Báo cáo cuối kì môn phân tích dữ liệu với Python

Họ và tên: Long Hoàng Vinh.

Mã sinh viên: 22022673.

Giới thiệu chung: Bài báo cáo này sẽ phân tích fanpage Basketball Forever. Đây là fanpage đưa tin về giải bóng rổ Nhà nghề Mĩ (NBA).

Link Github: link

Phần 1: Tiền xử lí dữ liệu

- Vì dữ liệu ở dạng thô có nhiều trường không cần dùng tới, vì vậy, để tối ưu bộ nhớ, ta sẽ xóa những trường này đi:

```
df = df.drop(list_col_delete, axis = 1)
```

- Lọc ra ngày trong tuần, giờ đăng từ ngày tháng phục vụ cho việc phân tích sau này thuận tiện hơn:

```
df['time'] = pd.to_datetime(df['time'])
#retrieve day name from datetime
df['Day_name'] = df['time'].dt.day_name()
# retrieve time_post(hour) from datetime
df['Time_post_hour'] = df['time'].dt.hour
# retrieve dd-mm-yy format
df['post_time'] = df['time'].dt.strftime('%d-%m-%y')
```

- Chỉ cần biết bài viết có ảnh/video hay không, ta sẽ chuyển về dạng link ảnh, video về dạng boolean:

```
df['Has_image'] = df['image'].astype(bool)
df['Has_video'] = df['video_id'].astype(bool)
```

- Tách từng cảm xúc của từng bài viết thành lượt Like, Care, Love, Haha, Wow, Sad, Angry:

```
df[['Like', 'Love', 'Haha', 'Wow', 'Sad', 'Care', 'Angry']] = df['reactions'].apply(pd.Series)

df = df.drop('reactions', axis=1)
```

- Điền những dữ liệu null bằng 0:

```
df.fillna(0, inplace = True)
```

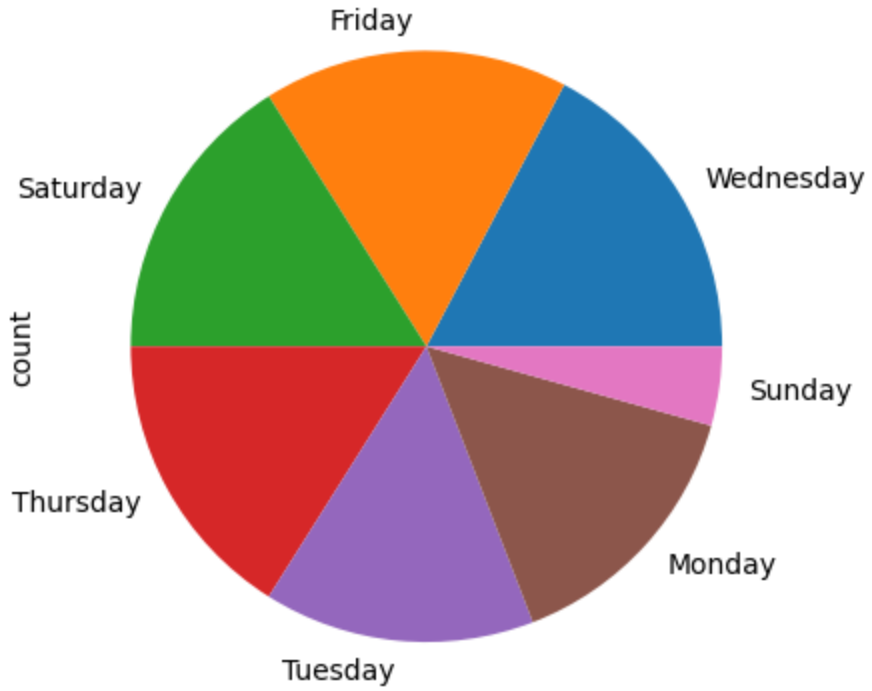
Phần 2: Phân tích

1. Phân tích cách hoạt động của Fanpage.

Ở phần này ta sẽ phân tích thói quen và tần suất đăng bài của page:

- Page thường đăng bài vào thời điểm nào trong ngày?
- Page thường đăng vào ngày nào trong tuần?
- Có sự liên quan giữa thời gian đăng bài và thời gian diễn ra trận đấu không?

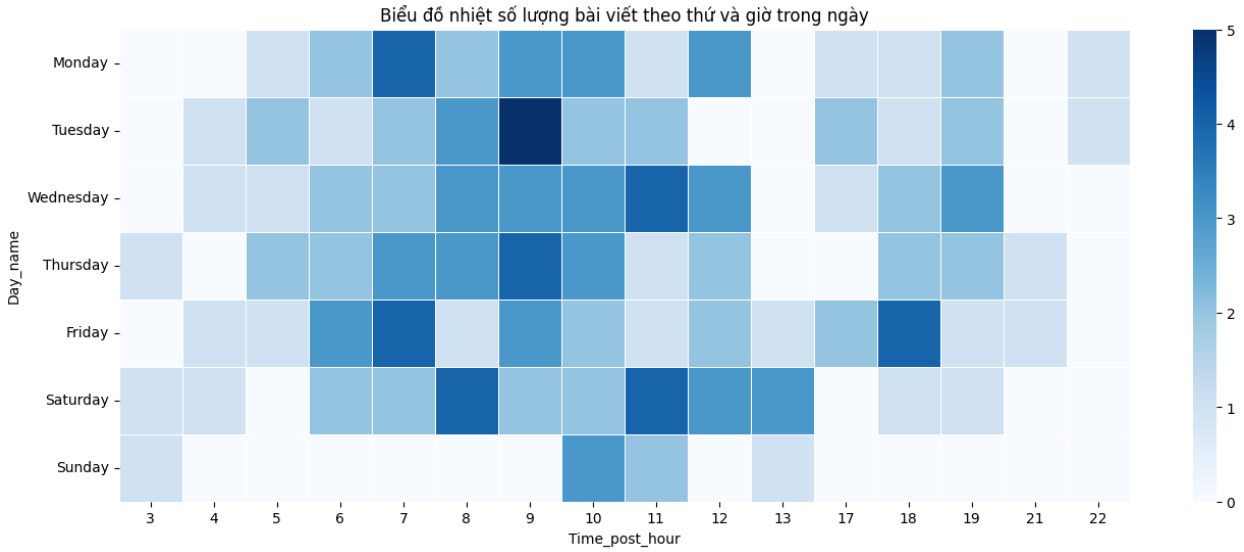
Đầu tiên, ta phân tích số lượng bài đăng của page theo các ngày trong tuần:



Qua biểu đồ trên, ta thấy được: trong ngày nghỉ cuối tuần, số lượng bài đăng ít hơn hẳn so với những ngày còn lại trong tuần, điều này khá vô lí vì cuối tuần hầu hết mọi người đều được nghỉ, lẽ ra page cần đăng nhiều bài hơn để tận dụng điều này. Có một vài giả thiết để giải thích điều này:

