

UỶ BAN NHÂN DÂN TP. HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

❑❑❑❑❑

BÁO CÁO CUỐI KỲ

MÔN HỌC:

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG (841447)

ĐỀ TÀI:

ỨNG DỤNG THUẬT TOÁN SVM PHÂN LOẠI CẢM XÚC DỰA TRÊN NỘI DUNG BÀI ĐĂNG TRÊN MXH

**Giảng** **viên:** ThS. Nguyễn Thanh Phước.

**Lớp** **:** DCT1223

**Thực hiện:** Trần Ngô Nhật Nam - 3122410253

*TPHCM, ngày 10 tháng 4 năm 2025*

***Mục lục***

[**Phần I: Giới thiệu** 5](#_Toc195392564)

[**1.** **Lý do chọn đề tài** 5](#_Toc195392565)

[**2.** **Mục đích của đề tài** 5](#_Toc195392566)

[**3.** **Phạm vi của đề tài** 5](#_Toc195392567)

[**4.** **Các bài toán tương tự trong thực tế** 5](#_Toc195392568)

[**Phần II: Mô tả dữ liệu** 6](#_Toc195392569)

[**1.** **Giới thiệu về bộ dữ liệu** 6](#_Toc195392570)

[**2.** **Cấu trúc bộ dữ liệu** 6](#_Toc195392571)

[**Phần III: Xử lý dữ liệu** 6](#_Toc195392572)

[**1.** **Chia lại bộ dữ liệu** 6](#_Toc195392573)

[**2.** **Xử lý dữ liệu thiếu** 9](#_Toc195392574)

[**3.** **Làm sạch và định dạng dữ liệu** 9](#_Toc195392575)

[**Phần IV: Phân tích dữ liệu** 11](#_Toc195392576)

[**1.** **Phân tích đơn biến** 11](#_Toc195392577)

[**1.1.** **Phân phối cảm xúc (Sentiment):** 11](#_Toc195392578)

[**1.2.** **Phân phối theo ngày, giờ (Timestamp):** 12](#_Toc195392579)

[**1.3.** **Phân phối theo nền tảng mạng xã hội (Platform):** 13](#_Toc195392580)

[**1.4.** **Phân phối theo quốc gia (Country):** 14](#_Toc195392581)

[**1.5.** **Phân phối theo người dùng (User):** 15](#_Toc195392582)

[**1.6.** **Phân phối theo lượt tương tác (Like/Retwwets):** 15](#_Toc195392583)

[**2.** **Phân tích đa biến** 17](#_Toc195392584)

[**3.** **Phân tích tương quan** 17](#_Toc195392585)

[**3.1.** **Phân phối theo cảm xúc/quốc gia (Sentiment/Country):** 17](#_Toc195392586)

[**3.2.** **Phân phối theo cảm xúc/thời gian (Sentiment/Timestamp):** 18](#_Toc195392587)

[**3.3.** **Phân phối theo cảm xúc/nền tảng MXH (Sentiment/Platform):** 20](#_Toc195392588)

[**3.4.** **Phân phối theo cảm xúc/lượt thích (Sentiment/Likes):** 21](#_Toc195392589)

[**3.5.** **Phân phối theo cảm xúc/lượt thích (Sentiment/Retweets):** 22](#_Toc195392590)

[**3.6.** **Phân phối theo lượt thích/lượt đăng lại (Likes/Retweets):** 23](#_Toc195392591)

[**Phần V: Khai phá dữ liệu** 23](#_Toc195392592)

[**1.** **Đánh giá tổng thể dữ liệu** 23](#_Toc195392593)

[**1.1.** **Kích thước và cấu trúc dữ liệu** 23](#_Toc195392594)

[**1.2.** **Chất lượng dữ liệu:** 24](#_Toc195392595)

[**1.3.** **Đặc điểm của dữ liệu văn bản:** 24](#_Toc195392596)

[**1.4.** **Các bước tiền xử lý đã thực hiện:** 26](#_Toc195392597)

[**1.5.** **Đánh giá tổng quan:** 26](#_Toc195392598)

[**Phần VI: Đánh giá và chọn thuật toán** 27](#_Toc195392599)

[**1.** **Các mô hình được áp dụng** 27](#_Toc195392600)

[**1.1.** **Logistic Regression** 27](#_Toc195392601)

[**1.2.** **Naive Bayes** 27](#_Toc195392602)

[**1.3.** **Support Vector Machine (SVM)** 27](#_Toc195392603)

[**1.4.** **XGBoost** 28](#_Toc195392604)

[**2.** **Các tiêu chí đánh giá** 28](#_Toc195392605)

[**3.** **Kết quả so sánh các mô hình** 28](#_Toc195392606)

[**Phần VII: Kết quả và thảo luận** 29](#_Toc195392607)

[**1.** **Kết quả trên tập kiểm tra (Test set)** 29](#_Toc195392608)

[**2.** **Thảo luận** 30](#_Toc195392609)

[**3.** **Nhận xét tổng quan** 30](#_Toc195392610)

[**Phần VIII: kết luận** 30](#_Toc195392611)

[**1.** **Kết luận sơ bộ** 30](#_Toc195392612)

[**2.** **Kết luận tổng quát** 31](#_Toc195392613)

[**3.** **Tài liệu tham khảo** 31](#_Toc195392614)

**Mục lục hình ảnh**

[**Hình 1: Kết quả kiểm tra dữ liệu thiếu 9**](#_Toc195389048)

[**Hình 2: Biểu đồ phân phôi bài đăng theo nền tảng MXH trước khi làm sạch dữ liệu 10**](#_Toc195389049)

[**Hình 3: Biểu đồ phân phôi bài đăng theo nền tảng MXH sau khi làm sạch dữ liệu 10**](#_Toc195389050)

[**Hình 4: Dữ liệu trước khi được xử lý 11**](#_Toc195389051)

[**Hình 5:Dữ liệu sau khi được xử lý 11**](#_Toc195389052)

[**Hình 6: Biểu đồ số lượng bài đăng theo từng loại cảm xúc 12**](#_Toc195389053)

[**Hình 7: Biểu đồ số lượng bài đăng theo các ngày trong tuần 12**](#_Toc195389054)

[**Hình 8: Biểu đồ số lượng bài đăng theo từng khung giờ trong ngày 13**](#_Toc195389055)

[**Hình 9: Biểu đồ số lượng bài đăng theo nền tảng MXH 14**](#_Toc195389056)

[**Hình 10: Biểu đồ số lượng bài đăng theo từng quốc gia 14**](#_Toc195389057)

[**Hình 11: Biểu đồ top 10 người dùng tích cực đăng bài nhất 15**](#_Toc195389058)

[**Hình 12: Top 10 bài viết có ít/nhiều lượt thích nhất 16**](#_Toc195389059)

[**Hình 13: Top 10 bài viết có ít/nhiều lượt đăng lại nhất 16**](#_Toc195389060)

[**Hình 14: Ma trận tương quan Likes, Retweets, Month, Hour, Day 17**](#_Toc195389061)

[**Hình 15: Biểu đồ top 10 quốc gia có nhiều bài đăng nhất 18**](#_Toc195389062)

[**Hình 16: Biểu đồ tỷ lệ cảm xúc trong các bài đăng tại top 10 quốc gia 18**](#_Toc195389063)

[**Hình 17: Biểu đồ số lượng bài đăng theo các ngày trong tuần 19**](#_Toc195389064)

[**Hình 18: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo các ngày trong tuần 19**](#_Toc195389065)

[**Hình 19: Biểu đồ số lượn bài đăng theo các khung giờ trong ngày 20**](#_Toc195389066)

[**Hình 20: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo giờ trong ngày 20**](#_Toc195389067)

[**Hình 21: Biểu đồ phân phối bài đăng theo nền tảng MXH 21**](#_Toc195389068)

[**Hình 22: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo nền tảng MXH 21**](#_Toc195389069)

[**Hình 23: Biểu đồ lượt tương tác trung bình Likes theo cảm xúc 22**](#_Toc195389070)

[**Hình 24: Biểu đồ lượt tương tác trung bình Retweets theo cảm xúc 22**](#_Toc195389071)

[**Hình 25: Biểu đồ mối quan hệ giữa Likes và Retweets 23**](#_Toc195389072)

***Mục lục bảng biểu***

**Bảng 1: Các thuộc tính của bộ dữ liệu 5**

**Bảng 2: Kết quả so sánh các thuật toán 25**

**Bảng 3: Kết quả trên tập kiểm tra (Test set) 25**

**Bảng 4: Ma trận nhầm lẫn SVM 26**

# **Phần I: Giới thiệu**

## **Lý do chọn đề tài**

Trong thời đại số hóa hiện nay, mạng xã hội đóng vai trò ngày càng quan trọng trong việc phản ánh quan điểm, cảm xúc và hành vi của người dùng. Việc phân tích cảm xúc từ các bài đăng trên mạng xã hội không chỉ giúp các doanh nghiệp nắm bắt xu hướng người tiêu dùng mà còn hỗ trợ trong việc phát hiện khủng hoảng truyền thông, cải thiện dịch vụ khách hàng, và đưa ra chiến lược tiếp thị hiệu quả.

## **Mục đích của đề tài**

Dự án này nhằm mục tiêu **phân tích và khai phá dữ liệu cảm xúc người dùng trên mạng xã hội**, thông qua việc áp dụng các kỹ thuật tiền xử lý, trích xuất đặc trưng văn bản (như **TF-IDF**), và mô hình học máy (như **XGBoost**) để phân loại cảm xúc của bài đăng thành các nhóm như **tích cực(positive)**, **tiêu cực(negative)** hoặc **trung tính(neutral)**. Từ đó có thể kiểm duyệt lượng lớn bài viết trên mạng xã hội mà không cần phải xem xét từng bài một các thủ công.

## **Phạm vi của đề tài**

Phạm vi của dự án tập trung vào việc xử lý dữ liệu văn bản, trích xuất đặc trưng, huấn luyện mô hình và đánh giá hiệu suất mô hình trong bài toán phân tích cảm xúc.

## **Các bài toán tương tự trong thực tế**

Một số bài toán tương tự với bài toán phân tích cảm xúc từ mạng xã hội. Các bài toán này đều nằm trong lĩnh vực **Khai phá dữ liệu (Data Mining)** và **Xử lý ngôn ngữ tự nhiên (NLP):**

1. **Phân loại bình luận sản phẩm (Product Review Classification)**

**Mô tả:** Dự đoán cảm xúc của người dùng thông qua đánh giá (review) trên các trang thương mại điện tử như Amazon, Tiki, Shopee,...

**Ứng dụng**: Xác định mức độ hài lòng của khách hàng, cải thiện chất lượng sản phẩm/dịch vụ.

1. **Phân loại tin tức (News Categorization)**

**Mô tả:** Phân loại bài báo hoặc tiêu đề tin tức vào các nhóm chủ đề như thể thao, chính trị, kinh tế, giải trí,...

**Ứng dụng:** Hệ thống gợi ý tin tức, phân luồng nội dung, lọc thông tin theo sở thích người dùng.

1. **Phát hiện phát ngôn thù ghét (Hate Speech Detection)**

**Mô tả:** Nhận diện các phát ngôn độc hại, phân biệt chủng tộc, giới tính, tôn giáo,... trên mạng xã hội.

**Ứng dụng:** Giúp các nền tảng mạng xã hội kiểm duyệt nội dung, bảo vệ cộng đồng.

Tuy các bài toán tương đối giống nhau nhưng quá trình rút trích đặc trưng lại có thể hoàn toàn khác tùy thuộc vào mô hình và đầu ra của bài toán.

# **Phần II: Mô tả dữ liệu**

## **Giới thiệu về bộ dữ liệu**

Bộ dữ liệu được cung cấp bởi người dùng **Kaggle có tên Kashish Parmar**, là tập hợp các bài đăng trên mạng xã hội được gán nhãn cảm xúc. Dữ liệu được chia sẻ công khai với mục đích phục vụ học thuật và nghiên cứu trong lĩnh vực khai phá dữ liệu và xử lý ngôn ngữ tự nhiên (NLP).

## **Cấu trúc bộ dữ liệu**

Bộ dữ liệu được xây dựng từ các bài đăng trên mạng xã hội, bao gồm nhiều thuộc tính mô tả nội dung, người dùng, thời gian và mức độ tương tác. Cụ thể, bộ dữ liệu bao gồm các trường sau:

|  |  |
| --- | --- |
| **Thuộc tính** | **Mô tả** |
| Text | Nội dung do người dùng tạo ra, phản ánh cảm xúc hoặc quan điểm. |
| Sentiment | Nhãn cảm xúc được gán cho bài đăng (tích cực, tiêu cực hoặc trung tính). |
| Timestamp | Thời gian cụ thể mà bài đăng được đăng tải. |
| User | Mã định danh duy nhất của người dùng đăng bài. |
| Platform | Nền tảng mạng xã hội nơi bài viết được đăng (như Twitter, Facebook, v.v.). |
| Hashtags | Các hashtag được sử dụng trong bài đăng, phản ánh chủ đề đang thịnh hành. |
| Likes | Số lượt thích, thể hiện mức độ tương tác từ người xem. |
| Retweets | Số lượt chia sẻ lại (retweet), thể hiện mức độ lan truyền của nội dung. |
| Country | Quốc gia nơi người dùng đăng bài. |
| Year | Năm bài viết được đăng. |
| Month | Tháng bài viết được đăng. |
| Day | Ngày bài viết được đăng. |
| Hour | Giờ bài viết được đăng. |

Bảng 1: Các thuộc tính của bộ dữ liệu

Bộ dữ liệu có kích thước khoảng hơn **700** dòng, với mỗi dòng tương ứng một bài đăng duy nhất. Các thuộc tính trong tập dữ liệu không chỉ phục vụ cho phân tích cảm xúc, mà còn hỗ trợ trong việc phân tích hành vi người dùng theo **thời gian, vị trí địa lý, nền tảng sử dụng, và mức độ tương tác với nội dung**. Chúng ta sẽ đặc biệt quan tâm 2 cột thuộc tính chính đó là **Text** và **Sentiment**, đây là 2 cột thuộc tính **input của bài toán.**

# **Phần III: Xử lý dữ liệu**

## **Chia lại bộ dữ liệu**

Vì đây là bộ dữ liệu được thu thập dựa trên cảm xúc của người dùng, số lượng cảm xúc lớn, đa dạng gồm **191** loại cảm xúc khác nhau. Vì vậy nên tôi quyết định chia lại bộ dữ liệu dựa theo 3 loại cảm xúc chính là **tích cực(Positive), tiêu cực(Negative), trung lập(Neutral).**

* 1. **Gom nhóm các các xúc với nhau thành 3 nhóm chính:**

**Tích cực (Positive)** gồm 113 cảm xúc:

positive = [  
 'Positive','Excitement', 'Happiness', 'Joy', 'Love', 'Amusement',

