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 MXH

Giảng viên: ThS. Nguyễn Thanh Phước.

Lớp : DCT1223; T7 - Tiết 6 - 10.

Thực hiện: Trần Ngô Nhật Nam - 3122410253

TPHCM, ngày 15 tháng 5 năm 2025

MỤC LỤC

Phần I: Giới thiệu 5
1. **Lý do chọn đề tài** 5
2. **Mục đích của đề tài** 5
3. **Phạm vi của đề tài** 5
4. **Các bài toán tương tự trong thực tế** 5

Phần II: Mô tả dữ liệu 6
1. **Giới thiệu về bộ dữ liệu** 6
2. **Cấu trúc bộ dữ liệu** 6

Phần III: Xử lý dữ liệu 7
1. **Chia lại bộ dữ liệu** 7
   1.1. **Gom nhóm các các xúc với nhau thành 3 nhóm chính:** 7
   1.2. **Ánh xạ các cảm xúc vào nhóm của nó** 8
   1.3. **Tiến hành phân chia tạo dữ liệu mới** 9
2. **Xử lý dữ liệu thiếu** 10
3. **Làm sạch và định dạng dữ liệu** 11

Phần IV: Phân tích dữ liệu 14
1. **Phân tích đơn biến** 14
   1.1. **Phân phối cảm xúc (Sentiment):** 14
   1.2. **Phân phối theo ngày, giờ (Timestamp):** 15
   1.3. **Phân phối theo nền tảng mạng xã hội (Platform):** 16
   1.4. **Phân phối theo quốc gia (Country):** 17
   1.5. **Phân phối theo người dùng (User):** 18
   1.6. **Phân phối theo lượt tương tác (Like/Retwwets):** 19
2. **Phân tích đa biến** 20
   2.1. **Phân phôi ma trận tương qua giữa các thuộc tính:** 20
3. **Phân tích tương quan** 21
   3.1. **Phân phối theo cảm xúc/quốc gia (Sentiment/Country):** 21
   3.2. **Phân phối theo cảm xúc/thời gian (Sentiment/Timestamp):** 22
   3.3. **Phân phối theo cảm xúc/nền tảng MXH (Sentiment/Platform):** 24
   3.4. **Phân phối theo cảm xúc/lượt thích (Sentiment/Likes):** 25
   3.5. **Phân phối theo cảm xúc/lượt thích (Sentiment/Retweets):** 26

3.6. Phân phối theo lượt thích/lượt đăng lại (Likes/Retweets): 27
Phần V: Khai phá dữ liệu 28
1. Đánh giá tổng thể dữ liệu 28
1.1. Kích thước và cấu trúc dữ liệu 28
1.2. Chất lượng dữ liệu: 28
1.3. Đặc điểm của dữ liệu văn bản: 29
1.4. Các bước tiền xử lý đã thực hiện: 30
1.5. Đánh giá tổng quan: 31
1.6. Xử lý mất cân bằng dữ liệu: 31
Phần VI: Đánh giá và chọn thuật toán 32
1. Các mô hình được áp dụng 32
1.1. Logistic Regression 32
1.2. Naive Bayes 32
1.3. Support Vector Machine (SVM) 33
1.4. XGBoost 33
2. Các tiêu chí đánh giá 33
3. Kết quả so sánh các mô hình 34
3.1. Logistic Regression 34
3.2. Naive Bayes 34
3.3. Support Vector Machine (SVM) 35
3.4. XGBoost 35
Phần VII: Kết quả và thảo luận 37
1. Kết quả trên tập kiểm tra 37
2. Thảo luận 38
3. Nhận xét tổng quan 39
4. Nhận xét thêm 39
Phần VIII: Kiểm thử mô hình 40
1. Đánh giá tổng thể 40
2. Những vấn đề nổi bật: 41
Phần IX: kết luận 42
1. Kết luận sơ bộ 42
2. Kết luận tổng quát 42
3. Tài liệu tham khảo 43

MỤC LỤC ẢNH

Hình 1: Dữ liệu trước khi chia lại cột Sentiment 9
Hình 2: Dữ liệu sau khi chia lại cột Sentiment 10
Hình 3: Kết quả kiểm tra dữ liệu thiếu 10
Hình 4: Biểu đồ phân phôi bài đăng theo nền tảng MXH trước khi làm sạch dữ liệu 12
Hình 5: Biểu đồ phân phôi bài đăng theo nền tảng MXH sau khi làm sạch dữ liệu 12
Hình 6: Dữ liệu trước khi được xử lý 13
Hình 7:Dữ liệu sau khi được xử lý 13
Hình 8: Biểu đồ số lượng bài đăng theo từng loại cảm xúc 14
Hình 9: Biểu đồ số lượng bài đăng theo các ngày trong tuần 15
Hình 10: Biểu đồ số lượng bài đăng theo từng khung giờ trong ngày 15
Hình 11: Biểu đồ số lượng bài đăng theo nền tảng MXH 16
Hình 12: Biểu đồ số lượng bài đăng theo từng quốc gia 17
Hình 13: Biểu đồ top 10 người dùng tích cực đăng bài nhất 18
Hình 14: Top 10 bài viết có ít/nhiều lượt thích nhất 19
Hình 15: Top 10 bài viết có ít/nhiều lượt đăng lại nhất 19
Hình 16: Ma trận tương quan Likes, Retweets, Month, Hour, Day 20
Hình 17: Biểu đồ top 10 quốc gia có nhiều bài đăng nhất 21
Hình 18: Biểu đồ tỷ lệ cảm xúc trong các bài đăng tại top 10 quốc gia 21
Hình 19: Biểu đồ số lượng bài đăng theo các ngày trong tuần 22
Hình 20: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo các ngày trong tuần 22
Hình 21: Biểu đồ số lượng bài đăng theo các khung giờ trong ngày 23
Hình 22: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo giờ trong ngày 23
Hình 23: Biểu đồ phân phối bài đăng theo nền tảng MXH 24
Hình 24: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo nền tảng MXH 24
Hình 25: Biểu đồ lượt tương tác trung bình Likes theo cảm xúc 25
Hình 26: Biểu đồ lượt tương tác trung bình Retweets theo cảm xúc 26
Hình 27: Biểu đồ mối quan hệ giữa Likes và Retweets 27

Hình 28: Top những từ được xuất hiện nhiều lần trong các văn bản.....29
Hình 29: Kết quả xử lý mất cân bằng nhãn. 31
Hình 30: Kết quả đánh giá mô hình Logistic Regression trên tập Validation. 34
Hình 31: Kết quả đánh giá mô hình Naive Bayes trên tập Validation.....34
Hình 32: Kết quả đánh giá mô hình Support Vector Machine trên tập Validation. 35
Hình 33: Kết quả đánh giá mô hình XGBoost trên tập Validation. 35
Hình 34: Kết quả đánh giá mô hình Support Vector Machine trên tập Test 37
Hình 35: Kết quả kiểm thử mô hình SVM.....40

MỤC LỤC BẢNG

Bảng 1: Các thuộc tính của bộ dữ liệu.....6
Bảng 2: Kết quả so sánh các thuật toán 36
Bảng 3: Kết quả trên tập kiểm tra (Test set) 37
Bảng 4: Ma trận nhầm lẫn SVM.....38

Phần I: Giới thiệu

1. 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.

2. 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 các thuật toán học máy như **Logistic Regression**, **Naive Bayes SVM, XGBoots** 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.

3. 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.

4. 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.

2. **Phát hiện phát ngôn từ 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 và huấn luyện lại có thể hoàn toàn khác tùy thuộc vào mô hình và đầu ra của bài toán.

Phần II: Mô tả dữ liệu

1. Giới thiệu về bộ dữ liệu

Bộ dữ liệu được cung cấp bởi người dùng **Kaggle có tên Kashish Parmar** [1], 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.

2. 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 thuộc tính 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, trung tính, buồn, vui,…).
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 (Instagram, Twitter, Facebook).
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, 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

1. 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 nhóm cảm xúc chính** là **tích cực(Positive), tiêu cực(Negative), trung lập(Neutral).**

1.1. 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',
]
```

1.2. 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 thuộc ba nhóm cảm xúc trên
```

1.3. Tiến hành phân chia tạo dữ liệu mới

```
def convert_sentiment_groups(input_file, output_file):
    try:
        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
```

11 rows 300 rows × 15 columns

	Text	Sentiment	Timesta…	User	Platform
0	Enjoying a beautiful day at the park! ...	Positive	2023-01-1…	User123	Twitter
1	Traffic was terrible this morning. ...	Negative	2023-01-1…	CommuterX	Twitter
2	Just finished an amazing workout! 💪 ...	Positive	2023-01-1…	FitnessFan	Instagram
3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-1…	AdventureX	Facebook
4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-1…	ChefCook	Instagram
...	...	...	...	...	...
295	A playful escapade in the carnival of life, c...	Playful	2018-08-2…	CarnivalDreamer	Facebook
296	Drowning in the abyss of despair, a heart sha...	Despair	2021-09-1…	ShatteredHeart	Instagram
297	Bitterness festering like a venomous vine, en...	Bitterness	2019-03-0…	VenomousSoul	Facebook
298	Wandering through the desert of loneliness, e...	Loneliness	2017-07-2…	DesertWanderer	Twitter
299	Yearning for a touch that's not there, echoes...	Yearning	2020-11-0…	ColdEmbrace	Instagram

Hình 1: Dữ liệu trước khi chia lại cột Sentiment

11 rows 300 rows × 15 columns

	Text	Sentiment	Timest…	User	Platform
0	Enjoying a beautiful day at the park! ...	Positive	2023-01-1…	User123	Twitter
1	Traffic was terrible this morning. ...	Negative	2023-01-1…	CommuterX	Twitter
2	Just finished an amazing workout! 💪 ...	Positive	2023-01-1…	FitnessFan	Instagram
3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-1…	AdventureX	Facebook
4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-1…	ChefCook	Instagram
...	...	...	...	...	...
295	A playful escapade in the carnival of life, c...	Positive	2018-08-2…	CarnivalDreamer	Facebook
296	Drowning in the abyss of despair, a heart sha...	Negative	2021-09-1…	ShatteredHeart	Instagram
297	Bitterness festering like a venomous vine, en...	Negative	2019-03-0…	VenomousSoul	Facebook
298	Wandering through the desert of loneliness, e...	Negative	2017-07-2…	DesertWanderer	Twitter
299	Yearning for a touch that's not there, echoes...	Negative	2020-11-0…	ColdEmbrace	Instagram

Hình 2: Dữ liệu sau khi chia lại cột Sentiment

2. 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())
```

```
Số lượng giá trị thiếu trong mỗi cột:
Unnamed: 0.1    0
Unnamed: 0      0
Text            0
Sentiment       0
Timestamp       0
User            0
Platform        0
Hashtags        0
Retweets        0
Likes           0
Country         0
Year            0
Month           0
Day             0
Hour            0
```

Hình 3: Kết quả kiểm tra dữ liệu thiếu

Nhận xét: dữ liệu đầy đủ, không bị thiếu.

3. 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)
```

