**TRƯỜNG ĐẠI HỌC PHENIKAA**

**KHOA CÔNG NGHỆ THÔNG TIN**

⸎⸎⸎⸎⸎

Ảnh có chứa biểu tượng

Mô tả được tạo tự động

**BÀI TẬP LỚN**

**PHÂN TÍCH TÌNH CẢM**

**TRUYỀN THÔNG XÃ HỘI**

(Social media sentiment analysis)

Lớp Lập trình phân tích dữ liệu với Python-1-2-22(N01)

Giảng viên hướng dẫn: ThS. Nguyễn Văn Thiệu

Sinh viên thực hiện: Nguyễn Nam Hải - 21012056

Vi Đăng Quang - 21010583

HÀ NỘI, THÁNG 5 NĂM 2023

**TRƯỜNG ĐẠI HỌC PHENIKAA**

**KHOA CÔNG NGHỆ THÔNG TIN**

⸎⸎⸎⸎⸎

Ảnh có chứa biểu tượng

Mô tả được tạo tự động

**BÀI TẬP LỚN**

**PHÂN TÍCH TÌNH CẢM**

**TRUYỀN THÔNG XÃ HỘI**

(Social media sentiment analysis)

Lớp Lập trình phân tích dữ liệu với Python-1-2-22(N01)

Giảng viên hướng dẫn: ThS. Nguyễn Văn Thiệu

Sinh viên thực hiện: Nguyễn Nam Hải 21012056

Vi Đăng Quang 21010583

HÀ NỘI, THÁNG 5 NĂM 2023

# MỤC LỤC

[MỞ ĐẦU 5](#_Toc135567014)

[NỘI DUNG 6](#_Toc135567015)

[I. Mục đích nghiên cứu 6](#_Toc135567016)

[II. Phương pháp nghiên cứu 7](#_Toc135567017)

[1. Cơ sở lý thuyết 7](#_Toc135567018)

[2. Mô tả dữ liệu 8](#_Toc135567019)

[3. Tiền xử lý dữ liệu 9](#_Toc135567020)

[3.1 Chuyển định dạng từ chữ hoa về chữ thường 10](#_Toc135567021)

[3.2 Loại bỏ các đường liên kết URL và tag @ 11](#_Toc135567022)

[3.3 Loại bỏ các kí tự đặc biệt 12](#_Toc135567023)

[3.4 Loại bỏ từ dừng (Stop words) 13](#_Toc135567024)

[3.5 Đưa các từ ngữ về dạng gốc Stemming & Lemmatizer 14](#_Toc135567025)

[4. Trực quan hóa dữ liệu 15](#_Toc135567026)

[4.1 Sử dụng Pie & Bar chart trực quan phân bổ dữ liệu 16](#_Toc135567027)

[4.2 Sử dụng biểu đồ đám mây (WordCloud) 18](#_Toc135567028)

[5. Sử dụng các mô hình học máy 20](#_Toc135567029)

[5.1 Giới thiệu kỹ thuật TF-IDF 21](#_Toc135567030)

[5.2 Áp dụng các mô hình học máy 22](#_Toc135567031)

[5.3 Đánh giá kết quả sau khi sử dụng mô hình học máy 29](#_Toc135567032)

[KẾT LUẬN 30](#_Toc135567033)

[TÀI LIỆU THAM KHẢO 31](#_Toc135567034)

## DANH MỤC HÌNH ẢNH

[Hình 1. Biểu đồ cột thể hiện số lượng dữ liệu theo 3 loại cảm xúc 16](file:///C:\Users\vidan\Documents\GitHub\Twitter_Sentiment_Analysis\Baocao_Nhom05_DataAnalysisWPython.docx#_Toc135567677)

[Hình 2. Biểu đồ tròn thể hiện tỷ lệ dữ liệu theo 3 loại cảm xúc 17](#_Toc135567678)

[Hình 3. Biểu đồ đám mây thể hiện các bình luận được đánh giá Positive 19](#_Toc135567679)

[Hình 4. Biểu đồ đám mây thể hiện các bình luận được đánh giá Neutral 19](#_Toc135567680)

[Hình 5. Bảng đánh giá độ chính xác của mô hình với 23](#_Toc135567681)

[Hình 6. Biểu đồ thể hiện hiệu suất của thuật toán Bernoulli Naive Bayes lần 1 24](#_Toc135567682)

[Hình 7. Bảng đánh giá độ chính xác của mô hình với 25](#_Toc135567683)

[Hình 8. Biểu đồ thể hiện hiệu suất của thuật toán Bernoulli Naive Bayes 25](#_Toc135567684)

[Hình 9. Bảng đánh giá độ chính xác của mô hình với 26](#_Toc135567685)

[Hình 10. Biểu đồ thể hiện hiệu xuất của thuật toán Linear Support Vector Classifier 27](#_Toc135567686)

[Hình 11. Bảng đánh giá độ chính xác của mô hình với thuật toán Random Forest 28](#_Toc135567687)

[Hình 12. Biểu đồ thể hiện hiệu suất của thuật toán Random Forest 28](#_Toc135567688)

## BẢNG PHÂN CÔNG CÔNG VIỆC

Giai đoạn 1:

Nhóm cùng nhau đưa ra các ý tưởng: tìm kiếm tập dữ liệu, phương pháp xử lý dữ liệu, cách thức ứng dụng …

Cuối cùng thống nhất đề tài với tập dữ liệu của twitter và sử dụng các mô hình học máy vào để đánh giá độ chính xác của dữ liệu sau khi được xử lý.

