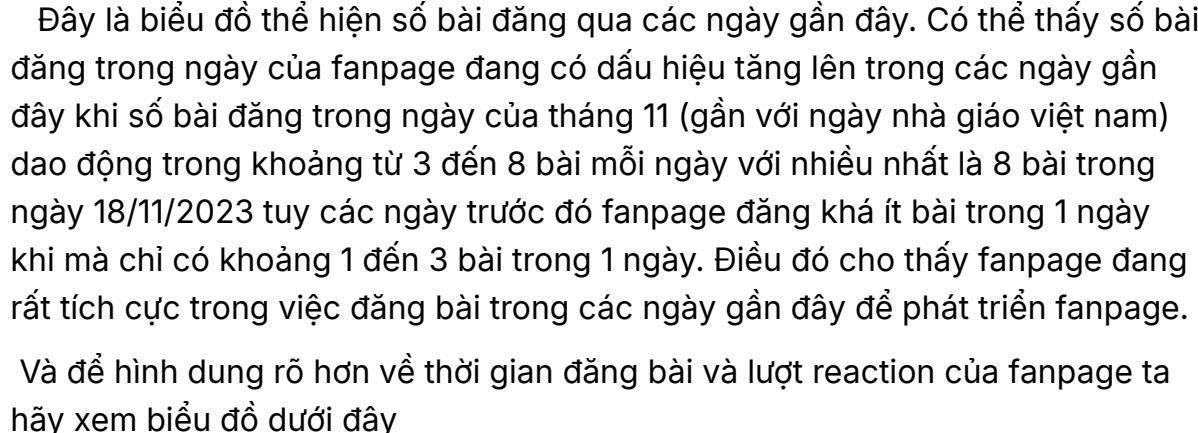
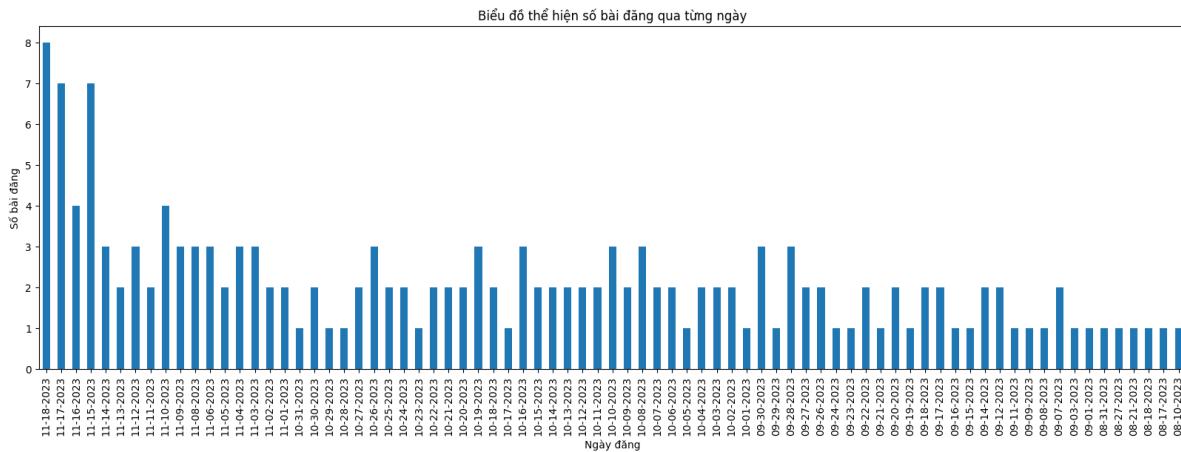
```
total_reactions
✓ 0.0s
{'like': 937735,
 'love': 641224,
 'haha': 409143,
 'wow': 3592,
 'care': 36406,
 'sad': 65709,
 'angry': 622}
```

Mọi người có vẻ thích 3 loại biểu tượng cảm xúc nhất là "like", "yêu thích", "haha" Chúng ta sẽ hình dung dễ hơn khi nhìn vào biểu đồ dưới đây.

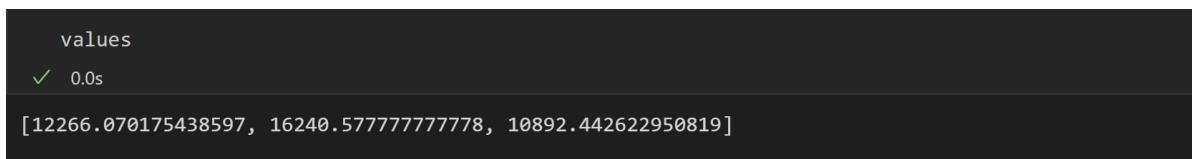


Có thể thấy 3 kiểu tương tác "like", "love", "haha" có số lượng áp đảo so với 4 kiểu còn lại. Đồng thời số lượt phản nô chỉ là 622. Có thể thấy đó là một tín hiệu tích cực về fanpage này khi mà lượt phản hồi về các bài post đang là khá tốt.