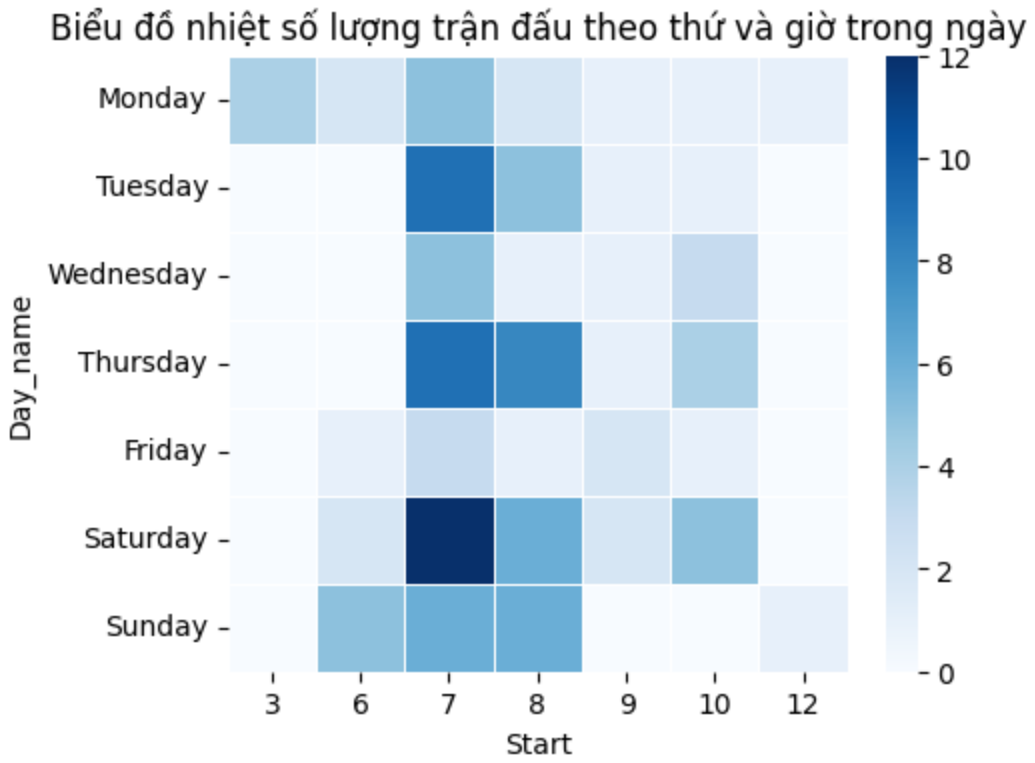
- Có thể page thuộc một công ty truyền thông, có thể có ngày nghỉ cuối tuần cho nhân viên, nên cuối tuần đăng ít bài hơn.
- Do chênh lệch múi giờ: dữ liệu thu thập theo giờ Việt Nam còn page thì ở Mỹ nên thực chất có thể khác so với dữ liệu.

Tiếp theo, ta phân tích thời điểm đăng bài trong ngày:



Đây là biểu đồ về thời gian đăng bài, ta thấy hầu hết các bài đăng đều được đăng từ 6h sáng tới 12h trưa theo giờ Việt Nam. Tuy nhiên, fanpage này ở Mỹ nên theo thời gian bên Mỹ thì sẽ là khoảng từ 6h tối đến 12h đêm hôm trước. Vì vậy, ta có thể giải thích điều vô lí phía trên là do chênh lệch múi giờ, hầu hết bài đăng vào ngày Chủ Nhật ở Mỹ được đăng vào Thứ Hai (theo giờ trong dữ liệu)

Dưới đây là biểu đồ về số lượng trận đấu và thời điểm diễn ra theo ngày(dữ liệu này thu thập trên website)



Qua heatmap trên, ta thấy các trận đấu diễn ra từ 3h sáng đến 12h, diễn ra chủ yếu vào 6h đến 10h sáng, thời điểm đăng bài nhiều nhất cũng nằm trong khoảng từ 6h đến 12h sáng. Vậy nên, ta có thể rút ra nhận xét sau:

- Page đưa thường đăng bài vào các thời điểm trận đấu diễn ra ⇒ Page theo dõi sát sao, đưa tin nhanh về diễn biến trận đấu.

💡 Tổng quan về thói quen đăng bài của page:

- Page thường đăng bài vào trong lúc các trận đấu diễn ra.
- Page thường đăng bài vào hầu hết các ngày trong tuần, trừ ngày Chủ Nhật.
- Page thường đăng vào buổi tối theo giờ Hoa Kì.

2. Phân tích lượt tương tác của page.

Ở phần này ta sẽ tìm hiểu:

- Những dạng bài viết được yêu thích hơn?
- Những nội dung nào được ưa chuộng hơn?
- Mối liên hệ giữa các lượt tương tác như thế nào?

1. Phân tích các dạng bài viết được yêu thích.

Cùng nhìn qua tổng lượt bày tỏ cảm xúc của page:

```
[120]: count       162.000000
       mean      20896.086420
       std       25207.374662
       min         205.000000
       25%        4107.500000
       50%       13632.500000
       75%       28599.500000
       max      164274.000000
       Name: reaction_count, dtype: float64
```

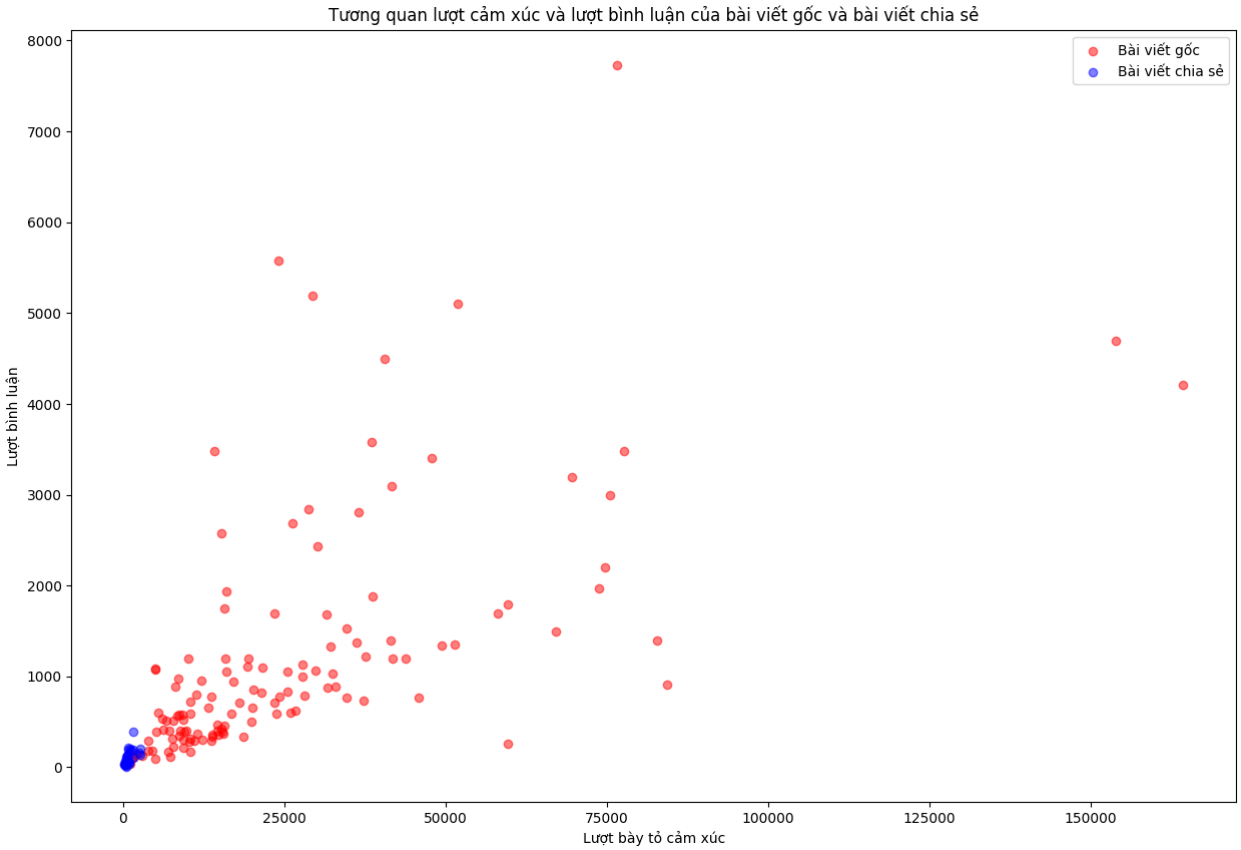
Có thể đưa ra nhận xét:

- Lượt tương tác trung bình của page ở mức tương đối cao: khoảng 20000 lượt mỗi bài, tuy nhiên chênh lệch giữa bài ít nhất và bài nhiều nhất rất lớn ⇒ một số dạng nội dung sẽ ăn khách hơn

Qua tìm hiểu thì thấy bài viết của page thường có 2 dạng:

- Bài viết gốc: page tự sáng tạo nội dung.
- Bài viết chia sẻ: bài viết được page chia sẻ từ nhiều nguồn, thường là gắn link.

