

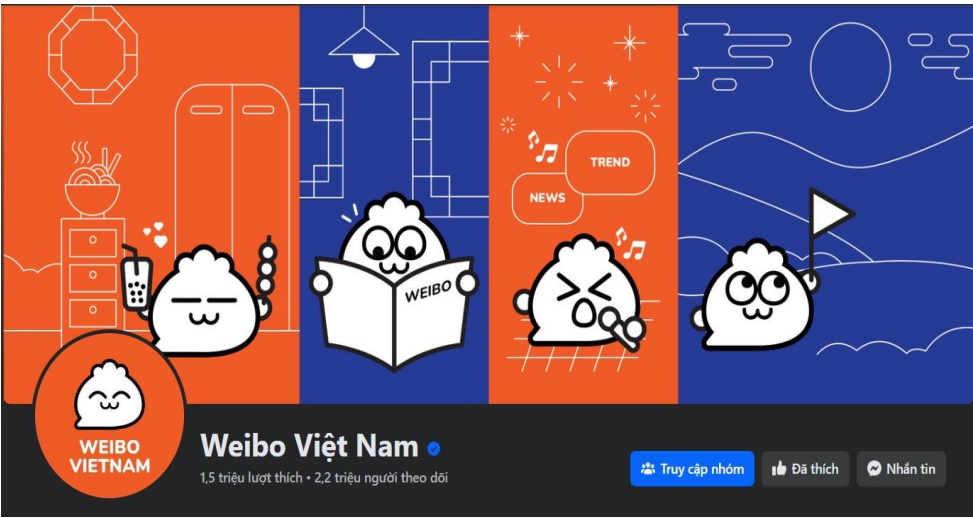
BÁO CÁO: PHÂN TÍCH DỮ LIỆU
TRANG FACEBOOK BẰNG PYTHON

CHƯƠNG 1. TỔNG QUAN VÀ THU THẬP DỮ LIỆU CỦA TRANG FACEBOOK.....	3
1.1. Tóm tắt tổng quan về trang Facebook đã chọn.....	3
1.2. Thực hiện chuẩn bị thu thập dữ liệu từ trang Weibo Việt Nam	3
1.3. Thực hiện thu thập dữ liệu từ trang Weibo Việt Nam	4
CHƯƠNG 2. LÀM SẠCH VÀ TIỀN XỬ LÝ DỮ LIỆU	5
2.1. Mô tả dữ liệu thô vừa thu thập được.....	5
2.2. Làm sạch và tiền xử lý dữ liệu thô.....	7
2.2.1. Loại bỏ các trường dữ liệu không sử dụng.....	7
2.2.2. Tái tổ chức bộ dữ liệu	8
2.2.3. Tiếp tục làm sạch và tiền xử lý dữ liệu	10
CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU	11
3.1. Cài đặt thư viện và tải lên dữ liệu.....	11
3.2. Phân tích dữ liệu đã thu thập.....	12
3.2.1. Phân tích tương tác của trang.....	12
3.2.2. Phân tích các bài đăng, nội dung đặc biệt	14
3.3.2.1. Bài đăng có lượt bình luận lớn nhất.....	14
3.3.2.2. Bài đăng có tổng lượt tương tác lớn nhất	15
3.3.2.3. Những người dùng có lượt bình luận lớn nhất	16
3.2.3. Phân tích hoạt động của trang	17
3.2.4. Một số phân tích về nội dung bài viết và bình luận	18
3.2.3. Một số mối liên hệ giữa các trường dữ liệu	21
3.3. Một số kết luận	23

CHƯƠNG 1. TỔNG QUAN VÀ THU THẬP DỮ LIỆU CỦA TRANG FACEBOOK

1.1. Tóm tắt tổng quan về trang Facebook đã chọn

Trong quá trình chọn lựa và tìm hiểu các trang, fanpage lớn ở trên nền tảng Facebook, em đã quyết định lựa chọn trang Weibo Việt Nam làm trang để thực hiện việc nghiên cứu. Nhìn chung, Weibo Việt Nam là trang Facebook lớn, lâu đời và có sức ảnh hưởng trên nền tảng Facebook. Cụ thể, Weibo Việt Nam được tạo và đi vào hoạt động vào tháng 11 năm 2010, sau này được đổi tên chính thức trở thành “Weibo Việt Nam” vào tháng 11 năm 2018. Đây là Fanpage hướng tới việc chia sẻ các bài đăng, câu chuyện cũng như tin tức có liên quan đến các Cộng đồng Hoa Ngữ và trong nước. Tính đến thời điểm hiện tại, trang đã có hơn 2,5 triệu người theo dõi và 1,5 triệu lượt thích.



Hình 1.1. Ảnh chụp trang Facebook Weibo Việt Nam

1.2. Thực hiện chuẩn bị thu thập dữ liệu từ trang Weibo Việt Nam

Sau khi lựa chọn trang, việc tiếp theo là cài đặt và thiết lập hệ thống để thực hiện thu thập dữ liệu từ trang Facebook. Ta sẽ thực hiện cài đặt môi trường hệ điều hành Linux, sử dụng phần mềm Visual Studio Code để thực hiện code. Ngôn ngữ lập trình sử dụng trong dự án là Python, được lập trình và chạy thông qua Jupiter Notebook. Các thiết lập về phiên bản như sau:

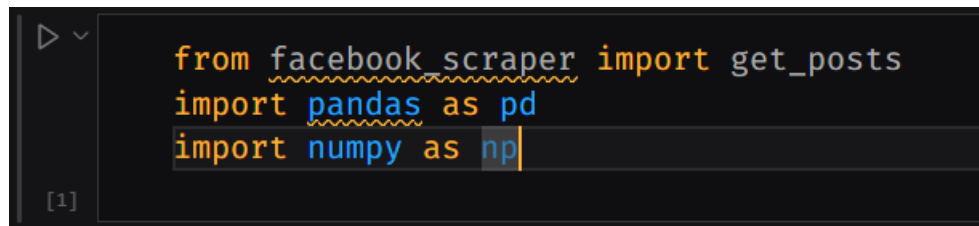
Tên	Phiên bản
Linux Ubuntu	22.04LTS x64
Visual Studio Code	1.80
Python Kernel	3.10

Bảng 1.1. Thông tin cài đặt môi trường

Toàn bộ nội dung thực hiện được đặt trong Folder “Facebook-Analytics-Project”. Folder này đã được Commit lên Github với link được đề cập trong trang tính ở bài nộp Final Project.

1.3. Thực hiện thu thập dữ liệu từ trang Weibo Việt Nam

Đầu tiên, tạo file Jupiter Notebook “Scrapping_Pages.ipynb” để thực hiện cài đặt và tiến hành Crawl dữ liệu từ trang. Thực hiện cài đặt các thư viện cần thiết:



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

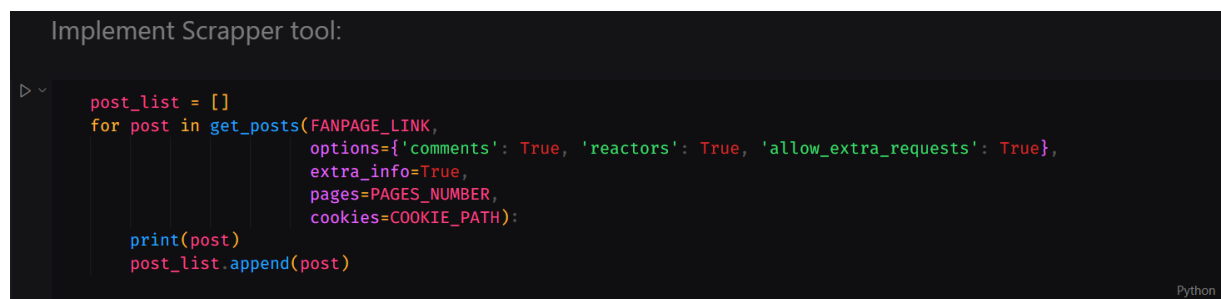
Hình 1.2. Cài đặt thư viện cho “Scrapping_Pages.ipynb”

Tiếp theo, định nghĩa các biến sử dụng cho mô đun thực hiện việc crawl dữ liệu từ trang Facebook, cụ thể như sau:

Tên biến	Kiểu dữ liệu	Giá trị	Mô tả
FANPAGE_LINK	String	'weibovietnam'	Tên đường dẫn đến trang
COOKIE_PATH	String	'www.facebook.com_cookies.txt'	Tên file chứa Cookie của trang
PAGES_NUMBER	Int	100	Số lượng ‘trang’ chứa các bài đăng

