TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẠI HỌC QUỐC GIA HÀ NỘI

Trần Tiến Nam

BÁO CÁO DỰ ÁN CUỐI KÌ – LẬP TRÌNH XỬ LÝ DỮ LIỆU VỚI PYTHON

Ngành: Trí Tuệ Nhân Tạo

HÀ NỘI – 2023

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẠI HỌC QUỐC GIA HÀ NỘI

Trần Tiến Nam

BÁO CÁO DỰ ÁN CUỐI KÌ – LẬP TRÌNH XỬ LÝ DỮ LIỆU VỚI PYTHON

Ngành: Trí Tuệ Nhân Tạo

Giảng Viên: Đặng Trần Bình

Đỗ Hải Sơn

Nguyễn Văn Phi

HÀ NỘI – 2023

TÓM TẮT

Phân tích tương tác và nội dung của một (nhiều) tài khoản Facebook. Từ dữ liệu của một Fanpage trên Facebook, chúng ta có thể nhận thấy được tất cả các yếu tố của trang Fanpage như nhóm người dùng, chủ đề chung của Fanpage, …Khi biết được nội dung mà người theo dõi quan tâm hay những mốc thời gian hay content có nhiều lượt tương tác, chủ Fanpage có thể lựa chọn mốc thời gian đăng bài, số lượng bài đăng mỗi ngày hay chủ đề của bài đăng để tăng số lượng lượt tương tác của bài đăng.

1. Thu thập dữ liệu

Sử dụng công cụ **facebook-scraper** để thu thập dữ liệu từ Fanpage. Các thông tin cần có để phân tích gồm có nội dung bài viết (post_text), thời gian đăng bài (time/timestamp), số lượt chia sẻ (shares), bình luận (comments), các loại reaction tương ứng như like, love, care, …(reactions) và tổng các lượt reaction (reaction_count)

Do khi sử dụng công cụ **facebook-scarper** để thu thập nên có một số ô dữ liệu bị miss, do vậy ta cần thu thập nhiều hơn dữ liệu cần thiết một chút để có thể có đủ dữ liệu để phân tích.

Sau khi thu thập dữ liệu thì ta có thể lưu dữ liệu dưới dạng một số file như **.csv**, **.xlsx**, …. Do trong dữ liệu có một trường đặc biệt chứa tất cả các bình luận nên ta phải lưu thêm một file dưới dạng **.npy** để có thể thuận tiện cho việc xử lý.

Link chứa code phân tích: Git Hub.

2. Làm sạch và tiền xử lý dữ liệu

1. Làm sạch dữ liệu

Như ở phần 1-Thu thập dữ liệu, do có một số dữ liệu bị miss có nên ta sẽ phải xử lý sao cho phù hợp nội dung phân tích mong muốn. Ví dụ như khi thu thập, có một số bài viết không lấy được bình luận, không lấy được số lượt reaction hay lượt chia sẻ vì vậy ta có thể xóa hàng dữ liệu đó, xử lý dữ liệu đó sao cho phù hợp hoặc có thể thu thập lại những dữ liệu đã bị miss đó.

Khi thu thập sẽ có một số trường không cần thiết cho việc phân tích dữ liệu, vì vậy ta sẽ xóa bớt đi các trường không liên quan để bộ dữ liệu được gọn gàng hơn

Trong dữ liệu thu thập có 217 hàng và 51 trường dữ liệu, trong đó có rất nhiều trường không có dữ liệu hoặc dữ liệu bị null.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 217 entries, 0 to 216
Data columns (total 51 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   post_id                         217 non-null    int64
 1   text                            217 non-null    object
 2   post_text                       217 non-null    object
 3   shared_text                     2 non-null      object
 4   original_text                   18 non-null     object
 5   time                            217 non-null    object
 6   timestamp                       217 non-null    int64
 7   image                           202 non-null    object
 8   image_lowquality                216 non-null    object
 9   images                          217 non-null    object
 10  images_description              217 non-null    object
 11  images_lowquality               217 non-null    object
 12  images_lowquality_description   217 non-null    object
 13  video                           7 non-null      object
 14  video_duration_seconds          0 non-null      float64
 15  video_height                    0 non-null      float64
 16  video_id                        7 non-null      float64
 17  video_quality                   0 non-null      float64
 18  video_size_MB                   0 non-null      float64
 19  video_thumbnail                 7 non-null      object
...
 49  was_live                        217 non-null    bool
 50  fetched_time                    174 non-null    object
dtypes: bool(3), float64(16), int64(8), object(24)
memory usage: 82.1+ KB
```

Sau khi loại bỏ các cột không cần thiết và các dữ liệu bị null thì dữ liệu mới có chứa 174 hàng và 13 trường dữ liệu.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 174 entries, 0 to 173
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   post_id         174 non-null    int64
 1   text            174 non-null    object
 2   post_text       174 non-null    object
 3   time            174 non-null    object
 4   timestamp       174 non-null    int64
 5   likes           174 non-null    int64
 6   comments        174 non-null    int64
 7   shares          174 non-null    int64
 8   comments_full   174 non-null    object
 9   reactors        174 non-null    object
 10  reactions       174 non-null    object
 11  reaction_count  174 non-null    int64
 12  fetched_time    174 non-null    object
dtypes: int64(6), object(7)
memory usage: 17.8+ KB
```

2. Tiền xử lý dữ liệu

Ta tổ chức lại các bộ dữ liệu theo mong muốn để phân tích.