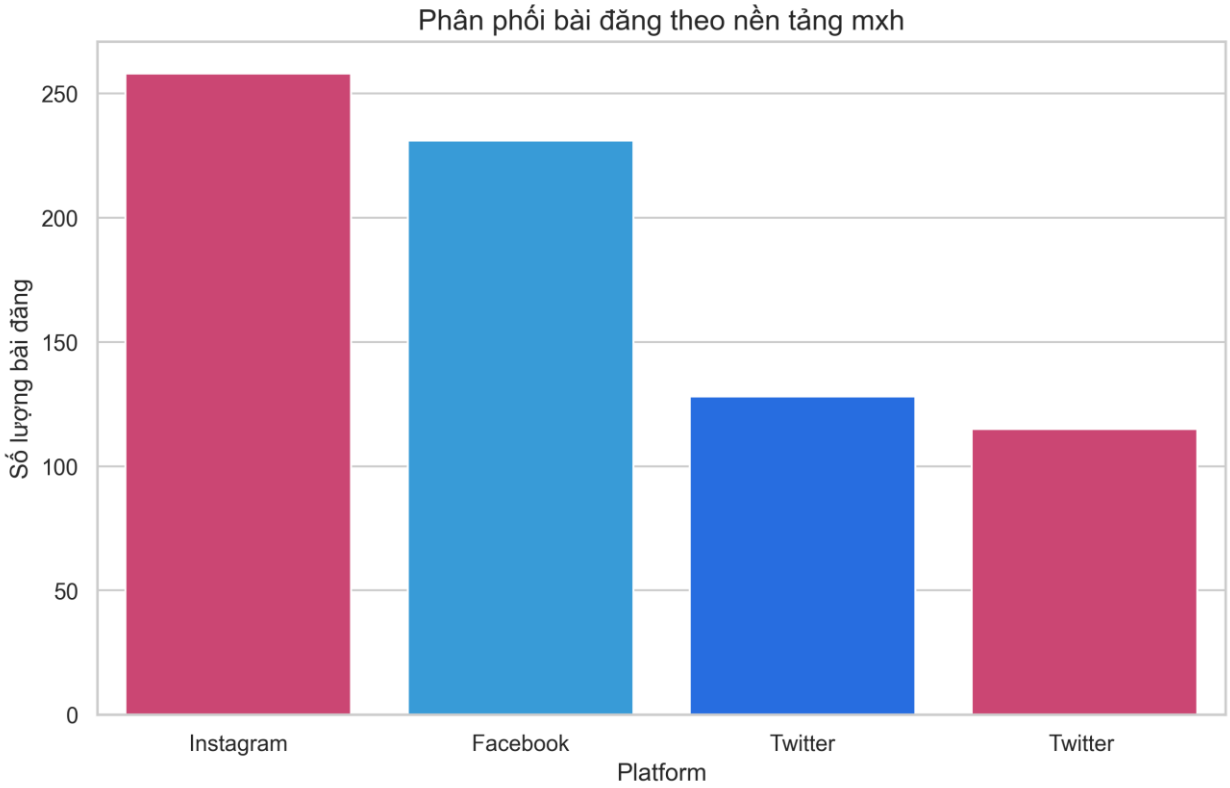
Cột Timestamp được chuyển đổi sang định dạng thời gian (datetime):

```
df['Timestamp'] = pd.to_datetime(df['Timestamp'])
df['DayOfWeek'] = df['Timestamp'].dt.dayofweek
```

