A white rectangular frame with black border

AI-generated content may be incorrect.BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC THĂNG LONG**

A black and white screen

AI-generated content may be incorrect.

**BÁO CÁO BÀI TẬP LỚN  
MÔN: CÔNG CỤ LẬP TRÌNH TRÍ TUỆ NHÂN TẠO**

**XÂY DỰNG HỆ THỐNG  
PHÂN TÍCH CẢM XÚC VĂN BẢN TIẾNG VIỆT**

**GIÁO VIÊN HƯỚNG DẪN:** Ngô Mạnh Cường

**LỚP:** 243\_AI220\_01

**SINH VIÊN THỰC HIỆN:** Đỗ Tiến Đạt A47331

**HÀ NỘI – 2025**

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn chân thành nhất đến thầy Ngô Mạnh Cường. Trong quá trình học tập, nghiên cứu và xây dựng bài tập lớn **“Xây dựng hệ thống đánh giá bình luận tích cực tiêu cực”,** nhóm em đã nhận được rất nhiều sự quan tâm, góp ý, hướng dẫn tận tình của thầy. Thầy đã giúp nhóm tích lũy thêm nhiều kiến thức bổ ích để không chỉ hoàn thành được bài tập lớn, mà còn học được nhiều kinh nghiệm về lĩnh vực thị giác máy tính. Tất cả những kiến thức này đều rất hữu ích sau này cho công việc của chúng em.

Có lẽ kiến thức là vô hạn mà sự tiếp nhận kiến thức của bản thân mỗi người luôn tồn tại những hạn chế nhất định. Do đó, trong quá trình hoàn thành bài tập lớn, chắc chắn không tránh khỏi những thiếu sót. Nhóm em rất mong nhận được những góp ý đến từ thầy để bài tiểu luận của nhóm được hoàn thiện hơn.

Kính chúc thầy sức khỏe, hạnh phúc và thành công trên con đường sự nghiệp giảng dạy.

MỤC LỤC

[CHƯƠNG 1 - TỔNG QUAN VỀ ĐỀ TÀI 5](#_Toc198866231)

[1.1. LÝ DO CHỌN ĐỀ TÀI 5](#_Toc198866232)

[1.2. MỤC ĐÍCH, ĐỐI TƯỢNG SỬ DỤNG 5](#_Toc198866233)

[1.2.1. Mục đích nghiên cứu 5](#_Toc198866234)

[1.2.2. Đối tượng sử dụng 6](#_Toc198866235)

[1.3. PHƯƠNG PHÁP NGHIÊN CỨU 6](#_Toc198866236)

[1.3.1. Thu thập và tiền xử lý dữ liệu 6](#_Toc198866237)

[1.3.2. Nghiên cứu và lựa chọn các mô hình 7](#_Toc198866238)

[1.3.3. Đánh giá và lựa chọn mô hình 7](#_Toc198866239)

[CHƯƠNG 2 - Cơ Sở lý thuyết 8](#_Toc198866240)

[2.1. Giới thiệu 8](#_Toc198866241)

[2.1.1. Giới thiệu về học máy 8](#_Toc198866242)

[2.1.2. Bài toán phân tích cảm xúc văn bản tiếng Việt 8](#_Toc198866243)

[2.2. CÁC công nghệ chính 9](#_Toc198866244)

[2.2.1. Tiền xử lý ngôn ngữ tự nhiên 9](#_Toc198866245)

[2.2.2. Kỹ thuật trích xuất đặc trưng văn bản 9](#_Toc198866246)

[2.2.3. TF-IDF – Chuyển văn bản thành vector số 11](#_Toc198866247)

[2.2.4. Support Vector Machine– Mô hình phân loại 11](#_Toc198866248)

[2.2.5. Flask - Xây dựng giao thức API và server 12](#_Toc198866249)

[CHƯƠNG 3 - Phát triển hệ thống 14](#_Toc198866250)

[3.1. Tổng quan kiến trúc hệ thống 14](#_Toc198866251)

[3.2. Lý do chọn mô hình Linearsvc 14](#_Toc198866252)