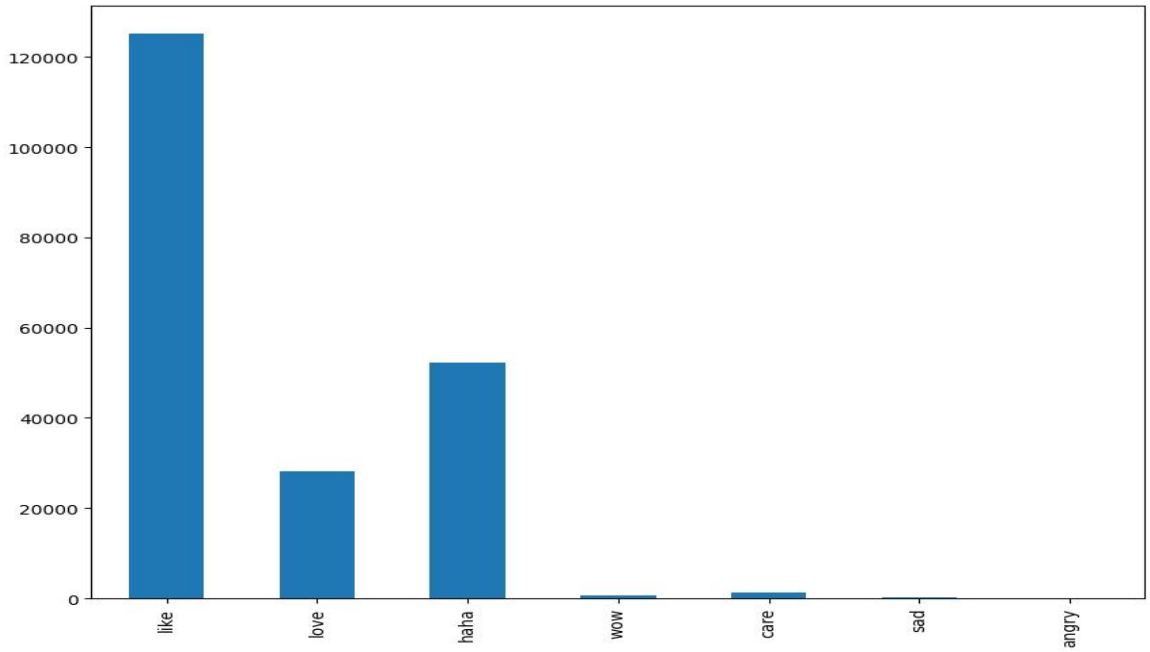
3. Phân tích dữ liệu

1. Đâu là bài viết có lượt tương tác lớn nhất trong bộ dữ liệu?

Từ dữ liệu thu thập ta có thể tìm được bài viết có lượt tương tác lớn nhất qua số liệu của reaction_count.

```
post_text        Người Tày chất 🫶\n\nTay people are cool
time                                 2023-09-22 00:14:28
comments                                            3500
shares                                               276
reaction_count                                    207645
Name: 157, dtype: object
```

Chúng ta cũng có thể thấy được số lượng của các loại reaction trong bài viết có lượt tương tác lớn nhất theo biểu đồ cột bên dưới.



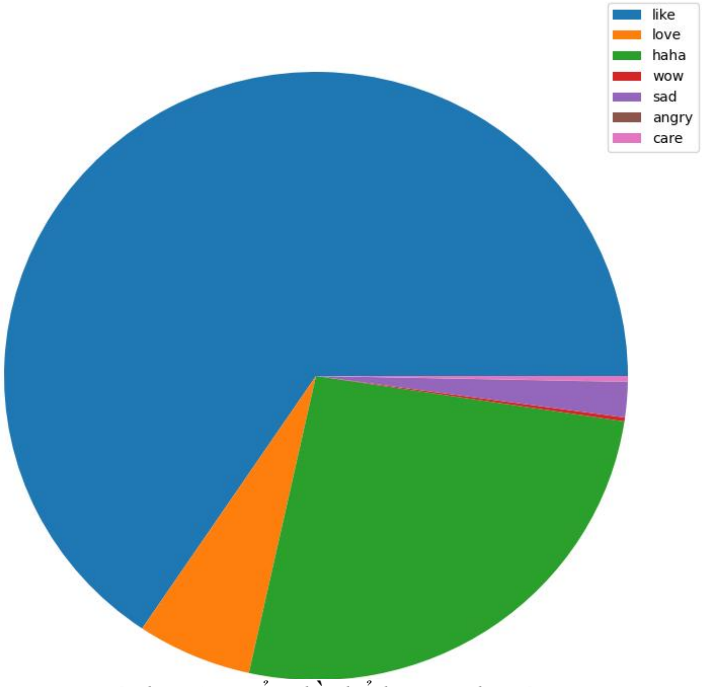
=> Vậy chủ yếu lượt tương tác đến từ lượt like, love, haha còn các loại reaction khác như wow, care, sad, angry rất thấp hay gần như không có.