'Enjoyment','Admiration', 'Affection', 'Awe', 'Acceptance',

'Adoration','Anticipation', 'Kind', 'Pride', 'Elation', 'Euphoria',  
 'Contentment', 'Serenity', 'Gratitude', 'Hope', 'Empowerment',  
 'Compassion', 'Tenderness', 'Arousal', 'Enthusiasm', 'Fulfillment',  
 'Reverence', 'Determination', 'Zest', 'Hopeful', 'Proud', 'Grateful',  
 'Empathetic', 'Compassionate', 'Playful', 'Free-spirited','Inspired',  
 'Confident', 'Thrill', 'Overjoyed', 'Inspiration', 'Motivation',  
 'Satisfaction', 'Blessed', 'Accomplishment', 'Wonderment','Optimism',  
 'Enchantment', 'Intrigue', 'PlayfulJoy', 'Mindfulness','DreamChaser',  
 'Elegance', 'Whimsy', 'Harmony', 'Creativity', 'Radiance', 'Wonder',  
 'Rejuvenation', 'Coziness', 'Adventure', 'Melodic', 'FestiveJoy',  
 'InnerJourney', 'Freedom', 'Dazzle', 'Adrenaline', 'ArtisticBurst',  
 'CulinaryOdyssey', 'Resilience', 'Immersion', 'Spark', 'Marvel',  
 'Success', 'Friendship', 'Romance', 'Tranquility', 'Grandeur',  
 'Energy', 'Celebration', 'Charm', 'Ecstasy', 'Colorful', 'Hypnotic',  
 'Connection', 'Iconic', 'Journey', 'Engagement', 'Touched',  
 'Triumph', 'Heartwarming', 'Breakthrough', 'Joy in Baking',  
 'Envisioning History', 'Imagination', 'Vibrancy', 'Mesmerizing',  
 'Culinary Adventure', 'Winter Magic', 'Thrilling Journey',  
 "Nature's Beauty", 'Celestial Wonder', 'Creative Inspiration',  
 'Runway Creativity', "Ocean's Freedom", 'Happy', 'Confidence',  
 'Kindness', 'Positivity', 'Amazement', 'Captivation', 'Emotion'  
]

**Tiêu cực (Negative)** gồm 54 cảm xúc:

negative = [  
 'Negative','Sad','Frustrated', 'Anger', 'Fear', 'Sadness', 'Disgust',   
 'Bitter', 'Shame', 'Despair', 'Grief', 'Loneliness', 'Jealousy',  
 'Resentment', 'Frustration', 'Boredom', 'Anxiety', 'Intimidation',  
 'Helplessness', 'Envy', 'Regret', 'Melancholy', 'Bitterness',  
 'Yearning', 'Fearful', 'Apprehensive', 'Overwhelmed', 'Jealous',  
 'Devastated', 'Envious', 'Dismissive', 'Heartbreak', 'Betrayal',  
 'Suffering', 'EmotionalStorm', 'Isolation', 'Disappointment',  
 'LostLove', 'Exhaustion', 'Sorrow', 'Darkness', 'Desperation',  
 'Ruins', 'Desolation', 'Loss', 'Heartache', 'Hate', 'Bad',  
 'Embarrassed', 'Pressure', 'Miscalculation', 'Obstacle', 'Challenge',

'Disappointed',  
]

**Trung lập (Neutral)** gồm 24 cảm xúc:

neutral = [  
 'Neutral','Bittersweet', 'Surprise', 'Calmness', 'Confusion',   
 'Numbness', 'Nostalgia', 'Ambivalence', 'Pensive', 'Reflection',  
 'Indifference', 'Contemplation', 'JoyfulReunion', 'Appreciation',  
 'Sympathy', 'Renewed Effort', 'Solace', 'Relief', 'Mischievous',  
 'Whispers of the Past', 'Solitude','Exploration','Suspense',  
 'Curiosity',  
]

* 1. **Ánh xạ các cảm xúc vào nhóm của nó**

def map\_sentiment\_to\_group(sentiment):  
 sentiment = sentiment.strip() *# Loại bỏ khoảng trắng thừa* if sentiment in positive:  
 return 'Positive'  
 elif sentiment in negative:  
 return 'Negative'  
 elif sentiment in neutral:  
 return 'Neutral'  
 else:  
 return 'unknown' *# Trường hợp không tìm thấy*

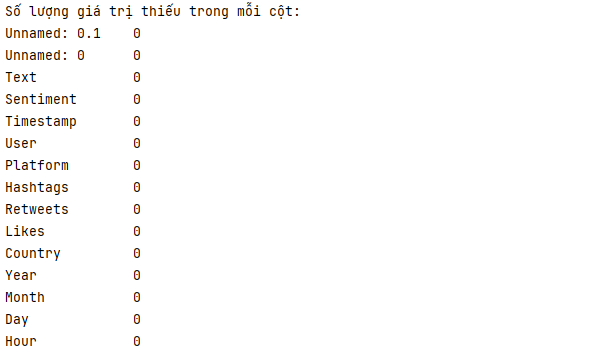
* 1. **Tiến hành phân chia tạo dữ liệu mới**

def convert\_sentiment\_groups(input\_file, output\_file):  
 try:  
 *# Đọc file CSV vào DataFrame* df = pd.read\_csv(input\_file)  
  
 *# Kiểm tra xem cột 'Sentiment' có tồn tại không* if 'Sentiment' not in df.columns:  
 print("Lỗi: Không tìm thấy cột 'Sentiment' trong file CSV")  
 return False  
  
 *# Tạo một bản sao của DataFrame để tránh thay đổi dữ liệu gốc* new\_df = df.copy()  
  
 *# Áp dụng hàm map\_sentiment\_to\_group cho mỗi giá trị trong cột 'Sentiment'* new\_df['Sentiment'] = new\_df['Sentiment'].apply(map\_sentiment\_to\_group)  
  
 *# Lưu DataFrame mới vào file CSV đầu ra* new\_df.to\_csv(output\_file, index=False)  
 print(f"Đã chuyển đổi thành công và lưu kết quả vào {output\_file}")  
 return True  
  
 except Exception as e:  
 print(f"Đã xảy ra lỗi: {str(e)}")  
 return False

## **Xử lý dữ liệu thiếu**

Vì đây là bộ dữ liệu đã được thu thập và tiền xử lý cơ bản nên không có dữ liệu bị thiếu. Tuy nhiên chúng ta cần phải kiểm tra lại để đảm bảo tính đầy đủ của dữ liệu.

print("\nSố lượng giá trị thiếu trong mỗi cột:")  
print(df.isnull().sum())



Hình 1: Kết quả kiểm tra dữ liệu thiếu

Dữ liệu đầy đủ, không bị thiếu.

## **Làm sạch và định dạng dữ liệu**

Loại bỏ các cột không cần thiết(không có ý nghĩa):

df.drop(columns=[col for col in df.columns if col.startswith('Unnamed:')]  
 ,inplace=True)

Cột Timestamp được chuyển đổi sang định dạng thời gian (datetime):

df['Timestamp'] = pd.to\_datetime(df['Timestamp'])  
df['DayOfWeek'] = df['Timestamp'].dt.dayofweek

Làm sạch các cột dữ liệu, đảm bảo tính nhất quán của dữ liệu:

df['Sentiment'] = df['Sentiment'].str.strip()  
df['Platform'] = df['Platform'].str.strip()  
df['Country'] = df['Country'].str.strip()  
df['User'] = df['User'].str.strip()  
df['Hashtags'] = df['Hashtags'].str.strip()

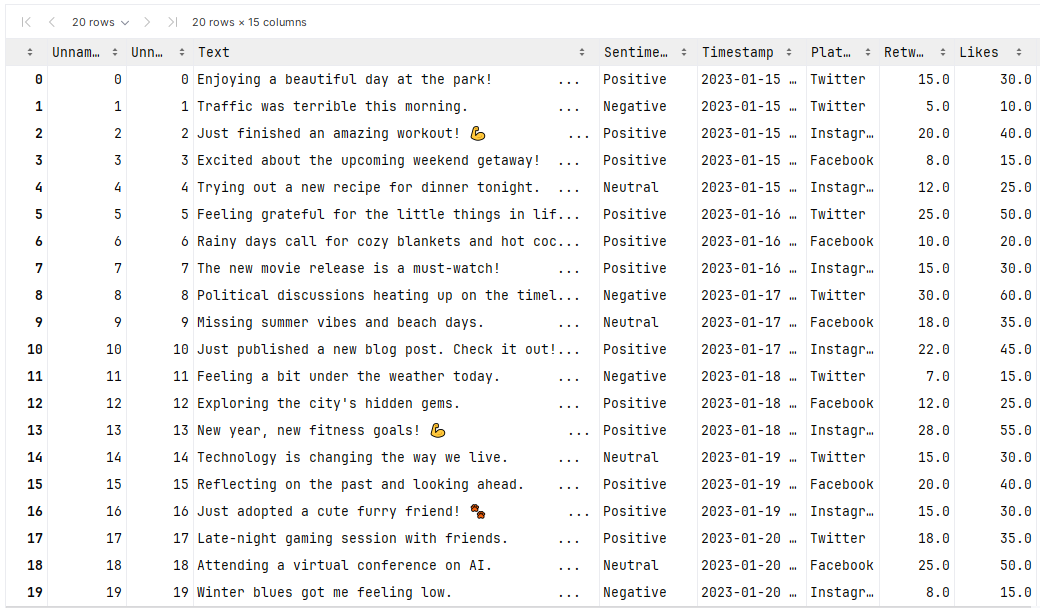
Tổng quan dữ liệu trước và sau khi được xử lý:



Hình 2: Biểu đồ phân phôi bài đăng theo nền tảng MXH trước khi làm sạch dữ liệu



Hình 3: Biểu đồ phân phôi bài đăng theo nền tảng MXH sau khi làm sạch dữ liệu



Hình 4: Dữ liệu trước khi được xử lý



Hình 5:Dữ liệu sau khi được xử lý

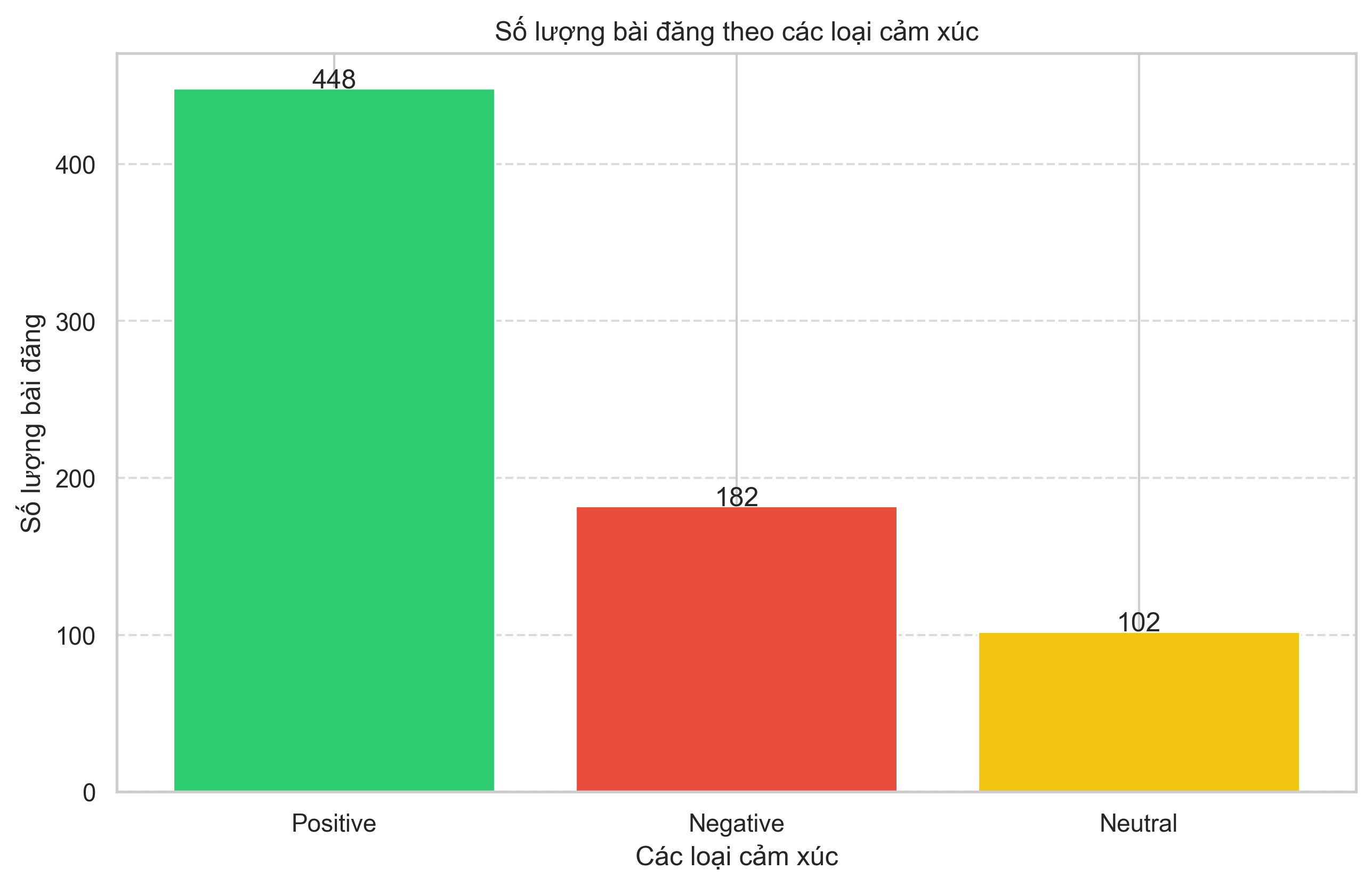
(lưu ý: chưa tiền xử lý cột **“Text”** vì phân tích **EDA** không dùng nhiêu tới cột **“Text”)**

# **Phần IV: Phân tích dữ liệu**

## **Phân tích đơn biến**

* 1. **Phân phối cảm xúc (Sentiment):**

Thống kê số lượng bài đăng tương ứng với từng loại cảm xúc (Positive, Negative, Neutral). Kết quả được thể hiện bằng biểu đồ cột.

