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

I. MỞ ĐẦU

1. Lý do chọn đề tài

Trong thời đại số hiện nay, việc hiểu và phân tích cảm xúc của con người thông qua văn bản không chỉ mang tính học thuật mà còn có ý nghĩa ứng dụng thực tiễn sâu sắc trong nhiều lĩnh vực như thương mại điện tử, truyền thông, chăm sóc sức khỏe tinh thần, và đặc biệt là giáo dục. Trong môi trường học đường, nhận diện cảm xúc từ văn bản của học sinh – bao gồm bài viết, phản hồi, hoặc nhật ký học tập – đóng vai trò quan trọng trong việc giúp giáo viên, chuyên viên tư vấn và nhà quản lý giáo dục nắm bắt được trạng thái tâm lý, mức độ hài lòng, hay các dấu hiệu tiêu cực để có những điều chỉnh và can thiệp kịp thời.

Tuy nhiên, phần lớn các nghiên cứu hiện tại trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và nhận diện cảm xúc chủ yếu tập trung vào các ngôn ngữ có tài nguyên phong phú như tiếng Anh. Tiếng Việt – với đặc trưng là ngôn ngữ đơn âm, ngữ pháp linh hoạt, hệ thống dấu thanh phong phú và cách biểu đạt giàu hình tượng – lại chưa được khai thác và hỗ trợ tương xứng. Các mô hình ngôn ngữ đa ngữ như Multilingual BERT không đủ khả năng nắm bắt sâu sắc các đặc điểm ngôn ngữ riêng biệt của tiếng Việt, dẫn đến hiệu quả xử lý thấp trong thực tế.

PhoBERT – một mô hình ngôn ngữ tiền huấn luyện dành riêng cho tiếng Việt – được phát triển bởi VinAI Research, là một bước tiến quan trọng trong việc khắc phục những hạn chế trên. Được huấn luyện trên tập dữ liệu văn bản tiếng Việt có quy mô lớn, PhoBERT có khả năng hiểu ngữ cảnh và biểu diễn ngôn ngữ tiếng Việt một cách hiệu quả. Khi kết hợp với bộ phân loại tuyến tính (Linear Classifier), mô hình có thể thực hiện nhận diện cảm xúc một cách đơn giản, nhanh chóng, nhưng vẫn đạt độ chính xác cao, phù hợp với các ứng dụng thực tiễn trong giáo dục.

Do đó, đề tài "Nhận diện cảm xúc trong văn bản tiếng Việt bằng mô hình học sâu" được lựa chọn nhằm xây dựng và thử nghiệm mô hình PhoBERT kết hợp Linear Classifier trong bối cảnh giáo dục, hướng tới việc phát triển một công cụ hỗ trợ giáo vi

2. Tổng quan vấn đề nghiên cứu

Nhận diện cảm xúc (Sentiment Analysis) là một trong những hướng nghiên cứu quan trọng của NLP, với mục tiêu xác định thái độ, quan điểm hoặc trạng thái cảm xúc được biểu đạt trong văn bản. Trong những năm gần đây, lĩnh vực này đã phát triển mạnh mẽ nhờ vào sự

tiến bộ của công nghệ học sâu (deep learning), đặc biệt là các mô hình ngôn ngữ tiền huấn luyện như BERT, GPT, RoBERTa.

Trong giáo dục, nhận diện cảm xúc từ văn bản học sinh đóng vai trò quan trọng trong việc:

- Phát hiện sớm các dấu hiệu căng thẳng, trầm cảm, hoặc thiếu động lực học tập;
- Cải thiện tương tác giữa giáo viên và học sinh;
- Hỗ trợ tư vấn tâm lý học đường;
- Cung cấp dữ liệu để phân tích và cải tiến phương pháp giảng dạy.

Tuy nhiên, tiếng Việt là một ngôn ngữ có nhiều thách thức trong xử lý tự động, bao gồm:

- Sự phụ thuộc mạnh vào ngữ cảnh để hiểu đúng nghĩa của từ hoặc cụm từ;
- Sự linh hoạt trong cấu trúc câu khiến việc phân tích cú pháp trở nên phức tạp;
- Hệ thống dấu thanh phong phú có thể thay đổi hoàn toàn nghĩa của từ;
- Sự phổ biến của văn bản không chuẩn trên mạng xã hội, diễn đàn (viết tắt, sai chính tả, ngôn ngữ teencode);
- Thiếu hụt bộ dữ liệu được gán nhãn cảm xúc đủ lớn và chất lượng cao cho tiếng Việt.