2. Số lượt tương tác trong các bài đăng thay đổi như thế nào?

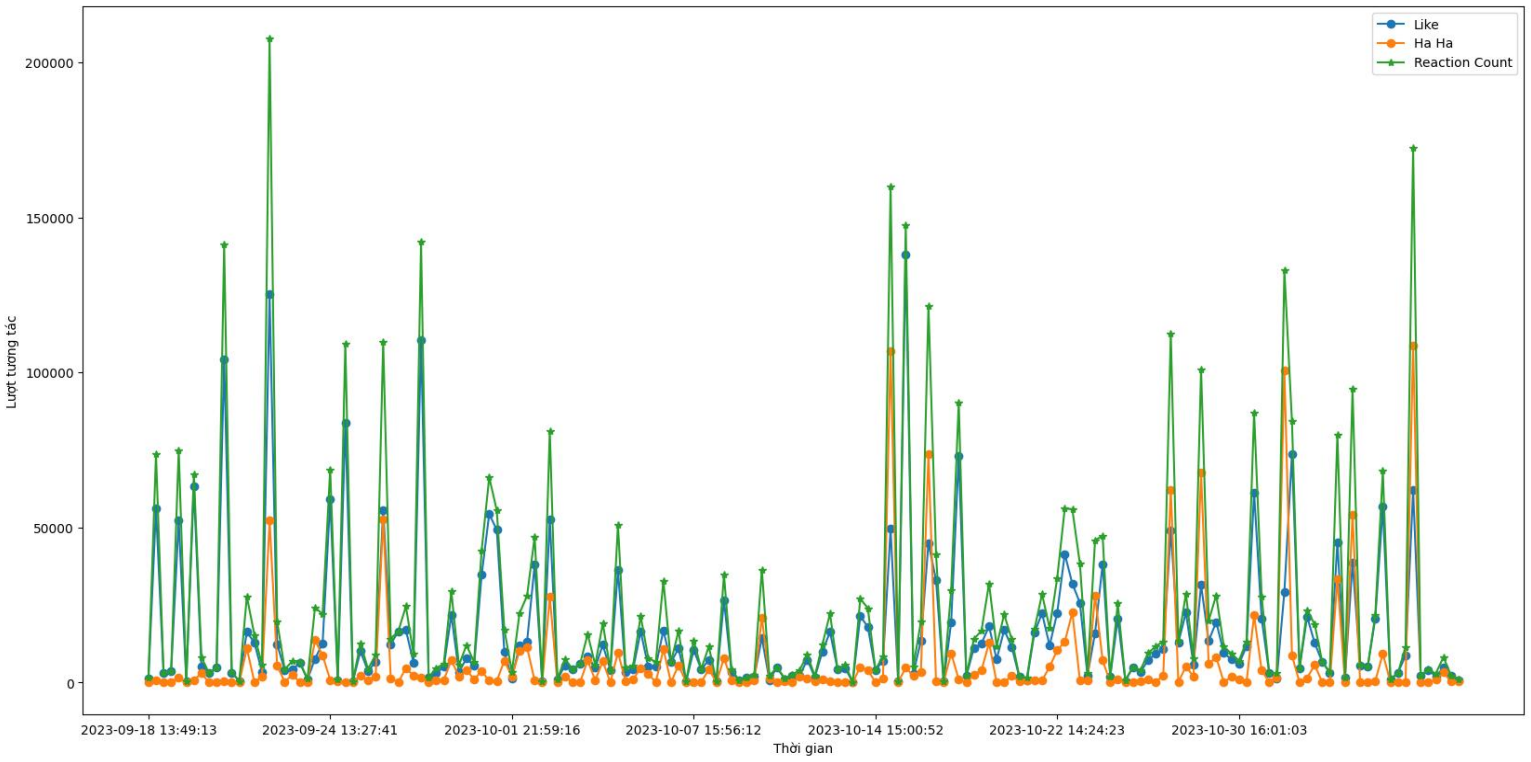
Qua *Hình 2.1* ta có thể thấy được tổng quan về tỷ lệ các loại reaction của Fanpage. Chủ yếu lượt tương tác là lượt like và haha.

=> Các bài viết của Fanpage không đa dạng về các lượt reaction (đa dạng cảm xúc) mà chủ yếu là lượt like cho thấy các bài viết hầu như là chứa thông tin, tin tức; lượt haha cũng chiếm khá nhiều cho thấy được các bài viết cũng mang tính giải trí cao.

.



Hình 2.1: Biểu đồ thể hiện tỷ lệ các loại reaction trong tất cả các bài viết



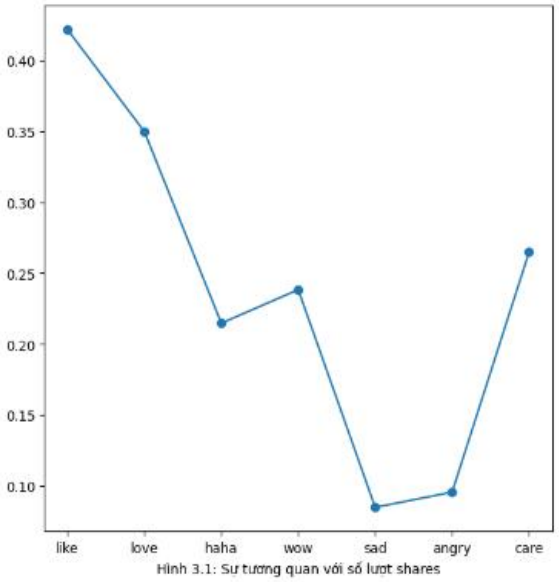
Hình 2.2 là biểu đồ cho thấy được sự phân bố loại reaction đặc trưng và tổng lượt reaction biến đổi theo thời gian

=> Lượt tương tác của các bài viết thay đổi liên tục. Có bài viết có lượng tương tác rất lớn nhưng có bài viết lại có lượng tương tác rất thấp.

Ngoài ra, ta có thể nhận ra do đây là Fanpage có các bài đăng hầu như mang thông tin nên khả năng bài viết có lượt tương tác cao nhất là một sự kiện đang nổi trội hay chứa thông tin thú vị, hài hước.