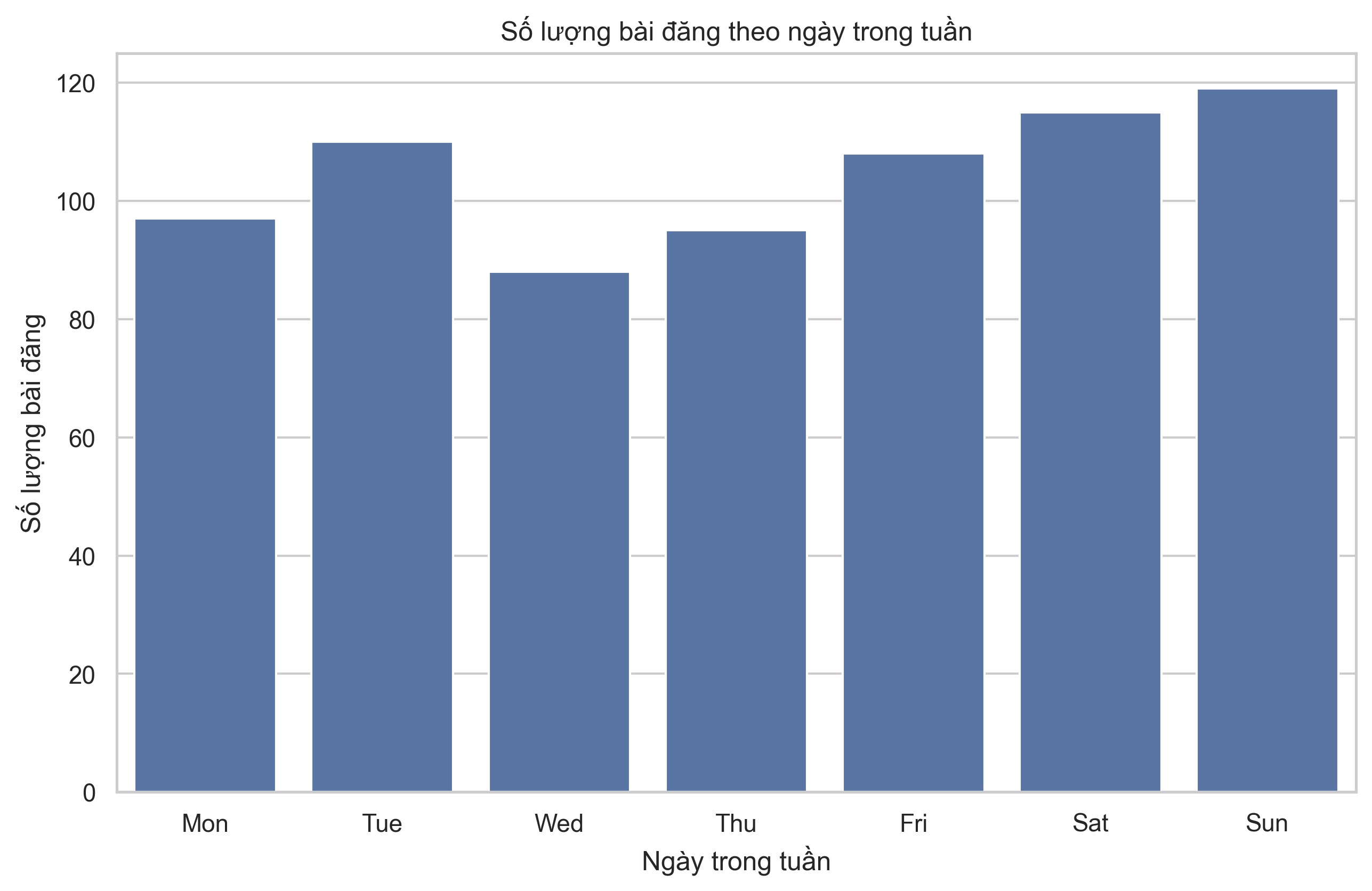


Hình 6: Biểu đồ số lượng bài đăng theo từng loại cảm xúc

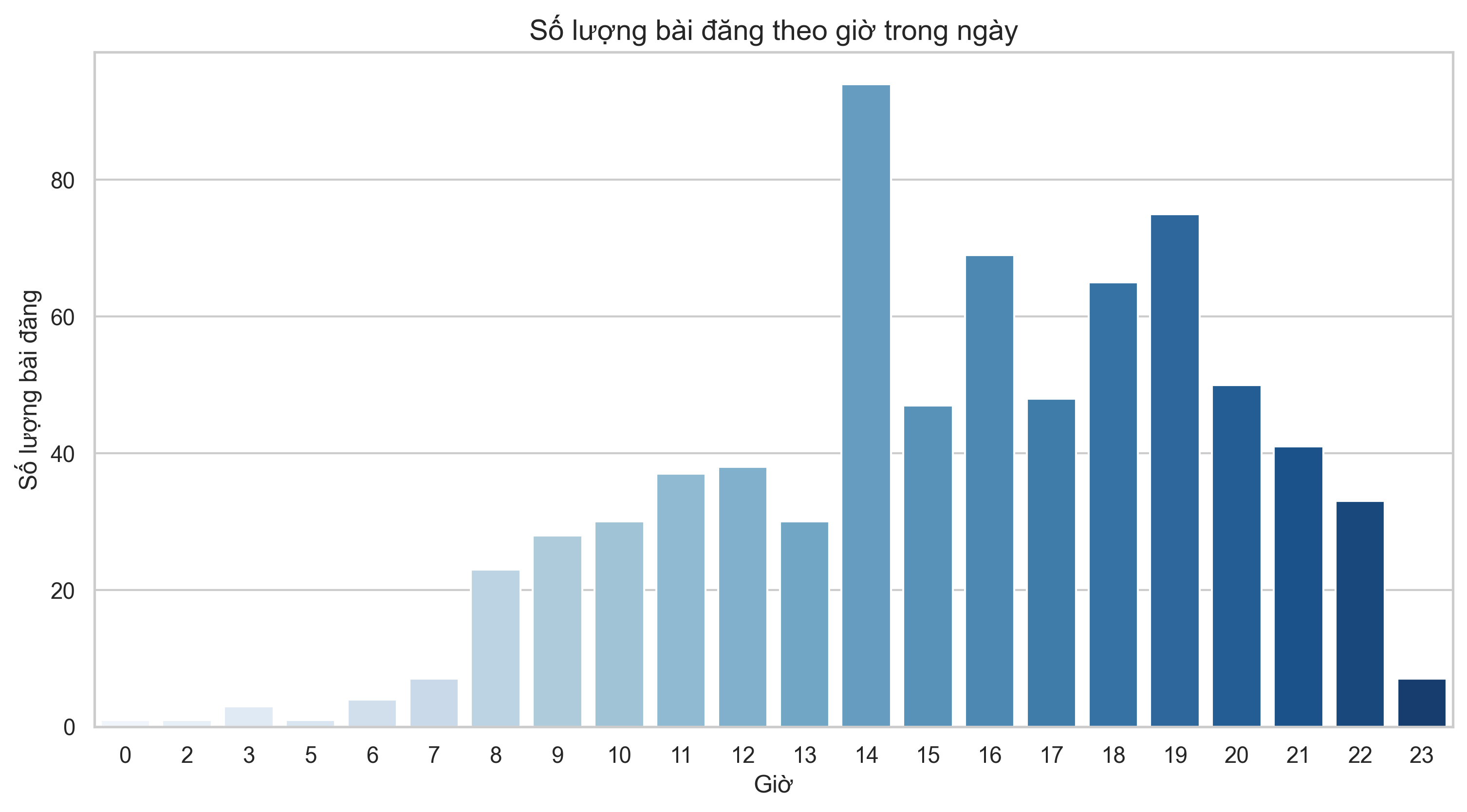
**Nhận xét:** có sự chênh lệch lớn giữa số lượng bài đăng phân loại theo cảm xúc. Nhìn chung các nền tảng mạng xã hội vẫn tích cực.

* 1. **Phân phối theo ngày, giờ (Timestamp):**

Thực hiện thống kê top số lượng bài đăng theo ngày trong tuần, theo giờ trong ngày.



Hình 7: Biểu đồ số lượng bài đăng theo các ngày trong tuần



Hình 8: Biểu đồ số lượng bài đăng theo từng khung giờ trong ngày

**Nhận xét:** Phân tích hai biểu đồ cho thấy có sự khác biệt đáng kể trong mô hình đăng bài cả theo giờ và theo ngày. Về thời gian trong ngày, mạng xã hội hoạt động sôi nổi nhất vào hai khung giờ chính: cao điểm đầu tiên vào 14h (có thể là giờ nghỉ trưa) và cao điểm thứ hai vào khoảng 18h-19h (có thể là giờ tan làm/tan học). Nhìn chung, khoảng thời gian từ 14h đến 21h có lượng bài đăng cao nhất, phản ánh thời gian người dùng có nhiều thời gian rảnh để tương tác trên mạng xã hội. Về ngày trong tuần, số lượng bài đăng tăng dần từ giữa tuần đến cuối tuần, với ngày cao nhất là Chủ nhật, tiếp theo là Thứ Bảy, cho thấy người dùng hoạt động nhiều hơn vào những ngày nghỉ.

* 1. **Phân phối theo nền tảng mạng xã hội (Platform):**

Sử dụng value\_counts() để thống kê số lượng bài đăng từ mỗi nền tảng. Sau đó vẽ biểu đồ cột với seaborn để minh họa số lượng bài đăng từ các nền tảng khác nhau (ví dụ: Twitter, Facebook...).

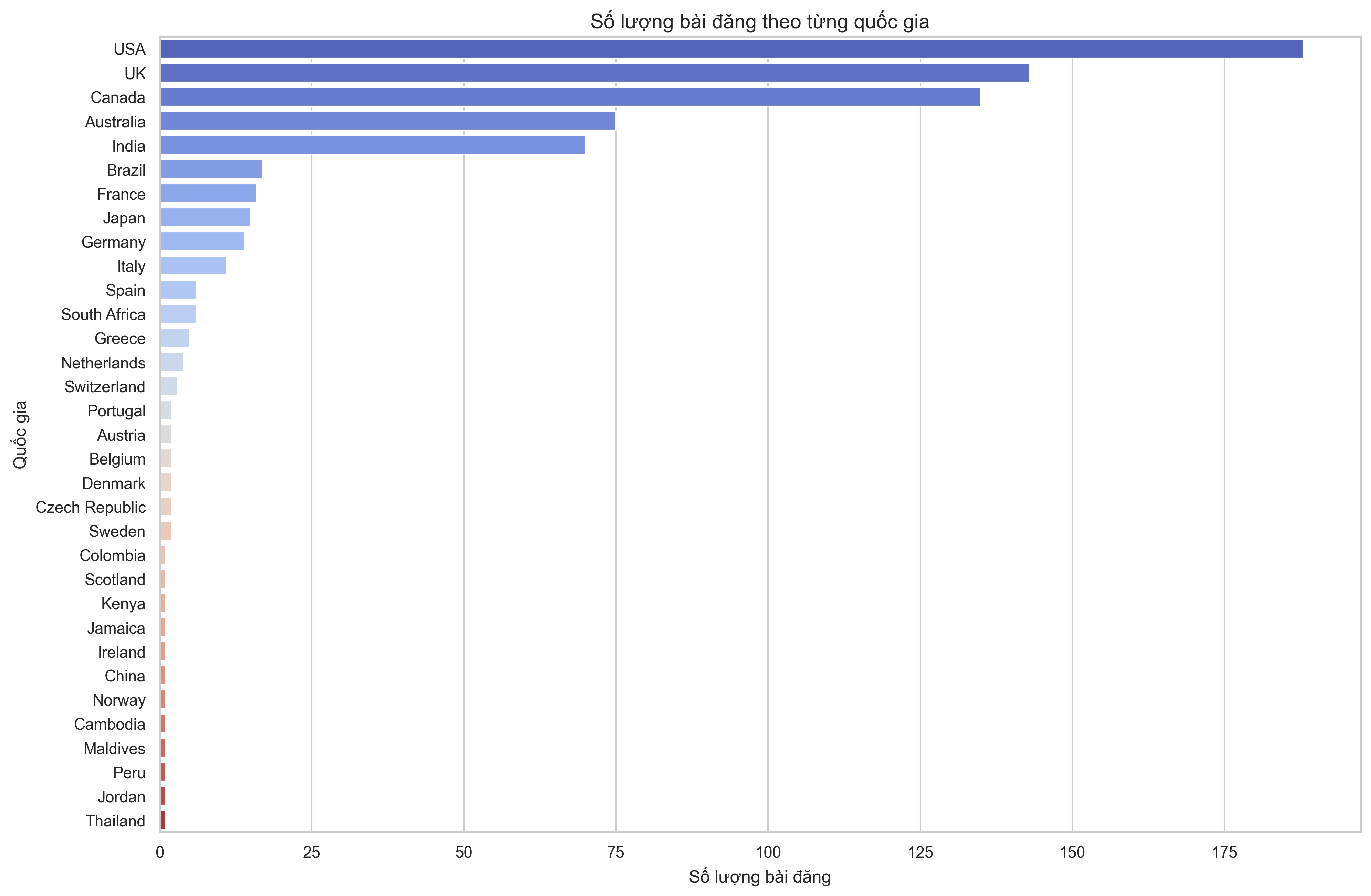


Hình 9: Biểu đồ số lượng bài đăng theo nền tảng MXH

**Nhận xét:** sự trên lệch số lượng bài đăng giữa các nền tảng là không quá lớn cho thấy quá trình thu thập dữ liệu diễn ra ổn định và đồng đều.

* 1. **Phân phối theo quốc gia (Country):**

Thực hiện thống kê số lượng bài đăng theo các quốc gia.

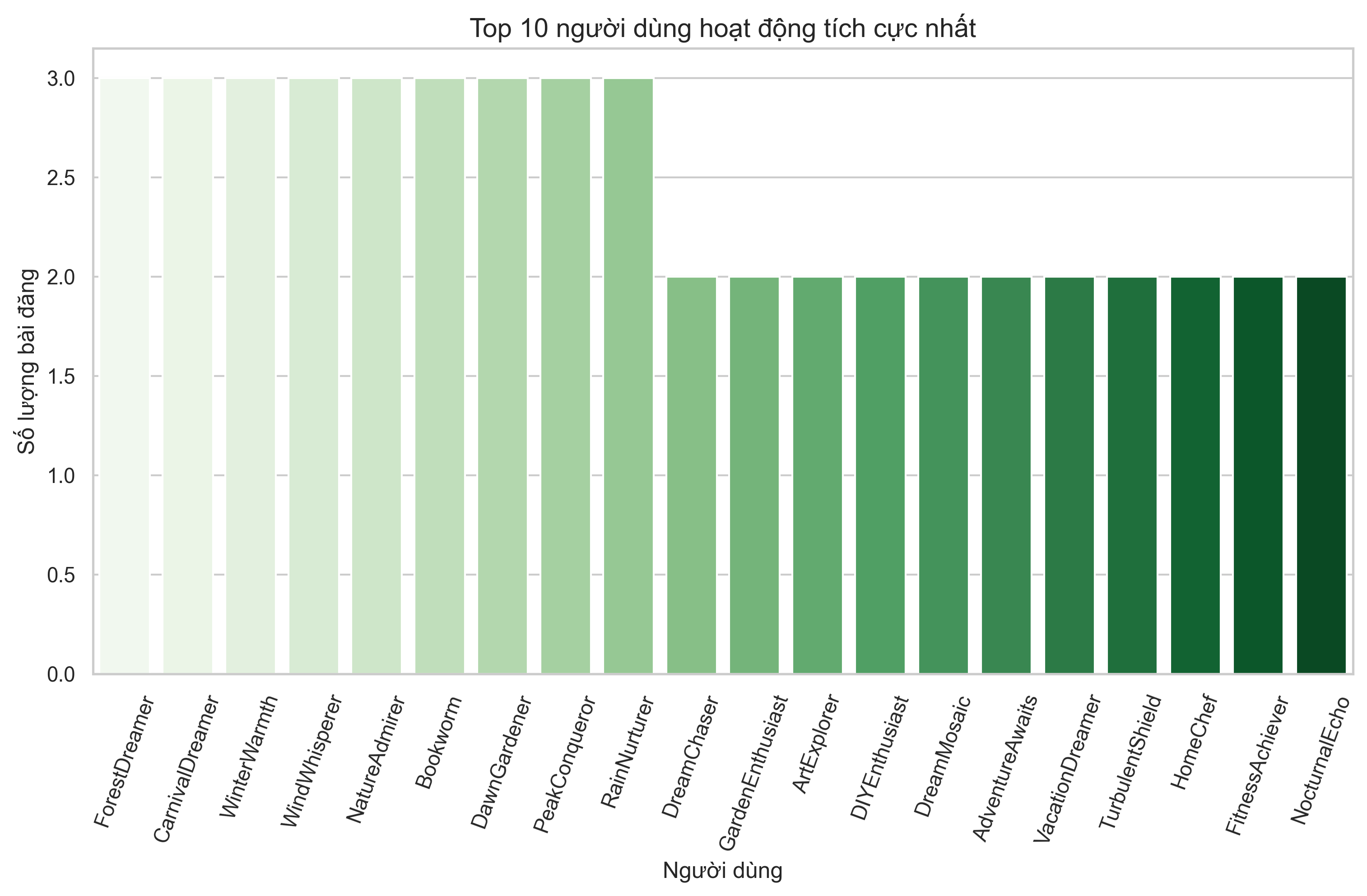


Hình 10: Biểu đồ số lượng bài đăng theo từng quốc gia

**Nhận xét:** Biểu đồ cho thấy sự chênh lệch rất lớn về số lượng bài đăng giữa các quốc gia. Các quốc gia nói tiếng Anh như Mỹ, Anh và Canada chiếm ưu thế áp đảo trong số lượng bài đăng. Điều này có thể cho thấy dữ liệu được thu thập chủ yếu từ các nền tảng mạng xã hội sử dụng tiếng Anh hoặc các nền tảng này phổ biến hơn ở các quốc gia nói tiếng Anh. Ngoài ra, sự hiện diện của các quốc gia đa dạng như Ấn Độ, Brazil, và Nhật Bản trong top 10 cho thấy tính toàn cầu của mạng xã hội, mặc dù với mức độ phân bố không đồng đều. Các quốc gia ở châu Phi, Đông Nam Á và Đông Âu có rất ít đại diện, phản ánh khả năng tiếp cận hoặc mức độ phổ biến của các nền tảng này ở các khu vực đó.