Ta có biểu đồ về lượt tương tác trung bình giữa hai dạng bài viết này:



Ta có nhận xét:

- Người theo dõi thường tương tác nhiều hơn với những bài viết do chính page viết, hơn là những bài viết chia sẻ từ nguồn khác.
- Bài viết chia sẻ ít thu hút sự chú ý từ người xem hơn.

Vấn đề: *Tại sao những bài viết chia sẻ từ nguồn khác vẫn được đăng lên dù người xem không thích?*

Ta cùng xem qua nội dung ở trong những bài viết chia sẻ:

```
post_share['shared_text']
```

```
15     BASKETBALLFOREVER.COM\nInstagram Model Exposes...
21     BASKETBALLFOREVER.COM\nLeaked Audio From Warri...
23     BASKETBALLFOREVER.COM\nKelly Oubre Jr.'s Car A...
24     BASKETBALLFOREVER.COM\nMark Jackson Loses MSG ...
30     BASKETBALLFOREVER.COM\nPat Beverley Calls Out ...
35     BASKETBALLFOREVER.COM\nNBA Trade Rumors Alread...
36     BASKETBALLFOREVER.COM\nRudy Gobert Fires Back ...
41     BASKETBALLFOREVER.COM\nNBA World Reacts to Wil...
50     BASKETBALLFOREVER.COM\nKawhi Leonard's Frustra...
51     BASKETBALLFOREVER.COM\nThe Brutal Stats That A...
55     BASKETBALLFOREVER.COM\nChris Paul Called Out F...
60     BASKETBALLFOREVER.COM\nThe Canadian Giant Who ...
61     BASKETBALLFOREVER.COM\nNBA World Outraged Over...
66     BASKETBALLFOREVER.COM\nEverything We Know Abou...
74     BASKETBALLFOREVER.COM\nJohn Wall Sets His Sigh...
75     BASKETBALLFOREVER.COM\nTaylor Jenkins Blasts R...
88     BASKETBALLFOREVER.COM\nJames Harden Responds t...
89     BASKETBALLFOREVER.COM\nJoel Embiid Takes Subtl...
92     BASKETBALLFOREVER.COM\nNBA World Reacts to Atl...
95     BASKETBALLFOREVER.COM\nJeanie Buss Opens Up Ab...
96     BASKETBALLFOREVER.COM\nJordan Poole Addresses ...
106    BASKETBALLFOREVER.COM\nLeBron James Responds t...
109    BASKETBALLFOREVER.COM\nThe Mavs Rookie Who Doe...
115    BASKETBALLFOREVER.COM\nLeBron Explains Why He'...
118    BASKETBALLFOREVER.COM\nLeBron Called Out for D...
122    BASKETBALLFOREVER.COM\nPaul George Addresses J...
123    BASKETBALLFOREVER.COM\nWizards Roasted Over Di...
125    BASKETBALLFOREVER.COM\nLeBron James Describes ...
136    BASKETBALLFOREVER.COM\nKnicks Players Divided ...
138    BASKETBALLFOREVER.COM\nThomas Bryant Explains ...
139    BASKETBALLFOREVER.COM\nNBA Fans React to Heat'...
157    BASKETBALLFOREVER.COM\nEmbiid Responds to Hard...
160    BASKETBALLFOREVER.COM\nShai Gilgeous-Alexander...
Name: shared_text, dtype: object
```

Ta thấy rõ đặc điểm chung của những bài viết này đều có nội dung chứa tên miền “BASKETBALLFOREVER.COM” cùng tên với tên fanpage. Có thể trang web và fanpage này đều cùng một công ty truyền thông. Ta có nhận xét:

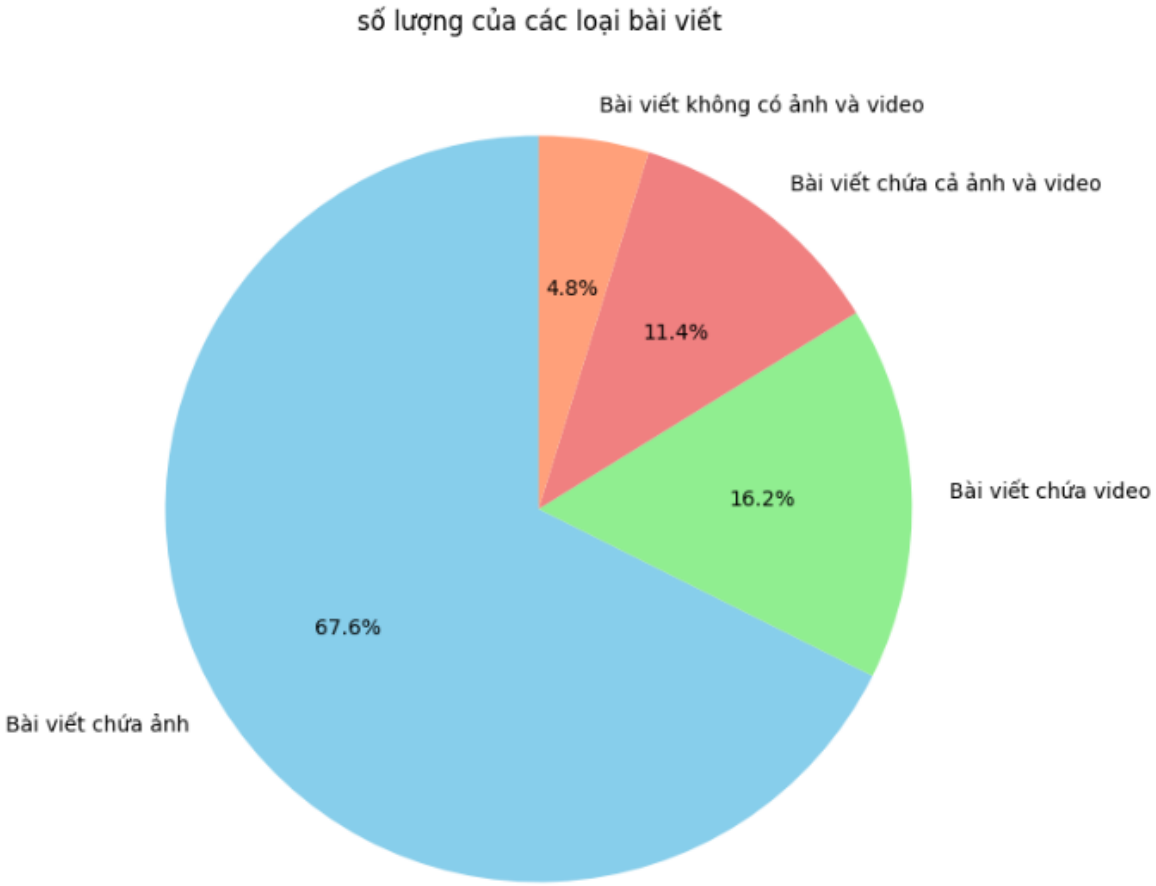
- Mặc dù những bài viết chia sẻ có ít lượt tương tác vẫn đăng vì đây là cách mà page quảng bá, thu hút lượt xem cho chính trang web của họ.