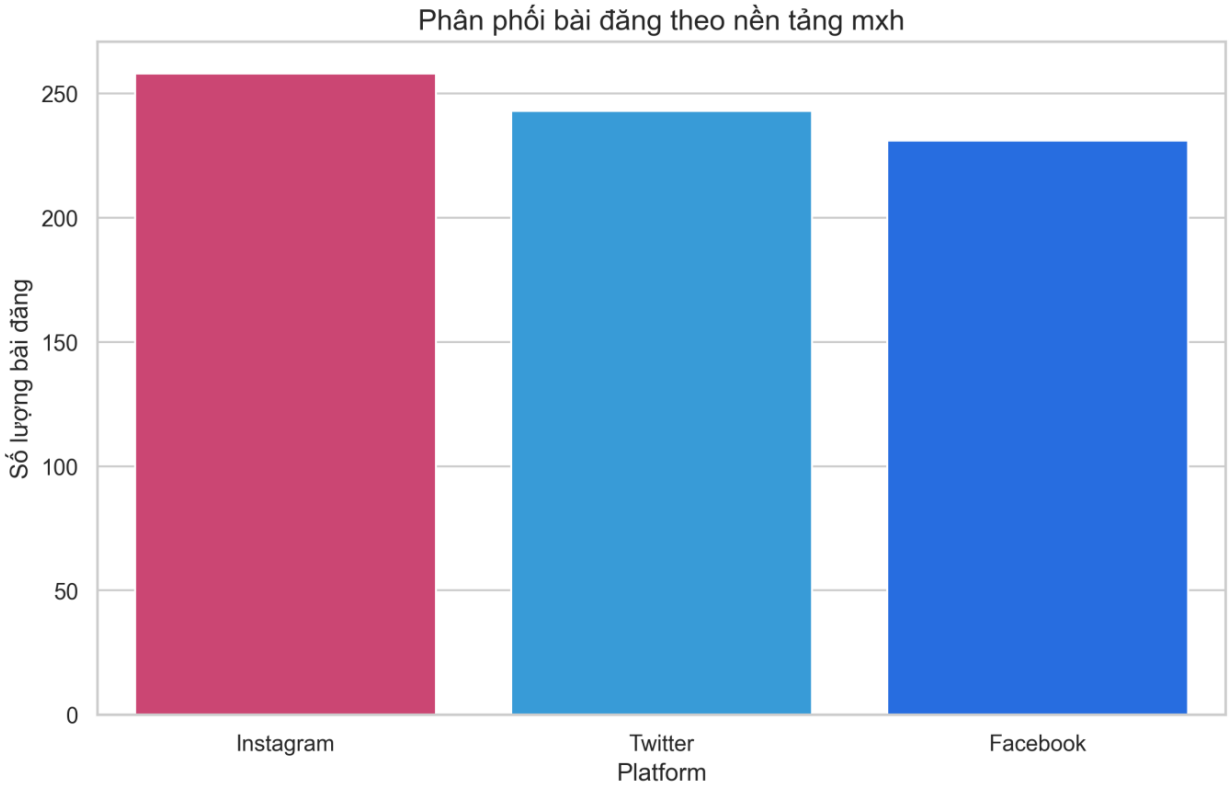
Làm sạch các cột dữ liệu, đảm bảo tính nhất quán của dữ liệu:

```
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 4: Biểu đồ phân phối bài đăng theo nền tảng MXH trước khi làm sạch dữ liệu



Hình 5: Biểu đồ phân phối bài đăng theo nền tảng MXH sau khi làm sạch dữ liệu

20 rows × 15 columns

	Unnam…	Unn…	Text	Sentime…	Timestamp	Plat…	Retw…	Likes
0	0	0	Enjoying a beautiful day at the park! ...	Positive	2023-01-15 …	Twitter	15.0	30.0
1	1	1	Traffic was terrible this morning. ...	Negative	2023-01-15 …	Twitter	5.0	10.0
2	2	2	Just finished an amazing workout! 💪 ...	Positive	2023-01-15 …	Instagr…	20.0	40.0
3	3	3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-15 …	Facebook	8.0	15.0
4	4	4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-15 …	Instagr…	12.0	25.0
5	5	5	Feeling grateful for the little things in lif...	Positive	2023-01-16 …	Twitter	25.0	50.0
6	6	6	Rainy days call for cozy blankets and hot coc...	Positive	2023-01-16 …	Facebook	10.0	20.0
7	7	7	The new movie release is a must-watch! ...	Positive	2023-01-16 …	Instagr…	15.0	30.0
8	8	8	Political discussions heating up on the timel...	Negative	2023-01-17 …	Twitter	30.0	60.0
9	9	9	Missing summer vibes and beach days. ...	Neutral	2023-01-17 …	Facebook	18.0	35.0
10	10	10	Just published a new blog post. Check it out!...	Positive	2023-01-17 …	Instagr…	22.0	45.0
11	11	11	Feeling a bit under the weather today. ...	Negative	2023-01-18 …	Twitter	7.0	15.0
12	12	12	Exploring the city's hidden gems. ...	Positive	2023-01-18 …	Facebook	12.0	25.0
13	13	13	New year, new fitness goals! 💪 ...	Positive	2023-01-18 …	Instagr…	28.0	55.0
14	14	14	Technology is changing the way we live. ...	Neutral	2023-01-19 …	Twitter	15.0	30.0
15	15	15	Reflecting on the past and looking ahead. ...	Positive	2023-01-19 …	Facebook	20.0	40.0
16	16	16	Just adopted a cute furry friend! 🐾 ...	Positive	2023-01-19 …	Instagr…	15.0	30.0
17	17	17	Late-night gaming session with friends. ...	Positive	2023-01-20 …	Twitter	18.0	35.0
18	18	18	Attending a virtual conference on AI. ...	Neutral	2023-01-20 …	Facebook	25.0	50.0
19	19	19	Winter blues got me feeling low. ...	Negative	2023-01-20 …	Instagr…	8.0	15.0

Hình 6: Dữ liệu trước khi được xử lý

20 rows × 15 columns

	Text	Senti…	Timestamp	DayOf…	Platf…	Retw…	Lik…	Engage…
0	Enjoying a beautiful day at the park! ...	Positive	2023-01-15 12:30:00	6	Twitter	15.0	30.0	45.0
1	Traffic was terrible this morning. ...	Negative	2023-01-15 08:45:00	6	Twitter	5.0	10.0	15.0
2	Just finished an amazing workout! 💪 ...	Positive	2023-01-15 15:45:00	6	Instagram	20.0	40.0	60.0
3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-15 18:20:00	6	Facebook	8.0	15.0	23.0
4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-15 19:55:00	6	Instagram	12.0	25.0	37.0
5	Feeling grateful for the little things in lif...	Positive	2023-01-16 09:10:00	0	Twitter	25.0	50.0	75.0
6	Rainy days call for cozy blankets and hot coc...	Positive	2023-01-16 14:45:00	0	Facebook	10.0	20.0	30.0
7	The new movie release is a must-watch! ...	Positive	2023-01-16 19:30:00	0	Instagram	15.0	30.0	45.0
8	Political discussions heating up on the timel...	Negative	2023-01-17 08:00:00	1	Twitter	30.0	60.0	90.0
9	Missing summer vibes and beach days. ...	Neutral	2023-01-17 12:20:00	1	Facebook	18.0	35.0	53.0
10	Just published a new blog post. Check it out!...	Positive	2023-01-17 15:15:00	1	Instagram	22.0	45.0	67.0
