**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: NHẬP MÔN HỌC MÁY**

**ĐỀ TÀI: PHÂN LOẠI CẢM XÚC SINH VIÊN**

**Giáo viên hướng dẫn: ThS. Trần Thu Trang**

**Sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Mã sinh viên** | **Họ và tên** | **Lớp** |
| **1** | **1771020618** | **Nguyễn Gia Thái** | **PTPM 17-12** |
| **2** | **1771020651** | **Lê Đức Thọ** | **PTPM 17-12** |

**Hà Nội, năm 2025**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: NHẬP MÔN HỌC MÁY**

**ĐỀ TÀI: PHÂN LOẠI CẢM XÚC SINH VIÊN**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Mã Sinh Viên | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 | 1771020049 | Nguyễn Gia Thái | 17/02/2005 |  |  |
| 2 | 1771020651 | Lê Đức Thọ | 26/01/2005 |  |  |

### 

**CÁN BỘ CHẤM THI**

**Hà Nội, năm 2025**

**LỜI NÓI ĐẦU**

Trong bối cảnh chuyển đổi số và sự phát triển mạnh mẽ của công nghệ trí tuệ nhân tạo, việc ứng dụng các thuật toán học máy (Machine Learning) vào nhiều lĩnh vực khác nhau đang ngày càng trở nên phổ biến. Một trong những hướng nghiên cứu nhận được nhiều sự quan tâm trong những năm gần đây là **phân tích và phân loại cảm xúc**, bởi cảm xúc đóng vai trò quan trọng trong quá trình học tập, giao tiếp và phát triển kỹ năng xã hội của mỗi cá nhân.

Đối với môi trường giáo dục, việc hiểu rõ cảm xúc của sinh viên có ý nghĩa đặc biệt quan trọng. Nó không chỉ giúp giảng viên nắm bắt được tâm lý và mức độ tiếp thu của người học, mà còn là cơ sở để điều chỉnh phương pháp giảng dạy, cải thiện chất lượng đào tạo và xây dựng môi trường học tập thân thiện, tích cực. Chính vì vậy, đề tài *“Phân loại cảm xúc sinh viên bằng các thuật toán học máy”* được lựa chọn với mong muốn góp phần mang lại một góc nhìn mới mẻ và thực tiễn trong việc ứng dụng công nghệ vào lĩnh vực giáo dục.

Trong quá trình thực hiện đề tài, nhóm tập trung nghiên cứu các thuật toán phân loại truyền thống có độ chính xác cao, đồng thời tiến hành phân tích, so sánh hiệu suất nhằm tìm ra mô hình phù hợp nhất với bộ dữ liệu cảm xúc của sinh viên. Kết quả nghiên cứu không chỉ mang ý nghĩa học thuật mà còn có khả năng ứng dụng trong thực tiễn, đặc biệt trong việc hỗ trợ các hệ thống quản lý học tập thông minh (Smart Learning Systems) và các nền tảng đánh giá trực tuyến.

Mặc dù đã có nhiều nỗ lực và cố gắng, song do hạn chế về kiến thức và kinh nghiệm, báo cáo khó tránh khỏi những thiếu sót. Nhóm rất mong nhận được sự đóng góp ý kiến của quý thầy cô và bạn bè để hoàn thiện hơn.

Mục lục

[**LỜI NÓI ĐẦU** 3](#_Toc209708268)

[**DANH MỤC CÁC TỪ VIẾT TẮT** 7](#_Toc209708269)

[**CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 8**](#_Toc209708270)

[**1.1. Giới thiệu đề tài 8**](#_Toc209708271)

[***1.1.1. Bối cảnh chung 8***](#_Toc209708272)

[***1.1.3. Ý nghĩa và mục tiêu của đề tài 8***](#_Toc209708273)

[***1.1.2. Vấn đề đặt ra 9***](#_Toc209708274)

[***1.1.4. Phạm vi nghiên cứu 9***](#_Toc209708275)

[***1.1.5. Giá trị thực tiễn 10***](#_Toc209708276)

[**1.2.Mô tả dữ liệu 10**](#_Toc209708277)

[***1.2.1. Nguồn dữ liệu 10***](#_Toc209708278)

[***1.2.2. Cấu trúc dữ liệu 10***](#_Toc209708279)

[***1.2.3. Quy mô dữ liệu 10***](#_Toc209708280)

[***1.2.4. Đặc điểm nội dung 11***](#_Toc209708281)

[***1.2.5. Nhận xét sơ bộ 11***](#_Toc209708282)

[**1.3.Trực quan hóa dữ liệu 11**](#_Toc209708283)

[***1.3.1. Phân phối nhãn 11***](#_Toc209708284)

[***1.3.2. Độ dài tin nhắn 12***](#_Toc209708285)

[***1.3.3. Tần suất từ ngữ 13***](#_Toc209708286)

[***1.3.4. Nhận xét 14***](#_Toc209708287)

[**CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 21**](#_Toc209708288)

[**2.1. Tiền xử lý văn bản (Text Preprocessing) 21**](#_Toc209708289)

[**2.2. Biểu diễn văn bản (Feature Extraction) 21**](#_Toc209708290)

[**2.3. Các thuật toán phân loại văn bản 22**](#_Toc209708291)

[**2.4. Các chỉ số đánh giá mô hình 25**](#_Toc209708292)

