HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

LƯU THANH TRÀ

NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT BẰNG MÔ HÌNH HỌC SÂU

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – NĂM 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

LƯU THANH TRÀ

NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN TIẾNG VIỆT BẰNG MÔ HÌNH HỌC SÂU

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:
TS. Quản Trọng Thế

HÀ NỘI – 2025

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong đề án tốt nghiệp là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Sinh viên thực hiện

Lưu Thanh Trà

MỤC LỤC

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
ML	Machine Learning	Học máy
DL	Deep Learning	Học sâu
BERT	Bidirectional Encoder Representations from Transformers	Biểu diễn hai chiều từ Transformer (BERT)
RoBERTa	Robustly Optimized BERT Approach	Phương pháp cải tiến tối ưu hóa mạnh mẽ từ mô hình BERT
PhoBERT	Vietnamese Pretrained BERT	BERT được huấn luyện cho tiếng Việt (VinAI)
LSTM	Long Short-Term Memory	Bộ nhớ ngắn dài hạn
Bi-LSTM	Bidirectional Long Short-Term Memory	LSTM hai chiều
SVM	Support Vector Machine	Máy vector hỗ trợ
BoW	Bag of Words	Mô hình túi từ
TF-IDF	Term Frequency – Inverse Document Frequency	Tần suất – Đảo nghịch tần suất tài liệu
RDR	RDR Segmenter	Bộ tách từ RDR
BPE	Byte Pair Encoding	Mã hóa ghép cặp byte
CLS	Classification Token	Token phân loại trong BERT
TPU	Tensor Processing Unit	Bộ xử lý tensor (phần cứng do Google phát triển)
GPU	Graphics Processing Unit	Bộ xử lý đồ họa
FP16	16-bit Floating Point	Số thực dấu phẩy động 16-bit
FP32	32-bit Floating Point	Số thực dấu phẩy động 32-bit
CSV	Comma-Separated Values	Dữ liệu phân tách bằng dấu phẩy
API	Application Programming Interface	Giao diện lập trình ứng dụng

Epoch	(Không viết tắt – đơn vị huấn luyện)	Một vòng lặp huấn luyện toàn bộ dữ liệu
F1-score	F1 Score	Trung bình điều hòa giữa độ chính xác và độ bao phủ
Precision	Precision	Độ chính xác (tỷ lệ dự đoán đúng trên tất cả dự đoán dương)
Recall	Recall	Độ bao phủ (tỷ lệ phát hiện đúng trên tất cả nhãn thực tế)
Accuracy	Accuracy	Độ chính xác tổng thể
Loss	Loss	Hàm mất mát
AdamW	Adaptive Moment Estimation with Weight Decay	Bộ tối ưu hóa Adam có triệt tiêu trọng số
FP	False Positive	Dự đoán sai dương tính
FN	False Negative	Dự đoán sai âm tính
TP	True Positive	Dự đoán đúng dương tính
TN	True Negative	Dự đoán đúng âm tính
UI	User Interface	Giao diện người dùng
Colab	Google Colaboratory	Môi trường notebook trực tuyến của Google
Kaggle	(Tên riêng – không viết tắt)	Nền tảng khoa học dữ liệu và máy học trực tuyến

DANH MỤC BẢNG

DANH MỤC HÌNH

I. MỞ ĐẦU

1. Lý do chọn đề tài:

Trong bối cảnh hiện đại, việc hiểu và phân tích cảm xúc của con người từ văn bản đang trở thành một nhu cầu cấp thiết trong nhiều lĩnh vực, đặc biệt là trong ngành giáo dục. Nhận diện cảm xúc không chỉ giúp cải thiện các dịch vụ hỗ trợ học tập mà còn nâng cao hiệu quả giao tiếp, hỗ trợ giáo viên trong việc đánh giá tâm lý học sinh, và cá nhân hóa nội dung giảng dạy. Tuy nhiên, các nghiên cứu hiện tại chủ yếu tập trung vào ngôn ngữ phổ biến như tiếng Anh, trong khi tiếng Việt – với đặc điểm ngữ pháp, cú pháp và sự phong phú trong ngữ nghĩa – chưa nhận được sự quan tâm xứng đáng.

PhoBERT – một mô hình ngôn ngữ tiền huấn luyện tối ưu hóa riêng cho tiếng Việt – được phát triển bởi VinAI và kế thừa kiến trúc RoBERTa, đã chứng minh hiệu quả cao trong nhiều tác vụ xử lý ngôn ngữ tự nhiên tiếng Việt. Khác với các mô hình BERT huấn luyện đa ngôn ngữ, PhoBERT được huấn luyện trên kho dữ liệu tiếng Việt quy mô lớn, giúp hiểu sâu sắc hơn các đặc trưng ngôn ngữ bản địa.

Trong nghiên cứu này, đề tài lựa chọn mô hình PhoBERT kết hợp với Linear Classifier để thực hiện nhận diện cảm xúc từ văn bản tiếng Việt. Mô hình hoạt động bằng cách sử dụng biểu diễn ngữ nghĩa do PhoBERT sinh ra cho từng câu đầu vào, sau đó đưa qua một hoặc vài lớp tuyến tính nhằm phân loại cảm xúc tương ứng. Cách tiếp cận này tuy đơn giản nhưng hiệu quả, đặc biệt khi kết hợp với khả năng biểu diễn ngữ cảnh mạnh mẽ của PhoBERT.

Trong ngành giáo dục, nhận diện cảm xúc từ văn bản sẽ mang lại nhiều lợi ích thực tiễn, chẳng hạn như:

Đánh giá ý kiến phản hồi: Phân tích cảm xúc trong phản hồi của học sinh, phụ huynh hoặc giáo viên giúp nhà trường hiểu rõ hơn về trải nghiệm của họ, từ đó cải thiện chất lượng dịch vụ giáo dục.

Hỗ trợ tâm lý học đường: Xác định các dấu hiệu cảm xúc tiêu cực từ các văn bản do học sinh viết (nhật ký, bài tập làm văn, tin nhắn, email) để kịp thời can thiệp và hỗ trợ.

Tăng cường học tập cá nhân hóa: Dựa trên cảm xúc của học sinh, giáo viên có thể điều chỉnh phương pháp giảng dạy phù hợp, tạo môi trường học tập tích cực hơn.

Việc áp dụng PhoBERT kết hợp Linear Classifier không chỉ mang lại hiệu quả trong việc phân loại cảm xúc văn bản tiếng Việt với chi phí tính toán thấp, mà còn mở ra khả năng triển khai thực tế trong môi trường giáo dục. Đây là một hướng tiếp cận phù hợp, vừa tận dụng được sức mạnh của mô hình tiền huấn luyện tiếng Việt, vừa đảm bảo tính khả thi và hiệu quả trong ứng dụng thực tiễn.

2. Tổng quan về vấn đề nghiên cứu:

Nhận diện cảm xúc trong văn bản (Sentiment Analysis) là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), tập trung vào việc xác định và phân loại cảm xúc, thái độ từ nội dung văn bản. Với sự phát triển nhanh chóng của công nghệ, việc hiểu cảm xúc từ văn bản đã được ứng dụng rộng rãi trong nhiều lĩnh vực như thương mại điện tử, mạng xã hội, và giáo dục. Đặc biệt, trong ngành giáo dục, việc nhận diện cảm xúc từ văn bản giúp cải thiện tương tác giữa học sinh và giáo viên, nâng cao chất lượng giảng dạy, và hỗ trợ học sinh về mặt tâm lý.

Tuy nhiên, nhận diện cảm xúc trong văn bản tiếng Việt đặt ra nhiều thách thức do đặc thù ngôn ngữ. Tiếng Việt là một ngôn ngữ đơn âm, ngữ nghĩa thường phụ thuộc mạnh mẽ vào ngữ cảnh và cách sử dụng từ vựng. Ngoài ra, văn phong tiếng Việt trong môi trường giáo dục cũng đa dạng, từ các bài luận văn, nhật ký, đến phản hồi ý kiến, đòi hỏi mô hình phải có khả năng xử lý linh hoạt và sâu sắc. Các nghiên cứu hiện tại về nhận diện cảm xúc chủ yếu tập trung vào ngôn ngữ Anh, trong khi tiếng Việt còn thiếu các hệ thống hiệu quả và chuyên biệt để giải quyết bài toán này.

Để khắc phục các hạn chế trên, mô hình PhoBERT đã được phát triển như một phiên bản tiền huấn luyện ngôn ngữ BERT dành riêng cho tiếng Việt. PhoBERT được huấn luyện trên tập dữ liệu tiếng Việt lớn và tận dụng kiến trúc Transformer hai chiều, giúp nắm bắt được ngữ cảnh hiệu quả hơn so với các phương pháp truyền thống. Trong nghiên cứu này, PhoBERT được sử dụng như một bộ trích xuất đặc trưng (feature extractor), kết hợp với một tầng phân loại tuyến tính (Linear Classifier) nhằm thực hiện nhiệm vụ phân loại cảm xúc. Việc sử dụng Linear Classifier giúp đơn giản hóa quá trình huấn luyện và triển khai

mô hình, đồng thời vẫn đạt được hiệu quả đáng kể nhờ vào chất lượng biểu diễn ngôn ngữ mà PhoBERT cung cấp.

Trong bối cảnh giáo dục, khả năng nhận diện cảm xúc hiệu quả sẽ mang lại những lợi ích to lớn, chẳng hạn như hỗ trợ giáo viên hiểu rõ tâm lý học sinh thông qua các bài viết hoặc phản hồi. Hơn nữa, các hệ thống này có thể được sử dụng để phân tích ý kiến phản hồi từ phụ huynh, học sinh nhằm cải thiện chất lượng giáo dục.

3. Mục đích nghiên cứu:

Nghiên cứu về nhận diện cảm xúc trong văn bản tiếng Việt nhằm giải quyết các vấn đề đặc thù của ngôn ngữ này và phát triển các phương pháp hiệu quả để phân tích cảm xúc một cách chính xác, đặc biệt trong lĩnh vực giáo dục. Đề tài tập trung vào việc xây dựng một mô hình học sâu dựa trên PhoBERT kết hợp với bộ phân loại tuyến tính (Linear Classifier), với kỳ vọng mang lại sự cải thiện đáng kể về độ chính xác và hiệu quả trong xử lý văn bản tiếng Việt.

Mục đích chính của nghiên cứu là phát triển một hệ thống nhận diện cảm xúc có khả năng:

Xác định cảm xúc trong văn bản tiếng Việt: Phân tích và nhận diện các cảm xúc như tích cực, tiêu cực, hoặc trung lập trong các văn bản thuộc môi trường giáo dục, từ đó hỗ trợ giáo viên và nhà quản lý giáo dục đưa ra các biện pháp phù hợp.

Tăng cường khả năng xử lý ngôn ngữ tự nhiên cho tiếng Việt: Bằng cách sử dụng PhoBERT – mô hình tiền huấn luyện dành riêng cho tiếng Việt – kết hợp với một tầng phân loại tuyến tính, nghiên cứu này tận dụng khả năng biểu diễn ngữ cảnh hai chiều mạnh mẽ của PhoBERT để phân loại cảm xúc chính xác hơn. Mô hình Linear giúp đơn giản hóa kiến trúc, tối ưu thời gian huấn luyện và vẫn giữ được hiệu quả cao trong bài toán phân loại.

Ứng dụng trong ngành giáo dục: Nghiên cứu không chỉ dừng lại ở mặt lý thuyết mà còn tập trung vào tính ứng dụng, cụ thể là phân tích các bài viết, phản hồi, hoặc nội dung liên quan đến học sinh và giáo viên, nhằm nâng cao chất lượng giảng dạy và quản lý giáo dục.

Ngoài ra, nghiên cứu này còn có mục tiêu thúc đẩy sự phát triển của các công nghệ xử lý ngôn ngữ tự nhiên (NLP) cho tiếng Việt, vốn chưa được quan tâm đầy đủ trong bối cảnh hiện nay. Bằng cách triển khai và đánh giá mô hình PhoBERT + Linear Classifier, đề tài hướng đến việc cung cấp một giải pháp khả thi, hiệu quả và dễ triển khai trong thực tế.

Trong lĩnh vực giáo dục, nghiên cứu này kỳ vọng tạo ra một công cụ hỗ trợ hiệu quả cho giáo viên và nhà trường trong việc phân tích cảm xúc của học sinh qua các văn bản. Điều này không chỉ giúp phát hiện sớm các vấn đề tâm lý hoặc cảm xúc tiêu cực mà còn góp phần xây dựng môi trường giáo dục lành mạnh, tích cực và đáp ứng được nhu cầu cá nhân hóa trong dạy học.

Với mục tiêu này, đề tài sẽ góp phần nâng cao hiệu quả ứng dụng công nghệ trong giáo dục, đồng thời mở ra nhiều hướng nghiên cứu tiếp theo trong lĩnh vực xử lý ngôn ngữ tự nhiên và nhận diện cảm xúc cho ngôn ngữ tiếng Việt.

4. Đối tượng và phạm vi nghiên cứu:

4.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là các văn bản tiếng Việt chứa đựng nội dung cảm xúc, tập trung chủ yếu vào các bài viết, phản hồi, và tài liệu trong lĩnh vực giáo dục. Cụ thể, nghiên cứu sẽ tập trung vào các loại văn bản sau:

Bài viết của học sinh: Bao gồm các bài tập làm văn, nhật ký học tập, hoặc các câu trả lời tự luận có yếu tố cảm xúc.

Phản hồi của giáo viên: Nhận xét, đánh giá hoặc ý kiến phản hồi về quá trình học tập và hành vi của học sinh.

Phản hồi từ phụ huynh: Ý kiến góp ý liên quan đến chất lượng giảng dạy và môi trường học tập của nhà trường.

Đối tượng nghiên cứu chính về mặt kỹ thuật là mô hình PhoBERT kết hợp với Linear Classifier, nhằm tăng cường khả năng nhận diện cảm xúc trong văn bản tiếng Việt. Mô hình này tận dụng biểu diễn ngôn ngữ hai chiều của PhoBERT để trích xuất đặc trưng văn bản, sau đó đưa qua tầng phân loại tuyến tính nhằm phân loại các cảm xúc thành tích cực, tiêu cực, hoặc trung lập. Mục tiêu là xây dựng một hệ thống phân loại cảm xúc đơn

giản, hiệu quả, dễ triển khai và có khả năng mở rộng ứng dụng thực tế.

