HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÙI HUY DŨNG

ÁP DỤNG KỸ THUẬT HỌC SÂU CHO PHÂN TÍCH Ý KIẾN KHÁCH HÀNG VỀ CÁC SẢN PHẨM CÔNG NGHỆ

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI - NĂM 2025

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS.TS. Trần Đình Quế

Phản biện 1: PGS.TS. Đỗ Trung Tuấn

Phản biện 2: PGS.TS. Nguyễn Hữu Quỳnh

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 8 giờ 15 phút ngày 24 tháng 01 năm 2026

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

1. Lý do lựa chọn đề tài

Trong những năm gần đây, tiến trình chuyển đổi số quốc gia tại Việt Nam đã tạo ra sự thay đổi mạnh mẽ trong cách thức người dân tiếp cận và sử dụng các ứng dụng công nghệ. Những nền tảng phổ biến như Zalo – ứng dụng nhắn tin, gọi điện và mạng xã hội có lượng người dùng lớn nhất Việt Nam – và iSpeed – ứng dụng đo tốc độ Internet của Trung tâm Internet Việt Nam (VNNIC) – đóng vai trò quan trọng trong đời sống số của người dân. Bên cạnh đó, các sản phẩm công nghệ quốc tế như iPhone 11 hay các dòng thiết bị điện tử trên Amazon cũng thu hút lượng lớn người dùng và phản hồi trên Internet.

Những phản hồi của người dùng trên các nền tảng như Google Play, App Store hay hệ thống thương mại điện tử chứa đựng thông tin giàu giá trị. Chúng không chỉ phản ánh mức độ hài lòng, mà còn thể hiện trực tiếp trải nghiệm thực tế, khó khăn và kỳ vọng của người dùng. Tuy nhiên, việc xử lý thủ công lượng lớn dữ liệu này là không khả thi, do số lượng bình luận quá lớn, ngôn ngữ phi chuẩn, chứa nhiều emoji, teencode, ký tự đặc biệt và cấu trúc câu thiếu thống nhất.

Trong bối cảnh đó, phân tích cảm xúc (Sentiment Analysis) – một bài toán quan trọng trong xử lý ngôn ngữ tự nhiên (NLP) – trở thành công cụ rất hữu ích. Với sự phát triển của học sâu (Deep Learning), đặc biệt là các mô hình ngôn ngữ tiền huấn luyện như BERT và PhoBERT, khả năng hiểu ngữ nghĩa trong bình luận đã được cải thiện đáng kể. Tuy nhiên, việc ứng dụng các mô hình này vào dữ liệu tiếng Việt, đặc biệt là dữ liệu thực tế từ kho ứng dụng, vẫn còn là thách thức.

Từ thực tế đó, đề tài lựa chọn hướng nghiên cứu xây dựng và đánh giá hệ thống phân tích cảm xúc đa ngôn ngữ trên bốn tập dữ liệu: Zalo, iSpeed, Amazon Electronics và iPhone 11 Reviews, với trọng tâm là dữ liệu tiếng Việt.

2. Tổng quan tình hình nghiên cứu

Phân tích cảm xúc (Sentiment Analysis) là một trong những hướng nghiên cứu quan trọng của xử lý ngôn ngữ tự nhiên (NLP) và đã đạt được nhiều tiến bộ trên thế giới. Các phương pháp truyền thống như TF-IDF, Naive Bayes hay SVM từng là nền tảng trong giai đoạn đầu, dựa trên thống kê tần suất từ và các đặc trưng thủ công. Tuy nhiên, những phương pháp này không nắm bắt được ngữ cảnh sâu, dẫn đến hạn chế khi xử lý các câu phức tạp hoặc mang tính đa nghĩa.

Sự phát triển của học sâu đã mở ra thế hệ mô hình hiệu quả hơn như CNN, LSTM, GRU, giúp mô hình hiểu được quan hệ tuần tự và cấu trúc ngôn ngữ trong câu. Đặc biệt, LSTM và BiLSTM chứng minh khả năng xử lý tốt chuỗi văn bản dài và khó. Những năm gần đây, xu hướng nghiên cứu chuyển mạnh sang các mô hình hiện đại dựa trên kiến trúc Transformer như BERT, RoBERTa và PhoBERT. Các mô hình này học biểu diễn ngữ nghĩa theo ngữ cảnh hai chiều, mang lại hiệu suất vượt trội trong hầu hết tác vụ NLP.

