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

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC:

PGS.TS. TRẦN ĐÌNH QUẾ

HÀ NỘI – Năm 2026

LỜI CAM ĐOAN

Tôi xin cam đoan rằng đề tài *"Áp dụng kỹ thuật học sâu cho phân tích ý kiến khách hàng về các sản phẩm công nghệ"* là công trình nghiên cứu do chính tôi thực hiện.

Toàn bộ số liệu, tập dữ liệu, quy trình xử lý, mô hình huấn luyện và kết quả thực nghiệm trình bày trong luận văn đều được tôi thu thập, xây dựng và triển khai một cách trung thực, đúng với thực trạng nghiên cứu. Các kết quả không sao chép từ bất kỳ công trình nào khác, trừ các nội dung đã được trích dẫn đầy đủ trong phần tài liệu tham khảo.

Tôi chịu hoàn toàn trách nhiệm về tính trung thực và chính xác của nội dung luận văn.

Tác giả luận văn



BÙI HUY DŨNG

MỤC LỤC

DANH MỤC BẢNG iv
DANH MỤC HÌNH v
DANH MỤC CHỮ CÁI VIẾT TẮT vi
MỞ ĐẦU 1
1. Lý do lựa chọn vấn đề 1
2. Tổng quan về vấn đề nghiên cứu 2
3. Mục tiêu, nghiên cứu 4
4. Đối tượng và phạm vi nghiên cứu 4
5. Phương pháp nghiên cứu 5
6. Kết cấu của luận văn 6
CHƯƠNG 1: CƠ SỞ LÝ LUẬN 7
1.1. Tổng quan về trí tuệ nhân tạo, học máy và học sâu 7
1.2. Xử lý ngôn ngữ tự nhiên 7
1.3. Bài toán phân tích cảm xúc trong tiếng việt 9
1.4. Các phương pháp và mô hình học sâu trong phân tích cảm xúc 10
1.5. Mô hình Transformer và các biến thể (BERT, RoBERTa, PhoBERT) 10
1.6. So sánh giữa các mô hình Word2Vec, BERT và PhoBERT 12
1.7. Kết luận chương 1 13
CHƯƠNG 2: PHƯƠNG PHÁP NGHIÊN CỨU VÀ MÔ HÌNH ĐỀ XUẤT 14
2.1. Quy trình tổng thể của hệ thống phân tích cảm xúc 14
2.2. Thu Thập và mô tả dữ liệu đánh giá người dùng 15
2.3. Tiền xử lý dữ liệu văn bản tiếng Việt 18
2.4. Phương pháp gán nhãn cảm xúc 19
2.5. Biểu diễn dữ liệu và trích xuất đặc trưng 20
2.6. Thiết kế và huấn luyện mô hình học sâu 21

2.7. Đánh giá mô hình và tiêu chí so sánh 22
2.8. Kết luận chương 2 23
CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 24
3.1. Mục tiêu thực nghiệm 24
3.2. Môi trường và công cụ thực nghiệm 24
3.3. Chuẩn bị dữ liệu và tiền xử lý 25
3.4. Tiền xử lý dữ liệu 26
3.5. Mô hình biểu diễn văn bản 26
3.6. Mô hình phân loại (MLP) 28
3.7. Kết quả thực nghiệm trên từng tập dữ liệu 30
3.8. So sánh tổng hợp giữa hai cấu hình mô hình (PhoBERT+MLP và Word2Vec+MLP) trên các tập dữ liệu 42
3.9. Nhận xét chung 43
3.10. Kết luận chương 3 44
KẾT LUẬN VÀ KIẾN NGHỊ 45
DANH MỤC TÀI LIỆU THAM KHẢO 47

DANH MỤC BẢNG

Tên	Nội dung
Bảng 1.1	So sánh 3 mô hình Word2Vec, BERT, PhoBERT
Bảng 2.1	Cấu trúc dữ liệu
Bảng 2.2	Đánh giá các mô hình
Bảng 3.1	Chuẩn bị dữ liệu xử lý
Bảng 3.2	So sánh tổng hợp giữa các tập dữ liệu

DANH MỤC HÌNH

Tên	Nội dung
Hình 2.1.	Quy trình tổng thể hệ thống
Hình 3.1.	Báo cáo phân loại – Zalo
Hình 3.2.	Biểu đồ Loss – Zalo
Hình 3.3.	Biểu đồ Accuracy – Zalo
Hình 3.4.	Ma trận nhầm lẫn – Zalo
Hình 3.5.	Báo cáo phân loại – iSpeed
Hình 3.6.	Biểu đồ Loss – iSpeed
Hình 3.7.	Biểu đồ Accuracy – iSpeed
Hình 3.8.	Ma trận nhầm lẫn – iSpeed
Hình 3.9.	Báo cáo phân loại – Kaggle
Hình 3.10.	Biểu đồ Loss – Kaggle
Hình 3.11.	Biểu đồ Accuracy – Kaggle
Hình 3.12.	Ma trận nhầm lẫn – Kaggle
Hình 3.13.	Báo cáo phân loại – iPhone 11 Reviews
Hình 3.14.	Biểu đồ Loss – iPhone 11 Reviews
Hình 3.15.	Biểu đồ Accuracy – iPhone 11 Reviews
Hình 3.16.	Ma trận nhầm lẫn – iPhone 11 Reviews

DANH MỤC CHỮ CÁI VIẾT TẮT

Từ viết tắt	Tiếng Anh	Nghĩa tiếng Việt
BERT	Bidirectional Encoder Representations from Transformers	Mô hình Transformer hai chiều
BiLSTM	Bidirectional Long Short-Term Memory	Mạng LSTM hai chiều
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
GRU	Gated Recurrent Unit	Đơn vị hồi quy có cổng
LSTM	Long Short-Term Memory	Mạng bộ nhớ dài ngắn hạn
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
SVM	Support Vector Machine	Máy hỗ trợ vector
VNNIC	Vietnam Internet Network Information Center	Trung tâm Internet Việt Nam

MỞ ĐẦU

1. Lý do lựa chọn vấn đề

Trong bối cảnh chuyển đổi số quốc gia đang được thúc đẩy mạnh mẽ tại Việt Nam, các nền tảng công nghệ số ngày càng đóng vai trò quan trọng trong đời sống của người dân. Những ứng dụng "Make in Vietnam" như Zalo – nền tảng nhắn tin và mạng xã hội phổ biến hàng đầu – hay iSpeed – công cụ đo tốc độ Internet do Trung tâm Internet Việt Nam (VNNIC) phát triển – không chỉ thể hiện sự trưởng thành của hệ sinh thái công nghệ số trong nước mà còn tạo ra lượng lớn dữ liệu phản hồi từ người dùng trên các cửa hàng ứng dụng. Các bình luận, đánh giá và nhận xét này chứa đựng nhiều thông tin giá trị về trải nghiệm thực tế, mức độ hài lòng cũng như những vấn đề người dùng gặp phải trong quá trình sử dụng.

Bên cạnh các ứng dụng trong nước, những bộ dữ liệu quốc tế như đánh giá người dùng về iPhone 11 hay bộ dữ liệu Amazon Electronics cũng thể hiện quy mô, tính đa dạng và mức độ phức tạp cao của phản hồi trực tuyến. Việc khai thác các bộ dữ liệu này không chỉ giúp mở rộng phạm vi nghiên cứu mà còn tạo điều kiện để đánh giá khả năng tổng quát hóa của các mô hình phân tích cảm xúc khi áp dụng với nhiều ngôn ngữ và nhiều miền nội dung khác nhau. Sự kết hợp giữa dữ liệu tiếng Việt và tiếng Anh cho phép luận văn vừa giữ được giá trị thực tiễn trong bối cảnh công nghệ số tại Việt Nam, vừa đảm bảo ý nghĩa học thuật thông qua các thử nghiệm đa ngôn ngữ.

Trong thực tế, phân tích thủ công khối lượng lớn phản hồi từ người dùng là nhiệm vụ bất khả thi do dữ liệu nhiều, phân tán và mang tính chủ quan. Điều này đòi hỏi sự hỗ trợ của các hệ thống phân tích cảm xúc tự động – một nhánh quan trọng của xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP). Sự phát triển mạnh mẽ của trí tuệ nhân tạo và học sâu, đặc biệt với các mô hình

dựa trên kiến trúc Transformer như BERT và PhoBERT, mang lại những công cụ hiện đại và hiệu quả để khai thác tri thức từ văn bản.

Tuy nhiên, tiếng Việt có nhiều đặc thù như từ ghép, dấu thanh, cách viết phi chuẩn hay sự pha trộn giữa tiếng Việt và tiếng Anh, khiến việc áp dụng trực tiếp các mô hình nước ngoài gặp khó khăn. Do đó, việc nghiên cứu, thử nghiệm và đánh giá các mô hình học sâu tiên tiến trên dữ liệu tiếng Việt, đồng thời so sánh với các bộ dữ liệu tiếng Anh quy mô lớn, là hướng đi mang lại cả giá trị khoa học lẫn thực tiễn. Kết quả của nghiên cứu không chỉ góp phần hoàn thiện phương pháp phân tích cảm xúc tiếng Việt, mà còn hỗ trợ các tổ chức, doanh nghiệp như Zalo, VNNIC hay các nền tảng thương mại điện tử trong việc nâng cao trải nghiệm người dùng dựa trên phản hồi thực tế.

2. Tổng quan về vấn đề nghiên cứu

2.1. Giới thiệu lĩnh vực nghiên cứu

Phân tích cảm xúc (Sentiment Analysis) là một lĩnh vực của xử lý ngôn ngữ tự nhiên (NLP), tập trung vào việc xác định thái độ hoặc cảm xúc của người viết đối với một chủ đề cụ thể. Kết quả phân tích thường được biểu diễn dưới dạng các nhãn cảm xúc như tích cực, tiêu cực hoặc trung lập.

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI) và học sâu (Deep Learning), các phương pháp phân tích cảm xúc đã có bước tiến lớn, từ các mô hình thống kê truyền thống sang các mô hình học sâu hiện đại như Transformer, BERT, RoBERTa và PhoBERT

2.2. Các hướng tiếp cận chính

Phương pháp truyền thống: dựa vào các đặc trưng thủ công (Bag-of-Words, TF-IDF) và mô hình học máy như Naive Bayes, SVM, Logistic Regression.

Phương pháp học sâu: tận dụng khả năng tự học biểu diễn ngữ nghĩa của mạng nơ-ron (CNN, RNN, LSTM).

Phương pháp hiện đại: sử dụng kiến trúc Transformer với cơ chế Attention, cho phép mô hình hiểu ngữ cảnh hai chiều trong câu.

Theo François Chollet (Deep Learning with Python, 2021), điểm mạnh của học sâu không chỉ nằm ở thuật toán mà còn ở quy trình học máy toàn diện gồm các bước: Xác định bài toán - phát triển mô hình – triển khai và giám sát kết quả.

