**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO MÔN HỌC**

**Phân tích dữ liệu**

**Đề tài: Phân tích cảm xúc phản hồi của sinh viên**

**Thành viên:**

**Trần Minh Triều MSSV: 2254052085**

**Đào Hùng Thuận MSSV: 2254052079**

**Trần Ngọc Tài MSSV: 2254052071**

**Giảng viên: Thầy Hồ Hướng Thiên**

***TP. Hồ Chí Minh, tháng 12 năm 2024***

[**Lời mở đầu**](#_6lnkbdlcvmo7) 1

[1. Giới thiệu về phân tích cảm xúc](#_61fx38d68vuo) 2

[2. Tổng quan về dữ liệu](#_gr9y971fsw5p) 2

[2.1 Nguồn dữ liệu](#_gz8oofrsyezc) 2

[2.2 Mô tả dữ liệu](#_6z87r37gpqp2) 3

2.3 Kiểm tra dữ liệu6

[2.4 Tiền xử lý](#_yrud0mosdes1)9

[3. Đánh giá các mô hình được sử dụng](#_fedtq8hk6g9x) 10

[3.1 Mô hình Logistic Regression](#_o40zna44ccmy)10

[3.2 Mô hình Random](#_izl9eeo84l) Forest13

3.3[. Mô hình SVM:](#_cc75aao4hdkt) 14

4.Kết luận 14

# Lời mở đầu

Trong bối cảnh giáo dục ngày càng hiện đại và phát triển, việc hiểu rõ cảm xúc và phản hồi của người học trở nên ngày càng quan trọng, đặc biệt là đối với các khóa học và hoạt động học tập. Phản hồi từ sinh viên không chỉ cung cấp cái nhìn sâu sắc về chất lượng giảng dạy mà còn phản ánh mức độ hài lòng và trải nghiệm học tập của họ. Phân tích cảm xúc từ phản hồi của sinh viên giúp nhận diện những điểm mạnh và hạn chế trong quá trình giảng dạy, từ đó đưa ra các biện pháp cải thiện phù hợp.

Báo cáo này tập trung vào việc phân tích dữ liệu phản hồi của sinh viên, áp dụng các kỹ thuật phân tích cảm xúc để xác định cảm xúc tích cực, tiêu cực hoặc trung tính. Qua đó, chúng em không chỉ hiểu rõ hơn về dữ liệu mà còn chế tạo các mô hình máy học, đồng thời khám phá nguyên nhân và bối cảnh đằng sau những cảm xúc đó.

Mặc dù nhóm đã nỗ lực hết sức, chắc chắn không thể tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự góp ý từ thầy để có thể hoàn thiện hơn trong tương lai. Chúng em xin chân thành cảm ơn thầy!

## 1. Giới thiệu về phân tích cảm xúc

Phân tích cảm xúc (Sentiment Analysis) là một nhánh quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), tập trung vào việc xác định, phân loại và hiểu cảm xúc hoặc thái độ được biểu đạt trong văn bản. Mục tiêu chính của phân tích cảm xúc là xác định xem nội dung văn bản mang tính tích cực, tiêu cực hay trung lập. Đây là một công cụ hữu ích trong nhiều lĩnh vực như tiếp thị, chăm sóc khách hàng, chính trị và nghiên cứu xã hội, giúp các tổ chức thu thập phản hồi từ người dùng, đo lường mức độ hài lòng, hoặc dự đoán xu hướng.

Phân tích cảm xúc thường được thực hiện thông qua các phương pháp như sử dụng từ điển cảm xúc (lexicon-based), mô hình học máy (machine learning), hoặc học sâu (deep learning). Đối với tiếng Việt, các thách thức như ngữ pháp phức tạp, từ ngữ đồng nghĩa hoặc ý nghĩa ngữ cảnh đặt ra nhiều khó khăn, nhưng đồng thời cũng mở ra cơ hội để phát triển các giải pháp kỹ thuật mới và hiệu quả.

Ứng dụng của phân tích cảm xúc không chỉ dừng lại ở việc phân loại cảm xúc, mà còn có thể mở rộng sang phân tích sâu hơn như đo lường cường độ cảm xúc, nhận diện đối tượng cụ thể được đề cập trong văn bản, hoặc dự đoán xu hướng xã hội dựa trên dữ liệu thu thập.

## 2. Tổng quan về dữ liệu

### 2.1 Nguồn dữ liệu

Dữ liệu được thực hiện bởi nhóm nghiên cứu NLP@UIT tại Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TP.HCM, nhằm phục vụ nghiên cứu phân tích cảm xúc và giáo dục. Dữ liệu được sử dụng trong nghiên cứu này là phản hồi của sinh viên từ một trường đại học tại Việt Nam. Các phản hồi được thu thập qua các khảo sát sinh viên do nhà trường thực hiện vào cuối mỗi học kỳ trong khoảng thời gian từ năm 2013 đến 2016. Tổng cộng thu thập được hơn 16.000 câu phản hồi từ sinh viên về nhiều vấn đề khác nhau liên quan đến giảng viên, chương trình học, cơ sở vật chất và các chủ đề khác.