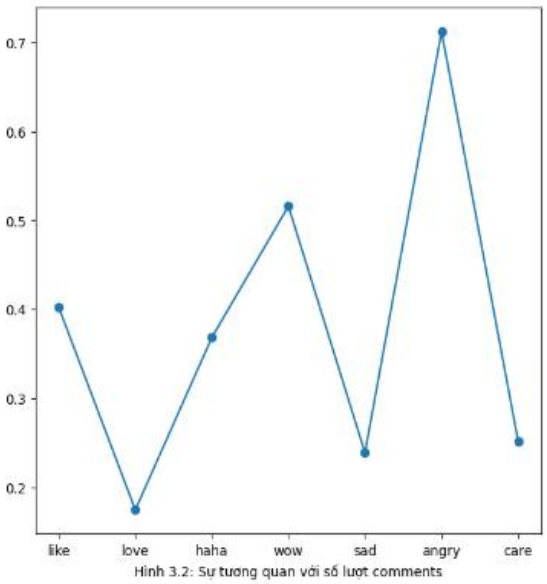
3. Sự tương quan giữa số lương reactions với các trường khác như số lượt bình luận, số lượt chia sẻ, độ dài bài viết, …?

Với biểu đồ về sự tương quan giữa các lượt reaction với số lượt chia sẻ: Tất cả các reaction đều có tương quan dương với số lượt chia sẻ nhưng chủ yếu vẫn là lượt like, love. Tức là bài viết có nhiều lượt reaction này thì sẽ có nhiều lượt chia sẻ.



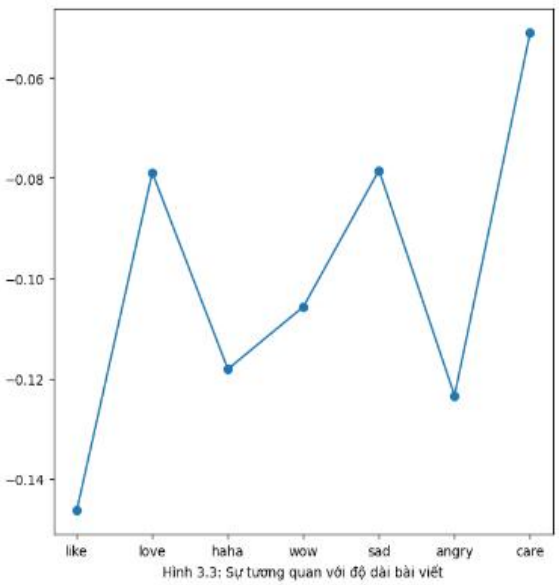
Hình 3.1: Sự tương quan với số lượt shares

Với biểu đồ về sự tương quan giữa các lượt reaction với số lượt bình luận: Tất cả các lượt reaction đều có tương quan dương so với số lượng bình luận. Ảnh hưởng nhất đó là lượt phẫn nộ, khi bài viết có nhiều lượt phẫn nộ thì bài viết đó sẽ có nhiều lượt bình luận.



Hình 3.2: Sự tương quan với số lượt comments

Với biểu đồ về sự tương quan giữa các lượt reaction với độ dài bài viết: Ngược lại với hai tương quan trên thì với độ dài bài viết, các lượt reaction đều có tương quan âm nhưng tương quan âm thấp nhất đó là lượt like. Chứng tỏ bài viết có càng nhiều lượt like thì độ dài bài viết càng ngắn.



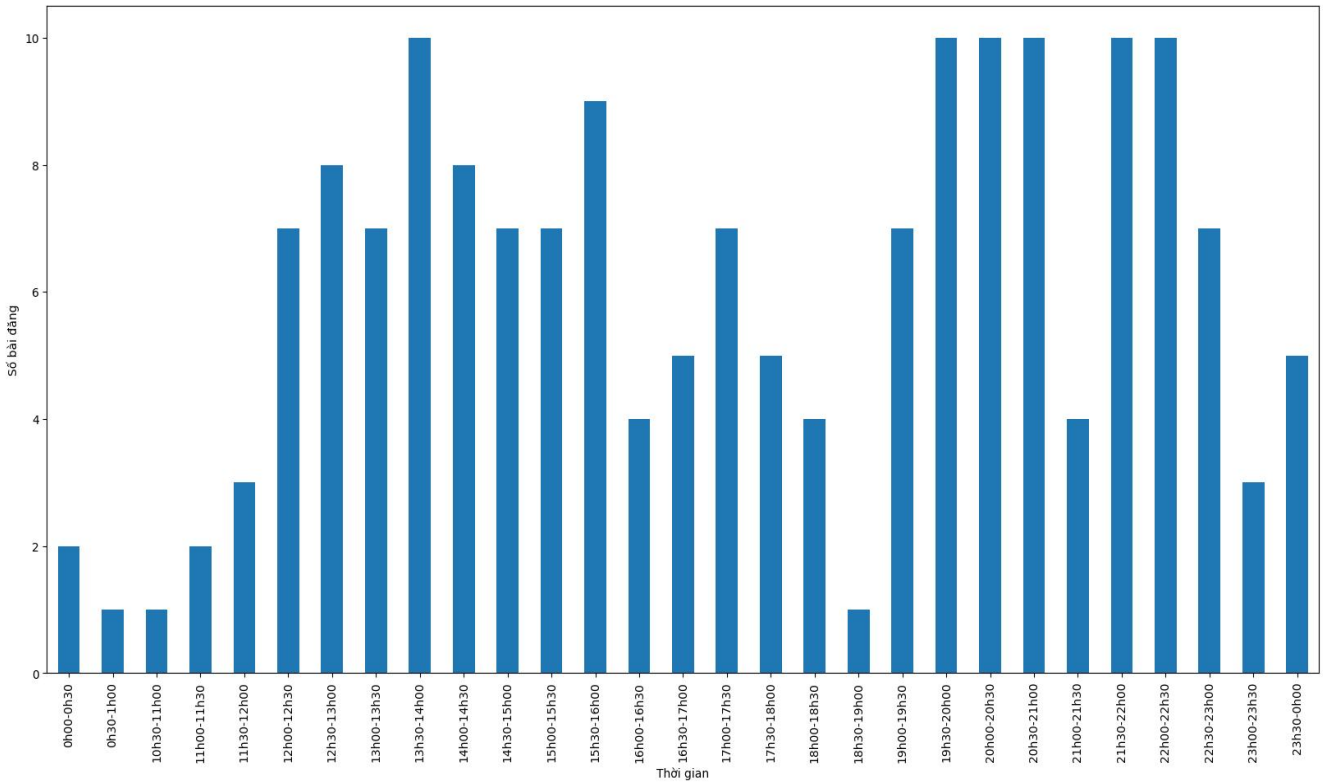
Hình 3.3: Sự tương quan với độ dài bài viết