Giai đoạn 2: Triển khai

|  |  |
| --- | --- |
| Nguyễn Nam Hải | Code: Tiền xử lý dữ liệu, Trực quan hóa dữ liệu.  Viết báo cáo. Thuyết trình. |
| Vi Đăng Quang | Code: Áp dụng kỹ thuật TF – IDF, Sử dụng mô hình học máy để đánh giá.  Tổng hợp, chỉnh sửa, hoàn thiện báo cáo. Làm slide. |

# MỞ ĐẦU

Từ những năm đầu của thế kỉ XXI, dưới sự phát triển bùng nổ của Internet và cuộc cách mạng công nghiệp 4.0 đã thay đổi hoàn toàn cách các doanh nghiệp vận hành thông qua các công nghệ. Phân tích dữ liệu đã trở thành một khía cạnh quan trọng của các doanh nghiệp và ngành công nghiệp hiện đại, cho phép các tổ chức đưa ra quyết định sáng suốt dựa trên những hiểu biết dựa trên dữ liệu. Một lĩnh vực mà phân tích dữ liệu đã đạt được sức hút đáng kể là phương tiện truyền thông xã hội, đặc biệt là Twitter. Với hàng triệu người dùng trên toàn thế giới, Twitter cung cấp một lượng lớn dữ liệu có thể được phân tích để thu được thông tin chi tiết có giá trị về hành vi của khách hàng, danh tiếng thương hiệu và phản ứng cảm xúc. Trong dự án này, chúng ta sẽ khám phá sử dụng kỹ thuật phân tích ngôn ngữ tự nhiên để phân tích dữ liệu và áp dụng xây dựng những mô hình học máy để đánh giá, dự đoán cảm xúc trong các bình luận mà chúng tôi thu thập được.

# NỘI DUNG

## I. Mục đích nghiên cứu

Dự án này giải quyết vấn đề phân tích tình cảm trên Twitter; đó là phân loại các Tweets theo tình cảm thể hiện trong đó: tích cực, tiêu cực hoặc trung lập.

Twitter là một nền tảng mạng xã hội và tiểu blog trực tuyến cho phép người dùng viết các cập nhật trạng thái ngắn có độ dài tối đa 140 ký tự. Đây là một dịch vụ đang mở rộng nhanh chóng với hơn 200 triệu người dùng đã đăng ký - trong số đó có 100 triệu người dùng đang hoạt động và một nửa trong số họ đăng nhập twitter hàng ngày - tạo ra gần 250 triệu tweet mỗi ngày. Do lượng sử dụng lớn này, chúng tôi hy vọng sẽ phản ánh được tình cảm của công chúng bằng cách phân tích tình cảm được thể hiện trong các tweet. Phân tích tình cảm của công chúng rất quan trọng đối với nhiều ứng dụng chẳng hạn như các công ty đang cố gắng tìm hiểu phản ứng của sản phẩm của họ trên thị trường, dự đoán các cuộc bầu cử chính trị và dự đoán các hiện tượng kinh tế xã hội như thị trường chứng khoán… Mục đích của dự án này là phát triển một bộ phân loại chức năng để phân loại tình cảm chính xác và tự động của một luồng tweet không xác định.

Chúng tôi đã chọn làm việc với Twitter vì chúng tôi cảm thấy đó là một phương tiện gần đúng hơn để thể hiện tình cảm của công chúng thay vì các bài báo trên internet và blog web thông thường. Lý do là lượng dữ liệu liên quan đối với twitter lớn hơn nhiều so với các trang viết blog truyền thống. Ngoài ra, phản hồi trên twitter nhanh chóng hơn và cũng đa dạng hơn (vì số lượng người dùng tweet nhiều hơn đáng kể so với những người viết blog hàng ngày).

Việc sử dụng các kỹ thuật phân tích ngôn ngữ tự nhiên và mô hình học máy sẽ giúp cải thiện độ chính xác và hiệu quả của việc phân tích tình cảm trên Twitter trong tương lai. Tuy nhiên, việc sử dụng các dữ liệu trên Twitter cũng cần được thực hiện với trách nhiệm và tôn trọng quyền riêng tư và bảo mật thông tin của người dùng.

## II. Phương pháp nghiên cứu

### 1. Cơ sở lý thuyết

1. Matplotlib: là một thư viện vẽ đồ thị trong Python. Nó cung cấp các công cụ cho việc tạo ra các biểu đồ, đồ thị, histogram và các loại biểu đồ khác để trình bày dữ liệu một cách trực quan và dễ hiểu.

2. Word Cloud: là một thư viện trong Python để thể hiện các từ xuất hiện thường xuyên trong văn bản dựa trên tần suất xuất hiện. Nó cho phép tạo ra các đám mây từ, với kích thước và màu sắc của từ được chỉ định dựa trên tần suất xuất hiện của chúng.

3. NLTK (Natural Language Toolkit): là một thư viện trong Python giúp xử lý ngôn ngữ tự nhiên (NLP). Nó cung cấp các công cụ để xử lý, phân tích và trích xuất thông tin từ văn bản tự nhiên, bao gồm các công cụ xử lý ngôn ngữ tự nhiên như tokenization, stemming, lemmatization, phân tích cú pháp (syntax analysis), phân tích ngữ nghĩa (semantic analysis), và phân loại văn bản (text classification).

4. Sklearn (Scikit-learn): là một thư viện trong Python cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling. Nó cung cấp các thuật toán phổ biến như hồi quy tuyến tính (linear regression), phân loại (classification), phân cụm (clustering), và rừng ngẫu nhiên (random forest). Sklearn cũng cung cấp các công cụ để tiền xử lý dữ liệu, chọn đặc trưng (feature selection), và đánh giá mô hình.