Tại Việt Nam, nghiên cứu phân tích cảm xúc vẫn chưa phát triển tương xứng, đặc biệt đối với dữ liệu từ nền tảng ứng dụng di động. Các bình luận từ App Store và Google Play thường ngắn, phi chuẩn, chứa nhiều emoji, teencode và lỗi chính tả, khiến việc xây dựng mô hình phân tích cảm xúc trở nên thách thức. Việc áp dụng mô hình hiện đại như PhoBERT vào dữ liệu thực tế của người dùng Việt Nam vẫn còn ít được khai thác, tạo ra khoảng trống nghiên cứu đáng kể.

3. Mục tiêu nghiên cứu

3.1. Mục tiêu tổng quát

Mục tiêu tổng quát của đề tài là xây dựng một hệ thống phân tích cảm xúc tự động áp dụng các kỹ thuật học sâu hiện đại, nhằm nhận diện cảm xúc của người dùng trong các bình luận tiếng Việt lẫn tiếng Anh trên những ứng dụng phổ biến như Zalo, iSpeed và các bộ dữ liệu quốc tế từ Kaggle. Hệ thống hướng tới khả năng hoạt động ổn định trên dữ liệu thực tế, có nhiều nhiễu và tính phi chuẩn cao.

3.2. Mục tiêu cụ thể

- Thu thập, tổng hợp và chuẩn hóa dữ liệu bình luận từ App Store và Google Play trong 12 tháng gần nhất, đảm bảo tính cập nhật và đại diện cho hành vi người dùng hiện tại.
- Xây dựng pipeline tiền xử lý tiếng Việt bao gồm làm sạch văn bản, chuẩn hóa ký tự, xử lý emoji, tách từ và loại bỏ stopword để tối ưu hóa đầu vào cho mô hình.
- Gán nhãn dữ liệu bằng phương pháp rule-based dựa trên tập từ điển cảm xúc tự xây dựng, làm cơ sở cho mô hình baseline.
- Huấn luyện và đánh giá mô hình Word2Vec + MLP như một mô hình cơ sở nhằm kiểmchứng quy trình xử lý và khả năng phân loại ban đầu.
- Ứng dụng PhoBERT – mô hình ngôn ngữ tiền huấn luyện mạnh mẽ cho tiếng Việt – để cải thiện độ chính xác, đồng thời so sánh hiệu năng với mô hình baseline và mô hình tiếng Anh.
- Đánh giá kết quả bằng bộ chỉ số toàn diện gồm Accuracy, Precision, Recall, F1-score và ma trận nhầm lẫn; phân tích nguyên nhân sai lệch và các yếu tố ảnh hưởng đến hiệu quả mô hình.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là các bình luận văn bản (textual reviews) được người dùng đăng tải công khai trên App Store, Google Play và các tập dữ liệu Kaggle. Cụ thể, đề tài tập trung vào bốn nhóm dữ liệu: bình luận tiếng Việt của người dùng Zalo và iSpeed, cùng với hai tập dữ liệu tiếng Anh là Amazon Electronics và iPhone 11 Reviews. Những bình luận này phản ánh trải nghiệm thực tế của người dùng đối với sản phẩm công nghệ, đa dạng về độ dài, mức độ nhiễu và ngữ cảnh.

Phạm vi nghiên cứu giới hạn ở dữ liệu được thu thập trong vòng 12 tháng gần nhất nhằm bảo đảm tính cập nhật. Đề tài chỉ phân tích nội dung văn bản bằng tiếng Việt và tiếng Anh, không thu thập thông tin cá nhân, dữ liệu nhạy cảm hoặc các thành phần đa phương tiện như hình ảnh, video. Việc sử dụng dữ liệu tuân thủ đầy đủ chính sách bảo mật và điều khoản của các nền tảng.