[**CHƯƠNG 3. THỰC NGHIỆM VÀ KẾT QUẢ 26**](#_Toc209708293)

[**3.1 Làm sạch và Chuẩn hóa dữ liệu 26**](#_Toc209708294)

[***3.1.1. Kiểm tra cấu trúc dữ liệu ban đầu 26***](#_Toc209708295)

[***3.1.2. Phân tích thống kê cơ bản 26***](#_Toc209708296)

[***3.1.3. Các bước tiền xử lý văn bản 26***](#_Toc209708297)

[***3.1.4. Đặc trưng bổ sung 27***](#_Toc209708298)

[***3.1.5. Kết luận bước tiền xử lý 27***](#_Toc209708299)

[**3.2. Chạy các thuật toán 28**](#_Toc209708300)

[***3.2.1. Thiết lập chung của thí nghiệm 28***](#_Toc209708301)

[***3.2.2. Danh sách thuật toán & tham số mặc định 29***](#_Toc209708302)

[***3.2.3. Quy trình huấn luyện & đánh giá (chi tiết từng bước) 30***](#_Toc209708303)

[***3.2.4. GridSearch / tham số gợi ý (ví dụ) 32***](#_Toc209708304)

[***3.2.5. Chạy thực nghiệm — tổ hợp khuyến nghị 32***](#_Toc209708305)

[***3.2.6. Mẫu bảng kết quả (bạn chèn số thực) 32***](#_Toc209708306)

[***3.2.7. Phân tích chi tiết theo thuật toán — mẫu văn bản để chèn 33***](#_Toc209708307)

[***3.2.8. Kiểm soát mất cân bằng & tuning ngưỡng 36***](#_Toc209708308)

[***3.2.9. Phân tích lỗi (tóm tắt để chèn) 37***](#_Toc209708309)

[***3.2.10. Kết luận tóm tắt phần II 37***](#_Toc209708310)

[**3.3. So sánh kết quả giữa các thuật toán 38**](#_Toc209708311)

[**KẾT LUẬN** 41](#_Toc209708312)

[**TÀI LIỆU THAM KHẢO 42**](#_Toc209708313)

**DANH MỤC CÁC TỪ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
| **STT** | **TỪ VIẾT TẮT** | **VIẾT ĐẦY ĐỦ** |
| 1 | DBMS | Database Management System |
| 2 | LINQ | Language Integrated Query |
| 3 |  |  |
| 4 |  |  |

# CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

## Giới thiệu đề tài

### 1.1.1. Bối cảnh chung

Trong bối cảnh chuyển đổi số và phát triển giáo dục đại học, **phản hồi từ sinh viên** là một nguồn dữ liệu vô giá để các cơ sở giáo dục cải thiện chất lượng giảng dạy, chương trình đào tạo và cơ sở vật chất. Tuy nhiên, việc xử lý và tổng hợp hàng ngàn nhận xét dưới dạng văn bản (ý kiến về giảng viên, môn học, môi trường học tập) theo phương pháp thủ công là bất khả thi.

A person holding a clipboard with a pen and smiley faces

AI-generated content may be incorrect.

### 1.1.3. Ý nghĩa và mục tiêu của đề tài

- **Ý nghĩa:** Đề tài giúp nhà trường có cái nhìn khách quan và tức thì về tâm lý, mức độ hài lòng của sinh viên, từ đó đưa ra các quyết định điều chỉnh kịp thời và hiệu quả.

**- Mục tiêu:** Xây dựng mô hình học máy phân loại cảm xúc cho văn bản tiếng Việt với ba lớp: **Positive (Tích cực)**, **Negative (Tiêu cực)**, và **Neutral (Trung lập)**, đạt hiệu suất **F1-score Macro** cao nhất khi so sánh các tổ hợp kỹ thuật đặc trưng (BoW, TF-IDF, Embeddings) và mô hình phân loại (LR, SVC).

### 1.1.2. Vấn đề đặt ra

Làm thế nào để tự động hóa quá trình phân tích các nhận xét này để xác định chính xác **mức độ cảm xúc** (Tích cực, Tiêu cực, Trung lập) của sinh viên đối với từng khía cạnh trong trường?

A person sitting at a desk with a clipboard and a pen

AI-generated content may be incorrect.

- Giải pháp Hướng Sử dụng underthesea để Tách từ chính xác, tiền xử lý và xây dựng các bộ đặc trưng hiệu quả . Thử nghiệm Sentence Embeddings để mã hóa ngữ nghĩa toàn bộ câu, khắc phục nhược điểm của BoW/TF-IDF. Triển khai và so sánh 6 tổ hợp (BoW/TF-IDF/Embeddings kết hợp với LR/SVC) để tìm ra tổ hợp có F1-score Macro cao nhất. Sử dụng các chỉ số Precision, Recall, F1-score Macro và phân tích Ma trận Nhầm lẫn để đánh giá hiệu suất đa chiều.

### 1.1.4. Phạm vi nghiên cứu

* **Dữ liệu**: Tập trung vào dữ liệu phản hồi sinh viên giả lập, bao gồm các nhận xét liên quan đến Giảng viên (lecturer), Chương trình học (curriculum), Cơ sở vật chất (facility), và các vấn đề khác (others).
* **Kỹ thuật:** Áp dụng các kỹ thuật Xử lý ngôn ngữ tự nhiên (NLP) cho tiếng Việt và thử nghiệm các mô hình học máy truyền thống (Logistic Regression, LinearSVC) trên các phương pháp biểu diễn văn bản khác nhau.