- Page có tính chuyên nghiệp, quy mô lớn.

Trong bài viết của page thường chứa ảnh, video, có cả hai hoặc không có cả ảnh lẫn video. Vậy thì:

- Page thường đăng ảnh hay video? Tại sao?
- Ảnh hay video được người xem yêu thích hơn?

Biểu đồ tròn về số lượng các dạng bài viết:

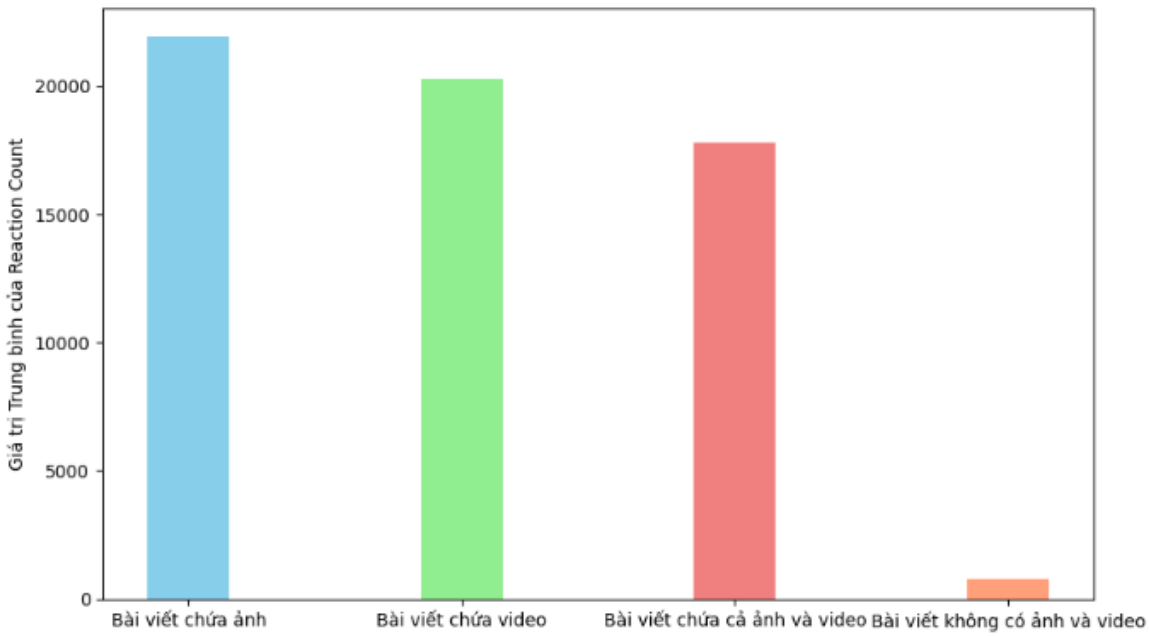


Qua biểu đồ trên, ta thấy những bài viết chứa ảnh chiếm đa số, có thể là do:

- Vì là trang đưa tin, ảnh dễ dàng truyền đạt nội dung hơn video.
- Ảnh dễ chỉnh sửa, thu thập hơn, video yêu cầu kĩ năng chỉnh sửa cao và tốn thời gian ⇒ không phù hợp với việc đưa tin nhanh.

- Đăng video có thể dễ bị vi phạm bản quyền.

Để biết người xem thường tương tác với nội dung nào hơn, ta cùng xem biểu đồ về lượng bày tỏ cảm xúc trung bình các dạng bài viết:



Có thể thấy, bài viết chứa ảnh và bài viết chứa video có lượng tương tác gần ngang nhau, ba cột đầu tiên không chênh lệch nhau nhiều, vậy ta có nhận xét:

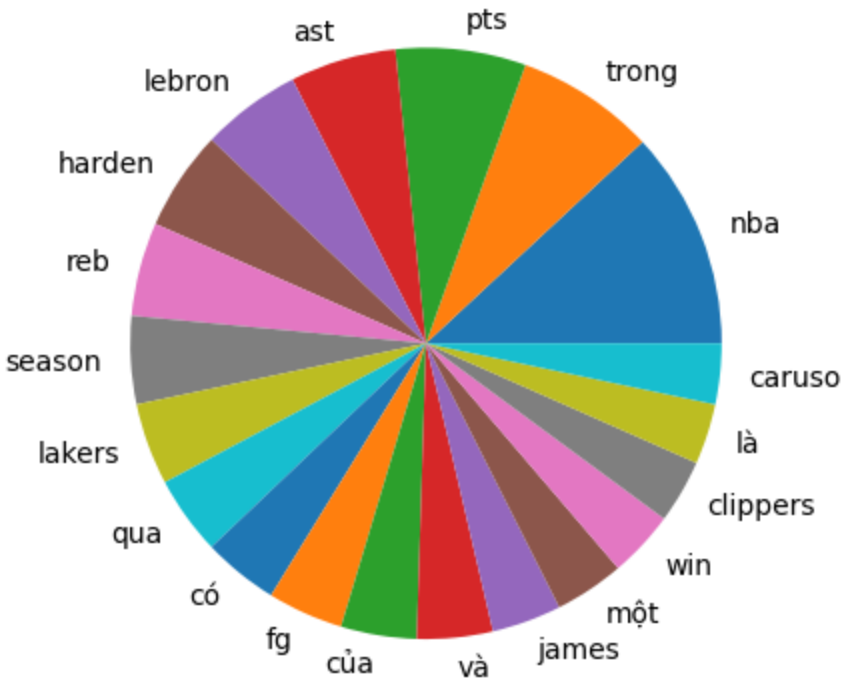
- Người xem thích cả nội dung chứa ảnh và nội dung chứa video, chênh lệch không nhiều.
- Chỉ cần trong bài viết có ảnh hoặc video, bài viết sẽ có lượng tương tác tương đối lớn.
- Người xem không thích những nội dung không có ảnh và video.

💡 Tổng quan về các dạng bài viết được yêu thích:

- Bài viết gốc nhiều lượt bày tỏ cảm xúc hơn bài viết chia sẻ.
- Bài viết chia sẻ luôn chia sẻ chính bài viết trang web của page này.
- Bài viết chứa hình ảnh có số lượng áp đảo.
- Bài viết chứa ảnh hoặc video thường có lượng tương tác cao hơn.

2. Phân tích nội dung bài viết.

Về phần nội dung bài đăng, sau đây là những từ khóa xuất hiện nhiều nhất trong các bài đăng:



Từ dữ liệu trên, ta có thể phân loại:

- Tên cầu thủ: Harden, Lebron, James, Caruso.
- Tên đội bóng: Lakers, Clippers.
- Chỉ số: pts(points = điểm số), ast(assists = kiến tạo), reb(rebounds = bắt bóng bật bảng), fg(field goals = số lần dứt điểm).
- Những từ còn lại có vẻ chỉ là những thành phần ngữ nghĩa trong câu.