[CHƯƠNG 4 - Thử nghiệm mô hình 17](#_Toc198866253)

[4.1. Mô Tả dữ liệu 17](#_Toc198866254)

[4.2. Kết quả mô hình 17](#_Toc198866255)

[CHƯƠNG 5 - Kết luận và hướng phát triển 18](#_Toc198866256)

[5.1. Kết luận 18](#_Toc198866257)

[5.1.1. Kết quả đạt được 18](#_Toc198866258)

[5.1.2. Hạn chế 18](#_Toc198866259)

[5.2. HƯớng phát triển trong tương lai 19](#_Toc198866260)

[DANH MỤC TÀI LIỆU THAM KHẢO 20](#_Toc198866261)

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **Viết tắt** | **Tiếng Anh** | **Tiếng Việt** |
| AI | Artificial Intelligence | Trí tuệ nhân tạo |
| NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên |
| NaN | Not a Number | Không phải là số |
| TF-IDF | term frequency – inverse document frequency | Tần suất thuật ngữ - Tần suất tài liệu ngược |
| SVM | Support Vector Machine | Thuật toán phân loại bằng phương pháp siêu phẳng tuyến tính |
| API | Application Programming  Interface | Giao diện lập trình ứng dụng |
| HTML | Hyper Text Markup Language | Ngôn ngữ đánh dấu siêu văn bản |
| KNN | K-Nearest Neighbors | Giải thuật k hàng xóm gần nhất |
| pkl | Pickle File | Tệp lưu ở định dạng Pickle Python |

DANH MỤC CÁC BẢNG

[Bảng 3.2.1. Bảng đánh giá các mô hình dựa trên tập test 14](#_Toc198865196)

DANH MỤC CÁC ẢNH VÀ HÌNH VẼ

[Ảnh 3.1.1. Kiến trúc mô hình hệ thống 14](#_Toc198866182)

[Ảnh 3.2.1. Ma trận nhầm lẫn của mô hình Logistic Regression 15](#_Toc198866183)

[Ảnh 3.2.2. Ma trận nhầm lẫn của mô hình LinearSVC 15](#_Toc198866184)

[Ảnh 3.2.3. Ma trận nhầm lẫn của mô hình Decision Tree 16](#_Toc198866185)

[Ảnh 3.2.4. Ma trận nhầm lẫn của mô hình Random Forest 16](#_Toc198866186)

[Ảnh 4.2.1. Ma trận nhầm lẫn sau dự đoán 17](#_Toc198866187)

# TỔNG QUAN VỀ ĐỀ TÀI

## LÝ DO CHỌN ĐỀ TÀI

Trong thời đại bùng nổ thông tin và mạng xã hội, mỗi ngày có hàng triệu dòng trạng thái, bình luận và đánh giá sản phẩm được tạo ra bởi người dùng trên toàn cầu. Những dữ liệu này không chỉ đơn thuần là văn bản, mà còn là nơi chứa đựng rất nhiều cảm xúc, quan điểm và thái độ của người viết. Tuy nhiên, việc phân tích thủ công một khối lượng dữ liệu khổng lồ như vậy là điều không khả thi.

Chính vì thế, **phân tích cảm xúc (Sentiment Analysis)** đã trở thành một công cụ quan trọng, giúp máy tính có thể tự động xác định và phân loại ý kiến của người dùng theo các chiều hướng **tích cực, tiêu cực**. Công nghệ này đang được ứng dụng rộng rãi trong nhiều lĩnh vực như **thương mại điện tử** (phân tích đánh giá sản phẩm), **mạng xã hội** (đo lường dư luận), hay **dịch vụ khách hàng** (hiểu rõ phản hồi của người dùng).

Không chỉ mang lại giá trị thực tiễn cao, Sentiment Analysis còn là một trong những bài toán tiêu biểu và đầy thách thức trong lĩnh vực **Xử lý ngôn ngữ tự nhiên (NLP)** và **Trí tuệ nhân tạo (AI)**. Vì vậy, em lựa chọn đề tài này nhằm tìm hiểu sâu hơn về cách thức hoạt động, quy trình xây dựng mô hình AI phân tích cảm xúc văn bản tiếng Việt, đồng thời khám phá tiềm năng ứng dụng của nó trong đời sống và kinh doanh hiện đại.