### 1.1.5. Giá trị thực tiễn

Kết quả nghiên cứu có thể áp dụng vào:

* **Đối với Nhà trường:** Cung cấp hệ thống tự động lọc và báo cáo cảm xúc, giúp Ban Giám hiệu nắm bắt nhanh chóng các vấn đề nóng (hot spots) cần cải thiện.
* **Đối với Giảng viên:** Nhận được phản hồi có cấu trúc về điểm mạnh và điểm yếu trong phương pháp giảng dạy.
* **Đối với Nghiên cứu:** Cung cấp kết quả thực nghiệm về hiệu suất của các kỹ thuật vector hóa khác nhau (BoW, TF-IDF, Embeddings) trên dữ liệu cảm xúc tiếng Việt, làm cơ sở cho các nghiên cứu tiếp theo.

## 1.2.Mô tả dữ liệu

### 1.2.1. Nguồn dữ liệu

Dữ liệu được sử dụng là bộ dữ liệu phản hồi sinh viên **tổng hợp (synthetic)**, mô phỏng các nhận xét thực tế tại Việt Nam.

### 1.2.2. Cấu trúc dữ liệu

Dữ liệu bao gồm các cột chính:

* sentence: Nội dung phản hồi (văn bản thô).
* sentiment: Nhãn cảm xúc (**positive, negative, neutral**).
* topic: Chủ đề của nhận xét (ví dụ: lecturer, curriculum, facility).

### 1.2.3. Quy mô dữ liệu

### A screenshot of a computer AI-generated content may be incorrect.

### 1.2.4. Đặc điểm nội dung

### - Nội dung của dữ liệu bao gồm các câu nhận xét về môi trường học tập, được phân loại rõ ràng theo cả Cảm xúc (Sentiment) và Chủ đề (Topic).

### A screenshot of a computer AI-generated content may be incorrect.

### 1.2.5. Nhận xét sơ bộ

Dữ liệu được xây dựng rất cân bằng giữa ba lớp cảm xúc (xấp xỉ 33.3% mỗi lớp), điều này là một lợi thế lớn, giúp mô hình tập trung vào việc học đặc trưng văn bản hơn là phải đối phó với vấn đề mất cân bằng lớp. Sự phân loại kép theo Chủ đề và Cảm xúc cũng cho phép phân tích chuyên sâu hơn về chất lượng mô hình sau này..

## 1.3.Trực quan hóa dữ liệu

Để hiểu rõ hơn về đặc điểm của bộ dữ liệu tin nhắn SMS, nhóm tiến hành trực quan hóa một số khía cạnh quan trọng như phân phối nhãn, độ dài tin nhắn, và tần suất từ ngữ xuất hiện.

### 1.3.1. Phân phối nhãn Cảm Xúc

### - Phân phối nhãn là yếu tố quan trọng nhất quyết định chiến lược huấn luyện mô hình. Sự cân bằng giữa các lớp là lý tưởng, còn mất cân bằng đòi hỏi các kỹ thuật xử lý đặc biệt (như SMOTE, thay đổi ngưỡng phân loại).A screenshot of a graph AI-generated content may be incorrect.

Dữ liệu cho thấy sự **cân bằng gần như hoàn hảo** giữa ba lớp cảm xúc (Positive, Negative, Neutral). Điều này cho phép chúng ta sử dụng **Accuracy** và **F1-score Macro** làm thước đo chính mà không cần quá lo lắng về việc mô hình sẽ bị thiên vị (bias) sang một lớp nào đó.

A graph of positive and negative

AI-generated content may be incorrect.

*Hình 1.1: Biểu đồ phân phối nhãn cảm xúc.*

### 1.3.2. Phân phối Tỷ lệ Cảm xúc theo chủ đề

### - Phân tích này giúp chúng ta nhận định liệu một chủ đề cụ thể (như Giảng viên, Cơ sở vật chất) có xu hướng gây ra cảm xúc tích cực hay tiêu cực hơn. Đây là dữ liệu nền tảng cho việc can thiệp cải tiến chất lượng.

### Phân tích: (Sau khi chạy code, bạn cần điền phân tích cụ thể, ví dụ:)

### Chủ đề lecturer (Giảng viên) có tỷ lệ cảm xúc Positive cao nhất, cho thấy sinh viên thường có cái nhìn hài lòng về chất lượng giảng dạy.

### Chủ đề facility (Cơ sở vật chất) có tỷ lệ Negative nhỉnh hơn các chủ đề khác, gợi ý rằng đây là lĩnh vực cần ưu tiên cải thiện.

### Chủ đề curriculum (Chương trình học) có tỷ lệ Neutral cao, thể hiện sự đánh giá trung lập hoặc mang tính mô tả.

**A graph of different colored squares

AI-generated content may be incorrect.**

*Hình 1.2: Histogram tỷ lệ cảm xúc theo chủ đề*

### 1.3.3. Tần suất từ ngữ (Word Cloud)

### Word Cloud là kỹ thuật trực quan hóa tần suất từ, giúp nhanh chóng nhận diện các đặc trưng ngôn ngữ (linguistic features) quan trọng nhất cho từng lớp cảm xúc. Các từ khóa này chính là nền tảng để các kỹ thuật BoW và TF-IDF xây dựng vector đặc trưng.