- **Hoạt động của fanpage**

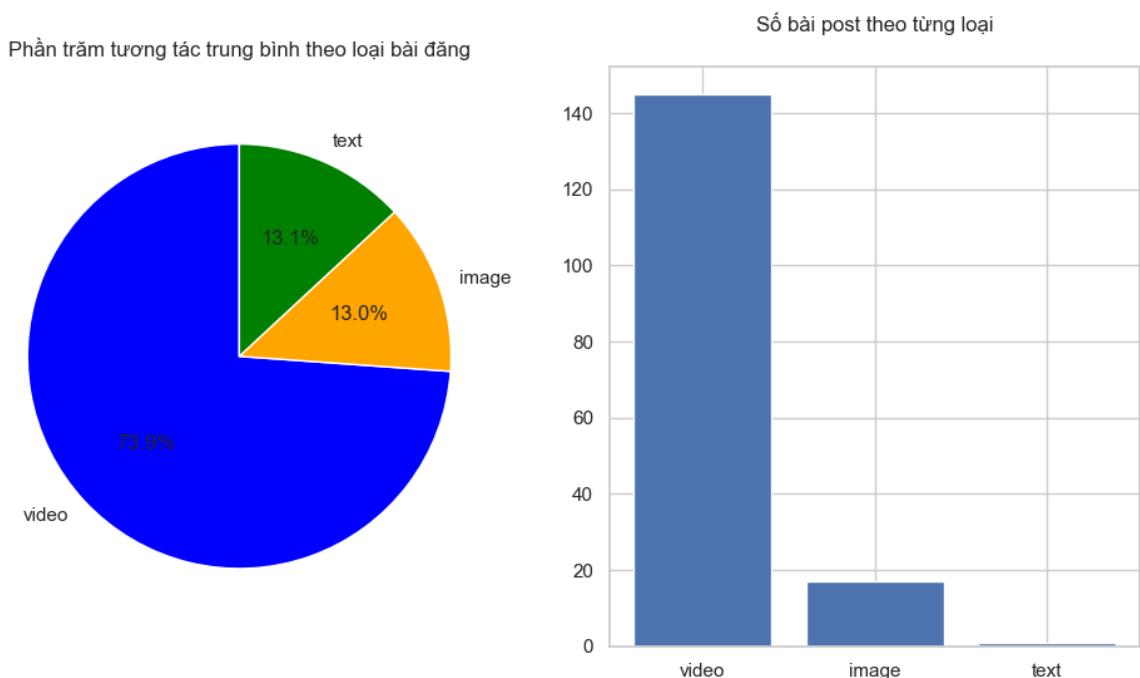


tương tác cũng rất lớn Với trung bình lượt reaction qua các buổi cũng rất cao cụ thể ta có Sáng : 12266,1; Chiều : 16240,6; Tối : 10892,4 Một lượng tương tác khá lớn đủ để thấy mức độ quan tâm của người dùng về trang fanpage này.



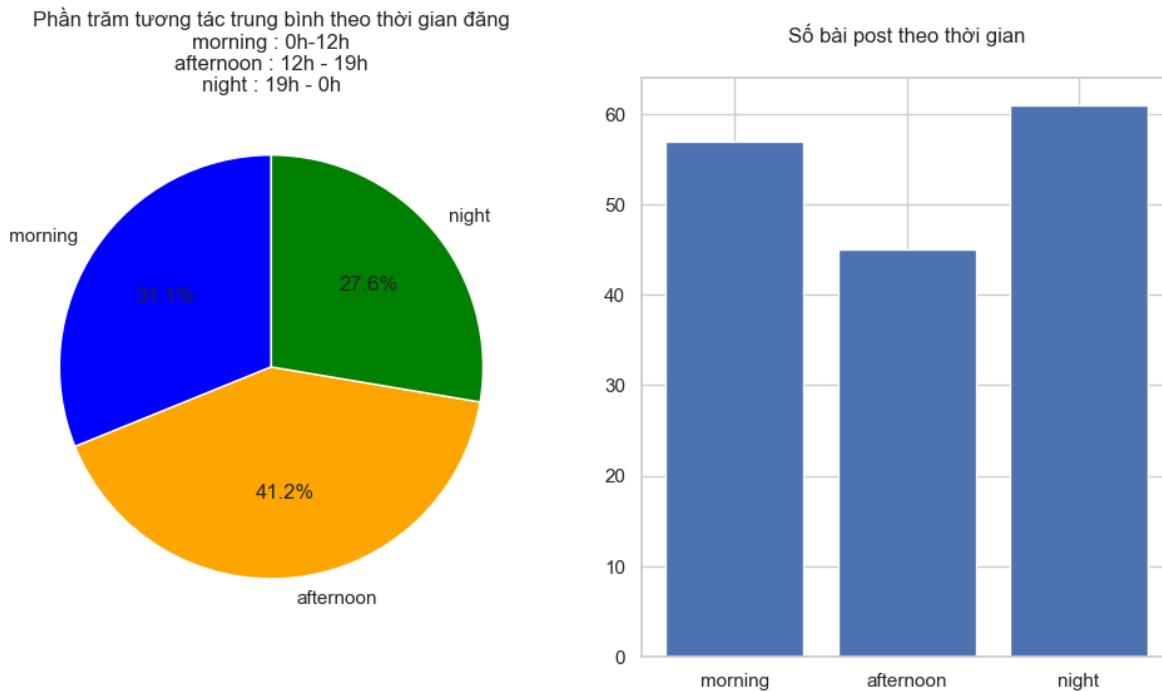
- **Ảnh hưởng của các yếu tố tới lượt tương tác**