5. Time: là một thư viện trong Python dùng để làm việc với thời gian. Nó cung cấp các công cụ để đọc và xử lý thời gian và ngày tháng, bao gồm các lớp và hàm để tạo ra các đối tượng thời gian, tính toán khoảng cách giữa các thời điểm, chuyển đổi định dạng thời gian, và tính toán thời gian chênh lệch giữa các múi giờ. Thư viện Time cũng cung cấp các công cụ để định dạng và hiển thị thời gian và ngày tháng theo định dạng mong muốn.

### 2. Mô tả dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Số sample | Số feature | Số class |
| [twitter.csv](https://drive.google.com/file/d/1mIG_1HzLhCaoSuXta5mvrSBPo0T718gm/view?usp=sharing) | 20000 | 5 | 3 |

Các đặc tính liên quan đến tập dữ liệu:

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Insight | Non-Null Count | Dtype |
| ids | ID Tweets | 20000 | int64 |
| date | Thời gian đăng tweets | 20000 | object |
| flag | Truy vấn | 20000 | object |
| user | Tên người đăng | 20000 | object |
| text | Nội dung bình luận | 20000 | object |
| Sentiment | Phân loại cảm xúc | 20000 | object |

### 3. Tiền xử lý dữ liệu

Trong nhiệm vụ lần này, text là một feature phổ biến nhất được sử dụng trong phân tích tình cảm truyền thông, vì nó cho phép phân tích ngôn ngữ tự nhiên và đưa ra đánh giá về tình cảm của các bình luận, đánh giá hoặc tin nhắn trên các nền tảng truyền thông xã hội.

Tuy nhiên, một thách thức lớn trong việc sử dụng text làm feature là việc thu thập và làm sạch dữ liệu. Khi thu thập dữ liệu, text thường chứa nhiều dấu câu, từ viết tắt, từ ngữ không chính thống và nhiều lỗi chính tả, ngữ pháp. Các vấn đề này gây ra khó khăn trong việc phân tích và dự đoán tình cảm của người dùng.

Do đó, việc làm sạch dữ liệu là một bước cần thiết và quan trọng trong quá trình phân tích tình cảm truyền thông. Quá trình này bao gồm việc xử lý dữ liệu thô bằng các công cụ như loại bỏ các ký tự đặc biệt, chuyển đổi chữ hoa thành chữ thường, loại bỏ các từ viết tắt và từ ngữ không chính thống. Sau đó, các từ trong text được phân tách và đưa vào một danh sách các từ để sử dụng làm feature cho các mô hình phân tích tình cảm.

Việc làm sạch dữ liệu là một quá trình phức tạp và tốn nhiều thời gian, nhưng nó là bước quan trọng để có thể thu được dữ liệu chính xác và hiệu quả.

Dưới đây là tập data ngẫu nhiên trong dataset, chúng ta sẽ tiến hành phân tích để chúng ta thấy rõ hơn trong việc xử lý dữ liệu dạng text.

|  |  |
| --- | --- |
| Text | Sentiment |
| @switchfoot I've been waiting for you for 6 hours. Why won't you show my location?! | Neutral |
| Where did u move to? @babiboo I thought u were already in sd. ?? Hmmm. Random u found me. Glad to hear you're doing well. | Positive |
| He's the reason for the teardrops on my guitar the only one who has enough of me to break my heart :<< | Negative |
| Sad, sad, sad. I don't know why but I hate this feeling. I wanna SLEEP and I still can't! | Negative |

#### 3.1 Chuyển định dạng từ chữ hoa về chữ thường

|  |
| --- |
| Text |
| @switchfoot i've been waiting for you for 6 hours. why won't you show my location?! http://twitpic.com/2y2es |
| where did u move to? @babiboo i thought u were already in sd. ?? hmmm. random u found me. glad to hear yer doing well. |
| he's the reason for the teardrops on my guitar the only one who has enough of me to break my heart :<< |
| "sad, sad, sad. i don't know why but i hate this feeling. i wanna sleep and i still can't! " |

Việc đồng bộ hóa về chữ thường giúp đồng nhất dữ liệu, giảm thiểu sự khác biệt và tạo ra một bộ dữ liệu thống nhất và dễ xử lý hơn.

Khi tìm kiếm hoặc so sánh văn bản, việc chuyển đổi tất cả các ký tự về cùng một dạng chữ thường giúp đảm bảo rằng các từ khóa và cụm từ được so sánh chính xác, không phụ thuộc vào việc ký tự được viết hoa hay viết thường.

#### 3.2 Loại bỏ các đường liên kết URL và tag @

Việc loại bỏ các đường liên kết URL và tag "@" trong phân tích tình cảm truyền thông là cần thiết vì những thông tin này không mang lại nhiều giá trị cho việc phân tích cảm xúc.

Đầu tiên, các đường liên kết URL có thể chứa nhiều thông tin không liên quan đến nội dung chính của văn bản, chẳng hạn như quảng cáo, thông tin về tác giả, hoặc những trang web không liên quan. Những thông tin này có thể làm nhiễu và làm giảm độ chính xác của phân tích tình cảm.

Thứ hai, tag "@" thường được sử dụng để đề cập đến một người dùng cụ thể trên mạng xã hội. Tuy nhiên, thông tin này thường không liên quan đến nội dung chính của văn bản và có thể không mang lại nhiều giá trị cho việc phân tích tình cảm.

|  |
| --- |
| Text |
| i've been waiting for you for 6 hours. why won't you show my location?! |
| where did u move to? i thought u were already in sd. ?? hmmm. random u found me. glad to hear yer doing well. |
| he's the reason for the teardrops on my guitar the only one who has enough of me to break my heart :<< |

#### 3.3 Loại bỏ các kí tự đặc biệt

Các kí tự đặc biệt hoặc emoji thường được sử dụng để biểu diễn tâm trạng, cảm xúc của con người trong giao tiếp trên mạng xã hội.