11	Feeling a bit under the weather today. ...	Negative	2023-01-18 10:30:00	2	Twitter	7.0	15.0	22.0
12	Exploring the city's hidden gems. ...	Positive	2023-01-18 14:50:00	2	Facebook	12.0	25.0	37.0
13	New year, new fitness goals! 💪 ...	Positive	2023-01-18 18:00:00	2	Instagram	28.0	55.0	83.0
14	Technology is changing the way we live. ...	Neutral	2023-01-19 09:45:00	3	Twitter	15.0	30.0	45.0
15	Reflecting on the past and looking ahead. ...	Positive	2023-01-19 13:20:00	3	Facebook	20.0	40.0	60.0
16	Just adopted a cute furry friend! 🐾 ...	Positive	2023-01-19 17:10:00	3	Instagram	15.0	30.0	45.0
17	Late-night gaming session with friends. ...	Positive	2023-01-20 00:05:00	4	Twitter	18.0	35.0	53.0
18	Attending a virtual conference on AI. ...	Neutral	2023-01-20 11:30:00	4	Facebook	25.0	50.0	75.0
19	Winter blues got me feeling low. ...	Negative	2023-01-20 15:15:00	4	Instagram	8.0	15.0	23.0

Hình 7:Dữ liệu sau khi được xử lý

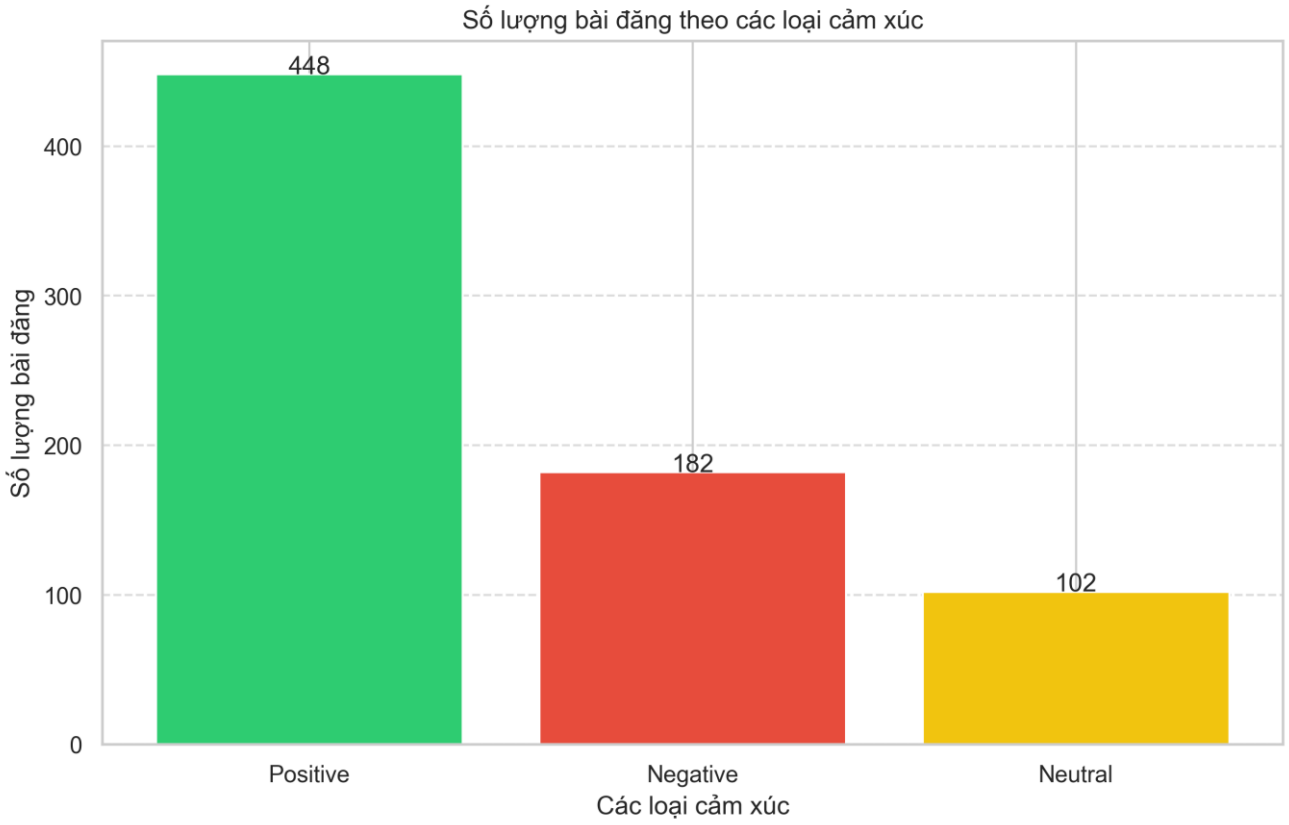
(**Lưu ý:** không tiền xử lý cột **"Text"** vì phân tích **EDA** không cần dùng tới cột **"Text"**)

Phần IV: Phân tích dữ liệu

1. Phân tích đơn biến

1.1. 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.

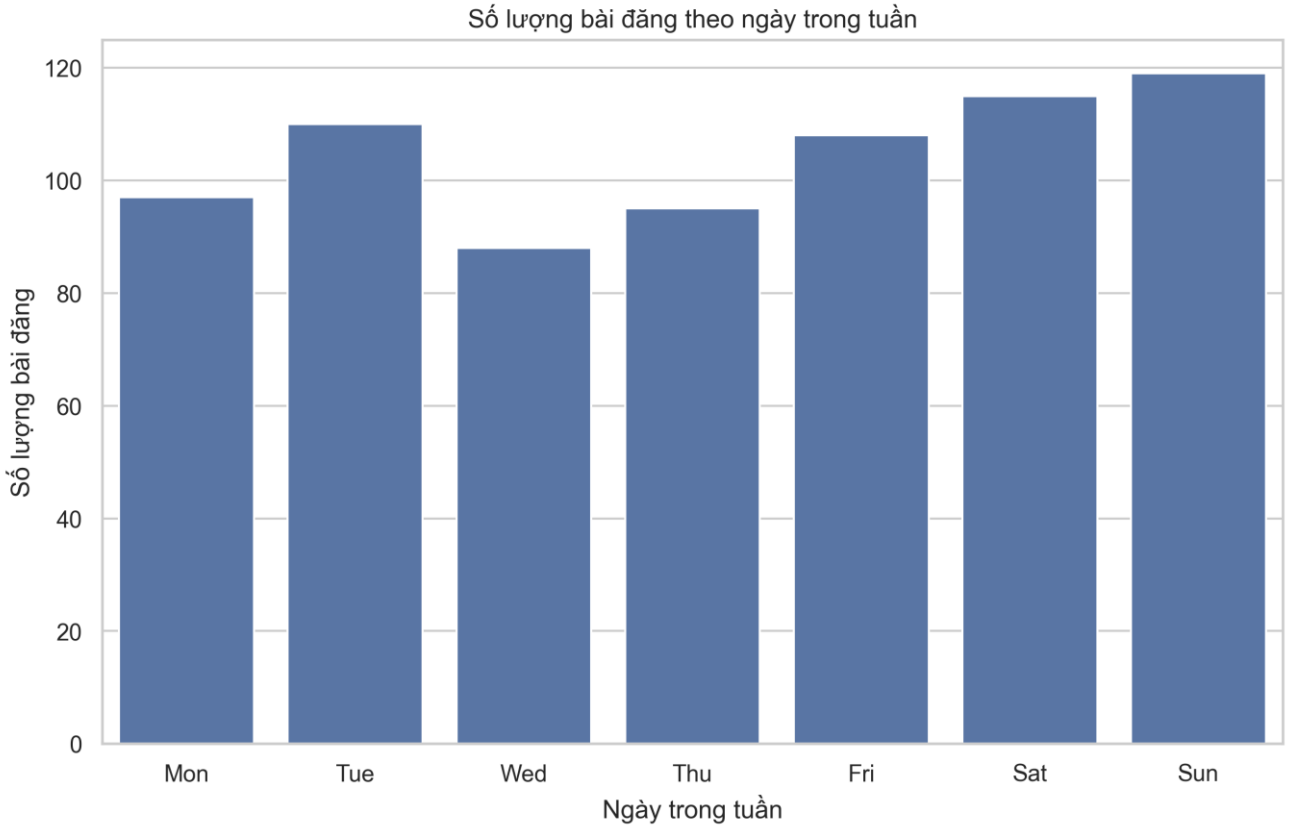


Hình 8: Biểu đồ số lượng bài đăng theo từng loại cảm xúc

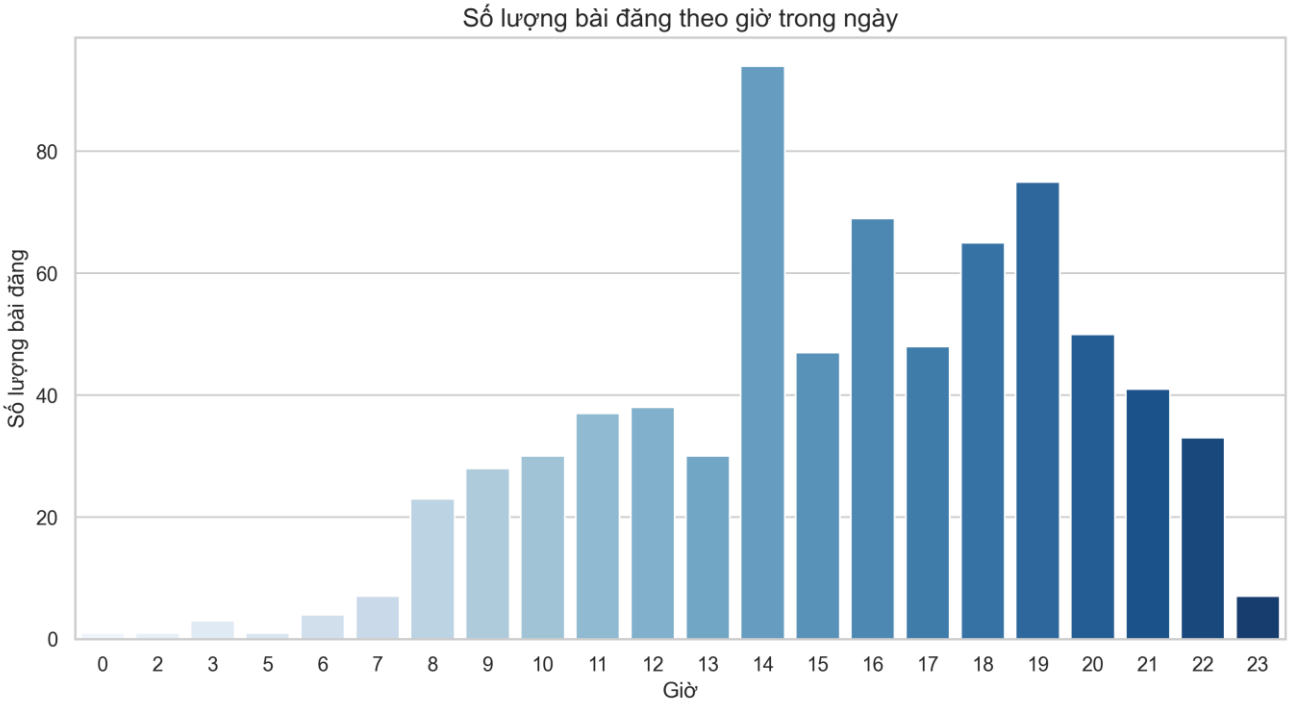
Nhận xét: có sự chênh lệch lớn giữa số lượng bài đăng phân loại theo cảm xúc. Nhìn chung các bài đăng trên các nền tảng mạng xã hội vẫn có xu hướng tích cực.

1.2. Phân phối theo ngày, giờ (Timestamp):

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



Hình 9: Biểu đồ số lượng bài đăng theo các ngày trong tuần

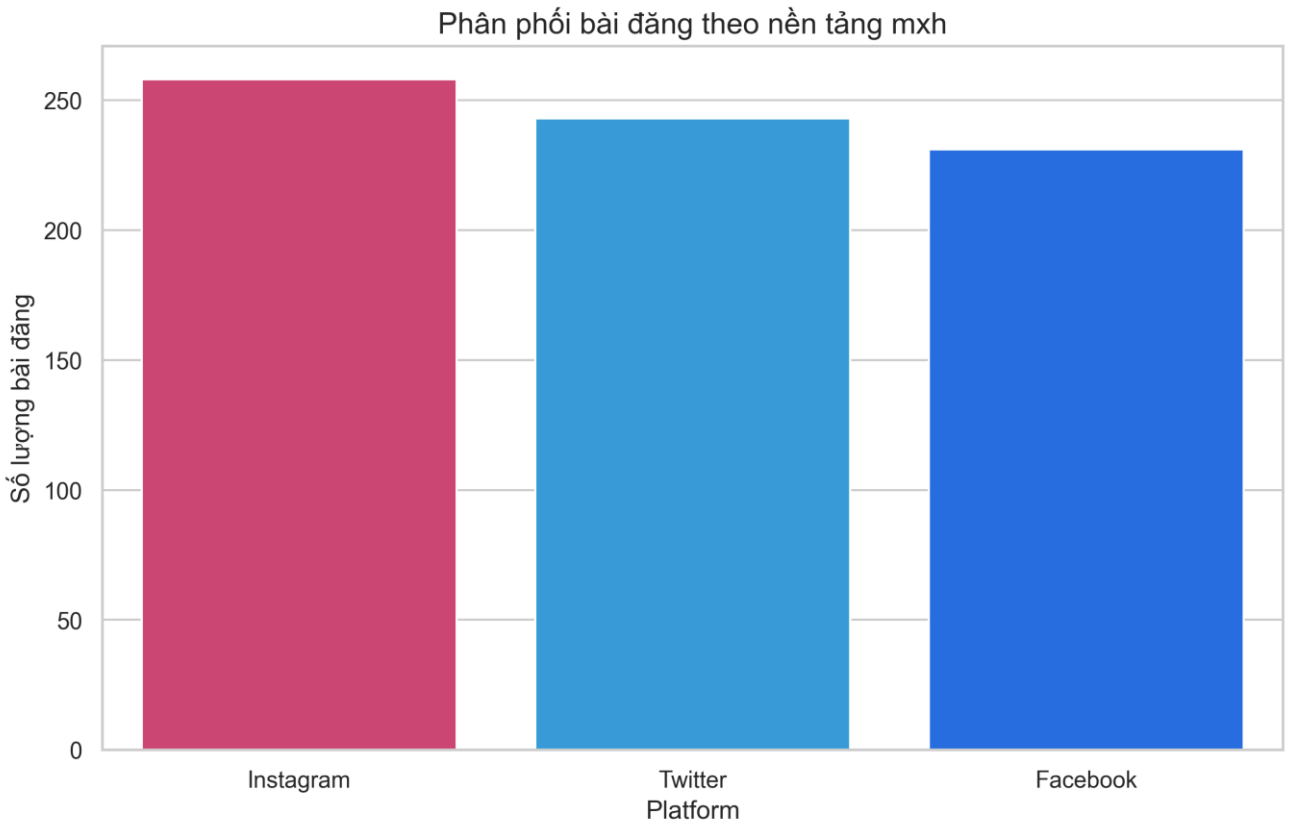


Hình 10: Biểu đồ số lượng bài đăng theo từng khung giờ trong ngày

Nhận xét: Phân tích hai biểu đồ cho thấy có sự khác biệt đáng kể trong việc 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.

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

Thống kê số lượng bài đăng từ mỗi nền tảng. Sử dụng biểu đồ cột để thể hiện số lượng bài đăng từ các nền tảng khác nhau **(Instagram, Twitter, Facebook)**.

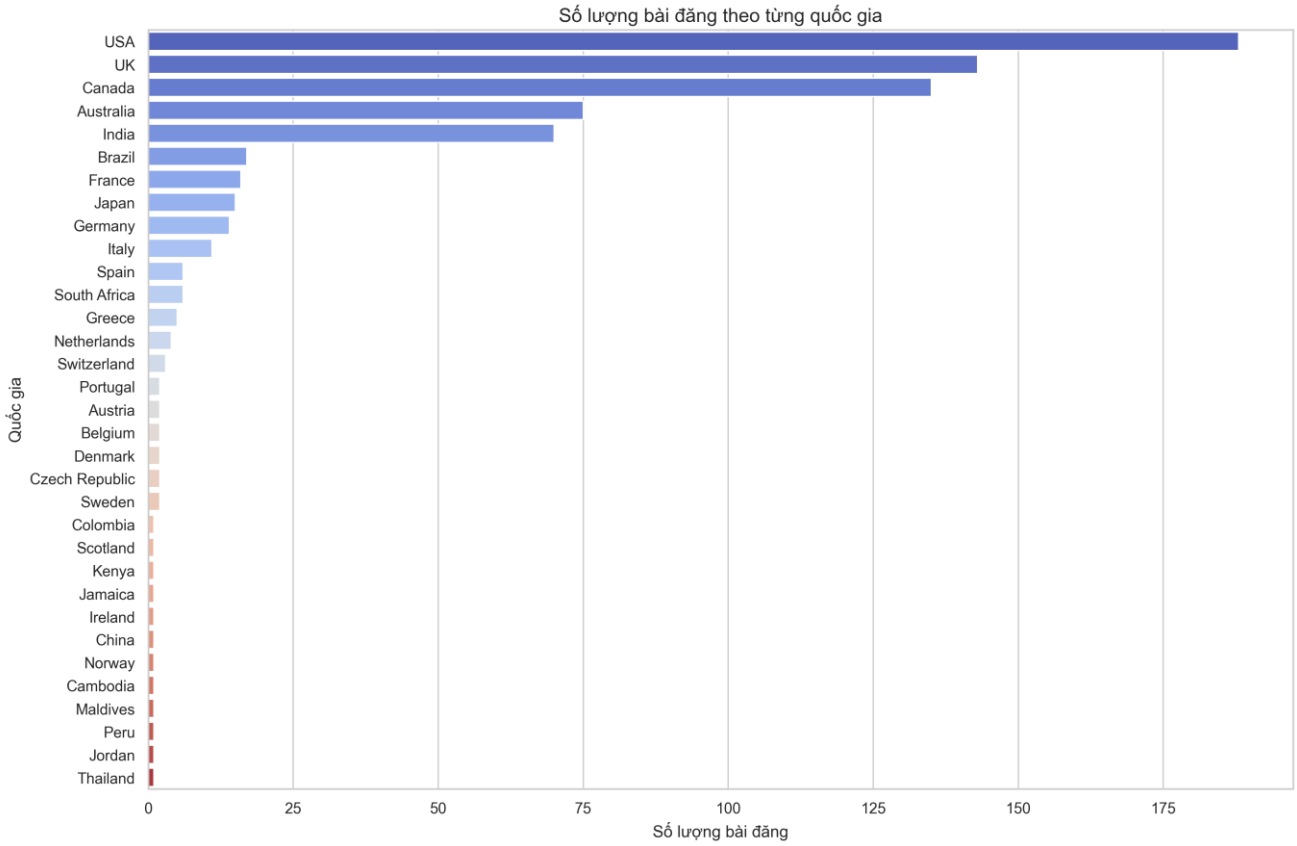


Hình 11: Biểu đồ số lượng bài đăng theo nền tảng MXH

Nhận xét: sự chênh 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.

1.4. Phân phối theo quốc gia (Country):

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

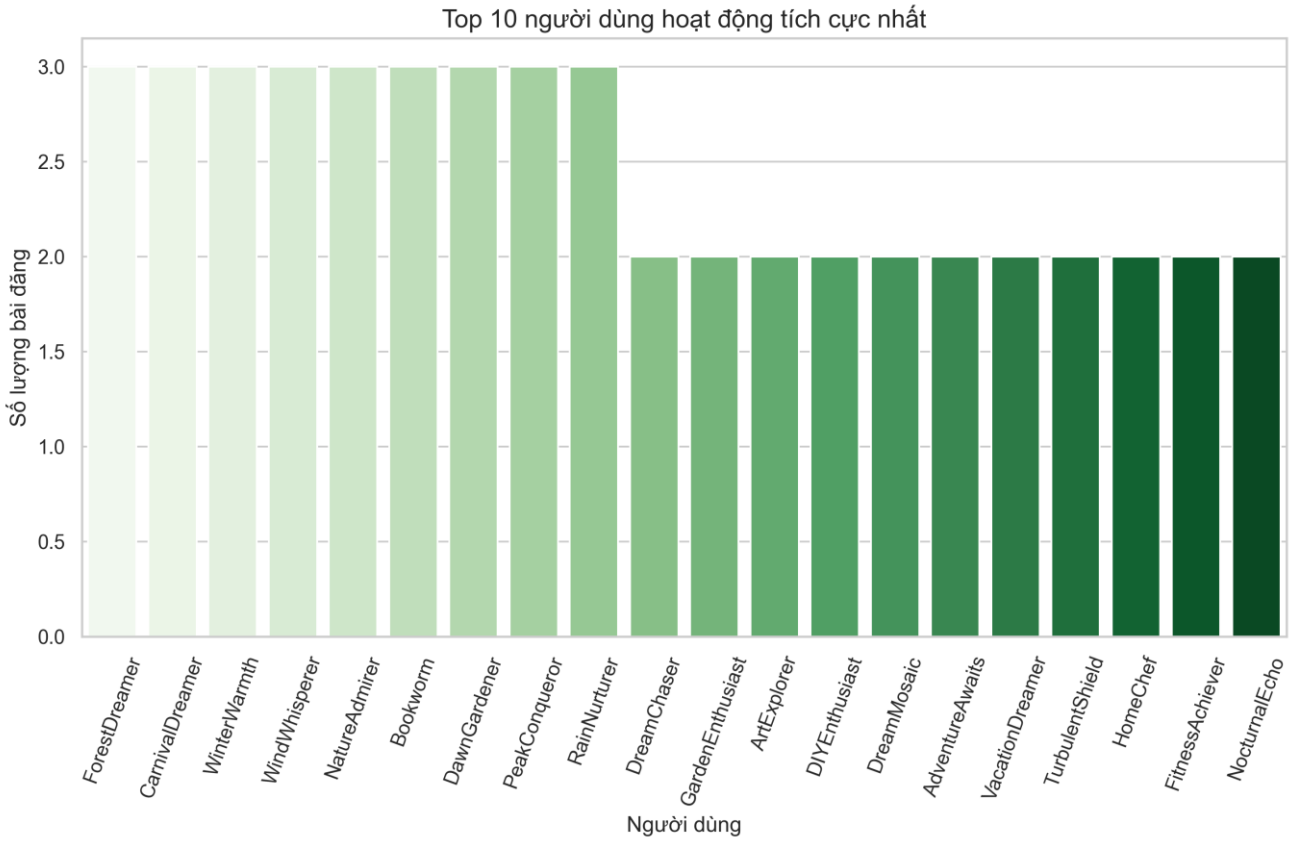


Hình 12: Biểu đồ số lượng bài đăng theo từng quốc gia

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 dữ liệu hoặc mức độ phổ biến của các nền tảng này ở các khu vực đó.

1.5. 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 13: 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.

1.6. 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:

Top 10 bài đăng có nhiều lượt thích nhất:

10 rows × 5 columns

	Text	Likes	Retweets	Sentiment	Platform
335	Thrilled to witness the grandeur of a cultural...	80.0	40.0	Positive	Instagram
345	Motivated to achieve fitness goals after an in...	80.0	40.0	Positive	Facebook
355	Anticipation for an upcoming adventure in an e...	80.0	40.0	Positive	Twitter
368	Elation over discovering a rare book in a quai...	80.0	40.0	Positive	Instagram
382	A sense of wonder at the vastness of the cosmo...	80.0	40.0	Positive	Instagram
402	Awe-inspired by the vastness of the cosmos on ...	80.0	40.0	Positive	Instagram
432	Heartache deepens, a solitary journey through ...	80.0	40.0	Negative	Instagram