Mô hình PhoBERT, với kiến trúc Transformer cải tiến từ BERT, được huấn luyện hoàn toàn bằng dữ liệu tiếng Việt, đã chứng minh hiệu quả cao trong nhiều tác vụ NLP như phân loại văn bản, nhận diện thực thể, phân tích cú pháp. Khi kết hợp với một tầng phân loại tuyến tính đơn giản, mô hình có thể được áp dụng để phân loại cảm xúc văn bản một cách hiệu quả, dễ huấn luyện và triển khai.

3. Mục đích nghiên cứu

Mục đích chính của đề tài là xây dựng và đánh giá một mô hình học sâu sử dụng PhoBERT kết hợp Linear Classifier nhằm thực hiện phân loại cảm xúc văn bản tiếng Việt, đặc biệt trong lĩnh vực giáo dục.

Các mục tiêu cụ thể bao gồm:

- Phân tích, tổng hợp lý thuyết liên quan đến nhận diện cảm xúc và xử lý ngôn ngữ tiếng Việt;

- Xây dựng tập dữ liệu cảm xúc tiếng Việt chất lượng từ phản hồi của học sinh, giáo viên, phụ huynh;
- Tiền xử lý dữ liệu, gán nhãn cảm xúc theo ba lớp: tích cực, tiêu cực, trung tính;
- Huấn luyện mô hình PhoBERT + Linear Classifier, điều chỉnh siêu tham số để tối ưu hóa kết quả;
- Đánh giá hiệu suất mô hình bằng các chỉ số như Độ chính xác, Precision, Recall, F1-score;
- Ứng dụng mô hình vào phần mềm phân tích phản hồi giáo dục, hỗ trợ nhà trường trong việc quản lý và cải thiện chất lượng giảng dạy.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Là các văn bản tiếng Việt chứa nội dung cảm xúc thuộc môi trường giáo dục, bao gồm:

- Bài viết, nhật ký, hoặc câu trả lời tự luận của học sinh;
- Nhận xét, đánh giá từ giáo viên;
- Góp ý, phản hồi từ phụ huynh hoặc người học trên các diễn đàn giáo dục.

Phạm vi nghiên cứu:

- Tập trung vào ba loại cảm xúc chính: tích cực, tiêu cực, trung tính;
- Ngữ liệu: văn bản tiếng Việt thu thập từ môi trường học tập thực tế, bao gồm các bài phản hồi được gán nhãn;
- Kỹ thuật áp dụng: mô hình học sâu PhoBERT, kết hợp với bộ phân loại tuyến tính;
- Thử nghiệm mô hình trong một phần mềm hỗ trợ phân tích cảm xúc phản hồi của sinh viên.

5. Phương pháp nghiên cứu

Đề tài sử dụng phương pháp kết hợp giữa lý thuyết và thực nghiệm:

Phương pháp lý thuyết:

- Nghiên cứu các tài liệu khoa học liên quan đến NLP, BERT, PhoBERT, và nhận diện cảm xúc;

- Phân tích đặc điểm ngôn ngữ tiếng Việt và các mô hình học sâu phù hợp;

- Tham khảo các công trình nghiên cứu quốc tế và trong nước làm cơ sở lý luận và kỹ thuật.

Phương pháp thực nghiệm:

- Thu thập, xử lý và gán nhãn bộ dữ liệu cảm xúc từ văn bản tiếng Việt;

- Cân bằng tập dữ liệu để đảm bảo độ tin cậy trong huấn luyện;

- Huấn luyện mô hình PhoBERT + Linear Classifier, tinh chỉnh siêu tham số;

- Đánh giá kết quả bằng các chỉ số chuẩn;

- Triển khai mô hình vào ứng dụng thực tiễn và đánh giá khả năng vận hành trong môi trường giáo dục.