4.2. Phạm vi nghiên cứu

4.2.1. Ngữ liệu

Tập dữ liệu nghiên cứu gồm các văn bản tiếng Việt thuộc lĩnh vực giáo dục, thu thập từ các nguồn như: bài tập học sinh, phản hồi từ giáo viên và phụ huynh, hoặc nội dung từ các diễn đàn giáo dục. Dữ liệu sẽ được tiền xử lý để loại bỏ nhiễu, chuẩn hóa ngôn ngữ và gán nhãn cảm xúc phục vụ cho việc huấn luyện và đánh giá mô hình.

4.2.2. Phạm vi ngôn ngữ

Nghiên cứu tập trung vào phản hồi bằng tiếng Việt của sinh viên – một ngôn ngữ có tính đặc thù cao về cú pháp, từ vựng và ngữ nghĩa.

Văn bản đầu vào đa dạng về độ dài: từ phản hồi ngắn (1–2 câu) cho đến các bài luận hoặc nhận xét dài hơn.

4.2.3. Kỹ thuật áp dụng

Kỹ thuật nền tảng sử dụng PhoBERT: Mô hình tiền huấn luyện dành riêng cho tiếng Việt, có khả năng biểu diễn ngữ nghĩa hai chiều hiệu quả.

Kết hợp Linear Classifier: Thay vì sử dụng các kiến trúc phức tạp như Bi-LSTM hay Attention, đề tài áp dụng một tầng phân loại tuyến tính để phân tích đầu ra của PhoBERT. Cách làm này vừa giữ được độ chính xác cao, vừa giảm tải tính toán và dễ triển khai thực tế.

4.2.4. Ứng dụng thực tiễn

Mô hình sau khi huấn luyện sẽ được tích hợp vào một ứng dụng hỗ trợ sinh viên phản hồi trực tuyến, giúp tự động phân tích cảm xúc trong nội dung phản hồi của sinh viên.

Ứng dụng sẽ hỗ trợ nhà trường và giáo viên trong việc phát hiện sớm các vấn đề cảm xúc thông qua phân tích ngôn ngữ viết của sinh viên, từ đó đề xuất các hướng tư vấn, hỗ trợ kịp thời.

Ngoài ra, hệ thống cũng có thể dùng để phân tích dữ liệu phản hồi theo thời gian, hỗ trợ cải tiến phương pháp giảng dạy và nâng cao chất lượng quản lý giáo dục.

5. Phương pháp nghiên cứu:

Để đạt được mục tiêu nghiên cứu đề ra, đề tài sử dụng kết hợp các phương pháp lý thuyết và thực nghiệm như sau:

5.1. Phương pháp lý thuyết

Nghiên cứu tài liệu: Khảo sát các tài liệu liên quan đến nhận diện cảm xúc trong văn bản, đặc biệt là những nghiên cứu áp dụng mô hình PhoBERT cho tiếng Việt. Phân tích các bài báo khoa học, luận văn, tài liệu kỹ thuật từ các nguồn uy tín và mã nguồn mở như Hugging Face, PyTorch, VNLPS,...Thu thập thông tin từ các nghiên cứu trước đây về ứng dụng học sâu trong xử lý ngôn ngữ tự nhiên (NLP), cũng như các đặc thù ngôn ngữ của tiếng Việt, như cú pháp linh hoạt, từ vựng đa nghĩa và văn phong đa dạng trong giáo dục.

5.2. Phương pháp thực nghiệm

5.2.1. Xây dựng dữ liệu

Thu thập dữ liệu văn bản tiếng Việt từ các nguồn thực tế như bài làm văn của học sinh, phản hồi giáo viên, ý kiến từ phụ huynh, hoặc dữ liệu công khai từ các diễn đàn giáo dục trực tuyến.

5.2.2. Tiền xử lý dữ liệu

Làm sạch văn bản (loại bỏ ký tự đặc biệt, lỗi chính tả, ký hiệu không cần thiết); Chuẩn hóa ngôn ngữ (viết tắt, kiểu gõ teencode…); Gán nhãn cảm xúc theo ba lớp: tích cực, tiêu cực, trung lập.

5.2.3. Xây dựng và huấn luyện mô hình

Phát triển mô hình học sâu sử dụng PhoBERT làm công cụ trích xuất đặc trưng ngữ nghĩa từ văn bản.

Kết nối đầu ra của PhoBERT với một tầng phân loại đơn giản – Linear Classifier – để thực hiện phân loại cảm xúc.

Tiến hành huấn luyện mô hình trên tập dữ liệu đã xử lý, sử dụng cross-entropy loss làm hàm mất mát và Adam optimizer để tối ưu hóa.

Tinh chỉnh các siêu tham số như learning rate, batch size, số epoch,... nhằm cải thiện độ chính xác.

5.2.4. Đánh giá và so sánh

Sử dụng các chỉ số sau để hỗ trợ phân tích đánh giá hiệu quả mô hình: Accuracy (độ chính xác tổng thể), Precision (độ chính xác theo từng lớp), Recall (khả năng phát hiện đầy đủ các nhãn cảm xúc), F1-score (trung bình hài hòa giữa Precision và Recall).

Sau khi có kết quả chỉ số độ chính xác, thực hiện so sánh mô hình PhoBERT + Linear Classifier với kết quả ác mô hình truyền thống như SVM, Naive Bayes sử dụng Bag-of-Words hoặc TF-IDF,…đã được nghiên cứu hoặc với phiên bản PhoBERT cơ bản không fine-tune (hoặc chỉ Linear từ embedding).

5.2.5. Thử nghiệm ứng dụng thực tiễn

Thực hiện tích hợp mô hình vào một ứng dụng thử nghiệm, cho phép sinh viên điền phản hồi (dưới dạng văn bản tự do) và hệ thống sẽ tự động phân tích cảm xúc. Sau đó thực hiện đánh giá khả năng ứng dụng mô hình trong môi trường thực tế như: Hỗ trợ giáo viên phát hiện học sinh có biểu hiện cảm xúc tiêu cực; Cung cấp tổng quan tâm lý học sinh theo từng lớp/học kỳ để phục vụ cải tiến phương pháp giảng dạy.

II. NỘI DUNG

CHƯƠNG 1: TỔNG QUAN VỀ NHẬN DIỆN CẢM XÚC TRONG VĂN BẢN

Đặt nền tảng lý thuyết và ngữ cảnh nghiên cứu bằng cách cung cấp một cái nhìn toàn diện về các chủ đề liên quan. Trước hết, chương giới thiệu khái niệm và đặc điểm của ngôn ngữ Tiếng Việt, nhằm xác định những yếu tố đặc thù có thể ảnh hưởng đến các phương pháp xử lý ngôn ngữ tự nhiên. Tiếp theo, phần tổng quan về nhận diện cảm xúc trong văn bản trình bày các phương pháp truyền thống và hiện đại, nhấn mạnh vai trò ngày càng tăng của học máy và học sâu. Đặc biệt, mô hình BERT - một bước đột phá trong xử lý ngôn ngữ tự nhiên - sẽ được giới thiệu chi tiết, tập trung vào khả năng hiểu ngữ cảnh và xử lý ngữ nghĩa.

1.1 Tổng quan về nghiên cứu

1.1.1 Giới thiệu bối cảnh nghiên cứu và tầm quan trọng của nhận diện cảm xúc trong văn bản.

Nhận diện cảm xúc trong văn bản đã trở thành một trong những lĩnh vực trọng tâm trong xử lý ngôn ngữ tự nhiên (NLP) nhờ khả năng mang lại giá trị lớn trong nhiều lĩnh vực.

Thương mại điện tử: Các doanh nghiệp sử dụng phân tích cảm xúc để hiểu phản hồi từ khách hàng. Những nhận xét mang cảm xúc tiêu cực có thể giúp phát hiện các vấn đề trong sản phẩm hoặc dịch vụ, trong khi cảm xúc tích cực giúp xác định yếu tố thu hút khách hàng.

Truyền thông xã hội: Nhận diện cảm xúc từ các bài đăng hoặc bình luận giúp theo dõi xu hướng xã hội, dự đoán tâm lý cộng đồng, và hỗ trợ chiến dịch truyền thông hiệu quả.

Giáo dục: Trong môi trường học đường, nhận diện cảm xúc từ văn bản như bài luận văn, phản hồi của học sinh, hoặc nhật ký học tập giúp giáo viên nắm bắt trạng thái tâm lý của học sinh. Việc này giúp tạo môi trường học tập cá nhân hóa, hỗ trợ tâm lý kịp thời và tăng cường chất lượng giảng dạy.

Chăm sóc sức khỏe tâm lý: Các hệ thống phân tích cảm xúc có thể phát hiện dấu hiệu căng thẳng hoặc trầm cảm từ tin nhắn hoặc bài đăng, giúp đưa ra cảnh báo sớm và hỗ trợ tư vấn tâm lý hiệu quả.

Với sự phát triển nhanh chóng của công nghệ học sâu, các mô hình tiên tiến như PhoBERT kết hợp Linear, GPT, đã nâng cao đáng kể độ chính xác trong phân tích cảm xúc. Trước đây, các phương pháp truyền thống dựa trên từ điển hoặc học máy gặp nhiều hạn chế về khả năng hiểu ngữ cảnh và xử lý ngữ nghĩa phức tạp. Ngày nay, các mô hình học sâu cho phép phân tích cảm xúc không chỉ dựa trên từ ngữ mà còn dựa trên toàn bộ ngữ cảnh văn bản.

Một ví dụ điển hình là trong thương mại điện tử, nhận xét "Sản phẩm giao nhanh, nhưng chất lượng không như mong đợi" có thể chứa cả yếu tố tích cực ("giao nhanh") và tiêu cực ("chất lượng không như mong đợi"). Các mô hình tiên tiến như BERT giúp phân tích và tách biệt các khía cạnh cảm xúc này một cách chính xác, hỗ trợ đưa ra phản hồi phù hợp.

1.1.2 Tổng quan về văn bản Tiếng Việt

Tiếng Việt là một ngôn ngữ đơn âm, có nhiều đặc điểm đặc thù ảnh hưởng đến việc xử lý ngôn ngữ tự nhiên (NLP). Một số đặc điểm và thách thức chính bao gồm:

Phụ thuộc vào ngữ cảnh: Tiếng Việt có nhiều từ đồng âm khác nghĩa, khiến cho việc phân tích ngữ nghĩa phụ thuộc lớn vào ngữ cảnh sử dụng. Ví dụ, từ "tôi" có thể chỉ bản thân hoặc chỉ hành động "tôi cơm" tùy thuộc vào ngữ cảnh.

Ngữ pháp linh hoạt: Trật tự từ trong câu tiếng Việt thường rất linh hoạt, và việc thay đổi vị trí các từ có thể không làm thay đổi nghĩa của câu. Ví dụ, câu "Tôi ăn cơm" và "Ăn cơm tôi" vẫn giữ nghĩa tương tự, điều này đặt ra thách thức lớn cho các mô hình NLP.

Hệ thống dấu thanh: Tiếng Việt sử dụng hệ thống dấu thanh phong phú, trong đó một từ có thể mang nhiều nghĩa khác nhau chỉ với sự thay đổi của dấu thanh. Ví dụ: "ma," "mà," "má," và "mã" đều là những từ khác biệt.

Sự đa dạng trong biểu đạt cảm xúc: Văn bản tiếng Việt thường sử dụng nhiều thành ngữ, tục ngữ, hoặc cách diễn đạt ẩn dụ để truyền tải cảm xúc. Ví dụ, câu "Buồn như mất sổ

gạo" không chỉ thể hiện sự buồn bã mà còn mang tính văn hóa, khó có thể giải mã chính xác nếu không có kiến thức văn hóa liên quan.

1.1.3 Ứng dụng và thách thức

Ứng dụng trong nhận diện cảm xúc: Các hệ thống xử lý văn bản tiếng Việt đang được áp dụng để phân tích cảm xúc trong mạng xã hội, đánh giá sản phẩm, và hỗ trợ giáo dục. Ví dụ, nhận diện cảm xúc từ phản hồi học sinh giúp giáo viên cải thiện phương pháp giảng dạy.

Thách thức trong xử lý ngôn ngữ tự nhiên: Việc thiếu dữ liệu chất lượng cao bằng tiếng Việt là một rào cản lớn. Các văn bản tiếng Việt cũng thường không tuân thủ chặt chẽ các quy tắc ngữ pháp, đặc biệt là trên mạng xã hội, khiến việc phân tích trở nên phức tạp.

Ví dụ thực tế

Câu "Tôi không thích" có thể mang nghĩa tiêu cực hoặc trung tính, tùy thuộc vào từ đi kèm hoặc ngữ cảnh cụ thể.

Trong mạng xã hội, các từ viết tắt hoặc sai chính tả, như "k bt nx" (không biết nữa), cũng đặt ra thách thức lớn cho các hệ thống NLP.

1.2 Khái niệm cảm xúc trong văn bản và nhận diện/phân tích cảm xúc

Cảm xúc trong văn bản có thể được hiểu là các trạng thái tâm lý hoặc phản ứng cảm xúc của con người, được biểu đạt qua ngôn từ. Thông thường, cảm xúc được chia thành ba loại chính:

Tích cực: Thể hiện sự hài lòng, hạnh phúc hoặc đồng thuận, ví dụ: "Tôi rất thích sản phẩm này."

Tiêu cực: Thể hiện sự không hài lòng, tức giận hoặc buồn bã, ví dụ: "Chất lượng quá tệ, tôi rất thất vọng."

Trung lập: Không nghiêng về tích cực hay tiêu cực, chỉ đưa ra thông tin, ví dụ: "Sản phẩm này giao hàng đúng hạn."

1.2.1 Nhận diện cảm xúc (Sentiment Analysis)

Nhận diện cảm xúc là quá trình sử dụng các phương pháp xử lý ngôn ngữ tự nhiên

và học máy để xác định và phân loại cảm xúc từ nội dung văn bản. Quá trình này bao gồm:

Thu thập dữ liệu: Từ các nguồn như đánh giá sản phẩm, bài đăng trên mạng xã hội, bài luận văn, hoặc các phản hồi ý kiến.

Tiền xử lý dữ liệu: Gồm loại bỏ nhiễu, chuẩn hóa văn bản (như sửa lỗi chính tả, chuẩn hóa viết tắt), và gắn nhãn cảm xúc.

Áp dụng mô hình học máy hoặc học sâu: Các mô hình như Logistic Regression, Random Forest, hoặc các mạng thần kinh như LSTM, BERT được sử dụng để dự đoán cảm xúc dựa trên văn bản đầu vào.