470	Dancing on sunshine, each step a celebration o...	80.0	40.0	Positive	Instagram
560	In the serene beauty of a sunset, nature unfol...	80.0	40.0	Positive	Instagram
570	Underneath the city lights, the dancer express...	80.0	40.0	Positive	Twitter

Top 10 bài đăng có ít lượt thích nhất:

10 rows × 5 columns

	Text	Likes	Retweets	Sentiment	Platform
1	Traffic was terrible this morning. ...	10.0	5.0	Negative	Twitter
163	Suffering from despair after another setback....	10.0	5.0	Negative	Twitter
164	Overwhelmed by grief, missing a loved one dea...	15.0	8.0	Negative	Instagram
175	Disgust at the sight of injustice and cruelty...	15.0	7.0	Negative	Twitter
179	Jealousy gnaws at my confidence, a toxic emot...	15.0	8.0	Negative	Instagram
185	Helplessness engulfs me, drowning in a sea of...	15.0	8.0	Negative	Instagram
188	Disgust at the corruption that stains society...	15.0	7.0	Negative	Instagram
195	Boredom lingers, a stagnant pool of indiffere...	15.0	7.0	Negative	Facebook
199	A numbness settles over me, a shield against ...	15.0	8.0	Neutral	Twitter
209	Numb to the chaos, emotions locked away, a st...	15.0	8.0	Neutral	Instagram

Hình 14: Top 10 bài viết có ít/nhiều lượt thích nhất

Retweets:

Top 10 bài đăng có nhiều lượt đăng lại nhất:

10 rows × 5 columns

	Text	Likes	Retweets	Sentiment	Platform
570	Underneath the city lights, the dancer express...	80.0	40.0	Positive	Twitter
560	In the serene beauty of a sunset, nature unfol...	80.0	40.0	Positive	Instagram
550	After a series of defeats, the soccer team fac...	80.0	40.0	Negative	Twitter
540	Celebrating a historic victory in the World Cu...	80.0	40.0	Positive	Instagram
530	Captivated by the spellbinding plot twists, th...	80.0	40.0	Positive	Twitter
520	At a Justin Bieber concert, the infectious bea...	80.0	40.0	Positive	Instagram
510	At the front row of Adele's concert, each note...	80.0	40.0	Positive	Instagram
481	Surrounded by the colors of joy, a canvas pain...	80.0	40.0	Positive	Instagram
402	Awe-inspired by the vastness of the cosmos on ...	80.0	40.0	Positive	Instagram
382	A sense of wonder at the vastness of the cosmo...	80.0	40.0	Positive	Instagram

Top 10 bài đăng có ít lượt đăng lại nhất:

10 rows × 5 columns

	Text	Likes	Retweets	Sentiment	Platform
1	Traffic was terrible this morning. ...	10.0	5.0	Negative	Twitter
163	Suffering from despair after another setback....	10.0	5.0	Negative	Twitter
182	Boredom settles like dust, life feels mundane...	15.0	7.0	Negative	Instagram
188	Disgust at the corruption that stains society...	15.0	7.0	Negative	Instagram
167	Resentment building up over past betrayals. ...	15.0	7.0	Negative	Instagram
175	Disgust at the sight of injustice and cruelty...	15.0	7.0	Negative	Twitter
195	Boredom lingers, a stagnant pool of indiffere...	15.0	7.0	Negative	Facebook
197	Lost in the vast sea of information, an indif...	15.0	7.0	Neutral	Instagram
11	Feeling a bit under the weather today. ...	15.0	7.0	Negative	Twitter
185	Helplessness engulfs me, drowning in a sea of...	15.0	8.0	Negative	Instagram

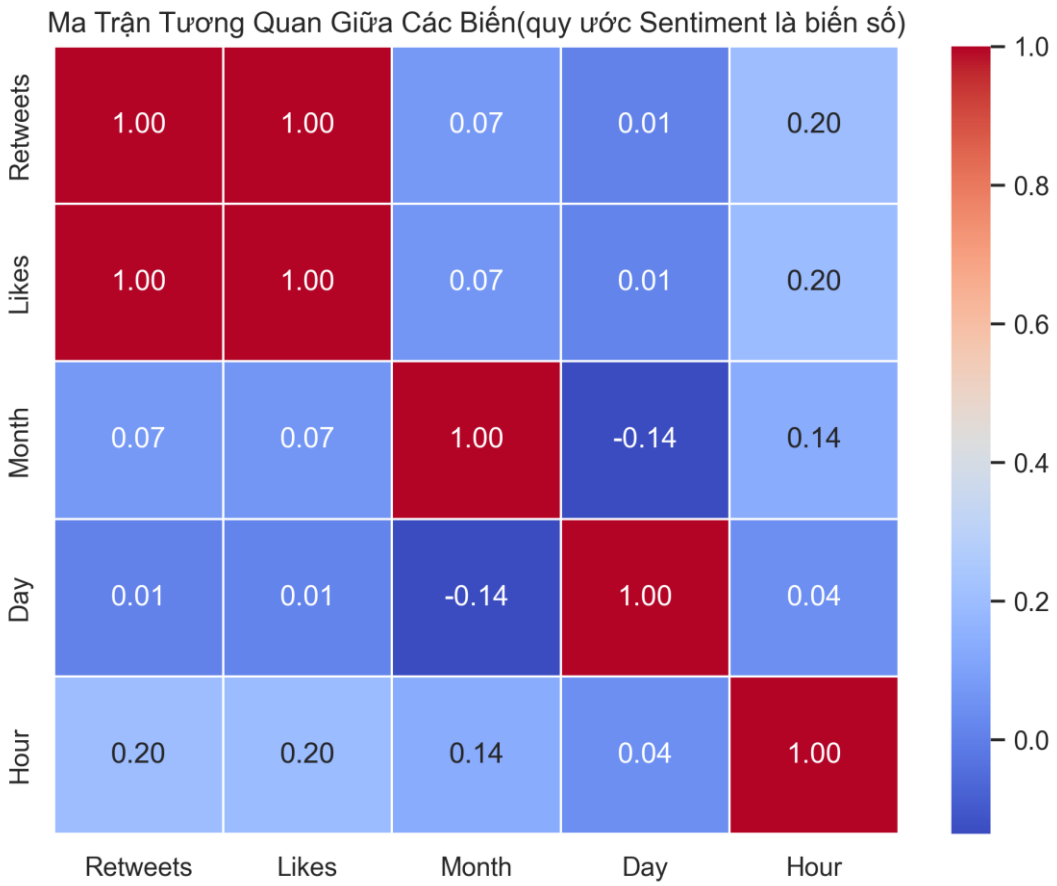
Hình 15: Top 10 bài viết có ít/nhiều lượt đăng lại nhất

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.

2. Phân tích đa biến

2.1. Phân phôi ma trận tương qua giữa các thuộc tính:

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 16: Ma trận tương quan Likes, Retweets, Month, Hour, Day