* 1. **Phân phối theo người dùng (User):**

Thực hiện thống kê số lượng bài đăng theo người dùng, lấy top 10 người dùng tích cực đăng bài nhất.



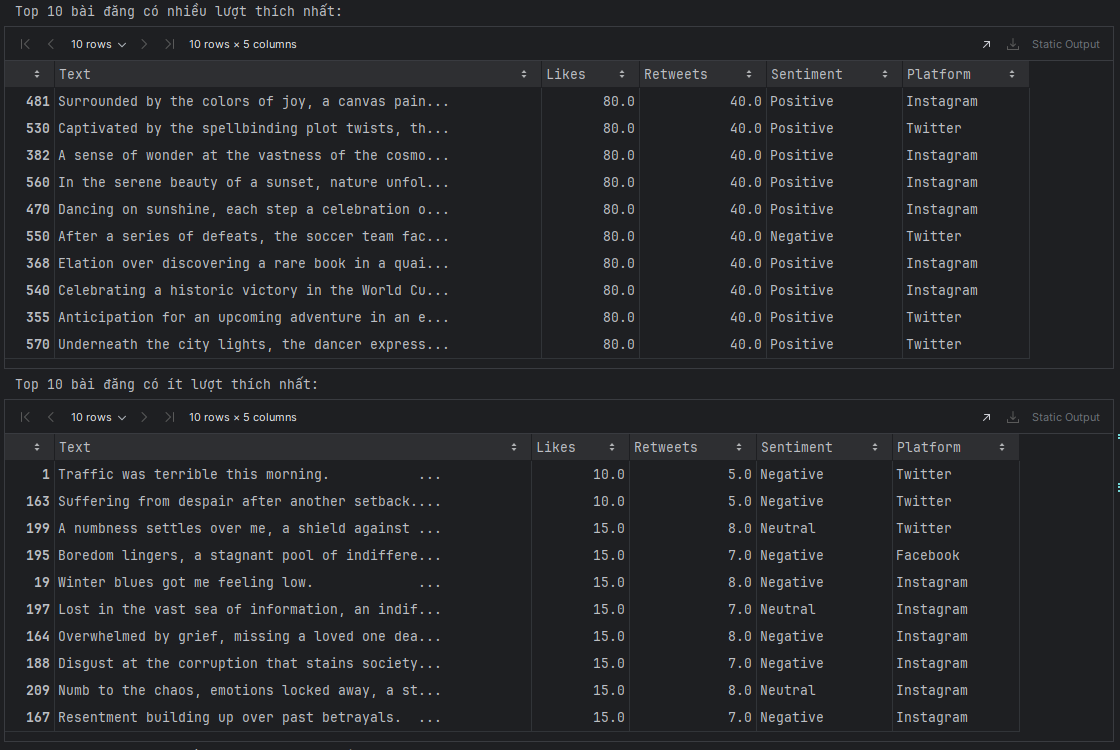
Hình 11: Biểu đồ top 10 người dùng tích cực đăng bài nhất

**Nhận xét:** Biểu đồ cho thấy số lượng bài đăng của từng người dùng là không nhiều (cao nhất là 3 bài), cho thấy quá trình thu thập dữ liệu diễn ra trên toàn bộ người dùng mạng xã hội chứ không tập trung vào một nhóm đối tượng riêng biệt nào, cột dữ liệu **“User”** gần như không có quá nhiều ý nghĩa trong bộ dữ liệu này.

* 1. **Phân phối theo lượt tương tác (Like/Retwwets):**

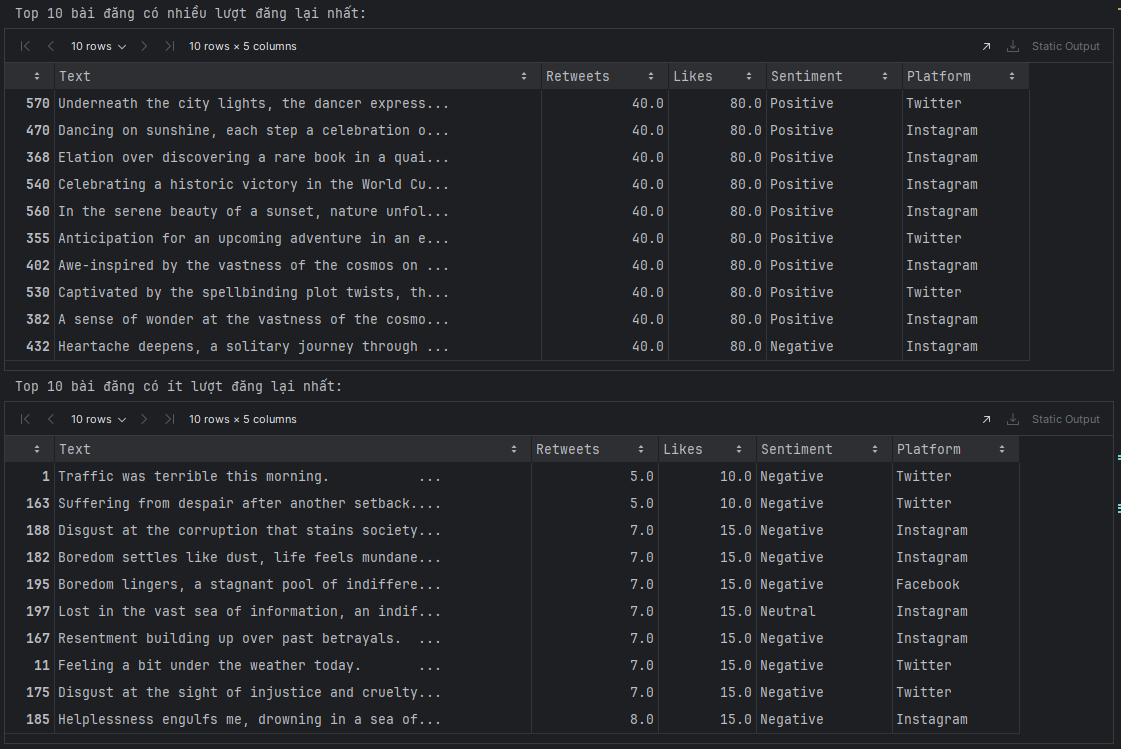
Thực hiện thống kê số lượng bài đăng theo lượt tương tác, lấy top 10 bài viết có lượt tương tác nhiều nhất và thấp nhất.

**Likes:**



Hình 12: Top 10 bài viết có ít/nhiều lượt thích nhất

**Retweets:**

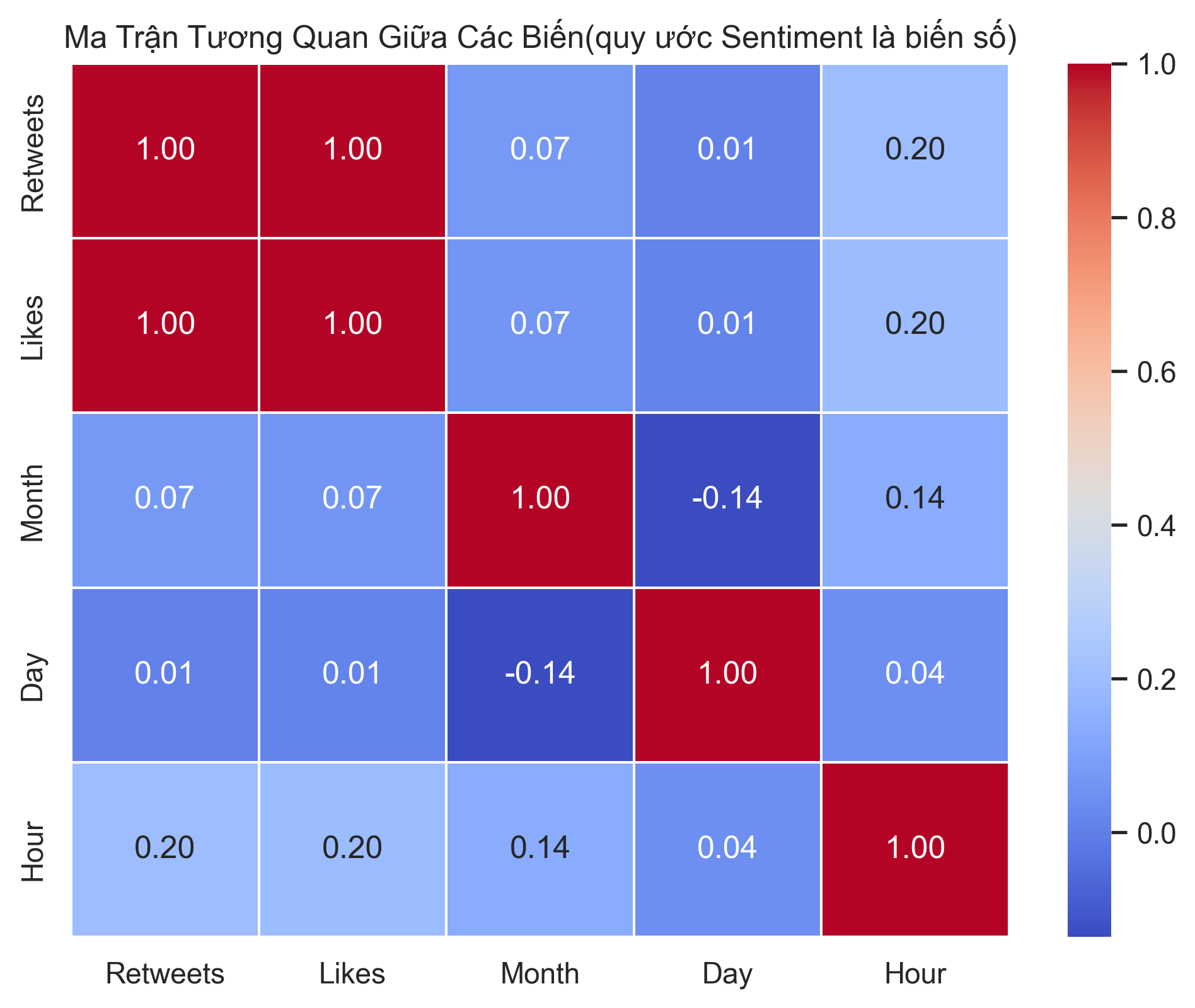


Hình 13: Top 10 bài viết có ít/nhiều lượt đăng lại nhất

**Nhận xét:** bảng dữ liệu cho thấy số lượng những bài đăng được nhiều lượt **likes/retweets** nhất chủ yếu nằm ở nhóm những bài đăng **tích cực(positive)**, những bài đăng ít lượt **likes/retweets** chủ yếu nằm ở nhóm những bài đăng **tiêu cực(negative)**, các bài đăng thuộc nhóm bài đăng có cảm xúc **trung lập(neutral)** gần như nằm ở tầm trung(không ít cũng không nhiều lượt tương tác). **Dữ liệu có sự phân hóa rõ ràng** theo từng nhóm cảm xúc.

## **Phân tích đa biến**

Vì dữ liệu gồm nhiều trường văn bản và phân loại, việc tính toán tương quan số học có thể không phù hợp. Tuy nhiên, tương quan giữa các trường số như Likes, Retweets, Month, Hour, Day có thể được khai thác tốt.



Hình 14: Ma trận tương quan Likes, Retweets, Month, Hour, Day

**Nhận xét:** Ma trận tương quan cho thấy **Retweets** và **Likes** có mối quan hệ chặt chẽ (tương quan 1.00), trong khi các yếu tố thời gian như **Month, Day**, và **Hour** không ảnh hưởng mạnh đến hành vi tương tác của người dùng.