1.2.2 Các mức độ phân tích cảm xúc

Phân tích cảm xúc ở cấp độ từ/cụm từ: Xác định cảm xúc tại các từ hoặc cụm từ cụ thể trong văn bản.

P*hân tích cảm xúc ở cấp độ câu:* Phân tích ý nghĩa cảm xúc của từng câu riêng lẻ.

Phân tích cảm xúc ở cấp độ tài liệu: Đánh giá cảm xúc tổng thể của toàn bộ văn bản.

1.2.3 Thách thức trong nhận diện cảm xúc

Một trong những thách thức lớn trong nhận diện cảm xúc tiếng Việt là tính đa nghĩa của từ ngữ, khi một từ hoặc cụm từ có thể biểu đạt những cảm xúc khác nhau tùy thuộc vào ngữ cảnh cụ thể. Bên cạnh đó, yếu tố văn hóa và ngôn ngữ cũng ảnh hưởng đáng kể, do các biểu đạt cảm xúc trong tiếng Việt thường mang tính bản địa và hàm ý sâu, gây khó khăn cho các mô hình xử lý ngôn ngữ không được thiết kế riêng cho tiếng Việt. Ngoài ra, sự không đồng nhất của dữ liệu cũng là vấn đề đáng kể, đặc biệt với các văn bản được thu thập từ mạng xã hội hoặc môi trường phi học thuật, thường xuyên xuất hiện lỗi chính tả, viết tắt, và ngôn ngữ không chuẩn, làm giảm độ chính xác của mô hình nhận diện cảm xúc.

Ví dụ ứng dụng: Trong giáo dục, việc nhận diện các từ như "mệt mỏi," "khó khăn" trong nhật ký học tập của học sinh có thể giúp giáo viên hiểu và hỗ trợ kịp thời. Trong thương mại, các cụm từ như "giá cả hợp lý" hoặc "không đáng tiền" cho phép doanh nghiệp cải thiện chiến lược kinh doanh dựa trên phản hồi khách hàng.

1.3 Các công trình nghiên cứu liên quan

1.3.1. Nghiên cứu quốc tế

Bo Pang & Lillian Lee (2004): A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts (*Cảm xúc trong giáo dục: Phân tích cảm xúc sử dụng tóm tắt chủ quan dựa trên phương pháp graph-based minimum-cut)*.

Xuất bản trong *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)* tại Barcelona, tháng 7 năm 2004, nghiên cứu đề xuất phương pháp tách câu chủ quan bằng graph-based minimum-cut để chỉ giữ lại những câu chứa nội dung cảm xúc và loại bỏ các câu khách quan. Sau đó, phần trích xuất này sẽ được đưa vào một bộ phân loại như Naive Bayes hoặc SVM để phân loại (cảm xúc tích cực/tiêu cực).

Khi chỉ sử dụng khoảng 60% các câu chủ quan, mô hình đạt độ chính xác lên đến 86.4%, so với khoảng 82.8% khi dùng toàn bộ văn bản, hoặc tương đương khi dùng toàn văn với 87.2% bằng SVM.

Kỹ thuật tách nội dung hữu ích về cảm xúc này có thể được tham khảo để xây dựng bước tiền xử lý dữ liệu: giúp mô hình PhoBERT kết hợp Linear chỉ tập trung vào các phần văn bản chứa biểu đạt cảm xúc, giảm nhiễu và tăng hiệu quả phân loại.

1.3.2. Nghiên cứu trong nước

Nguyen Quoc Dat & Nguyen Anh Tuan (2020): PhoBERT: Pre-trained language models for Vietnamese (PhoBERT: Mô hình ngôn ngữ tiền huấn luyện dành cho tiếng Việt)

Được đăng trong *Findings of the Association for Computational Linguistics: EMNLP 2020*, trang 1037–1042, xuất bản tháng 11 năm 2020, nghiên cứu tìm hiểu về mô hình tiền huấn luyện dành riêng cho tiếng Việt, có hai phiên bản PhoBERT-base và PhoBERT-large, được huấn luyện trên 20 GB văn bản tiếng Việt. PhoBERT đạt hiệu suất vượt trội hơn BERT đa ngôn ngữ (Multilingual BERT) và các biểu diễn truyền thống như fastText, đặc biệt trong các tác vụ như phân loại cảm xúc, nhận dạng thực thể và phân tích

cú pháp phụ thuộc.

Sử dụng PhoBERT-base-v2 trên tập dữ liệu triệu tập từ mạng xã hội và hội thoại tự nhiên, kết hợp kỹ thuật lấy mẫu bổ sung để cân bằng dữ liệu. Kết quả Accuracy đạt 94.22%, các chỉ số macro Precision/Recall/F1 đều vượt 94%. Hơn nữa, mô hình được triển khai giao diện Gradio để phân tích cảm xúc thời gian thực.

Những nghiên cứu này cung cấp đánh giá rõ ràng về hiệu suất của PhoBERT-v2 trong các tác vụ nhận diện cảm xúc đối với tiếng Việt. Bạn có thể tham khảo siêu tham số: learning rate khoảng 2e-5, batch size 16, training epoch = 10, dùng Adam optimizer và tokenization kết hợp byte-pair encoding và RDR Segmenter để phân tách từ tiếng Việt. Nghiên cứu này cung cấp nền tảng kỹ thuật đáng tin cậy để đề tài có thể tham khảo trực tiếp, đặc biệt trong phần tiền xử lý, cân bằng dữ liệu và lựa chọn kiến trúc mô hình.

1.4 Tổng quan mô hình học sâu trong NLP

1.4.1 PhoBERT – Mô hình ngôn ngữ tiền huấn luyện cho tiếng Việt

PhoBERT là một mô hình ngôn ngữ tiền huấn luyện (pre-trained language model) do VinAI Research phát triển, được thiết kế tối ưu hóa riêng cho tiếng Việt. PhoBERT kế thừa kiến trúc RoBERTa – một biến thể cải tiến của mô hình BERT (Bidirectional Encoder Representations from Transformers) do Google đề xuất [2], nhưng được huấn luyện hoàn toàn trên dữ liệu tiếng Việt. Cụ thể, PhoBERT được huấn luyện trên tập dữ liệu 20GB văn bản tiếng Việt chuẩn hóa, bao gồm nội dung từ báo chí, diễn đàn và mạng xã hội, nhằm giúp mô hình học được đầy đủ đặc trưng ngữ nghĩa và cú pháp của tiếng Việt [1].

Vì tiếng Việt là ngôn ngữ đơn âm và không sử dụng dấu cách để phân tách từ như tiếng Anh, PhoBERT sử dụng bộ tách từ RDRSegmenter và kỹ thuật mã hóa từ dựa trên ghép cặp byte (Byte Pair Encoding – BPE) để xử lý đầu vào. Điều này đảm bảo khả năng học biểu diễn ngữ nghĩa chính xác và ổn định hơn so với các phương pháp biểu diễn từ truyền thống như bag-of-words hay word2vec.

Về mặt kiến trúc, PhoBERT được xây dựng dựa trên bộ mã hóa chuỗi Transformer (Transformer encoder stack), gồm 12 tầng xử lý, mỗi tầng có cơ chế tự chú ý đa đầu (multi-head self-attention) và mạng neuron lan truyền về phía trước (feedforward neural network). Mỗi tầng giúp mô hình học được mối quan hệ giữa các từ trong chuỗi theo cả hai chiều

ngữ cảnh trái và phải, từ đó xây dựng biểu diễn ngữ nghĩa theo ngữ cảnh.

Trong cơ chế tự chú ý (self-attention), mỗi từ trong câu sẽ được ánh xạ thành ba vector là truy vấn (query – Q), khóa (key – K) và giá trị (value – V). Mô hình tính toán mức độ liên quan giữa các từ thông qua công thức:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Trong đó d_k là số chiều của vector khóa. Kết quả thu được là các trọng số thể hiện mức độ liên kết ngữ nghĩa giữa các từ trong câu, từ đó giúp mô hình tập trung hơn vào những thành phần có vai trò quan trọng trong biểu đạt ý nghĩa.

Khi nhận một câu đầu vào, PhoBERT sẽ mã hóa câu này thành một chuỗi các vector ngữ nghĩa. Vector biểu diễn của toàn bộ câu sẽ được trích xuất từ vị trí đặc biệt [CLS] (viết tắt của "classification"), là token được mô hình gán vào đầu câu và học để tóm tắt thông tin toàn bộ câu. Vector tại [CLS] có kích thước 768 chiều và được sử dụng như đầu vào cho bộ phân loại cảm xúc.

PhoBERT được lựa chọn trong nghiên cứu này vì các ưu điểm vượt trội: được huấn luyện chuyên biệt trên tiếng Việt, hỗ trợ xử lý tốt ngôn ngữ đơn âm, biểu diễn ngữ nghĩa theo ngữ cảnh hiệu quả hơn so với các phương pháp truyền thống, và đã chứng minh hiệu năng cao trong nhiều tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản, gán nhãn thực thể và phân tích cú pháp tiếng Việt [1].

1.4.2 Bộ phân loại tuyến tính – Linear Classifier

Sau khi trích xuất đặc trưng ngữ nghĩa của câu từ mô hình PhoBERT, vector đầu ra tại vị trí [CLS] sẽ được đưa vào một bộ phân loại tuyến tính (linear classifier) để xác định cảm xúc của câu. Bộ phân loại tuyến tính bao gồm một lớp mạng neuron đơn giản (fully connected layer), kết hợp với hàm kích hoạt softmax để chuyển đầu ra thành xác suất thuộc về từng nhãn cảm xúc, chẳng hạn như tích cực, tiêu cực hoặc trung lập.

Bộ phân loại này được mô tả bởi công thức:

$$y = softmax(Wx + b)$$

Trong đó x là vector 768 chiều từ mô hình PhoBERT, W là ma trận trọng số học

được trong quá trình huấn luyện, và *b* là vector độ lệch (bias). Kết quả *y* là một vector chứa xác suất tương ứng với từng lớp cảm xúc. Hàm softmax có vai trò chuyển đổi đầu ra thô thành phân phối xác suất, được định nghĩa như sau:

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Trong đó: *i = 1, 2,…,K* - *K* là số lớp cảm xúc.

Bộ phân loại tuyến tính có ưu điểm là đơn giản, dễ triển khai và huấn luyện nhanh. Khi đi kèm với biểu diễn mạnh mẽ từ PhoBERT, mô hình vẫn đạt độ chính xác cao trong các tác vụ phân loại mà không cần thêm các thành phần phức tạp như mạng LSTM hai chiều (Bi-LSTM) hay cơ chế tự chú ý độc lập (Attention). Điều này đặc biệt phù hợp với các bài toán thực tế trong giáo dục, nơi dữ liệu có thể không lớn nhưng yêu cầu xử lý nhanh và đáng tin cậy.

CHƯƠNG 2: XÂY DỰNG VÀ HUẤN LUYỆN MÔ HÌNH PHÂN TÍCH CẢM XÚC TIẾNG VIỆT DỰA TRÊN PHOBERT

Chương này trình bày quy trình xây dựng mô hình nhận diện cảm xúc tiếng Việt sử dụng PhoBERT kết hợp bộ phân loại tuyến tính (Linear Classifier). Bộ dữ liệu được mô tả về nguồn gốc, cách gán nhãn và chuẩn hóa. Các bước tiền xử lý nhằm đảm bảo đầu vào phù hợp cho mô hình huấn luyện. Sau đó, mô hình PhoBERT được sử dụng để trích xuất đặc trưng ngữ nghĩa, đầu ra [CLS] được đưa vào lớp phân loại tuyến tính để dự đoán cảm xúc. Cuối cùng, các chỉ số như Accuracy, Precision, Recall và F1-score được áp dụng để đánh giá hiệu suất mô hình một cách toàn diện.

2.1. Chuẩn bị dữ liệu huấn luyện

2.1.1. Nguồn dữ liệu

Để huấn luyện mô hình nhận diện cảm xúc trong văn bản tiếng Việt, đề tài sử dụng bộ dữ liệu Vietnamese Students Feedback được công bố trên nền tảng HuggingFace, với mã định danh uitnlp/vietnamese_students_feedback. Bộ dữ liệu này bao gồm các phản hồi thực tế của sinh viên Việt Nam về các khía cạnh khác nhau trong môi trường học tập như: giảng viên, chương trình đào tạo, cơ sở vật chất và các yếu tố liên quan khác.

Dữ liệu với khoảng 16.000 dòng được chia thành ba tập sẵn sàng cho huấn luyện và đánh giá mô hình:

- *Tập huấn luyện (train)* gồm 11.400 dòng dữ liệu trong đó:

+ Nhãn tích cực (positive) chiếm 49.4%;

+ Nhãn trung tính (neutral) chiếm 4%;

+ Nhãn tiêu cực (negative) chiếm 46.6%.

- *Tập kiểm định (validation)* gồm 1.580 dòng dữ liệu, trong đó:

+ Nhãn tích cực (positive) chiếm 50.9%;

+ Nhãn trung tính (neutral) chiếm 4.6%;

+ Nhãn tiêu cực (negative) chiếm 44.5%.

- *Tập kiểm tra (test)* gồm 3.170 dòng dữ liệu, trong đó:

+ Nhãn tích cực (positive) chiếm 50.2%;

+ Nhãn trung tính (neutral) chiếm 5.3%;

+ Nhãn tiêu cực (negative) chiếm 44.5%.

Mỗi dòng trong tập dữ liệu bao gồm ba trường thông tin chính:

- sentence: câu phản hồi bằng tiếng Việt;

- sentiment: nhãn cảm xúc (0: tiêu cực, 1: trung tính, 2: tích cực);

- topic: chủ đề phản hồi (0: giảng viên, 1: chương trình học, 2: cơ sở vật chất, 3: khác).

sentence string · lengths 4 718	sentiment class label 3 classes	topic class label 4 classes
giáo trình chưa cụ thể .	0 negative	1 training_program
giảng buồn ngủ .	0 negative	0 lecturer
giáo viên vui tính , tận tâm .	2 positive	0 lecturer
giảng viên nên giao bài tập nhiều hơn , chia nhóm để làm bài tập , giảng kỹ những vấn đề trọng tâm của môn học , đưa ra phương pháp để học hiệu quả hơn .	0 negative	0 lecturer
giảng viên cần giảng bài chi tiết hơn , đi sâu hơn code và chạy thử chương trình có trong bài giảng nếu được .	0 negative	0 lecturer
nên có giảng viên nước ngoài dạy để sinh viên có cơ hội thực hành giao tiếp .	0 negative	0 lecturer
nên có bài tập lớn đồ án môn học .	0 negative	1 training_program

Hình 2.1: Bộ dữ liệu Vietnamese Student Feedback

2.1.2. Tiền xử lý văn bản

Trước khi đưa vào mô hình, văn bản được xử lý qua các bước làm sạch và chuẩn hóa nhằm đảm bảo chất lượng ngữ liệu đầu vào. Cụ thể, toàn bộ văn bản được chuẩn hóa unicode và chuyển về chữ thường để giảm thiểu các biến thể không cần thiết. Các ký tự đặc biệt, dấu câu và chữ số được loại bỏ, chỉ giữ lại chữ cái và khoảng trắng. Sau đó, văn bản được tách từ bằng thư viện PyVi (ViTokenizer), giúp mô hình BERT tiếp cận hiệu quả hơn với các đơn vị từ vựng tiếng Việt. Ngoài ra, khoảng trắng dư thừa được xóa bỏ để đảm bảo định dạng thống nhất. Trong trường hợp cần mở rộng bước tiền xử lý, một danh sách stopwords tiếng Việt cũng có thể được áp dụng để loại bỏ các từ không mang giá trị cảm xúc rõ rệt.