Bảng 1.2. Các biến sử dụng trong mô đun crawl

Sau khi khai báo các biến chính, triển khai công cụ crawl có trong thư viện facebook_scrapper và lưu thông tin thu được vào trong một list “post_list”:



```
Implement Scraper tool:

post_list = []
for post in get_posts(FANPAGE_LINK,
                      options={'comments': True, 'reactors': True, 'allow_extra_requests': True},
                      extra_info=True,
                      pages=PAGES_NUMBER,
                      cookies=COOKIE_PATH):
    print(post)
    post_list.append(post)
```

Hình 1.3. Thực hiện cài đặt crawl dữ liệu từ Weibo Việt Nam

Khi chương trình đã chạy xong hoàn tất, thực hiện việc lưu dữ liệu đã thu thập được lưu ban đầu vào một dataframe tạm “post_df_full”, sau đó lưu dataframe vào file csv “weibovietnam.csv”. Đây chính là file csv chứa dữ liệu thô thu thập được từ trang Weibo Việt Nam thông qua facebook_scrapper. Việc thu thập dữ liệu kết thúc ở đây.

```
post_id,text,post_text,shared_text,original_text,time,timestamp,image,image_lowquality,images,
images_description,images_lowquality,images_lowquality_description,video,video_duration_seconds,
video_height,video_id,video_quality,video_size_MB,video_thumbnail,video_watches,video_width,likes,
comments,shares,post_url,link,links,user_id,username,user_url,is_live,factcheck,shared_post_id,
shared_time,shared_user_id,shared_username,shared_post_url,available,comments_full,reactors,w3_fb_url,
reactions,reaction_count,with,page_id,sharers,image_id,image_ids,was_live,fetched_time    You, 5 days
731043432389703,"NHẬT KỶ ĐI LẤY CHỒNG XA: KHÔNG DẬY SỚM BỊ CHỒNG HẮT THẮNG XÔ NƯỚC LÊN GIƯỜNG

Chi có thể nói số mệnh của tôi không tốt, từ Tứ Xuyên gả đến Giang Tây, hiện tại đang sông ở Quảng Tây.

Trước khi lấy chồng, người nhà tôi đã nhiều lần ngăn cản. Sau đó, vì tôi quá kiên quyết nên mọi người
dành chấp thuận. Lấy chồng được vài tháng, tôi đã thấy hối hận rất nhiều lần. Đứng là tự mình làm khổ
mình...

Hôm nay là ngày nghỉ cuối tuần, tôi muốn ngủ thêm 1 tiếng để bù cho những ngày đi làm. Nhưng chồng
liên tục gọi tôi dậy, không gọi được tôi dậy bèn kéo tay tôi kéo người tôi ra mép giường. Tôi bực mình
ngồi dậy, anh ta đẩy tôi xuống đất. Tôi tỏ ra bực tức ngồi lại lên giường. Được một lúc thì thấy người
vừa ướt vừa lạnh. Hoá ra anh ta đã hắt nguyên một xô nước lên người tôi.

Bây giờ tôi đang vừa khóc vừa gọi điện về nhà, mong bố mẹ đến đón mình về.

>> Coi như xô nước của chồng tím đã thực sự thức tỉnh được tím rồi đây.
```

Hình 1.4. Hình chụp từ file “weibovietnam.csv”.

CHƯƠNG 2. LÀM SẠCH VÀ TIỀN XỬ LÝ DỮ LIỆU

2.1. Mô tả dữ liệu thô vừa thu thập được

Trước tiên, tạo một Jupiter Notebook “Data_Cleaning.ipynb” để thực hiện việc làm sạch và tiền xử lý. Thiết lập các thư viện cần thiết:

```
import required library

import pandas as pd
import numpy as np
from pprint import pprint
import json
import re
import datetime
```

Hình 2.1. Các thư viện cần thiết

Sau khi đã cài đặt xong thư viện, ta thực hiện tạo dataframe “raw_df” lấy dữ liệu từ file “weibovietnam.csv”. Thực hiện hàm `raw_df.info()`, ta sẽ có được cái nhìn tổng quan về bộ dữ liệu thu thập được:

```
# information of the data
raw_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 51 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   post_id                               1000 non-null   int64
1   text                                  953 non-null    object
2   post_text                             953 non-null    object
3   shared_text                           1 non-null      object
4   original_text                         0 non-null      float64
5   time                                  1000 non-null   object
6   timestamp                             1000 non-null   int64
7   image                                 897 non-null    object
8   image_lowquality                     1000 non-null   object
9   images                                954 non-null    object
10  images_description                    954 non-null    object
11  images_lowquality                     1000 non-null   object
12  images_lowquality_description         1000 non-null   object
13  video                                 43 non-null     object
14  video_duration_seconds                0 non-null      float64
15  video_height                          0 non-null      float64
16  video_id                              43 non-null     float64
17  video_quality                         0 non-null      float64
18  video_size_MB                         0 non-null      float64
19  video_thumbnail                       43 non-null     object
...
49  was_live                             1000 non-null   bool
50  fetched_time                          239 non-null    object
dtypes: bool(3), float64(17), int64(7), object(24)
memory usage: 378.1+ KB
```

Hình 2.2. Mô tả dữ liệu thu thập được

Qua mô tả có thể thấy bộ dữ liệu có đến 50 trường với số lượng bản ghi lên đến 1000. Tuy nhiên, do khiếm khuyết của facebook_scrapper nên một số trường sẽ bị “nghèo” giá trị cũng như chứa các giá trị rỗng. Hơn nữa, trong bài phân tích này, em sẽ không sử dụng toàn bộ 50 trường dữ liệu này mà chỉ lấy các trường cần thiết cho việc phân tích cũng như biểu diễn dữ liệu.

2.2. Làm sạch và tiền xử lí dữ liệu thô

2.2.1. Loại bỏ các trường dữ liệu không sử dụng

Với số lượng trường lớn và để việc phân tích trở nên thuận lợi, ta phải thực hiện xoá bớt các hàng dữ liệu không sử dụng. Tên các trường bị xoá đi gồm có: "post_text", "image", "images_lowquality", "post_url", "timestamp", "username", "likes", "shares", "images_description", "images_lowquality_description", "shared_text", "original_text", "video_size_MB", "image_lowquality", "images", "video", "video_duration_seconds", "video_height", "video_id", "video_quality", "video_thumbnail", "video_watches", "video_width", "link", "links", "user_url", "is_live", "factcheck", "shared_post_id", "shared_time", "shared_user_id", "shared_username", "shared_post_url", "available", "reactors", "w3_fb_url", "with", "page_id", "sharers", "image_id", "user_id", "image_ids", "was_live", "fetched_time". Sau khi xoá đi các trường này, bộ dữ liệu mới gồm có:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   post_id               1000 non-null   int64
1   text                  953 non-null    object
2   time                  1000 non-null   object
3   comments              1000 non-null   int64
4   comments_full         1000 non-null   object
5   reactions             239 non-null    object
6   reaction_count        1000 non-null   int64
dtypes: int64(3), object(4)
memory usage: 54.8+ KB
```

Hình 2.3: Dữ liệu sau khi xoá đi các hàng không cần thiết

Trong các trường dữ liệu, trường “comments” với kiểu dữ liệu int64 thể hiện số bình luận có trong bài đăng “post_id”. Vì lí do facebook_scrapper không thể thu thập dữ liệu hoàn hảo, nên ta phải loại đi các bài đăng không có bình luận. Thực hiện tương tự với các bài đăng có trường “reaction_count” bằng không, hay không có tương tác. Đồng thời tạo ra một dataframe mới “clean_df” để lưu dữ liệu đã thay đổi.