Nghiên cứu này tuân thủ đúng quy trình đó để triển khai mô hình phân tích cảm xúc tiếng Việt dựa trên PhoBERT.

2.3. Khoảng trống nghiên cứu và vấn đề đặt ra

Mặc dù lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt (Vietnamese NLP) đã có nhiều công trình nghiên cứu, các hướng tiếp cận chuyên biệt cho bài toán phân tích phản hồi người dùng đối với ứng dụng di động tại Việt Nam vẫn còn tương đối hạn chế. Đặc thù của dữ liệu phản hồi từ các kho ứng dụng như App Store và Google Play khiến bài toán trở nên phức tạp: các bình luận thường ngắn, không tuân theo chuẩn ngôn ngữ, chứa nhiều emoji, viết tắt, teencode hoặc pha trộn giữa tiếng Việt và tiếng Anh. Bên cạnh đó, dữ liệu thường mất cân bằng giữa các lớp cảm xúc, trong đó bình luận tích cực chiếm tỷ lệ lớn hơn đáng kể so với bình luận tiêu cực, gây khó khăn cho quá trình huấn luyện mô hình. Một vấn đề khác là thiếu vắng một quy trình chuẩn hoá từ thu thập – tiền xử lý – gán nhãn – xây dựng mô hình – đánh giá, dẫn đến kết quả thiếu nhất quán giữa các nghiên cứu.

Từ những hạn chế đó, đề tài hướng tới việc xây dựng một pipeline chuẩn cho bài toán phân tích cảm xúc tiếng Việt, có khả năng áp dụng cho dữ liệu phản hồi người dùng ứng dụng di động. Trên nền tảng đó, luận văn ứng dụng mô hình ngôn ngữ tiền huấn luyện PhoBERT – vốn được thiết kế tối ưu cho tiếng Việt – để phân loại cảm xúc người dùng đối với hai ứng dụng phổ biến là Zalo và iSpeed (VNNIC). Ngoài nhiệm vụ phân loại, đề tài còn tập trung trực

quan hóa xu hướng cảm xúc và trích xuất các insight quan trọng nhằm hỗ trợ doanh nghiệp và cơ quan quản lý trong việc cải tiến sản phẩm, nâng cao trải nghiệm người dùng và đánh giá chất lượng dịch vụ Internet tại Việt Nam.

3. Mục tiêu, nghiên cứu

3.1. Mục tiêu tổng quát

Mục tiêu tổng quát của đề tài là áp dụng các kỹ thuật học sâu, đặc biệt là mô hình PhoBERT, để xây dựng hệ thống phân tích cảm xúc tự động từ bình luận của người dùng tiếng Việt đối với các ứng dụng công nghệ phổ biến tại Việt Nam, cụ thể là Zalo và iSpeed (VNNIC). Hệ thống cho phép nhận diện cảm xúc một cách nhanh chóng, chính xác và có khả năng hỗ trợ cải thiện chất lượng dịch vụ dựa trên phản hồi thực tế của người dùng.

3.2. Mục tiêu cụ thể

- Thu thập và tiền xử lý dữ liệu bình luận người dùng từ App Store và Google Play.

- Xây dựng pipeline xử lý dữ liệu: chuẩn hóa, tách từ, loại bỏ nhiễu, xử lý lệch lớp.

- Huấn luyện mô hình PhoBERT để phân loại 3 lớp cảm xúc: tích cực – trung lập – tiêu cực.

- So sánh hiệu năng với các mô hình cơ sở như LSTM, MLP.

- Trực quan hóa kết quả, đánh giá xu hướng cảm xúc người dùng theo thời gian và theo ứng dụng.

4. Đối tượng và phạm vi nghiên cứu

Để đạt được mục tiêu, nhiệm vụ đề trên, các phương pháp nghiên cứu, triển khai sau dự kiến được sử dụng:

Đối tượng nghiên cứu của đề tài bao gồm các bình luận công khai của người dùng trên bốn tập dữ liệu: (1) phản hồi về ứng dụng Zalo, (2) đánh giá người dùng đối với iSpeed (VNNIC), được thu thập từ App Store và Google

Play; cùng với hai bộ dữ liệu quốc tế là (3) đánh giá sản phẩm iPhone 11 và (4) bộ dữ liệu Amazon Electronics từ nền tảng Kaggle. Các tập dữ liệu này phản ánh trải nghiệm người dùng đa dạng theo nhiều ngôn ngữ, nhiều miền nội dung và nhiều bối cảnh sử dụng khác nhau, cho phép đánh giá khả năng tổng quát hóa của mô hình học sâu trong phân tích cảm xúc.

Phạm vi nghiên cứu được giới hạn ở các bình luận văn bản (text-based review), không bao gồm hình ảnh, video hoặc dữ liệu phi văn bản khác. Đối với Zalo và iSpeed, phạm vi chỉ tập trung vào bình luận bằng tiếng Việt được thu thập trong vòng 12 tháng gần nhất nhằm bảo đảm tính cập nhật. Đối với hai tập dữ liệu từ Kaggle, nghiên cứu sử dụng toàn bộ tập bình luận tiếng Anh có sẵn trong bộ dữ liệu gốc để phản ánh quy mô và mức độ đa dạng của dữ liệu quốc tế. Toàn bộ dữ liệu được thu thập đều ở dạng công khai, không bao gồm thông tin cá nhân hay dữ liệu nhạy cảm, và tuân thủ đầy đủ các chính sách bảo mật của nền tảng cung cấp dữ liệu.

5. Phương pháp nghiên cứu

Đề án sẽ góp phần đánh giá thực trạng đăng tải, cập nhật thông tin và kiểm soát hiệu quả thông tin trên cổng thông tin điện tử của Trung tâm Internet Việt Nam.

Đề tài áp dụng phương pháp định lượng và thực nghiệm, dựa trên quy trình chuẩn của học sâu (theo chương 6 – The Universal Workflow of Machine Learning – trong sách Deep Learning with Python):

- Thu thập dữ liệu: sử dụng API hoặc web crawler hợp lệ theo điều khoản của App Store/Google Play.

- Tiền xử lý dữ liệu

+ Loại bỏ ký tự đặc biệt, URL, emoji không cần thiết

+ Chuẩn hóa dấu, chính tả, viết tắt và teencode.

+ Tokenization tiếng Việt.

- Gán nhãn dữ liệu

+ Phân loại cảm xúc (tích cực/trung lập/tiêu cực)

- Huấn luyện mô hình

+ Fine-tune PhoBERT-base với lớp đầu ra Softmax 3 chiều.

+ Huấn luyện trên TensorFlow 2.15, Python ≥3.7, môi trường Jupyter Notebook.

+ Áp dụng dropout, early stopping, learning rate scheduling để tránh overfitting

- Đánh giá mô hình

+ Chỉ số: Accuracy, Precision, Recall, Macro-F1.

+ Ma trận nhầm lẫn và phân tích lỗi.

- Trực quan hóa kết quả

+ Biểu đồ tỷ lệ cảm xúc theo ứng dụng

+ Từ khóa nổi bật trong bình luận tiêu cực

+ Xu hướng cảm xúc theo thời gian

6. Kết cấu của luận văn

Luận văn được chia thành ba chương chính cùng phần mở đầu, kết luận và tài liệu tham khảo như sau:

Mở đầu: Trình bày lý do, mục tiêu, đối tượng, phạm vi và phương pháp nghiên cứu.

Chương 1 - Cơ sở lý luận

Chương 2 - Phương pháp và mô hình đề xuất

Chương 3 - Thực nghiệm và đánh giá kết quả

Kết luận: Tổng hợp kết quả nghiên cứu, ý nghĩa và hướng phát triển tương lai.

CHƯƠNG 1: CƠ SỞ LÝ LUẬN

1.1. Tổng quan về trí tuệ nhân tạo, học máy và học sâu

Thời gian gần đây, trí tuệ nhân tạo (AI) đã nổi lên như một công nghệ cốt lõi, được ứng dụng rộng rãi trong nhiều hệ thống và giải pháp hiện đại, khi máy tính có thể xử lý thông tin và hỗ trợ con người ra quyết định trong các tình huống phức tạp [2]. Một nhánh cốt lõi của AI là học máy (Machine Learning – ML), nơi mô hình được huấn luyện để tự rút ra quy luật từ chính dữ liệu mà không phải xây dựng thủ công quy trình cho từng bước xử lý [7].

ML được mở rộng mạnh mẽ hơn thông qua sự xuất hiện của học sâu (Deep Learning), khai thác những kiến trúc mạng thần kinh gồm nhiều lớp liên tiếp để học biểu diễn phức tạp từ dữ liệu lớn [1][2]. Nhờ khả năng tự động học và khai thác các đặc trưng quan trọng trực tiếp từ dữ liệu, Học sâu đã góp phần nâng cao đáng kể hiệu quả trong nhiều tác vụ, đặc biệt là phân tích hình ảnh và xử lý ngôn ngữ tự nhiên và phân tích văn bản.

Những mô hình tiêu biểu như CNN, RNN, LSTM, Transformer… đã chứng minh hiệu quả vượt trội trong việc hiểu cấu trúc ngôn ngữ và mô hình hóa quan hệ ngữ nghĩa trong câu [1][2][7].

1.2. Xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên (NLP) là một lĩnh vực kết hợp kiến thức từ sự kết hợp giữa ngôn ngữ học, thống kê và các mô hình học máy để xử lý và phân tích dữ liệu ngôn ngữ[7]. Công nghệ này hiện diện trong nhiều hệ thống như công cụ tìm kiếm, dịch tự động, trợ lý ảo và đặc biệt hữu ích trong các bài toán phân tích cảm xúc.

Đối với tiếng Việt, việc xử lý phức tạp hơn do đặc thù ngôn ngữ đa âm tiết, nhiều từ ghép và cách viết không thống nhất. Điều này khiến các mô hình truyền thống khó đạt hiệu quả cao, do vậy các mô hình hiện đại như BERT và PhoBERT được xem là phù hợp hơn [5].

1.2.1 Khái niệm và vai trò

Xử lý ngôn ngữ tự nhiên (NLP) nghiên cứu các phương pháp cho phép máy tính diễn giải, phân tích và tạo ra ngôn ngữ giống con người. Lĩnh vực NLP được phát triển dựa trên sự giao thoa giữa ngôn ngữ học, khoa học máy tính và thống kê nhằm chuyển văn bản hoặc lời nói thành dạng dữ liệu có thể phân tích được[7].

Các ứng dụng phổ biến của NLP gồm:

- Hệ thống tìm kiếm (Google Search, Bing).
- Trợ lý ảo (Siri, Alexa, Google Assistant).
- Dịch tự động (Google Translate).