## MỤC ĐÍCH, ĐỐI TƯỢNG SỬ DỤNG

### ****Mục đích nghiên cứu****

* Xây dựng một mô hình có khả năng phân loại văn bản tiếng Việt theo cảm xúc tích cực hoặc tiêu cực một cách tự động và nhanh chóng.
* Áp dụng kiến thức về học máy vào một bài toán thực tế trong xử lý ngôn ngữ.
* Hỗ trợ doanh nghiệp, cá nhân trong việc phân tích đánh giá người dùng, dư luận hoặc nội dung mạng xã hội.

### Đối tượng sử dụng

* Doanh nghiệp thương mại điện tử: Tự động phân tích đánh giá sản phẩm của khách hàng.
* Các nhà phát triển ứng dụng AI/NLP: Nghiên cứu, cải tiến mô hình phân tích cảm xúc.
* Nhà nghiên cứu, sinh viên: Học tập, ứng dụng bài toán thực tế về AI trong xử lý văn bản.

## PHƯƠNG PHÁP NGHIÊN CỨU

Để tiến hành xây dựng đề tài, nhóm sẽ thực hiện các nghiên cứu dưới đây:

### Thu thập và tiền xử lý dữ liệu

Dữ liệu được thu thập từ một tập văn bản chứa các bình luận tiếng Việt, mỗi bình luận được gán nhãn cảm xúc (tích cực, tiêu cực hoặc trung lập). Để đảm bảo chất lượng đầu vào, các bước tiền xử lý được thực hiện bao gồm:

**Xử lý thiếu dữ liệu**: Loại bỏ các dòng bị thiếu giá trị (NaN) để tránh gây nhiễu cho mô hình.

**Làm sạch văn bản**:

* Chuẩn hóa văn bản về chữ thường (lowercase).
* Xóa các dấu câu, ký tự đặc biệt không cần thiết.
* Chuẩn hóa tiếng Việt bằng cách loại bỏ dấu
* Xóa các khoảng trắng liên tiếp.

**Trích xuất đặc trưng (feature engineering)**:

* Đặc trưng từ văn bản: Áp dụng kỹ thuật TF-IDF cho cả từ (word) và ký tự (character), với n-gram từ 1 đến 6 để nắm bắt cả ngữ cảnh ngắn và dài.
* Đặc trưng hình thức văn bản: Trích xuất các đặc trưng như số từ, số ký tự, tỷ lệ chữ hoa, dấu câu (!, ?), chữ thường, emoji… để bổ sung thông tin định lượng cho mô hình.

### Nghiên cứu và lựa chọn các mô hình

Sau khi dữ liệu được xử lý và trích xuất đặc trưng, tiến hành huấn luyện và đánh giá các mô hình học máy phổ biến như:

* **SVM (Support Vector Machine)**: Phù hợp với bài toán phân loại văn bản nhờ khả năng phân tách rõ ràng các lớp dữ liệu.
* **Random Forest**: Tập hợp nhiều cây quyết định nhằm giảm thiểu overfitting và tăng tính khái quát của mô hình.
* **Logistic Regression**: Mô hình tuyến tính đơn giản nhưng hiệu quả, thường dùng làm baseline.
* **Decision Tree**: Mô hình phân loại dễ hiểu, trực quan, tuy nhiên dễ bị overfit.
* **KNN (K-Nearest Neighbors)**: Dựa trên khoảng cách giữa các văn bản đã học và văn bản cần phân loại.

Tất cả mô hình đều được tích hợp trong pipeline cùng với các bước tiền xử lý và trích xuất đặc trưng. Thêm vào đó, các mô hình hỗ trợ trọng số lớp (class weight) như SVM, Random Forest, Logistic Regression và Decision Tree được cấu hình với class\_weight = ‘balance’ để xử lý mất cân bằng dữ liệu.