### Phân tích:

### Word Cloud Positive: Các từ nổi bật như *tuyệt vời, nhiệt tình, chuyên nghiệp, hài lòng* xác nhận mô hình sẽ dễ dàng nhận diện các từ khóa chỉ sự đánh giá cao.

### Word Cloud Negative: Sự xuất hiện dày đặc của các từ như *kém, khó, không tốt, tù túng, thiếu* là các dấu hiệu tiêu cực mạnh mẽ, giúp mô hình phân loại chính xác các nhận xét phàn nàn.

A close up of words

AI-generated content may be incorrect.A close up of words

AI-generated content may be incorrect.

### 1.3.4. Nhận xét

- Quy trình Xử lý Tiếng Việt: Các bước Tiền xử lý là bắt buộc đối với dữ liệu phản hồi sinh viên tiếng Việt, đặc biệt là việc Tách từ chính xác.

- Chiến lược Biểu diễn Đa dạng: Việc thử nghiệm cả ba kỹ thuật biểu diễn (BoW, TF-IDF, và Sentence Embeddings) sẽ cho phép đánh giá toàn diện về sự khác biệt giữa các phương pháp dựa trên tần suất (Count-based) và phương pháp dựa trên ngữ nghĩa (Semantic-based).

- Lựa chọn Mô hình Phân loại: Các thuật toán tuyến tính (Logistic Regression và LinearSVC) được chọn vì chúng hiệu quả và mạnh mẽ khi xử lý các vector đặc trưng chiều cao, rất phù hợp cho giai đoạn so sánh ban đầu.

- Thước đo Đánh giá: Chỉ số F1-score Macro là thước đo chính, đảm bảo mô hình không chỉ đạt độ chính xác chung mà còn có khả năng phân loại tốt cả ba lớp cảm xúc cân bằng.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## 2.1. Phân loại và Tiền xử lý văn bản (Text Preprocessing)

**- Phân loại văn bản** : quá trình gán nhãn hoặc danh mục cho một đoạn văn bản dựa trên nội dung của nó. Trong đề tài này, bài toán được định nghĩa là Phân loại Cảm xúc (Sentiment Analysis) ba lớp: Tích cực (Positive), Tiêu cực (Negative), và Trung lập (Neutral). Đây là một bài toán học máy có giám sát (Supervised Learning), nơi mô hình được huấn luyện trên dữ liệu đã có nhãn.

**- Tiền xử lý** : bước bắt buộc để chuyển đổi văn bản thô, nhiễu thành định dạng sạch, có cấu trúc mà mô hình học máy có thể xử lý. Đối với tiếng Việt, bước này đặc biệt quan trọng:

* **Tách từ** : Do tiếng Việt là ngôn ngữ đơn lập, nhiều từ mang ý nghĩa cần được xử lý như một đơn vị thống nhất (ví dụ: "phương pháp giảng dạy", "cơ sở vật chất"). Kỹ thuật Tách từ được sử dụng để phân tách câu thành các đơn vị có nghĩa, giải quyết vấn đề từ ghép.
* **Loại bỏ Stopwords và chuẩn hóa** :

**Stopwords:** Các từ phổ biến như *và, của, là, thì* được loại bỏ vì chúng hầu như không mang thông tin cảm xúc, giúp giảm chiều dữ liệu.

**Chuẩn hóa:** Bao gồm chuyển chữ thường (lowercasing) để đảm bảo các từ được thống nhất, bất kể chúng xuất hiện ở đầu câu hay giữa câu.

## 2.2. Biểu diễn văn bản (Feature Extraction)

Để mô hình học máy có thể hoạt động, văn bản phải được mã hóa thành các **vector số (Numeric Vectors)**. Đề tài này so sánh ba phương pháp biểu diễn chính:

**(a) Bag-of-Words (BoW)**

**A diagram of a bag and bag

AI-generated content may be incorrect.**

BoW biểu diễn văn bản dưới dạng một vector đếm số lần xuất hiện của mỗi từ trong văn bản đó so với toàn bộ từ điển (vocabulary).

* **Ưu điểm:** Đơn giản, dễ tính toán.
* **Hạn chế:** Không quan tâm đến thứ tự từ và không phản ánh tầm quan trọng tương đối của từ.

**(b) TF-IDF (Term Frequency - Inverse Document Frequency)**

**A diagram of a diagram

AI-generated content may be incorrect.**

T F-IDF là một cải tiến của BoW, gán trọng số cho mỗi từ dựa trên hai yếu tố:

TFIDF(t,d)=TF(t,d)×IDF(t)

* TF(t,d) (Tần suất Từ): Số lần từ t xuất hiện trong văn bản d.
* IDF(t) (Tần suất Nghịch đảo Tài liệu): Mức độ hiếm của từ t trong toàn bộ tập dữ liệu (corpus).
* Ưu điểm: Giảm tầm quan trọng của các từ phổ biến (như Stopwords) và làm nổi bật các từ khóa có tính phân biệt cao (ví dụ: "tuyệt vời", "tồi tệ").

**(c) Sentence Embeddings (Word2Vec, BERT, Sentence Transformers)**

**A group of rectangular objects with numbers

AI-generated content may be incorrect.**

Đây là kỹ thuật hiện đại sử dụng các mô hình học sâu đã được huấn luyện trước (Pre-trained Models) để mã hóa toàn bộ câu thành một vector số **mật độ cao (Dense Vector)**.