5. Phương pháp nghiên cứu

Đề tài sử dụng phương pháp:

- Thu thập dữ liệu: Sử dụng crawler hoặc API để lấy dữ liệu từ App Store và Google Play, đồng thời khai thác dữ liệu sẵn có từ Kaggle. Dữ liệu được lưu trữ và xử lý thống nhất để phục vụ huấn luyện.
- Tiền xử lý dữ liệu: Làm sạch văn bản bằng cách loại bỏ ký tự đặc biệt, chuẩn hóa emoji, tách từ tiếng Việt bằng VnCoreNLP, chuyển văn bản về dạng thường và lọc bỏ stopword. Mục tiêu là tạo đầu vào nhất quán cho mô hình.
- Gán nhãn cảm xúc: Sử dụng bộ từ điển cảm xúc tự xây dựng để gán nhãn ban đầu theo phương pháp rule-based, phân chia bình luận thành ba lớp: tích cực, trung lập, tiêu cực. Đây là bước quan trọng để hình thành tập dữ liệu huấn luyện.
- Biểu diễn văn bản: Dùng Word2Vec để biến đổi văn bản thành vector ngữ nghĩa, phù hợp cho mô hình MLP. Với PhoBERT, văn bản được mã hóa trực tiếp bằng embedding theo ngữ cảnh.
- Huấn luyện mô hình: MLP được sử dụng làm baseline vì nhẹ, phù hợp môi trường không có GPU. PhoBERT được ứng dụng như mô hình đề xuất nhờ khả năng nắm bắt ngữ cảnh sâu và hiệu quả vượt trội trong tiếng Việt.
- Đánh giá mô hình: Sử dụng Accuracy, Precision, Recall, F1-score và Confusion Matrix để đo lường hiệu năng. Đồng thời so sánh kết quả giữa bốn tập dữ liệu để đánh giá mức độ tổng quát hóa.

6. Kết cấu của đề án

Đề án gồm ba chương:

- Chương 1: Cơ sở lý luận
- Chương 2: Phương pháp nghiên cứu và mô hình đề xuất
- Chương 3: Thực nghiệm và đánh giá kết quả

Cuối cùng là Kết luận và kiến nghị.

CHƯƠNG 1: CƠ SỞ LÝ LUẬN

1. Trí tuệ nhân tạo, học máy và học sâu

Trí tuệ nhân tạo (Artificial Intelligence – AI) là lĩnh vực nghiên cứu các phương pháp giúp máy tính có khả năng suy luận, học hỏi và đưa ra quyết định giống con người. Trong AI, học máy (Machine Learning – ML) đóng vai trò trung tâm, cho phép mô hình tự học từ dữ liệu mà không cần lập trình tường minh. ML khai thác các thuật toán như hồi quy, phân cụm, SVM nhằm nhận diện quy luật trong dữ liệu.

Học sâu (Deep Learning – DL) là bước tiến mở rộng của ML, sử dụng các mạng nơ-ron nhân tạo với nhiều tầng ẩn để học các biểu diễn trừu tượng và phức tạp của dữ liệu. DL hoạt động đặc biệt hiệu quả trong các lĩnh vực như thị giác máy tính, xử lý giọng nói và xử lý ngôn ngữ tự nhiên (NLP). Trong NLP, các kiến trúc mạng nơ-ron như CNN, RNN, LSTM, GRU và đặc biệt là Transformer đã chứng minh khả năng mô hình hóa ngữ nghĩa mạnh mẽ, giúp xử lý câu văn dài, ngữ cảnh phức tạp và quan hệ từ vựng đa dạng.

Nhờ khả năng học biểu diễn tự động từ dữ liệu lớn, học sâu trở thành nền tảng của các hệ thống phân tích cảm xúc hiện đại.