### Đánh giá và lựa chọn mô hình

Để lựa chọn mô hình tốt nhất cho bài toán cảm xúc văn bản, các mô hình học máy sau khi huấn luyện được đánh giá dựa trên tập kiểm tra với nhiều chỉ số hiệu suất khác nhau, bao gồm: Accuracy, F1-score trung bình của 2 lớp và tốc độ dự đoán trung bình một vector. Những chỉ số này phản ánh độ chính xác tổng thể, khả năng đánh giá tích cực tiêu cực, cũng như khả năng cân bằng giữa tỷ lệ phát hiện và tỷ lệ sai sót.

# Cơ Sở lý thuyết

## Giới thiệu

### Giới thiệu về học máy

Học máy (machine learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống “học” tự động từ dữ liệu để giải quyết những vấn đề cụ thể. Học máy nói đến một chương trình giúp cho máy tính có thể giải một bài toán cụ thể bằng những “kinh nghiệm” mà nó đã được “học” từ dữ liệu.

Ở các chương trình lập trình truyền thống, con người sẽ phải viết ra các quy tắc, điều kiện để máy tính thực hiện nhiệm vụ và đưa ra câu trả lời. Ngược lại, đối với Học máy, nhiệm vụ của máy tính là phải dựa vào dữ liệu quan sát để tìm ra được các quy tắc này.

Học máy đóng vai trò quan trọng trong nhiều ứng dụng thực tế như nhận dạng giọng nói (Siri, Google Assistant), hệ thống gợi ý (Netflix, Spotify), chẩn đoán y tế (phân tích hình ảnh y khoa), xe tự lái (Tesla, Waymo), phát hiện gian lận tài chính, xử lý hình ảnh (nhận diện khuôn mặt, biển số xe), và tối ưu hóa chuỗi cung ứng (Amazon, Walmart). Nó giúp hệ thống tự động học hỏi từ dữ liệu, tăng cường hiệu quả và chính xác trong các công việc hàng ngày, đồng thời tạo ra những ứng dụng mới thay đổi cuộc sống.

Trong dự án này, học máy được áp dụng để xây dựng hệ thống phân cảm xúc văn bản, giúp dễ dàng nhận diện được cảm xúc “tích cực” hoặc “tiêu cực” của văn bản được nhập vào.

### Bài toán phân tích cảm xúc văn bản tiếng Việt

Phân tích cảm xúc (Sentiment Analysis) là quá trình xác định cảm xúc chủ quan trong văn bản – thường là tích cực và tiêu cực. Với tiếng Việt, bài toán này gặp nhiều thách thức hơn so với tiếng Anh do:

* Ngữ pháp và cấu trúc câu linh hoạt.
* Sự đa dạng trong cách biểu đạt cảm xúc, từ trang trọng đến ngôn ngữ mạng.
* Thiếu tài nguyên xử lý ngôn ngữ như từ điển cảm xúc, dữ liệu gán nhãn lớn.

Do đó, việc xây dựng hệ thống phân tích cảm xúc tiếng Việt đòi hỏi quy trình tiền xử lý chuyên sâu và lựa chọn mô hình phù hợp.

## CÁC công nghệ chính

### Tiền xử lý ngôn ngữ tự nhiên

Để chuẩn hóa dữ liệu đầu vào, dự án áp dụng nhiều kỹ thuật tiền xử lý, bao gồm:

* Loại bỏ khoảng trắng dư thừa (RemoveConsecutiveSpaces)
* Chuyển về chữ thường (Lowercase)
* Loại bỏ dấu tiếng Việt (RemoveTone) để tăng tính khái quát
* Loại bỏ dấu câu (RemovePunct)

### Kỹ thuật trích xuất đặc trưng văn bản

Trong bài toán phân loại cảm xúc văn bản, đặc biệt với dữ liệu tiếng Việt có cấu trúc linh hoạt và giàu cảm xúc, việc thiết kế bộ đặc trưng đầu vào đóng vai trò then chốt trong việc đảm bảo hiệu quả học và khái quát của mô hình. Dự án đã lựa chọn kết hợp **hai nhóm đặc trưng chính**: (1) đặc trưng hình thức và (2) đặc trưng ngữ nghĩa-ngôn ngữ. Cách kết hợp này nhằm khai thác cả tín hiệu hình thức biểu cảm và nội dung từ ngữ mang ý nghĩa cảm xúc.

