BÁO CÁO ĐỀ TÀI

Tỉnh/thành phố: Hà Nội

ĐĂNG KÝ THÔNG TIN ĐỀ TÀI BÁO CÁO

2	**ĐỀ TÀI**	**Phần mềm**	**SP tích hợp phần cứng**
		☒	☐
3	**Tên đề tài báo cáo**	Mô hình học máy dự đoán cảm xúc văn bản	
4	**Ngôn ngữ lập trình hoặc nền tảng**	Python	
5	**Cấu hình cài đặt**	Để hoàn thiện dự án, chúng em triển khai mô hình trên máy tính cá nhân và nền tảng Google Colab. Cụ thể về cấu hình máy tính cục bộ: **Phần cứng:** + CPU: Intel Xeon E3-1220 v3 @ 3.10GHz + GPU: NVDIA GeForce GTX 1050 Ti - 4GB VRAM + RAM: 16GB DDR3 + SSD: 256GB SATA3 **Phần mềm:** + OS: Windows 10 Pro 22H2 + IDE: Visual Studio Code 1.92.0 + Python 3.12.4 + Pandas 2.2.2, Seaborn 0.13.2, Matplotlib 3.9.1, Scikit-Learn 1.5.1. - Mã nguồn của dự án được xuất bản trên Github Repository của nhóm. - Các tệp Jupyter Notebook của dự án cũng được triển khai trên Google Colab với 2 mô hình (Tiếng Anh và Tiếng Việt).	

THÔNG TIN TÁC GIẢ (NHÓM TÁC GIẢ)

Số lượng thành viên	**04 người**
Danh sách thành viên: Họ tên	
Thành viên 1	Phạm Công Hoàng

Thành viên 2	Nguyễn Đức Dũng
Thành viên 3	Nguyễn Gia Minh
Thành viên 4	Nguyễn Danh Bảo
Giáo viên hoặc chuyên gia hướng dẫn	
Họ và tên:	
Chức vụ, đơn vị công tác:	
Chức vụ:	
Điện thoại:	
Email:	

GIỚI THIỆU VỀ SẢN PHẨM

1. Ý tưởng của sản phẩm

Ý tưởng của "Dự án học máy dự đoán cảm xúc văn bản" là xây dựng một mô hình có thể xác định cảm xúc (tích cực hoặc tiêu cực) trong các văn bản như bình luận, bài đánh giá, nhận xét,... Hệ thống này sẽ đọc các đánh giá, phân tích nội dung văn bản, và sử dụng các thuật toán học máy để dự đoán cảm xúc của người viết.

2. Giới thiệu tổng quan

Mô hình phân tích cảm xúc văn bản là ứng dụng của trí tuệ nhân tạo, sử dụng phân tích văn bản, ngôn ngữ học tính toán để xác định, trích xuất, định lượng và nghiên cứu các trạng thái cảm xúc và thông tin chủ quan một cách có hệ thống. Phân tích cảm xúc được áp dụng rộng rãi cho các tài liệu tiếng nói của khách hàng như đánh giá và phản hồi khảo sát, truyền thông trực tuyến và mạng xã hội, và tài liệu chăm sóc sức khỏe cho các ứng dụng từ tiếp thị đến dịch vụ khách hàng đến y học lâm sàng.

Mô hình chúng em tạo ra dựa vào tiền xử lý dữ liệu nạp vào, từ đó máy tính có thể phân tích các từ khóa và dự đoán trạng thái cảm xúc là tích cực hay tiêu cực. Mô hình mang lại lợi ích nhờ khả năng xử lý nhiều loại thông tin văn bản một cách chính xác. Miễn là phần mềm được đào tạo với đầy đủ các ví dụ, phân tích cảm xúc có thể dự đoán chính xác sắc thái cảm xúc của tin nhắn.

MÔ TẢ SẢN PHẨM

1. **Mô tả nền tảng phát triển của sản phẩm**
   - **Ngôn ngữ lập trình: Python**, ngôn ngữ chính được sử dụng cho xử lý dữ liệu, xây dựng mô hình học máy, và phát triển API. Python nổi tiếng với các thư viện phong phú và dễ sử dụng trong lĩnh vực học máy và xử lý ngôn ngữ tự nhiên.