2.1.3. Cân bằng tập dữ liệu

Trong quá trình phân tích dữ liệu ban đầu, học viên nghiên cứu nhận thấy sự mất cân bằng rõ rệt giữa ba lớp cảm xúc, trong đó nhãn 1 (trung tính) chiếm tỷ lệ thấp hơn đáng kể so với hai nhãn còn lại (khoảng 4 – 5%). Sự chênh lệch này có thể ảnh hưởng tiêu cực đến quá trình học của mô hình, dẫn đến dự đoán lệch và giảm hiệu quả phân loại. Để khắc phục vấn đề trên, học viên đã tiến hành bổ sung dữ liệu trung tính nhân tạo thông qua ba phương pháp chính.

Thứ nhất, tạo câu trung tính bằng cách kết hợp các câu tích cực (nhãn 2) và tiêu cực (nhãn 0). Cụ thể, học viên trích một phần từ mỗi loại câu rồi ghép lại bằng các từ nối mang tính chuyển tiếp như “nhưng”, “tuy nhiên”, “mặc dù”… Đồng thời, chèn thêm các cụm từ thể hiện sự không chắc chắn như “có thể”, “chưa rõ”, “tôi nghĩ là…” để tạo sắc thái trung lập.

Thứ hai, sử dụng các câu mẫu (template) theo chủ đề để sinh văn bản. Các mẫu dạng điền khuyết như “phòng học {adj}”, “giảng viên {adj}” được xây dựng sẵn, sau đó điền vào các tính từ trung tính như “bình thường”, “tạm ổn”, “không đặc biệt” nhằm đảm bảo giữ nguyên ngữ nghĩa và cảm xúc trung lập. Mỗi câu được gán nhãn cảm xúc là 1 và gắn với một chủ đề cụ thể.

Thứ ba, xây dựng một danh sách thủ công gồm hơn 25 cụm phản hồi trung tính rút trích từ dữ liệu thực tế, chẳng hạn như “giảng viên dạy tạm được”, “khối lượng bài

tập vừa phải", "không rõ có hiệu quả hay không"… Các câu này được gán nhãn trung tính và phân bố chủ đề ngẫu nhiên theo trọng số phù hợp.

Thông qua việc bổ sung có kiểm soát các dữ liệu trung tính, học viên kỳ vọng mô hình sẽ học được đặc trưng tốt hơn cho lớp cảm xúc này, từ đó nâng cao độ cân bằng và độ chính xác tổng thể khi phân loại cảm xúc.

Sau khi tổng hợp các câu trung tính bằng các phương pháp trên, tập dữ liệu được cân bằng theo tỷ lệ tương đối giữa ba nhãn, đồng thời dữ liệu tổng được tăng lên 21.637 dòng dữ liệu hỗ trợ việc huấn luyện cho mô hình. Các tập huấn luyện (train), kiểm định (validation), và kiểm tra (test) mới được lưu dưới tên:

Tập huấn luyện: vietnamese_students_feedback_train_balanced.csv (15.338 dữ liệu sau cân bằng)

Tập kiểm định: vietnamese_students_feedback_validation_balanced.csv (2.113 dữ liệu sau cân bằng)

Tập kiểm tra: vietnamese_students_feedback_test_balanced.csv (gồm 4.186 dữ liệu sau cân bằng)

Việc cân bằng nhãn giúp mô hình học đều trên các loại cảm xúc, tăng độ chính xác và khả năng phân biệt trong môi trường phản hồi thực tế của sinh viên.

2.2. Xây dựng mô hình phân loại cảm xúc sử dụng PhoBERT

2.2.1. Lý do chọn mô hình PhoBERT

PhoBERT là một mô hình học sâu được tiền huấn luyện đặc biệt dành cho tiếng Việt, phát triển bởi nhóm nghiên cứu tại VinAI. Mô hình này kế thừa kiến trúc BERT nhưng được huấn luyện trên kho văn bản tiếng Việt khổng lồ, sử dụng token hóa dạng SentencePiece thay vì WordPiece để phù hợp với ngôn ngữ đơn âm như tiếng Việt. So với các mô hình tiền nhiệm như BERT multilingual, PhoBERT cho thấy hiệu suất vượt trội trong nhiều bài toán xử lý ngôn ngữ tự nhiên tiếng Việt, bao gồm phân loại văn bản, nhận diện thực thể, và đặc biệt là phân tích cảm xúc.

2.2.2. Cấu trúc mô hình đề xuất

Mô hình huấn luyện trong đề tài được xây dựng trên nền tảng vinai/phobert-base với kiến trúc tổng quát như sau:

- Sử dụng PhoBERT để trích xuất biểu diễn ngữ nghĩa của toàn bộ câu. Đầu ra là vector biểu diễn của token [CLS], được xem là đại diện toàn câu.

- Kết hợp Dropout giúp tránh hiện tượng overfitting bằng cách ngẫu nhiên loại bỏ một phần neuron trong quá trình huấn luyện.

- Phân loại Linear để ánh xạ đầu ra của PhoBERT thành phân phối xác suất trên ba lớp cảm xúc (0, 1, 2).

2.2.3 Tiền xử lý và tokenizer

PhoBERT yêu cầu văn bảng đầu vào phải được token hóa phù hợp với từ điển SentencePiece đã được huấn luyện trước. Trong đề tài, quá trình tokenization được thực hiện tự động thông qua AutoTokenizer từ thư viện HuggingFace Transformers.

Sau khi token hóa, mỗi câu được cắt/truncate đến độ dài tối đa 128 tokens, padding nếu cần và tạo ra các tensor input_ids và attention_mask để nạp vào mô hình.

Việc tiền xử lý dữ liệu lần 2 trước khi vào phần huấn luyện giúp mô hình xử lý nhanh hơn.

2.2.4 Datasets tùy chỉnh

Để tương thích với DataLoader của Pytorch, học viên thực hiện nghiên cứu xây dựng lớp Dataset riêng có tên SentimentDatasetBERT, trong đó mỗi phần tử là một câu phản hồi đã được mã hóa và nhãn cảm xúc tương ứng.

Lớp này đảm bảo dữ liệu được nạp tuần tự thơ từng batch với đầy đủ đầu vào cần thiết cho huấn luyện.

2.3 Huấn luyện mô hình

2.3.1. Thiết lập cấu hình huấn luyện

Quá trình huấn luyện mô hình PhoBERT được thiết lập dựa trên các thực nghiệm và tham khảo từ các nghiên cứu liên quan trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Các siêu tham số được lựa chọn theo tiêu chuẩn thực nghiệm phổ biến, đồng thời được

điều chỉnh phù hợp với quy mô tập dữ liệu và khả năng tính toán của hệ thống.

Mô hình nền tảng: Mô hình sử dụng trong nghiên cứu là PhoBERT-base, một phiên bản của BERT được tiền huấn luyện đặc biệt cho tiếng Việt. Cụ thể, phiên bản PhoBERT-base có:

- 12 tầng (transformer layers),

- 768 chiều vector ẩn (hidden size),

- 12 đầu attention (attention heads),

- Tổng cộng khoảng 135 triệu tham số.

Các siêu tham số chính

- Epoch: 20 – số lần lặp lại toàn bộ tập huấn luyện. Đây là mức đủ để mô hình hội tụ trên tập dữ liệu không quá lớn.

- Batch size: 8 – kích thước được lựa chọn nhằm cân bằng giữa tốc độ huấn luyện và hạn chế tràn bộ nhớ (out-of-memory) khi huấn luyện trên GPU.

- Max_length: 128 token – đảm bảo đủ thông tin cho các câu phản hồi ngắn đến trung bình trong lĩnh vực giáo dục.

- Learning rate: 2e-5 – tốc độ học nhỏ giúp mô hình cập nhật ổn định, đặc biệt quan trọng khi fine-tuning các mô hình tiền huấn luyện lớn như BERT.

- Dropout rate: 0.3 – nhằm giảm hiện tượng overfitting trong giai đoạn fine-tuning.

- Epsilon (eps) in optimizer: 1e-8 – đảm bảo tính ổn định trong quá trình cập nhật trọng số.

Trình tối ưu hóa và lịch trình học

- Optimizer: AdamW – một cải tiến của Adam, phù hợp với mô hình transformer do khả năng xử lý tốt các tham số có điều kiện khác nhau.

- Scheduler: get_linear_schedule_with_warmup – giảm tuyến tính tốc độ học (learning rate) sau một số bước khởi động (warmup), giúp mô hình học ổn định hơn và tránh cập nhật đột ngột khi bắt đầu huấn luyện.

Chiến lược tăng tốc huấn luyện: Trong trường hợp mô hình được huấn luyện trên GPU, đề tài áp dụng kỹ thuật mixed precision training (FP16) thông qua thư viện torch.cuda.amp. Kỹ thuật này cho phép sử dụng độ chính xác hỗn hợp giữa 16-bit và 32-bit, góp phần:

- Giảm mức sử dụng bộ nhớ GPU, từ đó cho phép batch size lớn hơn hoặc mô hình phức tạp hơn.

- Tăng tốc quá trình huấn luyện, đồng thời vẫn duy trì được độ chính xác tương đương với FP32.

Gradient clipping

Để tránh hiện tượng bùng nổ gradient (exploding gradients), gradient của mô hình được giới hạn ở mức max_norm = 1.0. Việc này giúp quá trình huấn luyện ổn định hơn, đặc biệt trong các mô hình mạng sâu như BERT.

Thiết bị huấn luyện

Toàn bộ quá trình huấn luyện mô hình được triển khai trên nền tảng Kaggle Notebooks, sử dụng phần cứng hỗ trợ TPU (Tensor Processing Unit). Cụ thể, đề tài sử dụng TPU v3–8 với cấu hình 2 cores, cho phép huấn luyện nhanh chóng các mô hình học sâu có số lượng tham số lớn như PhoBERT.

TPU là phần cứng chuyên biệt được thiết kế bởi Google để tối ưu hóa các phép toán tensor thường gặp trong học sâu, đặc biệt là khi sử dụng các thư viện như PyTorch/XLA hoặc TensorFlow. Trong bối cảnh đề tài, việc sử dụng TPU mang lại một số lợi ích đáng kể:

- Tăng tốc đáng kể tốc độ huấn luyện so với GPU truyền thống;

- Cho phép xử lý song song các batch lớn, từ đó rút ngắn thời gian hội tụ mô hình;

- Ổn định và tương thích với hạ tầng của HuggingFace Transformers, thông qua tích hợp với thư viện accelerate hoặc torch_xla.

Mô hình được huấn luyện hoàn toàn trên TPU với đầy đủ hỗ trợ từ thư viện PyTorch và HuggingFace Transformers, đảm bảo tính tương thích, hiệu quả và khả năng mở rộng cho các bài toán học sâu khác trong tương lai.

2.3.2. Tạo tập dữ liệu và DataLoader

Để phục vụ cho quá trình huấn luyện mô hình học sâu với kiến trúc BERT, tập dữ liệu đầu vào cần được xử lý và chuyển đổi sang định dạng tensor, đồng thời tổ chức thành từng batch tối ưu hóa việc truyền dữ liệu vào mô hình.

Chuẩn hóa văn bản và gán nhãn

Sau bước cân bằng và tiền xử lý sơ bộ được trình bày tại Mục 2.1, các câu phản hồi tiếng Việt trong ba tập dữ liệu (huấn luyện, kiểm định, kiểm thử) tiếp tục được chuẩn hóa nhằm đảm bảo định dạng đầu vào phù hợp cho quá trình huấn luyện mô hình. Cụ thể, hàm normalize_text() được áp dụng để loại bỏ dấu câu, ký tự đặc biệt và chuyển toàn bộ văn bản về chữ thường. Đồng thời, dữ liệu văn bản (sentence) và nhãn (sentiment) được tách riêng phục vụ cho bước mã hóa và đưa vào mô hình.

Việc phân tách quá trình tiền xử lý thành hai giai đoạn được thực hiện có chủ đích. Giai đoạn đầu nhằm xử lý dữ liệu thô để phục vụ mục tiêu xây dựng tập dữ liệu huấn luyện hoàn chỉnh, bao gồm các thao tác như tách từ, cân bằng nhãn và làm sạch sơ bộ. Giai đoạn sau tập trung vào chuẩn hóa ngữ liệu đầu vào ở mức độ thống nhất, nhằm đảm bảo dữ liệu đáp ứng đầy đủ yêu cầu kỹ thuật của tokenizer PhoBERT và giảm thiểu tối đa các yếu tố nhiễu còn tồn tại. Cách tiếp cận này giúp nâng cao hiệu quả huấn luyện và đảm bảo tính nhất quán trong toàn bộ quy trình xử lý ngôn ngữ tự nhiên.

Token hóa bằng PhoBERT Tokenizer

PhoBERT sử dụng mô hình token hóa đặc thù SentencePiece, khác với BERT thông thường. Việc mã hóa dữ liệu văn bản được thực hiện thông qua AutoTokenizer tương ứng với mô hình vinai/phobert-base. Các câu phản hồi được chuyển thành:

- input_ids: danh sách các chỉ số từ trong từ điển PhoBERT;

- attention_mask: mặt nạ đánh dấu các token có ý nghĩa (1) và padding (0).

Tokenization được thực hiện với các cấu hình sau:

- add_special_tokens = True: tự động thêm token [CLS] và [SEP] cần thiết cho BERT.

- max_length = 128: giới hạn độ dài token để đảm bảo đồng đều và phù hợp với

dung lượng tính toán.