```
# clear rows that 0 comments, 0 reaction_count
clean_df = clean_df[clean_df["comments"] != 0]
clean_df = clean_df['reaction_count' != 0]
clean_df.info()
```

```
[7]
... <class 'pandas.core.frame.DataFrame'>
Int64Index: 237 entries, 0 to 988
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   post_id               237 non-null   int64
1   text                  237 non-null   object
2   time                  237 non-null   object
3   comments              237 non-null   int64
4   comments_full         237 non-null   object
5   reactions             237 non-null   object
6   reaction_count        237 non-null   int64
dtypes: int64(3), object(4)
memory usage: 14.8+ KB
```

Hình 2.4. Dữ liệu sau khi xoá các bài đăng không tương tác

Có thể thấy, sau khi thực hiện xoá thì có khá nhiều cột đã bị lược bỏ. Dữ liệu lúc này đã một phần đảm bảo để thực hiện các bước tiếp theo.

2.2.2. Tái tổ chức bộ dữ liệu

Để có thể dễ dàng thao tác và phân tích, ta nên tổ chức lại bộ dữ liệu trên. Để có thể làm điều này, ta làm các thao tác sau:

- Từ trường “comments_full” có dạng danh sách các từ điển chỉ thông tin cụ thể về bình luận của bài đăng, thực hiện chọn lọc và lấy ra một vài thuộc tính chính để tạo nên một dataframe mới là “Comment_frame”.

	comment_id	commenter_id	comment_text	post_id
0	659041829740311	100004523313287	Chả bù mình đi lấy chồng ngủ đến 11h trưa. Từ ...	731043432389703
1	1058980365296051	100004428425175	Đa số mà bố mẹ ngăn cản thì là bố mẹ đúng đấy	731043432389703
2	1049705363008340	100005874492328	Đúng là lấy chồng như canh bạc	731043432389703
3	847616997155316	100000337745111	Hải Phong tốt nhất là lấy chồng cũng mê ngủ gi...	731043432389703
4	3561634410773586	1251407235	Cãi ba mẹ lấy cho đã rồi lúc về lại nhà ba mẹ ...	731043432389703
...
3622	3405980696397936	100035876087877	Mặc quần đùi đi đá bóng có coi là lộ chân ko nhỉ?	6563791543670813
3623	1345872649536668	100017682734580	cái post này mà cmt ảnh được là phần cmt toàn ...	6563791543670813
3624	1363567874435450	100012723275222	Này là chân, còn của tui là cột đình👀	6563791543670813
3625	752935386491017	100045700756125	Trang Le chân kia nhỏ hơn bắp tay mình luôn (=)))	6563791543670813
3626	772792927414635	100009969248642	Lê Ngọc Huyền thon	6563791543670813

3627 rows × 4 columns

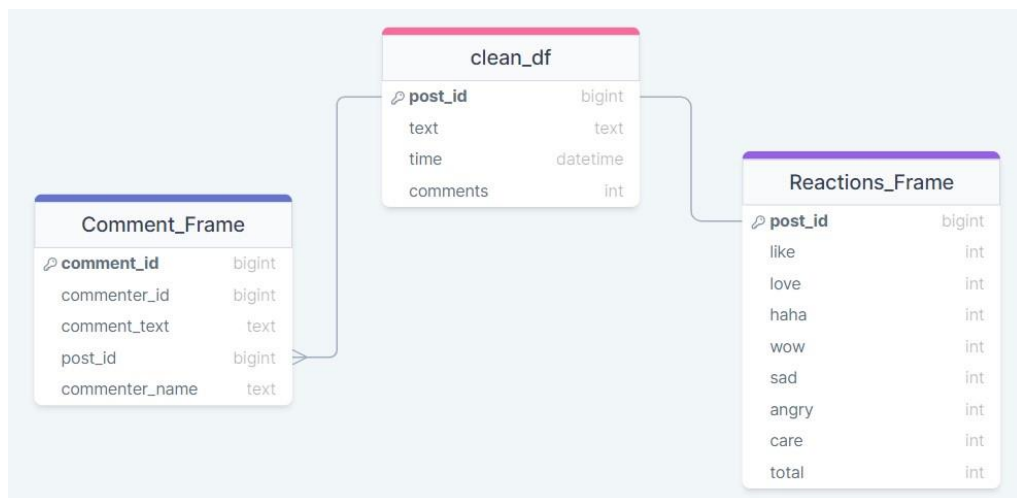
Hình 2.5. Dataframe “Comment_frame”

- Tương tự, từ trường “reactions” ta thực hiện tạo ra dataframe mới là “Reactions_Frame” lưu trữ số lượng tương tác với các bài viết.

	like	love	haha	wow	care	sad	angry	post_id	total
0	3003	6	21	13	4	1089	19	731043432389703	4155
1	699	3	564	6	1	1	1	731015475725832	1275
2	2846	737	21	2	14	0	1	730966445730735	3621
3	256	5	164	0	1	0	1	730945852399461	427
4	1014	7	1271	0	1	3	1	730934009067312	2297

Hình 2.6. Dataframe “Reactions_Frame”

Sau khi tổ chức lại, ta sẽ có mô hình dữ liệu như sau:



Hình 2.7. Cấu trúc tổ chức của dữ liệu

Cụ thể, thông tin của các trường trong ba dataframe như sau:

- Dataframe “clean_df”:

Tên trường	Kiểu dữ liệu	Mô tả
post_id	Int64	Id của bài viết
text	String	Nội dung của bài viết
time	Datetime	Thời điểm đăng bài
comments	Int64	Số lượng bình luận bài viết

Bảng 2.1. Cấu trúc dataframe “clean_df”

- Dataframe “Comments_Frame”:

Tên trường	Kiểu dữ liệu	Mô tả
comment_id	Int64	Id của bình luận
commenter_id	Int64	Id của người bình luận
commenter_name	String	Tên tài khoản bình luận
comment_text	String	Nội dung bình luận
post_id	Int64	Bài viết chứa bình luận

Bảng 2.2. Cấu trúc dataframe “Comments_Frame”

- Dataframe “Reactions_Frame”:

Tên trường	Kiểu dữ liệu	Mô tả
like	Int64	Số lượt thích
love	Int64	Số lượt yêu thích
haha	Int64	Số lượt haha
wow	Int64	Số lượt wow
care	Int64	Số lượt thương thương
sad	Int64	Số lượt buồn
angry	Int64	Số lượt phẫn nộ
post_id	Int64	Id của bài viết
total	Int64	Tổng số lượt tương tác

Bảng 2.3. Cấu trúc dataframe “Reactions_Frame”

2.2.3. Tiếp tục làm sạch và tiền xử lý dữ liệu

Ta thực hiện hiệu chỉnh các giá trị rỗng của các bản ghi ở các bảng, thay thế các giá trị NaN bằng giá trị hợp lệ. Đối với văn bản, loại bỏ các kí tự ‘thừa’ như xuống hàng, hàng rỗng và các Emoji có trong chuỗi. Vì các loại kí tự này không những thể hiện quá

nhiều ý nghĩa mà còn gây nhập nhằng sau này. Cuối cùng, ta thực hiện lưu lại các dataframe vừa làm sạch vào ba file csv để thực hiện việc phân tích dữ liệu.

Save the dataframe

```
# save to csv
clean_df.to_csv("Data/clean_df.csv", index=False)
Comment_frame.to_csv("Data/Comment_frame.csv", index=False)
Reactions_Frame.to_csv("Data/Reactions_Frame.csv", index=False)
```

Hình 2.8. Lưu trữ dataframe vào file csv

CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU

3.1. Cài đặt thư viện và tải lên dữ liệu

Sau khi dữ liệu đã được làm sạch, ta bắt đầu đi phân tích về dữ liệu đã thu thập. Đầu tiên tạo một Jupiter Notebook mới “Analytics_Data.ipynb” để thực hiện phân tích và cài đặt các thư viện cần thiết, sau đó gọi lại các dataframe đã được lưu.

Import required Library

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import json
import re
import datetime
import seaborn as sns
import plotly.express as px
from wordcloud import WordCloud
import matplotlib.ticker as ticker
import squarify
```

Hình 3.1. Các thư viện sử dụng

```
# read clean_df
clean_df = pd.read_csv('Data/clean_df.csv')
# read Comment_frame
Comment_frame = pd.read_csv('Data/Comment_frame.csv')
# read Reactions_frame
Reactions_frame = pd.read_csv('Data/Reactions_Frame.csv')
```

Hình 3.2. Đọc file lưu dữ liệu đã làm sạch

3.2. Phân tích dữ liệu đã thu thập

3.2.1. Phân tích tương tác của trang

Nhìn chung, Weibo Việt Nam là một trang hoạt động khá đều đặn và thường xuyên trong vòng nhiều năm trở về đây. Đây cũng chính là một trong những lí do khiến cho trang có sức ảnh hưởng và được nhiều người biết đến. Thực hiện phân tích thời gian và số lượng của các bài đăng có trong dataframe, ta thu được kết quả:

This dataset contains data from 2023-04-03 15:55:49 to 2023-11-29 23:58:05, with total 237 posts.