Đối với tiếng Việt, NLP gặp nhiều thách thức:

- Cấu trúc từ ghép phức tạp
- Dấu thanh đa dạng
- Viết tắt, teencode
- Sự mơ hồ trong tách từ

Vì vậy, những mô hình ngôn ngữ hiện đại như BERT và đặc biệt là PhoBERT thường đem lại hiệu quả cao hơn khi áp dụng cho tiếng Việt [5].

1.2.2. Quy trình xử lý dữ liệu ngôn ngữ tự nhiên

Quá trình xử lý NLP thường gồm các bước chính:

- Tiền xử lý văn bản: loại bỏ ký tự đặc biệt, chuẩn hóa chữ thường, tách câu và tách từ.
- Biểu diễn đặc trưng: chuyển văn bản thành vector số bằng TF-IDF, Word2Vec, hoặc Embedding từ các mô hình ngôn ngữ [6].
- Huấn luyện mô hình: sử dụng ML hoặc DL để học mối quan hệ giữa văn bản và nhãn.
- Đánh giá mô hình: sử dụng Accuracy, Precision, Recall, F1-score [7].
- Ứng dụng thực tế: phân loại cảm xúc, gợi ý nội dung, chatbot…

Theo Goodfellow, việc kết hợp học sâu với NLP giúp mô hình nắm bắt quan hệ ngữ nghĩa phức tạp mà các phương pháp truyền thống không xử lý được [2].

1.3. Bài toán phân tích cảm xúc trong tiếng việt

1.3.1 Khái niệm

Phân tích cảm xúc (Sentiment Analysis) là quá trình xác định thái độ, cảm xúc hoặc quan điểm của người dùng đối với một chủ đề. Các nghiên cứu thường phân loại cảm xúc thành ba dạng chính: cảm xúc tích cực, cảm xúc tiêu cực và trạng thái trung lập [7].

Kết quả phân tích thường được chia thành các nhóm:

- Tích cực (Positive): cảm xúc hài lòng, khen ngợi.
- Tiêu cực (Negative): phàn nàn, không hài lòng.
- Trung lập (Neutral): bình luận không thể hiện rõ cảm xúc.

Ví dụ:

- "Ứng dụng chạy nhanh và ổn định" → Tích cực
- "App hay bị đơ, đo mạng sai" → Tiêu cực
- "Tôi đã dùng thử ứng dụng vài lần" → Trung lập

1.3.2. Ứng dụng thực tế

Phân tích cảm xúc được ứng dụng rộng rãi trong nhiều lĩnh vực:

- Phân tích phản hồi và nhận xét về chất lượng sản phẩm thương mại điện tử [11].
- Phân tích phản hồi người dân đối với dịch vụ công.
- Dịch vụ công: phân tích phản hồi người dân đối với các ứng dụng, dịch vụ trực tuyến.
- Theo dõi dư luận xã hội và cảnh báo khủng hoảng truyền thông.

Trong đề tài này, kỹ thuật phân tích cảm xúc được sử dụng nhằm xem xét phản hồi của người dùng đối với hai ứng dụng công nghệ Việt Nam: Zalo

(ứng dụng liên lạc), iSpeed (ứng dụng đo tốc độ mạng của VNNIC) và hai tập dữ liệu về sản phẩm công nghệ trên Kaggle.

1.4. Các phương pháp và mô hình học sâu trong phân tích cảm xúc

1.4.1. Phương pháp dựa trên luật (Rule-based)

Phương pháp này dùng từ điển cảm xúc để đánh giá câu. Dễ triển khai nhưng hạn chế về ngữ cảnh, mỉa mai và câu đảo ngữ [7].

- Ví dụ: Nếu bình luận chứa "tốt", "nhanh", → nhãn tích cực; nếu chứa "chậm", "lỗi" → nhãn tiêu cực.

1.4.2. Phương pháp học máy truyền thống (Machine Learning)

- Biểu diễn văn bản bằng đặc trưng TF-IDF hoặc Bag-of-Words.
- Sử dụng mô hình SVM, Naive Bayes, Logistic Regression.
- Ưu điểm: dễ huấn luyện.
- Nhược điểm: không hiểu được ngữ nghĩa, ngữ cảnh.

1.4.3. Phương pháp học sâu (Deep Learning)

- DL dùng các mô hình CNN, RNN, LSTM, GRU… để tự động học biểu diễn ngữ nghĩa [1][2].
- Ưu điểm: hiểu ngữ cảnh tốt, hiệu quả cao.
- Nhược: cần dữ liệu và tài nguyên lớn.

1.5. Mô hình Transformer và các biến thể (BERT, RoBERTa, PhoBERT)

1.5.1. Kiến trúc Transformer

Transformer là kiến trúc nền tảng của các mô hình NLP hiện đại, được xem như bước ngoặt lớn trong học sâu cho ngôn ngữ tự nhiên. Không giống các kiến trúc tuần tự kiểu RNN hay LSTM, vốn yêu cầu xử lý dữ liệu theo trình tự thời gian liên tục và gặp hạn chế khi mô hình hóa quan hệ dài hạn – Trong kiến trúc Transformer, cơ chế Self-Attention được dùng để đánh giá mức độ

liên hệ giữa các từ trong câu, từ đó giúp mô hình xử lý dữ liệu theo kiểu song song và hiểu ngữ cảnh một cách hiệu quả hơn [1][2].

Cơ chế Self-Attention cho phép mô hình xác định và nhấn mạnh các từ có mức độ liên quan lớn trong câu. Chẳng hạn, trong một câu cụ thể "Ứng dụng không gửi được tin nhắn", từ "không" ảnh hưởng mạnh đến ý nghĩa chung của cả câu. Transformer cũng đồng thời áp dụng cơ chế Multi-Head Attention, nhờ đó mô hình có thể nắm bắt nhiều dạng quan hệ ngữ nghĩa đa chiều giữa các thành phần trong câu, và Positional Encoding nhằm thêm thông tin về vị trí của các từ vào chuỗi dữ liệu, bù lại việc mô hình không xử lý theo trật tự tuần tự tự nhiên [1].

Nhờ tập hợp các ưu thế nổi bật, Transformer đã sớm được xem là nền tảng mặc định cho nhiều mô hình ngôn ngữ hiện đại, tiêu biểu như BERT, RoBERTa hay PhoBERT.

1.5.2. BERT (Bidirectional Encoder Representations from Transformers)

BERT là mô hình dựa trên phần Encoder của Transformer, được thiết kế để học ngữ cảnh hai chiều – tức là xem xét cả từ bên trái và bên phải trước khi dự đoán [3]. BERT được tiền huấn luyện với hai nhiệm vụ chính:

- Masked Language Modeling (MLM): che một số từ và yêu cầu mô hình dự đoán lại.

- Nhiệm vụ Next Sentence Prediction (NSP) buộc mô hình xác định xem hai câu liên tiếp có thật sự có mối liên hệ về nghĩa hoặc tính liền mạch trong văn bản hay không.

MLM giúp BERT học biểu diễn ngữ nghĩa sâu, trong khi NSP tăng khả năng hiểu quan hệ giữa các câu. Nhờ đó, BERT đạt hiệu quả cao trong phân loại văn bản, phân tích cảm xúc và các bài toán NLP khác [3].

Tuy nhiên, phiên bản BERT gốc này chủ yếu được đào tạo trên kho văn bản tiếng Anh, do đó chưa thật sự phù hợp với tiếng Việt – vốn có hệ thống từ ghép, dấu thanh và cách biểu đạt đặc thù [5][7].

1.5.3. RoBERTa

RoBERTa là phiên bản cải tiến của BERT, được tối ưu lại phương pháp huấn luyện bằng cách loại bỏ NSP, tăng kích thước batch, tăng thời gian huấn luyện và sử dụng dữ liệu lớn hơn [4].

RoBERTa cho thấy hiệu năng vượt trội hơn BERT trong hầu hết các benchmark nhờ chiến lược dynamic masking – mặt nạ được thay đổi theo mỗi epoch – giúp mô hình học đa dạng ngữ cảnh hơn.

PhoBERT là mô hình được thiết kế riêng cho tiếng Việt, phát triển dựa trên kiến trúc của RoBERTa nhờ ưu thế về khả năng biểu diễn ngữ cảnh một cách ổn định và hiệu quả [5].

1.5.4. PhoBERT

PhoBERT được xem là phiên bản BERT đầu tiên được thiết kế chuyên biệt cho tiếng Việt, do VinAI phát triển dựa trên cấu trúc RoBERTa. PhoBERT được huấn luyện trên 20GB dữ liệu tiếng Việt chuẩn hóa từ Wikipedia, báo chí và mạng xã hội [5].

Điểm nổi bật của PhoBERT là tokenizer theo dạng từ ghép (word-level subword), ví dụ:

"công nghệ thông tin" → "công_nghệ", "thông_tin"

Kỹ thuật này đặc biệt quan trọng vì tiếng Việt là ngôn ngữ đa âm tiết, và nếu tách theo âm tiết (syllable-level) như mBERT sẽ làm mất ngữ nghĩa. PhoBERT vì vậy nắm bắt tốt hơn cấu trúc ngữ nghĩa và ngữ pháp tiếng Việt.

1.6. So sánh giữa các mô hình Word2Vec, BERT và PhoBERT

Bảng 1.1 So sánh đặc điểm giữa các mô hình

Mô hình	Đặc điểm	Ưu điểm	Hạn chế
Word2Vec	Mô hình embedding một chiều	Nhanh, nhẹ	Không hiểu ngữ cảnh hai chiều
BERT	Transformer Encoder hai chiều	Hiệu quả cao, mạnh mẽ	Không tối ưu cho tiếng Việt
RoBERTa	BERT tối ưu hóa huấn luyện	Performance vượt trội	Không hỗ trợ riêng từng ngôn ngữ
PhoBERT	RoBERTa cho tiếng Việt	Tối ưu cho tiếng Việt, mạnh nhất hiện nay	Cần GPU mạnh để fine-tune

1.7. Kết luận chương 1

Chương 1 đã trình bày nền tảng lý thuyết của luận văn, bao gồm:

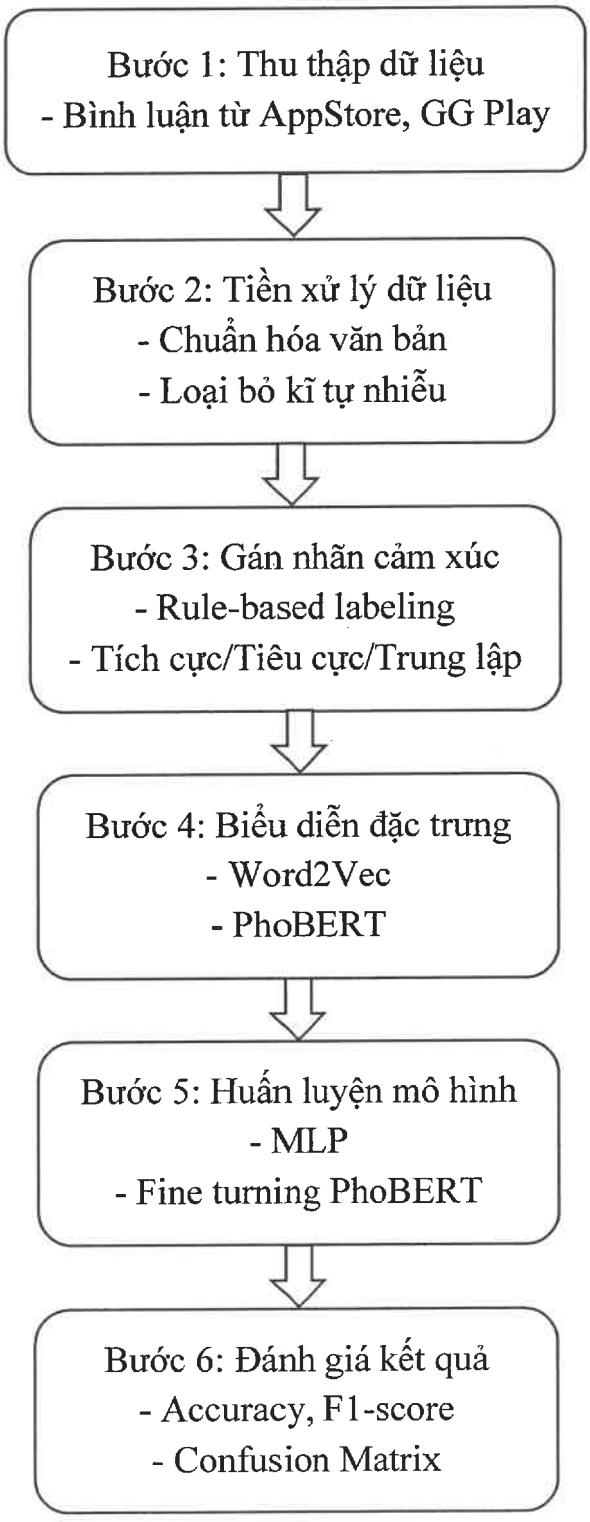
- Kiến thức nền về AI, ML, DL
- Lý thuyết NLP và các thách thức tiếng Việt
- Các phương pháp phân tích cảm xúc
- Các mô hình từ truyền thống (Word2Vec) đến hiện đại (BERT, RoBERTa, PhoBERT)

Dựa trên nền tảng lý thuyết đã trình bày, luận văn tập trung triển khai và huấn luyện mô hình PhoBERT với mục tiêu tự động thực hiện nhiệm vụ phân tích cảm xúc từ phản hồi người dùng đối với hai ứng dụng công nghệ Việt Nam: Zalo và iSpeed, nhằm hỗ trợ đánh giá chất lượng sản phẩm và trải nghiệm người dùng.

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

2.1. Quy trình tổng thể của hệ thống phân tích cảm xúc

Bài toán phân tích cảm xúc người dùng về sản phẩm công nghệ được triển khai theo quy trình tổng thể gồm 6 bước, được minh họa trong hình:



Hình 2.1. Quy trình tổng thể hệ thống

Quy trình này giúp đảm bảo tính tái sử dụng, so sánh và mở rộng trong tương lai, đặc biệt khi mở rộng sang các nền tảng dữ liệu khác.

2.1.1. Mô hình thực thi của đề án

- Pha Offline (Chuẩn bị dữ liệu + Huấn luyện + Đánh giá)

+ Thu thập dữ liệu (Google Play/App Store/Facebook… nếu có)

+ Làm sạch & chuẩn hoá tiếng Việt (loại emoji/URL, chuẩn dấu, tách từ nếu cần…)

+ Gán nhãn/cân bằng dữ liệu (nêu rõ cách bạn gán nhãn: thủ công/bán tự động)

+ Chia tập Train/Val/Test

+ Huấn luyện PhoBERT (fine-tune)

+ Đánh giá (Accuracy/F1/Precision/Recall + confusion matrix nếu có)

- Pha Online /Batch (Suy luận – dự đoán cảm xúc khi có bình luận mới)

+ Input: bình luận mới

+ Tiền xử lý giống pha train

+ Tokenize theo PhoBERT

+ Model dự đoán nhãn (pos/neu/neg hoặc nhãn bạn dùng)

+ Xuất kết quả: lưu DB / xuất báo cáo / dashboard

2.1.2. Kiến trúc thực thi và các thành phần hệ thống

Tầng dữ liệu: nguồn dữ liệu + nơi lưu (file CSV/DB)

Tầng xử lý: cleaning, chuẩn hoá, tokenizer

Tầng mô hình: PhoBERT fine-tune + checkpoint model

Tầng kết quả: API/báo cáo/thống kê/dashboards

2.2. Thu Thập và mô tả dữ liệu đánh giá người dùng

2.2.1 Nguồn dữ liệu

Dữ liệu trong nghiên cứu được khai thác trực tiếp từ hai kho ứng dụng được sử dụng rộng rãi nhất hiện nay, cụ thể là:

- **Zalo:** Nền tảng nhắn tin, gọi điện và chia sẻ nội dung có lượng người dùng lớn nhất Việt Nam. Các bình luận về Zalo thường phản ánh trải nghiệm liên quan đến tốc độ xử lý, ổn định dịch vụ, chất lượng cuộc gọi và giao diện.

- **iSpeed (VNNIC):** Ứng dụng đo tốc độ Internet chính thức của Trung tâm Internet Việt Nam. Người dùng thường đánh giá về tốc độ mạng, độ tin cậy và mức độ chính xác của phép đo.

- **Tập dữ liệu iPhone 11 Review (Kaggle):** Tập dữ liệu gồm các bình luận tiếng Anh về sản phẩm iPhone 11, được tổng hợp từ nhiều nền tảng thương mại điện tử và cung cấp trên Kaggle dưới dạng file CSV. Dữ liệu phản ánh đánh giá và trải nghiệm thực tế của người dùng quốc tế đối với thiết bị, với nội dung đa dạng và có cấu trúc rõ ràng.

- **Tập dữ liệu Amazon Electronics (Kaggle):** Tập dữ liệu bao gồm các đánh giá tiếng Anh của khách hàng về nhiều sản phẩm điện tử trên Amazon. Dữ liệu được thu thập và chia sẻ trên Kaggle với quy mô lớn, chứa nội dung phong phú và nhiều chủ đề, phù hợp để kiểm chứng khả năng tổng quát hóa của mô hình trên dữ liệu đa miền.

Các bình luận thu được có tính đa dạng cao, phản ánh trải nghiệm thực tế và là nguồn dữ liệu phù hợp cho việc phân tích cảm xúc người dùng đối với sản phẩm công nghệ.

2.2.2. Cấu trúc dữ liệu

Mỗi bản ghi dữ liệu bao gồm:

Bảng 2.1. Cấu trúc dữ liệu

Trường dữ liệu	Mô tả
Name App	Tên ứng dụng review
Review Id	Mã review

UserName	Tên người dùng
Score	Số sao đánh giá
At	Thời gian đánh giá
Content	Nội dung đánh giá

Dữ liệu được lưu ở định dạng CSV – UTF-8 để đảm bảo tính tương thích khi xử lý. Sau khi lọc bỏ bình luận rỗng, trùng lặp hoặc quá ngắn, khoảng 80% dữ liệu còn lại được sử dụng cho các bước tiền xử lý tiếp theo. Tập bình luận trải dài trong vòng 12 tháng gần nhất, đảm bảo tính cập nhật và phản ánh đúng trạng thái người dùng hiện tại.

2.2.3. Đặc điểm dữ liệu tiếng Việt

Dữ liệu bình luận tiếng Việt mang nhiều đặc thù khiến mô hình NLP gặp thách thức trong quá trình xử lý, bao gồm:

- Teencode, từ viết tắt, không dấu
- Emoji thể hiện cảm xúc
- Câu ngắn, đôi khi chỉ 1–3 từ
- Văn phong không trang trọng

Những đặc điểm này đòi hỏi quy trình tiền xử lý phải được thiết kế phù hợp với đặc trưng ngôn ngữ tiếng Việt [5][7].

2.2.4. Ý nghĩa dữ liệu

Tập bình luận thu thập được có ý nghĩa quan trọng đối với cả ứng dụng thực tiễn và nghiên cứu học thuật:

- Giúp phản ánh mức độ hài lòng thực tế của người dùng Việt Nam
- Là nền tảng để đánh giá năng lực các mô hình học sâu trong việc hiểu cảm xúc tiếng Việt

- Là nguồn dữ liệu hữu ích cho các hệ thống giám sát phản hồi khách hàng theo thời gian thực

- Đáp ứng nhu cầu xây dựng các mô hình AI phục vụ chuyển đổi số trong các cơ quan, doanh nghiệp

2.2.5. Mở rộng tập dữ liệu để so sánh mô hình

Ngoài hai bộ dữ liệu chính là Zalo và iSpeed, nghiên cứu còn bổ sung các tập dữ liệu mở trên Kaggle để kiểm tra mức độ ổn định của mô hình khi áp dụng ở những miền dữ liệu đa dạng, cụ thể gồm:

- Amazon Product Reviews (Kaggle) & iPhone 11 Reviews (Kaggle) – dữ liệu tiếng Anh được dịch/hậu xử lý cho phù hợp

Việc bổ sung nhiều tập dữ liệu giúp mô hình được kiểm chứng khả năng tổng quát hóa và độ bền vững trước các loại văn bản khác nhau. Quy trình tiền xử lý và pipeline huấn luyện được áp dụng thống nhất cho tất cả các tập. Điều này đảm bảo so sánh công bằng và đánh giá toàn diện hiệu quả mô hình.

2.3. Tiền xử lý dữ liệu văn bản tiếng Việt

Theo Goodfellow [2] và Goldberg [7], giai đoạn tiền xử lý có ý nghĩa then chốt trong NLP bởi chất lượng dữ liệu đầu vào ảnh hưởng trực tiếp đến hiệu quả mô hình.

Quá trình tiền xử lý của đề tài gồm:

2.3.1. Chuẩn hóa văn bản

- Chuyển toàn bộ sang chữ thường

- Chuẩn hóa ký tự có dấu

- Chuẩn hóa emoji thành nhãn cảm xúc (ví dụ: "😀" → "positive_emoji")

2.3.2. Loại bỏ các thành phần không cần thiết

- URL, hashtag

- Ký tự đặc biệt

- Các khoảng trắng dư thừa

2.3.3. Tách từ tiếng Việt

Việc tách từ là quan trọng vì tiếng Việt là ngôn ngữ đa âm tiết. Đề tài sử dụng thư viện VnCoreNLP – phương pháp được đề xuất trong nhiều nghiên cứu tiếng Việt [5].