- padding = 'max_length': bổ sung padding khi cần thiết.

- truncation = True: cắt bớt nếu câu dài hơn ngưỡng.

Xây dựng lớp Dataset tùy chỉnh: Nhằm tương thích với DataLoader của PyTorch, đề tài xây dựng lớp SentimentDatasetBERT kế thừa torch.utils.data.Dataset. Mỗi mẫu dữ liệu bao gồm:

- input_ids: tensor chứa mã token,

- attention_mask: tensor chứa mặt nạ chú ý,

- label: nhãn cảm xúc dạng LongTensor.

Lớp Dataset này cho phép tùy biến linh hoạt việc tiền xử lý văn bản và ánh xạ các phần tử thành định dạng đầu vào của mô hình.

Khởi tạo DataLoader: Sau khi xây dựng Dataset cho ba tập dữ liệu, đề tài sử dụng DataLoader để tổ chức dữ liệu theo từng batch, hỗ trợ shuffling và tăng tốc truy xuất. Cấu hình DataLoader như sau:

- batch_size = 8 (tương ứng với cấu hình siêu tham số tại mục 2.3.1),

- shuffle = True đối với tập huấn luyện để đảm bảo đa dạng hóa các batch qua mỗi epoch,

- num_workers = 2 khi sử dụng TPU hoặc GPU để tận dụng khả năng xử lý song song,

- pin_memory = True để tăng hiệu suất truyền dữ liệu (nếu sử dụng thiết bị CUDA/TPU).

Việc khởi tạo DataLoader cho ba tập train, validation và test đảm bảo luồng xử lý dữ liệu mượt mà, chuẩn bị sẵn sàng cho quá trình huấn luyện trong các bước tiếp theo.

2.3.3. Huấn luyện qua các epoch

Sau khi chuẩn bị dữ liệu và xây dựng mô hình như đã trình bày ở các mục trước, mô hình PhoBERT được huấn luyện qua nhiều vòng lặp (epoch) nhằm tối ưu hóa trọng

số sao cho sai số giữa dự đoán và nhãn thực tế được giảm thiểu tối đa.

Quy trình huấn luyện: Trong mỗi epoch, dữ liệu huấn luyện được chia thành các mini-batch và đưa vào mô hình theo trình tự sau:

- Forward pass: mô hình nhận input_ids và attention_mask, tính toán đầu ra logits thông qua kiến trúc PhoBERT và lớp phân loại tuyến tính.

- Tính toán hàm mất mát (loss): sử dụng hàm CrossEntropyLoss – hàm mất mát phổ biến cho bài toán phân loại nhiều lớp:

$$£= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} yij \times \log(Pij)$$

Trong đó:

- N: số mẫu trong batch,
- C: số lớp cảm xúc (3 lớp),
- yij: nhãn thực tế (one-hot),
- Pij: xác suất mô hình dự đoán cho lớp j với mẫu i.

- Backward pass: lan truyền ngược đạo hàm của hàm mất mát để cập nhật trọng số.

- Gradient clipping: giới hạn giá trị gradient không vượt quá một ngưỡng nhất định (ở đây là $|g|_2 \leq 1.0$) nhằm tránh hiện tượng bùng nổ gradient (exploding gradients).

- Cập nhật trọng số: thực hiện bởi tối ưu hóa AdamW, cùng với lịch trình điều chỉnh learning rate tuyến tính (get_linear_schedule_with_warmup).

- Tích hợp mixed precision (nếu có): sử dụng thư viện torch.cuda.amp để tăng tốc độ huấn luyện và giảm bộ nhớ bằng việc tự động chuyển đổi giữa FP16 và FP32.

Theo dõi các chỉ số trong quá trình huấn luyện: Trong mỗi epoch, ba chỉ số chính được tính toán để đánh giá quá trình học của mô hình:

- Độ chính xác (Accuracy):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- Điểm F1 trung bình có trọng số (Weighted F1-score):

$$F1_{Weighted} = \sum_{i=1}^{C} \frac{n_i}{N} \times \frac{2 \times Precision_{i \times Recall_i}}{Precision_i + Recall_i}$$

Trong đó:

C: là số lớp

n_i: là số mẫu của lớp i, N là tổng số mẫu

Precision và Recall của lớp i tính theo công thức

$$Precision_i = \frac{TP_i}{TP_i + FP_i}, Recall = \frac{TP_i}{TP_i + FN};$$

Chiến lược lưu mô hình

Sau mỗi epoch, mô hình được đánh giá trên tập validation. Nếu điểm F1-score trên validation đạt mức cao hơn so với các epoch trước, mô hình sẽ được lưu lại vào file best_model.pt. Việc lưu mô hình theo tiêu chí F1 giúp đảm bảo hiệu suất phân loại đồng đều giữa các lớp, đặc biệt trong trường hợp dữ liệu có sự mất cân bằng nhẹ giữa các nhãn.

2.3.4. Triển khai mô hình dự đoán cảm xúc văn bản mới

Sau khi hoàn tất quá trình huấn luyện, mô hình PhoBERT có thể được sử dụng như một công cụ phân tích cảm xúc cho các văn bản tiếng Việt mới. Việc triển khai chức năng dự đoán không chỉ giúp kiểm chứng tính ứng dụng của mô hình mà còn phục vụ các tình huống thực tế như: phân tích phản hồi sinh viên, hỗ trợ giáo viên đánh giá tâm lý học sinh, hoặc phản hồi tự động trong hệ thống giáo dục thông minh.

Cấu trúc hàm dự đoán: Hàm predict_sentiment() được xây dựng để nhận đầu vào là một chuỗi văn bản tiếng Việt (phản hồi từ người dùng), thực hiện các bước tiền xử lý và mã hóa tương tự như trong quá trình huấn luyện, và trả về kết quả dự đoán thuộc một trong ba lớp cảm xúc:

- 0: Tiêu cực
- 1: Trung tính

- 2: Tích cực

Quy trình dự đoán

Quy trình dự đoán cảm xúc từ văn bản đầu vào được triển khai theo bốn bước chính, nhằm đảm bảo tính chính xác và nhất quán trong suốt quá trình suy diễn của mô hình học sâu:

Chuẩn hóa văn bản: Trước khi đưa vào mô hình, văn bản đầu vào được xử lý bằng hàm normalize_text() để làm sạch và chuẩn hóa nội dung. Cụ thể, các thao tác bao gồm:

- Chuyển toàn bộ văn bản về chữ thường nhằm giảm thiểu sự phân tán từ vựng;

- Loại bỏ dấu câu, ký tự đặc biệt không cần thiết;

- Chuẩn hóa mã hóa Unicode và loại bỏ các khoảng trắng dư thừa.

Mã hóa bằng Tokenizer: Văn bản sau khi chuẩn hóa được mã hóa thành các định dạng đầu vào phù hợp với mô hình học sâu thông qua tokenizer đi kèm với mô hình vinai/phobert-base. Tokenizer thực hiện:

- Chuyển văn bản thành chuỗi các chỉ số input_ids và attention_mask đặc trưng cho kiến trúc BERT;

- Áp dụng cơ chế padding và truncation để đảm bảo mọi câu đầu vào có độ dài thống nhất, không vượt quá 128 tokens.

Suy diễn (Inference): Quá trình dự đoán được thực hiện trong chế độ đánh giá (model.eval()) để đảm bảo mô hình không cập nhật trọng số. Các bước cụ thể:

- Dữ liệu đầu vào được truyền qua mô hình đã huấn luyện để tạo ra một vector đầu ra logits;

- Lớp cảm xúc được dự đoán là lớp tương ứng với xác suất cao nhất, xác định bằng hàm argmax sau khi áp dụng softmax trên logits.

Hiển thị kết quả: Chỉ số lớp đầu ra được ánh xạ ngược về nhãn cảm xúc thông qua một từ điển ánh xạ (sentiment_map). Cuối cùng, hệ thống trả lại kết quả cho người dùng dưới dạng chuỗi mô tả tương ứng với ba trạng thái cảm xúc: “Tiêu cực”**,** “Trung tính”**,** hoặc “Tích cực”**.**

Ứng dụng thực tiễn

Hàm dự đoán cảm xúc được xây dựng trong nghiên cứu có tiềm năng tích hợp vào các hệ thống giáo dục thông minh với nhiều ứng dụng thực tiễn. Cụ thể, hệ thống có thể:

- Tự động phân tích cảm xúc từ các phản hồi ngôn ngữ tự nhiên của học sinh nhằm hỗ trợ đánh giá mức độ hài lòng, căng thẳng hoặc thái độ học tập;

- Gợi ý hành động can thiệp kịp thời trong bối cảnh tâm lý học đường, qua đó hỗ trợ công tác tư vấn và chăm sóc sức khỏe tinh thần học sinh;

- Hỗ trợ giáo viên và nhà quản lý giáo dục trong việc điều chỉnh nội dung, phương pháp giảng dạy hoặc quản lý lớp học dựa trên các chỉ số cảm xúc thu thập được.

Bên cạnh lĩnh vực giáo dục, mô hình cũng có thể được mở rộng áp dụng trong các bối cảnh khác như: phân tích dư luận trên mạng xã hội, khảo sát xã hội học, hệ thống tổng đài hoặc đối thoại tự động bằng tiếng Việt, từ đó góp phần nâng cao hiệu quả tương tác người–máy và hiểu biết ngôn ngữ tự nhiên trong tiếng Việt.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH

3.1. Cấu hình thực nghiệm

Để kiểm tra hiệu quả của mô hình PhoBERT trong bài toán nhận diện cảm xúc tiếng Việt, các thực nghiệm được triển khai trong môi trường tính toán hỗ trợ tăng tốc phần cứng, cùng với hệ thống cấu hình mô hình và siêu tham số được xác định nhất quán trong suốt quá trình huấn luyện và đánh giá.

3.1.1. Môi trường thực thi

Toàn bộ quá trình huấn luyện và thử nghiệm được thực hiện trên nền tảng Kaggle Notebooks, sử dụng phần cứng hỗ trợ TPU v3–8 (2 cores). Đây là hệ thống tính toán chuyên dụng do Google phát triển, tối ưu cho các tác vụ học sâu, đặc biệt là các mô hình transformer như BERT.

TPU được lựa chọn thay vì GPU thông thường vì những lợi thế:

- Tốc độ tính toán vượt trội với độ chính xác hỗn hợp (mixed precision),

- Khả năng huấn luyện mô hình lớn trên tập dữ liệu vừa và lớn với thời gian hội tụ

ngắn,

- Tương thích tốt với thư viện torch_xla trong PyTorch và bộ công cụ Transformers từ HuggingFace.

3.1.2. Mô hình và thư viện sử dụng

- Mô hình nền tảng: vinai/phobert-base, với 12 tầng Transformer, 768 chiều ẩn, 12 đầu attention, khoảng 135 triệu tham số.

- Thư viện chính:

+ transformers (HuggingFace): tải mô hình, tokenizer, scheduler.

+ torch, torch.nn, torch.utils.data: xây dựng mô hình, Dataset, DataLoader.

+ torch.cuda.amp: hỗ trợ huấn luyện với mixed precision.

+ sklearn.metrics: đánh giá mô hình với classification report, confusion matrix.

+ matplotlib, seaborn: trực quan hóa kết quả.

3.1.3. Cấu hình siêu tham số

Các siêu tham số được chọn sau khi tham khảo các nghiên cứu học sâu với BERT và được điều chỉnh phù hợp với dung lượng dữ liệu:

Bảng 3.1: Bảng cài đặt siêu tham số mô hình học sâu BERT

Tham số	Giá trị
Epochs	20
Batch size	8
Learning rate	2e-5
Max sequence length	128 tokens
Optimizer	AdamW
Scheduler	Linear schedule

Dropout	0.3
Gradient clipping	Max_norm = 1.0
Mixed Precision	FP16

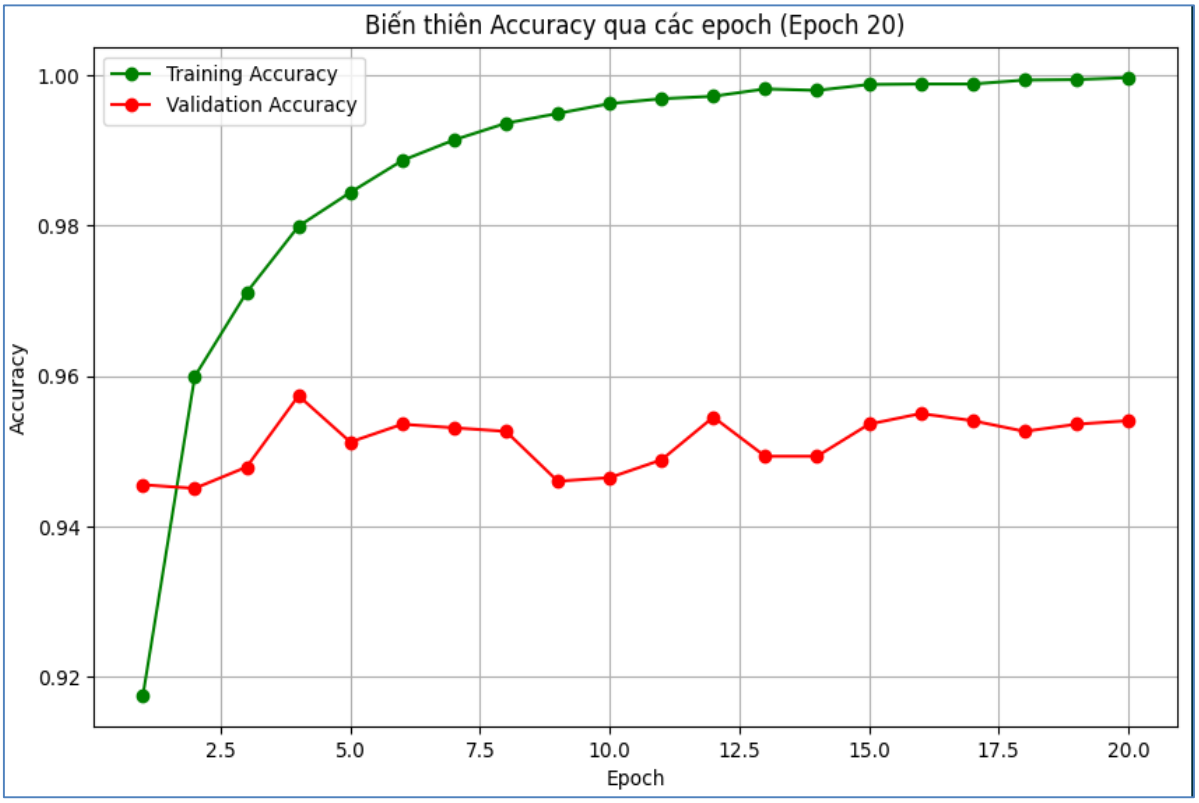
Việc huấn luyện được tiến hành với số lượng epoch tương đối cao (20 vòng) để quan sát rõ quá trình hội tụ và hiện tượng quá khớp (overfitting). Các chỉ số như độ mất mát (loss), độ chính xác (accuracy), trung bình điều hòa (F1-score) được ghi nhận và lưu trữ qua từng epoch để phục vụ cho việc đánh giá và trực quan hóa ở các mục sau.