Mỗi câu phản hồi thường là một câu ngắn, được viết tự do, chứa nhiều dạng thông tin phong phú như từ viết tắt, lỗi chính tả, biểu tượng cảm xúc và các ký tự đặc biệt. Mỗi câu có thể mang tính tích cực, tiêu cực hoặc mang tính trung lập.

Dữ liệu từ năm học 2014-2015 có 175 giảng viên, 2.235 sinh viên, 143 môn học và 6.038 phản hồi. Đến năm học 2015-2016, số liệu tăng lên với 184 giảng viên, 2.856 sinh viên, 160 môn học và 6.288 phản hồi. Đặc biệt, trong năm học 2016-2017, số liệu thu thập đạt 227 giảng viên, 3.789 sinh viên, 175 môn học và 13.417 phản hồi.

Các phản hồi này được phân tích và tiền xử lý kỹ lưỡng. Quá trình tiền xử lý bao gồm phân tách câu, thay thế từ viết tắt bằng từ đầy đủ và sửa lỗi chính tả. Những từ viết tắt và lỗi chính tả được lưu trữ trong một từ điển để hỗ trợ các bước khai thác dữ liệu sau này. Để bảo mật thông tin cá nhân, tên riêng được ẩn danh bằng các ký hiệu thay thế.

Kết quả của quá trình tiền xử lý là hơn 16.000 câu phản hồi được chuẩn hóa, sẵn sàng cho các nhiệm vụ phân loại dựa trên cảm xúc và chủ đề. Dữ liệu này là một nguồn tài nguyên giá trị, phục vụ nghiên cứu trong lĩnh vực phân tích cảm xúc và giáo dục.

### 2.2 Mô tả dữ liệu

|  |  |  |
| --- | --- | --- |
| topics | sentences | sentiments |
| 1 | slide giáo trình đầy đủ . | 2 |
| 0 | nhiệt tình giảng dạy , gần gũi với sinh viên . | 2 |
| 1 | đi học đầy đủ full điểm chuyên cần . | 0 |
| 0 | chưa áp dụng công nghệ thông tin và các thiết bị... | 0 |
| 0 | thầy giảng bài hay , có nhiều bài tập ví dụ ngắn. | 2 |

#### Giải thích từng cột:

1. topics: Giá trị nhãn chủ đề là một trong bốn chủ đề:

* '0': Chủ đề liên quan đến giảng viên.
* '1': Chủ đề liên quan đến chương trình đào tạo.
* '2': Chủ đề liên quan đến cơ sở vật chất.
* '3': Các chủ đề khác.

2. sentences: Cột này chứa các câu nhận xét hoặc mô tả ngắn gọn về trải nghiệm của người dùng.

3. sentiments: Cột này chứa các nhãn cảm xúc:

* 2: Cảm xúc tích cực.
* 1: Cảm xúc trung tính.
* 0: Cảm xúc tiêu cực

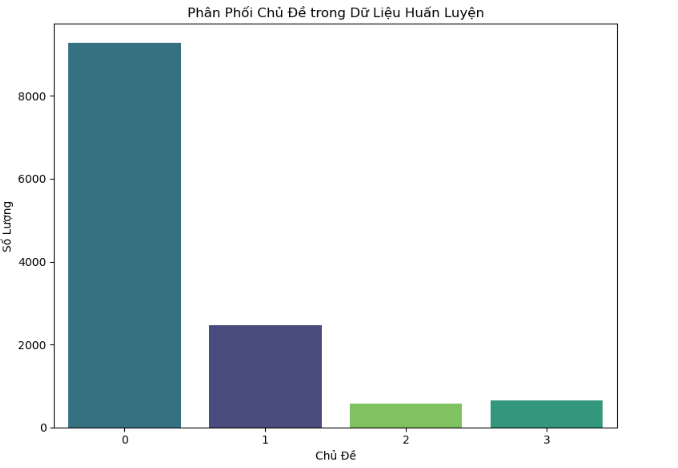
### Mô tả chi tiết dữ liệu xuất ra

#### 1. Tập dữ liệu train\_df: Gồm 11,426 dòng dữ liệu.

#### 2. Tập dữ liệu test\_df: Gồm 3,166 dòng dữ liệu.

#### 3. Tập dữ liệu val\_df: Tổng số dòng: Gồm 1,583 dòng dữ liệu.

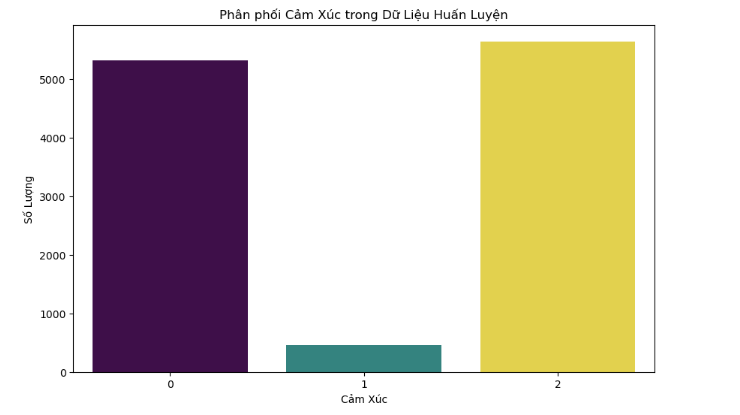
### Biểu đồ phân phối chủ đề trong dữ liệu train