2. Xử lý ngôn ngữ tự nhiên tiếng Việt

Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) tập trung vào việc cho phép máy tính hiểu và xử lý ngôn ngữ của con người. Tuy nhiên, tiếng Việt có nhiều đặc thù khiến bài toán NLP trở nên khó khăn hơn so với tiếng Anh:

- Tiếng Việt là ngôn ngữ đa âm tiết, mỗi từ có thể bao gồm nhiều âm tiết tách rời bằng dấu cách.
- Dấu thanh phong phú (sắc, huyền, hỏi, ngã, nặng) làm thay đổi nghĩa từ.
- Nhiều biến thể phi chuẩn: teencode, viết tắt, viết không dấu.
- Sự pha trộn giữa tiếng Việt – tiếng Anh trong bình luận trên mạng.

Tiếng Việt có đặc thù đa âm tiết, dấu thanh phức tạp, nhiều teencode, khiến việc xử lý khó khăn hơn so với tiếng Anh. Quy trình NLP gồm:

- Tiền xử lý văn bản: chuẩn hóa chữ viết, xoá ký tự đặc biệt, chuẩn hóa unicode, loại bỏ emoji hoặc giữ lại tùy mục đích.
- Tách từ tiếng Việt: sử dụng các công cụ như VnCoreNLP hoặc RDRSegmenter.
- Biểu diễn đặc trưng: biến đổi văn bản sang dạng vector bằng TF-IDF, Word2Vec hoặc embedding theo ngữ cảnh.
- Huấn luyện mô hình: sử dụng học máy hoặc học sâu.
- Đánh giá: đo lường hiệu quả qua Accuracy, F1-score, Precision, Recall.

Những mô hình hiện đại như BERT và PhoBERT thể hiện ưu thế vượt trội vì có khả năng hiểu ngữ cảnh tốt hơn, phù hợp với ngữ pháp và cấu trúc tiếng Việt.

3. Bài toán phân tích cảm xúc

Phân tích cảm xúc (Sentiment Analysis) là nhiệm vụ xác định thái độ hoặc quan điểm của người viết đối với một đối tượng cụ thể. Kết quả phân tích thường được biểu diễn dưới dạng các nhãn:

- Tích cực
- Tiêu cực
- Trung lập

Bài toán này có nhiều ứng dụng thực tế, bao gồm:

- Đánh giá mức độ hài lòng của khách hàng đối với sản phẩm.
- Hỗ trợ doanh nghiệp cải thiện dịch vụ hoặc phát hiện vấn đề.
- Phân tích dư luận xã hội, truyền thông và mạng xã hội.
- Tổng hợp ý kiến người dùng để đưa ra quyết định chiến lược.

Trong bối cảnh các ứng dụng di động và thương mại điện tử phát triển mạnh, phân tích cảm xúc trở thành công cụ quan trọng để hiểu rõ hành vi và trải nghiệm người dùng.

4. Các phương pháp phân tích cảm xúc

- Rule-based: dựa trên từ điển cảm xúc
- Học máy truyền thống: TF-IDF + SVM
- Học sâu: CNN, LSTM
- Transformer-based: BERT, RoBERTa, PhoBERT

5. Mô hình Transformer, BERT, RoBERTa, PhoBERT

- Transformer: nền tảng với cơ chế self-attention.
- BERT: mô hình song song hai chiều đầu tiên.
- RoBERTa: tối ưu phương pháp huấn luyện BERT.
- PhoBERT: mô hình RoBERTa huấn luyện trên 20GB dữ liệu tiếng Việt.

6. So sánh Word2Vec – BERT – PhoBERT

Mô hình	Ưu điểm	Hạn chế
Word2Vec	Nhẹ, nhanh	Không hiểu ngữ cảnh
BERT	Hiểu ngữ cảnh tốt	Không tối ưu tiếng Việt
PhoBERT	Tối ưu tiếng Việt, mạnh nhất hiện nay	Cần nhiều tài nguyên

CHƯƠNG 2: PHƯƠNG PHÁP NGHIÊN CỨU VÀ MÔ HÌNH ĐỀ XUẤT

1. Quy trình hệ thống phân tích cảm xúc

Hệ thống phân tích cảm xúc được xây dựng dựa trên một pipeline gồm nhiều bước liên tiếp, đảm bảo tính nhất quán từ khâu thu thập dữ liệu đến giai đoạn đánh giá cuối cùng. Quy trình tổng thể bao gồm:

- Thu thập dữ liệu: lấy các bình luận của người dùng từ nhiều nguồn khác nhau, bao gồm ứng dụng Zalo, iSpeed cũng như các bộ dữ liệu quốc tế từ Kaggle.
- Tiền xử lý văn bản: làm sạch dữ liệu thô, xử lý emoji, chuẩn hóa văn bản và tách từ (đối với tiếng Việt).
- Gán nhãn cảm xúc: áp dụng phương pháp rule-based theo bộ từ điển tự xây dựng để phân loại cảm xúc ban đầu.
- Trích xuất đặc trưng (embedding): chuyển đổi văn bản thành dạng vector số để mô hình học sâu có thể xử lý.
- Huấn luyện mô hình: mô hình MLP được huấn luyện trên các vector embedding để dự đoán cảm xúc.
- Đánh giá mô hình: sử dụng các chỉ số Accuracy, Precision, Recall, F1-score và ma trận nhầm lẫn để xác định chất lượng mô hình.

Quy trình này đảm bảo hệ thống có khả năng xử lý dữ liệu tiếng Việt lẫn tiếng Anh, đồng thời hỗ trợ so sánh giữa các mô hình và các tập dữ liệu khác nhau.

2. Thu thập và mô tả dữ liệu

Nguồn:

- Bình luận Zalo (Google Play): dữ liệu tiếng Việt tương đối rõ ràng, số lượng lớn, nội dung dài và ít nhiễu, phù hợp cho việc huấn luyện mô hình.
- Bình luận iSpeed (Google Play): dữ liệu tiếng Việt nhưng có đặc trưng bình luận ngắn, chứa nhiều từ viết tắt, từ phủ định (“không”), emoji và lỗi chính tả, gây khó khăn cho quá trình phân tích.

- Amazon Electronics (Kaggle): dữ liệu tiếng Anh với chủ đề đa dạng liên quan đến thiết bị điện tử, được dùng để đánh giá khả năng tổng quát hóa của mô hình.
- iPhone 11 Reviews (Kaggle): dữ liệu tiếng Anh tập trung vào một sản phẩm cụ thể, bình luận dài, ngữ cảnh rõ ràng, ít nhiễu hơn Amazon.

Các tập dữ liệu có sự khác biệt lớn về cấu trúc câu và độ nhiễu, giúp kiểm tra mô hình trong nhiều điều kiện khác nhau. Đặc điểm chung là hầu hết bình luận có độ dài ngắn, xuất hiện nhiều teencode, emoji, từ viết tắt và lỗi chính tả – đặc biệt ở dữ liệu tiếng Việt.

3. Tiền xử lý dữ liệu tiếng Việt

Tiền xử lý đóng vai trò quan trọng trong NLP, đặc biệt với tiếng Việt – ngôn ngữ đa âm tiết và chứa nhiều biến thể phi chuẩn trong giao tiếp trực tuyến. Các bước tiền xử lý được thực hiện bao gồm:

- Chuẩn hóa văn bản: chuyển toàn bộ văn bản về chữ thường, chuẩn hóa unicode tránh lỗi font.
- Loại bỏ ký tự đặc biệt: xoá URL, ký tự thừa, ký hiệu cảm xúc, emoji (hoặc thay thế bằng token nếu cần giữ lại).
- Tách từ tiếng Việt: sử dụng VnCoreNLP để phân tách câu thành các cụm từ có nghĩa, giúp mô hình hiểu đúng đơn vị từ vựng.
- Loại bỏ stopword: loại các từ không mang nhiều giá trị ngữ nghĩa như “và”, “là”, “thì”, “của”,…
- Xử lý chữ viết tắt và teencode: chuẩn hóa các dạng như “ko → không”, “j → gì”, “ad → admin”.

Kết quả của bước tiền xử lý là văn bản sạch và thống nhất, phù hợp cho các phương pháp embedding.