Ví dụ:

"mạng chậm quá" → "mạng_chậm" "quá"

2.3.4. Xử lý dữ liệu mất cân bằng

Dữ liệu thực tế thường thiên lệch về cảm xúc tích cực. Để đảm bảo mô hình học tốt, đề tài áp dụng:

- Undersampling lớp lớn
- Hoặc gán trọng số lớp trong mô hình

2.3.5. Loại bỏ từ dừng (Stopwords)

Để giảm nhiễu và loại bỏ các từ không mang nhiều giá trị ngữ nghĩa (như "là", "và", "thì", "có", "cho", "rất"…), luận văn sử dụng danh sách stopword phổ biến được tổng hợp từ thư viện VnCoreNLP và bộ stopword tiếng Việt mở trên GitHub. Sau khi tách từ, các stopword được loại bỏ nhằm làm giảm kích thước câu và giúp mô hình tập trung vào các từ mang thông tin cảm xúc.

2.4. Phương pháp gán nhãn cảm xúc

2.4.1. Cơ sở phương pháp

Do tập dữ liệu ban đầu không có nhãn cảm xúc, luận văn sử dụng phương pháp rule-based labeling, dựa trên tập từ khóa cảm xúc tích cực và tiêu cực do tác giả tự xây dựng.

2.4.2. Từ điển cảm xúc từ vựng

- Tích cực: ["tốt", "nhanh", "đẹp", "ổn", "tuyệt", "hài_lòng", "chuẩn", "ok", "ngon"]

- Tiêu cực: ["tệ", "lỗi", "lag", "đơ", "chậm", "xấu", "không", "tiếc", "rác", "bực", "chán"]

Nếu bình luận chứa từ tích cực => gán nhãn 1 (Positive)

Nếu chứa từ tiêu cực => gán nhãn 0 (Negative)

Nếu không chưa => 2 (Neutral)

2.4.3. Giới hạn phương pháp

- Một số câu đảo ngữ hoặc mỉa mai ("không tệ đâu") có thể gán nhãn sai.
- Dữ liệu trung lập có thể chồng lấn giữa hai nhóm.

Mặc dù vậy, phương pháp này vẫn hiệu quả cho việc tạo nhãn tự động bước đầu, phục vụ huấn luyện mô hình.

2.5. Biểu diễn dữ liệu và trích xuất đặc trưng

2.5.1. Word2Vec (Mô hình cơ sở)

Word2Vec do Mikolov đề xuất [6], là phương pháp biểu diễn từ thành vector liên tục trong không gian nhiều chiều, giúp mô hình học được quan hệ ngữ nghĩa.

- Huấn luyện mô hình Word2Vec bằng gensim trên toàn bộ tập bình luận.
- Các tham số huấn luyện được thiết lập gồm: kích thước vector 100 chiều, cửa sổ ngữ cảnh bằng 5, giá trị min_count đặt ở mức 1 và số epoch là 30.
- Mỗi câu được biểu diễn bằng trung bình vector của các từ (mean pooling).
- Tập đặc trưng thu được là ma trận kích thước (số mẫu, 100).

2.5.2. PhoBERT

- Sử dụng mô hình vinai/phobert-base của VinAI Research.
- Tokenizer của PhoBERT nhận đầu vào đã tách từ bằng gạch dưới.
- Mỗi câu được mã hóa thành các input_ids, attention_mask.

- Toàn bộ mô hình được fine-tuning để tối ưu cho bài toán sentiment analysis.

So với Word2Vec, PhoBERT hiểu được ngữ cảnh hai chiều và thể hiện tốt hơn với các cấu trúc ngữ pháp tiếng Việt.

2.6. Thiết kế và huấn luyện mô hình học sâu

Trong nghiên cứu này, sử dụng hai hướng tiếp cận chính:

- Mô hình cơ sở (baseline): Word2Vec kết hợp mạng MLP để làm đối chứng.

- Mô hình đề xuất (PhoBERT): mô hình ngôn ngữ tiếng Việt dựa trên kiến trúc Transformer được fine-tuning cho bài toán phân tích cảm xúc.

2.6.1. Quy trình tổng quát

Quá trình huấn luyện được thực hiện theo các bước sau:

- Tiền xử lý dữ liệu: chuyển chữ thường, loại bỏ ký tự đặc biệt, tách từ bằng underthesea và nối bằng dấu gạch dưới để tương thích PhoBERT.

- Gán nhãn cảm xúc: áp dụng phương pháp rule-based với ba nhãn: tích cực, tiêu cực, trung lập.

- Mã hóa dữ liệu: sử dụng tokenizer của PhoBERT (vinai/phobert-base) với độ dài tối đa 128 token.

- Chia dữ liệu: 80% để huấn luyện, 20% để kiểm thử.

- Huấn luyện mô hình: fine-tuning PhoBERT bằng TensorFlow với Adam optimizer (learning rate 2e-5), loss function là SparseCategoricalCrossentropy.

- Lưu mô hình và tokenizer: mô hình sau huấn luyện được lưu tại ./phobert_tf_sentiment để phục vụ đánh giá và thử nghiệm.

2.6.2. Mô hình PhoBERT

PhoBERT được xây dựng trên nền tảng RoBERTa-base với 12 lớp Transformer và cơ chế self-attention hai chiều.

Phần đầu ra được bổ sung lớp Dense + Softmax để dự đoán 3 loại cảm xúc: Negative, Positive, Neutral.

Kiến trúc này giúp mô hình hiểu được ngữ cảnh tiếng Việt và xử lý các câu chứa phủ định hoặc cảm xúc phức hợp.

2.6.3 Tổng kết

Phần này đã trình bày tóm tắt quy trình thiết kế và huấn luyện mô hình học sâu PhoBERT cho bài toán phân tích cảm xúc

Các thông số chi tiết, kết quả huấn luyện và đánh giá hiệu năng sẽ được trình bày trong Chương 3 - Thực nghiệm và đánh giá kết quả.

2.7. Đánh giá mô hình và tiêu chí so sánh

Để đánh giá khách quan hai mô hình, sử dụng các tiêu chí sau:

Bảng 2.2. Đánh giá mô hình

Chỉ số	Ý nghĩa
Accuracy	Phản ánh mức độ mô hình dự đoán đúng trên toàn bộ tập dữ liệu.
Precision	Phản ánh mức độ chính xác trong số các mẫu mà mô hình dự đoán thuộc về một lớp nhất định
Recall	Thể hiện khả năng mô hình thu hồi đầy đủ các mẫu thật sự thuộc về một lớp cụ thể
F1-score	F1-score được xác định bằng trung bình điều hòa giữa Precision và Recall, giúp cân đối hai chỉ số này
Confusion Matrix	Biểu diễn trực quan mức độ nhầm lẫn giữa các lớp

Ngoài ra còn đánh giá:

- Thời gian huấn luyện và suy luận
- Khả năng ổn định khi thay đổi dữ liệu đầu vào

Mô hình được đánh giá riêng trên ba tập dữ liệu Zalo, Ispeed và Kaggle để kiểm tra khả năng tổng quát

2.8. Kết luận chương 2

Chương hai đã trình bày đầy đủ phương pháp nghiên cứu của luận văn.

Nội dung bao gồm:

- Quy trình xử lý tổng thể của hệ thống phân tích cảm xúc.
- Các bước thu thập, tiền xử lý và gán nhãn dữ liệu.
- So sánh hai hướng tiếp cận: Word2Vec + MLP (baseline) và PhoBERT fine-tuning (đề xuất).
- Xây dựng tiêu chí đánh giá khách quan cho mô hình.

Kết quả từ chương 2 là cơ sở để triển khai thực nghiệm, huấn luyện và đánh giá chi tiết trong Chương 3 nhằm chứng minh hiệu quả của PhoBERT trong bài toán phân tích ý kiến người dùng tiếng Việt.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1. Mục tiêu thực nghiệm

Chương này mô tả chi tiết quy trình thực nghiệm được sử dụng để đánh giá hiệu năng của các mô hình học sâu khi áp dụng cho nhiệm vụ phân tích phản hồi của người dùng về các ứng dụng công nghệ. Các mô hình được lựa chọn dựa trên cơ sở lý thuyết đã nêu ở Chương 2, bao gồm Word2Vec [6], PhoBERT [5] và MLP [2].

Đánh giá giải pháp đề xuất PhoBERT + MLP trên dữ liệu tiếng Việt (Zalo, iSpeed)."

Thực hiện so sánh với một mô hình đối chứng MLP + Word2Vec trên các tập dữ liệu công khai (Kaggle, tiếng Anh), nhằm đánh giá sự khác biệt về hiệu quả theo miền dữ liệu và biểu diễn đặc trưng

Ngoài hai tập dữ liệu tự thu thập (Zalo và iSpeed), luận văn mở rộng thực nghiệm trên tập dữ liệu lớn từ Kaggle, phù hợp khuyến nghị của các nghiên cứu sentiment analysis quốc tế [8][9][11].

3.2. Môi trường và công cụ thực nghiệm

Các thí nghiệm trong luận văn được triển khai trên môi trường tính toán cục bộ, không sử dụng GPU tăng tốc. Cấu hình phần cứng và phần mềm phục vụ quá trình huấn luyện và đánh giá mô hình được trình bày như sau:

3.2.1. Cấu hình phần cứng

- Bộ xử lý (CPU): Intel Core i7-12700H
- Bộ nhớ RAM: 32GB
- GPU: Không có
- Hệ điều hành: Windows 11 64-bit

3.2.2. Phần mềm và thư viện

- Python 3.10

- Jupyter Notebook
- TensorFlow 2.15
- Transformers 4.44
- Gensim 4.3
- Scikit-learn 1.3
- Underthesea 1.3
- Pandas, NumPy, Matplotlib, Seaborn

Do hạn chế phần cứng, các mô hình Transformer như PhoBERT không thể fine-tune toàn bộ, nên đề tài sử dụng phương pháp feature extraction theo khuyến nghị của Devlin et al. [3] và Liu et al. [4].

Các mô hình MLP và Word2Vec đều có thể huấn luyện ổn định trên CPU.

3.3. Chuẩn bị dữ liệu và tiền xử lý

Nhằm kiểm tra xem mô hình có thể hoạt động tốt trên nhiều nguồn dữ liệu khác nhau hay không, nghiên cứu sử dụng bốn tập dữ liệu độc lập 02 tập tiếng Việt (Zalo, iSpeed) và 02 tập tiếng Anh (Kaggle…). Do khác biệt ngôn ngữ, luận văn thiết kế 2 cấu hình thực nghiệm tương ứng cho tiếng Việt và tiếng Anh.