=> Ý nghĩa:

- Sự tương quan với số lượt chia sẻ:
  - Lượt like, love là bài viết thể hiện sự đồng tình, yêu thích. Những bài viết nhận được nhiều lượt like hoặc love thì có thể là những bài viết có nội dung thú vị, hữu ích hoặc có liên quan đến sở thích của người đọc nên họ muốn chia sẻ bài viết với người khác.
  - Lượt sad, angry là cảm xúc tiêu cực cho thấy đó là bài viết gây cảm xúc tiêu cực hay gây tranh cãi khiến người đọc không muốn chia sẻ với mọi người.
- Sự tương quan với số lượt chia sẻ:
  - Lượt angry là cảm xúc tiêu cực cho thấy bài viết gây nhiều tranh cãi, nhiều luồng ý kiến nên mọi người muốn bình luận để bày tỏ ý kiến và tranh luận về vấn đề của bài viết.
- Sự tương quan với độ dài bài viết:
  - Khi độ dài bài viết ngắn đi thì các lượt reaction lại tăng. Đây cũng có thể là do thông tin người viết muốn truyền đạt lại nằm trong hình ảnh hay video chứ không nằm trong đoạn text bài viết.

4. Đâu là thời gian Fanpage thường đăng bài, đâu là khoảng thời gian đăng bài được nhiều lượt tương tác nhất, số bài viết mà Fanpage đăng mỗi ngày, một ngày Fanpage đăng tối đa bao nhiêu bài đăng?

**** Đâu là thời gian Fanpage thường đăng bài.***



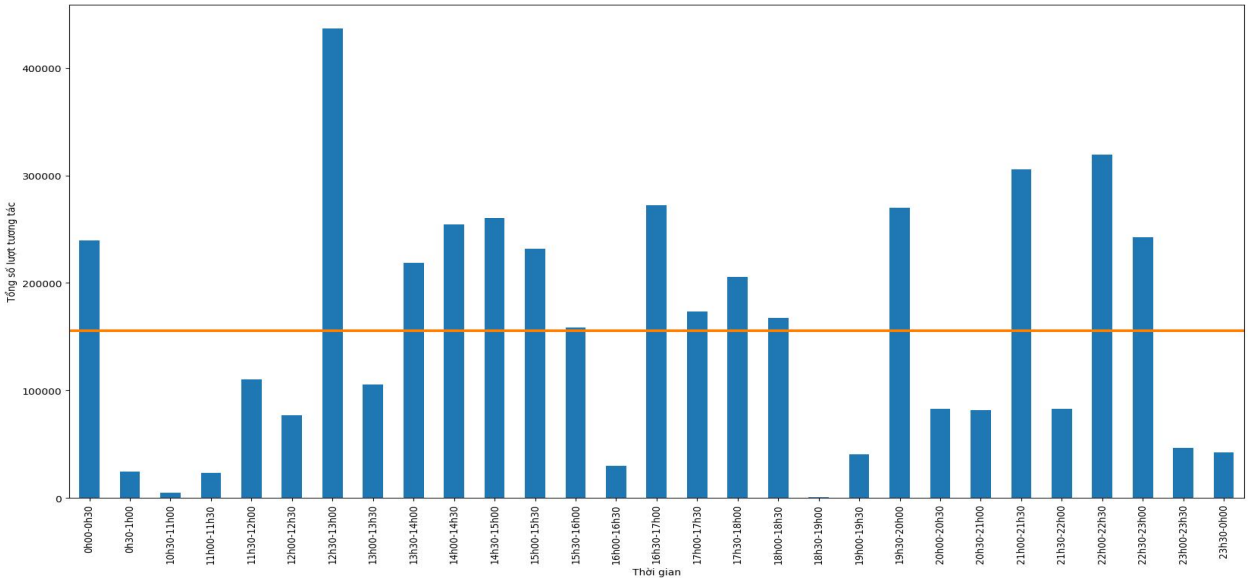
Hình 4.1: Biểu đồ về sự tương quan giữa thời gian đăng bài với số lượng bài đăng

Từ biểu đồ ta thấy được số lượng bài viết được đăng trong từng khoảng thời gian trong ngày. Ở biểu đồ thì khỏng thời gian nào không có thì chứng tỏ không có bài đăng nào được đăng trong khoảng thời gian đó.