LeBron James trở thành từ khóa phổ biến cũng có vẻ dễ hiểu vì đây từng là cầu thủ xuất sắc nhất trong gần 2 thập kỉ qua, là cầu thủ ghi nhiều điểm nhất mọi thời , là biểu tượng của giải đấu, cho nên LeBron James trở thành nội dung thu hút người theo dõi là điều dễ hiểu.

LeBron James đang thi đấu cho đội bóng Lakers, vậy nên, đội bóng này cũng trở thành một từ khóa phổ biến trong nội dung các bài viết.

Cầu thủ xuất hiện nhiều không kém chính là James Harden, có thể là do vụ chuyển nhượng gần đây vào ngày 31/10/2023, James Harden đã được nhắc tới nhiều hơn.

https://www.hoopshype.com › 2023 › 10 › 31 › james-har…

James Harden traded to Clippers | HoopsHype

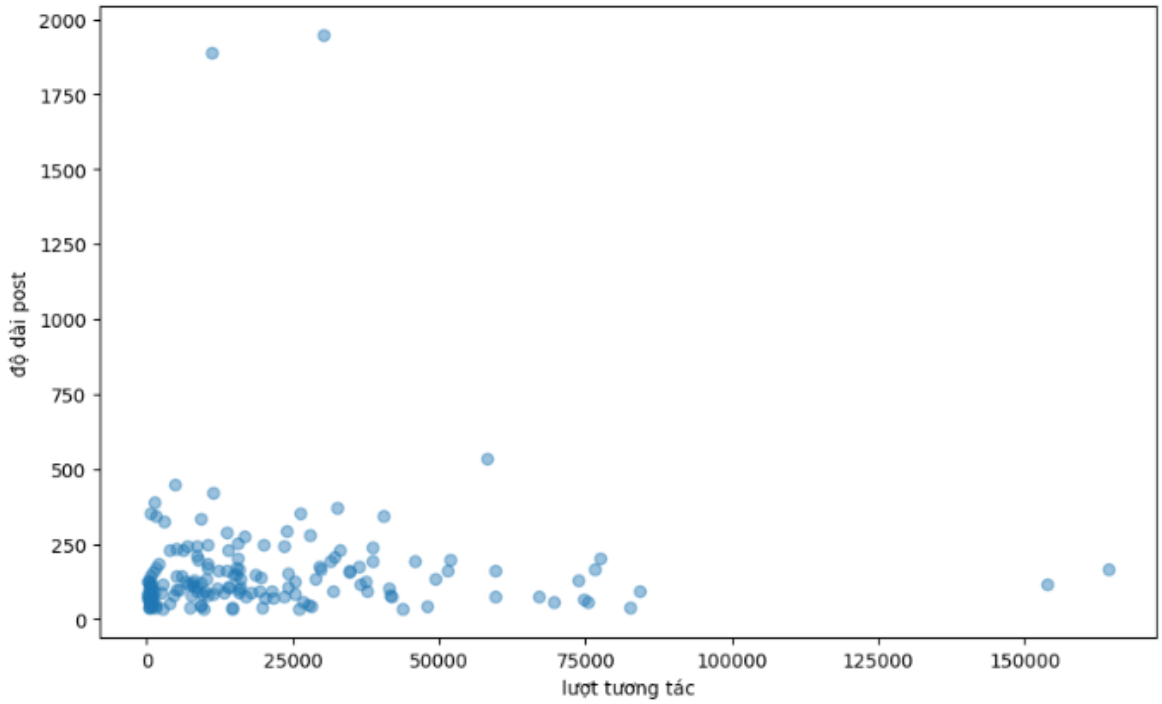
Adrian Wojnarowski: BREAKING: The Philadelphia 76ers have agreed on a trade to send guard James Harden to the Los Angeles Clippers, sources tell ESPN.Source: Twitter @wojespn Whats the buzz ...

Đội bóng mà James Harden chuyển đến là Clippers cũng nằm trong top từ khóa đăng bài.

Từ những kết luận trên, ta có thể nhận xét về nội dung bài viết của page:

- Lebron James có nhiều người hâm mộ và quan tâm.
- Nội dung xoay quanh cầu thủ, đội bóng NBA.
- Page tận dụng những nội dung nhiều người quan tâm để đăng bài, thu hút tương tác.

Dưới đây là biểu đồ về lượng bày tỏ cảm xúc và độ dài nội dung bài đăng:



qua biểu đồ, ta nhận xét:

- Có 2 bài viết có lượng kí tự vượt trội nhưng lượt tương tác chỉ ở mức khá.
- Hầu hết những bài viết đều dưới 250 kí tự.

Từ nhận xét trên, ta có thể kết luận:

- Nội dung của page thường ngắn gọn xúc tích, phù hợp với mục đích đưa tin.

- Người theo dõi page thường thích những nội dung ngắn gọn, dễ hiểu hơn.

> 💡 Tổng quan về nội dung các bài viết trên page:
>
> - Mặc dù đưa tin về toàn bộ giải đấu, nhưng có vẻ Lebron James vẫn được ưu ái với nhiều bài đăng hơn.
> - Nội dung xoay quanh những chủ đề nhiều người quan tâm.
> - Nội dung ngắn gọn, dễ truyền tải, dễ tiếp cận.

3. Phân tích mối liên hệ giữa lượt tương tác và các yếu tố khác:

Bài viết tương tác nhiều nhất có 164274 lượt tương tác, ta hãy cùng xem qua post này:

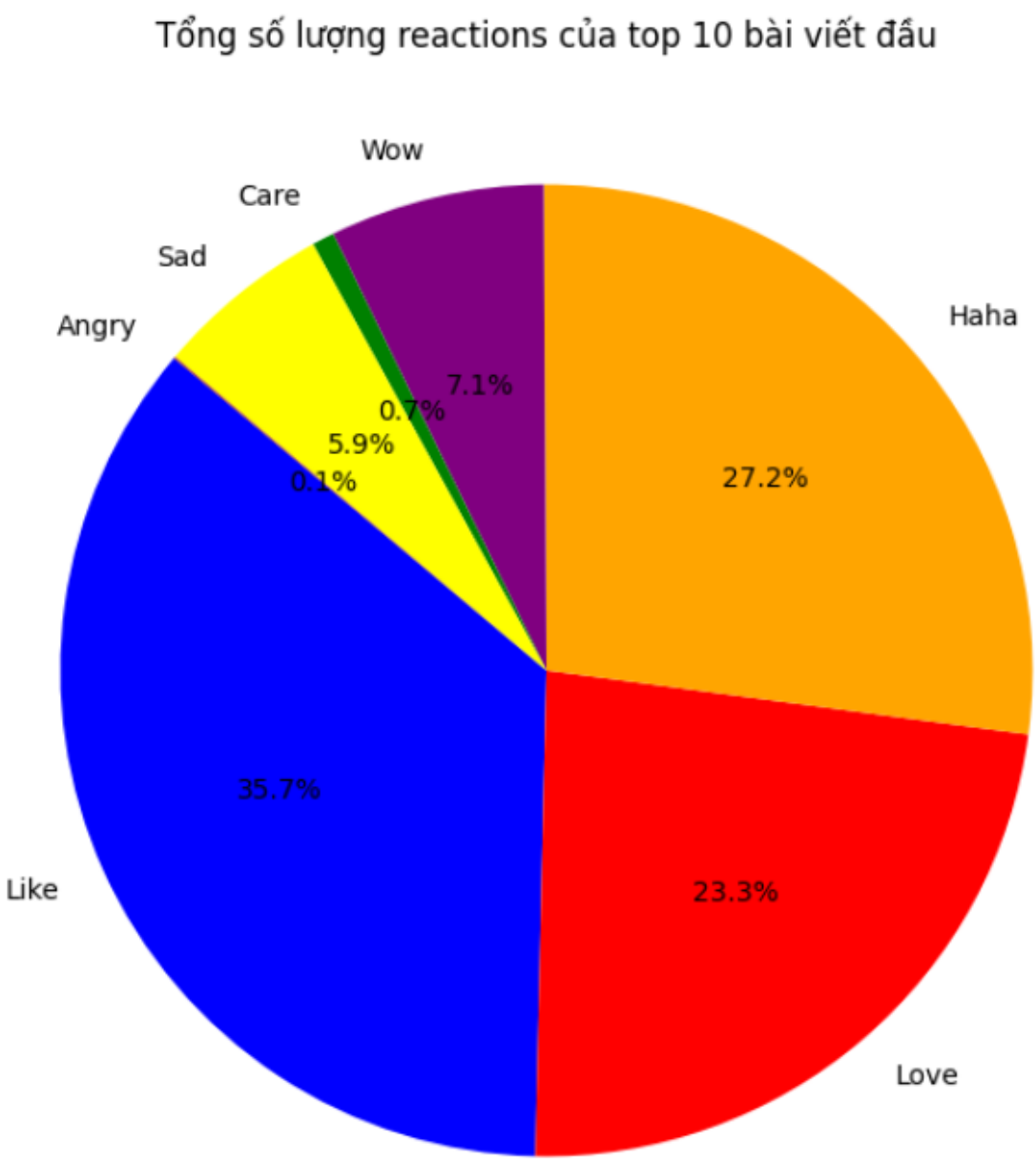
```
post_text         The Warriors CHOKE away a 12-point lead as the...
shared_text                                                       0
time                                              2023-11-15 12:32:49
comments                                                       4205
shares                                                        13131
comments_full     [{'comment_id': '308864348616150', 'comment_ur...
reaction_count                                                164274
Like                                                           24371
Love                                                            1296
Haha                                                          136622
Wow                                                             561
Care                                                            220
Sad                                                             956
Angry                                                           248
Day_name                                                  Wednesday
Time_post_hour                                                   12
post_time                                                 15-11-23
Has_image                                                     True
Has_video                                                    False
Name: 39, dtype: object
```

Có một điều khá bất ngờ là số lượng Haha ở bài viết này rất lớn, hơn rất nhiều so với lượt Like - cảm xúc phổ biến nhất. Ngoài ra, lượt bình luận và chia sẻ cũng lớn.

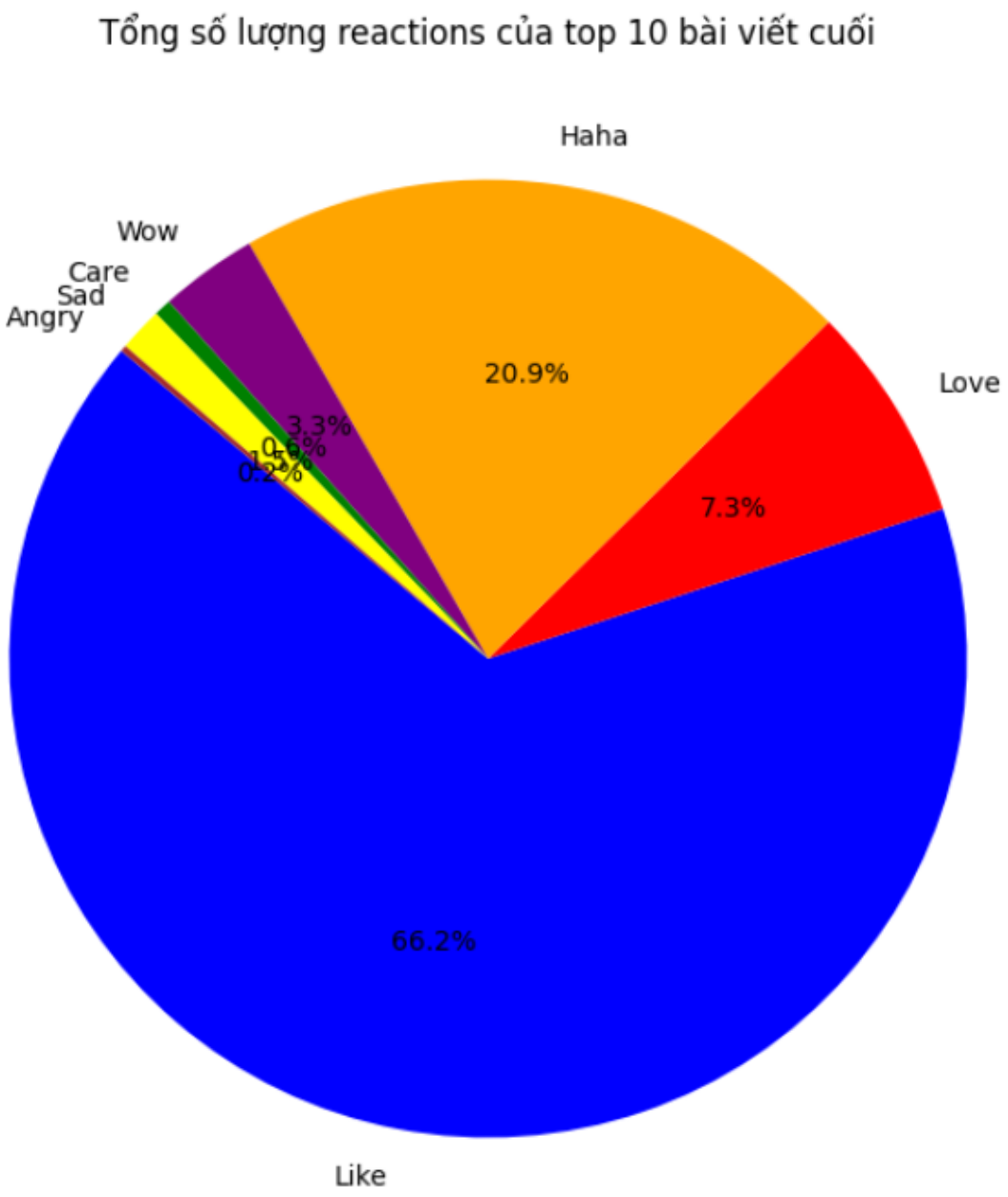
Ta có thể đặt ra câu hỏi:

- Ở những bài viết nhiều lượt bày tỏ cảm xúc thì lượt Like không quá áp đảo? Hoặc tỉ trọng các loại cảm xúc ở các bài viết nhiều tương tác và các post ít tương tác khác nhau như thế nào?
- Bài đăng vào thời điểm nào thì có nhiều lượt tương tác?
- Mối liên hệ giữa lượt bày tỏ cảm xúc, bình luận và chia sẻ là như thế nào?