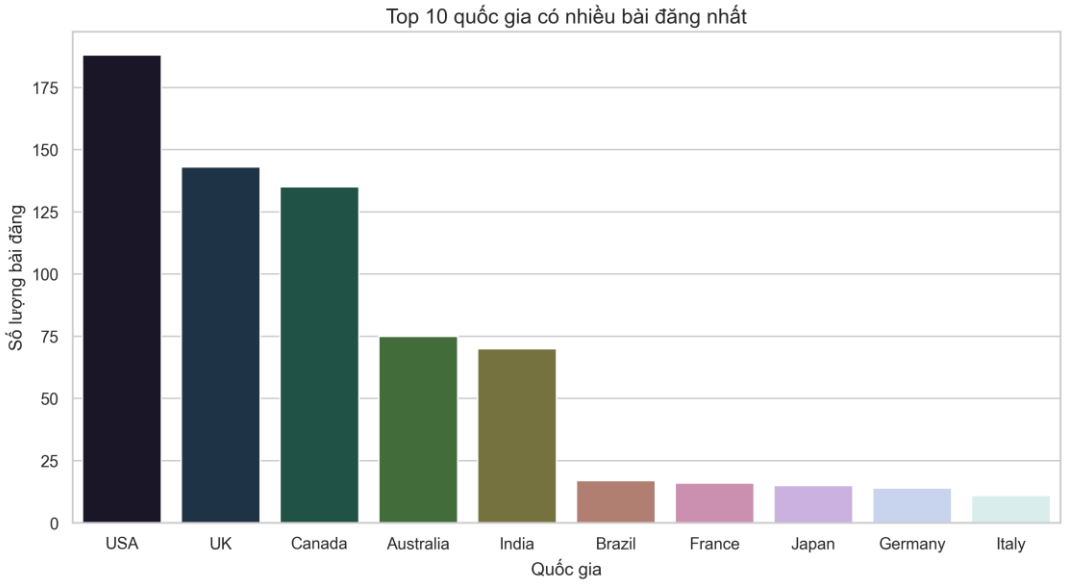
Nhận xét: Ma trận tương quan cho thấy **Retweets** và **Likes** có mối quan hệ chặt chẽ (tương quan 1.00), trong khi các yếu tố thời gian như **Month, Day**, và **Hour** không ảnh hưởng mạnh đến hành vi tương tác của người dùng.

3. Phân tích tương quan

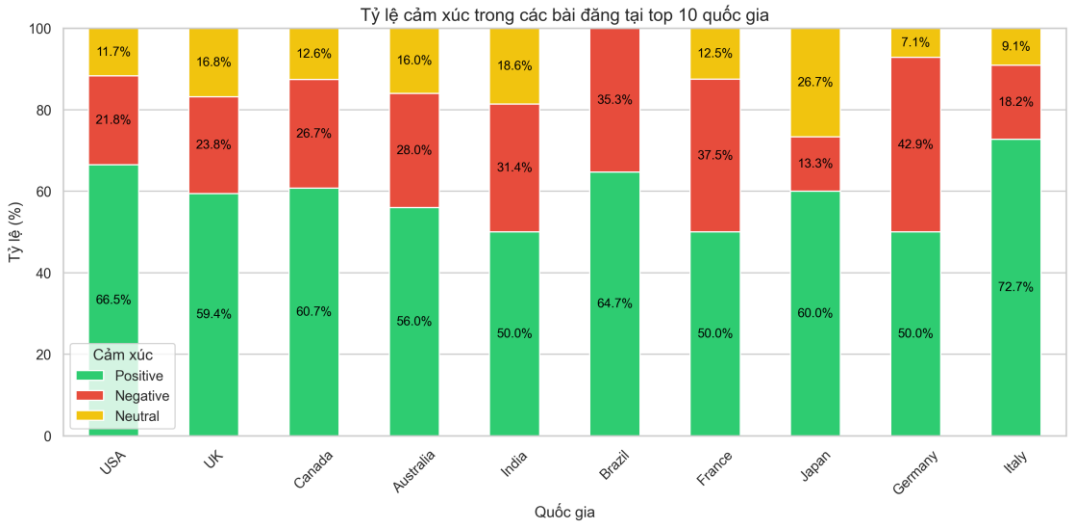
Kết hợp các biến như **Sentiment với Platform, Country, hoặc Timestamp, Likes, Retweets** để tìm hiểu cảm xúc theo từng nhóm nền tảng, khu vực, thời gian hoặc lượt tương tác.

3.1. 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 17: Biểu đồ top 10 quốc gia có nhiều bài đăng nhất



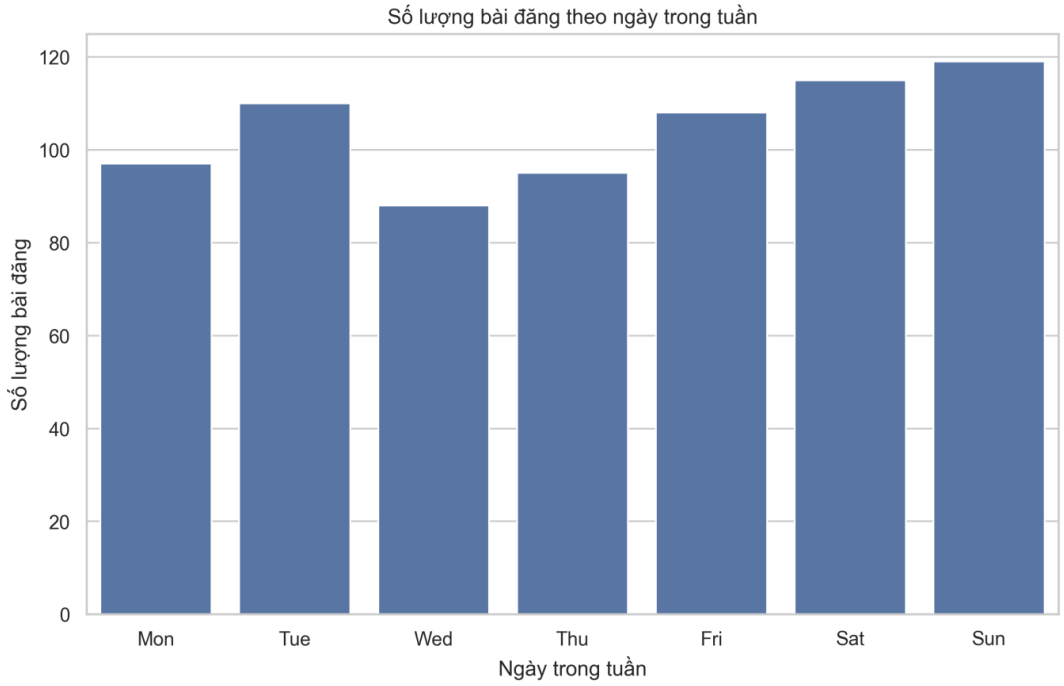
Hình 18: Biểu đồ tỷ lệ cảm xúc trong các bài đăng tại top 10 quốc gia

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**

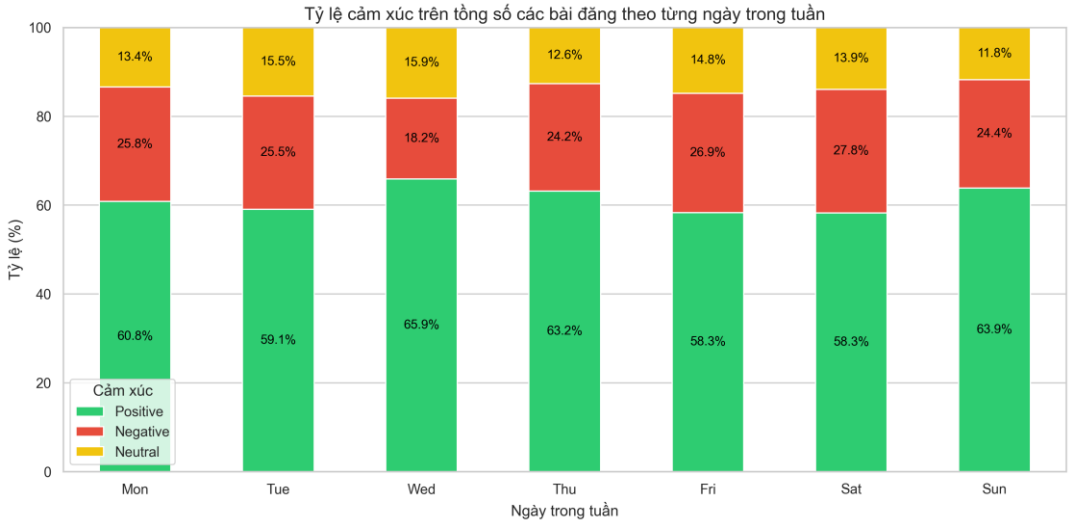
(dưới 25 bài mỗi nước). Tuy nhiên, về chất lượng cảm xúc, **Ý lại nổi bật với tỷ lệ nội dung tích cực cao nhất (72,7%)**, trong khi **Đức có tỷ lệ tiêu cực cao nhất (42,9%).** Đáng chú ý là không có mối liên hệ rõ ràng giữa số lượng và chất lượng bài đăng – tuy nhiên các nước có nhiều bài đăng thể hiện sự phân hóa dữ liệu rõ rệt hơn.

3.2. 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.

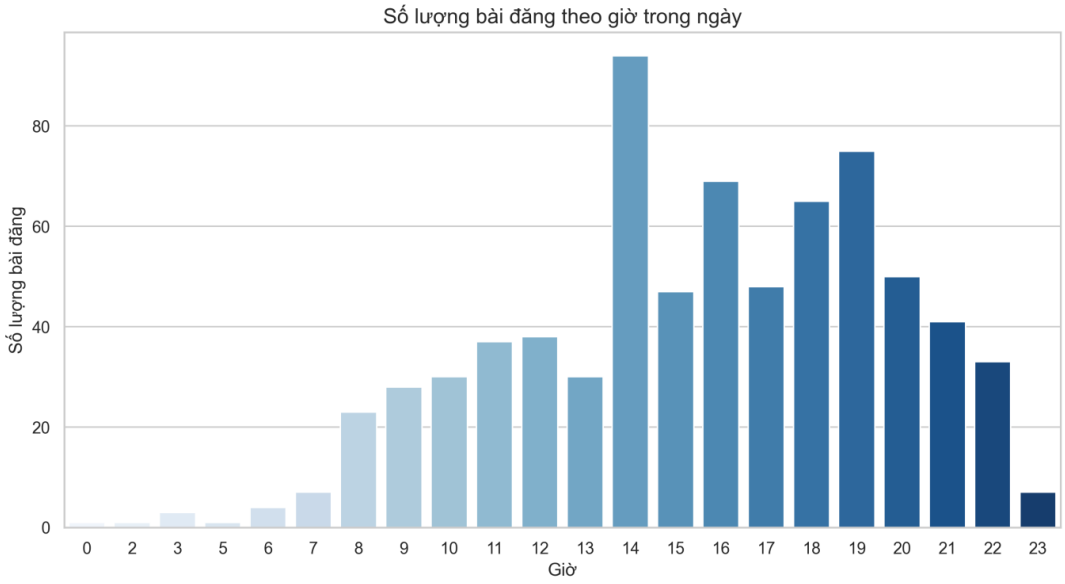


Hình 19: Biểu đồ số lượng bài đăng theo các ngày trong tuần

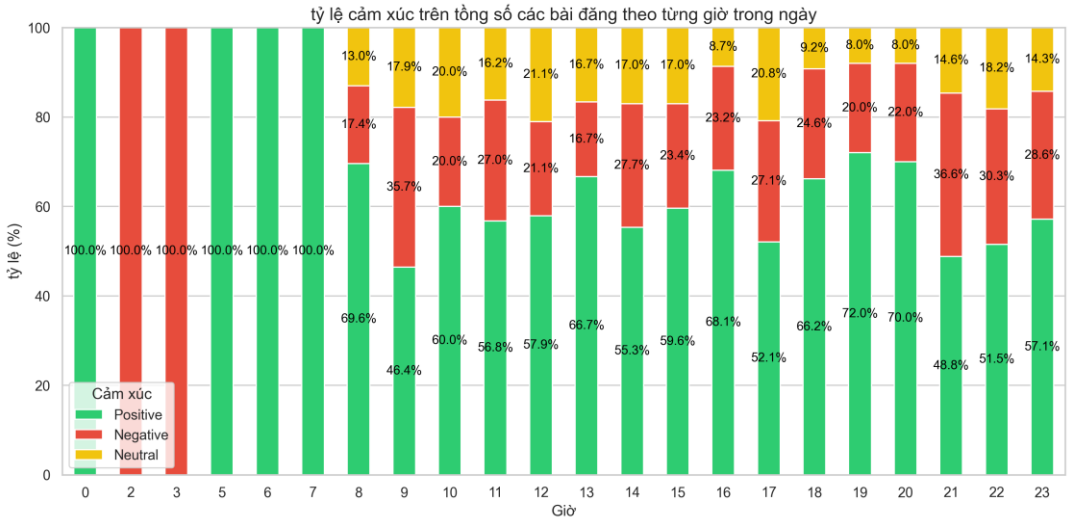


Hình 20: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo các ngày trong tuần

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 dùng 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 21: Biểu đồ số lượng bài đăng theo các khung giờ trong ngày

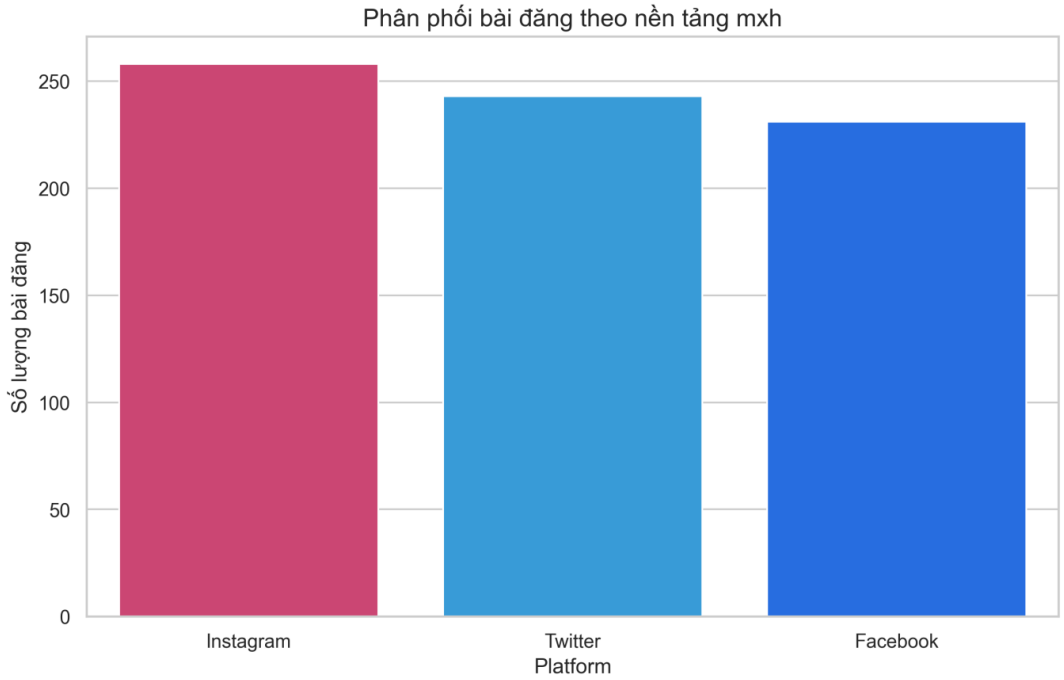


Hình 22: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo giờ trong ngày

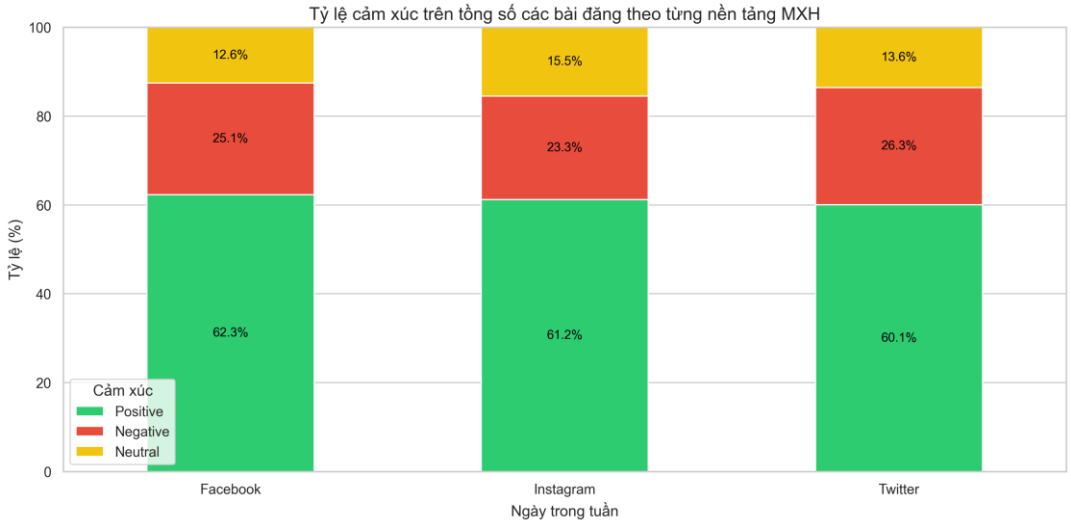
Nhận xét: Phân tích hai biểu đồ cho thấy lượng bài đăng **cao nhất** vào lúc **14h và 19h**, trong khi gần như **không có bài từ 0–5h sáng**. Về cảm xúc, các khung giờ **6h, 7h và 19h** có tỷ lệ nội dung **tích cực cao nhất (khoảng 72%)**, ngược lại **21h** ghi nhận tỷ lệ **tiêu cực cao nhất (36.6%)**. 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.

3.3. 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 của từng nền tảng MXH.



Hình 23: Biểu đồ phân phối bài đăng theo nền tảng MXH

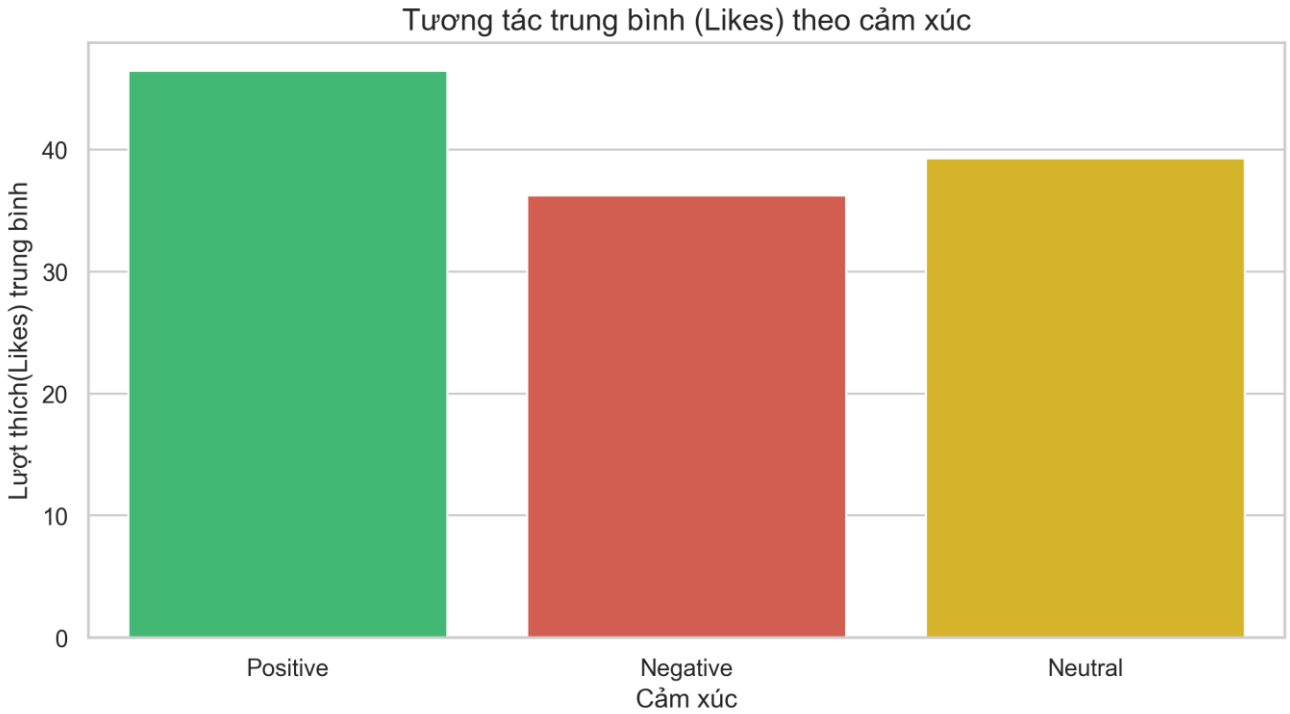


Hình 24: Biểu đồ tỷ lệ cảm xúc trên các bài đăng theo nền tảng MXH

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 nhưng không đều về mặt cảm xúc.

3.4. 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**.

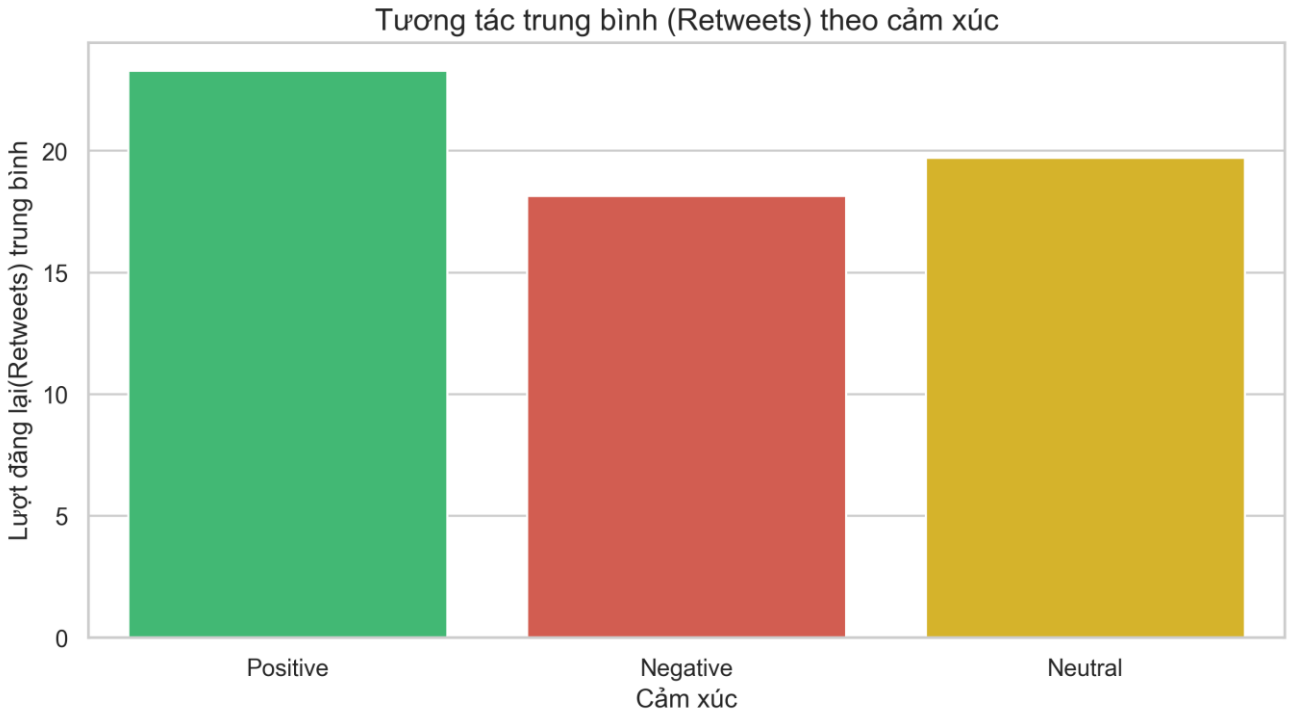


Hình 25: Biểu đồ lượt tương tác trung bình Likes theo cảm xúc

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.

3.5. 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**.

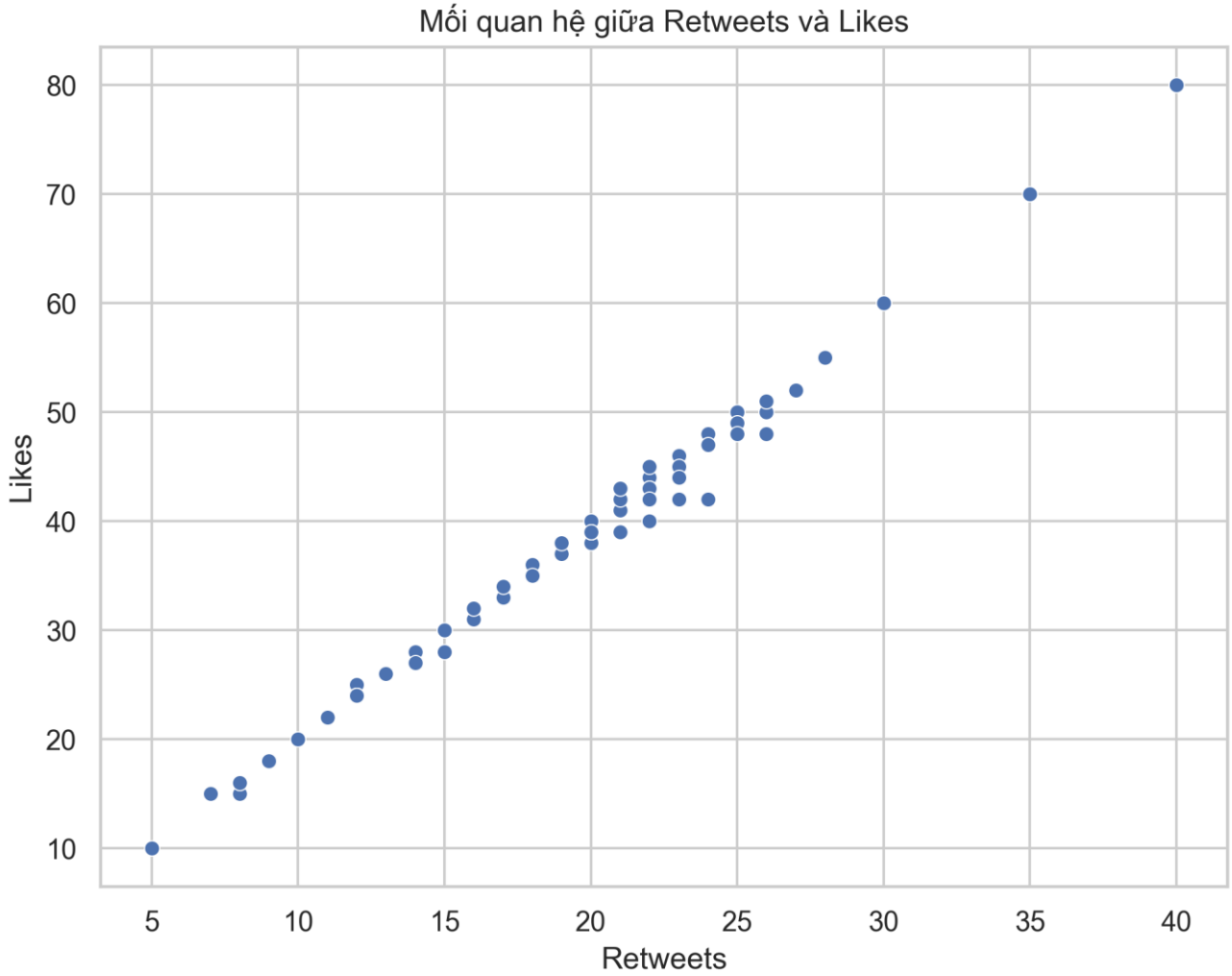


Hình 26: Biểu đồ lượt tương tác trung bình Retweets theo cảm xúc

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.

3.6. 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 27: 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

1. 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.

1.1. 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.

1.2. 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: ~61.2%
- 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 ngoại trừ các ký tự **icon**.

1.3. 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")
```