## **Phân tích tương quan**

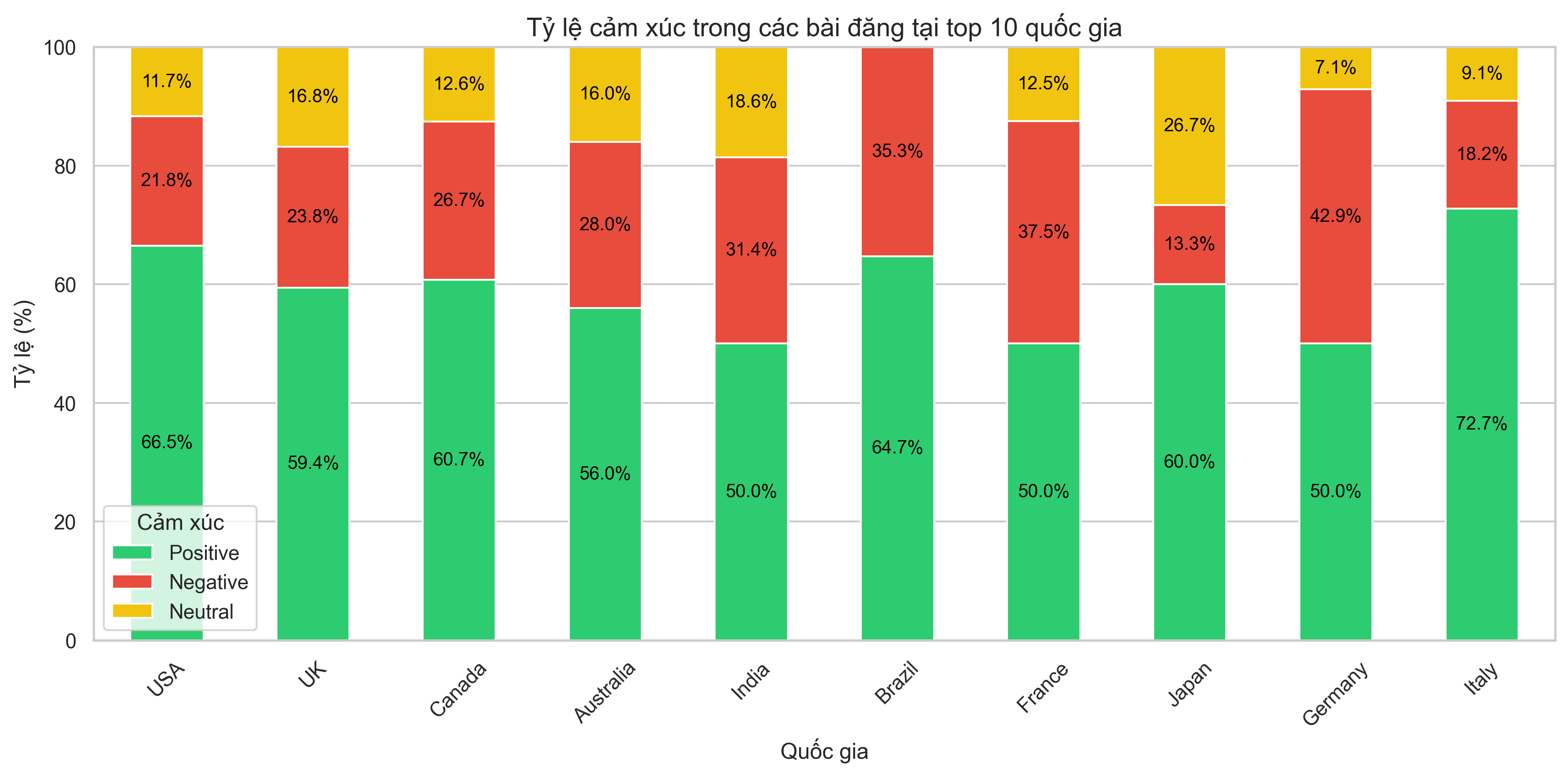
Kết hợp các biến như Sentiment với Platform, Country, hoặc Timestamp, Likes, Retweets để tìm hiểu cảm xúc theo từng nhóm nền tảng, khu vực, thời gian hoặc lượt tương tác.

* 1. **Phân phối theo cảm xúc/quốc gia (Sentiment/Country):**

Thực hiện thống kê số lượng bài đăng theo cảm xúc trên top 10 quốc gia có số lượng bài đăng nhiều nhất. Tính tỷ lệ bài đăng tích cực/tiêu cực/trung lập trên tổng số lượng bài đăng của từng quốc gia.



Hình 15: Biểu đồ top 10 quốc gia có nhiều bài đăng nhất

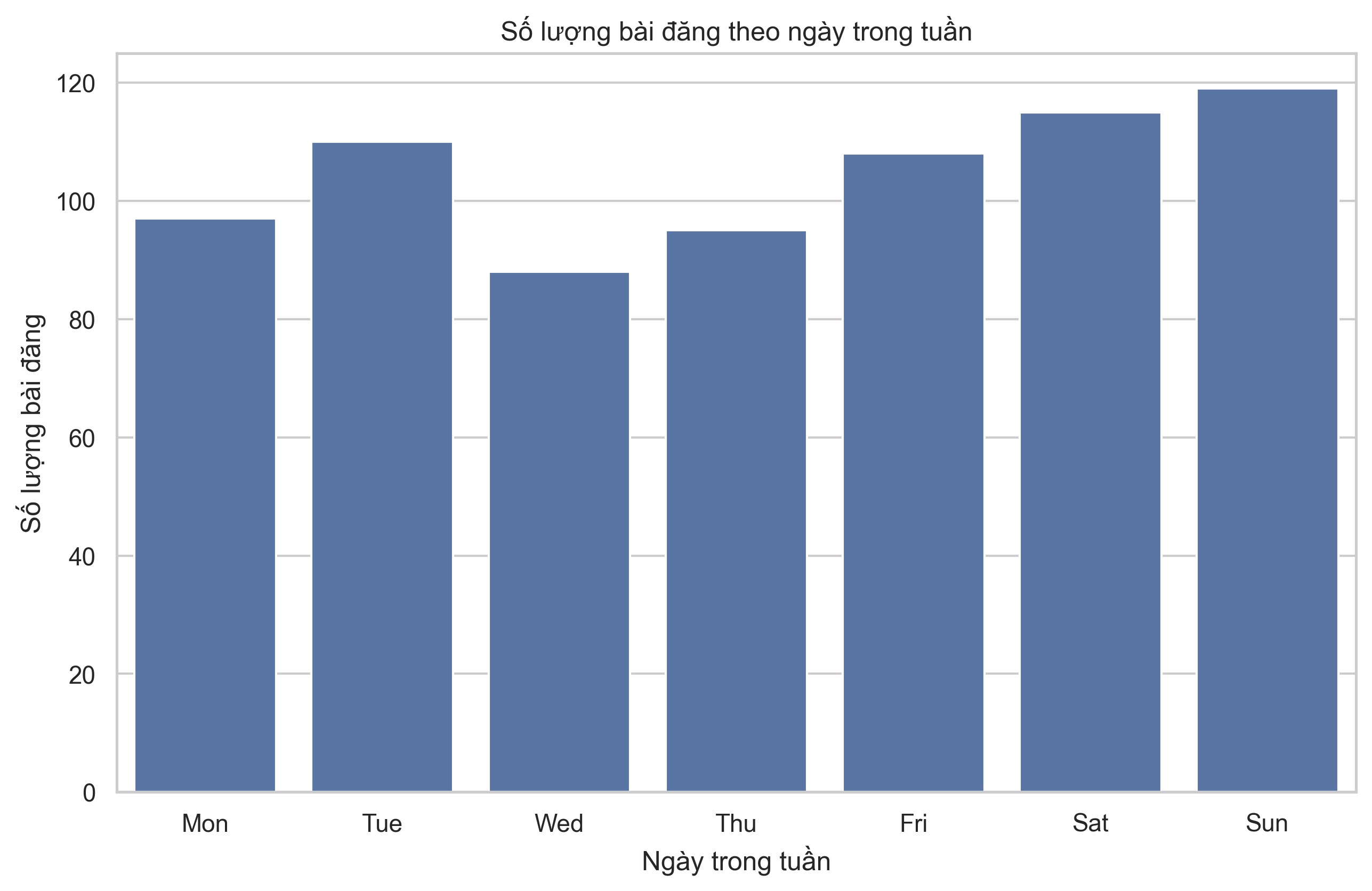


Hình 16: Biểu đồ tỷ lệ cảm xúc trong các bài đăng tại top 10 quốc gia

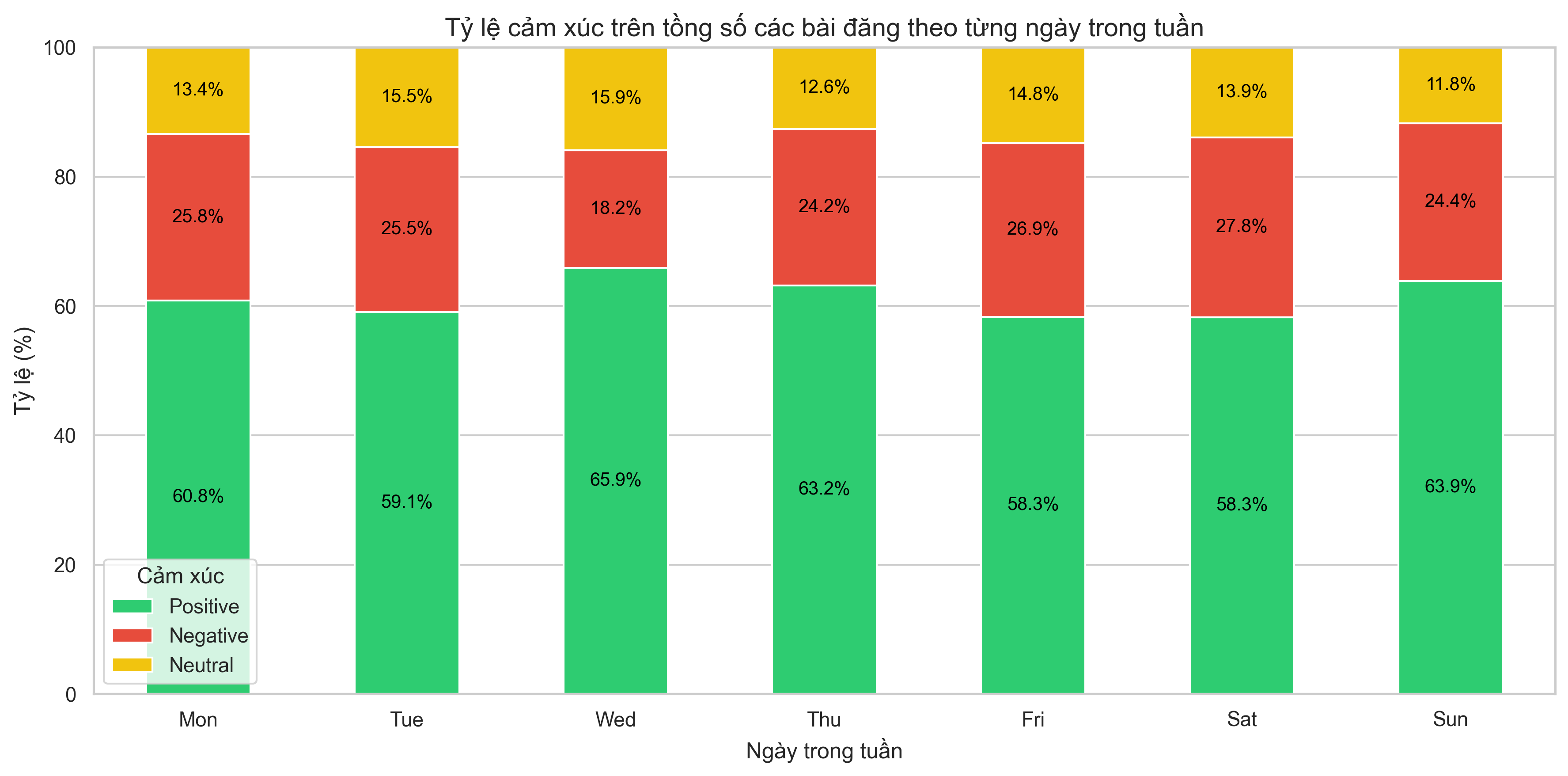
**Nhận xét:** Qua phân tích hai biểu đồ, ta thấy Hoa Kỳ dẫn đầu về số lượng bài đăng (khoảng 180 bài), cao hơn nhiều so với các nước như Brazil, Pháp, Nhật, Đức và Ý (dưới 25 bài mỗi nước). Tuy nhiên, về chất lượng cảm xúc, **Ý lại nổi bật với tỷ lệ nội dung tích cực cao nhất (72,7%)**, trong khi **Đức có tỷ lệ tiêu cực cao nhất (42,9%).** Đáng chú ý là không có mối liên hệ rõ ràng giữa số lượng và chất lượng bài đăng - các nước có nhiều bài đăng thể hiện sự phân hóa dữ liệu rõ rệt hơn.

* 1. **Phân phối theo cảm xúc/thời gian (Sentiment/Timestamp):**

Thực hiện thống kê số lượng bài đăng theo cảm xúc theo các ngày trong tuần và theo giờ trong ngày. Tính tỷ lệ bài đăng tích cực/tiêu cực/trung lập trên tổng số lượng bài đăng của từng ngày trong tuần và từng khung giờ trong ngày.