Nhận xét:

Dữ liệu huấn luyện có sự phân bố không đồng đều giữa các chủ đề, với chủ đề 0 chiếm tỷ lệ áp đảo. Chủ đề 2 và 3: Có số lượng mẫu thấp hơn nhiều, mỗi chủ đề chỉ khoảng vài trăm

### Biểu đồ phân phối cảm xúc trong dữ liệu train



#### **Phân phối cụ thể của các cảm xúc:**

#### Cảm Xúc 0:

#### Có số lượng mẫu lớn, khoảng 5000 - 5200 mẫu.

#### Đây là một trong hai cảm xúc chiếm tỷ lệ cao nhất trong tập dữ liệu.

* Cảm Xúc 1:
* Có số lượng mẫu rất thấp, chỉ khoảng 500 mẫu, thấp hơn đáng kể so với các cảm xúc khác.
* Sự chênh lệch lớn cho thấy cảm xúc này ít phổ biến trong tập dữ liệu.
* Cảm Xúc 2:
* Có số lượng mẫu tương tự cảm xúc 0, khoảng 5200 - 5500 mẫu.
* Đây cũng là một trong hai cảm xúc chiếm tỷ lệ lớn nhất trong tập dữ liệu.

#### **Nhận xét:**

#### Sự thiên lệch dữ liệu:

#### Tập dữ liệu có sự mất cân bằng lớn giữa các cảm xúc, với cảm xúc 0 và 2 chiếm hơn 90% dữ liệu. ưu thế, còn cảm xúc 1 rất ít, là cảm xúc hiếm gặp nhất trong dữ liệu, chiếm chưa đến 10% tổng số mẫu.

* Mô hình học máy có thể bị thiên lệch về cảm xúc 0 và 2, dẫn đến khả năng dự đoán cảm xúc 1 không tốt.
* Ảnh hưởng đến mô hình:
* Các cảm xúc có số lượng lớn (0 và 2) sẽ được mô hình ưu tiên học hơn.
* Cảm xúc 1, do có số lượng nhỏ, có thể không được nhận diện tốt hoặc bị bỏ qua trong quá trình huấn luyện.

### 2.3 Kiểm tra dữ liệu

**1. Kiểm tra dữ liệu bị thiếu (isna().sum()):**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên file | topics | sentences | sentiments |
| File test | 0 | 0 | 0 |
| File train | 0 | 0 | 0 |
| File val | 0 | 0 | 0 |

* Từ kết quả trên, chúng ta thấy không có giá trị nào bị thiếu (NA) trong các cột "topics", "sentences", và "sentiments" của tập test, train, và valid. Không cần phải xóa những dòng bị thiếu trong dữ liệu.

**2. Kiểm tra kiểu dữ liệu (dtypes):**

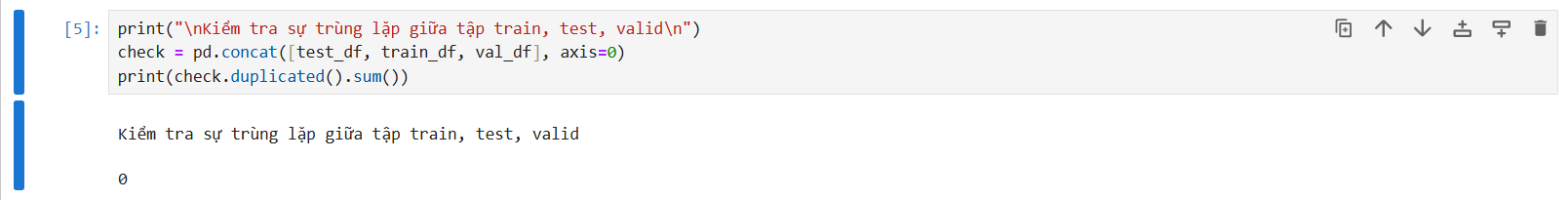
topics object

sentences object

sentiments object

Kiểu dữ liệu của các cột topics, sentences, sentiments có kiểu dữ liệu object, nghĩa là chúng là các chuỗi văn bản. Ta cần phải tiền xử lý cột sentiment sang kiểu dữ liệu số.

**3. Kiểm tra sự trùng lặp dữ liệu giữa các file (duplicated().sum()):**



Không có dòng dữ liệu trùng lặp trong tập dữ liệu được ghép lại từ ba tập dữ liệu test, train, và valid.

## 2.4 Tiền xử lý

### Chuẩn hóa văn bản:

Chuyển tất cả văn bản về chữ thường để đảm bảo tính nhất quán và loại bỏ khoảng trắng thừa thãi.