Top từ phổ biến trong bài viết cảm xúc Positive:

- 31, 'new': 'Mới',
- 22, 'laughter': 'Tiếng cười',
- 21, 'joy': 'Niềm vui',
- 19, 'beauty': 'Vẻ đẹp',
- 17, 'sky': 'Bầu trời',
- 16, 'friends': 'Những người bạn',
- 16, 'dreams': 'Những giấc mơ',
- 15, 'feeling': 'Cảm giác',
- 15, 'world': 'Thế giới',
- 15, 'concert': 'Buổi hòa nhạc',

Top từ phổ biến trong bài viết cảm xúc Negative:

- 15, 'like': 'Thích',
- 14, 'despair': 'Tuyệt vọng',
- 14, 'echoes': 'Tiếng vang',
- 13, 'shattered': 'Tan vỡ',
- 12, 'heart': 'Trái tim',
- 11, 'feeling': 'Cảm giác',
- 10, 'lost': 'Lạc lõng',
- 9, 'loneliness': 'Cô đơn',
- 8, 'grief': 'Đau buồn',
- 8, 'frustration': 'Bực bội',

Top từ phổ biến trong bài viết cảm xúc Neutral:

- 11, 'new': 'Mới',
- 10, 'curiosity': 'Sự tò mò',
- 8, 'old': 'Cũ',
- 8, 'nostalgia': 'Hoài niệm',
- 8, 'exploring': 'Khám phá',
- 7, 'day': 'Ngày',
- 7, 'confusion': 'Bối rối',
- 7, 'lifes': 'Cuộc sống',
- 7, 'emotions': 'Cảm xúc',
- 6, 'knowledge': 'Kiến thức',

Hình 28: 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.

1.4. 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.
- Áp dụng n-gram từ (1,1) đến (1,3)
- Giới hạn số đặc trưng: 1.000.

1.5. 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: 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.

1.6. Xử lý mất cân bằng dữ liệu:

Áp dụng phương pháp **Random Over-sampling**. Đây là một kỹ thuật xử lý mất cân bằng dữ liệu (**imbalanced data**) bằng cách nhân bản ngẫu nhiên các mẫu từ lớp thiểu số cho đến khi các lớp có số lượng gần bằng nhau.

```
# In ra kích thước dữ liệu trước khi cân bằng
print(f"\nKích thước dữ liệu trước khi cân bằng: {X.shape}")
# Áp dụng Random Over-sampling để cân bằng dữ liệu
ros = RandomOverSampler(random_state=42)
X, y = ros.fit_resample(X, y)
# In ra kích thước dữ liệu sau khi cân bằng
print(f"Kích thước dữ liệu sau khi over-sampling: {X.shape}")
```

```
Kích thước dữ liệu trước khi cân bằng: (732, 1000)
Kích thước dữ liệu sau khi over-sampling: (1344, 1000)
```

Hình 29: Kết quả xử lý mất cân bằng nhãn.

Nhận xét: Sau khi áp dụng **Random Over-Sampling**, tổng số mẫu huấn luyện đã **tăng từ 732 lên 1,344**, tương đương với mức **tăng khoảng 83.6%**. **Kỹ thuật này không làm thay đổi số lượng đặc trưng**, nhưng đã tăng cường số lượng mẫu thuộc các lớp thiểu số ("Negative" và "Neutral"), giúp đưa chúng về mức cân bằng với lớp đa số ("Positive").

Phần VI: Đánh giá và chọn thuật toán

1. 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**

1.1. Logistic Regression

Logistic Regression [2] 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

1.2. Naive Bayes

Naive Bayes [3] 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

1.3. Support Vector Machine (SVM)

SVM [4] 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)

1.4. XGBoost

XGBoost [5] **(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**.

2. 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.**

3. Kết quả so sánh các mô hình

3.1. Logistic Regression

Mô hình đạt hiệu suất **tốt hơn Naive Bayes** với **Accuracy 90.5%, Precision 91%, Recall 91% và F1-Score 91%**. Với đặc điểm đơn giản, dễ giải thích và hiệu suất ổn định, Logistic Regression là sự cân bằng giữa độ phức tạp và hiệu quả, phù hợp khi cần triển khai nhanh với tính minh bạch cao.

```
Kết quả trên tập Validation:

Validation Accuracy: 0.9054726368159204
              precision    recall  f1-score   support

    Negative       0.97      0.94      0.95        67
     Neutral       0.91      0.87      0.89        67
    Positive       0.85      0.91      0.88        67

    accuracy                           0.91       201
   macro avg       0.91      0.91      0.91       201
weighted avg       0.91      0.91      0.91       201
```

Hình 30: Kết quả đánh giá mô hình Logistic Regression trên tập Validation.

3.2. Naive Bayes

Mô hình có hiệu suất thấp nhất trong nhóm với **Accuracy 90.0%, Precision 90%, Recall 90% và F1-Score 90%**. Tuy nhiên, Naive Bayes có ưu điểm là huấn luyện nhanh, tiết kiệm tài nguyên và vẫn cho kết quả khá tốt làm baseline. Mô hình có thể bị ảnh hưởng bởi giả định độc lập giữa các đặc trưng trong dữ liệu văn bản.

```
Kết quả trên tập validation:

Validation Accuracy: 0.900497512437811
              precision    recall  f1-score   support

    Negative       0.94      0.99      0.96        67
     Neutral       0.82      0.90      0.86        67
    Positive       0.95      0.82      0.88        67

    accuracy                           0.90       201
   macro avg       0.90      0.90      0.90       201
weighted avg       0.90      0.90      0.90       201
```

Hình 31: Kết quả đánh giá mô hình Naive Bayes trên tập Validation.

3.3. Support Vector Machine (SVM)

Đây là mô hình có hiệu suất cao nhất trong các thuật toán được so sánh, với **Accuracy 92.5%, Precision 93%, Recall 93% và F1-Score 93%**. SVM hoạt động hiệu quả trên dữ liệu văn bản có nhiều chiều và thể hiện sự cân bằng tốt giữa các chỉ số. Mô hình này là lựa chọn phù hợp nhất để triển khai trong thực tế cho bài toán phân loại cảm xúc này.

```
Kết quả trên tập validation:

Validation Accuracy: 0.9253731343283582
              precision    recall  f1-score   support

    Negative       0.98      0.94      0.96        67
     Neutral       0.93      0.93      0.93        67
    Positive       0.87      0.91      0.89        67

    accuracy                           0.93       201
   macro avg       0.93      0.93      0.93       201
weighted avg       0.93      0.93      0.93       201
```

Hình 32: Kết quả đánh giá mô hình Support Vector Machine trên tập Validation.