Hình 3.3. Khoảng thời gian và số lượng bài đăng

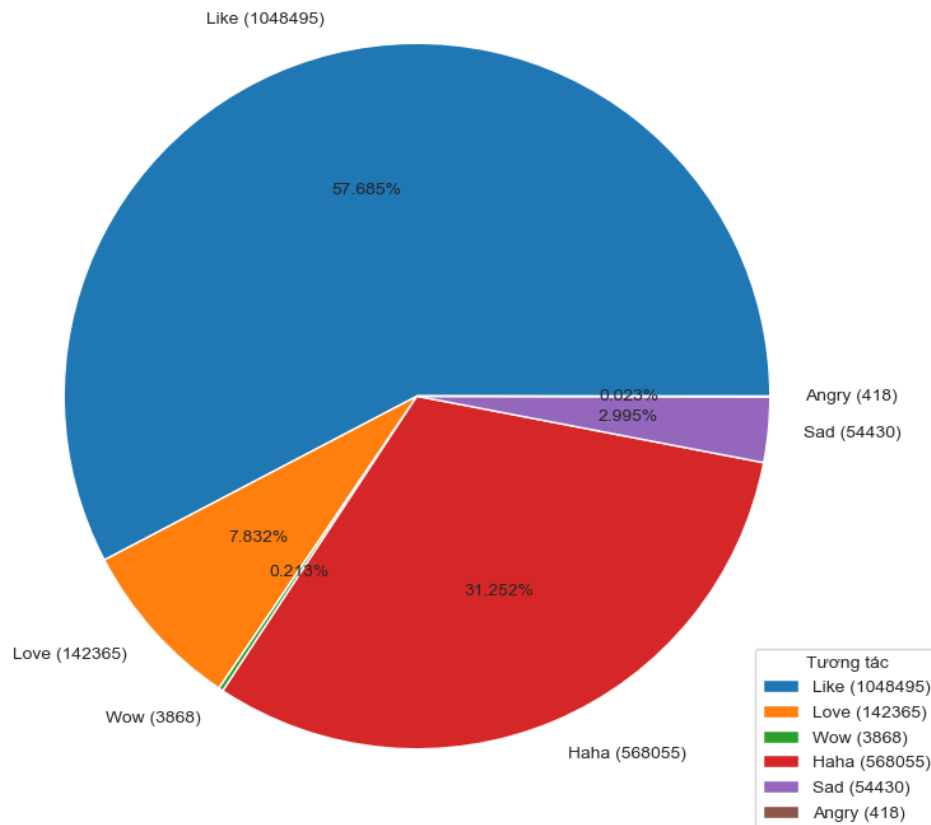
```
The number of posts per month:
11      174
4         42
10        21
Name: month, dtype: int64, all are in 2023.
```

Hình 3.4. Số lượng bài đăng theo các tháng

Ta có thể thấy các bài đăng được phân tích của trang sẽ chủ yếu tập trung vào ba tháng là tháng Tư, tháng Mười và tháng Mười Một năm 2023 của trang. Các bài đăng của tháng khác có thể đã bị xoá đi khỏi dữ liệu trong lúc xử lí.

Để phân tích lượt tương tác của người dùng đối với trang, ta sử dụng biểu đồ tròn để biểu thị phần trăm các loại tương tác với các bài viết như sau:

Biểu đồ: Tỷ lệ các loại tương tác bài viết

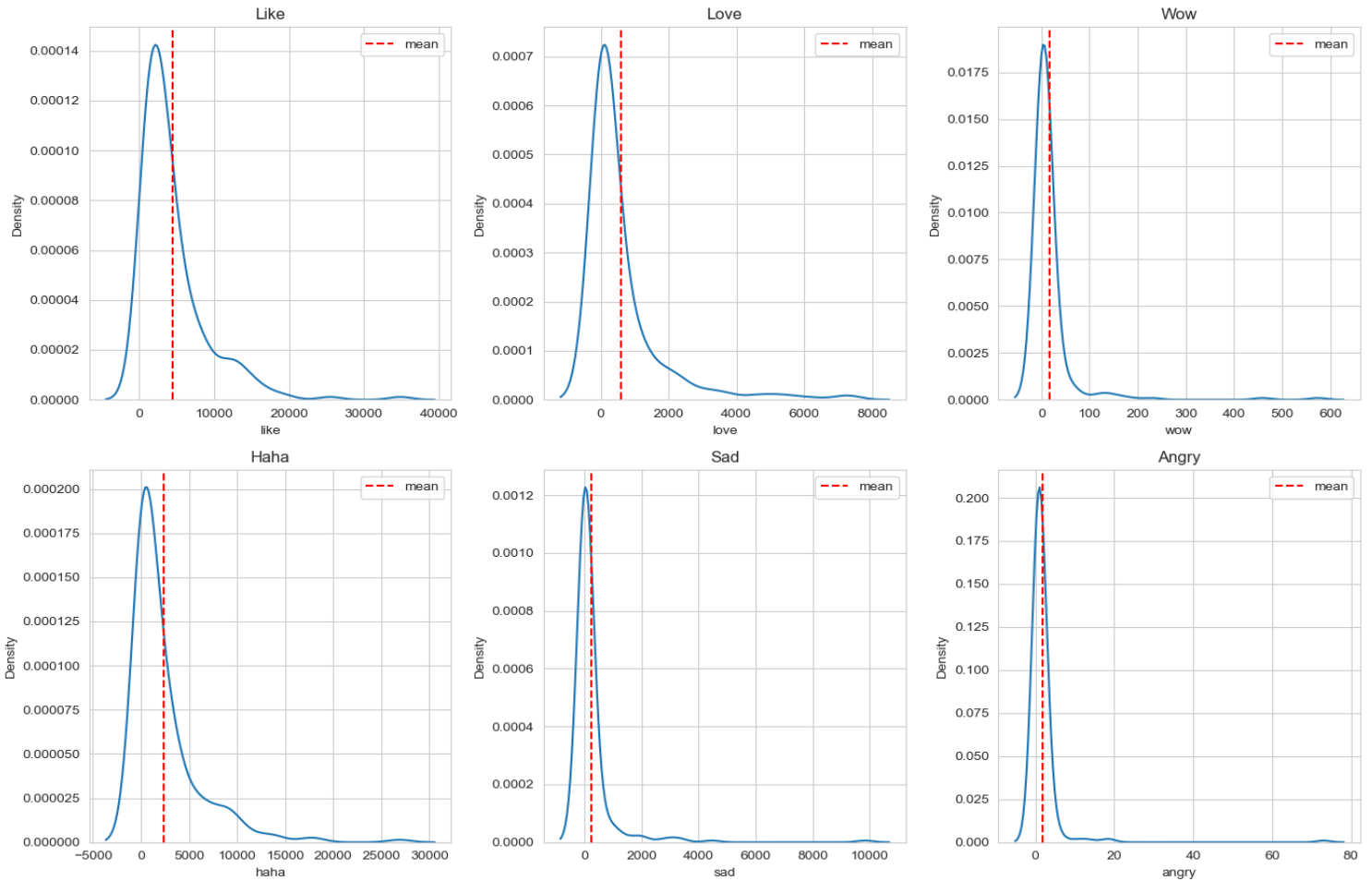


Biểu đồ 3.1. Tỷ lệ các loại tương tác bài viết

Qua biểu đồ có thể thấy người dùng có xu hướng “thả tương tác” like, haha và love là chủ yếu, trong đó lượng tương tác lớn nhất đến từ nút ‘like’. Nó cũng phản ánh đúng về tính chất của trang Weibo Việt Nam là hướng đến sự thông tin và giải trí, và cũng có một lượng ít tương tác đến từ ‘sad’, ‘wow’ và angry. Có vẻ trang không thường đưa những bài ‘giật gân’ hay ‘cảm động’.

Với mỗi loại tương tác thì sẽ có những loại phân bố về giá trị như sau:

Biểu đồ: Phân bố cụ thể của các loại tương tác



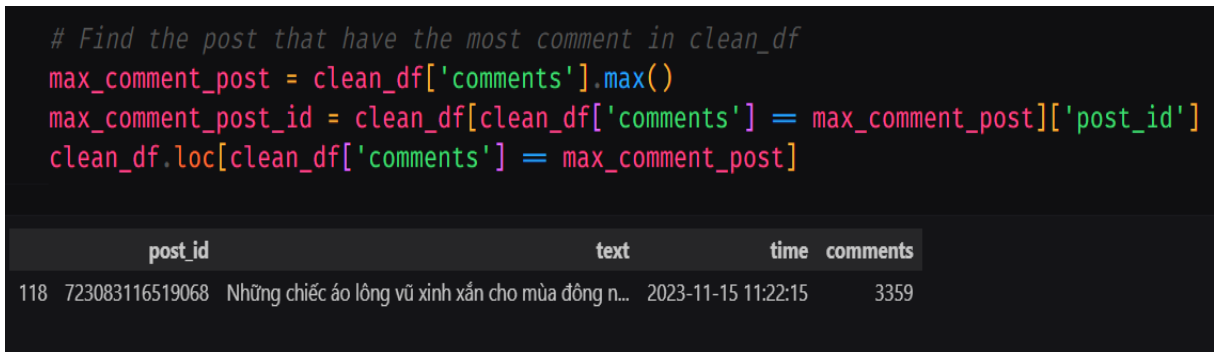
Biểu đồ 3.2. Biểu đồ phân bố của các loại tương tác

Thông qua thể hiện phân bố trên biểu đồ **kernel density estimation (KDE)**, lượng tương tác phân bố khá đẹp theo hình chuông và không lệch quá đáng kể quanh trung bình của nó. Mặc dù vậy, vẫn xuất hiện các giá trị ngoại vi trong các loại tương tác.