Ví dụ: 😀 biểu thị cho mặt cười, :-O là sự ngạc nhiên há hốc mồm và <3 là hình trái tim.

Tuy nhiên, đối với nghiên cứu phân tích cảm xúc trên Twitter, chúng tôi sẽ loại bỏ những ký tự này vì chúng khó để phân tích và có thể gây nhiễu cho kết quả. Thay vào đó, chúng tôi sẽ chỉ dựa vào nội dung văn bản của các bài đăng để xác định cảm xúc của người dùng.

|  |
| --- |
| Text |
| ive been waiting for you for 6 hours why wont you show my location |
| where did u move to i thought u were already in sd hmmm random u found me glad to hear yer doing well |
| hes the reason for the teardrops on my guitar the only one who has enough of me to break my heart |
| sad sad sad i dont know why but i hate this feeling i wanna sleep and i still cant |

#### 3.4 Loại bỏ từ dừng (Stop words)

|  |
| --- |
| Text |
| ive waiting 6 hours wont show location |
| u move thought u already sd hmmm random u found glad hear yer well |
| hes reason teardrops guitar one enough break heart |
| sad sad sad dont know hate feeling wanna sleep still cant |

Khi phân tích văn bản, stop words là các từ phổ biến trong tiếng anh như 'who', 'about', "hadn't", 'this', 'themselves', 'a', 'doesn',,... những từ này không mang lại nhiều ý nghĩa trong câu và thường xuất hiện rất nhiều trong các văn bản.

Vì vậy, trong quá trình xử lý dữ liệu, loại bỏ các stop words là một phương pháp thường được sử dụng để giảm kích thước của dữ liệu và tăng tốc độ xử lý, giảm độ nhiễu trong quá trình phân tích tình cảm.

Tuy nhiên, việc loại bỏ stopwords cũng có thể làm mất đi một số thông tin quan trọng trong văn bản. Chẳng hạn, trong một số trường hợp, các stopwords có thể mang lại ý nghĩa quan trọng trong việc phân tích cảm xúc, ví dụ như trong trường hợp đối tượng phân tích nói về sự so sánh giữa hai hoặc nhiều đối tượng.

Vì vậy, việc loại bỏ stopwords hay không phụ thuộc vào mục đích phân tích và dữ liệu được sử dụng. Nếu mục đích của phân tích là tìm hiểu ý nghĩa và nội dung cụ thể của một văn bản, thì loại bỏ stop word có thể không phù hợp.

Tuy nhiên, nếu mục đích là phân tích cảm xúc trên quy mô lớn và dữ liệu rất nhiều, việc loại bỏ stop word có thể giúp tăng độ chính xác và hiệu quả của phân tích.

#### 3.5 Đưa các từ ngữ về dạng gốc Stemming & Lemmatizer

Trong phân tích văn bản, Stemming và Lemmatization là hai phương pháp để đưa các từ trong văn bản về dạng gốc, để thuận tiện cho việc xử lý và phân tích dữ liệu.

Stemming là quá trình loại bỏ các hậu tố (suffix) của một từ để đưa nó về dạng gốc (root form) của từ đó. Ví dụ: từ "running" có thể được đưa về dạng gốc là "run".

Lemmatization là quá trình đưa các từ về dạng gốc (lemma) của nó bằng cách sử dụng từ điển và quy tắc ngữ pháp để xác định các dạng khác nhau của từ đó. Ví dụ: từ "ran" và "running" đều có thể được đưa về dạng gốc là "run".

Việc đưa các từ về dạng gốc bằng Stemming và Lemmatization là quan trọng trong phân tích văn bản vì nó giúp giảm độ phức tạp của dữ liệu và làm cho việc phân tích dữ liệu dễ dàng hơn. Nó cũng giúp tăng độ chính xác và độ tin cậy của các kết quả phân tích.

Việc đưa các từ về dạng gốc cũng giúp loại bỏ các từ trùng lặp và tăng khả năng phát hiện các từ khóa (keywords) quan trọng trong văn bản. Nó cũng giúp đồng bộ hóa các từ khóa và giúp các công cụ tìm kiếm hoạt động hiệu quả hơn.

Tuy nhiên, cũng cần lưu ý rằng quá trình Stemming và Lemmatization có thể gây mất mát thông tin trong quá trình đưa các từ về dạng gốc. Do đó, cần phải lựa chọn phương pháp phù hợp và cân nhắc trước khi sử dụng chúng để đảm bảo tối ưu hoá kết quả phân tích dữ liệu.

|  |
| --- |
| Text |
| ive wait 6 hour wont show locat |
| u move thought u alreadi sd hmmm random u found glad hear yer well |
| he reason teardrop guitar one enough break heart |
| sad sad sad dont know hate feel wanna sleep still cant |

Và đó bước cuối cùng sau khi thực hiện bước tiền xử lý dữ liệu để đưa về định dạng phù hợp với các mô hình học máy.

### 4. Trực quan hóa dữ liệu

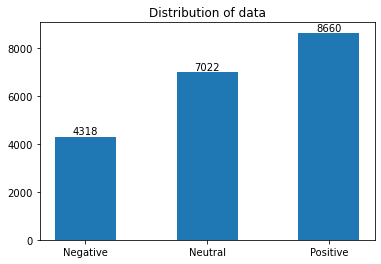
Trực quan hóa là một công cụ quan trọng trong phân tích dữ liệu tình cảm truyền thông, đặc biệt là trong việc phân loại các bình luận, đánh giá hoặc tin nhắn thành các loại cảm xúc tích cực, tiêu cực hoặc trung tính. Trực quan hóa giúp chúng ta hiểu rõ hơn về phân bố của các loại cảm xúc trong dữ liệu, từ đó đưa ra những quyết định và hành động phù hợp.