4. Phương pháp gán nhãn

Sử dụng từ điển:

- Tích cực: tốt, đẹp, nhanh, tuyệt_vời,…
- Tiêu cực: lỗi, chậm, tệ, lag, đơ,…

Trường hợp không có từ khóa → nhãn trung lập.

5. Biểu diễn dữ liệu (embedding)

- Word2Vec: vector 100 chiều, window 5, epochs 80
- PhoBERT: dùng feature extraction (do hạn chế CPU)

6. Mô hình phân loại (MLP)

- Dense 64 → Dense 32 → Softmax
- Epochs = 200
- Optimizer = Adam
- Loss = sparse categorical crossentropy

7. Tiêu chí đánh giá mô hình

- Accuracy
- Precision – Recall – F1-score
- Confusion Matrix

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

1. Tập dữ liệu thực nghiệm

Dataset	Accuracy	Nhận xét
Zalo	~90%	Dữ liệu rõ ràng, ít nhiễu
iSpeed	~73%	Nhiều bình luận ngắn, khó phân loại
Amazon Electronics	~82%	Khả năng tổng quát hóa tốt
iPhone 11 Reviews	~95%	Dữ liệu tiếng Anh rõ ràng, ít nhiễu, câu dài và có ngữ nghĩa đầy đủ

2. Kết quả phân tích

Kết quả thực nghiệm trên bốn tập dữ liệu cho thấy sự khác biệt đáng kể về hiệu năng mô hình giữa các miền dữ liệu và ngôn ngữ. Đối với tập **Zalo**, mô hình đạt kết quả cao nhất trong toàn bộ quá trình đánh giá. Nguyên nhân chính là các bình luận Zalo có nội dung tương đối dài, rõ ràng và chứa nhiều từ biểu thị cảm xúc, giúp Word2Vec học được các đặc trưng ngữ nghĩa ổn định. Đây cũng là tập dữ liệu ít nhiễu nhất trong bốn tập, do phần lớn bình luận được viết đúng chính tả và theo ngữ pháp phổ thông, nên mô hình có điều kiện học tốt hơn.

Ngược lại, dữ liệu iSpeed thể hiện độ khó cao hơn. Các bình luận thường chỉ gồm 1–3 từ, nhiều câu đơn, từ viết tắt, hoặc thông tin quá ngắn khiến mô hình phân loại gặp khó khăn. Đặc biệt, tần suất xuất hiện dày đặc của từ "không" – một từ phủ định đa nghĩa – làm Word2Vec dễ bị hiểu sai ngữ cảnh nếu không xử lý negation cẩn thận. Điều này dẫn đến độ chính xác thấp hơn đáng kể so với tập Zalo, phản ánh tính chất đặc thù và mức độ nhiễu cao của dữ liệu iSpeed.

Đối với hai tập dữ liệu tiếng Anh từ Kaggle (Amazon Electronics và iPhone 11 Reviews), kết quả cho thấy mô hình vẫn duy trì khả năng tổng quát hóa tốt khi chuyển sang ngôn ngữ khác. Dữ liệu Amazon có tính đa miền (multi-domain), do

bao gồm nhiều loại thiết bị điện tử khác nhau, vì vậy độ nhiễu cao và độ chính xác trung bình. Trong khi đó, tập iPhone 11 Reviews có chủ đề thống nhất, cấu trúc câu đầy đủ, từ vựng rõ ràng hơn nên kết quả đạt mức cao hơn, thậm chí vượt một số tập dữ liệu tiếng Việt. Điều này chứng minh pipeline Word2Vec + MLP vẫn có thể hoạt động hiệu quả với tiếng Anh.

Kết quả so sánh giữa Word2Vec và PhoBERT cho thấy PhoBERT vượt trội hoàn toàn nhờ cơ chế embedding theo ngữ cảnh hai chiều. PhoBERT nắm bắt được cấu trúc và ngữ nghĩa tinh tế trong tiếng Việt mà Word2Vec không thể hiện được, đặc biệt với các câu ngắn hoặc câu chứa từ phủ định. Điều này khẳng định tính phù hợp của các mô hình Transformer trong phân tích cảm xúc hiện đại.