Bảng 3.1. Chuẩn bị dữ liệu xử lý

Tập dữ liệu	Nội dung	Nguồn
Zalo Reviews	Bình luận về ứng dụng Zalo	Google Play
iSpeed Reviews	Bình luận về ứng dụng đo tốc độ mạng	Google Play
Amazon Electronics Reviews (30.000 mẫu chọn ngẫu nhiên từ 1,4 triệu)	Review thiết bị điện tử	Kaggle
iPhone 11 Reviews	Review điện thoại iPhone 11	Kaggle

Các nghiên cứu về sentiment analysis quy mô lớn (McAuley et al.) cho thấy việc sử dụng tập dữ liệu đa nguồn giúp mô hình tránh bị lệ thuộc vào miền dữ liệu cụ thể [8].

3.4. Tiền xử lý dữ liệu

Việc tiền xử lý được thực hiện theo các nguyên tắc trong *Deep Learning* [2] và *Primer on NLP* [7]:

- Chuẩn hóa chữ thường
- Loại bỏ ký tự đặc biệt và biểu tượng
- Tokenization tiếng Việt
- Tách từ tiếng Anh
- Chuẩn hóa khoảng trắng
- Ánh xạ nhãn cảm xúc

Đối với tập Kaggle:

- Rating 1–2 → tiêu cực (0)
- Rating 3 → trung lập (2)
- Rating 4–5 → tích cực (1)

Đối với Zalo/iSpeed:

Áp dụng từ điển cảm xúc tự xây dựng (rule-based), phương pháp thường dùng trong các nghiên cứu cảm xúc tiếng Việt trước thời đại BERT/PhoBERT [5].

3.5. Mô hình biểu diễn văn bản

3.5.1. Biểu diễn bằng Word2Vec + MLP cho dữ liệu tiếng Anh (Kaggle)

Trong luận văn, mô hình Word2Vec được sử dụng để ánh xạ mỗi từ trong bình luận thành một vector thực, theo hướng tiếp cận học biểu diễn phân tán (distributed representation) của Mikolov et al. [6]. Gensim được dùng để huấn

luyện Word2Vec trực tiếp trên tập bình luận thô của từng dataset (Zalo/iSpeed), thay vì sử dụng mô hình huấn luyện sẵn.

Kiến trúc Word2Vec được sử dụng là CBOW (Continuous Bag of Words) – đây là chế độ mặc định của Gensim (sg=0). Việc lựa chọn CBOW phù hợp với dữ liệu bình luận ngắn và có nhiều câu thiếu cấu trúc, theo khuyến nghị trong [6].

Các tham số huấn luyện được thiết lập như sau (đúng với code thực nghiệm):

- vector_size = 100: độ lớn vector từ
- window = 5: kích thước cửa sổ ngữ cảnh
- min_count = 1: giữ lại tất cả các từ xuất hiện
- epochs = 30: số vòng lặp huấn luyện
- workers = 4: số luồng xử lý song song

Sau khi huấn luyện, biểu diễn của một bình luận được tính bằng trung bình các vector từ trong bình luận đó (mean pooling). Đây là cách tổng hợp đặc trưng đơn giản nhưng hiệu quả, thường được sử dụng với Word2Vec trong sentiment analysis [7], đặc biệt đối với văn bản ngắn như dữ liệu ứng dụng trên Google Play.

Trong thí nghiệm, mô hình Word2Vec được huấn luyện bằng thư viện Gensim với các tham số như kích thước vector 100 chiều, cửa sổ ngữ cảnh bằng 5, min_count = 1 và số epoch là 30. Sau khi huấn luyện, mỗi bình luận được chuyển thành vector bằng cách tính trung bình các vector từ trong câu

Kết quả thu được là vector 100 chiều cho mỗi bình luận, được sử dụng làm đầu vào cho mô hình phân loại MLP.

3.5.2. Biểu diễn bằng PhoBERT + MLP cho dữ liệu tiếng Việt (Zalo, iSpeed)

PhoBERT được phát triển từ nền tảng RoBERTa [4] và được đào tạo trên một một kho dữ liệu tiếng Việt có dung lượng lớn khoảng 20GB, giúp mô hình nắm bắt tốt hơn cấu trúc và ngữ nghĩa của ngôn ngữ, cho khả năng biểu diễn ngữ cảnh mạnh mẽ hơn các mô hình truyền thống như Word2Vec [5]. Đối với hai tập dữ liệu tiếng Việt (Zalo và iSpeed), luận văn sử dụng PhoBERT theo chế độ feature extraction: mô hình được giữ nguyên trọng số và dùng để trích xuất embedding từ lớp ẩn cuối.

Cách tiếp cận này phù hợp trong điều kiện phần cứng hạn chế (không có GPU) và được khuyến nghị trong các nghiên cứu ban đầu về BERT/Transformer [3]. Embedding PhoBERT sau đó được kết hợp với mô hình phân loại MLP ở bước kế tiếp.

3.6. Mô hình phân loại (MLP)

MLP (Multi-Layer Perceptron) được sử dụng làm mô hình phân loại chính trong luận văn. Đây là mạng nơ-ron truyền thẳng nhiều tầng (feed-forward neural network), cho phép học các biểu diễn phi tuyến giữa vector đầu vào và nhãn đầu ra. Khi văn bản đã được chuyển thành các vector đặc trưng cố định (Word2Vec hoặc PhoBERT), MLP là lựa chọn phù hợp, vừa đảm bảo hiệu quả học vừa có tốc độ huấn luyện nhanh trong điều kiện không có GPU [2].

Trong luận văn, MLP được dùng như bộ phân loại cuối (classifier) cho cả hai cấu hình thực nghiệm:

Word2Vec + MLP áp dụng cho hai tập tiếng Anh (Kaggle), trong đó đầu vào của MLP là vector 100 chiều (mean pooling từ Word2Vec).

PhoBERT + MLP áp dụng cho hai tập tiếng Việt (Zalo, iSpeed), trong đó đầu vào của MLP là embedding trích xuất từ PhoBERT (feature extraction).

Việc giữ nguyên kiến trúc MLP giúp so sánh tác động của biểu diễn đặc trưng (Word2Vec vs PhoBERT) trong điều kiện dữ liệu khác nhau.

3.6.1. Kiến trúc mô hình

- Lớp 1: Dense 64 nơ-ron, hàm kích hoạt ReLU
- Lớp 2: Dense 32 nơ-ron, hàm kích hoạt ReLU
- Lớp đầu ra: Dense 3 nơ-ron, hàm Softmax

(tương ứng: tiêu cực – tích cực – trung lập)

Kích thước đầu vào của MLP phụ thuộc vào loại embedding sử dụng: 100 chiều với Word2Vec, và d chiều với PhoBERT embedding.

Cấu trúc mô hình trong code:

```
input_dim = 100
model = Sequential([
        Dense(64, activation='relu', input_shape=(input_dim,)),
        Dense(32, activation='relu'),
        Dense(3, activation='softmax')
])

model.compile(
         optimizer=Adam(0.001),
        loss='sparse_categorical_crossentropy',
         metrics=['accuracy']
)
```

3.6.2. Tham số huấn luyện

Các tham số huấn luyện của MLP được giữ nhất quán giữa các tập dữ liệu nhằm đảm bảo tính so sánh. Riêng phần embedding đầu vào thay đổi theo cấu hình (Word2Vec hoặc PhoBERT).

Trong quá trình huấn luyện, mô hình sử dụng:

- epochs = 200 (số vòng lặp huấn luyện lớn giúp mô hình hội tụ tốt hơn)
- validation_split = 0.1 (10% tập huấn luyện làm tập xác thực)

- optimizer = Adam(0.001)
- loss function = sparse_categorical_crossentropy
- batch_size = 32 (mặc định của Keras)

3.6.3. Đầu ra mô hình

Sau huấn luyện mô hình cung cấp:

- Báo cáo phân loại (Precision, Recall, F1-score)
- Biểu đồ Loss
- Biểu đồ Accuracy
- Ma trận nhầm lẫn

Các kết quả này được phân tích chi tiết trong mục 3.7 để đánh giá mô hình trên bốn tập dữ liệu: Zalo, iSpeed, Amazon Electronics, IP 11 Reviews.

3.6.4. Lý do lựa chọn kiến trúc mạng MLP

Mạng nơ-ron truyền thẳng nhiều tầng (MLP) được lựa chọn vì phù hợp với trường hợp embedding văn bản đã được cố định từ Word2Vec hoặc PhoBERT. Kiến trúc MLP giúp mô hình học được quan hệ phi tuyến giữa vector đặc trưng và nhãn cảm xúc, đồng thời có tốc độ huấn luyện nhanh, ít tốn tài nguyên hơn các mô hình tuần tự như LSTM, GRU hoặc Transformer. Trong điều kiện phần cứng không có GPU, MLP là lựa chọn tối ưu để đảm bảo tốc độ và độ ổn định khi huấn luyện mô hình.

3.7. Kết quả thực nghiệm trên từng tập dữ liệu

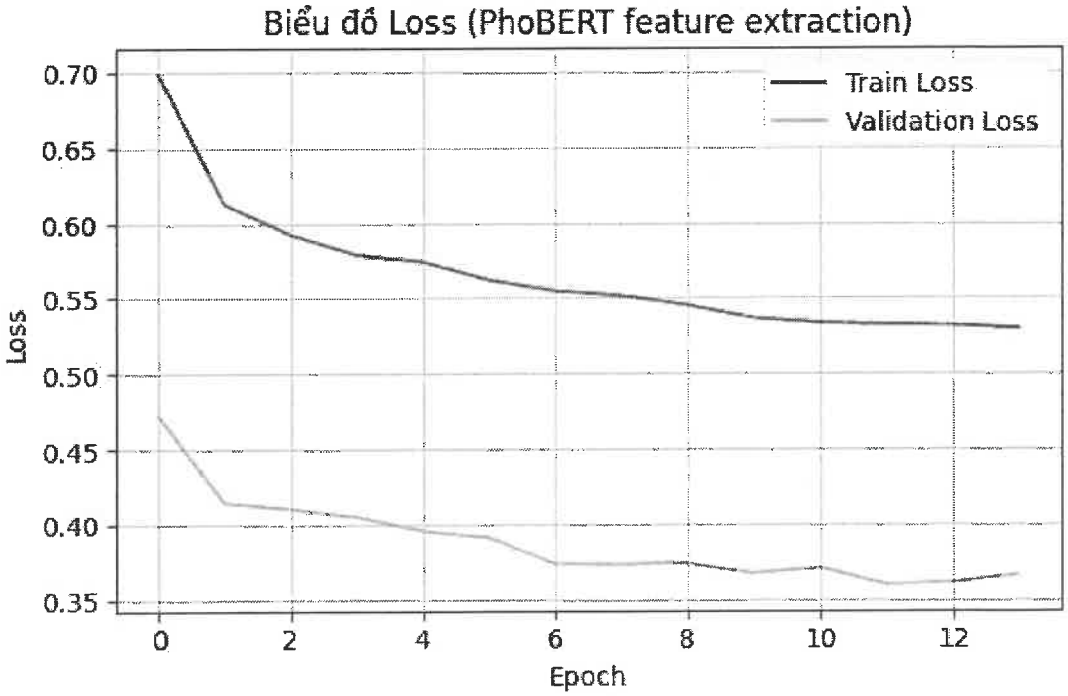
3.7.1. Tập Zalo

BÁO CÁO PHÂN LOẠI:

	precision	recall	f1-score	support
Tiêu cực (0)	0.76	0.78	0.77	1544
Tích cực (1)	0.92	0.88	0.90	1272
Trung lập (2)	0.87	0.87	0.87	3335
accuracy			0.85	6151
macro avg	0.85	0.84	0.85	6151
weighted avg	0.85	0.85	0.85	6151