Sau đây là biểu đồ thể hiện tỉ trọng các loại cảm xúc của 10 bài viết nhiều tương tác nhất:



Và 10 bài viết ít tương tác nhất:

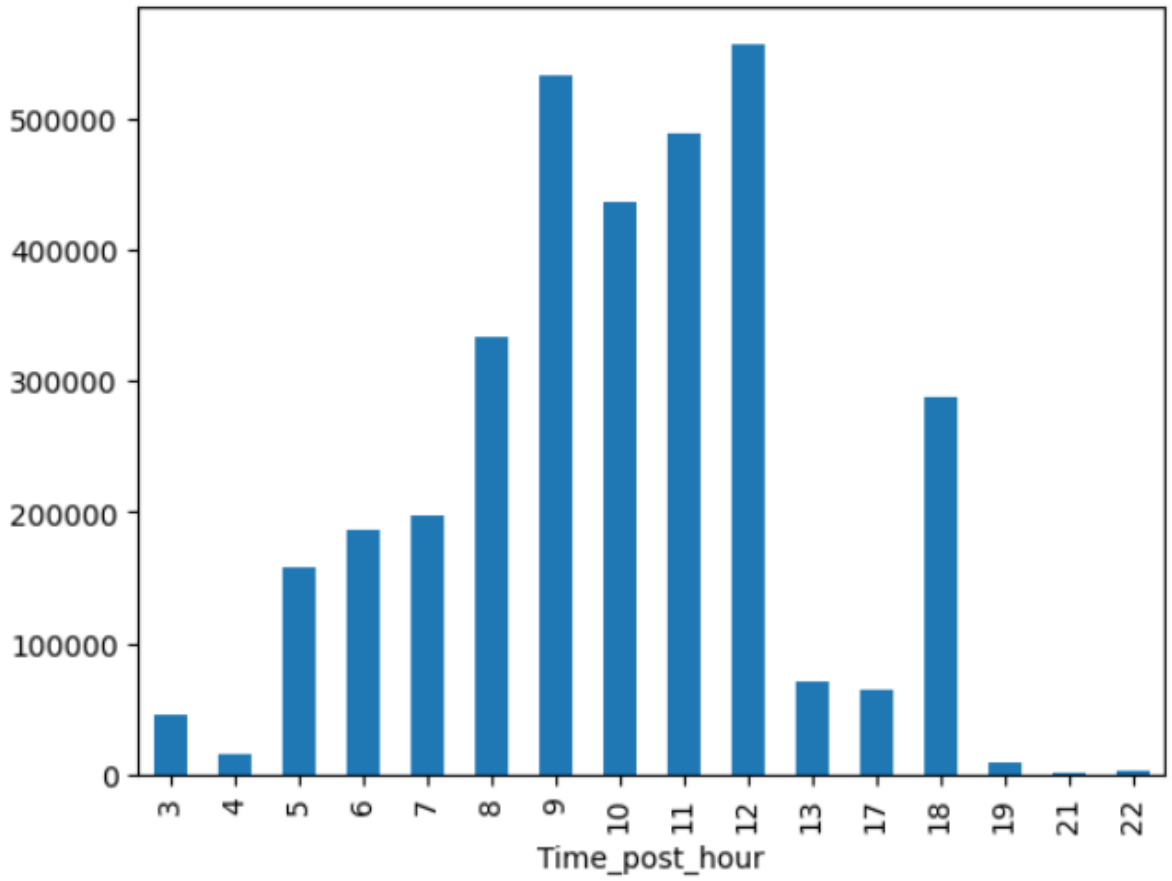


Qua hai biểu đồ trên ta có thể đưa ra kết luận: Ở những bài viết nhiều lượt bày tỏ cảm xúc, lượt Like không chiếm áp đảo như những bài viết ít tương tác.

Ta có thể nhận xét: Những bài viết có yếu tố hài hước, bất ngờ thường thu hút nhiều người theo dõi hơn.

Bài viết đăng vào khoảng thời gian nào thu hút được nhiều tương tác nhất?

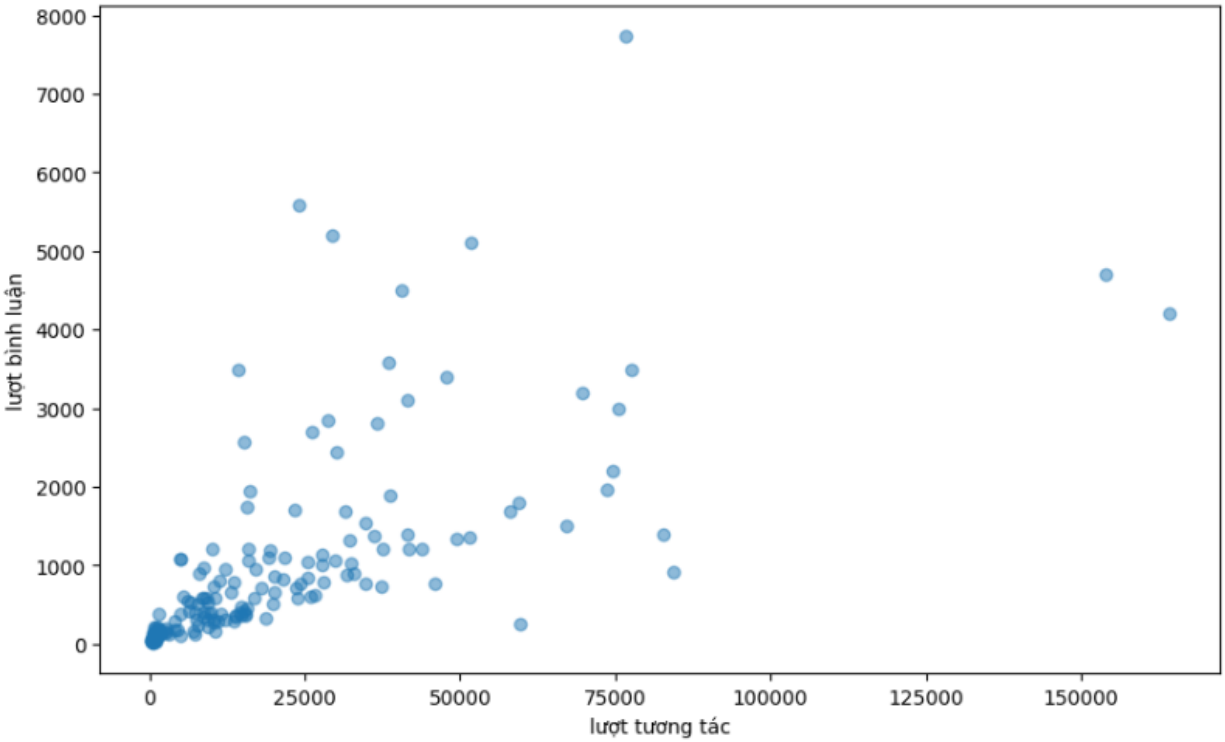
Sau đây là biểu đồ lượt tương tác của các bài đăng theo giờ đăng bài:



qua biểu đồ, ta thấy, bài đăng nhận được nhiều tương tác nhất khi đăng vào khoảng từ 9h-12h sáng theo giờ Việt Nam, còn theo giờ Mĩ, khoảng thời gian này sẽ thuộc vào buổi tối (khoảng thời gian từ 6h tối đến 12h đêm) điều này dễ hiểu bởi vì thời gian buổi tối thì hầu hết mọi người sẽ có thời gian rảnh để giải trí.

Vậy nên: Những bài viết vào thời gian người dùng rảnh rỗi thường có lượt tương tác cao hơn.