**Đặc trưng hình thức (Statistical / Surface-based Features)**

Đây là những đặc trưng được trích xuất từ hình thức hiển thị của câu, độc lập với nội dung từ vựng. Chúng phản ánh cách người viết thể hiện cảm xúc thông qua cấu trúc hình thức, vốn rất quan trọng trong các ngữ cảnh như mạng xã hội, e-mail hoặc các nền tảng bình luận. Cụ thể:

* **Số lượng từ và ký tự**: Đặc trưng này cho phép mô hình phân biệt các văn bản ngắn gọn (ví dụ: "Tệ quá!") với các văn bản dài hơn, nhiều khả năng diễn đạt đầy đủ cảm xúc hoặc đánh giá có chiều sâu.
* **Tỷ lệ chữ in hoa trên tổng ký tự**: Người viết thường sử dụng chữ in hoa để nhấn mạnh hoặc thể hiện sự tức giận (ví dụ: "QUÁ TỆ!"). Do đó, đặc trưng này phản ánh cường độ cảm xúc.
* **Tỷ lệ ký tự thường**: Bổ sung cho đặc trưng chữ in hoa để phân biệt các câu văn mang tính trung tính, khách quan hơn.
* **Tần suất xuất hiện dấu cảm thán (!) và dấu hỏi (?)**: Đây là chỉ báo mạnh cho cảm xúc tiêu cực (hoang mang, tức giận) hoặc tích cực (phấn khích, ngạc nhiên).
* **Tần suất dấu câu hoặc ký tự đặc biệt**: Một số bình luận chứa nhiều dấu “!!!”, “...”, hoặc các biểu tượng như “@”, “#” phản ánh các cách nhấn mạnh phi truyền thống.
* **Tỷ lệ emoji xuất hiện trên số từ**: Các biểu tượng cảm xúc như 😡, 😍, 😭 là chỉ báo trực tiếp của trạng thái cảm xúc, đặc biệt trong văn bản không chính thức. Việc đưa emoji vào đặc trưng giúp mô hình xử lý hiệu quả hơn các dữ liệu hiện đại.

Các đặc trưng này được chuẩn hóa bằng phương pháp **StandardScaler (không dùng trung bình)** để đảm bảo đồng nhất về thang đo, giảm ảnh hưởng của độ dài văn bản.

**Đặc trưng ngôn ngữ (Linguistic / Semantic Features)**

Các đặc trưng này khai thác trực tiếp vào nội dung từ ngữ của văn bản, đóng vai trò nhận diện và phân loại cảm xúc dựa trên nghĩa của từ/cụm từ. Trong hệ thống đề xuất, nhóm đặc trưng này được xây dựng qua các tầng TF-IDF đa cấp như sau:

* **TF-IDF theo từ (word n-grams)**: Với n-gram từ 1 đến 4, đặc trưng này cho phép mô hình nhận diện không chỉ các từ đơn mà còn cả các cụm từ gắn liền với cảm xúc như “rất tốt”, “không hài lòng”, “cực kỳ tệ hại” … Đây là nguồn tín hiệu chính về mặt ngữ nghĩa.
* **TF-IDF theo ký tự (character n-grams và character\_wb)**: Sử dụng n-gram ký tự từ 1 đến 6 giúp mô hình nhận diện được các kiểu viết không chuẩn (ví dụ: "tuyệtttt", "xấccc"), viết tắt, hoặc chơi chữ – những đặc điểm thường thấy trong ngôn ngữ mạng xã hội.
* **Kết hợp có dấu và không dấu**: Dữ liệu được xử lý ở cả hai dạng có dấu (chuẩn tiếng Việt) và không dấu (phổ biến khi gõ nhanh). Điều này giúp mô hình **tăng tính khái quát** và **giảm lỗi** khi gặp người dùng gõ không đúng chính tả hoặc cố tình bỏ dấu.