Nhìn vào biểu đồ ta nhận ra ngay được các khoảng thời gian Fanpage thường đăng bài là: '13h30-14h00', '19h30-20h00', '20h00-20h30', '20h30-21h00', '21h30-22h00', '22h00-22h30' hay là vào các khoảng đầu giờ chiều(13h30-14h00) và vào buổi tối muộn(19h30-21h00, 21h30-22h30).

Những khoảng thời gian Fanpage thường đăng bài này đều là thời gian rảnh hoặc thời gian giải lao. Đặc biệt là buổi tối muộn là khoảng thời gian mọi người hay giải trí.

**** Đâu là khoảng thời gian đăng bài được nhiều lượt tương tác nhất***

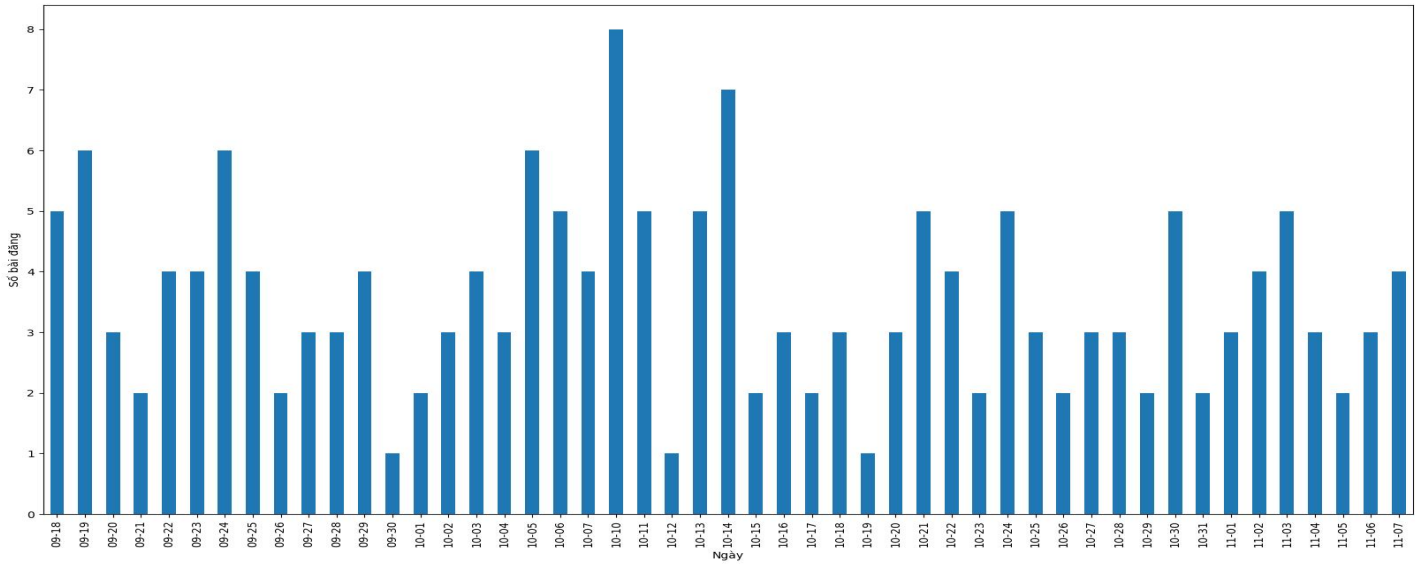


Hình 4.2: Biểu đồ về sự tương quan giữa thời gian đăng bài với số lượt tương tác

Dựa vào biểu đồ trên ta thấy được những số lượt tương tác theo từng khoảng thời gian trong ngày. Khoảng thời gian không có trong biểu đồ là khoảng thời gian Fanpage không đăng bài. Đường màu cam là đường thẳng chỉ giá trị trung bình của tổng lượt tương tác.

Khoảng thời gian có lượt tương tác lớn nhất là 12h30-13h00. Dựa biểu đồ trên thì Fanpage có thể biết đăng bài vào các khoảng thời gian nào giúp đạt được nhiều lượt tương tác nhất

**** Số bài viết Fanpage đăng mỗi ngày***



Hình 4.3: Biểu đồ về số lương bài đăng theo từng ngày (T9-T11 năm 2023)

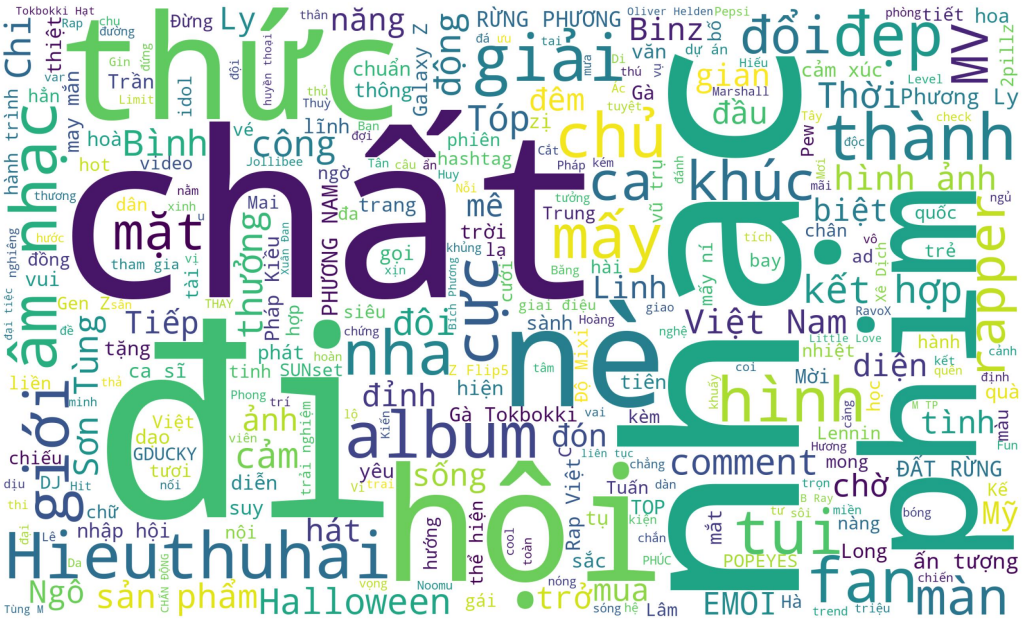
Nhìn biểu đồ một cách tổng quan thì Fanpage rất thường xuyên đăng bài.