Việc sử dụng các công cụ trực quan hóa giúp người dùng dễ dàng theo dõi và phân tích sự phân bố của các loại cảm xúc trong dữ liệu. Ví dụ, biểu đồ tròn và biểu đồ cột có thể được sử dụng để hiển thị tỷ lệ phần trăm hoặc số lượng các loại cảm xúc trong các bình luận, đánh giá hoặc tin nhắn. Các biểu đồ đường có thể được sử dụng để theo dõi sự thay đổi của các loại cảm xúc qua thời gian.

#### 4.1 Sử dụng Pie & Bar chart trực quan phân bổ dữ liệu

Sau bước tiền xử lý dữ liệu, chúng ta có bảng phân loại dữ liệu:

|  |  |  |
| --- | --- | --- |
|  | Số lượng | Tỉ lệ |
| Positive | 8660 | 43.30% |
| Neutral | 7022 | 35.11% |
| Negative | 4318 | 21.59% |



Hình 1. Biểu đồ cột thể hiện số lượng dữ liệu theo 3 loại cảm xúc

Ảnh có chứa vòng tròn, ảnh chụp màn hình, Đồ họa, Nhiều màu sắc

Mô tả được tạo tự động

Hình 2. Biểu đồ tròn thể hiện tỷ lệ dữ liệu theo 3 loại cảm xúc

Dựa vào dữ liệu và những biểu đồ trực quan hóa, chúng ta có thể thấy những bình luận tích cực chiếm tỉ lệ lớn hơn trung tính và tiêu cực. Số lượng bình luận được phân loại là Positive chiếm tỉ lệ lớn hơn so với Neutral và Negative,chiếm tỉ lệ là 43.30%. Trong khi đó, số lượng bình luận phân loại Neutral chỉ chiếm 35,11% và Negative là 21,59%. Điều này có thể cho thấy rằng trong tập dữ liệu, người dùng Twitter có xu hướng tích cực hơn là tiêu cực hoặc trung tính.

Từ đó chúng ta có thể đặt ra 1 số câu hỏi:

1. Tại sao tỉ lệ tweet tích cực lại cao hơn so với tweet tiêu cực và trung tính?

Điều này có thể do người dùng Twitter thường có xu hướng chia sẻ những trải nghiệm tích cực, bày tỏ cảm xúc tốt hoặc khen ngợi về sản phẩm, dịch vụ hoặc chủ đề đang được xét. Ngoài ra, có thể có những chiến lược marketing hoặc chiến dịch quảng cáo tích cực nhắm đến người dùng Twitter, làm cho tỉ lệ tweet tích cực tăng lên.

1. Tại sao số lượng tweet trung tính lại chiếm tỉ lệ cao?

Số lượng tweet trung tính chiếm tỉ lệ cao có thể do một số người dùng Twitter không có cảm xúc rõ ràng về chủ đề đang được xét, hoặc không muốn chia sẻ cảm xúc của mình. Ngoài ra, có thể có những tweet chỉ đơn giản là chia sẻ thông tin hoặc tin tức mà không mang tính chất cảm xúc cao.

1. Tại sao cần quan tâm đến tweet tiêu cực?

Tweet tiêu cực có thể cho thấy các vấn đề hoặc thách thức đang tồn tại trong chủ đề đang được xét, hoặc cảm giác tiêu cực của người dùng về chủ đề đó. Việc phân tích các tweet tiêu cực này có thể giúp bạn hiểu rõ hơn về các vấn đề này và đưa ra các giải pháp hoặc cải thiện. Ngoài ra, tweet tiêu cực cũng có thể ảnh hưởng đến hình ảnh và uy tín của sản phẩm hoặc thương hiệu đang được xét, do đó cần được quan tâm và giải quyết kịp thời.

#### 4.2 Sử dụng biểu đồ đám mây (WordCloud)

Biểu đồ đám mây (Word Cloud) là một công cụ trực quan để trình bày các từ hoặc cụm từ được sử dụng nhiều nhất trong một tập dữ liệu văn bản. Nó thường được sử dụng để trực quan hóa các khái niệm quan trọng hoặc các từ khóa trong một bài viết, một trang web, hoặc một tập dữ liệu khác.

Trong biểu đồ đám mây, các từ được sắp xếp ngẫu nhiên và kích thước của từ phụ thuộc vào tần suất xuất hiện của từ đó trong tập dữ liệu. Những từ có tần suất xuất hiện cao hơn sẽ được trình bày lớn hơn và nổi bật hơn trong biểu đồ.

Biểu đồ đám mây là một công cụ hữu ích để giúp người dùng có cái nhìn tổng quan về tập dữ liệu văn bản, nhận diện các chủ đề quan trọng và các từ khóa liên quan đến chủ đề đó. Nó cũng có thể giúp trực quan hóa các khái niệm và thông tin khó hiểu trở nên dễ hiểu hơn.

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, Đồ họa

Mô tả được tạo tự động

Hình 3. Biểu đồ đám mây thể hiện các bình luận được đánh giá Positive

Chúng ta dễ dàng nhận thầy những từ khóa được xuất hiện lớn và nổi bật trong biểu đồ đám mây của những bình luận được phân loại tích cực như: “good”. “love”, “new”, “thank”, “nice”, “lol”, “better”,… những từ này thường mang ý nghĩa tích cực sâu sắc, chỉ những biểu cảm, tính từ của người bình luận thể hiện cảm xúc vui tươi, phấn khởi.