II. NỘI DUNG

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

1.1. Bối cảnh và tầm quan trọng

1.1.1. Tổng quan về nhận diện cảm xúc

Trong thời đại số hiện nay, việc hiểu và phân tích cảm xúc của con người thông qua văn bản không chỉ mang tính học thuật mà còn có ý nghĩa ứng dụng thực tiễn sâu sắc trong nhiều lĩnh vực như thương mại điện tử, truyền thông, chăm sóc sức khỏe tinh thần, và đặc biệt là giáo dục. Trong môi trường học đường, nhận diện cảm xúc từ văn bản của học sinh – bao gồm bài viết, phản hồi, hoặc nhật ký học tập – đóng vai trò quan trọng trong việc giúp giáo viên, chuyên viên tư vấn và nhà quản lý giáo dục nắm bắt được trạng thái tâm lý, mức độ hài lòng, hay các dấu hiệu tiêu cực để có những điều chỉnh và can thiệp kịp thời.

Tuy nhiên, phần lớn các nghiên cứu hiện tại trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và nhận diện cảm xúc chủ yếu tập trung vào các ngôn ngữ có tài nguyên phong phú như tiếng Anh. Tiếng Việt – với đặc trưng là ngôn ngữ đơn âm, ngữ pháp linh hoạt, hệ thống dấu thanh phong phú và cách biểu đạt giàu hình tượng – lại chưa được khai thác và hỗ trợ tương xứng. Các mô hình ngôn ngữ đa ngữ như Multilingual BERT không đủ khả năng nắm bắt sâu sắc các đặc điểm ngôn ngữ riêng biệt của tiếng Việt, dẫn đến hiệu quả xử lý thấp trong thực tế.

PhoBERT – một mô hình ngôn ngữ tiền huấn luyện dành riêng cho tiếng Việt – được phát triển bởi VinAI Research, là một bước tiến quan trọng trong việc khắc phục những hạn chế trên. Được huấn luyện trên tập dữ liệu văn bản tiếng Việt có quy mô lớn, PhoBERT có khả năng hiểu ngữ cảnh và biểu diễn ngôn ngữ tiếng Việt một cách hiệu quả. Khi kết hợp với bộ phân loại tuyến tính (Linear Classifier), mô hình có thể thực hiện nhận diện cảm xúc một cách đơn giản, nhanh chóng, nhưng vẫn đạt độ chính xác cao, phù hợp với các ứng dụng thực tiễn trong giáo dục.

Do đó, đề tài "Nhận diện cảm xúc trong văn bản tiếng Việt bằng mô hình học sâu" được lựa chọn nhằm xây dựng và thử nghiệm mô hình PhoBERT kết hợp Linear Classifier trong bối cảnh giáo dục, hướng tới việc phát triển một công cụ hỗ trợ giáo vi

1.1.2. Đặc điểm văn bản tiếng Việt

Tiếng Việt là một ngôn ngữ đơn âm, mỗi âm tiết mang một ý nghĩa riêng, kết hợp với hệ thống dấu thanh để tạo ra các từ khác nhau. Ví dụ, "ma" (ma quỷ), "má" (mẹ), "mà" (liên từ), "mã" (mã số), "mạ" (lúa non) có ý nghĩa hoàn toàn khác dù chỉ khác dấu thanh. Ngữ pháp tiếng Việt linh hoạt, không có chia động từ theo thì, khiến ngữ nghĩa phụ thuộc mạnh vào ngữ cảnh và trật tự từ. Ví dụ, "Tôi ăn cơm" và "Cơm tôi ăn" có thể mang ý nghĩa khác tùy ngữ cảnh.

Ngoài ra, văn bản tiếng Việt chứa nhiều thành ngữ, ẩn dụ, và biểu đạt cảm xúc gián tiếp, đặc biệt trong giáo dục. Ví dụ, học sinh có thể viết "Nước mắt cá sấu" để ám chỉ sự giả tạo (tiêu cực) hoặc "Mặt trời nhỏ của em" để bày tỏ sự yêu thương (tích cực). Văn bản không chuẩn trên mạng xã hội hoặc diễn đàn giáo dục càng phức tạp hơn, với lỗi chính tả ("hk" thay "không"), viết tắt ("z" thay "vậy"), và ngôn ngữ lai (kết hợp tiếng Việt và tiếng Anh, ví dụ "Cảm ơn teacher!"). Những đặc điểm này đòi hỏi mô hình NLP phải có khả năng xử lý dấu thanh, hiểu ngữ cảnh, và khái quát hóa trên dữ liệu không chuẩn.