Mục đích:

* Đảm bảo tính nhất quán dữ liệu: khi xử lý ngôn ngữ tự nhiên, hệ thống cần phải hiểu rằng các từ như "Hà Nội" và "hà nội" là cùng một từ, không phân biệt chữ hoa hay chữ thường. Bằng cách chuyển tất cả về chữ thường, cần loại bỏ sự khác biệt không cần thiết, giúp dữ liệu trở nên đồng nhất và dễ xử lý hơn.
* Giảm nhiễu trong dữ liệu: Khoảng trắng thừa (như dư ở đầu, giữa, hoặc cuối câu) không mang thông tin ý nghĩa nhưng lại có thể làm tăng độ phức tạp của mô hình, ảnh hưởng đến kết quả phân tích. Loại bỏ chúng giúp dữ liệu "sạch" hơn, giảm nhiễu và cải thiện hiệu suất của các thuật toán.
* Cải thiện chất lượng mô hình: Dữ liệu nhất quán và sạch sẽ làm giảm lỗi trong việc so khớp từ vựng hoặc ánh xạ vector từ, từ đó giúp mô hình học được đúng các quy luật trong dữ liệu, nâng cao độ chính xác trong các bài toán như phân tích cảm xúc, phân loại văn bản, v.v.

1. **Chuyển đổi cột sentiments từ object sang số**

Việc chuyển đổi cột sentiments (nhãn cảm xúc) từ kiểu dữ liệu object sang kiểu số là một bước quan trọng để chuẩn bị dữ liệu cho mô hình học máy hoặc các thuật toán phân tích.

1. **Loại bỏ dấu câu và ký tự đặc biệt:**

Mục đích:

* Loại bỏ nhiễu trong dữ liệu: Dấu câu và ký tự đặc biệt (như !, @, #, ...) thường không mang ý nghĩa ngữ cảnh hoặc cảm xúc trong văn bản, đặc biệt đối với các bài toán như phân tích cảm xúc. Những ký tự này có thể gây nhiễu cho mô hình học máy, khiến quá trình học kém hiệu quả.
* Đơn giản hóa tập từ vựng: Khi xử lý ngôn ngữ tự nhiên, hệ thống cần tạo ra một tập từ vựng để biểu diễn dữ liệu. Nếu không loại bỏ các ký tự đặc biệt, tập từ vựng sẽ phình to không cần thiết, làm tăng độ phức tạp của mô hình và giảm hiệu quả trong việc học các mối quan hệ giữa từ ngữ.
* Loại bỏ các ký tự không liên quan đến ngôn ngữ tự nhiên: Nhiều ký tự đặc biệt như @, #, %, hoặc \* không phải là một phần của ngôn ngữ tự nhiên và thường không mang ý nghĩa trong văn bản (ngoại trừ trường hợp đặc biệt như phân tích hashtag trên mạng xã hội). Việc loại bỏ chúng giúp mô hình tập trung vào những từ ngữ có giá trị phân tích thực sự.
* Đảm bảo tương thích với công cụ xử lý văn bản: Một số công cụ hoặc thư viện NLP có thể không xử lý tốt các ký tự đặc biệt hoặc dấu câu, dẫn đến lỗi hoặc kết quả không mong muốn. Loại bỏ các ký tự này trước khi sử dụng giúp quá trình xử lý ổn định hơn.

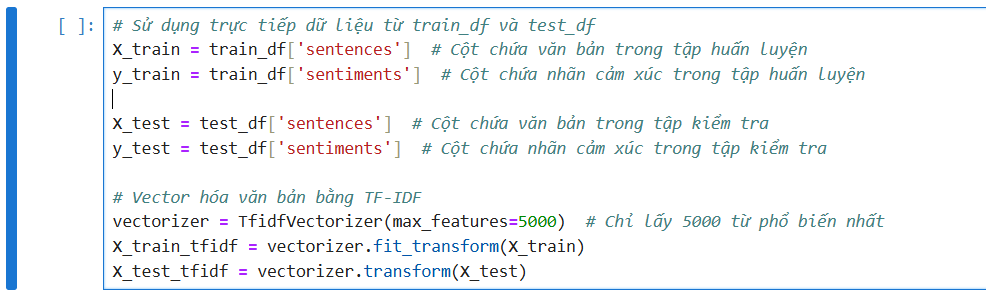
1. **Loại bỏ trùng lặp:**

Trong phân tích cảm xúc, dữ liệu văn bản thường chứa các câu hoặc mẫu dữ liệu bị lặp lại. Loại bỏ các bản ghi trùng lặp là một bước quan trọng trong quá trình tiền xử lý. Việc này đảm bảo chất lượng dữ liệu đầu vào và cải thiện hiệu suất của mô hình.

Mục đích:

* Tránh thiên lệch dữ liệu (Bias): Dữ liệu trùng lặp có thể làm cho một loại cảm xúc (tiêu cực, trung lập hoặc tích cực) xuất hiện nhiều hơn, dẫn đến mô hình bị thiên lệch.
* Giảm kích thước dữ liệu: Loại bỏ các bản ghi trùng lặp làm giảm khối lượng dữ liệu, giúp giảm thời gian huấn luyện mô hình.
* Tăng tính tổng quát: Dữ liệu trùng lặp không cung cấp thêm thông tin mới mà có thể khiến mô hình trở nên quá khớp với dữ liệu huấn luyện (overfitting).
* Giữ lại một bản ghi duy nhất: Chỉ giữ lại một bản ghi trong các dòng trùng lặp.