Ảnh có chứa văn bản, Phông chữ, ảnh chụp màn hình, Đồ họa

Mô tả được tạo tự động

Hình 4. Biểu đồ đám mây thể hiện các bình luận được đánh giá Neutral

Trong trường hợp biểu đồ đám mây của những bình luận được phân loại là Neutral, chúng ta gặp khó khăn khi rút ra những kết luận từ các từ khóa nổi bật. Những từ như "go", "work", "im", "day", "today",... thường không mang nhiều ý nghĩa để phân loại thành tích cực hay tiêu cực.



Hình 5. Biểu đồ đám mây thể hiện các bình luận đánh giá Negative:

Các từ được xuất hiện nhiều và kích thước lớn trong biểu đồ đám mây được phân loại là Negative như “bad”, “hate”, “sad”, “sorri”, “tired”, … thường mang cảm xúc, trạng thái của người dùng thể hiện tính tiêu cực cao.

### 5. Sử dụng các mô hình học máy

Trong lĩnh vực phân tích văn bản, sự sử dụng các mô hình học máy đã mang lại những tiến bộ đáng kể. Với khả năng xử lý và hiểu dữ liệu văn bản tự động, các mô hình học máy đã trở thành công cụ mạnh mẽ cho text analysis.

Các mô hình học máy có khả năng phân loại văn bản, giúp tự động gán nhãn và nhận diện chủ đề hoặc loại văn bản khác nhau. Chúng có thể phân loại email, tin nhắn, bài viết, hay bất kỳ loại văn bản nào dựa trên nội dung và ngữ cảnh. Điều này giúp tối ưu hóa quy trình và tăng hiệu suất công việc.

Mô hình học máy cũng đóng vai trò quan trọng trong phân tích ý kiến và cảm xúc từ văn bản. Chúng có khả năng tự động phân tích và đánh giá ý kiến, đánh giá tích cực hay tiêu cực của người viết. Điều này giúp các doanh nghiệp hiểu được sự phản hồi của khách hàng, đo lường hiệu quả sản phẩm hoặc dịch vụ, và đưa ra các biện pháp cải thiện.

#### 5.1 Giới thiệu kỹ thuật TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) là một phương pháp quan trọng trong xử lý ngôn ngữ tự nhiên và phân tích văn bản. Nó được sử dụng để đánh giá mức độ quan trọng của một từ trong một tài liệu dựa trên tần suất xuất hiện của từ đó trong tài liệu và tần suất nghịch đảo xuất hiện của từ đó trong toàn bộ tập văn bản.

Phương pháp TF-IDF tính toán giá trị cho mỗi từ trong văn bản bằng cách nhân hai yếu tố chính:

* Term Frequency (TF): Đo lường tần suất xuất hiện của một từ trong một tài liệu cụ thể. TF có thể được tính bằng cách đếm số lần xuất hiện của từ đó trong tài liệu hoặc sử dụng các phương pháp chuẩn hóa để tránh ảnh hưởng của độ dài văn bản.
* Inverse Document Frequency (IDF): Đo lường tần suất nghịch đảo xuất hiện của một từ trong toàn bộ tập văn bản. IDF tính toán bằng cách lấy logarit của tổng số tài liệu chia cho số tài liệu chứa từ đó. Điều này giúp giảm trọng số của các từ phổ biến và tăng trọng số của các từ hiếm.

Bằng cách kết hợp giữa TF và IDF, phương pháp TF-IDF đánh giá được sự quan trọng của một từ trong một tài liệu cụ thể so với toàn bộ tập văn bản. Giá trị TF-IDF càng cao cho thấy từ đó có sự đặc trưng và quan trọng trong tài liệu.

Phương pháp TF-IDF được sử dụng rộng rãi trong nhiều ứng dụng, bao gồm tìm kiếm thông tin, phân loại văn bản, khai phá dữ liệu và trích xuất thông tin. Nó giúp cải thiện hiệu suất và chính xác trong việc xử lý và đánh giá văn bản, và là một công cụ quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên.

#### 5.2 Áp dụng các mô hình học máy

Trong phần lần này, chúng ta sẽ đưa dữ liệu về dạng ma trận TF-IDF (Term Frequency-Inverse Document Frequency) sau đó sử dụng 3 mô hình học máy để train và test tập dữ liệu:

1. Bernoulli Naive Bayes
2. Linear Support Vector Classifier
3. Random Forest

Bộ dữ liệu được chia làm 2 phần huấn luyện (training) và tập kiểm tra (testing) với tỉ lệ tương ứng 80% - 20% dựa trên phương pháp lấy mẫu ngẫu nhiên.

|  |  |  |
| --- | --- | --- |
|  | X\_train | X\_test |
| Độ lớn tập dữ liệu | 16000 | 4000 |

Sau đó, chúng ta đưa dữ liệu về dạng ma trận TF-IDF để đánh giá cho biết mức độ quan trọng của từ trong tài liệu. Giá trị càng cao thì từ càng quan trọng và đặc biệt trong tài liệu đó.

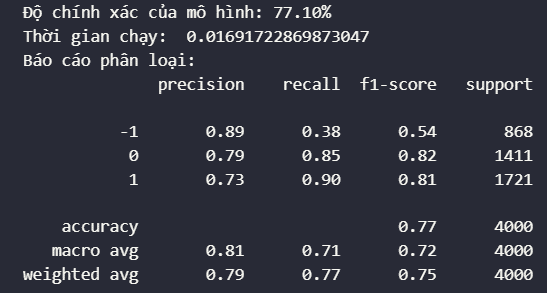
Số lượng từ được lấy làm thuộc tính trong ma trận TF-IDF: 15671

##### 5.2.1 Bernoulli Naive Bayes

Bernoulli Naive Bayes là một thuật toán trong Machine Learning và xử lý ngôn ngữ tự nhiên được sử dụng trong bài toán phân loại văn bản. Nó là một biến thể của thuật toán Naive Bayes dựa trên giả định rằng các đặc trưng (features) đều là nhị phân (binary), tức chỉ nhận giá trị 0 hoặc 1.