1.1.3. Ứng dụng và thách thức

- Ứng dụng:

+ Giáo dục: Phân tích bài viết học sinh để phát hiện cảm xúc tiêu cực, hỗ trợ tâm lý, ví dụ phát hiện sự căng thẳng từ câu "Em thấy áp lực quá".

+ Thương mại điện tử: Đánh giá sản phẩm, cải thiện dịch vụ khách hàng, ví dụ phân tích "Hàng đẹp, giá hợp lý" (tích cực).

+ Mạng xã hội: Theo dõi dư luận, phát hiện khủng hoảng truyền thông, ví dụ phân tích bình luận về một sự kiện công cộng.

+ Chăm sóc sức khỏe: Phát hiện dấu hiệu trầm cảm hoặc lo âu qua văn bản, ví dụ "Tôi không còn hứng thú với bất cứ điều gì".

- Thách thức:

+ Ngữ nghĩa mơ hồ: "Đẹp thật" có thể là khen hoặc mỉa mai tùy ngữ cảnh.

+ Yếu tố văn hóa: Cảm xúc trong tiếng Việt chịu ảnh hưởng bởi ngữ điệu, quan hệ xã hội, ví dụ "Cảm ơn" có thể mang ý tiêu cực nếu nói với giọng mỉa mai.

+ Dữ liệu hạn chế: Thiếu bộ dữ liệu lớn, chất lượng cao cho tiếng Việt, đặc biệt trong giáo dục.

+ Văn bản không chuẩn: Lỗi chính tả, viết tắt, ngôn ngữ lai trên mạng xã hội và diễn đàn.

1.2. Khái niệm và phân tích cảm xúc

- Cảm xúc được chia thành ba loại chính:

+ Tích cực: Vui vẻ, hài lòng, phấn khởi, ví dụ "Hôm nay em được điểm cao, thật hạnh phúc!".

+ Tiêu cực: Buồn bã, tức giận, thất vọng, ví dụ "Em thấy môn học này quá khó, chán lắm".

+ Trung tính: Không cảm xúc rõ ràng, ví dụ "Hôm nay em học bài mới".

- Nhận diện cảm xúc gồm ba bước:

+ Thu thập dữ liệu: Từ bài viết, phản hồi, bình luận, hoặc khảo sát.

+ Tiền xử lý: Làm sạch (loại ký tự đặc biệt, chuẩn hóa Unicode), sửa lỗi chính tả, gán nhãn cảm xúc.

+ Áp dụng mô hình: Từ phương pháp truyền thống (Logistic Regression, SVM) đến học sâu (LSTM, BERT, PhoBERT).

- Mức độ phân tích:

+ Cấp từ/cụm từ: Phân tích từ "vui" hoặc cụm "rất tệ".

+ Cấp câu: Phân tích câu "Hôm nay tôi rất vui".

+ Cấp tài liệu: Phân tích toàn bộ bài luận hoặc nhật ký.

- Thách thức:

+ Đa nghĩa: Từ "hay" mang nhiều nghĩa tùy ngữ cảnh (thú vị, mỉa mai).

+ Yếu tố văn hóa: Biểu đạt gián tiếp, ví dụ "Cười ra nước mắt" (tiêu cực).

+ Dữ liệu không chuẩn: Viết tắt, lỗi chính tả, ngôn ngữ lai.

1.3. Công trình nghiên cứu liên quan

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

- 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.

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.

- Kết quả Độ chính xác đạt 94.22%, các chỉ số macro Precision/Recall/F1 đều vượt 94%.

1.4. Mô hình học sâu trong NLP

- BERT (Bidirectional Encoder Representations from Transformers): Dựa trên kiến trúc Transformer, sử dụng pre-Huấn luyện với hai nhiệm vụ:

+ Masked Language Model (MLM): Che 15% từ trong câu, yêu cầu mô hình dự đoán dựa trên ngữ cảnh. Ví dụ: "Tôi [MASK] vui" → dự đoán "rất".

+ Next Sentence Prediction (NSP): Phân biệt cặp câu liên tiếp hoặc không liên tiếp, giúp hiểu quan hệ giữa các câu. Ví dụ: “Hôm nay tôi vui.” và “Bài học rất thú vị.” là liên tiếp.