3.4. XGBoost

Mô hình đạt hiệu suất tốt thứ hai với **Accuracy 91.5%, Precision 92%, Recall 92% và F1-Score 92%**. XGBoost là mô hình ensemble mạnh mẽ, có khả năng xử lý dữ liệu phức tạp và quan hệ phi tuyến tính. Mô hình này còn tiềm năng cải thiện nếu được điều chỉnh tham số kỹ lưỡng hơn.

```
Kết quả trên tập validation:

Validation Accuracy: 0.9154228855721394
              precision    recall  f1-score   support

    Negative       0.98      0.91      0.95        67
     Neutral       0.91      0.93      0.92        67
    Positive       0.86      0.91      0.88        67

    accuracy                           0.92       201
   macro avg       0.92      0.92      0.92       201
weighted avg       0.92      0.92      0.92       201
```

Hình 33: Kết quả đánh giá mô hình XGBoost trên tập Validation.

Thuật toán	Accuracy	Precision	Recall	F1-Score	Nhận xét sơ bộ
Naive Bayes	90.0%	90%	90%	90%	Mô hình đơn giản, huấn luyện nhanh với hiệu suất khá tốt, phù hợp làm baseline hoặc ứng dụng yêu cầu tốc độ cao, tuy nhiên có độ chính xác thấp nhất trong các mô hình so sánh.
Logistic Regression	90.5%	91%	91%	91%	Mô hình dễ giải thích với hiệu suất cao hơn Naive Bayes, cân bằng tốt giữa đơn giản và hiệu quả, phù hợp cho triển khai thực tế khi cần tính minh bạch.
XGBoost	91,5%	92%	92%	92%	Mô hình ensemble mạnh mẽ với hiệu suất cao, có tiềm năng cải thiện thêm nếu được điều chỉnh tham số kỹ lưỡng, phù hợp cho các ứng dụng yêu cầu độ chính xác cao.
SVM	92.5%	93%	93%	93%	Mô hình hiệu suất cao nhất với precision, recall và F1-score đều đạt 93%, cân bằng tốt giữa các lớp, phù hợp làm mô hình chính cho hệ thống phân loại cảm xúc.

Bảng 2: Kết quả so sánh các thuật toán

Phần VII: Kết quả và thảo luận

1. Kết quả trên tập kiểm tra

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.

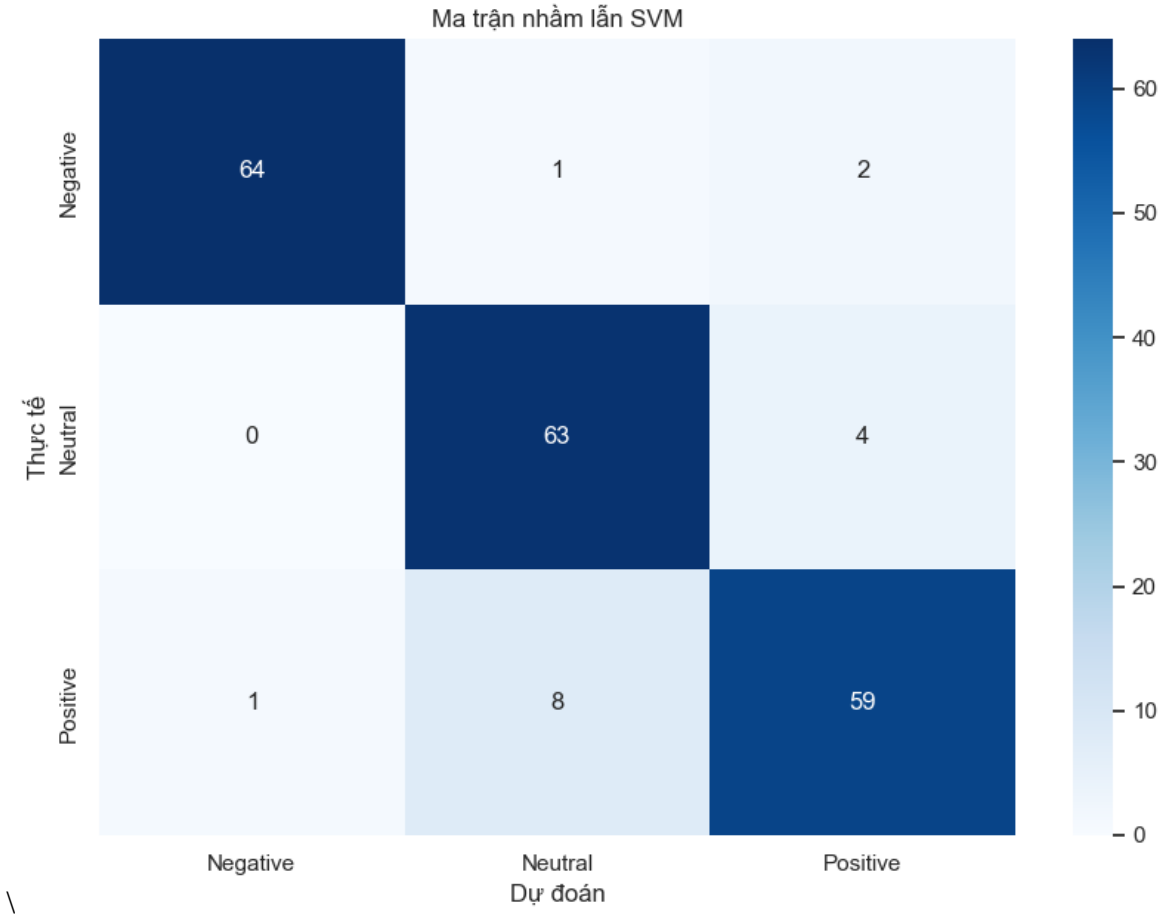
```
Kết quả trên tập test:

Test Accuracy: 0.9207920792079208
              precision    recall  f1-score   support

    Negative       0.98      0.96      0.97        67
     Neutral       0.88      0.94      0.91        67
    Positive       0.91      0.87      0.89        68

    accuracy                           0.92       202
   macro avg       0.92      0.92      0.92       202
weighted avg       0.92      0.92      0.92       202
```

Hình 34: Kết quả đánh giá mô hình Support Vector Machine trên tập Test

- **Accuracy:** 92.07%
- **Weighted Precision:** 82%
- **Weighted Recall:** 82%
- **Weighted F1-Score:** 82%

Nhãn	Precision	Recall	F1-score	Số mẫu
Negative	98%	96%	97%	67
Neutral	88%	94%	91%	67
Positive	91%	87%	89%	68

Bảng 3: Kết quả trên tập kiểm tra.



\

Bảng 4: Ma trận nhầm lẫn Support Vector Machine.

2. 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:

- Có 59/68 mẫu được phân loại đúng là Positive
- 1 mẫu Positive bị nhầm thành Negative
- 8 mẫu Positive bị nhầm thành Neutral
- **Độ nhớ (Recall):** 59/(1+8+59) = 59/68 ≈ **86,8%**

Lớp Negative:

- Có 64/67 mẫu được phân loại đúng là Negative
- 1 mẫu Negative bị nhầm thành Neutral
- 2 mẫu Negative bị nhầm thành Positive
- **Độ nhớ (Recall):** 64/(64+1+2) = 64/67 ≈ **95,5%**

Lớp Neutral:

- Có 63/67 mẫu được phân loại đúng là Neutral
- 0 mẫu Neutral bị nhầm thành Negative
- 4 mẫu Neutral bị nhầm thành Positive
- **Độ nhớ (Recall):** 63/(0+63+4) = 63/67 ≈ **94%**

Độ chính xác (Precision) cho từng lớp:

- **Negative:** 64/(64+0+1) = 64/65 ≈ **98,5%**
- **Neutral:** 63/(1+63+8) = 63/72 ≈ **87,5%**
- **Positive:** 59/(2+4+59) = 59/65 ≈ **90,8%**

3. **Nhận xét tổng quan**
   1. **Lớp Negative** có hiệu suất tốt nhất với độ nhớ và độ chính xác đều cao (**95,5% và 98,5%**)
   2. **Lớp Neutral** có độ nhớ khá tốt (**94%**) nhưng độ chính xác thấp hơn do nhiều mẫu Positive bị phân loại nhầm thành Neutral
   3. **Lớp Positive** có độ nhớ thấp nhất (**86,8%**), chủ yếu do nhiều mẫu bị phân loại nhầm thành Neutral
4. **Nhận xét thêm**
   - **Mô hình SVM hoạt động khá tốt với tổng thể độ chính xác:** (64+63+59)/(67+67+68) = 186/202 ≈ **92,1%**
   - **Điểm yếu của mô hình** là phân biệt giữa lớp **Positive** và **Neutral**, với 8 mẫu Positive bị phân loại nhầm thành Neutral
   - Mô hình có xu hướng phân loại chính xác nhất các mẫu **Negative**

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 **(Neutral)** và tiêu cực **(Negative)** một cách chính xác hơn.

Phần VIII: Kiểm thử mô hình

```
sample_texts = [
    # Positive
    "The customer service here is truly excellent, the staff are very helpful and friendly.",
    "I've been using this product for 6 months and I'm completely satisfied. Totally worth the money.",
    "The app runs smoothly, has a beautiful interface, and is very easy to use. 10 out of 10!"    ,

    # Negative
    "I can't believe how bad this product is, I'm extremely disappointed.",
    "The app keeps crashing and it's basically unusable. # Negative",
    "Delivery was almost a week late and nobody answered the customer service hotline.",

    # Neutral
    "The product matches the description, packaging was okay, nothing special.",
    "Attending a virtual conference on AI.",
    "Confusion surrounds me as I navigate through life's choices.",
]
```

1. Đánh giá tổng thể

```
==== Kết quả với mô hình: Support_Vector_Machine ====

"The customer service here is truly excellent, the staff are very helpful and friendly." => Positive: 75.66%
"I've been using this product for 6 months and I'm completely satisfied. Totally worth the money." => Positive: 75.66%
"The app runs smoothly, has a beautiful interface, and is very easy to use. 10 out of 10!" => Positive: 95.02%

"I can't believe how bad this product is, I'm extremely disappointed." => Negative: 99.62%
"The app keeps crashing and it's basically unusable." => Positive: 75.66%
"Delivery was almost a week late and nobody answered the customer service hotline." => Positive: 75.66%

"The product matches the description, packaging was okay, nothing special." => Positive: 75.66%
"Attending a virtual conference on AI." => Neutral: 96.73%
"Confusion surrounds me as I navigate through life's choices." => Neutral: 97.10%
```

Hình 35: Kết quả kiểm thử mô hình SVM

3 câu đầu (Positive): Hệ thống dự đoán chính xác cả 3 câu, với xác suất khá cao (76–95%).

3 câu giữa (Negative):

- **Câu đầu tiên** mô hình dự đoán đúng là **Negative**, với **99.62%**, rất tốt.
- Tuy nhiên **2 câu còn lại** bị **dự đoán sai** là **Positive**, mặc dù **nội dung rất tiêu cực**.

3 câu cuối (Neutral):

- **2/3 câu đầu** mô hình dự đoán đúng là **Neultral** với xác suất rất cao (96,73% và 97,10%)
- Chỉ **câu đầu tiên** dự đoán nhầm là **Positive**.

2. Những vấn đề nổi bật:

Sai lệch với câu Negative và Posetive:

Những câu như:

- "the app keeps crashing and its basically unusable negative"
- "delivery was almost a week late and nobody answered the customer service hotline"

Nội dung rõ ràng **phản ánh sự bất mãn**, nhưng lại bị dự đoán là **Positive**.

Điều này cho thấy mô hình đang: **Bị lệch nhãn (label bias)**, có thể thiên về **"Positive"** do phân bố dữ liệu huấn luyện không cân bằng. **Chưa hiểu tốt ngữ cảnh tiêu cực nếu không có từ khóa tiêu cực rõ ràng** (như **"bad"**, **"terrible"**, **"worst"**,...).

Neutral dễ bị nhầm là Positive:

Câu như **"the product matches the description packaging was okay nothing special"** là rất trung tính, nhưng bị gán là Positive.

Mô hình có vẻ đánh giá trung lập + có chút tích cực là Positive.

Phần IX: kết luận

1. 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**.

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 92.5%** và **F1-score đạt 93%** trên tập **validation**, **92% 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) và** tiêu cực **(Negative).** 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.

2. 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 lẫn 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).

3. Tài liệu tham khảo

[1] K. Parmar, "Social Media Sentiments Analysis Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset.

[2] D. W. Hosmer, S. Lemeshow and R. X. Sturdivant, Applied logistic regression (3rd ed.), Wiley, 2013.

[3] 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.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, p. 273–297, 1995.

[5] 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.