Việc áp dụng đồng thời **3 dạng TF-IDF (word-based, char-based và word-boundary char-based)**, trên cả dữ liệu **có dấu và không dấu**, giúp mô hình nắm bắt được ngữ nghĩa, cách thể hiện biến thể và dấu hiệu cảm xúc tiềm ẩn từ nhiều góc độ khác nhau.

### TF-IDF – Chuyển văn bản thành vector số

TF-IDF thực chất được phát triển dựa trên nền tảng của mô hình Bag of Words (BoW). Ở mô hình BoW, mỗi tài liệu được biểu diễn như một vector đếm từ: mỗi chiều tương ứng với một từ trong từ vựng, và giá trị tại mỗi chiều là số lần từ đó xuất hiện trong tài liệu.

TF (Term Frequency): đánh giá độ quan trọng của một từ trong chính tài liệu đó.

IDF (Inverse Document Frequency): đánh giá độ quan trọng của từ trên toàn bộ tập tài liệu – nghĩa là từ nào xuất hiện ở nhiều tài liệu thì bị giảm trọng số, vì nó không giúp phân biệt tài liệu này với tài liệu khác.

Trong đó:

* N là số lượng văn bản.
* DF(t) là số văn bản chứa từ t.

### Support Vector Machine– Mô hình phân loại

Support Vector Classifier (SVC) là một mô hình học máy thuộc họ Support Vector Machine (SVM), được sử dụng chủ yếu trong các bài toán phân loại. Mục tiêu của SVC là tìm ra một siêu phẳng (hyperplane) tối ưu để phân tách các điểm dữ liệu thuộc hai lớp khác nhau. Siêu phẳng tối ưu này được xác định sao cho khoảng cách (gọi là margin) từ siêu phẳng đến các điểm dữ liệu gần nhất của mỗi lớp là lớn nhất và bằng nhau. Những điểm nằm gần siêu phẳng nhất và có vai trò xác định ranh giới phân chia được gọi là các vector hỗ trợ (support vectors) – chính là lý do tên gọi của phương pháp. Trong trường hợp dữ liệu có thể phân tách tuyến tính, SVC tìm một đường thẳng (trong 2D), mặt phẳng (trong 3D), hoặc siêu phẳng (trong không gian nhiều chiều) để phân chia hai lớp. Tuy nhiên, trên thực tế, nhiều bài toán có dữ liệu phức tạp không thể phân tách tuyến tính. Khi đó, SVC sử dụng một kỹ thuật gọi là kernel trick để ánh xạ dữ liệu từ không gian đầu vào sang một không gian đặc trưng có chiều cao hơn. Trong không gian mới này, dữ liệu có thể được phân tách tuyến tính dễ dàng hơn.

Trong mô hình Support Vector Classifier (SVC), tham số C đóng vai trò điều chỉnh mức độ đánh đổi giữa độ chính xác trên tập huấn luyện và độ tổng quát hóa của mô hình. Nói cách khác, C kiểm soát mức phạt đối với các điểm dữ liệu bị phân loại sai.

* Khi C lớn: Mô hình sẽ cố gắng phân loại đúng tất cả các điểm dữ liệu trong tập huấn luyện, kể cả nếu phải chọn một siêu phẳng có margin nhỏ hơn. Điều này có thể dẫn đến hiện tượng overfitting, tức là mô hình hoạt động rất tốt trên dữ liệu huấn luyện nhưng kém hiệu quả trên dữ liệu mới.
* Khi C nhỏ: Mô hình cho phép một số điểm bị phân loại sai để có thể tìm được một siêu phẳng có margin rộng hơn, từ đó tăng khả năng tổng quát hóa trên tập kiểm tra. Tuy nhiên, nếu C quá nhỏ, mô hình có thể underfit (không học đủ thông tin từ dữ liệu).

Một số hàm kernel phổ biến trong SVC bao gồm: linear kernel (hạt nhân tuyến tính), polynomial kernel (đa thức), và RBF kernel (Radial Basis Function) – một trong những lựa chọn phổ biến nhất cho dữ liệu phi tuyến. Việc lựa chọn kernel phù hợp là yếu tố quan trọng quyết định đến hiệu quả phân loại của mô hình.