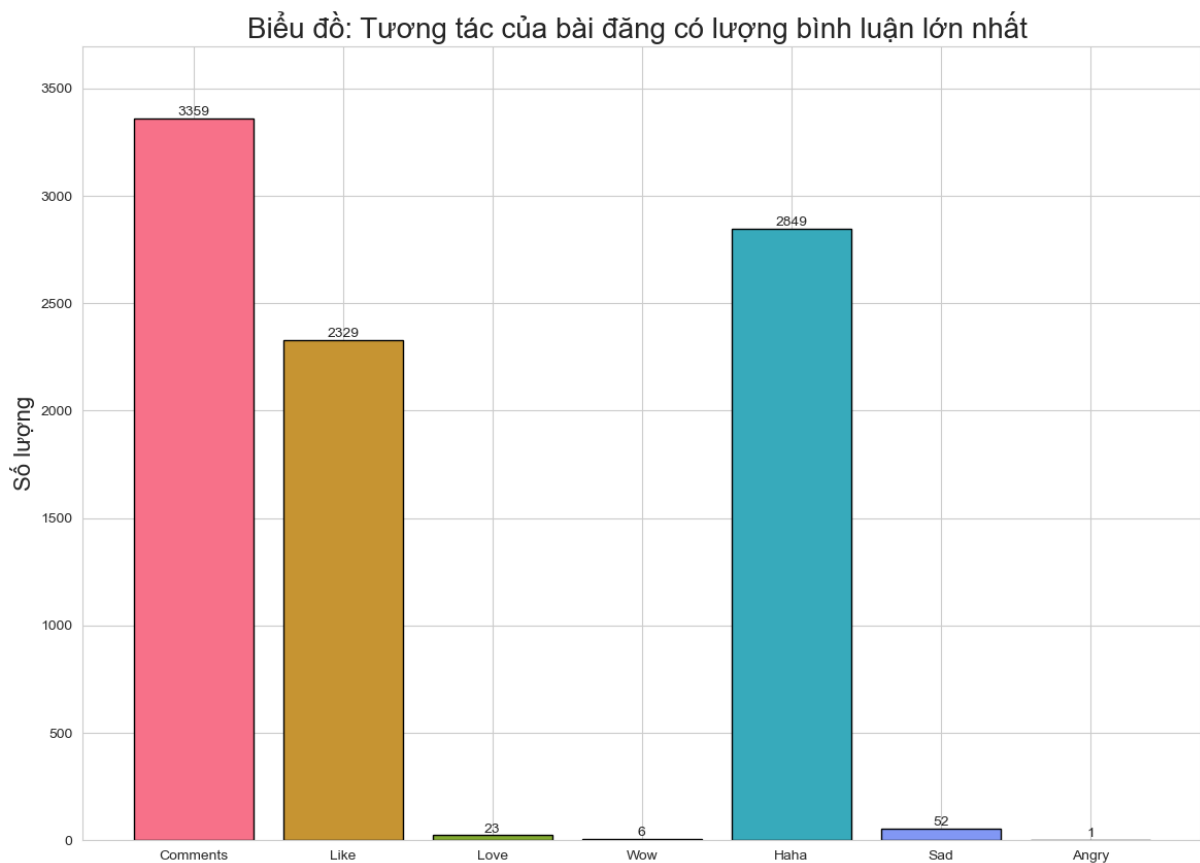
3.2.2. Phân tích các bài đăng, nội dung đặc biệt

3.3.2.1. Bài đăng có lượt bình luận lớn nhất

Để tìm bài đăng có lượt bình luận lớn nhất, ta tìm bài đăng có giá trị của trường 'comments' lớn nhất. Kết quả như sau:



Hình 3.5. Thông tin về bài đăng có lượt bình luận nhiều nhất



Biểu đồ 3.3. Tương tác của bài đăng có nhiều bình luận nhất

Từ biểu đồ có thể thấy rằng số lượt bình luận nhiều hơn bất kỳ một loại tương tác nào khác, trong đó lượt like và haha chiếm đa số. Đây có thể là một bài đăng có nội dung hài hước.

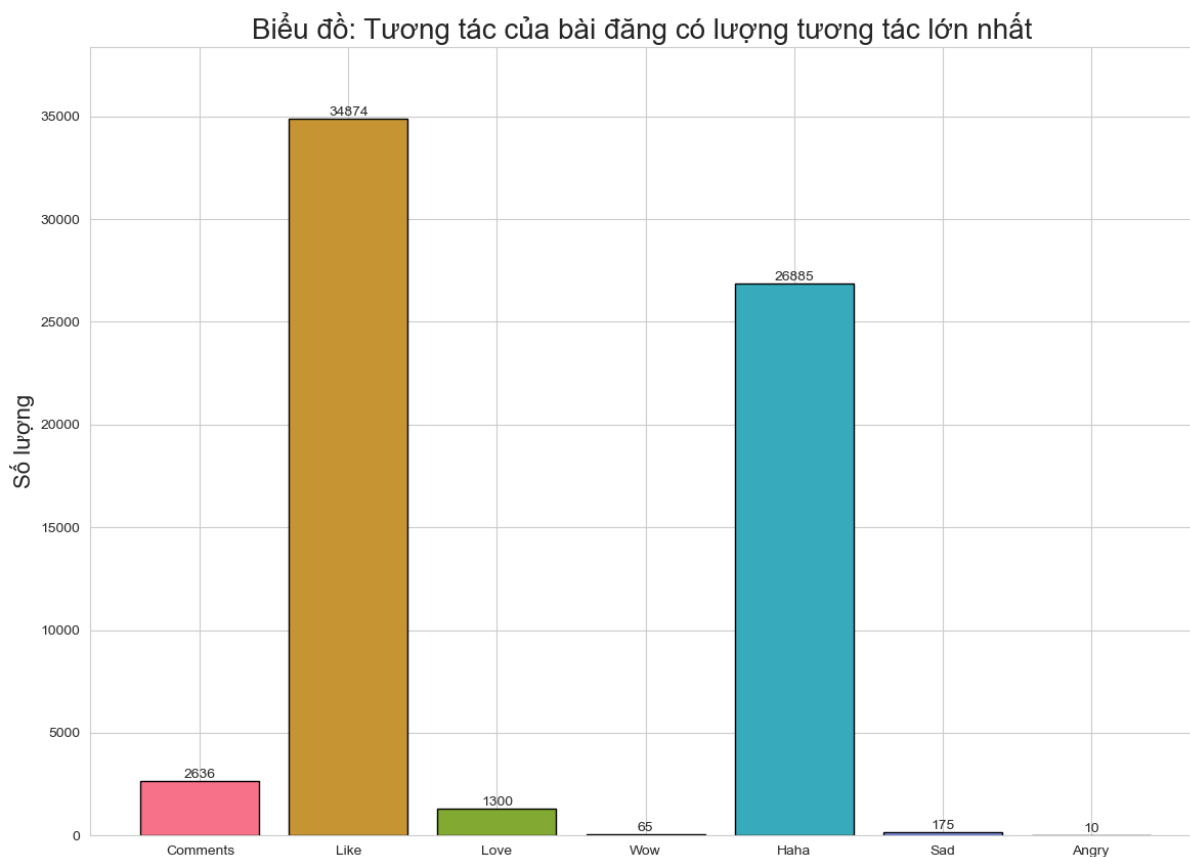
3.3.2.2. Bài đăng có tổng lượt tương tác lớn nhất

Để tìm bài đăng có tổng lượt tương tác lớn nhất, ta xác định bài đăng có trường ‘total’ lớn nhất.

```
# Find the post that have the most reactions in clean_df by total in Reactions_frame
max_reactions_post = Reactions_frame['total'].max()
max_reactions_post_id = Reactions_frame.loc[Reactions_frame['total'] == max_reactions_post]['post_id'].values[0]
clean_df.loc[clean_df['post_id'] == max_reactions_post_id]
```

	post_id	text	time	comments
175	715226890638024	Hôm nay tan ca, tôi về nhà khá sớm chứ không ở...	2023-10-31 21:25:11	2636

Hình 3.6. Thông tin về bài đăng có lượt tương tác lớn nhất



Biểu đồ 3.4. Tương tác của bài đăng có lượng tương tác nhiều nhất

Từ biểu đồ có thể thấy rằng sự vượt trội của hai tương tác like, haha so với lượt bình luận. Một lần nữa số lượng tương tác like và haha lại chiếm đa số trong các kiểu tương tác bài viết. Trên thực tế, đây là bài viết truyện mang yếu tố gây cười nên thu hút lượt tương tác lớn, kể cả số lượt bình luận nói riêng.