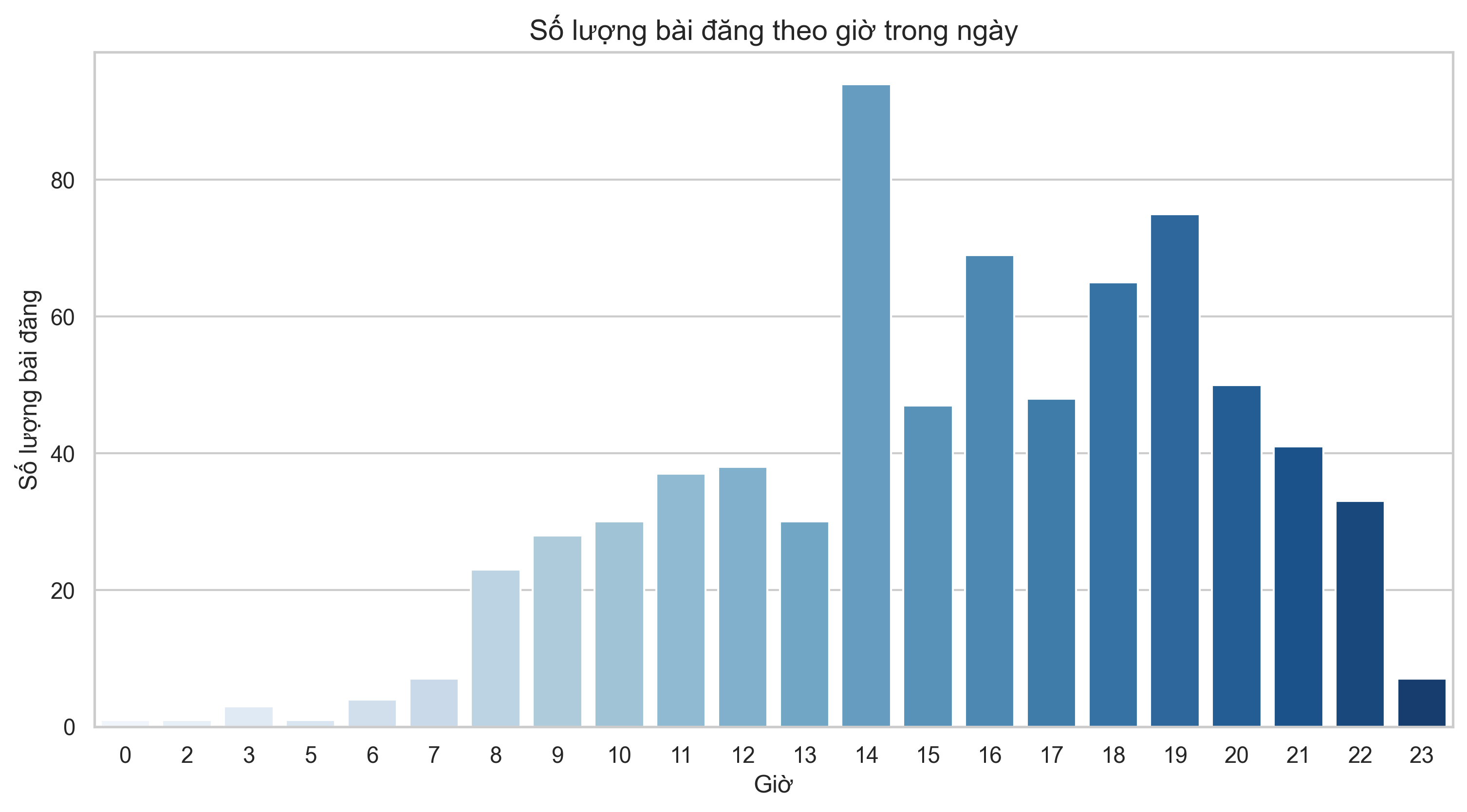


Hình 17: Biểu đồ số lượng bài đăng theo các ngày trong tuần

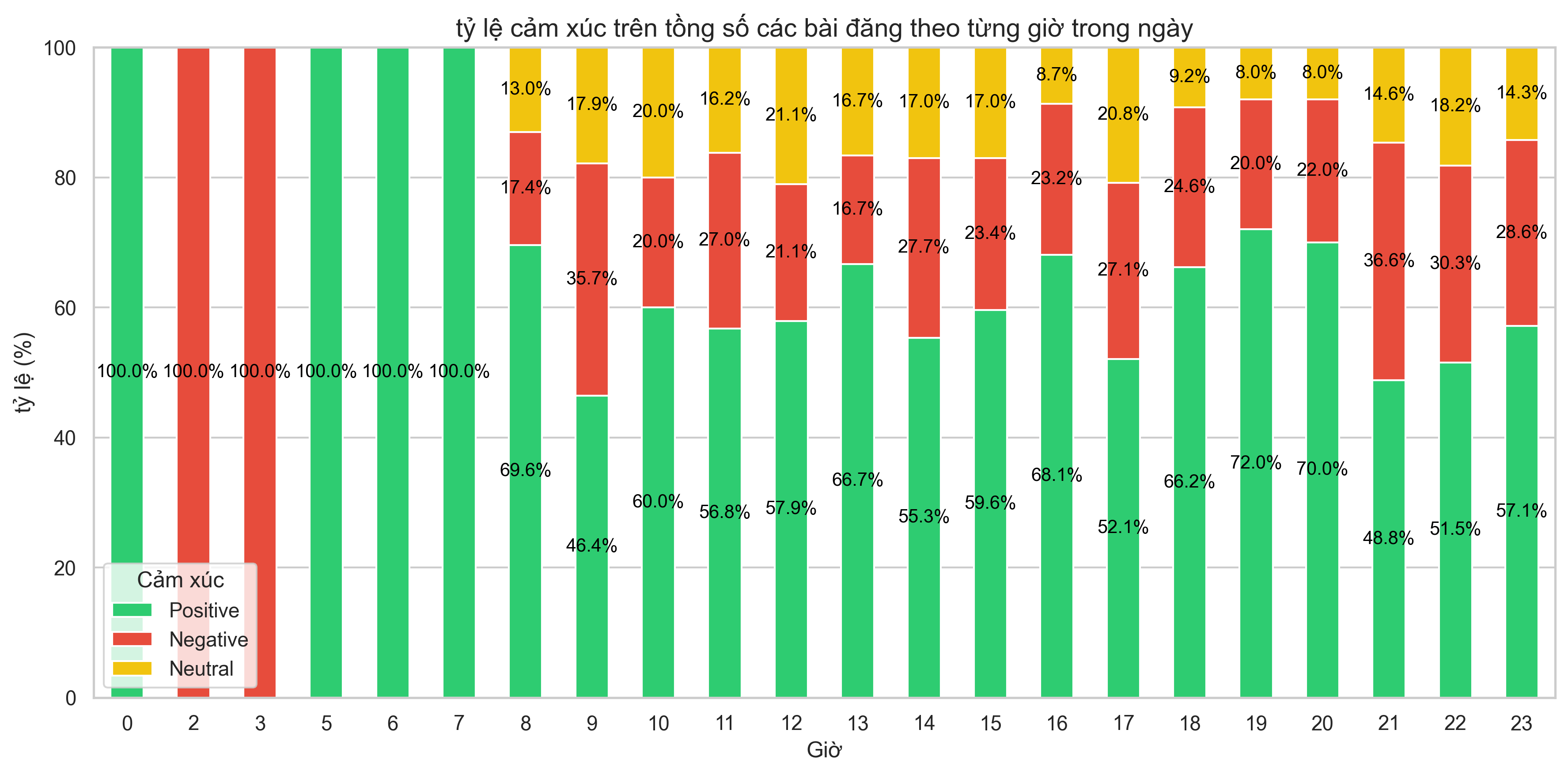


Hình 18: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo các ngày trong tuần

**Nhận xét:** Sự phân bổ cảm xúc bài đăng theo các ngày trong tuần dường như không có sự chênh lệch đáng kể, mặc dù người dung hoạt động mạnh vào các ngày cuối tuần nhưng biểu đồ tỷ lệ vẫn không thể hiện được sự khác biệt giữa các ngày cuối tuần và các ngày đầu tuần.



Hình 19: Biểu đồ số lượn bài đăng theo các khung giờ trong ngày



Hình 20: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo giờ trong ngày

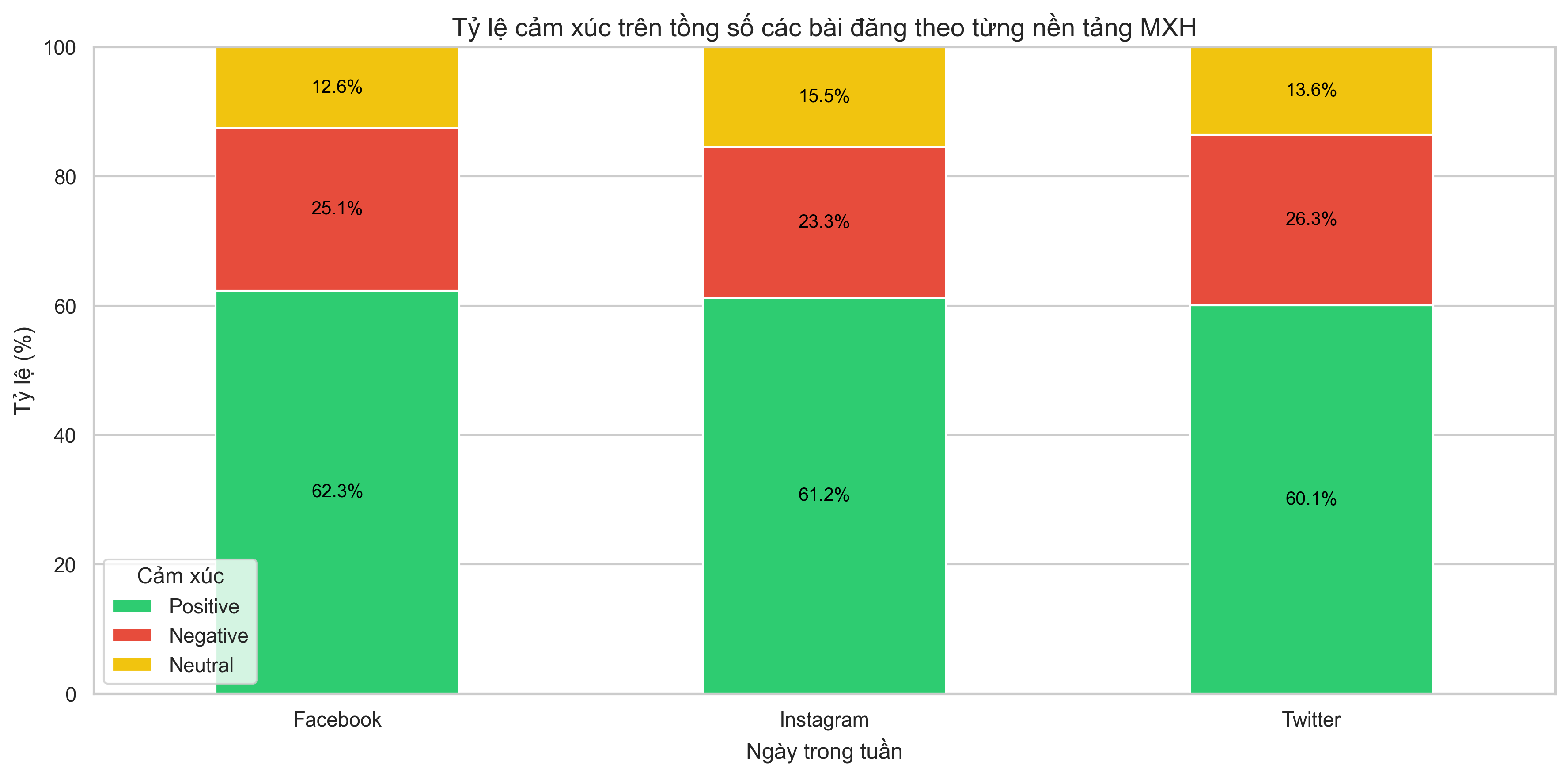
**Nhận xét:** Phân tích hai biểu đồ cho thấy lượng bài đăng cao nhất vào lúc 14h và 19h, trong khi gần như không có bài từ 0–5h sáng. Về cảm xúc, các khung giờ 6h, 7h và 19h có tỷ lệ nội dung tích cực cao nhất (khoảng 72%), ngược lại 21h ghi nhận tỷ lệ tiêu cực cao nhất (36.6%). Đáng chú ý, những thời điểm có nhiều bài đăng (14h, 19h) cũng là lúc cảm xúc tích cực chiếm ưu thế (55.3% và 72%), cho thấy người dùng có xu hướng đăng bài nhiều hơn khi tâm trạng tốt. Tuy nhiên, tỷ lệ nội dung tích cực dao động khá lớn theo giờ, từ 46% đến 72%, điều này rất đáng chú ý về mặt xu hướng người dùng.

* 1. **Phân phối theo cảm xúc/nền tảng MXH (Sentiment/Platform):**

Thực hiện thống kê số lượng bài đăng theo cảm xúc trên các nền tảng MXH. Tính tỷ lệ bài đăng tích cực/tiêu cực/trung lập trên tổng số lượng bài đăng tích cực/tiêu cực/trung lập của từng nền tảng MXH.



Hình 21: Biểu đồ phân phối bài đăng theo nền tảng MXH

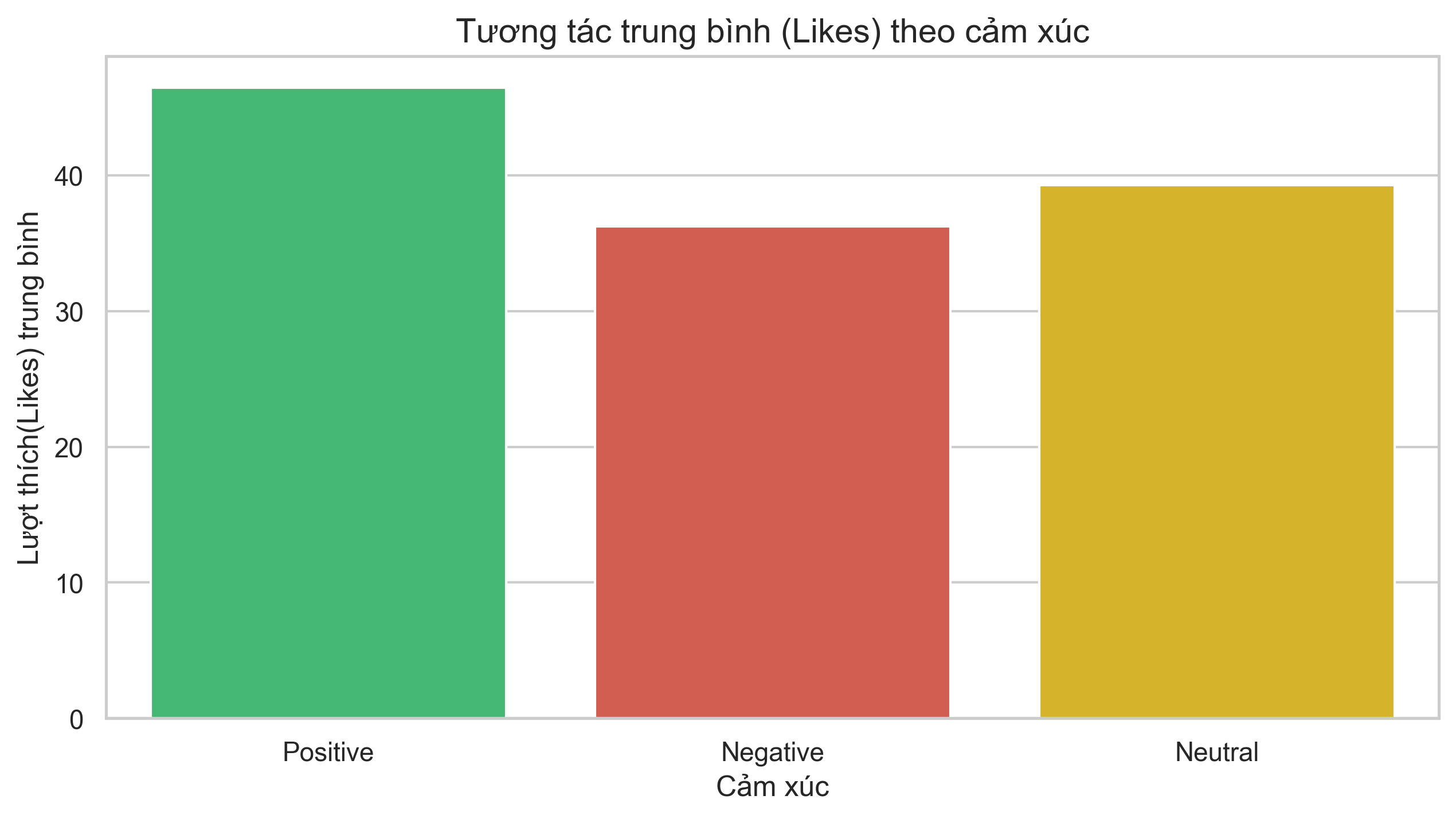


Hình 22: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo nền tảng MXH

**Nhận xét:** Như đã phân tích trước đó, lượng bài đăng và tỷ lệ các cảm xúc giữa các nên tảng mạng xã hội gần như không có sự chênh lệch, tỷ lệ bài đăng tích cực vẫn chiếm ưu thế. Điều này cho thấy dữ liệu được thu thấp một cách đồng đều giữa các nền tảng mạng xã hội.

* 1. **Phân phối theo cảm xúc/lượt thích (Sentiment/Likes):**

Thực hiện thống kê số lượng lượt thích dựa trên các bài đăng. Tính trung bình số lượt thích trên các bài đăng tích cực/tiêu cực/trung lập.