+ Fine-tuning: Tùy chỉnh mô hình cho bài toán cụ thể, như phân loại cảm xúc, bằng cách thêm linear classifier.

+ Hạn chế: Yêu cầu tài nguyên tính toán lớn (GPU, bộ nhớ cao).

- PhoBERT:

+ Phiên bản BERT tối ưu cho tiếng Việt, pre-trained trên bộ dữ liệu lớn (Wikipedia tiếng Việt, báo chí).

+ Token hóa cấp âm tiết: “Học sinh” thành “học_sinh”, phù hợp với đặc trưng đơn âm.

+ Hỗ trợ dấu thanh và ngữ pháp linh hoạt, đạt hiệu suất cao hơn BERT cơ bản trên tiếng Việt.

1.5. Mô hình PhoBERT + Linear Classifier

PhoBERT: Là mô hình ngôn ngữ tiền huấn luyện dành riêng cho tiếng Việt, kế thừa kiến trúc RoBERTa, huấn luyện trên 20GB dữ liệu tiếng Việt. Token hóa sử dụng SentencePiece và RDRSegmenter phù hợp với tiếng Việt đơn âm.

Linear Classifier: Sau khi PhoBERT trích xuất biểu diễn ngữ nghĩa tại token [CLS], vector đầu ra được đưa vào một tầng phân loại tuyến tính để xác định cảm xúc. Ưu điểm là đơn giản, nhanh, hiệu quả, phù hợp với ứng dụng giáo dục.

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

2.1. Chuẩn bị dữ liệu huấn luyện

- Nguồn dữ liệu: Bộ dữ liệu phản hồi của sinh viên bằng Tiếng Việt - Vietnamese Students Feedback, bao gồm:

+ Tập huấn luyện (train) gồm 11.400 dòng dữ liệu

+ Tạp kiểm định (kiểm định) gồm 1.580 dòng dữ liệu

+ Tập kiểm tra (test) gồm 3.170 dòng dữ liệu

+ Tổng cộng: khoảng 16.000 mẫu ban đầu

- Tiền xử lý:

+ Làm sạch: Loại bỏ ký tự đặc biệt (emoji, ký hiệu HTML), chuẩn hóa Unicode (VD: “hoà” thành “hòa”).

+ Sửa lỗi chính tả: Sử dụng VnCoreNLP để sửa “hk” thành “không”, “z” thành “vậy”.

+ Gán nhãn:

Tích cực: ví dụ “Em rất thích bài học hôm nay!”.

Tiêu cực: ví dụ “Em thấy môn này quá khó, chán lắm”.

Trung tính: ví dụ “Hôm nay em học bài mới”.

+ Cân bằng dữ liệu: Sử dụng SMOTE (Synthetic Minority Oversampling Technique) để sinh mẫu cho lớp trung tính, tránh việc mô hình học nghiêng về lớp tích cực/tiêu cực bỏ qua lớp trung tính

- Kết quả: Bộ dữ liệu sau cân bằng tăng lên tổng 21.637 mẫu, chia thành:

+ Huấn luyện: 15.348 mẫu

+ Kiểm định: 2.113 mẫu.

+ Kiểm tra: 4.156 mẫu.

2.2. Xây dựng mô hình PhoBERT

- Lý do chọn PhoBERT:

+ Tối ưu cho tiếng Việt với token hóa cấp âm tiết, xử lý tốt dấu thanh và ngữ pháp linh hoạt.

+ Pre-trained trên bộ dữ liệu lớn (20GB văn bản tiếng Việt từ Wikipedia, báo chí), đảm bảo hiểu ngữ cảnh.

+ Hiệu suất cao hơn BERT cơ bản trên các bài toán tiếng Việt, ví dụ F1-score 0.92 so với 0.88 của BERT.

- Cấu trúc mô hình:

+ PhoBERT-base: 12 layer, 110 triệu tham số.

+ Thêm linear classifier: Đầu ra 3 lớp (tích cực, tiêu cực, trung tính).

+ Fine-tune: Tùy chỉnh trọng số mô hình trên dữ liệu giáo dục để tối ưu hóa hiệu suất.

- Tokenizer:

+ Chuyển văn bản thành token cấp âm tiết, ví dụ “học sinh” thành “học_sinh”.

+ Chuẩn hóa đầu vào: Độ dài tối đa 128 token, padding (thêm 0) hoặc cắt nếu cần.