3.3.2.3. Những người dùng có lượt bình luận lớn nhất

Để tìm những người có lượt bình luận lớn nhất, ta tìm số lần xuất hiện của “commenter_id” nhiều nhất. Ta thu được năm tài khoản có lượt bình luận nhiều nhất như sau:


```

commenter_name
Phượng Thảo          18
Weibo Việt Nam        16
Nguyễn Kim Bài Vũ     13
Góc nhỏ của Tiểu Nguyệt 12
Giang Thảo Huyền      10
Name: commenter_name, dtype: int64

```

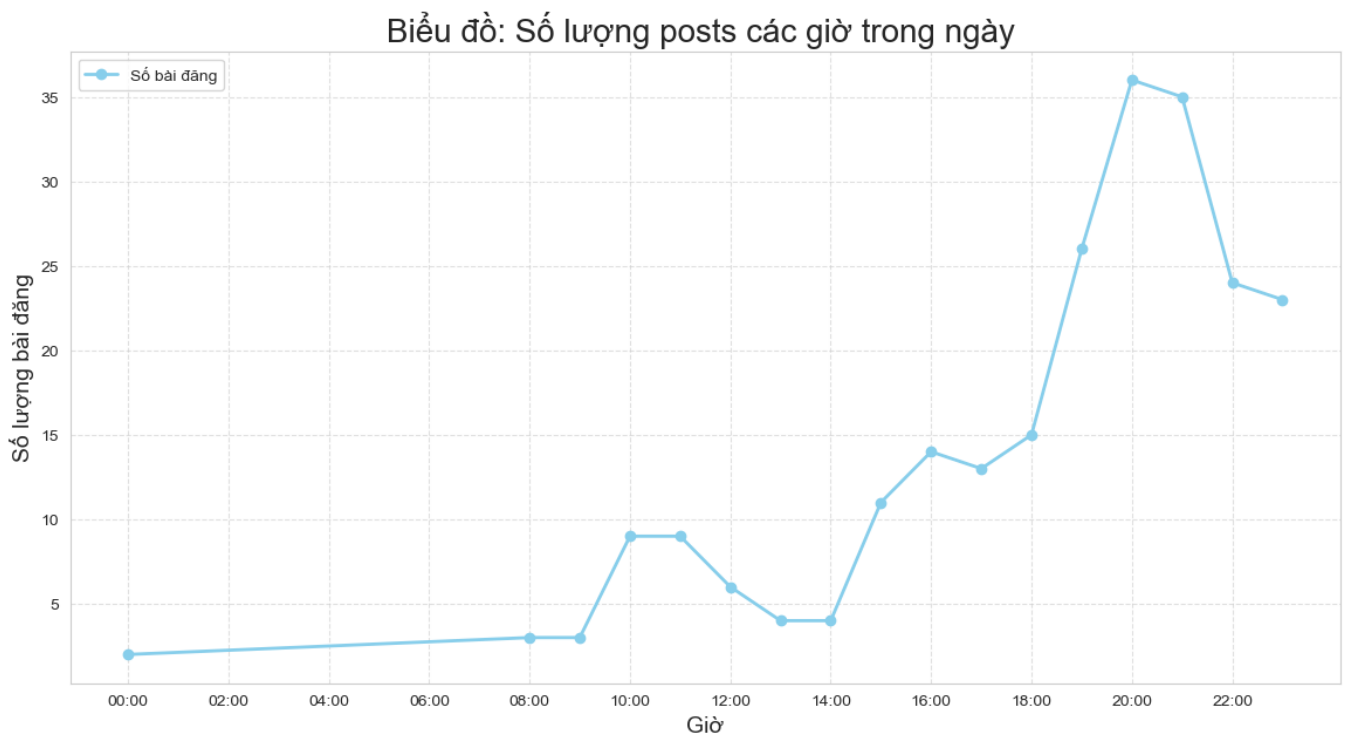
Hình 3.7. Những tài khoản bình luận nhiều nhất

Ngoài các tài khoản khác có thể là ‘fan cứng’ của trang, thì ta còn thấy trang tự bình luận vào những bài đăng của mình. Thường thì những bình luận như thế này sẽ được ghim để gây nổi bật trong phần bình luận.

3.2.3. Phân tích hoạt động của trang

Để phân tích sự hoạt động của trang, ta sẽ theo dõi sự hoạt động của trang qua thời gian đăng các bài viết lên theo những thước đo khác nhau.

Đối với thời gian hoạt động trong ngày, ta xem xét số lượng đăng bài hàng giờ của trang. Báo cáo về thời gian hoạt động được thể hiện qua biểu đồ đường dưới đây:

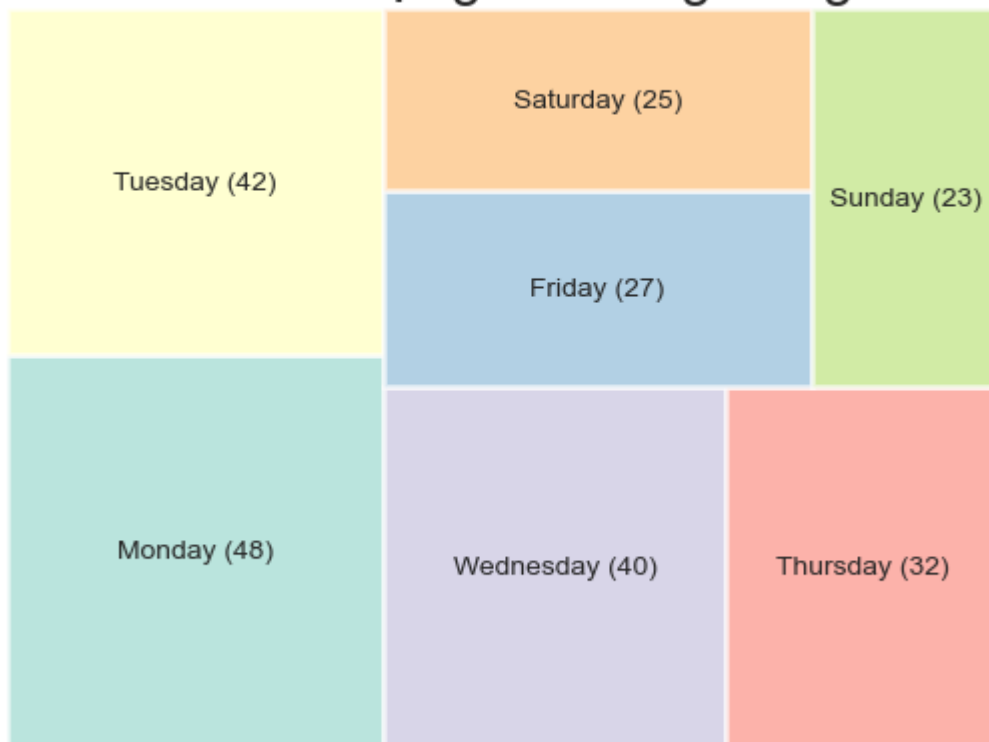


Biểu đồ 3.5. Số lượng bài đăng trong ngày theo giờ

Qua biểu đồ có thể thấy trang thường đăng bài bắt đầu từ giờ trưa và duy trì đến buổi tối. Thời gian từ 18:00 đến 24:00 là thời gian trang hoạt động tích cực nhất. Các bài đăng có lượt tương tác lớn nhất cũng được đăng trong khoảng thời gian này. Đây có thể là khoảng thời gian lí tưởng để trang đăng bài.