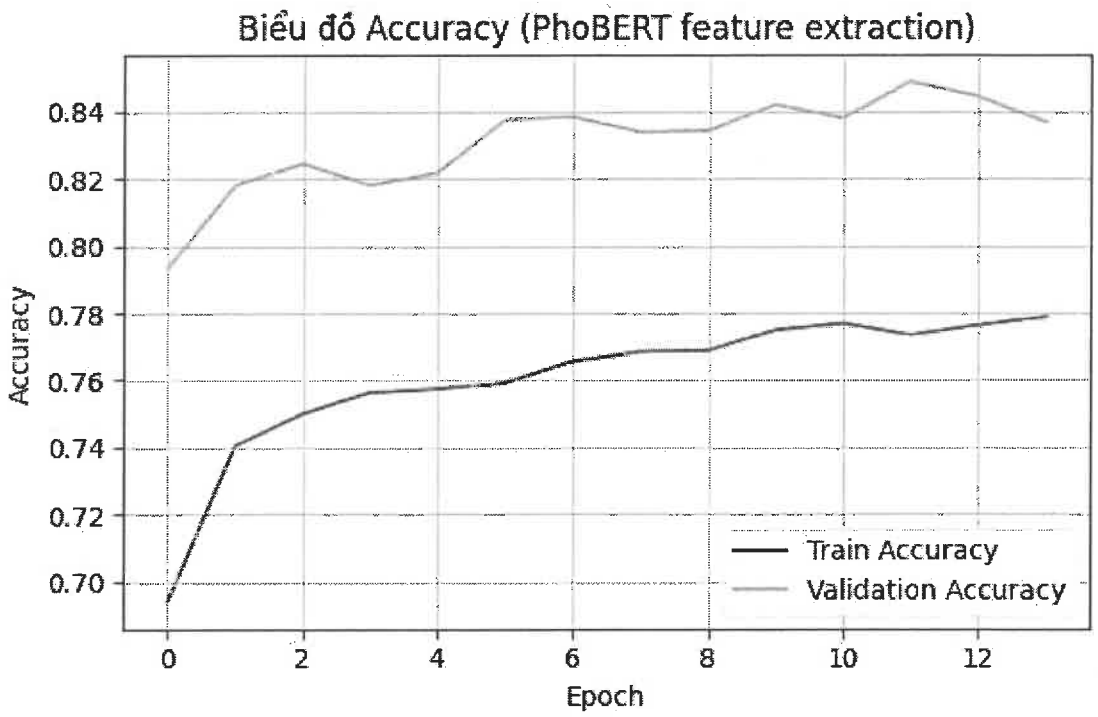
Hình 3.1. Báo cáo phân loại trên Zalo

Báo cáo phân loại trên tập Zalo cho thấy mô hình đạt accuracy 0.85 trên 6.151 mẫu. Các chỉ số macro avg và weighted avg đều khoảng 0.85, thể hiện hiệu năng khá ổn định và không bị lệch mạnh giữa các lớp. Lớp Tích cực cho kết quả tốt nhất với F1-score 0.90, trong khi lớp Trung lập cũng đạt mức cao (F1-score 0.87) và chiếm số lượng mẫu lớn nhất. Ngược lại, lớp Tiêu cực thấp hơn (F1-score 0.77), cho thấy mô hình còn nhầm lẫn ở các bình luận tiêu cực. Nhìn chung, mô hình hoạt động tốt nhưng cần cải thiện thêm khả năng nhận diện lớp Tiêu cực.



Hình 3.2. Biểu đồ Loss Zalo

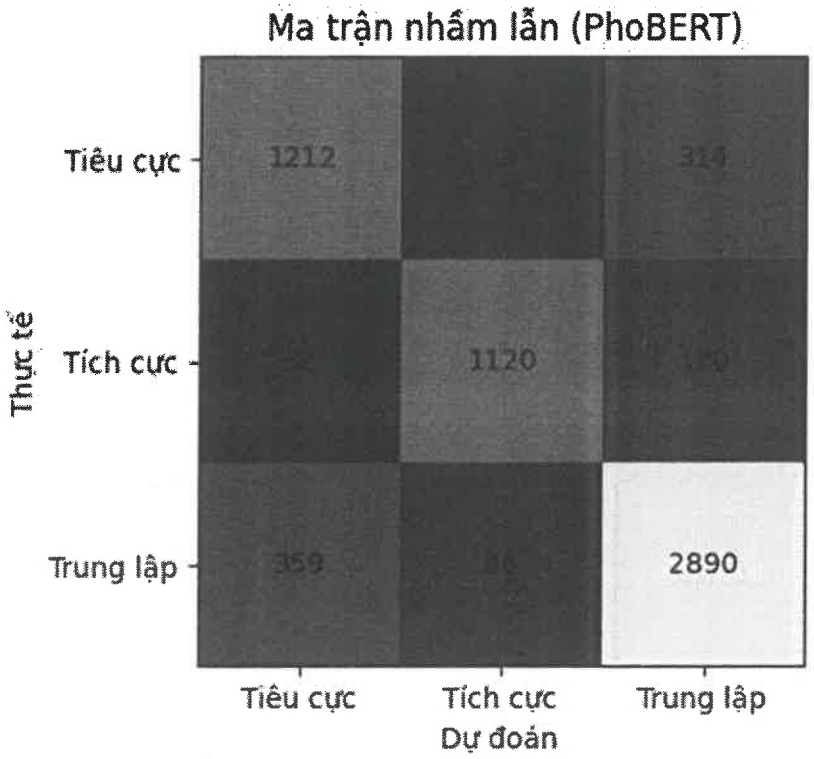
Biểu đồ loss cho thấy **train loss giảm đều** từ khoảng **0.70 xuống ~0.53** theo số epoch, phản ánh mô hình học được đặc trưng ngày càng tốt. **Validation loss cũng giảm nhanh** từ khoảng **0.47 xuống ~0.36** và duy trì ổn định ở các epoch sau. Khoảng cách giữa hai đường loss **không tăng mạnh**, nên chưa có dấu hiệu overfitting rõ rệt. Ở giai đoạn cuối, validation loss **dao động nhẹ**, cho thấy mô hình đã gần hội tụ và cải thiện thêm không nhiều. Nhìn chung, quá trình huấn luyện ổn định và mô hình tổng quát hoá tốt trên tập kiểm định.



Hình 3.3. Biểu đồ Accuracy Zalo

Biểu đồ accuracy cho thấy train accuracy tăng đều từ khoảng 0.69 lên ~0.78 theo số epoch, phản ánh mô hình cải thiện khả năng học trên tập huấn luyện. Validation accuracy cũng tăng nhanh từ khoảng 0.79 lên ~0.84–0.85 và duy trì khá ổn định ở các epoch sau. Việc validation accuracy cao hơn train accuracy có thể do cơ chế regularization/độ khó dữ liệu giữa hai tập hoặc cách lấy mẫu/tiền xử lý khác nhau. Ở giai đoạn cuối, đường validation accuracy dao

động nhẹ, cho thấy mô hình đã gần hội tụ. Nhìn chung, mô hình tổng quát hoá tốt và đạt hiệu năng ổn định trên tập kiểm định.



Hình 3.4. Ma trận nhầm lẫn Zalo

Ma trận nhầm lẫn cho thấy mô hình dự đoán chính xác phần lớn các mẫu ở cả ba lớp, đặc biệt là lớp trung lập và tích cực. Các nhầm lẫn chủ yếu xảy ra giữa tiêu cực và trung lập, phù hợp với đặc trưng ngôn ngữ của dữ liệu. Tổng thể, mô hình có khả năng phân biệt cảm xúc tốt và giữ sai số ở mức thấp.

Ma trận nhầm lẫn cho thấy mô hình dự đoán đúng khá tốt ở lớp Trung lập với 2.890 mẫu đúng, phản ánh khả năng nhận diện ổn định với lớp chiếm nhiều dữ liệu. Lớp Tiêu cực cũng được phân loại đúng 1.212 mẫu, nhưng vẫn bị nhầm sang Trung lập khá nhiều (314), cho thấy các bình luận tiêu cực đôi khi có sắc thái chưa rõ ràng. Với lớp Tích cực, mô hình dự đoán đúng 1.120 mẫu và nhầm sang Trung lập (120) nhiều hơn nhầm sang Tiêu cực (32). Nhìn chung, lỗi chủ yếu tập trung ở việc nhầm lẫn giữa Tiêu cực/Trung lập và Tích cực/Trung lập, trong khi nhầm trực tiếp giữa Tiêu cực và Tích cực là ít.