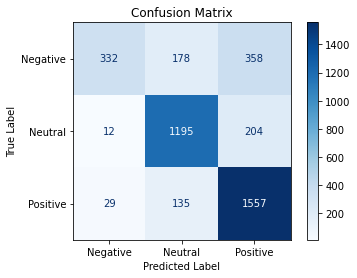
Thuật toán Bernoulli Naive Bayes dựa trên lý thuyết xác suất và nguyên lý Bayes để xác định xác suất của một văn bản thuộc vào một lớp (class) cụ thể. Nó được gọi là "Bernoulli" vì phân phối Bernoulli được sử dụng để mô hình hóa xác suất của các đặc trưng.

Ta thu được Output:



Hình 5. Bảng đánh giá độ chính xác của mô hình với

thuật toán Bernoulli Naive Bayes lần 1



Hình 6. Biểu đồ thể hiện hiệu suất của thuật toán Bernoulli Naive Bayes lần 1

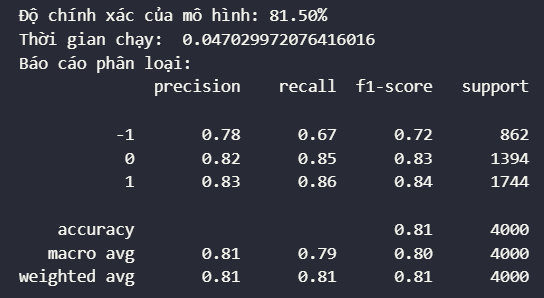
Ta có thể thấy, độ chính xác vẫn còn thấp, thuật toán chạy chậm vì số lượng features quá lớn: 15671. Chúng ta có thể giảm số lượng features để thuật toán học nhanh hơn.

- Thiết lập tham số max\_features=5000 trong lúc tạo đối tượng TfidfVectorizer. Tuy nhiên cách này sẽ mất đi nhiều từ có ý nghĩa trong từ điển.

- Thiết lập tham số min\_df=4 để loại bỏ những feature có số lượng từ tần suất xuất hiện ít hơn 4 lần.

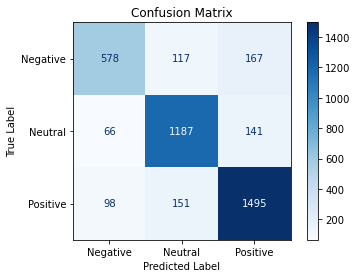
  Số feature ban đầu 15671. Sau khi tối ưu, số feature là 3080.

Chúng ta cùng chạy model lại 1 lần nữa.



Hình 7. Bảng đánh giá độ chính xác của mô hình với

thuật toán Bernoulli Naive Bayes sau khi giảm số feature



Hình 8. Biểu đồ thể hiện hiệu suất của thuật toán Bernoulli Naive Bayes

sau khi giảm số feature

Chúng ta có thể loại bỏ đáng kể feature xuất hiện ít dẫn đến gây nhiễu dữ liệu. Độ chính xác thuật toán tăng từ 77.1% lên 81.5%.

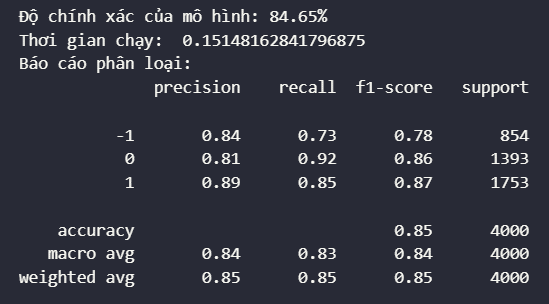
Thời gian chạy thuật toán cũng nhanh hơn. Khác biệt khi chạy thuật toán với số lượng feature lớn và số lần chạy thuật toán lớn.

=> Không phải lúc nào dùng nhiều data để học cũng tốt.

##### 5.2.2 Linear Support Vector Classifier

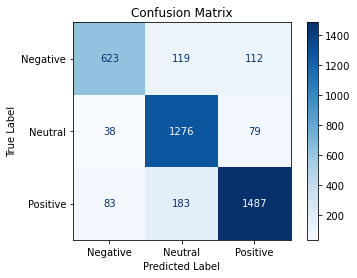
Linear SVC là một phiên bản đơn giản của SVM, trong đó phân phối tuyến tính (linear) được sử dụng để xác định ranh giới phân chia giữa các lớp. Mục tiêu của thuật toán là tìm một siêu phẳng (hyperplane) trong không gian đặc trưng sao cho các điểm dữ liệu thuộc vào các lớp khác nhau nằm xa nhau nhất.

Kết quả sau khi đưa dữ liệu vào thuật toán:



Hình 9. Bảng đánh giá độ chính xác của mô hình với

thuật toán Linear Support Vector Classifier

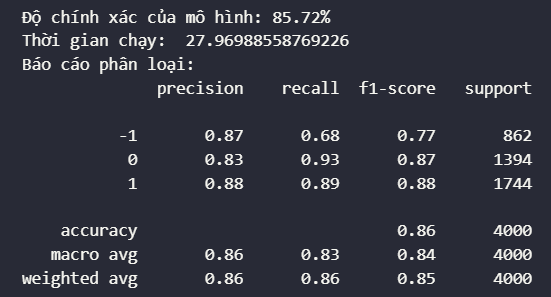


Hình 10. Biểu đồ thể hiện hiệu xuất của thuật toán Linear Support Vector Classifier