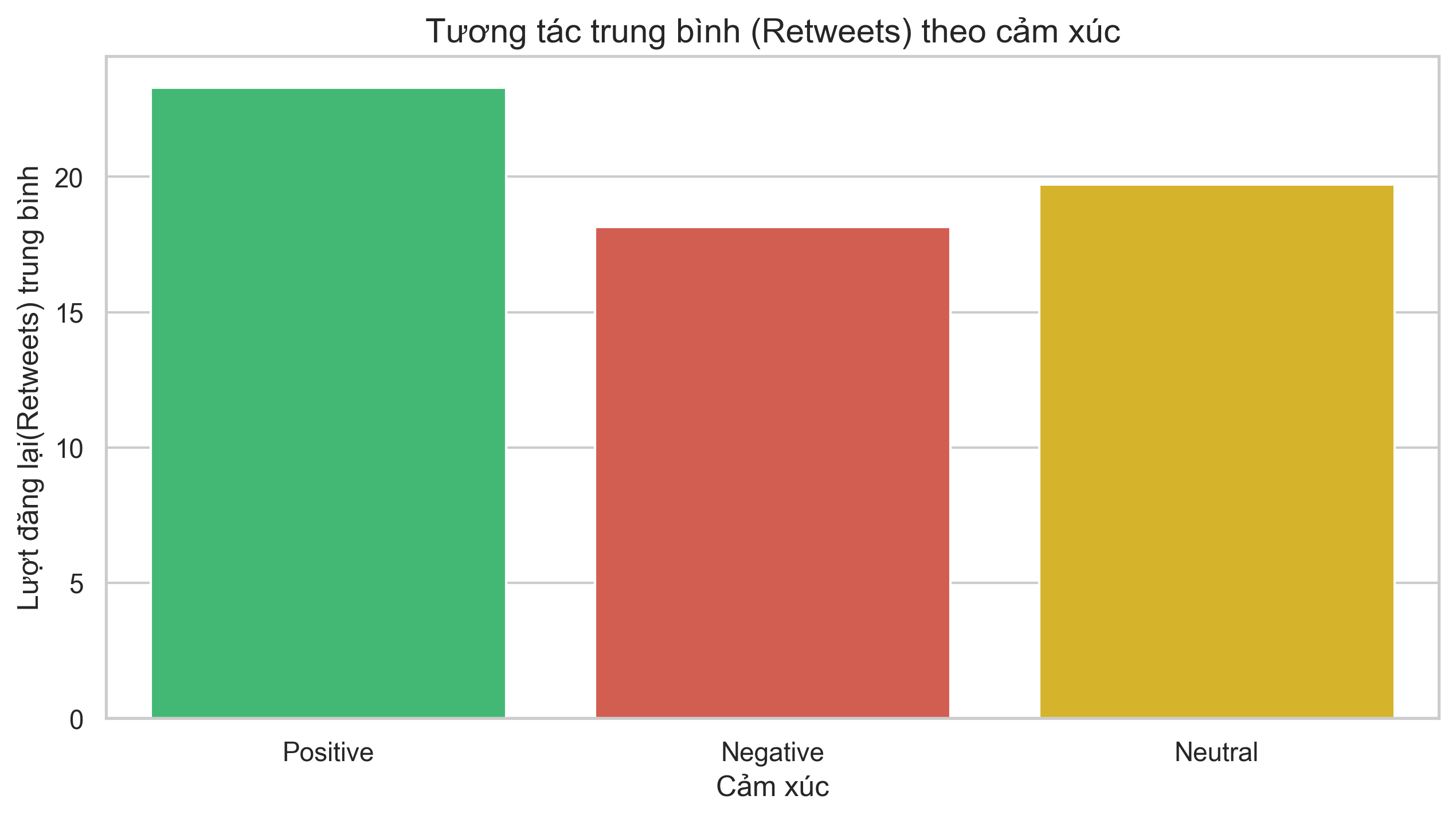


Hình 23: Biểu đồ lượt tương tác trung bình Likes theo cảm xúc

**Nhận xét:** Biểu đồ cho thấy **bài viết có cảm xúc tích cực (Positive)** nhận được lượng **like trung bình cao nhất**, trong khi **cảm xúc tiêu cực (Negative)** **có số lượt like thấp nhất**. Cảm xúc **trung lập (Neutral**) **nằm ở mức giữa**. Điều này cho thấy người dùng có xu hướng tương tác tích cực hơn với nội dung mang cảm xúc tích cực.

* 1. **Phân phối theo cảm xúc/lượt thích (Sentiment/Retweets):**

Thực hiện thống kê số lượng lượt đăng lại dựa trên các bài đăng. Tính trung bình số lượt đăng lại trên các bài đăng tích cực/tiêu cực/trung lập.

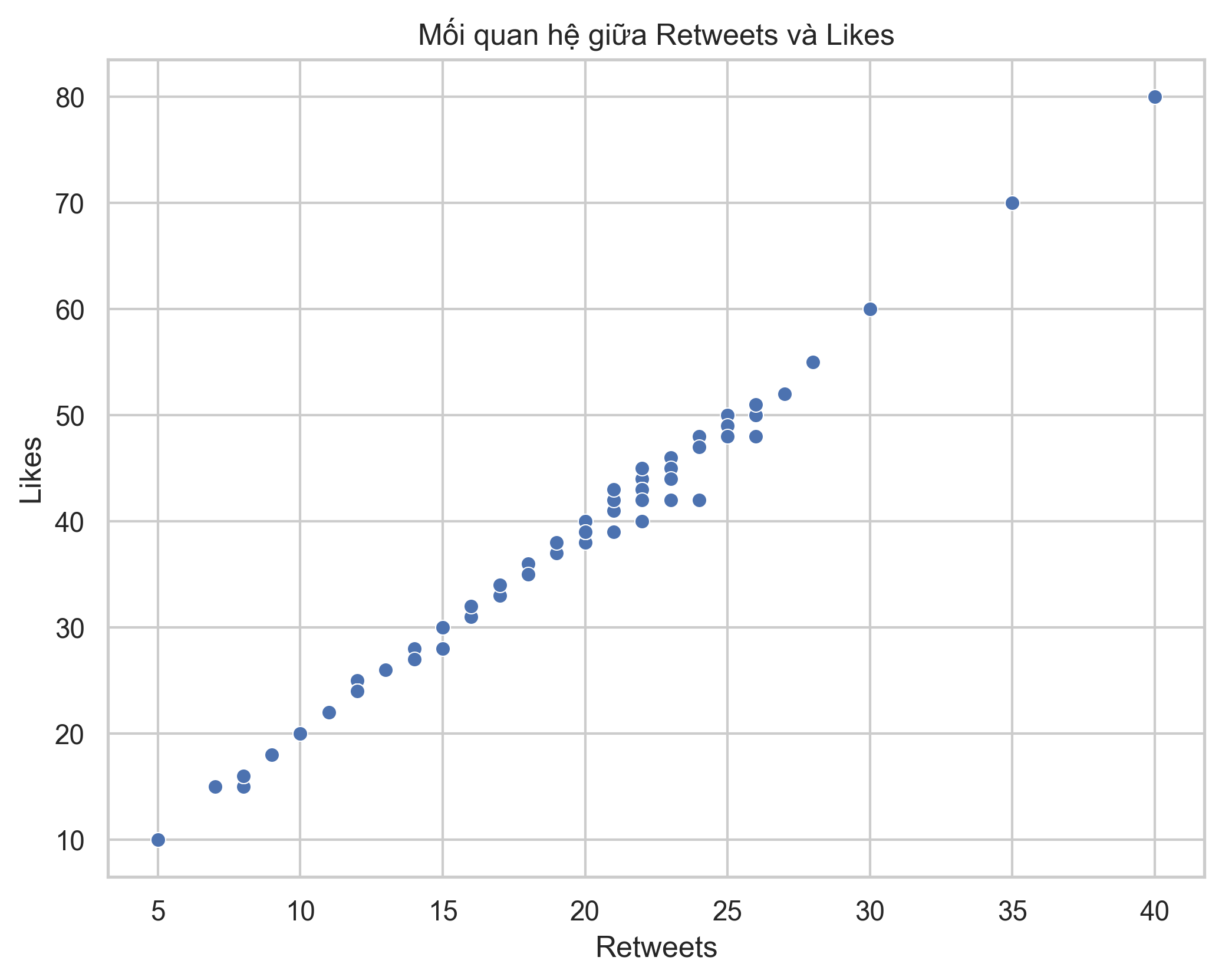


Hình 24: Biểu đồ lượt tương tác trung bình Retweets theo cảm xúc

**Nhận xét:** Tương tự như biểu đồ **Sentiment/Likes** biểu đồ cho thấy **bài viết có cảm xúc tích cực (Positive)** nhận được lượng **retweet trung bình cao nhất**, trong khi **cảm xúc tiêu cực (Negative)** **có số lượt retweet thấp nhất**. Cảm xúc **trung lập (Neutral**) **nằm ở mức giữa**. Điều này cho thấy người dùng có xu hướng tương tác tích cực hơn với nội dung mang cảm xúc tích cực.

* 1. **Phân phối theo lượt thích/lượt đăng lại (Likes/Retweets):**

Thực hiện thống kê số lượt thích/đăng lại của các bài đăng. Tìm mối tương quan giữa lượt thích và lượt đăng lại.



Hình 25: Biểu đồ mối quan hệ giữa Likes và Retweets

**Nhận xét:** Biểu đồ cho thấy mối tương quan tuyến tính rõ rệt giữa số lượt Retweets và Likes: khi Retweets tăng thì Likes cũng tăng đều, cho thấy những bài viết được chia sẻ nhiều thường cũng được yêu thích nhiều, phản ánh hiệu ứng lan tỏa tích cực từ người dùng.

# **Phần V: Khai phá dữ liệu**

## **Đánh giá tổng thể dữ liệu**

Phần này chúng ta sẽ tiến hành đánh giá lại dữ liệu sau khi đã phân tích EDA để tìm ra hướng đi và mô hình phù hợp.

### **Kích thước và cấu trúc dữ liệu**

**Tổng số mẫu sau khi làm sạch**: 732 dòng dữ liệu hợp lệ.

**Cấu trúc dữ liệu** gồm 2 cột chính mà chúng ta cần quan tâm:

* **Text:** chứa nội dung văn bản tiếng Anh.
* **Sentiment:** chứa nhãn cảm xúc tương ứng, gồm 3 giá trị:
  + **Positive:** 448 mẫu.
  + **Negative:** 182 mẫu.
  + **Neutrail:** 102 mẫu.

### **Chất lượng dữ liệu:**

**Không có giá trị thiếu nghiêm trọng:** các dòng thiếu Text hoặc Sentiment đã được loại bỏ trong quá trình thu thập dữ liệu.

Dữ liệu **mất cân bằng nhãn** đáng kể:

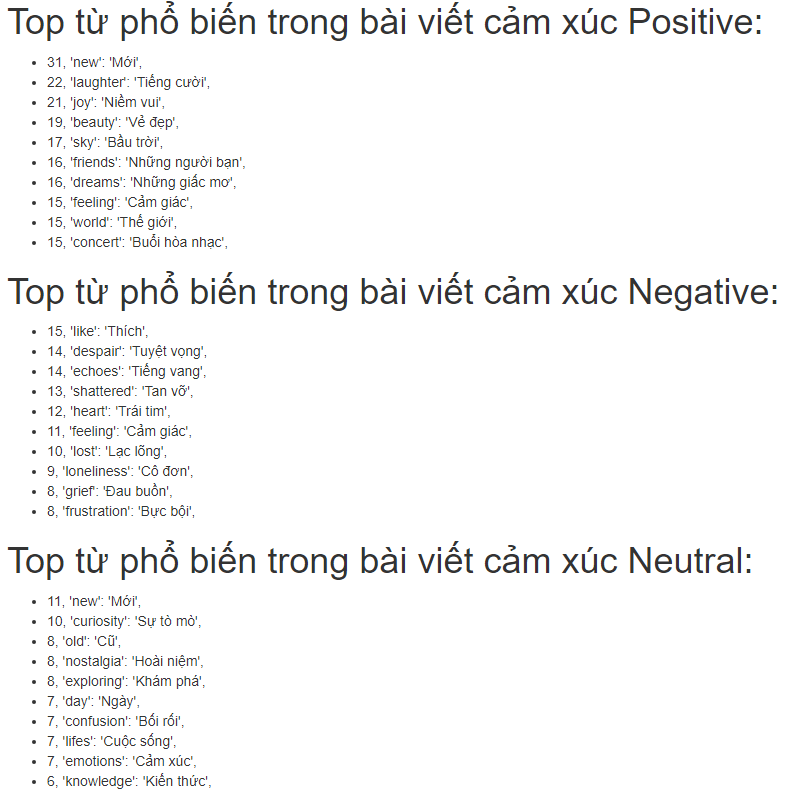
* Positive: ~62.7%
* Negative: ~24.9%
* Neutral: ~13.9%

**Không phát hiện lỗi định dạng:** Văn bản ở dạng chuỗi, không chứa ký tự lỗi hoặc ngoại lệ lớn..

### **Đặc điểm của dữ liệu văn bản:**

Tiến hành phân tích các từ thường xuất hiện trong dữ liệu

for sentiment in ['Positive', 'Negative', 'Neutral']:  
 subset = df[df['Sentiment'] == sentiment]  
 *# Tách văn bản thành danh sách các từ* all\_words = ' '.join(subset['clean\_text']).split()  
 counter = Counter(all\_words)  
 common\_words = counter.most\_common(10)  
   
 print(f"\nTop từ phổ biến trong bài viết cảm xúc {sentiment}:")  
 for word, freq in common\_words:  
 print(f"- {word}: {freq} lần")



Hình 26: Top những từ được xuất hiện nhiều lần trong các văn bản

*(các số đầu dòng đại diện cho tần suất xuất hiện của từ)*

Đặc điểm chính của cột **“Text”**:

* **Ngôn ngữ:** **tiếng Anh**, chủ yếu là các đoạn văn ngắn, mang tính đánh giá hoặc phản hồi.
* **Độ dài văn bản:** dao động từ **một vài từ đến vài câu**, phần lớn là **câu** **đơn giản** hoặc **cụm từ ngắn**.
* Nội dung thường liên quan đến **cảm xúc hoặc quan điểm**, nhưng không sử dụng nhiều biểu thức đặc trưng như **"great product"** hay **"waste of time".**
* Chỉ có **một vài từ** trong top các từ phổ biến **thể hiện cảm xúc** của đoạn văn, còn lại đều là những từ không liên quan hoặc thể hiện nhóm cảm xúc khác, từ **“new”** xuất hiện trong cả nhóm văn bản **Posivite** và **Neutral**, có thể xem xét loại bỏ khi huấn luyện.
* Dữ liệu mang tính thực tế, thể hiện rõ sự đa dạng trong cách biểu đạt cảm xúc của người dùng.

### **Các bước tiền xử lý đã thực hiện:**

Chuyển văn bản về chữ thường (lowercase).

Loại bỏ:

* URL, email, emoji, ký tự đặc biệt, số.
* Khoảng trắng dư thừa.
* Loại bỏ những từ không cần thiết

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Từ** | **Positive** | **Negative** | **Neutral** | **Nhận xét** |
| **new** | 31 | - | 11 | Rất chung, không mang cảm xúc rõ |
| **feeling** | 15 | 11 | - | Trung tính, không đặc trưng |
| **day** | - | - | 7 | Không mang cảm xúc, nên bỏ |
| **world** | 15 | - | - | Có thể giữ nếu mang sắc thái |
| **like** | - | 15 | - | Mang nghĩa rộng, mơ hồ, nên bỏ |

Vector hóa văn bản bằng:

* TF-IDF hoặc CountVectorizer.
* Áp dụng n-gram từ (1,1) đến (1,3)
* Giới hạn số đặc trưng: 1.000 – 5.000.