- Biểu đồ thể hiện sự ảnh hưởng của loại bài đăng tới lượt reaction :



Nhìn vào biểu đồ trên ta có thể thấy các bài đăng có video thì lượt tương tác trung bình cao hơn so với các bài đăng dạng text hay có hình ảnh. Tuy bài đăng dạng text và image có lượt tương tác trung bình đều nhau tuy vậy số bài đăng dạng text fanpage đang ít hơn rất nhiều so với bài đăng có hình ảnh. Có thể thấy người xem rất thích xem những video và hình ảnh hơn là đọc những đoạn văn bản nhất là bài đăng có video đang được mọi người quan tâm nhiều hơn hẳn.

- Biểu đồ thể hiện sự ảnh hưởng của thời gian đăng bài tới lượt reaction :



Biểu đồ trên cho ta thấy sự ảnh hưởng của thời gian đăng bài tới lượt reaction. Như ta có thể thấy các bài đăng vào buổi chiều đang có lượt reaction cao nhất tuy vậy fanpage lại ít đăng vào buổi chiều nhất. Và đồng thời buổi tối có lượt reaction thấp nhất nhưng fanpage lại đăng nhiều nhất. Tuy sự chênh lệch giữa số bài đăng và số lượt reaction là không đáng kể và khá đồng đều với nhau nhưng qua đó ta cũng thấy 1 cái nhìn tổng quan về sự ảnh hưởng của thời gian đăng bài tới lượt reaction.

▼ Kết luận

Trong quá trình nghiên cứu và phân tích trang fanpage trên Facebook của chúng ta, chúng ta đã đặt ra và giải quyết nhiều vấn đề quan trọng nhằm hiểu rõ hơn về tương tác của cộng đồng và hiệu suất chung của trang. Dưới đây là một số điểm kết luận quan trọng:

- Tương Tác Người Dùng:** Thông qua việc phân tích dữ liệu, chúng ta đã nhận thấy xu hướng tương tác của người dùng có sự biến động theo thời

gian. Các bài đăng chất lượng cao và có tính tương tác cao đã thu hút sự chú ý đặc biệt.

2. **Chiến Lược Nội Dung:** Việc đánh giá chiến lược nội dung hiện tại đã đưa ra nhận định về sự đa dạng của nội dung. Các bài đăng có nội dung hay, sáng tạo, đồng thời được đính kèm video có lượt tương tác cao hơn cả
3. **Tăng Trưởng và Độ Phổ Biến:** Sự tăng trưởng trong lượng người theo dõi và tương tác có dấu hiệu giảm nhưng không nhiều. Điều này cho thấy chiến lược tiếp thị đang đem lại hiệu quả và sự chấp nhận tích cực từ cộng đồng.
4. **Chiến Lược Thời Gian:** Thời gian đăng bài cũng ảnh hưởng tới số lượng tương tác của các bài viết. Các bài viết được đăng vào buổi chiều có lượt tương tác cao hơn các buổi còn lại. Điều đó cho thấy sự ảnh hưởng của thời gian tới lượt tương tác là khá cao.

⇒ Qua báo cáo trên ta có cái nhìn khái quát về fanpage [if.blue](#) về hoạt động, sự ảnh hưởng của các yếu tố tới lượng tương tác của các bài viết. từ đó có những giải pháp để tăng tương tác cũng như quản lí nội dung và thời gian sao cho hợp lý nhất với người dùng trên nền tảng mạng xã hội hiện nay.