* **Nguyên lý:** Vector này được đặt trong một không gian ngữ nghĩa, nơi các câu có ý nghĩa tương tự nhau sẽ có vector nằm gần nhau hơn.
* **Ưu điểm:** Bảo toàn được **ngữ cảnh và ý nghĩa tổng thể** của câu, vượt trội hơn BoW/TF-IDF trong việc xử lý các sắc thái ngôn ngữ phức tạp.

## 2.3. Các thuật toán phân loại văn bản

**- Logistic Regression (LR)**

LR là một mô hình tuyến tính, sử dụng hàm **Sigmoid** để ánh xạ đầu ra từ tổng trọng số thành xác suất. Mặc dù tên gọi là "Regression", nó được sử dụng rộng rãi cho các bài toán phân loại nhị phân và đa lớp.

* **Ưu điểm**: Dễ huấn luyện, dễ diễn giải, và hiệu quả với các vector thưa (BoW, TF-IDF).

**A diagram of a mathematical equation

AI-generated content may be incorrect.**

**- Linear Support Vector Classification (LinearSVC)**

LinearSVC là phiên bản tuyến tính của thuật toán Support Vector Machine (SVM). Nó hoạt động bằng cách tìm kiếm một **siêu phẳng (Hyperplane)** có biên độ (margin) lớn nhất để phân chia các lớp dữ liệu.

* **Ưu điểm:** Rất mạnh mẽ với dữ liệu chiều cao, thường mang lại độ chính xác cao và ít bị ảnh hưởng bởi vấn đề quá khớp (overfitting).

A diagram of a graph

AI-generated content may be incorrect.

## 2.4. Các chỉ số đánh giá mô hình

Để đánh giá 6 tổ hợp mô hình một cách khách quan, các chỉ số sau đây được sử dụng, dựa trên Ma trận Nhầm lẫn (Confusion Matrix):

A black background with white text

AI-generated content may be incorrect.

# CHƯƠNG 3. THỰC NGHIỆM VÀ KẾT QUẢ

## 3.1 Làm sạch và Chuẩn hóa dữ liệu

### 3.1.1. Kiểm tra cấu trúc dữ liệu ban đầu

Bộ dữ liệu được cung cấp dưới dạng file CSV với 3 cột chính:

* **sentence**: Chứa nội dung phản hồi của sinh viên.
* **sentiment**: Chứa nhãn phân loại (0 – Neutral, 1 – Negative, 2 – Positive).
* **topic**: Chứa chủ đề của phản hồi (ví dụ: lecturer, curriculum, facility).

Sau khi nạp dữ liệu, tiến hành kiểm tra kích thước và 5 dòng đầu tiên:

* **Kích thước tập huấn luyện:** N = **4000** mẫu.
* **Kích thước tập kiểm tra:** N = **1000** mẫu.
* **Số lượng nhãn:** Phân phối nhãn **rất cân bằng**, với tỷ lệ xấp xỉ 33.3% cho mỗi lớp positive, negative, và neutral.

Điều này gợi ý rằng khi huấn luyện mô hình cần chú ý đến vấn đề **class imbalance**.

### 3.1.2. Phân tích thống kê cơ bản

### Phân phối nhãn cảm xúc:

### Positive: 1340 mẫu (33.50%)

### Negative: 1332 mẫu (33.30%)

### Neutral: 1328 mẫu (33.20%)

### Độ dài nhận xét: Tính độ dài câu nhận xét (số từ) của mỗi phản hồi.

### Nhận xét dài thường chứa nhiều thông tin chi tiết về trải nghiệm hoặc đánh giá.

### Nhận xét ngắn gọn thường mang tính chất nhận định hoặc cảm thán.

### Biểu đồ trực quan:

### Countplot phân phối nhãn → minh họa tình trạng cân bằng dữ liệu.

### Barplot tỷ lệ cảm xúc theo chủ đề → giúp xác định xu hướng cảm xúc đối với từng khía cạnh (giảng viên, cơ sở vật chất, chương trình học).

### 3.1.3. Các bước tiền xử lý văn bản

Dữ liệu tiếng Việt được làm sạch và chuẩn hóa qua các bước:

* **Chuyển chữ thường (Lowercasing):** Tất cả các từ được chuyển về dạng chữ thường.
* **Tách từ (Tokenization):** Sử dụng thư viện **underthesea** để tách chính xác các từ ghép tiếng Việt.
* **Loại bỏ Stopwords:** Các từ chức năng (ví dụ: *là, của, và*) không mang ý nghĩa cảm xúc được loại bỏ.

### 3.1.4. Vector hóa Dữ liệu

Sau tiền xử lý, văn bản được chuyển thành các vector số:

* **Bag-of-Words (BoW):** Biểu diễn văn bản dựa trên tần suất từ.
* **TF-IDF (Term Frequency-Inverse Document Frequency):** Biểu diễn văn bản dựa trên trọng số từ, ưu tiên các từ có tính phân biệt cao.
* **Sentence Embeddings (SBERT):** Sử dụng mô hình ***paraphrase-multilingual-MiniLM-L12-v2*** để mã hóa toàn bộ câu thành vector ngữ nghĩa **384 chiều**.

### 3.1.5. Kết luận bước tiền xử lý

* Sau khi làm sạch và chuẩn hóa, dữ liệu đã sẵn sàng cho bước **vector hóa văn bản**.
* Các biểu đồ thống kê và wordcloud cung cấp bằng chứng rằng có sự khác biệt rõ ràng về **đặc trưng ngôn ngữ và độ dài** giữa Spam và Ham.