Đối với hoạt động các ngày trong tuần, ta sẽ biểu diễn số lượng bài đăng từ thứ Hai đến Chủ Nhật bằng bản đồ cây như sau:

Biểu đồ: Số lượng bài đăng trong tuần



Biểu đồ 3.6. Số lượng bài đăng các ngày trong tuần

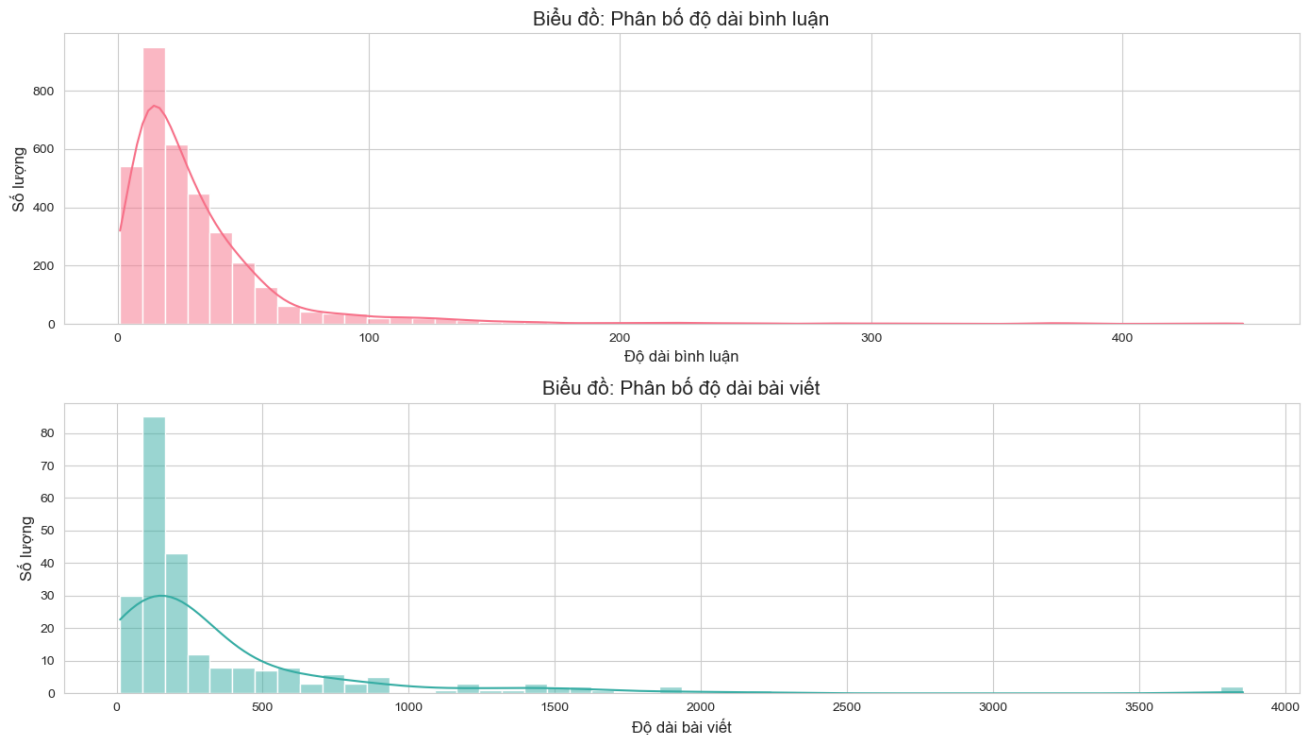
Qua biểu đồ có thể thấy trang hoạt động sôi nổi hơn vào các ngày đầu tuần, giảm dần về cuối tuần. Lí do có thể đến từ tần suất làm việc của đội ngũ quản trị trang.

Tóm lại, có thể thấy thời gian ‘cao điểm’ hoạt động của trang là tầm 18h đến 23h của ngày thứ Hai.

3.2.4. Một số phân tích về nội dung bài viết và bình luận

Weibo Việt Nam hướng tới trang có nội dung giải trí, thông tin đơn thuần nên khi đăng bài thường kết hợp việc kèm ảnh minh họa. Hiện nay, rất nhiều trang đã theo đuổi lối đăng bài chỉ đăng hình ảnh có nội dung được ghi lên trên ảnh. Đối với Weibo Việt Nam, có thể là do ảnh hưởng từ lối viết đầu đề dài đặc trưng của Weibo – Mạng xã hội lớn nhất tại Trung Quốc, trang này vẫn duy trì lối viết khá dài và vẫn kèm hình ảnh.

Sử dụng biểu đồ KDE, ta sẽ xem phân bố của độ dài các bài viết và độ dài của bình luận.



Biểu đồ 3.7. Phân bố độ dài của nội dung bài viết và bình luận

Ta có thể thấy rằng phân độ dài bình luận có phân bố khá đều, ở mức trung bình. Còn với phần nội dung các bài đăng, có một số bài có dung lượng khá dài nhưng nhìn chung là giữ ở mức ổn định.

Đối với các bài đăng, một số từ sẽ có xu hướng xuất hiện nhiều ở trong nội dung. Điều tương tự cũng có thể xảy ra với phần bình luận. Để xác định điều đó, ta có thể biểu diễn Wordcloud của phần nội dung cũng như phần bình luận của bài viết.

[illegible]

Wordcloud của bình luận



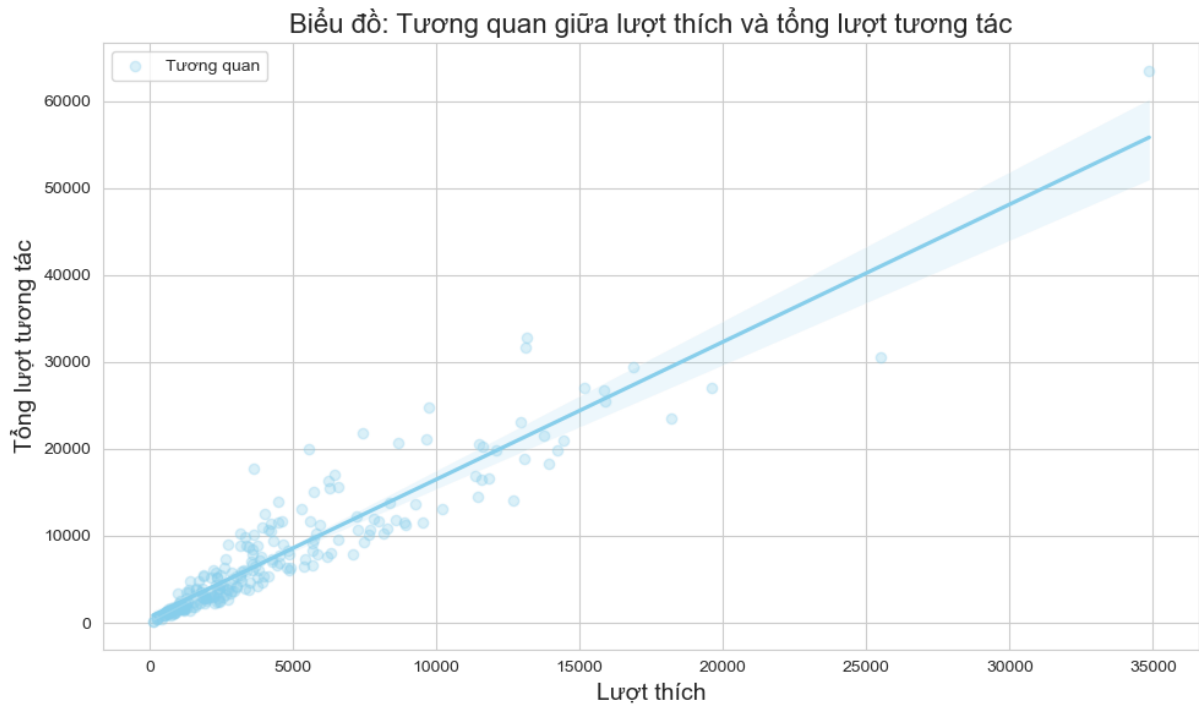
Có thể thấy rằng, đối với phần nội dung bài viết, ta có thể thấy rằng các từ thường xuất hiện (có kích cỡ lớn) là các từ như: không, tôi, là, của,...Những từ này thường xuất hiện trong văn nói và văn viết của tiếng Việt. Ngoài ra còn có sự xuất hiện của tagname WeiboVietNam, đây có thể là phần tác giả của bài viết.

Còn về phần nội dung bình luận, ta thấy rằng những từ ngữ nổi bật xuất hiện là những Tên Họ của Việt Nam. Vậy có thể kết luận rằng nhiều người tham gia bình luận

bằng cách tag tên của người khác vào, giúp người đợng tag có thể tiếp cận được với nội dung bài viết. Và kiểu bình luận như thế này cũng khá phổ biến.