3.2. Kết quả huấn luyện

3.2.1. Độ chính xác huấn luyện và kiểm định

Độ chính xác (accuracy) là một trong những chỉ số cơ bản nhưng quan trọng để đánh giá hiệu quả của mô hình phân loại. Trong quá trình huấn luyện, độ chính xác được ghi nhận sau mỗi epoch đối với cả hai tập: tập huấn luyện (training) và tập kiểm định (validation).

Hình dưới đây thể hiện sự biến thiên độ chính xác qua từng epoch trong 20 vòng lặp đầu tiên:



Hình 3.1: Accuracy mô hình qua 20 epoch

Phân tích kết quả

Đối với tập huấn luyện: Độ chính xác tăng đều đặn từ khoảng 91.8% (epoch 1) lên đến gần 100% tại epoch 20. Điều này phản ánh quá trình hội tụ tốt của mô hình trên tập dữ liệu đã học, cho thấy mô hình có khả năng ghi nhớ đặc trưng dữ liệu rất hiệu quả.

Đối với tập kiểm định: Ngược lại, độ chính xác trên tập kiểm định dao động trong khoảng 94.5% – 95.7% và không tăng trưởng rõ rệt theo thời gian huấn luyện. Đường biểu diễn cho thấy sự ổn định nhưng có biến động nhẹ, phản ánh mô hình giữ được khả năng tổng quát hóa ở mức nhất định.

Nhận xét

Khoảng cách giữa hai đường biểu diễn – đặc biệt từ epoch thứ 10 trở đi – bắt đầu nới rộng, khi độ chính xác huấn luyện tiến dần về 1.0, trong khi độ chính xác của tập kiểm định dừng lại ở mức dưới 96%. Hiện tượng này gợi ý rằng mô hình có thể đã bắt đầu học thuộc (overfitting) dữ liệu huấn luyện, tức là mô hình ghi nhớ dữ liệu huấn luyện quá kỹ nhưng không còn cải thiện đáng kể trên dữ liệu chưa thấy.

Mặc dù vậy, việc giữ được độ chính xác của tập kiểm định ổn định và duy trì ở mức cao cho thấy mô hình vẫn đảm bảo khả năng khái quát trong phạm vi tập kiểm định.

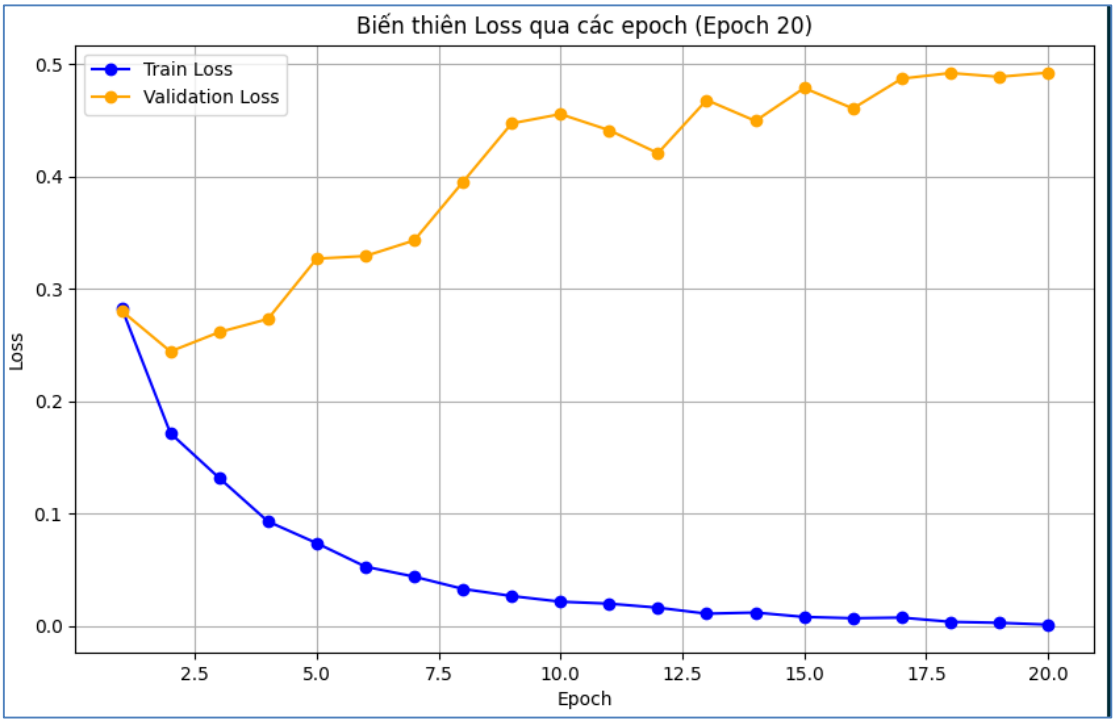
Kết luận sơ bộ

Mô hình hiện tại cho thấy khả năng học tốt và ổn định trên tập huấn luyện, tuy nhiên độ chính xác trên tập kiểm định lại không tăng tương ứng, cho thấy dấu hiệu của hiện tượng overfitting – tức là mô hình học quá kỹ dữ liệu huấn luyện nhưng không tổng quát hóa tốt cho dữ liệu mới. Để khắc phục vấn đề này, cần cân nhắc áp dụng các kỹ thuật giảm overfitting như: early stopping (dừng huấn luyện sớm khi hiệu suất trên tập kiểm định không còn cải thiện), tăng tỷ lệ dropout để làm ngẫu nhiên một phần kết nối giữa các tầng, sử dụng regularization (ví dụ L2) để giảm độ phức tạp của mô hình, hoặc thử nghiệm với một mô hình đơn giản hơn nhằm tìm được cấu trúc phù hợp và hiệu quả hơn với tập dữ liệu hiện có.

3.2.2. Độ mất mát (Loss) trong huấn luyện

Bên cạnh độ chính xác, hàm mất mát (loss) là một chỉ số then chốt giúp đánh giá mức độ sai lệch giữa dự đoán của mô hình và nhãn thực tế. Trong nghiên cứu này, hàm mất mát được sử dụng là Cross Entropy Loss, phù hợp cho các bài toán phân loại đa lớp. Giá trị của hàm mất mát càng nhỏ đồng nghĩa với việc mô hình dự đoán càng gần với giá trị mục tiêu.

Biểu đồ dưới đây thể hiện quá trình biến thiên của loss trong 20 epoch huấn luyện đầu tiên, với hai đường biểu diễn tương ứng cho tập huấn luyện và tập kiểm định:



Hình 3.2: Độ mất mát của mô hình qua 20 epoch

Phân tích kết quả

- Độ mất mát trên tập huấn luyện:

+ Có dấu hiệu giảm rõ rệt từ xấp xỉ 0.28 (epoch 1) xuống gần 0.00 (epoch 20).

+ Quá trình giảm đều và ổn định qua từng epoch, cho thấy mô hình học được đặc trưng dữ liệu huấn luyện rất hiệu quả.

+ Đường biểu diễn loss gần như tiến sát 0 ở các epoch sau, phản ánh việc mô hình gần như không còn sai số trên tập huấn luyện, là một dấu hiệu của quá khớp (overfitting).

- Độ mất mát trên tập kiểm định:

+ Ban đầu giảm nhẹ, sau đó bắt đầu tăng trở lại từ khoảng epoch 5–6.

+ Từ epoch 10 trở đi, độ mất mát trong tập kiểm định tăng dần và đạt gần 0.49 tại epoch 20, cao hơn đáng kể so với loss huấn luyện.

+ Sự gia tăng liên tục của độ mất mát trong tập kiểm định, đồng thời với việc độ mất mát trong tập huấn luyện tiếp tục giảm, là dấu hiệu rõ rệt của hiện tượng quá khớp (overfitting) dẫn đến việc mô hình chỉ đang học thuộc.

So sánh với biểu đồ Độ chính xác (Hình 3.1 mục 3.2.1)

Sự tương phản giữa hai chỉ số:

- Độ chính xác của tập kiểm định giữ ở mức ổn định (~95%),

- Trong khi độ mất mát trên tập kiểm định lại tăng dần qua các epoch

Gợi ý rằng mô hình tuy vẫn phân loại được đúng nhãn, nhưng mức độ "tự tin" trong dự đoán bị suy giảm (độ lệch xác suất so với nhãn đúng lớn hơn). Đây là đặc trưng phổ biến trong các mô hình BERT fine-tuning khi huấn luyện quá sâu hoặc quá lâu trên tập dữ liệu nhỏ hoặc vừa.

Nhận xét và đề xuất

Kết quả này cho thấy mô hình có khả năng ghi nhớ tốt dữ liệu huấn luyện nhưng cũng bắt đầu thể hiện dấu hiệu quá khớp trên tập kiểm định sau khoảng epoch thứ 6–7. Một số hướng cải thiện được đề xuất:

- Áp dụng kỹ thuật dừng quá trình huấn luyện mô hình trước khi hoàn tất tất cả epoch (early stopping) dựa trên độ lệch độ mất mát trên tập kiểm định;

- Tăng cường “bỏ qua” một số neuron (dropout rate) hoặc sử dụng regularization (L2) – thêm vào hàm mất mat một phần phạt (penalty);

- Sử dụng thêm kỹ thuật tăng cường dữ liệu (data augmentation) cho tiếng Việt để mở rộng tính đa dạng dữ liệu huấn luyện.

3.3. Đánh giá mô hình trên tập kiểm tra

Sau khi huấn luyện hoàn chỉnh, mô hình PhoBERT với điểm mốc tốt nhất được chọn dựa trên điểm F1-score cao nhất trên tập kiểm định. Mô hình này được sử dụng để đánh giá cuối cùng trên tập kiểm tra (test set) – một tập dữ liệu độc lập, chưa từng được sử dụng trong quá trình huấn luyện hoặc tinh chỉnh mô hình.

3.3.1. Báo cáo chi tiết theo từng lớp cảm xúc

Kết quả đánh giá mô hình trên tập kiểm tra được trình bày dưới dạng báo cáo phân loại (classification report), bao gồm các chỉ số chính: precision, recall, và F1-score cho từng lớp cảm xúc. Cụ thể:

Bảng 3.2: Kết quả đánh giá mô hình huấn luyện trên tập kiểm tra

Lớp cảm xúc	Precision	Recall	F1-score	Số mẫu (test)
0-Tiêu cực	0.93	0.97	0.95	1.409
1-Trung tính	0.97	0.92	0.94	1.186
2-Tích cực	0.95	0.95	0.95	1.590
Trung bình cộng	0.95	0.95	0.95	4.185

Độ chính xác tổng thể (Accuracy): 95%

Bảng kết quả đánh giá mô hình trên tập kiểm tra (test set) cho thấy mô hình PhoBERT kết hợp Linear đạt hiệu suất cao và ổn định ở cả ba lớp cảm xúc. Cụ thể:

- Lớp "Tiêu cực" (0) đạt Precision = 0.93, Recall = 0.97, và F1-score = 0.95, cho thấy mô hình có khả năng nhận diện tốt các phản hồi mang tính tiêu cực, với tỷ lệ bỏ sót thấp (Recall cao).

- Lớp "Trung tính" (1) có Precision = 0.97, cho thấy mô hình ít nhầm lẫn các phản hồi khác là trung tính. Tuy nhiên, Recall = 0.92 lại thấp hơn một chút, nghĩa là có một số phản hồi trung tính bị mô hình gán nhãn sai (thường bị nhầm với tích cực hoặc tiêu cực). Điều này phản ánh đúng thực tế rằng các phản hồi trung tính thường mang ngữ nghĩa mơ hồ, khó phân biệt rõ ràng.

- Lớp "Tích cực" (2) duy trì hiệu suất cân bằng ở cả ba chỉ số (Precision = 0.95, Recall = 0.95, F1-score = 0.95), thể hiện sự ổn định trong việc nhận diện các phản hồi tích cực.

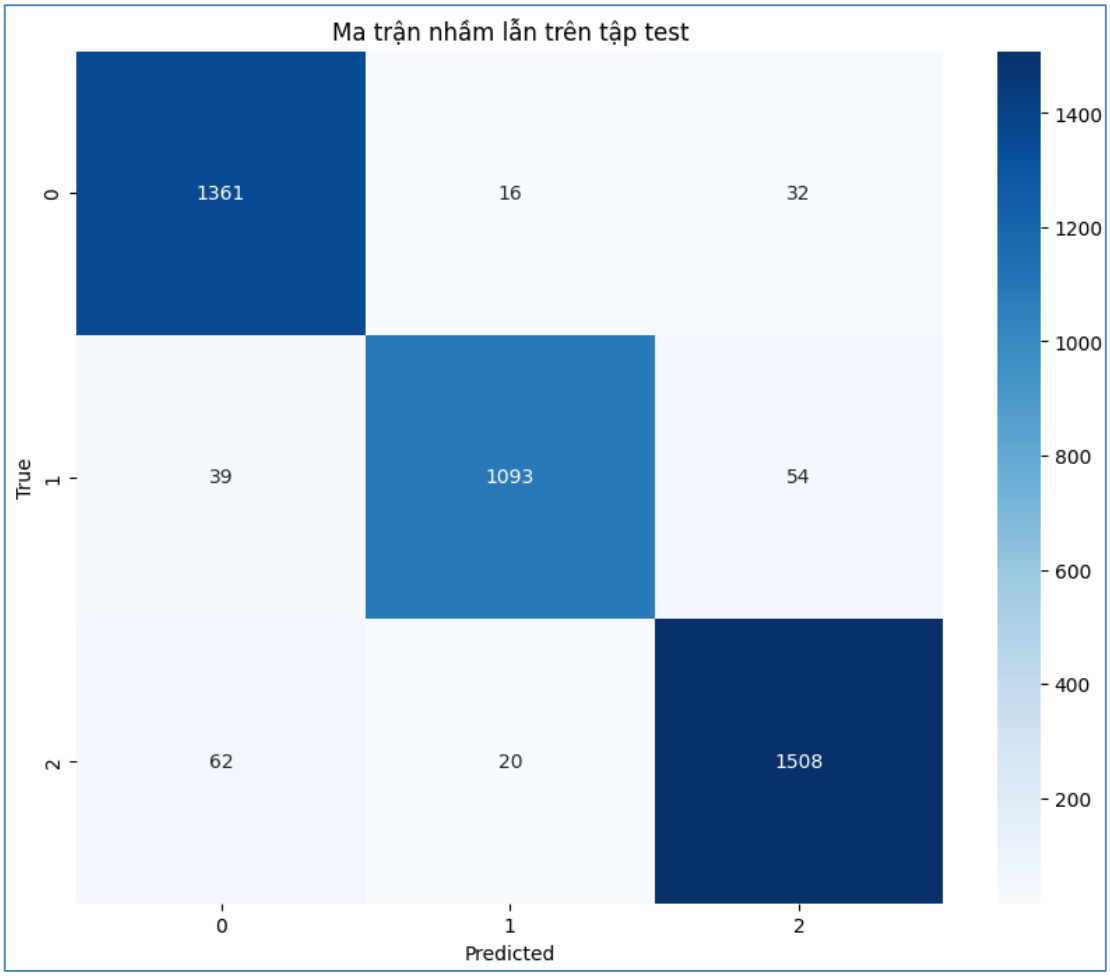
- Trung bình cộng (macro-average) cho ba lớp đạt 0.95 ở cả ba chỉ số đánh giá. Điều này cho thấy mô hình không chỉ đạt hiệu suất cao mà còn có độ cân bằng tốt giữa các lớp, tránh hiện tượng mô hình thiên lệch về một nhãn cụ thể.