## 3.2. Chạy các thuật toán

### 3.2.1. Thiết lập chung của thí nghiệm

**A screen shot of a computer program

AI-generated content may be incorrect.**

Nhận xét**:** Đoạn code này thực hiện các bước chuẩn bị ban đầu bao gồm: định nghĩa danh sách VIETNAMESE\_STOPWORDS và hàm vietnamese\_tokenizer tùy chỉnh để xử lý văn bản tiếng Việt. Sau đó, nó tải dữ liệu huấn luyện và kiểm tra, sử dụng LabelEncoder để chuyển đổi các nhãn cảm xúc dạng chuỗi thành số nguyên, và cuối cùng phân tách dữ liệu thành các tập X(văn bản) và y (nhãn) cho quá trình huấn luyện và đánh giá.

**3.2.2. Quy trình Vector hóa Dữ liệu**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**A graph of blue and orange bars

AI-generated content may be incorrect.**

Nhận xét: Đoạn code này áp dụng ba kỹ thuật vector hóa khác nhau. CountVectorizer và TfidfVectorizer tạo ra các vector BoW và TF-IDF bằng cách sử dụng vietnamese\_tokenizer đã định nghĩa và giới hạn 5000 đặc trưng, bao gồm cả bi-gram. Đối với Sentence Embeddings, mô hình paraphrase-multilingual-MiniLM-L12-v2 được tải và sử dụng để mã hóa các câu, chuyển đổi văn bản thành các vector số có ngữ nghĩa.

**3.2.3. Danh sách Thuật toán & Tham số mặc định**

**A computer screen with text and numbers

AI-generated content may be incorrect.**

Nhận xét: Đoạn code này thiết lập các thuật toán phân loại và tập dữ liệu đầu vào. models chứa hai mô hình chính: Logistic Regression (LR) và LinearSVC (SVC), cùng với các tham số đã được tối ưu trước (ví dụ: class\_weight='balanced' cho LR để xử lý tiềm năng mất cân bằng lớp). vectorizers\_map ánh xạ tên của kỹ thuật vector hóa với các tập dữ liệu đặc trưng đã được tạo ở bước trước. Biến results được khởi tạo để thu thập hiệu suất của tất cả các tổ hợp.

**3.2.4. Quy trình Huấn luyện và Đánh giá Mô hình**

**A screen shot of a computer program

AI-generated content may be incorrect.**

Nhận xét: Đây là vòng lặp chính thực hiện huấn luyện và đánh giá. Nó lặp qua từng kỹ thuật vector hóa và từng mô hình phân loại. Với mỗi tổ hợp, mô hình được huấn luyện trên dữ liệu huấn luyện, sau đó dự đoán trên tập kiểm tra. Các chỉ số hiệu suất như F1-score Macro, Accuracy, Precision và Recall (weighted) được tính toán và lưu vào danh sách results. Cuối cùng, results được chuyển đổi thành DataFrame report\_table để hiển thị kết quả một cách rõ ràng và sắp xếp theo F1-score Macro.

**3.2.5. Bảng Tổng hợp Kết quả**

**A screenshot of a graph

AI-generated content may be incorrect.**

**3.2.6. Phân tích Tổng quan Hiệu suất**

### A computer screen shot of text AI-generated content may be incorrect.

### Nhận xét chi tiết: Biểu đồ này trực quan hóa hiệu suất F1-score Macro của các tổ hợp mô hình. Nó củng cố nhận định về sự vượt trội của Sentence Embeddings so với BoW/TF-IDF, cũng như việc LinearSVC hoạt động hiệu quả hơn Logistic Regression trên bộ dữ liệu này. Tổ hợp Embeddings + SVC là lựa chọn tối ưu nhất.

### 3.2.7. Phân tích Chi tiết theo Thuật toán

### 3.2.7.1. Phân tích Mô hình Tốt nhất (Embeddings + SVC)

### A screenshot of a graph AI-generated content may be incorrect.

### A screen shot of a computer program AI-generated content may be incorrect.

### Nhận xét: Ma trận nhầm lẫn cho thấy mô hình hoạt động rất tốt với lớp positive (Recall 0.90) và có độ tin cậy cao khi dự đoán negative (Precision 0.89). Tuy nhiên, lớp neutral vẫn là thách thức lớn nhất, với nhiều trường hợp bị nhầm lẫn thành positive (34 trường hợp), do sự mơ hồ ngữ nghĩa giữa các nhận xét trung lập và tích cực nhẹ.

### 3.2.7.2. Phân tích các thuật toán khác

### Logistic Regression (LR):

### Hiệu suất: Nhìn chung kém hơn LinearSVC một chút trong mọi trường hợp vector hóa.

### Độ phức tạp: Đơn giản, dễ diễn giải, nhưng có thể không mạnh mẽ bằng SVC trong việc tìm biên phân cách tối ưu khi dữ liệu có nhiều chiều.

### Với Embeddings: Đạt F1-Macro 0.8649, vẫn rất tốt và có thể xem xét nếu cần một mô hình đơn giản hơn.

### LinearSVC (SVC):

### Hiệu suất: Vượt trội hơn Logistic Regression, đặc biệt khi kết hợp với Embeddings.

### Độ phức tạp: Mạnh mẽ hơn trong việc tìm kiếm siêu phẳng phân chia tốt nhất, đặc biệt hiệu quả với dữ liệu có chiều cao.