### Flask - Xây dựng giao thức API và server

Flask là một framework Python nhẹ, cho phép phát triển các ứng dụng web một cách nhanh chóng. Trong dự án, Flask được sử dụng để:

* Xây dựng giao diện người dùng đơn giản, nơi người dùng có thể nhập văn bản và nhận kết quả phân tích cảm xúc.
* Tích hợp mô hình đã huấn luyện (dạng .pkl) để đưa vào sử dụng thực tế thông qua API.

# Phát triển hệ thống

## Tổng quan kiến trúc hệ thống

Hệ thống phân tích cảm xúc văn bản được thiết kế theo mô hình **client-server**, trong đó có sự phối hợp chặt chẽ giữa **Frontend** và **Backend** để đảm bảo trải nghiệm người dùng mượt mà và chính xác.

Frontend đóng vai trò là cầu nối trực tiếp giữa người dùng và hệ thống. Đây là nơi người dùng nhập vào văn bản cần phân tích và nhận lại kết quả dự đoán cảm xúc. Frontend được xây dựng bằng **HTML** kết hợp với **Flask template engine** để tương tác với backend. Giao diện tối giản và thân thiện, đảm bảo phù hợp với người dùng phổ thông.

Backend là trái tim của hệ thống, thực hiện các chức năng cốt lõi. Backend được xây dựng bằng **Flask**, một framework web nhẹ và linh hoạt trong Python, dễ dàng tích hợp với mô hình học máy đã huấn luyện bằng Scikit-learn.

A diagram of a computer program

AI-generated content may be incorrect.

Ảnh .. Kiến trúc mô hình hệ thống

## Lý do chọn mô hình Linearsvc

| **Mô hình** | **Thời gian dự đoán trung bình 1 vector** | **Accuracy** | **F1-Score trung bình** |
| --- | --- | --- | --- |
| Logistic Regression | 0.015 giây | 0.89 | 0.89 |
| LinearSVC | 0.131 giây | 0.89 | 0.90 |
| Decision Tree | 0.025 giây | 0.81 | 0.81 |
| Random Forest | 0.010 giây | 0.88 | 0.88 |

Bảng .. Bảng đánh giá các mô hình dựa trên tập test

Sau đây là biểu đồ confusion matrix minh họa số lượng dự đoán đúng và sai của từng mô hình:

A chart with different colored squares

AI-generated content may be incorrect.

Ảnh .. Ma trận nhầm lẫn của mô hình Logistic Regression

A chart with numbers and labels

AI-generated content may be incorrect.

Ảnh .. Ma trận nhầm lẫn của mô hình LinearSVC

A diagram of a tree

AI-generated content may be incorrect.

Ảnh .. Ma trận nhầm lẫn của mô hình Decision Tree

A chart of different colors

AI-generated content may be incorrect.

Ảnh .. Ma trận nhầm lẫn của mô hình Random Forest

# Thử nghiệm mô hình

## Mô Tả dữ liệu

Tổng quan dữ liệu

* Sau khi chọn được mô hình tốt nhất, nhóm đã triển khai thực tế, sau khi thử nghiệm với 2574 văn bản thực tế với nội dung đa dạng. Mỗi văn bản có độ dài trung bình khoảng 3 đến 15 từ để đảm bảo việc dữ liệu được kiểm tra một cách đúng nhất
* Tập dữ liệu đã được phân loại sẵn tích cực (positive) và tiêu cực(negative) với tỷ lệ phân bố ở mức chấp nhận được để đảm bảo dữ liệu cân bằng và hiệu quả trong việc đánh giá mô hình.

## Kết quả mô hình

Sau khi nhóm sử dụng mô hình LinearSVC đã cài đặt để tiến hành đánh giá hiệu suất. Dưới đây đây là ma trận nhầm lẫn của kết quả dự đoán, có thể thấy mô hình dự đoán chính xác trên dữ liệu thực tế.

Mô tả kết quả:

* Accuracy tổng thể đạt: 90.2%
* F1-Score trung bình đạt: 90%

A chart with numbers and labels

AI-generated content may be incorrect.