Tổng số mẫu trong tập kiểm tra là 4.185, với phân phối lớp khá đồng đều. Điều này góp phần giúp đánh giá mô hình khách quan, không bị ảnh hưởng nhiều bởi mất cân bằng dữ liệu.

Nhận xét: Mô hình PhoBERT kết hợp Linear classifier thể hiện độ chính xác và khả năng khái quát hóa tốt trên tập kiểm tra. Điểm F1 cao ở cả ba lớp cho thấy mô hình không chỉ dự đoán đúng mà còn duy trì sự ổn định giữa Precision và Recall. Tuy nhiên, kết quả của lớp "trung tính" cũng gợi ý rằng mô hình có thể hưởng lợi từ việc mở rộng tập dữ liệu huấn luyện với nhiều ví dụ trung lập rõ ràng hơn, nhằm cải thiện khả năng phân biệt các phản hồi mơ hồ.

3.3.2. Ma trận nhầm lẫn

Để phân tích sâu hơn hiệu suất mô hình, một ma trận nhầm lẫn (confusion matrix) được xây dựng để chỉ ra cách mô hình phân loại từng nhãn cảm xúc**:**



Hình 3.3: Ma trận nhầm lẫn của mô hình PhoBERT

Diễn giải:

- Đường chéo chính đại diện cho các dự đoán chính xác (True Positives).

- Ô ngoài đường chéo biểu thị các dự đoán sai:

- Số lượng dự đoán đúng (theo đường chéo chính):

Cảm xúc thực tế	Dự đoán đúng	Ý nghĩa
0 (tiêu cực)	1361	1361 văn bản tiêu cực được dự đoán đúng
1 (trung tính)	1093	1093 văn bản trung tính được đoán đúng
2 (tích cực)	1508	1508 văn bản trung tính được đoán đúng

Mô hình có hiệu suất cao với cả 3 lớp, đặc biệt là lớp 2 (tích cực).

- Các lỗi dự đoán (off-diagonal)

Cảm xúc thật	Dự đoán sai thành lớp	Số lượng	Nhận xét
0	1	16	Một số tiêu cực bị nhầm là trung tính
0	2	32	Bị nhầm thành tích cực (tác động tiêu cực)
1	0	39	Trung tính bị nhầm là tiêu cực
1	2	54	Trung tính bị nhầm là tích cực
2	0	62	Tích cực bị nhầm thành tiêu cực (ảnh hưởng lớn)
2	1	20	Tích cực bị nhầm là trung tính

Nhận xét:

Lỗi nghiêm trọng nhất là 62 văn bản tích cực bị nhầm thành tiêu cực, dễ gây hiểu nhầm lớn về cảm xúc người dùng.

Lớp trung tính có tỷ lệ nhầm lẫn nhiều hơn cả, đặc biệt là bị nhầm lẫn sang hai chiều (với lớp 0 và 2), đúng với vấn đề mất cân bằng dữ liệu trung tính như bạn đề cập trước đó.

Kết luận

Một số trường hợp trung tính bị nhầm sang tiêu cực hoặc tích cực, phản ánh sự khó phân biệt trong nhóm phản hồi trung dung.

Các nhầm lẫn giữa "tiêu cực" và "tích cực" là rất hiếm, cho thấy mô hình phân biệt rõ ràng hai thái cực cảm xúc này.

3.3.3. Đánh giá tổng thể

Hiệu suất của mô hình PhoBERT trên tập kiểm tra đạt 95% độ chính xác**,** cùng với F1-score cao và cân bằng trên cả ba nhãn cảm xúc. Đây là kết quả ấn tượng đối với một bài toán phân tích cảm xúc tiếng Việt, đặc biệt khi các phản hồi mang tính ngôn ngữ tự nhiên, chủ quan và có yếu tố ẩn dụ, biểu cảm đặc thù văn hóa.

Kết quả này cũng phản ánh tính ổn định của mô hình trên dữ liệu mới, khẳng định khả năng tổng quát hóa và tiềm năng ứng dụng trong các hệ thống phản hồi thông minh trong môi trường giáo dục.

3.4. Đánh giá tổng quan và nhận xét

Sau quá trình huấn luyện, kiểm định và đánh giá trên tập dữ liệu kiểm tra, mô hình PhoBERT fine-tuned cho bài toán nhận diện cảm xúc tiếng Việt đã đạt được kết quả khả quan cả về hiệu suất định lượng lẫn khả năng khái quát hóa trên dữ liệu mới.

Hiệu suất tổng thể

Dựa trên các chỉ số được trình bày tại mục 3.3, có thể tóm lược hiệu quả mô hình như sau:

- Độ chính xác (Accuracy) trên tập kiểm tra đạt 95%, cho thấy mô hình có khả năng phân loại đúng phần lớn văn bản phản hồi.

- F1-score trung bình đạt 0.95, phản ánh sự cân bằng tốt giữa độ chính xác (precision) và độ bao phủ (recall) trong từng lớp cảm xúc.

- Precision cao ở lớp trung tính (0.97) nhưng recall thấp hơn (0.92) cho thấy mô hình vẫn còn thách thức trong việc phát hiện đầy đủ các văn bản có sắc thái trung tính.

Phân tích từ đồ thị và ma trận nhầm lẫn

Biểu đồ độ chính xác cho thấy độ chính xác huấn luyện tăng liên tục và đạt ngưỡng gần tuyệt đối sau 20 epoch, trong khi độ chính xác trên tập kiểm định giữ ở mức ổn định (~95%). Điều này phản ánh quá trình hội tụ tốt nhưng đồng thời cũng gợi ý mô hình có thể đã quá khớp (overfitting) ở giai đoạn sau.

Độ mất mát trên tập kiểm định tăng đều từ khoảng epoch thứ 6 trở đi trong khi độ mất mát trên tập huấn luyện tiếp tục giảm cho thấy dấu hiệu điển hình của quá khớp (overfitting). Tuy nhiên, ảnh hưởng này không nghiêm trọng do hiệu suất trên tập kiểm tra vẫn rất cao.

Ma trận nhầm lẫn cho thấy phần lớn các dự đoán tập trung đúng vào các nhãn thực tế (đường chéo chính), trong đó hai lớp "tiêu cực" và "tích cực" được phân biệt rất tốt. Nhầm lẫn chủ yếu xảy ra ở lớp "trung tính", cho thấy tính chất ngôn ngữ và ngữ nghĩa mơ hồ của lớp này vẫn là thách thức với mô hình.

So sánh với mục tiêu đề tài và các mô hình đã được nghiên cứu

Bảng 3.3: Tổng kết đánh giá mô hình BERT

Tổng kết đánh giá	Kết quả đạt được
Accuracy (test set)	95%
F1-score (trung bình)	0.95
Hiện tượng overfitting	Có, nhưng kiểm soát được
Mức độ phân biệt cảm xúc	Cao
Tính ứng dụng thực tiễn	Rất tốt

Mô hình thực nghiệm sử dụng trong nghiên cứu kết hợp PhoBERT và Linear đã đạt được hiệu suất nổi bật: độ chính xác 95% và F1-score trung bình đạt 0.95 trên tập kiểm thử. Điều này không chỉ thể hiện khả năng học mạnh mẽ từ tập huấn luyện, mà còn cho thấy tính tổng quát hóa tốt trên dữ liệu chưa từng thấy, ngay cả trong môi trường thực tế như phân tích cảm xúc tiếng Việt trong giáo dục.

So với các mô hình truyền thống trong các nghiên cứu trước đây, kết quả đạt được thể hiện sự vượt trội rõ rệt. Cụ thể, mô hình sử dụng Word2Vec kết hợp CNN của Dat Quoc Nguyen và Anh Tuan Nguyen chỉ đạt độ chính xác 86% trong bài toán phân loại cảm xúc tiếng Việt [1]. Trong khi đó, Le và Nguyen triển khai mô hình BiLSTM trên cùng lĩnh vực, nhưng độ chính xác cũng chỉ dừng ở mức 89% [8]. Cả hai mô hình đều sử dụng các kỹ thuật truyền thống dựa trên embedding tĩnh, vốn không thể hiện tốt ngữ cảnh như các mô hình ngôn ngữ tiền huấn luyện hiện đại.

Ngược lại, PhoBERT – được huấn luyện đặc thù cho tiếng Việt – khi kết hợp với Linear đơn giản trong nghiên cứu này vẫn đạt hiệu năng cao, đồng thời giảm thiểu chi phí tính toán và thời gian huấn luyện so với các mô hình phức tạp như Transformer đầy đủ. Điều này khẳng định tiềm năng ứng dụng rộng rãi của PhoBERT trong các bài toán xử lý ngôn ngữ tự nhiên tiếng Việt, đồng thời cho thấy việc fine-tuning mô hình ngôn ngữ tiền huấn luyện là một hướng đi hiệu quả và thiết thực.

3.5. Mô hình thử nghiệm

Sau khi xây dựng và đánh giá thành công mô hình PhoBERT cho bài toán nhận diện cảm xúc tiếng Việt, đề tài hướng đến việc ứng dụng mô hình vào thực tiễn thông qua một phần mềm có tính tương tác cao. Mục tiêu chính là triển khai áp dụng mô hình trên vào một ứng dụng tự xây trên desktop mang tên *EduSentiment* – phục vụ quản lý và phân tích phản hồi sinh viên trong các hoạt động giảng dạy, học tập tại các trường đại học, cao đẳng.

Ứng dụng nhằm hiện thực hóa các chức năng sau:

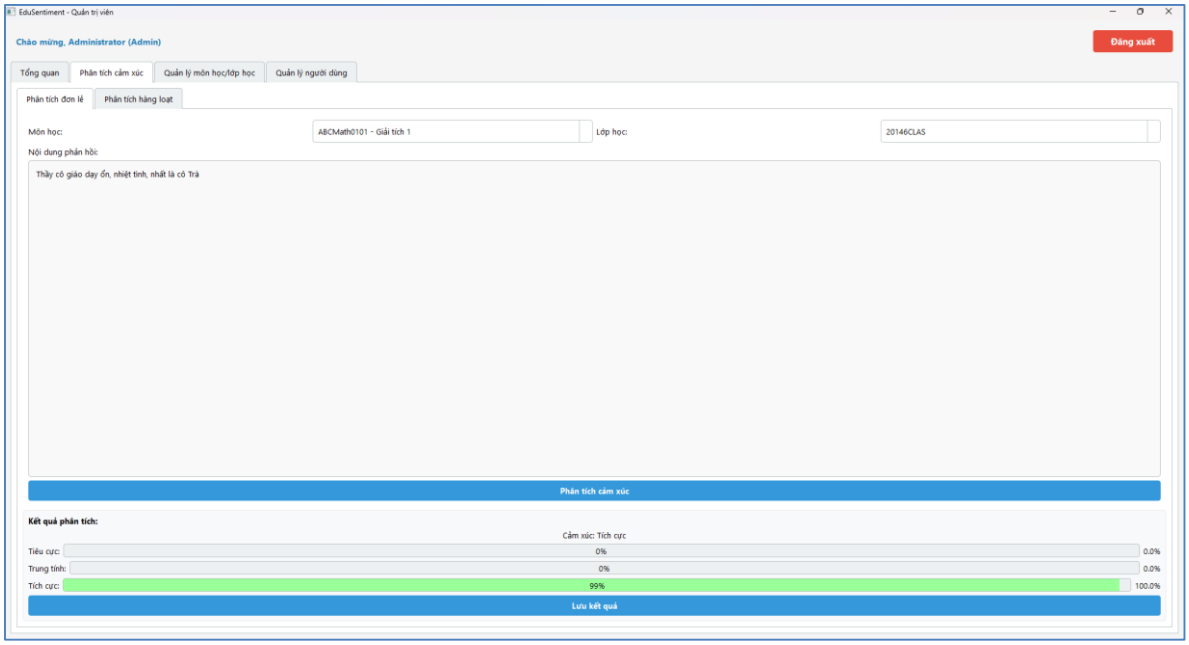
- Tự động phân tích cảm xúc phản hồi sinh viên theo ba cấp độ: tiêu cực, trung tính, tích cực.

- Hỗ trợ quản lý lớp học, môn học và người dùng, với phân quyền rõ ràng (quản trị viên – sinh viên).

- Lưu trữ toàn bộ kết quả phản hồi và phân tích cảm xúc theo thời gian, cho phép thống kê và truy xuất linh hoạt.

- Trực quan hóa dữ liệu phản hồi qua biểu đồ, phục vụ công tác cải tiến chất lượng giảng dạy, đánh giá môn học, và chăm sóc sinh viên.

Việc phát triển ứng dụng giúp mô hình học sâu không chỉ dừng lại ở nghiên cứu lý thuyết mà còn góp phần giải quyết một bài toán thực tiễn trong lĩnh vực giáo dục bằng công nghệ AI.