### Với BoW/TF-IDF: Mặc dù không bằng Embeddings, SVC vẫn cho kết quả tốt hơn LR trên BoW/TF-IDF.

### 3.2.8. Threshold tuning (Điều chỉnh ngưỡng)

### Mặc dù không được áp dụng chính thức trong báo cáo này do mục tiêu là tối ưu F1-Macro và dữ liệu đã cân bằng, việc điều chỉnh ngưỡng phân loại là một kỹ thuật quan trọng:

### Mục đích: Khi cần ưu tiên một chỉ số cụ thể (ví dụ: tăng Recall của lớp Negative để không bỏ sót phản hồi tiêu cực), có thể điều chỉnh ngưỡng phân loại của mô hình.

### Cách thực hiện: Phân tích đường cong Precision-Recall hoặc ROC (Receiver Operating Characteristic) để tìm ngưỡng tối ưu cho từng lớp.

### 3.2.10. Ma trận nhầm lẫn

### A blue squares with white text AI-generated content may be incorrect.

### 3.2.10. So sánh các phương pháp (tổng hợp)

### Dưới đây là bảng so sánh tổng hợp hiệu suất của các phương pháp:

### A screenshot of a graph AI-generated content may be incorrect.

### Nhận xét tổng quát:

### Các kỹ thuật dựa trên tần suất (BoW/TF-IDF): Hiệu quả ở mức cơ bản, dễ hiểu, nhanh. TF-IDF có phần nhỉnh hơn BoW nhờ khả năng gán trọng số từ.

### Kỹ thuật Embeddings: Đem lại hiệu suất vượt trội, đặc biệt khi kết hợp với LinearSVC, chứng minh khả năng nắm bắt ngữ nghĩa sâu sắc.

**3.2.11. Kết luận**

Chương này đã trình bày quá trình thực nghiệm một cách chi tiết và có hệ thống. Kết quả cho thấy **Sentence Embeddings kết hợp với LinearSVC** là tổ hợp mô hình hiệu quả nhất cho bài toán phân loại cảm xúc phản hồi sinh viên tiếng Việt, với khả năng nắm bắt ngữ nghĩa và phân loại chính xác ở mức cao. Mặc dù vẫn còn một số thách thức trong việc phân biệt các sắc thái trung lập, hiệu suất tổng thể của mô hình đã đạt được mục tiêu đề ra.

## 3.3. So sánh kết quả giữa các thuật toán

**A black background with white text

AI-generated content may be incorrect.**

**3.3.1. So sánh hiệu quả giữa các phương pháp**

A computer screen with text

AI-generated content may be incorrect.

**Nhận xét chi tiết:**

1. **Hiệu suất của BoW:**
   * **Logistic Regression:** Đạt Accuracy khoảng 0.8170. Mặc dù đơn giản, BoW đã có thể phân loại cơ bản, nhưng bỏ qua ngữ cảnh và mối quan hệ giữa các từ.
   * **LinearSVC:** Nhỉnh hơn một chút với Accuracy 0.8250, cho thấy khả năng tìm biên phân cách tốt hơn ngay cả với biểu diễn tần suất từ thô.
   * Nhận xét chung: Cả hai mô hình đều có hiệu suất ở mức chấp nhận được nhưng hạn chế do thiếu ngữ cảnh.
2. **Hiệu suất của TF-IDF:**
   * **Logistic Regression:** Đạt Accuracy 0.8350. TF-IDF cải thiện hơn BoW nhờ việc gán trọng số, giúp các từ khóa quan trọng có ảnh hưởng lớn hơn.
   * **LinearSVC:** Đạt Accuracy 0.8420. Mô hình này tiếp tục thể hiện sự vượt trội, đặc biệt khi TF-IDF đã lọc bớt nhiễu và làm nổi bật các đặc trưng quan trọng.
   * Nhận xét chung: TF-IDF giúp tăng cường khả năng phân biệt của các thuật toán truyền thống.
3. **Hiệu suất của Sentence Embeddings:**
   * **Logistic Regression:** Đạt Accuracy 0.8650. Việc sử dụng Embeddings mang lại cải thiện đáng kể về hiệu suất, cho thấy khả năng nắm bắt ngữ nghĩa sâu sắc của Embeddings.
   * **LinearSVC:** Đạt Accuracy cao nhất 0.8710. Đây là sự kết hợp tối ưu nhất, nơi Embeddings cung cấp biểu diễn ngữ nghĩa phong phú và LinearSVC tận dụng hiệu quả không gian đặc trưng này để tìm ranh giới phân loại.
   * Nhận xét chung: Kỹ thuật Embeddings vượt trội hoàn toàn, mang lại hiệu suất phân loại cao nhất, khẳng định vai trò quan trọng của ngữ cảnh và ngữ nghĩa trong xử lý ngôn ngữ tự nhiên.