Ảnh .. Ma trận nhầm lẫn sau dự đoán

# Kết luận và hướng phát triển

## Kết luận

### Kết quả đạt được

Trong quá trình thực hiện đề tài, chúng tôi đã xây dựng và đánh giá các mô hình học máy nhằm phân loại cảm xúc từ dữ liệu văn bản (ví dụ: phân tích sentiment). Các mô hình bao gồm SVM, Random Forest, Logistic Regression và Decision Tree đã được huấn luyện và đánh giá trên tập dữ liệu thực tế.

* Các mô hình SVM và Logistic Regression đạt được hiệu quả cao nhất với độ chính xác (accuracy) và F1-score đều ở mức trên 0.89, cho thấy khả năng phân loại cảm xúc khá chính xác.
* Mô hình Random Forest cũng cho kết quả tốt, tuy có phần thấp hơn một chút so với SVM và Logistic Regression.
* Mô hình Decision Tree có độ chính xác và hiệu quả thấp hơn, phù hợp với các bài toán yêu cầu giải thích mô hình rõ ràng hơn nhưng hiệu suất chưa cao.

Như vậy, các kết quả thu được cho thấy các kỹ thuật học máy truyền thống vẫn phát huy hiệu quả tốt trong bài toán phân loại cảm xúc từ văn bản tiếng Việt khi được kết hợp với các bước xử lý đặc thù như loại bỏ dấu câu, chuẩn hóa ngôn ngữ và trích xuất đặc trưng.

### Hạn chế

* Dữ liệu huấn luyện còn hạn chế về kích thước và độ đa dạng, có thể ảnh hưởng đến khả năng tổng quát của mô hình.
* Việc xử lý ngôn ngữ tiếng Việt phức tạp hơn do ngữ pháp và các dấu câu, từ ngữ đa dạng gây khó khăn cho bước tiền xử lý.
* Chưa áp dụng các mô hình deep learning tiên tiến như LSTM, Transformer, PhoBert vốn có tiềm năng nâng cao hiệu quả nhận dạng cảm xúc.

## HƯớng phát triển trong tương lai

* Mở rộng tập dữ liệu huấn luyện với nhiều nguồn dữ liệu khác nhau, đa dạng về thể loại và ngữ cảnh, giúp mô hình học được nhiều đặc trưng hơn.
* Ứng dụng các mô hình deep learning hiện đại như BiLSTM, CNN, Transformer (BERT, PhoBERT) để cải thiện khả năng hiểu ngữ cảnh và ngôn ngữ phức tạp.
* Tối ưu pipeline tiền xử lý dữ liệu, tăng cường kỹ thuật xử lý ngôn ngữ tự nhiên cho tiếng Việt như word segmentation, POS tagging, entity recognition.
* Xây dựng hệ thống phân tích cảm xúc thời gian thực và tích hợp vào các ứng dụng như chatbot, giám sát mạng xã hội.
* Nghiên cứu mở rộng sang các bài toán phân loại đa lớp hoặc phân tích sâu hơn như nhận dạng cảm xúc đa chiều, phát hiện xu hướng cảm xúc.

# DANH MỤC TÀI LIỆU THAM KHẢO

[1] **Manning, C. D., Raghavan, P., & Schütze, H.** (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[2] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.

[3] **Luong, M. T., Pham, H., & Manning, C. D.** (2015). Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1412-1421.

[4] **Pham, H., Nguyen, T. H., & Nguyen, H. V.** (2019). PhoBERT: Pre-trained language models for Vietnamese. arXiv preprint arXiv:2003.00744.

[5] **Nguyen, D. Q., Nguyen, T. H., & Nguyen, D. Q.** (2020). ViET: Vietnamese Emotional Text Dataset for Sentiment Analysis. Proceedings of the 28th International Conference on Computational Linguistics.

[6] **Vapnik, V. N.** (1998). Statistical Learning Theory. Wiley.

[7] **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É.** (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[8] **Jurafsky, D., & Martin, J. H.** (2021). Speech and Language Processing (3rd ed.). Draft available at <https://web.stanford.edu/~jurafsky/slp3/>