1. **Vector hóa văn bản bằng TF-IDF**

****

**5.1 Chuẩn bị dữ liệu đầu vào:**

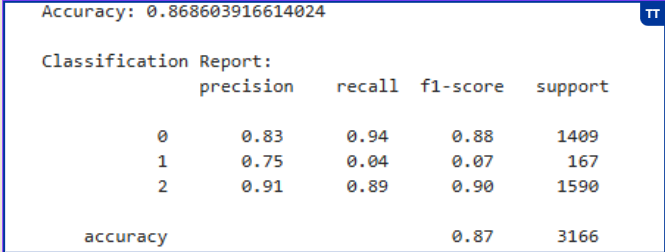
* X\_train và y\_train với:
* X\_train chứa các câu văn bản cần phân tích.
* y\_train chứa nhãn cảm xúc tương ứng (ví dụ: 0, 1, 2).
* X\_test và y\_test: Dữ liệu kiểm tra, được sử dụng để đánh giá mô hình.

### 5.2 Vector hóa văn bản bằng TF-IDF với max\_features=5000

X\_train\_tfidf và X\_test\_tfidf: Dữ liệu đầu vào (các câu) đã được chuyển đổi thành ma trận số, sẵn sàng để huấn luyện mô hình.

## 3. Đánh giá mô hình sử dụng

### 3.1 Mô hình Logistic Regression



### 1. Accuracy

* **Accuracy:** 0.87 (~87%) là một chỉ số tốt, cho thấy mô hình có khả năng dự đoán đúng 87% trên toàn bộ tập dữ liệu. Tuy nhiên, chỉ số này vẫn có thể bị ảnh hưởng bởi sự chênh lệch giữa số lượng mẫu của các lớp (class imbalance).

### 2. Classification Report

#### **Class 0 (Tiêu cực):**

* **Precision:** 0.83 (rất tốt, tức là trong những dự đoán là tiêu cực, 83% là đúng).
* **Recall:** 0.94 (rất cao, mô hình nhận diện hầu hết các câu tiêu cực).
* **F1-score:** 0.88 (thể hiện sự cân bằng giữa precision và recall, hiệu quả cao).

#### **Class 1 (Trung lập):**

* **Precision:** 0.75 (khá tốt, nhưng chưa cao như lớp tiêu cực và tích cực).
* **Recall:** 0.04 (rất thấp, nghĩa là mô hình bỏ sót rất nhiều câu trung lập).
* **F1-score:** 0.07 (rất thấp, hiệu quả kém khi xử lý các câu trung lập).
* **Nguyên nhân:** Số lượng mẫu cho lớp trung lập (167) ít hơn rất nhiều so với các lớp còn lại, dẫn đến hiện tượng mất cân bằng lớp (class imbalance).

#### **Class 2 (Tích cực):**

* **Precision:** 0.91 (rất cao, dự đoán tích cực chính xác).
* **Recall:** 0.89 (rất tốt, mô hình nhận diện hầu hết các câu tích cực).
* **F1-score:** 0.90 (rất hiệu quả trong việc phân loại câu tích cực).

### 3. Nhận xét

#### **Hiệu suất tổng thể:**

* Mô hình có kết quả tốt, phản ánh khả năng dự đoán chính xác trên toàn bộ dữ liệu với **Accuracy** 87%.
* Tuy nhiên, chỉ số **Accuracy** có thể gây hiểu lầm do sự chênh lệch lớn về số lượng mẫu giữa các lớp (class imbalance).

#### **Hiệu suất theo từng lớp:**

* **Lớp tiêu cực (0):** Được phân loại rất tốt, với các chỉ số precision, recall, và f1-score cao.
* **Lớp tích cực (2):** Hiệu quả cũng rất cao, gần như hoàn hảo trong việc nhận diện các câu tích cực.
* **Lớp trung lập (1):** Gặp vấn đề nghiêm trọng, với recall rất thấp (0.04) và f1-score chỉ đạt 0.07. Mô hình hầu như không thể nhận diện được các câu trung lập.

#### **Tác động của mất cân bằng lớp:**

* Mô hình chịu ảnh hưởng lớn từ sự mất cân đối số lượng mẫu giữa các lớp. Lớp trung lập có số mẫu rất ít (167), dẫn đến hiệu suất thấp và kéo giảm các chỉ số trung bình như **macro recall** và **macro f1-score**.

### 3.2 Mô hình Random Forest

**Mô hình Random Forest:**

*Độ chính xác: 0.8543903979785218*

*Classification Report:*

*precision recall f1-score support*

*0 0.84 0.90 0.87 1409*

*1 0.50 0.14 0.22 167*

*2 0.88 0.89 0.88 1590*

### 1. Accuracy

* **Accuracy** là 0.854 (~85.4%), cho thấy mô hình phân loại đúng 85.4% trong tổng số trường hợp. Đây là một kết quả khá tốt, đặc biệt nếu dữ liệu có sự cân bằng giữa các lớp.