3. Nhận xét chung

Từ các kết quả thực nghiệm, có thể đưa ra một số nhận xét quan trọng:

Trước hết, mô hình Word2Vec + MLP chứng minh là giải pháp khả thi trong điều kiện hạn chế phần cứng, đặc biệt khi không có GPU. Mô hình có tốc độ huấn luyện nhanh, ít tiêu tốn tài nguyên và vẫn đạt độ chính xác tốt trên những tập dữ liệu có mức độ nhiễu thấp như Zalo hoặc iPhone 11 Reviews. Đây là một lựa chọn hợp lý cho các hệ thống real-time hoặc các doanh nghiệp nhỏ muốn triển khai phân tích cảm xúc nhưng không có hạ tầng mạnh.

Tuy nhiên, PhoBERT cho kết quả vượt trội rõ rệt so với Word2Vec, nhờ khả năng hiểu ngữ cảnh sâu và biểu diễn ngôn ngữ giàu thông tin hơn. Hạn chế lớn nhất của PhoBERT là yêu cầu tài nguyên tính toán lớn cả trong huấn luyện và suy luận. Trong điều kiện phần cứng hạn chế (chỉ CPU), việc fine-tune PhoBERT gần như không khả thi. Dù vậy, ngay cả khi chỉ dùng dạng feature extraction, PhoBERT vẫn cho kết quả tốt hơn, chứng tỏ tiềm năng mạnh mẽ trong các ứng dụng NLP tiếng Việt.

Một kết luận quan trọng khác là pipeline xử lý tiếng Việt ảnh hưởng rất mạnh đến hiệu quả của mô hình. Các thao tác như chuẩn hóa văn bản, xử lý từ phủ định, tách từ và loại bỏ stopword đóng vai trò quyết định. Những tập dữ liệu có nhiều

bình luận phi chuẩn, ngắn hoặc thiếu thông tin (như iSpeed) sẽ làm giảm đáng kể chất lượng phân loại nếu pipeline không được thiết kế kỹ lưỡng.

Cuối cùng, các biểu đồ Loss và Accuracy cho thấy mô hình hội tụ tốt, ổn định và không có hiện tượng overfitting đáng kể. Điều này phản ánh rằng kiến trúc mạng MLP đang được sử dụng hợp lý, kết hợp với dropout, regularization và early stopping giúp mô hình đạt sự cân bằng giữa độ phức tạp và khả năng tổng quát hóa.

KẾT LUẬN VÀ KIẾN NGHỊ

1. Kết luận

Đề tài đã xây dựng một hệ thống phân tích cảm xúc tiếng Việt hoàn chỉnh, thử nghiệm trên ba tập dữ liệu độc lập. Kết quả cho thấy:

Đề tài đã xây dựng thành công một hệ thống phân tích cảm xúc cho tiếng Việt dựa trên các kỹ thuật học sâu, áp dụng cho dữ liệu thực tế từ các ứng dụng Zalo, iSpeed và các bộ dữ liệu tiếng Anh trên Kaggle. Hệ thống được thiết kế theo một pipeline hoàn chỉnh bao gồm các bước: thu thập dữ liệu, tiền xử lý văn bản, gán nhãn cảm xúc, biểu diễn đặc trưng bằng Word2Vec/PhoBERT và huấn luyện mô hình phân loại bằng MLP. Quy trình này đảm bảo tính chặt chẽ, có thể tái sử dụng và có khả năng mở rộng.

Kết quả thực nghiệm cho thấy mô hình đạt hiệu năng cao, với Accuracy tối đa lên đến 95% trên tập iPhone 11 Reviews và khoảng 90% trên tập Zalo. Điều này chứng minh pipeline đề xuất hoạt động hiệu quả trên dữ liệu có ngữ nghĩa rõ ràng và độ nhiễu thấp. Ngược lại, kết quả trên tập iSpeed – nơi dữ liệu ngắn và chứa nhiều ký tự nhiễu – thấp hơn đáng kể, phản ánh đúng vai trò then chốt của chất lượng dữ liệu đầu vào.