- **Bộ dữ liệu sử dụng:**
  + **Tiếng Anh:** IMDb 50k movie reviews dataset
  + **Tiếng Việt:** Vietnamese Sentiment Analysis
- **Nền tảng để phát triển các ứng dụng:** Google Colab, viết tắt của Google Colaboratory, là một dịch vụ cung cấp môi trường Jupyter Notebook hoàn toàn trực tuyến. Nó cho phép người dùng tạo, chia sẻ và chỉnh sửa các tệp notebook một cách dễ dàng mà không cần cài đặt bất kỳ phần mềm nào.
- **Mô tả hoạt động của sản phẩm:** sản phẩm giúp phân loại cảm xúc là tích cực hay tiêu cực qua một đoạn văn bản cho trước.

2. Đánh giá

- **Ưu điểm:**
  + Độ chính xác của mô hình xử lý văn bản tiếng Anh khá cao
  + Có thể đánh giá các văn bản ở nhiều ngữ cảnh khác nhau
  + Sử dụng Logistic Regression và Naive Bayes giúp hiệu quả với nguồn dữ liệu nhỏ nhưng đem lại tính chính xác cao, tăng tốc độ tính toán, hiệu quả trong việc xử lý các lớp dữ liệu không cân bằng.
- **Nhược điểm:** Bên cạnh những tính năng và hiệu quả mà sản phẩm mang lại, vẫn còn tồn tại một số vấn đề:
  + Chưa xử lý được đa dạng các ngôn ngữ: Mô hình này chỉ phân loại được cảm xúc của văn bản tiếng Anh và tiếng Việt, thậm chí với văn bản tiếng Việt thì có độ chính xác khá thấp do hạn chế và số lượng dữ liệu nạp vào.
  + Một số văn bản dài, phức tạp có khả năng đánh giá sai
  + Chưa học được icon, tiếng lóng, viết tắt, teencode,...

3. Mở rộng

Sau khi phân tích mặt tốt và hạn chế của sản phẩm, chúng em sẽ có một số định hướng phát triển và cải tiến sản phẩm như sau:

- Cải thiện accuracy của văn bản tiếng Việt bằng cách tăng lượng dữ liệu nạp vào.
- Đào tạo mô hình hiểu được thêm nhiều thứ tiếng, một số kí hiệu đặc biệt,...
- Đào tạo mô hình có thể sử dụng được trong đa dạng lĩnh vực hơn.
- Nâng cấp mô hình có thể phân loại được tính khách quan và chủ quan của văn bản.

KẾT LUẬN

Đây là sơ bộ báo cáo giúp mọi người có thể nắm được sơ qua ý tưởng và cách triển khai của bọn em. Chi tiết báo cáo đã được trình bày ở file pdf đính kèm. Chúng em đã thực hiện các bước từ tiền xử lý dữ liệu đến xây dựng và xây dựng một số mô hình khác nhau để đem tới hiệu suất tối ưu. Quá trình tiền xử lý giúp tối ưu hóa dữ liệu đầu vào, đồng thời các kỹ thuật xây dựng mô hình khác như Logistic Regression và Naive Bayes đã được áp dụng để phân tích ngữ nghĩa và cảm xúc của văn bản. Hơn nữa, mô hình đạt được mức độ chính xác cao, cho phép nhận diện cảm xúc với độ tin cậy đáng kể. Tuy nhiên, một số thách thức vẫn tồn tại, như khả năng xử lý cảm xúc văn bản phức tạp còn hạn chế và thiếu đa dạng ngữ cảnh và ngôn ngữ. Những vấn đề này chỉ ra rằng vẫn còn nhiều điểm để có thể phát triển và nâng cấp mô hình.

Kết quả nghiên cứu không chỉ chứng minh tính khả thi của mô hình phân tích cảm xúc mà còn mở ra hướng phát triển cho các ứng dụng thực tiễn trong nhiều lĩnh vực khác. Nhìn chung, báo cáo này đóng góp một cái nhìn sâu sắc vào lĩnh vực phân tích cảm xúc văn bản và cung cấp nền tảng cho các nghiên cứu và ứng dụng tiếp theo trong tương lai.

TÀI LIỆU THAM KHẢO

- **WIKIPEDIA:** https://en.wikipedia.org/wiki/Sentimentanalysis
- **Logistic regression:** https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- **Naive Bayes:** https://scikit-learn.org/stable/modules/naivebayes.html
- **Nguồn dữ liệu:**
  + https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
  + https://www.kaggle.com/datasets/linhlpv/vietnamese-sentiment-analyst