Hình 3.4: Giao diện phân tích cảm xúc của EduSentiment

3.5.1. Tích hợp mô hình PhoBERT vào phần mềm

Một trong những điểm cốt lõi tạo nên tính thông minh cho hệ thống EduSentiment là khả năng tích hợp trực tiếp mô hình học sâu PhoBERT vào phần mềm desktop để phân tích cảm xúc theo thời gian thực. Việc tích hợp này được thực hiện hoàn toàn cục bộ (offline), không phụ thuộc vào kết nối mạng hoặc dịch vụ bên ngoài.

a. Tải mô hình và tokenizer

Ngay khi ứng dụng khởi động, lớp ModelManager sẽ thực hiện quá trình:

- Tải tokenizer và mô hình đã huấn luyện từ thư viện transformers (HuggingFace)

- Chuyển mô hình sang thiết bị tính toán (CPU hoặc GPU)

- Mô hình được đặt ở chế độ suy luận (eval()) để tăng tốc và tiết kiệm bộ nhớ.

b. Tiền xử lý văn bản đầu vào

Trước khi đưa văn bản vào mô hình, phản hồi từ người dùng (sinh viên hoặc admin nhập) được chuẩn hóa bằng hàm normalize_text():

- Loại bỏ ký tự đặc biệt, dấu câu, số và biểu tượng.

- Chuyển toàn bộ văn bản thành chữ thường.

- Xóa khoảng trắng thừa và chuẩn Unicode.

c. Ưu điểm của tích hợp cục bộ

Việc tích hợp mô hình trực tiếp vào phần mềm mang lại nhiều lợi ích:

- Không phụ thuộc Internet: có thể hoạt động trong môi trường nội bộ của trường học.

- Bảo mật dữ liệu: không truyền văn bản sinh viên ra ngoài hệ thống.

- Hiệu năng thời gian thực: phản hồi phân tích gần như ngay lập tức.

- Mở rộng linh hoạt: dễ thay thế mô hình khác (như PhoBERT-LSTM) trong tương lai.

3.5.2. Cơ sở dữ liệu và tổ chức lưu trữ

Để đảm bảo hệ thống có thể lưu trữ, quản lý và truy xuất hiệu quả dữ liệu phản hồi, người dùng và kết quả phân tích cảm xúc, ứng dụng EduSentiment sử dụng cơ sở dữ liệu nhẹ SQLite3, tích hợp trực tiếp trong ứng dụng, không yêu cầu cài đặt thêm phần mềm máy chủ.

a.. Cấu trúc cơ sở dữ liệu

Cơ sở dữ liệu được khởi tạo khi ứng dụng chạy lần đầu, với các bảng chính sau:

Bảng 3.4: Chức năng từng bảng dữ liệu

Bảng	Mô tả
User	Lưu thông tin người dùng hệ thống: quản trị viên và sinh viên
Courses	Danh sách các môn học có thể phản hồi
Classes	Lớp học cụh thể trong từng môn học
Feedback	Phản hồi văn bảng từ sinh viên, kèm kết quả cảm xúc và thông tin

	liên quan
Enrollments	(Tùy chọn mở rộng), ánh xạ giữa sinh viên và lớp học

b. Thiết kế cơ sở dữ liệu và ràng buộc

Bảng 3.5: User – Thông tin người dùng

Tên trường	Kiểu dữ liệu	Ràng buộc	Mô tả
Id	INTEGER	PRIMARY	ID người dùng (tự tăng)
Username	TEXT	NOT NULL, UNIQUE	Tên đăng nhập duy nhất
Password	TEXT	NOT NULL	Mật khẩu mã hóa SHA-256
Full_name	TEXT	NOT NULL	Họ và tên
Email	TEXT	NULL	Địa chỉ email
Role	TEXT	CHECK (‘admin’, ‘student’)	Vai trò người dùng
Student_id	TEXT	NULL	Mã sinh viên (nếu có)
Created_at	TIMESTAMP	DEFAULT CURRENT_TIMESTAMP	Ngày tạo tài khoản
Last_login	TIMESTAMP	NULL	Lần đăng nhập gần nhất

Bảng 3.6: Courses – Danh sách môn học

Tên trường	Kiểu dữ liệu	Ràng buộc	Mô tả
Id	INTEGER	PRIMARY KEY, AUTOINCREMENT	ID Môn học
Name	TEXT	NOT NULL	Tên môn học
code	TEXT	UNIQUE, NOT NULL	Mã môn học
description	TEXT	NULL	Mô tả môn học

Bảng 3.7: Classes – Danh sách lớp học

Tên trường	Kiểu dữ liệu	Ràng buộc	Mô tả
Id	INTEGER	PRIMARY KEY, AUTOINCREMENT	ID lớp học
Name	TEXT	NOT NULL	Tên lớp học
Course_id	INTEGER	FOREIGN KEY -> courses(id)	Thuộc môn học nào
Semester	TEXT	NULL	Học kỳ (nếu có)
Year	INTEGER	NULL	Năm học (nếu có)

Bảng 3.8: Feedback - Phản hồi và kết quả phân tích cảm xúc

Tên trường	Kiểu dữ liệu	Ràng buộc	Mô tả
Id	INTEGER	PRIMARY KEY, AUTOINCREMENT	ID phản hồi

Content	TEXT	NOT NULL	Nội dung phản hồi văn bảng
Sentiment	INTEGER	CHECK (0,1,2), NOT NULL	Nhãn cảm xúc (0: Tiêu cực, 1: Trung tính, 2:Tích cực)
Neg_prob	REAL	NULL	Xác suất lớp tích cực
Neu_prob	REAL	NULL	Xác suất lớp trung tính
Pos_prob	REAL	NULL	Xác suất lớp tích cực
User_id	INTEGER	FOREIGN KEY ->users(id), NULLABLE	Người gửi (ẩn danh nếu NULL)
Course_id	INTEGER	FOREIGN KEY -> courses(id)	Gắn với môn học
Class_id	INTEGER	FOREIGN KEY -> classes(id)	Gắn với lớp học
Created_at	TIMESTAMP	DEFAULT CURRENT_TIMETSTAMP	Ngày gửi phản hồi
Is_anonymous	BOOLEAN	DEFAULT 0	Gửi ẩn danh hay không

III. KẾT LUẬN

1. Các kết quả đạt được của đề án tốt nghiệp

Trong khuôn khổ đề án, sinh viên đã triển khai một quy trình đầy đủ và có hệ thống, bao gồm khảo sát tài liệu, xây dựng tập dữ liệu, thiết kế kiến trúc mô hình, huấn luyện và đánh giá hiệu suất, đồng thời kiểm chứng khả năng ứng dụng trong thực tiễn. Các kết quả cụ thể như sau:

- Xây dựng thành công mô hình nhận diện cảm xúc tiếng Việt dựa trên kiến trúc PhoBERT (vinai/phobert-base), một mô hình ngôn ngữ tiền huấn luyện được phát triển tối ưu hóa cho tiếng Việt. Mô hình này được lựa chọn và tinh chỉnh nhằm phù hợp với đặc điểm cú pháp và ngữ nghĩa của ngôn ngữ đơn âm.

- Tập dữ liệu được sử dụng là bộ phản hồi của học sinh Việt Nam (Vietnamese Students Feedback). Dữ liệu đầu vào được xử lý, chuẩn hóa và cân bằng giữa các nhãn cảm xúc thông qua nhiều phương pháp sinh mẫu trung tính như: trích xuất từ các câu có ý nghĩa trái chiều, xây dựng mẫu thủ công, và sử dụng các mẫu sinh từ khuôn mẫu (template).

- Quá trình huấn luyện được tối ưu hóa với nhiều kỹ thuật hiện đại:

+ Áp dụng huấn luyện với độ chính xác hỗn hợp (mixed precision – FP16) nhằm tăng tốc độ và tiết kiệm bộ nhớ;

+ Sử dụng trình tối ưu hóa AdamW cùng với lịch trình giảm tốc độ học tuyến tính có khởi động (linear scheduler with warmup) để đảm bảo quá trình huấn luyện ổn định;

+ Thực hiện theo dõi liên tục các chỉ số quan trọng như độ mất mát (loss), độ chính xác (accuracy), và điểm F1-score trên tập validation; đồng thời lưu lại trọng số mô hình tốt nhất trong quá trình huấn luyện.

- Mô hình đạt hiệu suất cao trên tập dữ liệu kiểm thử, đặc biệt là khả năng phân biệt hiệu quả ba nhãn cảm xúc (tích cực, tiêu cực, trung tính). Kết quả cho thấy độ chính xác tổng thể và điểm F1 ở mức đáng tin cậy, với sai số chủ yếu tập trung ở lớp "trung tính" – điều này phản ánh đặc trưng phổ biến của tiếng Việt, khi nhiều biểu đạt cảm xúc mang tính mơ hồ hoặc khó phân loại rõ ràng.

2. Nhận xét, đề xuất và khuyến nghị

2.1. Nhận xét

Đề án đã hoàn thành đầy đủ các mục tiêu nghiên cứu đã đề ra, bao gồm cả khía cạnh xây dựng mô hình học sâu áp dụng cho tiếng Việt và khả năng triển khai vào các tình huống thực tế. Mô hình PhoBERT cho thấy khả năng biểu diễn ngữ nghĩa mạnh mẽ, đặc biệt hiệu quả khi xử lý các văn bản ngắn, chứa nhiều ngữ nghĩa tiềm ẩn như phản hồi của sinh viên. Kết quả thực nghiệm chứng minh rằng mô hình có thể nhận diện cảm xúc với độ chính xác cao, đáp ứng yêu cầu ứng dụng trong môi trường giáo dục.

2.2. Đề xuất

Trong các nghiên cứu tiếp theo, cần tăng cường thu thập và tích hợp nhiều dữ liệu phản hồi trung tính từ thực tế để cải thiện khả năng phân biệt giữa ba nhóm cảm xúc, đặc biệt là trong các trường hợp mơ hồ hoặc thiếu rõ ràng. Ngoài ra, hướng phát triển tiềm năng là mở rộng mô hình phân tích không chỉ dừng lại ở cảm xúc tổng thể, mà tiến tới phân tích cảm xúc theo từng cụm chủ đề cụ thể như: nội dung môn học, phương pháp giảng dạy, giảng viên, hoặc cơ sở vật chất. Điều này sẽ nâng cao tính ứng dụng và giá trị phân tích của hệ thống trong bối cảnh quản lý giáo dục.

2.3. Hướng nghiên cứu tiếp theo

Trong tương lai, đề tài có thể được tiếp tục phát triển theo nhiều hướng nhằm nâng cao hiệu quả và mở rộng phạm vi ứng dụng. Trước hết, việc kết hợp PhoBERT với các kiến trúc mạng sâu khác như Attention hoặc Bi-LSTM sẽ giúp tăng cường khả năng nhận diện ngữ cảnh sâu, đặc biệt trong các văn bản dài hoặc chứa nhiều tầng cảm xúc phức tạp. Bên cạnh đó, hệ thống phân loại cảm xúc cũng cần được mở rộng để bao gồm nhiều trạng thái cụ thể hơn ngoài ba nhãn cơ bản "tích cực", "tiêu cực" và "trung tính". Việc bổ sung các nhãn như "hài lòng", "bối rối" hay "phẫn nộ" sẽ cho phép mô hình nhận diện sắc thái cảm xúc chi tiết hơn, phục vụ tốt hơn cho các phân tích chuyên sâu trong giáo dục và xã hội. Ngoài ra, đề tài có thể tích hợp chức năng phân tích xu hướng cảm xúc theo thời gian (ví dụ theo từng tuần, học kỳ), từ đó hỗ trợ các đơn vị đào tạo theo dõi và cải tiến nội dung giảng dạy một cách linh hoạt. Đồng thời, hệ thống có thể được phát triển thành một ứng dụng web thay vì chạy cục bộ (local), nhằm phục vụ quy mô người dùng lớn hơn, cho phép truy cập từ nhiều thiết bị và triển khai trong môi trường thực tế. Cuối cùng, mô hình cũng

có tiềm năng ứng dụng rộng rãi trong các lĩnh vực khác như chăm sóc khách hàng, phân tích đánh giá sản phẩm hoặc khảo sát dư luận xã hội, với điều chỉnh phù hợp về dữ liệu và ngữ cảnh sử dụng. Các hướng mở rộng này không chỉ giúp nâng cao giá trị học thuật mà còn tăng tính thực tiễn của đề tài trong bối cảnh chuyển đổi số hiện nay.

TÀI LIỆU THAM KHẢO

[1] Nguyen D.Q., Nguyen A.T. – 2020 - PhoBERT: Pre-trained language models for Vietnamese. *Findings of the Association for Computational Linguistics*

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. – 2018 - BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding

[3] Bo Pang, Lillian Lee – 2004 – A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts

[4] Phan, H., et al. – 2020 – PhoBERT: Pretrained Language Models for Vietnamese

[5] S. Ruder – 2019 – How to Train Your BERT: A Detailed Guide

[6] Hugging Face – 2020 – Transformers Documentation

[7] Phạm Văn Nam. (2019). Ứng dụng mô hình học sâu trong phân loại văn bản tiếng Việt. Tạp chí Khoa học Công nghệ.

[8] T. Q. Le and H. L. Nguyen – 2021 - Applying BiLSTM for Vietnamese Sentiment Classification - *Journal of Computer Science and Cybernetics*

[9] Wikipedia – Tiếng Việt - https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Vi%E1%BB%87t.

[10] Nguyễn Thanh Huy – 2022 – Nhận diện cảm xúc trong văn bản Tiếng Việt trong mô hình máy học - https://ptithcm.edu.vn/wp-content/uploads/2023/07/2020_HTTT_NguyenThanhHuy_TTLV.pdf

[11] Nguyễn Chiến Thắng – 2020 – Thử nhận diện cảm xúc văn bản Tiếng Việt với phoBert - https://miai.vn/2020/12/29/bert-series-chuong-3-thu-nhan-dien-cam-xuc-van-ban-tieng-viet-voi-phobert-cach-1/

[10] Cambria, E., &White,B. - 2014 - Jumping NLP Curves: A Review of Natural Language Processing Research

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung đề án qua phần mềm https://kiemtratailieu.vn một cách trung thực và đạt kết quả mức độ tương đồng 11% toàn bộ nội dung đề án. Đề án kiểm tra qua phần mềm là bản cứng đề án đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

……., ngày….. tháng…..năm……

HỌC VIÊN CAO HỌC