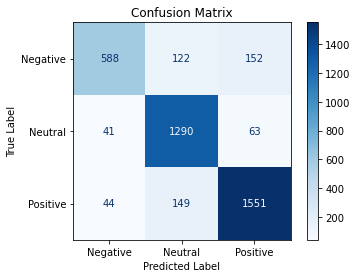
##### 5.2.3 Random Forest

Cấu trúc chính của Random Forest bao gồm một tập hợp các cây quyết định (decision trees). Mỗi cây quyết định được xây dựng dựa trên một phần dữ liệu huấn luyện được lấy mẫu ngẫu nhiên từ tập dữ liệu ban đầu. Khi xây dựng một cây quyết định, quá trình chia dữ liệu được thực hiện bằng cách chọn các đặc trưng ngẫu nhiên từ tập đặc trưng ban đầu. Điều này giúp tạo ra sự đa dạng và độc lập giữa các cây trong rừng.

Khi dự đoán một điểm dữ liệu mới, Random Forest sẽ đưa ra dự đoán bằng cách kết hợp dự đoán của tất cả các cây trong rừng. Trong bài toán phân loại, dự đoán cuối cùng được xác định bằng cách áp dụng "bầu cử đa số" (majority voting) trên các dự đoán cá nhân. Trong bài toán hồi quy, dự đoán cuối cùng được tính bằng giá trị trung bình của các dự đoán cá nhân.



Hình 11. Bảng đánh giá độ chính xác của mô hình với thuật toán Random Forest



Hình 12. Biểu đồ thể hiện hiệu suất của thuật toán Random Forest

#### 5.3 Đánh giá kết quả sau khi sử dụng mô hình học máy

Đối với tập dữ liệu đã được xử lý bên trên, sau khi đưa vào 3 mô hình học máy chúng ta thu được kết quả khá tốt. Độ chính xác của thuật toán Bernoulli Naive Bayes là 81,5%. Độ chính xác của thuật toán Bernoulli Naive Bayes là 84,65%. Độ chính xác của thuật toán Random Forest đạt giá trị cao nhất là 85,72%.

Tuy nhiên, thời gian chạy của RandomForest lại rất lớn, khoảng 27,96s cho 1 lần chạy mô hình trong khi thời gian chạy của Linear SVC là khoảng 0.15148s và thời gian Bernoulli Naive Bayes nhanh nhất là 0.047s

Kết luận, mỗi thuật toán có những ưu nhược điểm khác nhau. Tùy vào từng bài toán và yêu cầu của người thực hiện thì sẽ áp dụng vào những thuật toán khác nhau để đưa ra kết quả phù hợp.

# KẾT LUẬN

Bằng cách sử dụng các công cụ phân tích dữ liệu và xử lý ngôn ngữ tự nhiên, chúng ta đã phân tích tình cảm truyền thông và tìm hiểu cách mà các phương tiện truyền thông ảnh hưởng đến khán giả. Kết quả phân tích cho thấy rằng truyền thông đóng vai trò quan trọng trong việc tạo ra và truyền tải các tình cảm, nhận thức và ý kiến của khán giả.

Tuy nhiên, cần lưu ý rằng social media sentiment analysis không phải là một công cụ hoàn hảo và vẫn tồn tại những hạn chế. Việc đánh giá tình cảm là một quá trình phức tạp và cần sự đánh giá chính xác và đầy đủ từ phía các chuyên gia thực hiện. Một số vấn đề thường gặp bao gồm sự đa nghĩa và ngữ cảnh, đặc biệt là trong các ngôn ngữ khác nhau, cũng như sự ảnh hưởng của định kiến và sự tiếp nhận cá nhân đối với các thông điệp.

Do đó, social media sentiment analysis chỉ nên được coi là một phần trong quá trình đánh giá và quản lý truyền thông và tiếp thị của một tổ chức. Để đưa ra các quyết định chiến lược tốt nhất cho tổ chức của mình, các chuyên gia cần phải kết hợp các phương pháp khác nhau và sử dụng sự đánh giá chính xác và khách quan. Ngoài ra, việc áp dụng các phương tiện truyền thông phù hợp cũng là một yếu tố quan trọng để nâng cao chất lượng những cảm xúc tích cực mà người dùng mang lại. Từ đó, các công ty có thể tăng cường sự hài lòng của khách hàng, tạo ra mối quan hệ tốt hơn với họ và tăng doanh số bán hàng.

Tóm lại, phân tích tình cảm truyền thông là những công cụ hữu ích giúp các tổ chức đánh giá và cải thiện chiến lược truyền thông và tiếp thị của mình. Tuy nhiên, để sử dụng hiệu quả các công cụ này, cần phải có sự kết hợp giữa việc sử dụng công nghệ và sự đánh giá chuyên môn từ các chuyên gia. Từ đó, các tổ chức có thể tối ưu hóa chiến lược truyền thông và tiếp thị, tăng cường sự hài lòng của khách hàng và tạo ra những mối quan hệ tốt hơn với họ.

# TÀI LIỆU THAM KHẢO

[1] Nguyễn Văn Hiếu . “Giới thiệu về TF - IDF" [Posted 08/2021]

<https://blog.luyencode.net/tf-idf-la-gi/>

[2] Nguyễn Văn Thiệu . “DAP-14-Text-Analysis.ipynb”

<https://colab.research.google.com/drive/17WudyoYF1b6ZwYlqvKQW4V0B-HozBqM8?usp=sharing#scrollTo=DS9uuHwUKuX4>

[3] MILOUD BELAREBIA. “Sentiment140 dataset (1,600,000 tweets)” – Kaggle [Updated 2021]

<https://www.kaggle.com/datasets/milobele/sentiment140-dataset-1600000-tweets>