- Sử dụng thư viện HuggingFace Transformers để token hóa tự động.

2.3. Huấn luyện mô hình

- Cấu hình huấn luyện:

+ Optimizer: AdamW với learning rate 2e-5 (đã thử nghiệm 1e-5, 5e-5).

+ Scheduler: Tuyến tính với warmup (10% bước đầu tiên để ổn định huấn luyện).

+ Mixed precision (FP16): Giảm bộ nhớ, tăng tốc độ huấn luyện trên GPU.

+ Batch size: 8.

+ Epoch: 20.

- Dataloader:

+ Chia dữ liệu thành batch, shuffle để đảm bảo tính ngẫu nhiên.

+ Sử dụng PyTorch DataLoader để tối ưu hóa tải dữ liệu, giảm thời gian chờ.

- Quá trình huấn luyện:

+ Huấn luyện 20 epoch, mỗi epoch kiểm tra Độ mất mát và Độ chính xác trên tập kiểm định.

+ Theo dõi Độ mất mát (Loss):

Độ mất mát trên tập huấn luyện: Giảm từ 0.28 xuống gần 0 sau 15 epoch.

Độ mất mát trên tập kiểm định: Giảm từ 0.25 xuống 0.15, tăng nhẹ từ epoch 6-7 (dấu hiệu overfitting).

+ Lưu checkpoint tốt nhất dựa trên F1-score kiểm định (0.95).

- Triển khai:

+ Tích hợp mô hình vào phần mềm EduSentiment.

+ Phân tích cảm xúc thời gian thực trên văn bản đầu vào (bài viết, phản hồi).

+ Độ trễ: <1 giây cho văn bản ngắn (<100 từ).

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH

3.1. Cấu hình thử nghiệm

- Môi trường:

+ Nền tảng: Kaggle

+ Phần cứng: GPU NVIDIA RTX 3060 (12GB VRAM), CPU Intel Core i7-10700, RAM 32GB.

+ Phần mềm: Python 3.8, PyTorch 1.9, HuggingFace Transformers 4.12.

- Siêu tham số:

+ Epoch: 20.

+ Learning rate: 2e-5 (tốt nhất sau khi thử 1e-5, 5e-5).

+ Batch size: 8.

+ Dropout: 0.1 ở các layer fully-connected để giảm overfitting.

+ Early stopping: Dừng huấn luyện nếu độ mất mát trên tập kiểm định không cải thiện sau 3 epoch.

3.2. Kết quả huấn luyện

- Độ chính xác (Accuracy):

+ Huấn luyện: Gần 100% sau 15 epoch, cho thấy mô hình học tốt trên dữ liệu huấn luyện.

+ Kiểm định: 95% sau 10 epoch, ổn định từ epoch 10-20.

- Độ mất mát (Loss):

+ Độ mất mát trên tập huấn luyện: Giảm từ 0.28 xuống gần 0, cho thấy mô hình tối ưu hóa tốt.

+ Độ mất mát trên tập kiểm định: Giảm từ 0.25 xuống 0.15, tăng nhẹ từ epoch 6-7 (dấu hiệu overfitting nhẹ).

- F1-score:

+ Kiểm định: 0.95 (trung bình 3 lớp).

+ Cao nhất ở lớp tích cực (0.96), thấp nhất ở lớp trung tính (0.94).

3.3. Đánh giá mô hình

- Báo cáo phân loại (Bảng 3.2):

Lớp	**Precision**	**Recall**	**F1-score**
Tiêu cực	0.93	0.97	0.95
Trung tính	0.97	0.92	0.94
Tích cực	0.95	0.95	0.95
Trung bình	0.95	0.95	0.95

- Ma trận nhầm lẫn (Hình 3.3):

+ Mô hình phân biệt tốt lớp tích cực và tiêu cực (98% chính xác).

+ Nhầm lẫn chủ yếu ở lớp trung tính: 5% mẫu trung tính bị nhầm thành tích cực hoặc tiêu cực, ví dụ “Bài học hôm nay ổn” bị phân loại sai thành tích cực.

+ Nguyên nhân: Ngữ nghĩa mơ hồ của lớp trung tính, thiếu dữ liệu đa dạng.

- Tổng thể:

+ Độ chính xác: 95% trên tập kiểm tra.