### 2. Classification Report

#### **Class 0 (Tiêu cực):**

* + **Precision**: 0.84 (84%) nghĩa là trong tất cả các dự đoán là lớp 0, có 84% là đúng.
  + **Recall**: 0.90 (90%) nghĩa là trong tất cả các mẫu thực sự thuộc lớp 0, mô hình nhận diện đúng 90%.
  + **F1-score**: 0.87 (87%) cho thấy sự cân bằng tốt giữa precision và recall đối với lớp 0.
  + **Nhận xét**: Mô hình hoạt động rất tốt với lớp 0.

#### **Class 1 (Trung lập):**

* + **Precision**: 0.50 (50%) khá thấp, cho thấy trong các dự đoán là lớp 1, chỉ một nửa là chính xác.
  + **Recall**: 0.14 (14%) cực thấp, nghĩa là chỉ 14% các mẫu thực sự thuộc lớp 1 được nhận diện đúng.
  + **F1-score**: 0.22 (22%) rất thấp, cho thấy mô hình gặp khó khăn lớn với lớp này.
  + **Nhận xét**: Hiệu suất trên lớp trung lập rất kém, có thể do dữ liệu lớp này mất cân bằng (167 mẫu, chỉ chiếm ~5.3% tổng dữ liệu).

#### **Class 2 (Tích cực):**

* + **Precision**: 0.88 (88%) nghĩa là trong các dự đoán là lớp 2, có 88% là đúng.
  + **Recall**: 0.89 (89%) nghĩa là trong tất cả các mẫu thực sự thuộc lớp 2, mô hình nhận diện đúng 89%.
  + **F1-score**: 0.88 (88%) cho thấy mô hình xử lý lớp tích cực rất tốt.

### 3. Nhận xét tổng quan

* Mô hình hoạt động rất tốt với các lớp 0 (tiêu cực) và 2 (tích cực), nhưng hiệu suất trên lớp 1 (trung lập) rất kém. Điều này có thể do:
  + **Dữ liệu không cân bằng**: Lớp trung lập chỉ chiếm một phần rất nhỏ trong tổng số dữ liệu.
  + **Đặc điểm dữ liệu**: Có thể lớp trung lập có các đặc trưng không rõ ràng, dễ bị nhầm lẫn với lớp tiêu cực hoặc tích cực.

### 3.3 Mô hình SVM

**Mô hình SVM (Support Vector Machine):**

*Độ chính xác: 0.8698673404927353*

*Classification Report:*

*precision recall f1-score support*

*0 0.84 0.94 0.89 1409*

*1 0.69 0.05 0.10 167*

*2 0.91 0.89 0.90 1590*

**1. Accuracy:**

* Độ chính xác 86.99% là khá cao, cho thấy mô hình hoạt động tốt trên toàn bộ dữ liệu kiểm thử.

### 2. Classification Report

* **Class 0 (tiêu cực):**
  + Precision: 0.84 (Tỷ lệ dự đoán đúng trên tất cả các dự đoán là nhãn 0).
  + Recall: 0.94 (Tỷ lệ dự đoán đúng trên tổng số mẫu thực sự là nhãn 0).
  + F1-score: 0.89 (cân bằng giữa precision và recall, rất tốt).
  + Nhãn 0 có hiệu suất rất cao, đặc biệt recall tốt, tức là mô hình nhận diện rất tốt các câu tiêu cực.
* **Class 1 (trung lập):**
  + Precision: 0.69 (hợp lý).
  + Recall: 0.05 (rất thấp, mô hình bỏ sót gần như toàn bộ các mẫu nhãn 1).
  + F1-score: 0.10 (thấp).
  + Nhãn trung lập có hiệu suất kém, nguyên nhân chính là recall rất thấp. Điều này có thể do mất cân bằng dữ liệu (số lượng mẫu nhãn 1 chỉ chiếm 167 trên tổng 3166).
* **Class 2 (tích cực):**
  + Precision: 0.91 (rất cao).
  + Recall: 0.89 (rất tốt).
  + F1-score: 0.90 (rất tốt).
  + Nhãn 2 được mô hình xử lý hiệu quả, gần tương đương với nhãn 0.

**3. Nhận xét tổng quan**

* **Hiệu suất tổng thể:**
  + Mô hình đạt độ chính xác cao (**86.99%**), cho thấy nó hoạt động tốt trên tập dữ liệu kiểm thử.
  + Tuy nhiên, độ chính xác không phản ánh toàn bộ hiệu suất do tập dữ liệu không cân bằng.
* **Hiệu suất theo nhãn:**
  + Mô hình dự đoán tốt với các nhãn chiếm đa số:
    - **Nhãn 0 (tiêu cực):** Dự đoán rất tốt với **F1-score: 0.89**.
    - **Nhãn 2 (tích cực):** Hiệu suất cao với **F1-score: 0.90**.
  + **Nhãn 1 (trung lập):** Hiệu suất rất kém, đặc biệt là **recall: 0.05**. Điều này cho thấy mô hình bỏ qua hầu hết các mẫu thuộc nhãn trung lập.
* **Tính cân bằng:**
* Chỉ số **Macro Average (F1-score: 0.63)** cho thấy hiệu suất không đồng đều giữa các nhãn.
* Mất cân bằng dữ liệu (nhãn 1 quá ít) là vấn đề lớn, dẫn đến mô hình thiên vị các nhãn chiếm đa số (0 và 2).
* **Tính cân bằng:**
  + Mô hình SVM phù hợp và hiệu quả trong việc phân loại các nhãn có số lượng lớn, nhưng gặp khó khăn với nhãn ít dữ liệu.
  + Hiệu suất tốt trên các nhãn tiêu cực và tích cực có thể hữu ích trong nhiều ứng dụng, nhưng việc bỏ sót nhãn trung lập sẽ làm giảm giá trị tổng thể trong các kịch bản yêu cầu phân tích chi tiết.