Số bài đăng trong vòng một ngày nhiều nhất là 8 bài. Trong biểu đồ có hai ngày bị thiếu đó là ngày 8/10 và 9/10, hai ngày này là hai này Fanpage không đăng bài.

5. Đâu là từ khóa xuất hiện nhiều nhất trong các bài đăng đã thu thập hay đâu là từ khóa xuất hiện nhiều nhất trong các bình luận đã thu thập?

**** Từ khóa xuất hiện nhiều nhất trong các bài đăng của Fanpage.***

Với ảnh trên ta thấy có một số từ là các từ không có nghĩa trong tiếng việt, hay còn gọi là các stopword trong tiếng việt. Do đó ta phải loại bỏ các stopword để lấy các từ chuẩn xác.



Sau khi loại bỏ các stopword thì các từ đã có có tính chủ đề hơn. Đặc biệt chủ đề chủ yếu về âm nhạc, phim ảnh, rapper hay các thành viên trong showbiz.

Từ khóa xuất hiện nhiều trong các bình luận.

Cũng như trên ta cũng áp dụng loại bỏ các từ stopword thì ta sẽ có hình sau:

Sau khi loại bỏ các stopword thì chúng ta thấy hầu như các từ đều là tên hay họ mà mọi người tag trong bình luận. Qua hình ta cũng thấy được họ Nguyễn là chủ yếu sau đó là họ Trần.

6. Thử dự đoán số lượng bài đăng trung bình trong một ngày của page là bao nhiêu, số lượt tương tác. Từ đó so sánh với dữ liệu thực tế khi thu thập.

Theo dữ liệu ban đầu trong vòng 49 ngày thì ta thấy được giá trị trung bình của các loại reaction, số bài viết một ngày hay số lượt tương tác như sau:

```
Số bài viết 1 ngày: 3.5510204081632653
comments             685.097701
shares                88.660920
reaction_count     25926.994253
like               17005.201149
love                1545.385057
haha                6757.597701
wow                   53.482759
care                  77.166667
sad                  483.925287
angry                  4.235632
dtype: float64
```

Ta có thể thấy số bài đăng trung bình trong vòng 1 ngày là 3,55 bài. Với các Fanpage thì trung bình 1 ngày đăng 3,55 bài là bình thường. Vì vậy trong vòng 10 ngày tiếp theo khả năng Fanpage cũng sẽ đăng bài khá thường xuyên và số bài viết trung bình 1 ngày cũng nằm trong khoảng 3-4.

Vì các bài viết có các loại reaction khá thất thường nên dữ liệu dự đoán chênh lệch sẽ khá cao. Với dữ liệu này ta có thể dự đoán trong 10 ngày tới thì các dữ liệu sẽ có giá trị lần lượt sau: bình luận 270, chia sẻ 35, reaction_count 10400, like 6800, love 600, haha 2700, wow 21, care 31, sad 96, angry 1-2.

Ta có thể so sánh tỷ lệ dự đoán so với thực tế:

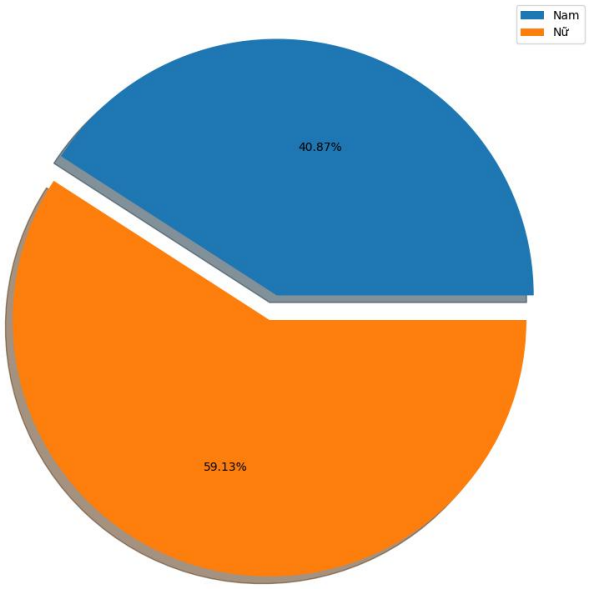
```
Số bài viết 1 ngày: 92.93%
comments: 97.67%
shares: 115.15%
reaction_count: 170.42%
like: 171.20%
love: 105.90%
haha: 196.95%
wow: 192.50%
care: 125.22%
sad: 49.34%
angry: 153.54%
```

```
Số bài viết 1 ngày: 3.3
comments             263.696970
shares                40.303030
reaction_count     17723.363636
like               11641.363636
love                 635.393939
haha                5317.696970
wow                   40.424242
care                  38.818182
sad                   47.363636
angry                  2.303030
dtype: float64
```

Các giá trị >100% chứng tỏ dữ liệu dự đoán vượt quá so với dữ liệu thực tế còn giá trị <100% chứng tỏ dữ liệu dự đoán ít hơn dữ liệu thực tế.