+ F1-score: 0.95, cho thấy mô hình cân bằng giữa Precision (độ chính xác) và Recall (độ bao phủ).

3.4. Đánh giá tổng quan

- Hiệu suất:

+ PhoBERT đạt hiệu quả cao, khái quát hóa tốt trên dữ liệu giáo dục, đặc biệt với văn bản ngắn (<100 từ).

+ Hiệu suất vượt trội so với các mô hình truyền thống như SVM (F1-score ~0.85) và LSTM (F1-score ~0.90).

- Quá khớp (Overfitting) được kiểm soát bằng:

+ Early stopping: Dừng huấn luyện khi độ mất mát trên tập kiểm định không cải thiện.

+ Dropout: 0.1.

+ Data augmentation: Sinh mẫu bằng SMOTE để cân bằng dữ liệu.

- Hạn chế:

+ Nhầm lẫn lớp trung tính do ngữ nghĩa mơ hồ.

+ Hiệu suất giảm nhẹ trên văn bản dài (>200 từ) do giới hạn độ dài đầu vào (128 token).

- Cải tiến đề xuất:

+ Tích hợp Bi-LSTM để xử lý văn bản dài, giữ ngữ cảnh dài hạn.

+ Attention để tập trung vào từ/cụm từ quan trọng, ví dụ "chán" trong "Bài học này chán quá".

+ Mở rộng dữ liệu trung tính bằng cách thu thập thêm từ diễn đàn giáo dục.

- Khả năng kết nối với các ứng dụng để thực tiễn hóa

3.5. Mô hình thử nghiệm

Học viên thực hiện xây dựng hệ thống EduSentiment trên ứng dụng Visual Code Studio, thực hiện tích hợp mô hình đã được huấn luyện hoàn thiện trên nền tảng Kaggle để thực tiễn hóa khả năng phân tích cảm xúc.

Hệ thống hỗ trợ phân tích các câu cảm xúc ngắn của sinh viên, phân theo lớp, ký học và đưa ra kết quả phân tích theo các nhãn "Tích cực", "Trung tính", "Tiêu cực". Ngoài ra hệ thống còn hỗ trợ lưu lại các đánh giá cảm xúc để hỗ trợ báo cáo đánh giá cảm xúc của sinh viên đối với môn học.

III. KẾT LUẬN

Đề án đã xây dựng thành công mô hình nhận diện cảm xúc dựa trên PhoBERT, đạt Độ chính xác 95% và F1-score 0.95 trên dữ liệu giáo dục tiếng Việt. Mô hình phân loại hiệu quả cảm xúc tích cực và tiêu cực, nhưng cần cải thiện nhầm lẫn ở lớp trung tính do ngữ nghĩa mơ hồ. Phần mềm EduSentiment tích hợp mô hình, hỗ trợ phân tích thời gian thực từ bài viết và phản hồi học sinh, giúp giáo viên hiểu rõ tâm lý, năng lượng học tập, và điều chỉnh phương pháp giảng dạy. Phần mềm cung cấp giao diện thân thiện, bảo mật cao, và khả năng trực quan hóa dữ liệu, góp phần nâng cao chất lượng giáo dục.

- Hạn chế của đề án:

+ Nhầm lẫn lớp trung tính do thiếu dữ liệu đa dạng và ngữ nghĩa mơ hồ.

+ Hiệu suất giảm trên văn bản dài (>200 từ) do giới hạn độ dài đầu vào của PhoBERT.

- Hướng phát triển:

+ Tích hợp Bi-LSTM và Attention để xử lý văn bản dài, cải thiện hiểu ngữ cảnh phức tạp.

+ Mở rộng tập dữ liệu, đặc biệt là mẫu trung tính, bằng cách thu thập thêm từ các nguồn giáo dục và mạng xã hội.

+ Mở rộng việc huấn luyện khi mở rộng thêm các nhãn cảm xúc khác như tức giận, buồn bã, vui vẻ….

+ Ứng dụng mô hình vào các lĩnh vực khác như thương mại điện tử (phân tích đánh giá sản phẩm trên Shopee) và mạng xã hội (theo dõi dư luận trên Facebook), mở rộng phạm vi tác động của nghiên cứu.

+ Mở rộng khả năng của ứng dụng khi triển khai trên website thực tế.