### **Đánh giá tổng quan:**

Bộ dữ liệu có chất lượng tốt về mặt nội dung và tính khả thi huấn luyện.

**Tuy nhiên, sự mất cân bằng nhãn** **là điểm yếu lớn**, có thể **gây thiên lệch (bias)** khi huấn luyện.

Dữ liệu phù hợp với các bài toán phân tích cảm xúc cơ bản, nhưng **chưa có thông tin ngữ cảnh hay đặc trưng người dùng**, do đó chưa thể áp dụng cho các hệ thống khuyến nghị hoặc phân tích nâng cao.Nhận xét đánh giá.

**Nhận xét:** Bộ dữ liệu có quy mô nhỏ (732 mẫu) và nội dung văn bản tương đối ngắn, phù hợp với các mô hình học máy truyền thống. Dữ liệu đã được làm sạch kỹ lưỡng và chuẩn hóa nhãn rõ ràng. Tuy nhiên, nhược điểm lớn nhất là sự mất cân bằng nhãn nghiêm trọng, khi số mẫu Positive chiếm hơn 60% toàn bộ dữ liệu, gây nguy cơ thiên lệch trong quá trình huấn luyện mô hình.

Ngoài ra, nội dung văn bản tuy mang tính cảm xúc nhưng không có nhiều mẫu mang biểu thức đặc trưng hoặc phức tạp. Điều này vừa là thuận lợi cho việc huấn luyện mô hình đơn giản, vừa là hạn chế nếu muốn áp dụng vào các hệ thống phân tích cảm xúc nâng cao hơn.

Nhìn chung, bộ dữ liệu đủ tốt cho mục tiêu khảo sát và thử nghiệm các mô hình học máy cơ bản, nhưng cần cải thiện về mặt cân bằng nhãn và mở rộng quy mô nếu muốn áp dụng vào thực tế.

# **Phần VI: Đánh giá và chọn thuật toán**

## **Các mô hình được áp dụng**

Trong quá trình thực nghiệm, tôi đã triển khai và huấn luyện 4 mô hình học máy khác nhau trên cùng một tập dữ liệu đã được tiền xử lý:

* **Logistic Regression**
* **Naive Bayes**
* **Support Vector Machine (SVM)**
* **XGBoost**
  1. **Logistic Regression**

**Logistic Regression [1]** là một thuật toán phân loại tuyến tính phổ biến dùng để dự đoán xác suất xảy ra của một sự kiện. Dù tên gọi là **“Regression”**, nhưng thuật toán được sử dụng chủ yếu trong các bài toán **phân loại nhị phân** và **đa lớp**.

Ưu điểm:

* Đơn giản, dễ triển khai
* Tốc độ huấn luyện nhanh
* Dễ giải thích và trực quan hóa kết quả
  1. **Naive Bayes**

**Naive Bayes [2]** là một nhóm các thuật toán phân loại dựa trên định lý **Bayes**, với giả định “**naive**” rằng các đặc trưng là độc lập với nhau. Mô hình thường được dùng trong các bài toán **phân loại văn bản** do tốc độ nhanh và hiệu quả với dữ liệu rời rạc như từ ngữ.

**Ưu điểm:**

* Huấn luyện nhanh
* Tốt với dữ liệu nhiều chiều như văn bản
* Hiệu quả cả khi dữ liệu ít
  1. **Support Vector Machine (SVM)**

**SVM [3]** là một thuật toán phân loại mạnh mẽ, hoạt động bằng cách tìm siêu phẳng **(hyperplane)** tối ưu để phân tách các lớp dữ liệu trong không gian đặc trưng. **SVM** hoạt động tốt với dữ liệu có chiều cao **(high-dimensional)** như **TF-IDF**.

**Ưu điểm:**

* Hiệu suất cao trên dữ liệu phân lớp rõ
* Phù hợp với dữ liệu nhiều chiều
* Có thể mở rộng qua kernel (phi tuyến)
  1. **XGBoost**

**XGBoost [4] (Extreme Gradient Boosting)** là một thuật toán tăng cường **(boosting)** tối ưu hóa hiệu năng cho các bài toán phân loại và hồi quy. Nó xây dựng nhiều cây quyết định **(decision trees)** liên tiếp, trong đó mỗi cây mới học từ lỗi của cây trước đó.

**Ưu điểm:**

* Hiệu suất mạnh mẽ, thường đứng đầu trong các cuộc thi **(Kaggle, v.v.)**
* Có cơ chế **chống overfitting** (regularization)
* Hỗ trợ xử lý thiếu dữ liệu và song song tốt

Các mô hình này đều sử dụng biểu diễn văn bản theo dạng **TF-IDF với n-gram từ 1 đến 3**, giới hạn số chiều từ **1000**.

## **Các tiêu chí đánh giá**

Hiệu quả của mô hình được đánh giá thông qua các chỉ số sau:

* **Accuracy (Độ chính xác)**
* **Precision / Recall / F1-Score** cho từng nhãn: **Positive, Negative, Neutral.**
* **Thời gian huấn luyện / dự đoán**
* **Độ ổn định khi kiểm thử nhiều lần.**

## **Kết quả so sánh các mô hình**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Thuật toán** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Nhận xét sơ bộ** |
| **Logistic Regression** | 71% | 65% | 72% | 64% | Tốt, nhẹ, ổn định |
| **Naive Bayes** | 75% | 81% | 75% | 70% | Nhanh nhưng kém cân bằng |
| **SVM** | 80% | 85% | 81% | 79% | Hiệu suất tốt nhất |
| **XGBoost** | 68% | 66% | 68% | 63% | Hiệu suất kém, có dấu hiệu overfitting. |

Bảng 2: Kết quả so sánh các thuật toán

Dựa vào kết quả trong bảng, nhận xét sơ bộ cho từng thuật toán:

**Logistic Regression**: Hiệu suất trung bình với Accuracy 71%, F1-Score 64%. Cân bằng giữa tốc độ và hiệu suất, phù hợp làm baseline.

**Naive Bayes**: Hiệu suất khá tốt với Accuracy 75%, Precision 81%, và F1-Score 70%. Mô hình đơn giản, nhanh nhưng có thể gặp vấn đề khi giả định về tính độc lập của đặc trưng không đúng.

**SVM**: Hiệu suất cao nhất trong các mô hình với Accuracy 80%, Precision 85%, Recall 81% và F1-Score 79%. Thích hợp cho dữ liệu có nhiều chiều. Đây sẽ là thuật toán phù hợp để giải quyết bài toán đã đề ra

**XGBoost**: Hiệu suất thấp nhất với Accuracy 68%, F1-Score 63%. Bất ngờ kém hiệu quả trong trường hợp này, có dấu hiệu overfitting. Cần điều chỉnh tham số để tối ưu hóa.

# **Phần VII: Kết quả và thảo luận**

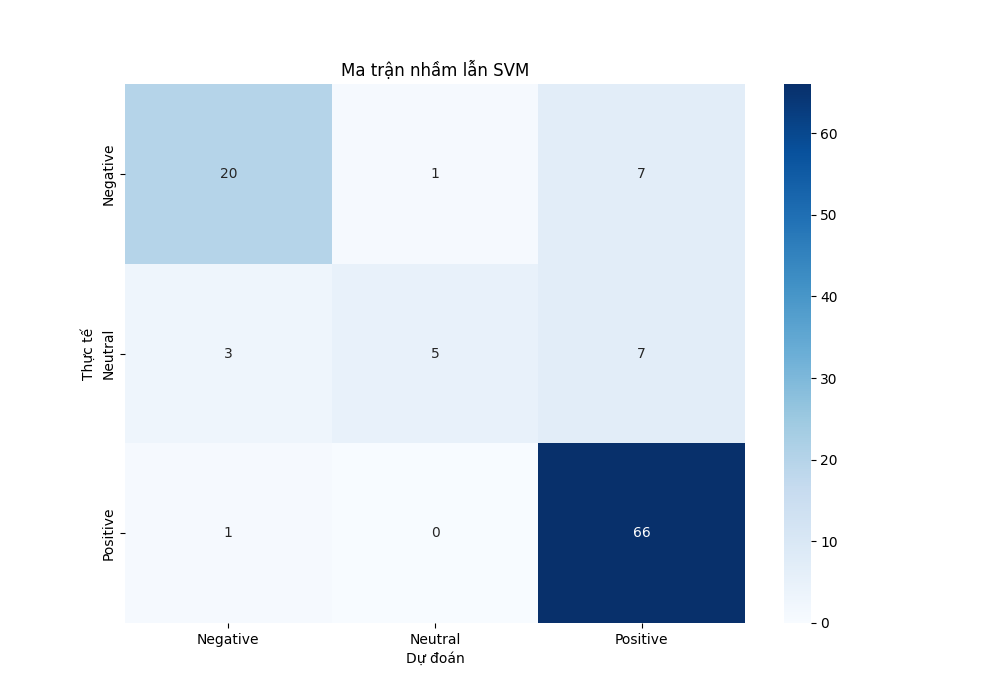
## **Kết quả trên tập kiểm tra (Test set)**

Sau khi lựa chọn mô hình **Support Vector Machine (SVM**) là mô hình phù hợp nhất dựa trên hiệu suất trên tập validation, mô hình đã được đánh giá lại trên tập kiểm tra (test set) – tập dữ liệu hoàn toàn chưa từng được sử dụng trong quá trình huấn luyện hoặc chọn mô hình.

* **Accuracy:** 82.7%
* **Weighted Precision:** 83%
* **Weighted Recall:** 83%
* **Weighted F1-Score:** 81%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nhãn** | **Precision** | **Recall** | **F1-score** | **Số mẫu** |
| **Negative** | 83% | 71% | 77% | 28 |
| **Neutral** | 83% | 33% | 48% | 15 |
| **Positive** | 82% | 99% | 90% | 67 |

Bảng 3: Kết quả trên tập kiểm tra (Test set)



Bảng 4: Ma trận nhầm lẫn SVM

## **Thảo luận**

Dựa trên ma trận nhầm lẫn thu được từ mô hình SVM trên tập kiểm tra, có thể rút ra các nhận định sau:

**Lớp Positive:**

* Mô hình hoạt động rất tốt với lớp này khi **66/67 mẫu được phân loại chính xác.**
* Recall đạt khoảng 98.5%, cho thấy mô hình gần như không bỏ sót các văn bản mang cảm xúc tích cực.
* Đây là lớp có hiệu suất cao nhất trong cả 3 lớp.

**Lớp Negative:**

* Hiệu suất ở mức khá, với **20/28 mẫu được phân loại đúng**, đạt Recall khoảng 71.4%.
* Tuy nhiên, vẫn còn **7 mẫu Negative bị nhầm thành Positive**, phản ánh việc mô hình đôi khi đánh giá cảm xúc tiêu cực là tích cực nếu văn bản không rõ ràng.

**Lớp Neutral:**

* Là lớp khó phân loại nhất đối với mô hình này.
* Chỉ **5/15 mẫu được phân loại đúng** → Recall chỉ đạt khoảng 33.3%.

Phần lớn các mẫu Neutral bị nhầm sang Positive hoặc Negative, cho thấy mô hình gặp khó khăn khi xác định cảm xúc trung lập, có thể do:

Văn bản trung lập thường không có từ ngữ cảm xúc rõ rệt.

Số lượng mẫu Neutral trong tập huấn luyện ít hơn đáng kể so với các lớp còn lại.C

## **Nhận xét tổng quan**

**Mô hình SVM thể hiện hiệu suất rất tốt với lớp Positive**, giữ được độ chính xác cao và khả năng phát hiện chính xác.

**Lớp Negative được nhận diện tương đối tốt**, nhưng vẫn có tỷ lệ nhầm lẫn sang Positive.

**Lớp Neutral là điểm yếu chính**, cho thấy mô hình chưa thể hiện được khả năng phân biệt cảm xúc không rõ ràng hoặc trung lập.

**Mất cân bằng dữ liệu** là một trong những nguyên nhân dẫn đến hiện tượng này, do số lượng mẫu Neutral ít hơn đáng kể.

Nhìn chung, mô hình **SVM** có **hiệu suất tổng thể cao**, khả năng tổng quát hóa tốt, nhưng cần được cải tiến thêm nếu muốn xử lý các văn bản mang cảm xúc trung tính một cách chính xác hơn.

# **Phần VIII: kết luận**

## **Kết luận sơ bộ**

Trong dự án khai phá dữ liệu này, tôi đã thực hiện xây dựng hệ thống phân loại cảm xúc từ văn bản trên mạng xã hội bằng cách ứng dụng các thuật toán học máy. Dữ liệu đầu vào là tập văn bản tiếng Anh chứa cảm xúc người dùng được thu thập và xử lý kỹ lưỡng thông qua các bước làm sạch, chuẩn hóa và biểu diễn bằng TF-IDF/CountVectorizer.

Triển khai và so sánh hiệu quả của 4 thuật toán: **Naive Bayes, Logistic Regression, XGBoost và Support Vector Machine (SVM)**. Qua quá trình đánh giá hiệu suất trên tập validation và kiểm tra lại trên tập test, **SVM** được lựa chọn là mô hình tối ưu nhất với **accuracy đạt 80%** và **F1-score đạt 79%** trên tập validation, **81% trên tập test (weighted average).**

Tuy mô hình hoạt động rất tốt ở lớp **Positive**, nhưng gặp khó khăn trong việc phân loại các văn bản mang cảm xúc trung lập **(Neutral).** Điều này đến từ **sự mất cân bằng dữ liệu**, một yếu tố cần cải thiện trong các phiên bản tiếp theo của hệ thống.

## **Kết luận tổng quát**

Quá trình khai phá dữ liệu và ứng dụng học máy đã giúp xây dựng được một hệ thống phân loại cảm xúc hiệu quả với độ chính xác cao. Tuy nhiên vẫn còn nhầm lần do sự mất cân bằng trong dữ liệu.

Các mô hình học máy cổ điển như **SVM** vẫn cho thấy khả năng áp dụng mạnh mẽ trong các bài toán **NLP** nếu được tiền xử lý đúng cách.

**Dự án có thể được mở rộng bằng cách:**

* Bổ sung thêm dữ liệu để cân bằng các lớp cảm xúc.
* Áp dụng các mô hình ngôn ngữ hiện đại như **BERT** hoặc **Transformer-based models** để nâng cao độ hiểu ngữ nghĩa.
* Tích hợp mô hình vào các ứng dụng thực tế như phân tích phản hồi người dùng, chatbot cảm xúc, hệ thống gợi ý, hoặc hệ thống kiểm duyệt bài viết trên mạng xã hội (đây có lẽ là 1 trong những hệ thống phù hợp nhất với mô hình này).

## **Kết luận tổng quát**

|  |  |
| --- | --- |
| [1] | D. W. Hosmer, S. Lemeshow and R. X. Sturdivant, Applied logistic regression (3rd ed.), Wiley, 2013. |
| [2] | J. D. M. Rennie, L. Shih, J. Teevan and D. R. Karger, "Tackling the poor assumptions of naive Bayes text classifiers," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003. |
| [3] | C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, p. 273–297, 1995. |
| [4] | T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. |