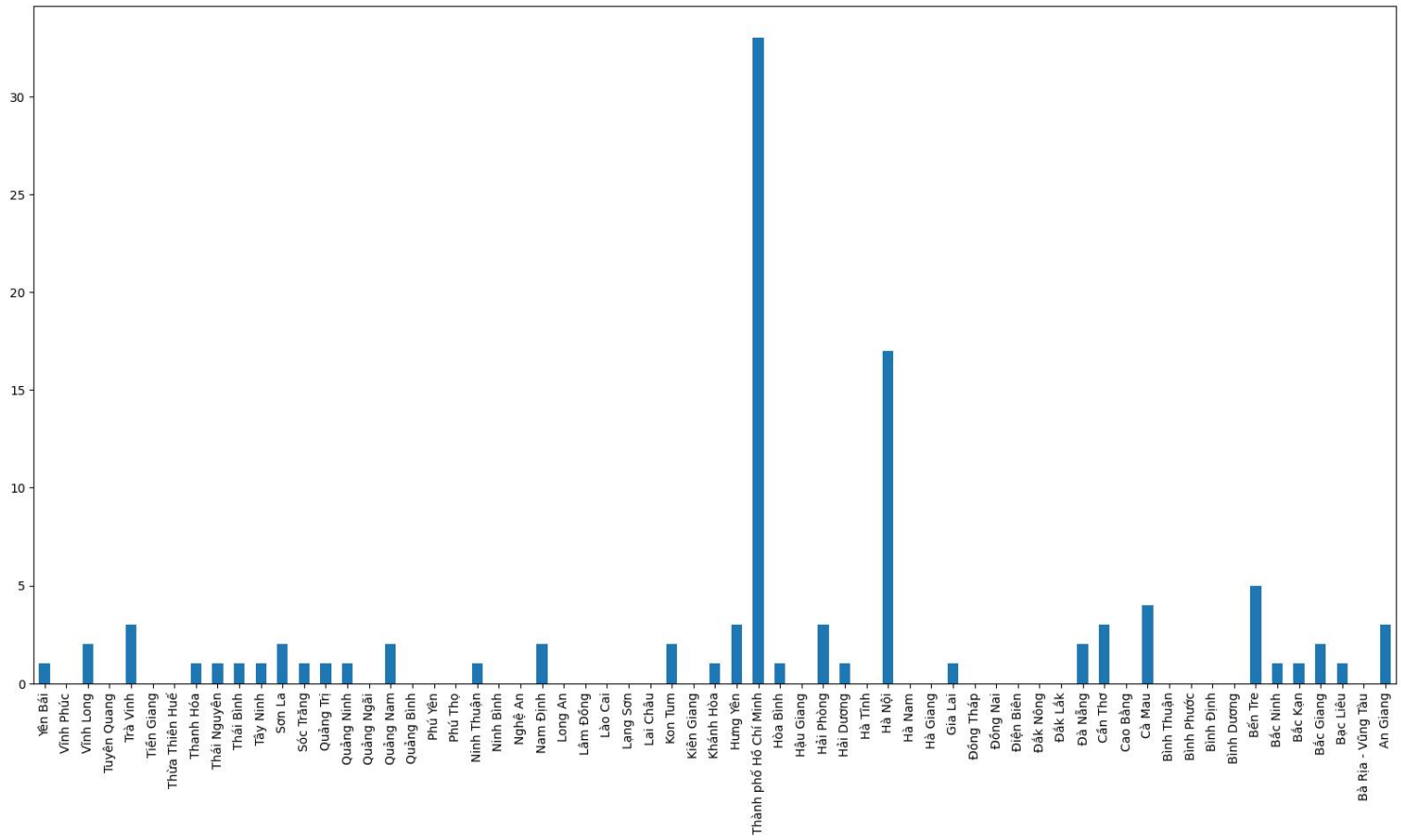
7. Những người tương tác với Fanpage có tỷ lệ giới tính như thế nào?

Với biểu đồ bên ta thấy được tỷ lệ nam nữ khá rõ rệt. Với người bình luận ta thấy tỷ lệ nữ giới chiếm phần hơn so với nam giới. Nam giới chỉ chiếm 40.87%, Nữ giới chiếm 59.13%. (Fanpage chủ yếu đưa tin tức về người nổi tiếng trên các thể loại mạng xã hội hay trên showbiz Việt).



8. Tỷ lệ phân bố người theo dõi ở 63 tỉnh thành hay ở 3 miền

Biểu đồ về số lượng người theo dõi Fanpage ở 63 tỉnh thành



Với biểu đồ trên ta thấy được chủ yếu người dùng đều ở 2 thành phố lớn là Hà Nội và Hồ Chí Minh (Đặc biệt Hồ Chí Minh chiếm phần lớn và nổi trội hơn so với các tỉnh thành khác).

***Biểu đồ phân bố người dùng ở 3 miền**

Ta thấy tỷ lệ 3 miền Bắc Trung, Nam. Miền Bắc và Miền Nam chiếm khá lớn lần lượt là 53,85% và 35,58% còn Miền Trung chỉ chiếm 10,58%.