3.2.3. Một số mối liên hệ giữa các trường dữ liệu

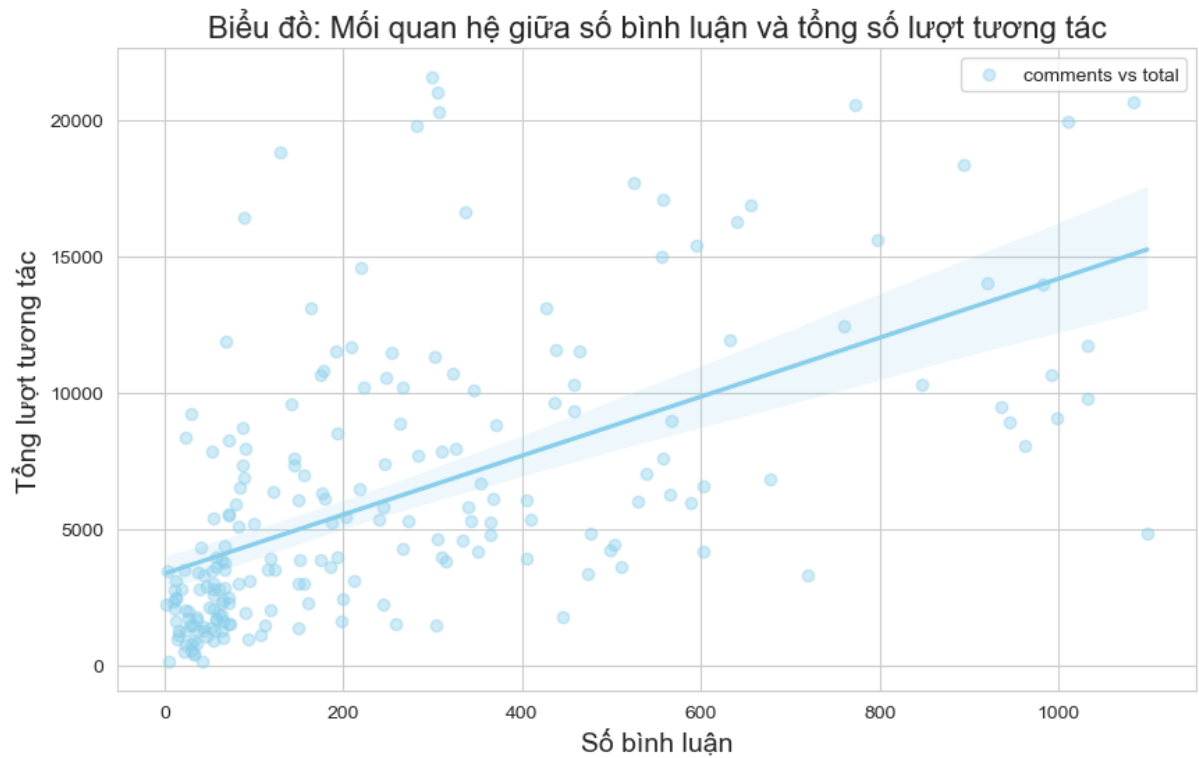
Theo thống kê ở trên, số lượt like chiếm số đông tổng lượt tương tác, vậy có thể nói rằng: Có mối liên hệ giữa số lượt like và tổng lượt tương tác hay không?. Để trả lời, ta thực hiện biểu diễn mối quan hệ bằng biểu đồ điểm và vẽ đường hồi quy tuyến tính của đồ thị.



Biểu đồ 3.10. Tương quan giữa lượt thích và tổng lượt tương tác

Có thể thấy rằng, việc có càng nhiều lượt thích sẽ làm tăng tổng lượt tương tác.

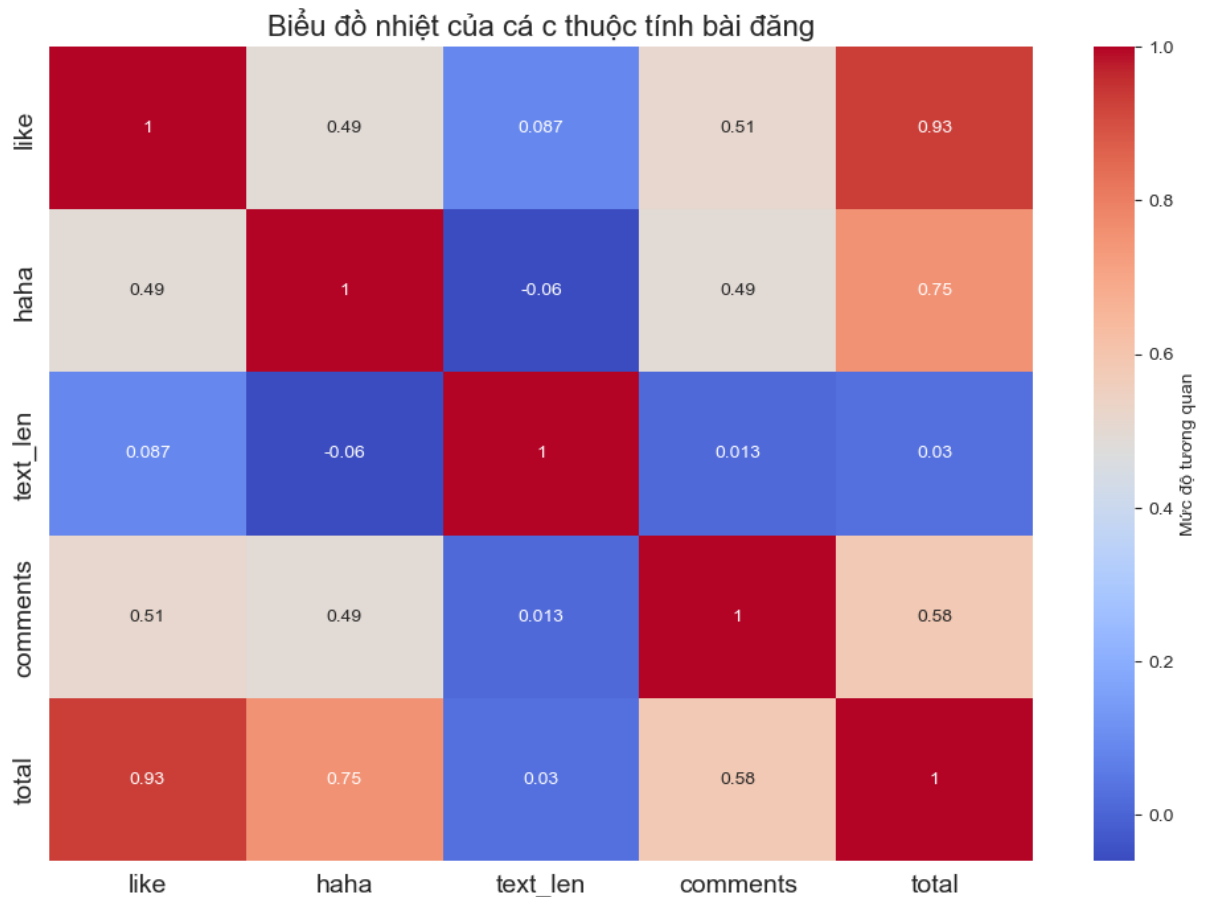
Ta sẽ xét sự tương quan giữa lượt bình luận và lượt tương tác như sau:



Biểu đồ 3.11. Tương quan giữa lượt tương tác và lượng bình luận

Quan đồ thị có thể thấy, sự tương quan giữa số lượng tương tác và bình luận là không quá rõ nét. Vậy nên chưa thể kết luận rằng: Càng nhiều lượt tương tác sẽ có càng nhiều lượt bình luận.

Một cách tổng quát, ta sẽ mô tả sự tương quan của các trường dữ liệu thông qua biểu đồ nhiệt dưới đây:



Biểu đồ 3.12. Tương quan giữa các trường dữ liệu

3.3. Một số kết luận

Sau khi thực hiện phân tích dữ liệu thu được, ta có thể đưa ra một vài kết luận như sau:

- Trang Weibo Việt Nam là một trang hoạt động và có ảnh hưởng khá tích cực trên mạng xã hội.
- Trang có một số bài đăng có lượng tương tác đột biến, nhìn chung trang duy trì được lượt tương tác khá ổn định.
- Thời gian hoạt động thường xuyên là đầu tuần vào các buổi tối, giảm dần về cuối tuần và buổi sáng.
- Trang tập trung vào các nội dung giải trí hoặc kể những câu chuyện
- Phần bình luận của trang cũng thể hiện một phần thói quen bình luận là tag tên người khác vào bài viết.
- Một số mối liên hệ giữa nội dung, độ dài bài viết và lượt tương tác mặc dù là có nhưng chưa quá rõ ràng, việc này yêu cầu cần có nhiều dữ liệu cũng như các mô hình phức tạp hơn.