Bên cạnh đó, kết quả cho thấy PhoBERT vượt trội hoàn toàn so với Word2Vec nhờ khả năng nắm bắt ngữ cảnh hai chiều và hiểu sâu hơn cấu trúc tiếng Việt. Mặc dù hạn chế phần cứng không cho phép fine-tuning mô hình, việc sử dụng PhoBERT theo hướng feature extraction vẫn mang lại chất lượng tốt hơn đáng kể so với mô hình baseline. Điều này khẳng định tính phù hợp của các mô hình Transformer trong bài toán phân tích cảm xúc tiếng Việt hiện đại.

Ngoài ra, pipeline tiền xử lý được xây dựng trong đề tài đã chứng minh hiệu quả cao và có thể tái sử dụng cho nhiều tác vụ NLP khác như phân loại chủ đề, nhận diện ý định người dùng hay phát hiện bình luận độc hại. Các biểu đồ Loss/Accuracy trong quá trình huấn luyện cho thấy mô hình học ổn định, có tốc độ hội tụ nhanh và

không xảy ra overfitting, nhờ việc kết hợp hợp lý giữa dropout, regularization và early stopping.

Tổng thể, đề tài đã chứng minh tính khả thi và hiệu quả của việc áp dụng kỹ thuật học sâu vào phân tích cảm xúc tiếng Việt và tiếng Anh, góp phần vào việc phát triển các ứng dụng phân tích phản hồi người dùng trong chuyển đổi số.

2. Kiến nghị

Để hệ thống có thể được hoàn thiện và ứng dụng rộng rãi hơn trong thực tế, đề tài đưa ra một số kiến nghị sau:

- Mở rộng quy mô và tính đa dạng của bộ dữ liệu: Việc thu thập thêm bình luận từ nhiều nhóm ứng dụng, nhiều lĩnh vực (tài chính, giáo dục, thương mại điện tử, viễn thông) sẽ giúp mô hình học được nhiều đặc trưng ngôn ngữ hơn, giảm hiện tượng lệch miền và tăng khả năng tổng quát hóa.
- Cải thiện bước tiền xử lý tiếng Việt: Quá trình chuẩn hóa teencode, xử lý emoji, phân tách từ viết liền và phát hiện negation hiện nay vẫn còn hạn chế. Hoàn thiện bộ từ điển stopword và áp dụng các thuật toán xử lý ngữ cảnh sẽ giúp nâng cao chất lượng dữ liệu đầu vào.
- Thử nghiệm các mô hình embedding tiên tiến hơn: Bên cạnh Word2Vec, có thể tích hợp thêm các mô hình như FastText, BERT-multilingual, PhoBERT-large, hoặc các embedding dựa trên Sentence-BERT để cải thiện khả năng biểu diễn ngữ nghĩa.
- Fine-tuning PhoBERT khi có hạ tầng GPU: Nếu có điều kiện sử dụng GPU hoặc TPU, việc fine-tune trực tiếp PhoBERT sẽ mang lại hiệu quả vượt trội so với feature extraction. Đây là hướng quan trọng để khai thác tối đa sức mạnh của các mô hình Transformer.
- Ứng dụng mô hình vào hệ thống giám sát phản hồi người dùng theo thời gian thực: Hệ thống có thể được triển khai để tự động thu thập và phân tích bình luận từ App Store/Google Play, từ đó hỗ trợ các doanh nghiệp hoặc tổ chức như VNNIC, Zalo trong việc theo dõi mức độ hài lòng và phát hiện sớm các vấn đề dịch vụ.

- Mở rộng bài toán sang phân tích khía cạnh (Aspect-based Sentiment Analysis): Đây là hướng nghiên cứu sâu hơn, cho phép mô hình không chỉ phân loại cảm xúc tổng thể mà còn đánh giá cảm xúc theo từng khía cạnh của sản phẩm (ví dụ: tốc độ, giao diện, tính năng,…).
- Tích hợp dashboard trực quan: Xây dựng giao diện trực quan giúp hiển thị biểu đồ cảm xúc, xu hướng theo thời gian, tỷ lệ phản hồi tích cực/tiêu cực để hỗ trợ ra quyết định.