**Ứng dụng và lợi ích của phân tích cảm xúc:**

**Ứng dụng:**

* Phân tích phản hồi khách hàng: Các công ty có thể sử dụng phân tích cảm xúc để hiểu cảm nhận của khách hàng đối với sản phẩm hoặc dịch vụ của họ thông qua các đánh giá, bài viết trên mạng xã hội, phản hồi email, hoặc khảo sát. Điều này giúp cải thiện chất lượng dịch vụ và sản phẩm.
* Giám sát thương hiệu: Các công ty có thể theo dõi và đánh giá cảm xúc của người dùng về thương hiệu của mình trên các nền tảng như Twitter, Facebook, và các diễn đàn trực tuyến, từ đó có thể phản ứng kịp thời với các vấn đề hoặc cơ hội.
* Phân tích dữ liệu mạng xã hội: Các nhà nghiên cứu và công ty có thể sử dụng phân tích cảm xúc để nghiên cứu các xu hướng xã hội, chính trị, hoặc văn hóa thông qua các bài đăng và bình luận trên mạng xã hội.
* Cải thiện dịch vụ khách hàng: Các chatbot và trợ lý ảo có thể sử dụng phân tích cảm xúc để hiểu cảm xúc của người dùng và đưa ra phản hồi phù hợp hơn, từ đó tạo ra một trải nghiệm người dùng tốt hơn.
* Phân tích nội dung truyền thông: Các công ty truyền thông có thể sử dụng phân tích cảm xúc để đánh giá phản ứng của khán giả đối với các chương trình truyền hình, phim ảnh, quảng cáo, v.v.
* Ứng dụng trong tài chính: Các nhà đầu tư có thể sử dụng phân tích cảm xúc để đánh giá các tin tức tài chính, báo cáo và các bài viết để nhận diện các xu hướng và dự báo sự biến động của thị trường chứng khoán.

**Lợi ích:**

* Cải thiện hiểu biết về khách hàng: Phân tích cảm xúc giúp doanh nghiệp hiểu rõ hơn về nhu cầu, thái độ và cảm xúc của khách hàng, từ đó cung cấp sản phẩm và dịch vụ phù hợp hơn.
* Ra quyết định nhanh chóng và chính xác: Phân tích cảm xúc có thể giúp các công ty, tổ chức đưa ra quyết định kịp thời và hiệu quả bằng cách nhận diện các tín hiệu và xu hướng từ dữ liệu cảm xúc.
* Tăng cường mối quan hệ khách hàng: Bằng cách hiểu được cảm xúc của khách hàng, doanh nghiệp có thể cải thiện các chiến lược giao tiếp, giải quyết các vấn đề và gia tăng sự hài lòng của khách hàng.
* Dự báo xu hướng và sự thay đổi trong thị trường: Phân tích cảm xúc có thể giúp dự đoán các sự kiện hoặc thay đổi trong hành vi tiêu dùng, ví dụ như sự chuyển biến trong thái độ đối với sản phẩm hoặc dịch vụ nào đó.
* Tự động hóa và tiết kiệm thời gian: Phân tích cảm xúc tự động hóa quá trình xử lý các phản hồi và ý kiến của khách hàng, giúp tiết kiệm thời gian và tài nguyên cho các tổ chức.
* Cải thiện chiến lược marketing: Hiểu rõ cảm xúc của khách hàng giúp tối ưu hóa chiến lược marketing, từ việc chọn lựa thông điệp quảng cáo phù hợp đến việc xác định các nhóm đối tượng tiềm năng.

**4.Kết luận**

Qua quá trình thực hiện đề tài, chúng em đã đạt được những kết quả đáng khích lệ, đồng thời cũng nhận thấy một số hạn chế cần khắc phục và những hướng nghiên cứu mở rộng trong tương lai.

Đầu tiên, việc áp dụng các phương pháp xử lý ngôn ngữ tự nhiên (NLP) đã giúp chúng em xử lý và phân loại cảm xúc của dữ liệu văn bản tiếng Việt một cách hiệu quả. Dữ liệu được phân tích, xử lý và gán nhãn (0: tiêu cực, 1: trung lập, 2: tích cực) đã phản ánh rõ ràng các cảm xúc cơ bản trong ngôn ngữ tiếng Việt. Với việc áp dụng các thuật toán học máy như mô hình Logistic Regression, Random Forest, SVM và chúng em đã đạt được độ chính xác khoảng 87%. Điều này cho thấy tiềm năng của các mô hình trong việc hỗ trợ phân tích cảm xúc trong thực tế.