Nếu một bài viết thu hút nhiều sự chú ý, có khả năng cao rằng người đọc sẽ tương tác thông qua lượt bày tỏ cảm xúc hoặc bình luận. Dưới đây là biểu đồ thể hiện mối quan hệ giữa hai thuộc tính trên:

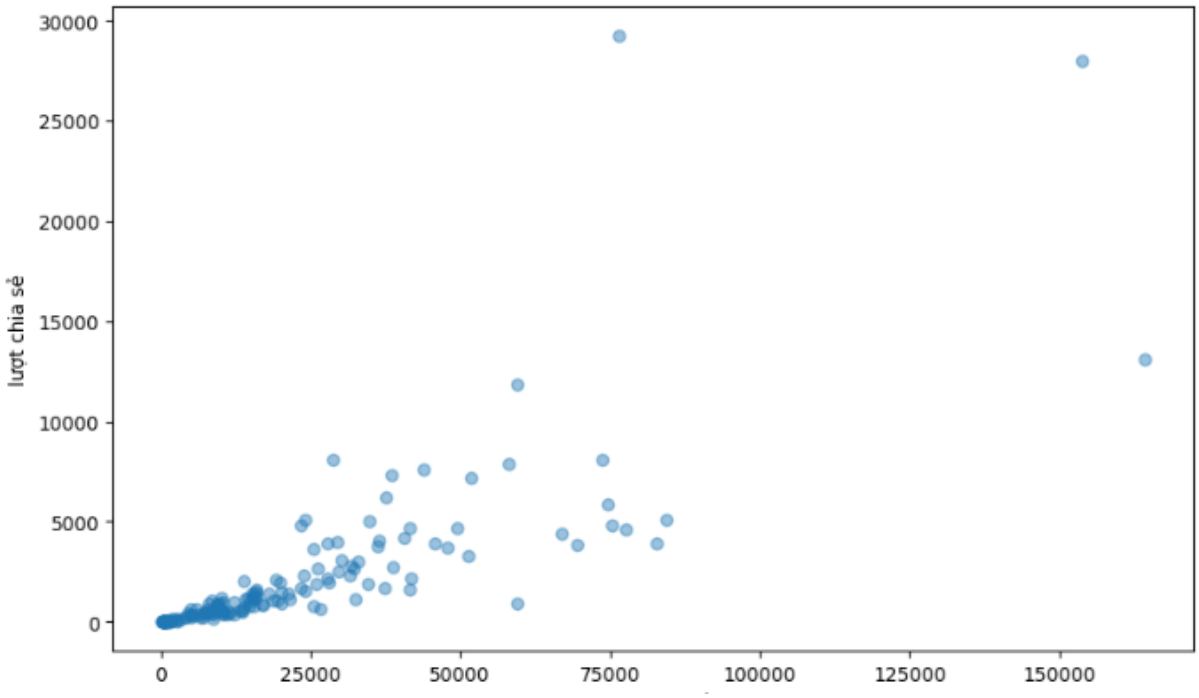


Ta thấy các điểm dữ liệu có xu hướng tạo thành một đám đông đi lên dần từ góc dưới bên trái lên góc trên bên phải ⇒ Có xu hướng tăng tương đồng.

Vậy nên, ta kết luận: Bài viết nhiều lượt bày tỏ cảm xúc thì cũng có nhiều bình luận.

Một bài viết được nhiều lượt chia sẻ thì tương tác có cao không? thông thường, khi một bài viết có nhiều lượt bày tỏ cảm xúc, nó có thể trở thành "hot topic" và lan truyền nhanh chóng trên mạng xã hội. Người dùng thích thú với nội dung này có thể chia sẻ để đưa ra ý kiến hoặc thảo luận với cộng đồng của họ ⇒ bài viết có nhiều lượt thả cảm xúc hơn.

Theo biểu đồ dưới đây thì dự đoán là đúng, do biểu đồ thể xu hướng tăng nếu một giá trị tăng thì cái còn lại cũng tăng.



Vậy nên: Bài viết nhiều lượt chia sẻ thì lượt tương tác cũng nhiều.

💡 Tổng quan về mối liên hệ giữa các lượt tương tác:

- Những bài viết có lượng bày tỏ cảm xúc lớn thì lượng Like chiếm tỉ trọng không quá áp đảo.
- Quan hệ giữa lượt tương tác và thời gian đăng bài: Những bài đăng vào thời gian người theo dõi rảnh thì có lượt tương tác cao hơn.
- Bài viết nhiều lượt bày tỏ cảm xúc thì lượt bình luận và lượt chia sẻ cũng cao.

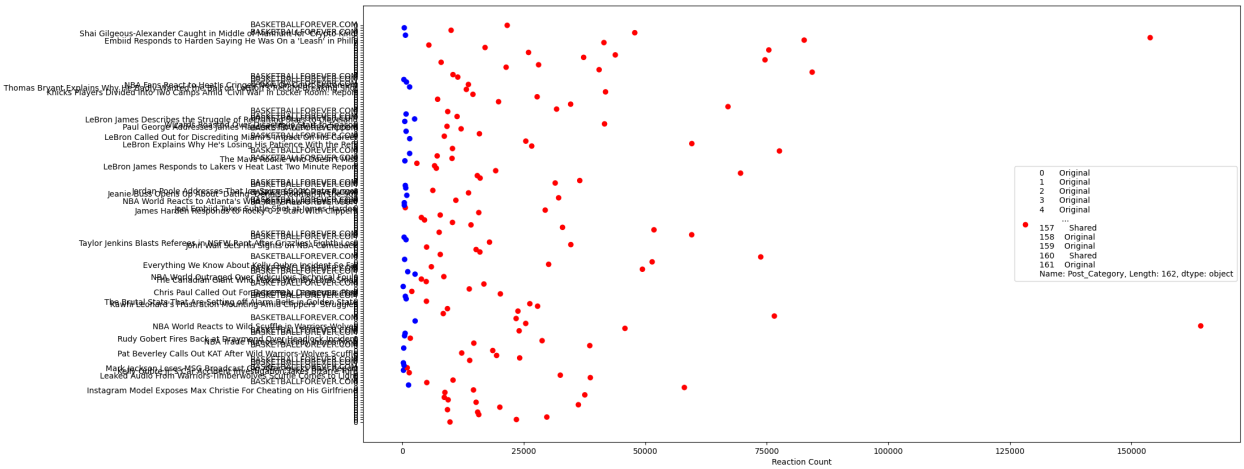
3. Bài toán phân loại bài viết gốc và bài viết chia sẻ dựa vào lượt bày tỏ cảm xúc.

1. Bài toán đặt ra và input:

- Dùng bài toán phân loại dựa vào số lượng bày tỏ cảm xúc để lọc ra những bài viết gốc và bài viết chia sẻ.
- Ứng dụng: Khi gặp một số trường hợp xấu, dữ liệu bị thiếu, có thể dùng lượt bày tỏ cảm xúc để phân loại 2 dạng bài trên thay vì bỏ đi toàn bộ dữ liệu.
- Input: dạng bài viết và số lượng lượt bày tỏ cảm xúc

2. Triển khai:

- Ý tưởng: có thể đặt ra một ngưỡng(số lượng lượt bày tỏ cảm xúc) phù hợp với dữ liệu để phân loại.
- Ta thấy dữ liệu phân cụm khá mạnh, màu xanh là bài viết chia sẻ, màu đỏ là bài viết gốc, có thể áp dụng ý tưởng trên:



- Ta chia dữ liệu thành 2 phần, 98 % cho việc huấn luyện, 2% cho test. Sử dụng thư viện Sklearn và mô hình hồi quy logistic.

3. Kết quả:

- Kết quả thu được khá tích cực, tuy nhiên nếu thu thập được nhiều dữ liệu hơn thì sẽ chính xác hơn.

```
Accuracy: 0.9696969696969697
Classification Report:
              precision    recall  f1-score   support

    Original       0.96      1.00      0.98        23
      Shared       1.00      0.90      0.95        10

    accuracy                           0.97        33
   macro avg       0.98      0.95      0.96        33
weighted avg       0.97      0.97      0.97        33
```