**KẾT LUẬN**

**1. Ưu điểm**

* Đa dạng phương pháp thử nghiệm: Đề tài đã thử nghiệm một loạt các phương pháp biểu diễn văn bản (Bag-of-Words, TF-IDF, Sentence Embeddings) và kết hợp với các thuật toán phân loại hiệu quả (Logistic Regression, LinearSVC). Điều này giúp có cái nhìn toàn diện về các giải pháp khả thi.
* Hiệu quả cao: Các mô hình dựa trên Sentence Embeddings, đặc biệt khi kết hợp với LinearSVC, đã đạt được hiệu suất cao với F1-score Macro là 0.8709. Mức độ chính xác này đủ tin cậy để áp dụng trong việc phân tích cảm xúc từ phản hồi sinh viên.
* Nắm bắt ngữ nghĩa: Việc sử dụng Sentence Embeddings đã chứng minh khả năng vượt trội trong việc nắm bắt ngữ cảnh và sắc thái ngữ nghĩa của tiếng Việt, điều mà các phương pháp truyền thống như BoW hay TF-IDF khó đạt được.
* Dễ triển khai: Các thuật toán học máy cổ điển như Logistic Regression và LinearSVC tương đối nhẹ, dễ huấn luyện và triển khai, phù hợp cho các hệ thống cần tốc độ xử lý nhanh.

**2. Nhược điểm**

* Thách thức với nhãn trung lập: Mặc dù đạt hiệu suất tốt, mô hình vẫn còn một số khó khăn trong việc phân biệt chính xác các phản hồi có sắc thái neutral, đặc biệt là nhầm lẫn chúng với positive do sự mơ hồ ngữ nghĩa trong một số trường hợp.
* Tài nguyên cho Embeddings: Việc tạo ra Sentence Embeddings có thể tốn nhiều tài nguyên hơn (RAM/CPU/GPU) và thời gian xử lý ban đầu so với BoW/TF-IDF, đặc biệt khi làm việc với tập dữ liệu rất lớn. Điều này cần được cân nhắc khi triển khai trên các thiết bị hoặc môi trường hạn chế.
* Khả năng diễn giải của LinearSVC: Tương tự như SVM, LinearSVC là mô hình "hộp đen" hơn Logistic Regression, khiến việc diễn giải lý do đằng sau một dự đoán cụ thể trở nên khó khăn hơn.
* Dữ liệu huấn luyện: Bộ dữ liệu phản hồi sinh viên tuy cân bằng nhưng kích thước còn khiêm tốn (5000 mẫu). Việc mở rộng tập dữ liệu sẽ giúp mô hình học được nhiều sắc thái ngôn ngữ và tổng quát hóa tốt hơn.

**3. Hướng phát triển**

1. **Mở rộng và làm giàu dữ liệu:**

Thu thập thêm phản hồi sinh viên từ nhiều trường, nhiều khóa học, đa dạng hóa các chủ đề và phong cách ngôn ngữ (bao gồm cả teencode, từ lóng).

Xem xét các phương pháp Data Augmentation (tăng cường dữ liệu) để mở rộng tập huấn luyện một cách nhân tạo.

1. **Thử nghiệm các mô hình học sâu (Deep Learning) mạnh mẽ hơn:**

Áp dụng các kiến trúc mạng nơ-ron như LSTM (Long Short-Term Memory) hoặc BiLSTM (Bidirectional LSTM) để tận dụng tốt hơn mối quan hệ tuần tự của từ trong câu.

Fine-tuning các mô hình Transformer đã huấn luyện trước cho tiếng Việt như PhoBERT hoặc các mô hình đa ngôn ngữ như BERT Multilingual để đạt hiệu suất cao hơn nữa.

1. **Kỹ thuật giải quyết sự mơ hồ của nhãn trung lập:**

Sử dụng các phương pháp phân tích sắc thái tinh vi hơn, có thể phân biệt giữa "không có cảm xúc" và "cảm xúc rất nhẹ".

Xem xét việc sử dụng Ensemble learning (kết hợp nhiều mô hình) hoặc hierarchical classification (phân loại phân cấp) để xử lý các lớp nhãn khó.

1. **Triển khai và tích hợp thực tế:**

Xây dựng API RESTful cho mô hình đã huấn luyện để dễ dàng tích hợp vào các hệ thống quản lý phản hồi sinh viên hoặc ứng dụng khác.

Tối ưu hóa mô hình Embeddings (ví dụ: bằng quantization hoặc distillation) để giảm kích thước và tài nguyên cần thiết, cho phép triển khai trên các thiết bị di động hoặc môi trường tài nguyên hạn chế.

1. **Giải thích kết quả (Explainable AI - XAI):**

Tích hợp các công cụ XAI như LIME (Local Interpretable Model-agnostic Explanations) hoặc SHAP (SHapley Additive exPlanations) để giải thích lý do mô hình dự đoán một phản hồi có cảm xúc cụ thể. Điều này giúp người dùng cuối tin tưởng và hiểu rõ hơn về hoạt động của mô hình.

# 

# TÀI LIỆU THAM KHẢO

[1] Alan Beaulieu, “Learning SQL: Generate, Manipulate, and Retrieve Data”, O’Reilly Media, 2020

[2] Adam Freeman, “Pro ASP.NET Core 7”, Manning, 2023.

[3] Mark J. Price, “C# 13 and .NET 9 Modern Cross-Platform Development Fundamentals: Start building websites and services with ASP.NET Core 9, Blazor, and EF Core 9”, Packt Publishing, 2024.

[4] RB Whitaker, “The C# Player’s Guide”, RB Whitaker, 2022.

[5] Phạm Văn Tiệp,***Thiết kế,lập trình Back-End***,Đại học Đại Nam,2025