Tuy nhiên, trong quá trình thực hiện, chúng em cũng đối mặt với một số thách thức. Dữ liệu tiếng Việt có đặc thù riêng với cấu trúc ngữ pháp phức tạp, từ vựng phong phú và nhiều sắc thái biểu cảm, dẫn đến việc xử lý dữ liệu và xây dựng mô hình trở nên khó khăn hơn so với một số ngôn ngữ khác. Bên cạnh đó, tập dữ liệu ban đầu có kích thước hạn chế và chưa thể bao quát hết các trường hợp đặc thù, làm giảm hiệu quả của mô hình khi áp dụng cho các tập dữ liệu thực tế đa dạng hơn.

Trong tương lai, chúng em dự kiến mở rộng tập dữ liệu, bao gồm nhiều ngữ cảnh và lĩnh vực khác nhau để tăng tính đa dạng và tính bao quát. Đồng thời, việc tích hợp các kỹ thuật tiên tiến như mạng neural sâu (Deep Learning) hoặc sử dụng các mô hình ngôn ngữ lớn (Large Language Models) như BERT hoặc GPT cho tiếng Việt có thể giúp cải thiện đáng kể hiệu suất của mô hình phân tích cảm xúc.

Nhóm chúng em tin rằng, những kết quả đạt được trong đề tài không chỉ có ý nghĩa về mặt học thuật mà còn có tiềm năng ứng dụng cao trong các lĩnh vực như chăm sóc khách hàng, marketing, và quản lý mạng xã hội. Đây là động lực để chúng em tiếp tục nghiên cứu và phát triển sâu hơn trong lĩnh vực này.

Cuối cùng, chúng em xin chân thành cảm ơn thầy Hồ Hướng Thiên đã hướng dẫn và tạo điều kiện để chúng em hoàn thành đề tài này.

**Định hướng tương lai**

Định hướng tương lai của bài phân tích cảm xúc có thể được hiểu theo nhiều cách khác nhau, phụ thuộc vào mục tiêu nghiên cứu và ứng dụng cụ thể. Tuy nhiên, một số định hướng chính có thể được đề cập bao gồm:

* **Ứng dụng trong công nghệ trí tuệ nhân tạo (AI) và học máy (Machine Learning)**: Phân tích cảm xúc ngày càng trở thành một phần quan trọng trong việc phát triển các hệ thống AI, giúp các máy móc hiểu được cảm xúc của người dùng để tương tác tốt hơn. Trong tương lai, khả năng phân tích cảm xúc sẽ tiếp tục được cải thiện, từ nhận diện cảm xúc trong văn bản, giọng nói, đến hình ảnh và video.
* **Phân tích cảm xúc trong marketing và chăm sóc khách hàng**: Các doanh nghiệp đang sử dụng phân tích cảm xúc để hiểu sâu hơn về cảm nhận của khách hàng đối với sản phẩm, dịch vụ, hay thương hiệu của họ. Trong tương lai, việc tích hợp phân tích cảm xúc vào các chiến dịch marketing và dịch vụ khách hàng sẽ ngày càng phổ biến, giúp tăng cường trải nghiệm người dùng và tối ưu hóa các chiến lược.
* **Phân tích cảm xúc trong tâm lý học và xã hội học**: Nghiên cứu về cảm xúc và hành vi con người luôn là một lĩnh vực quan trọng trong tâm lý học và xã hội học. Phân tích cảm xúc có thể giúp chúng ta hiểu rõ hơn về các yếu tố tác động đến hành vi và quyết định của con người, từ đó xây dựng các phương pháp can thiệp, hỗ trợ tâm lý hiệu quả hơn.
* **Phân tích cảm xúc trong giáo dục và phát triển cá nhân**: Trong lĩnh vực giáo dục, phân tích cảm xúc có thể giúp giáo viên hiểu rõ hơn về tình trạng tâm lý của học sinh, từ đó điều chỉnh phương pháp giảng dạy sao cho phù hợp. Việc sử dụng công cụ phân tích cảm xúc trong phát triển cá nhân cũng giúp mọi người nhận diện và điều chỉnh cảm xúc của mình tốt hơn.
* **Cải tiến thuật toán và công nghệ nhận diện cảm xúc**: Các thuật toán phân tích cảm xúc, đặc biệt trong lĩnh vực học sâu (Deep Learning), vẫn còn rất nhiều tiềm năng để cải thiện. Những nghiên cứu mới sẽ giúp các hệ thống trở nên chính xác hơn trong việc nhận diện và phân loại cảm xúc trong nhiều ngữ cảnh khác nhau, từ lời nói đến các dấu hiệu phi ngôn ngữ.

Tóm lại, tương lai của phân tích cảm xúc sẽ không chỉ dừng lại ở việc phát triển các công cụ phân tích thông minh mà còn mở rộng ra nhiều lĩnh vực khác nhau như marketing, giáo dục, tâm lý học, và công nghệ AI, mang lại những thay đổi tích cực cho xã hội.