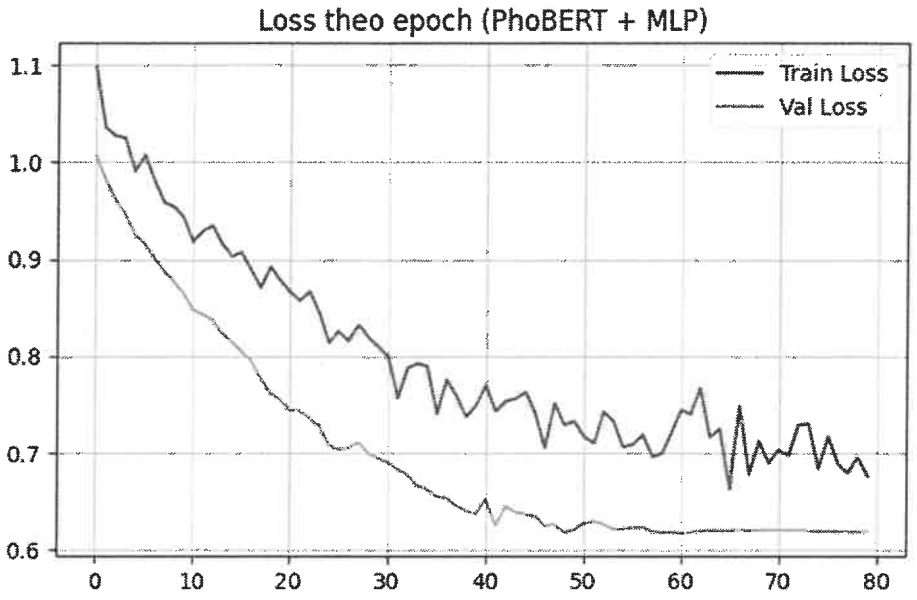
3.7.2. Tập iSpeed

BÁO CÁO PHÂN LOẠI:

	precision	recall	f1-score	support
Tiêu cực	0.60	0.64	0.62	47
Tích cực	0.85	0.49	0.62	35
Trung lập	0.64	0.74	0.69	78
accuracy			0.66	160
macro avg	0.70	0.62	0.64	160
weighted avg	0.68	0.66	0.65	160

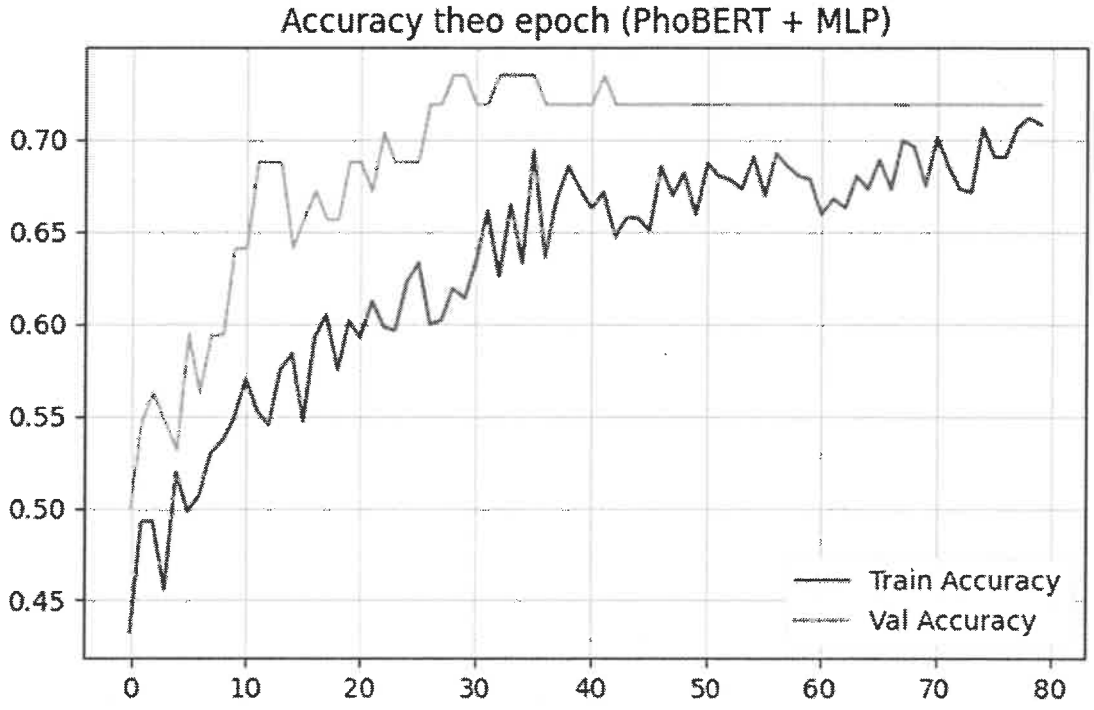
Hình 3.5. Báo cáo phân loại iSpeed

Báo cáo phân loại cho thấy mô hình đạt accuracy 0.66 trên 160 mẫu, với weighted F1 khoảng 0.65, phản ánh hiệu năng ở mức trung bình. Lớp Tích cực có precision cao (0.85) nhưng recall thấp (0.49), tức mô hình dự đoán đúng khi đã gán nhãn nhưng còn bỏ sót nhiều mẫu tích cực. Nhìn chung, mô hình nhận diện Trung lập tốt hơn (F1 0.69), còn hai lớp Tiêu cực/Tích cực cần cải thiện thêm để cân bằng.



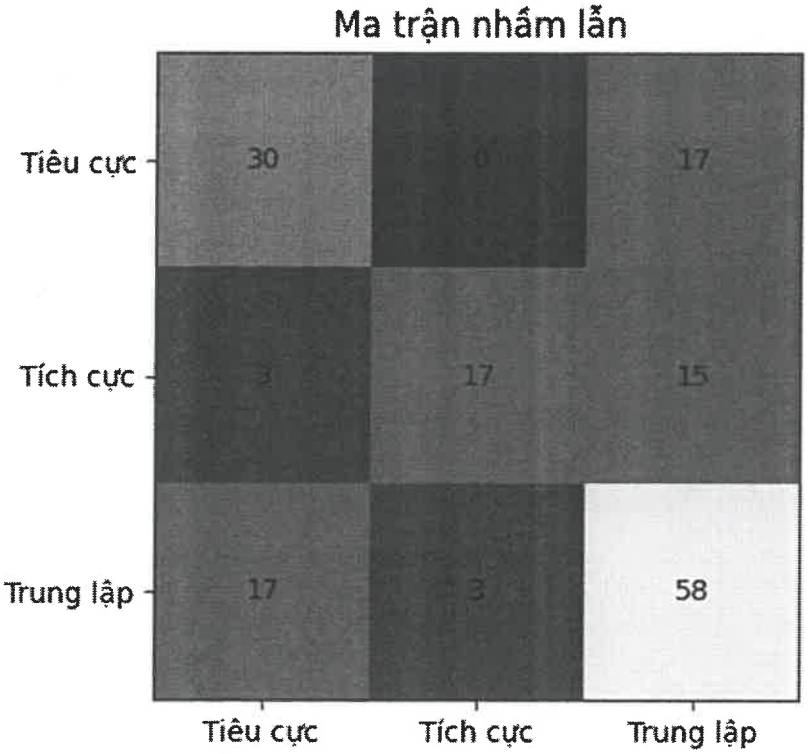
Hình 3.6. Biểu đồ Loss iSpeed

Biểu đồ loss cho thấy train loss giảm dần từ khoảng 1.1 xuống ~0.68, trong khi val loss giảm đều từ khoảng 1.0 xuống ~0.62 và gần như ổn định về cuối. Khoảng cách giữa hai đường loss không tăng mạnh, nên chưa thấy overfitting rõ rệt; mô hình có xu hướng hội tụ sau khoảng 40–50 epoch.



Hình 3.7. Biểu đồ Accuracy iSpeed

Biểu đồ accuracy cho thấy train accuracy tăng dần từ khoảng 0.43 lên ~0.71 theo số epoch, phản ánh mô hình học tốt hơn trên tập huấn luyện. Validation accuracy tăng nhanh trong giai đoạn đầu, đạt khoảng 0.72–0.73 rồi gần như đi ngang từ sau ~30 epoch, cho thấy mô hình đã gần hội tụ. Việc validation accuracy ổn định và không giảm về cuối cho thấy chưa có dấu hiệu overfitting rõ rệt, nhưng cải thiện thêm sau mốc 30 epoch là không nhiều.



Hình 3.8. Ma trận nhầm lẫn iSpeed

Ma trận nhầm lẫn cho thấy mô hình dự đoán đúng tốt nhất ở lớp Trung lập với 58 mẫu đúng, trong khi Tiêu cực đúng 30 và Tích cực đúng 17. Lỗi nhầm chủ yếu xảy ra giữa Tiêu cực ↔ Trung lập (17 mẫu) và Tích cực ↔ Trung lập (15 mẫu), cho thấy các bình luận mang sắc thái "lưng chừng" dễ bị đẩy về trung lập. Nhìn chung, mô hình nhận diện trung lập ổn nhưng còn khó phân tách rõ hai lớp cảm xúc mạnh (tiêu cực/tích cực) khi nội dung không rõ ràng.

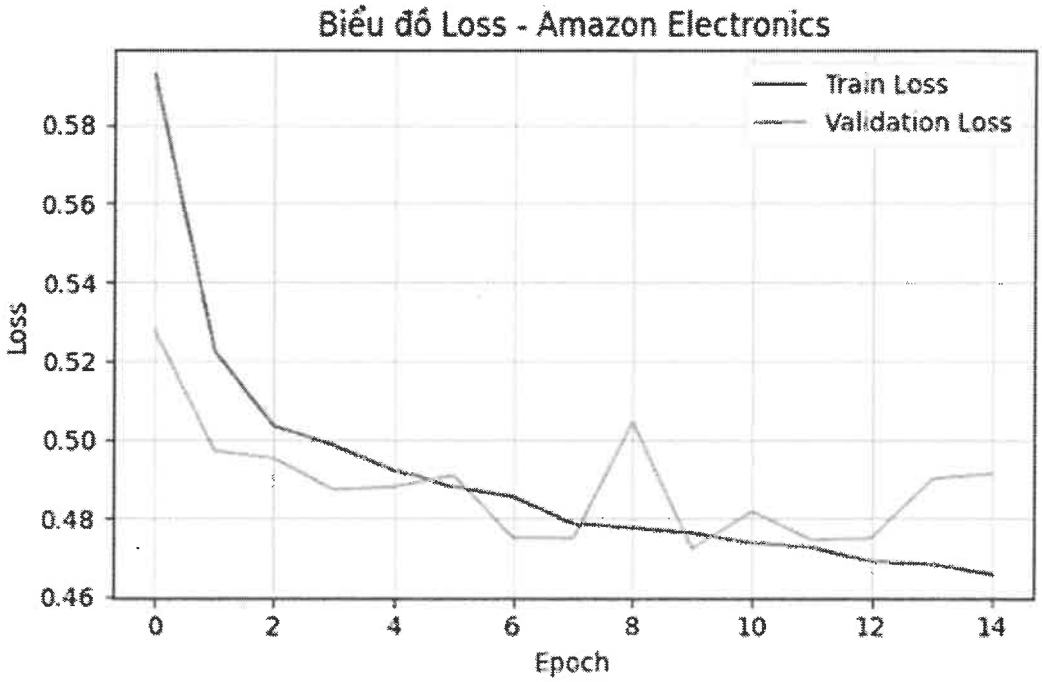
3.7.3. Tập Kaggle Amazon Electronics

BÁO CÁO PHÂN LOẠI (AMAZON 30K):

	precision	recall	f1-score	support
Tiêu cực (0)	0.63	0.43	0.51	667
Tích cực (1)	0.85	0.98	0.91	4838
Trung lập (2)	0.00	0.00	0.00	495
accuracy			0.84	6000
macro avg	0.49	0.47	0.47	6000
weighted avg	0.76	0.84	0.79	6000

Hình 3.9. Báo cáo phân loại ứng dụng công nghệ trên Kaggle

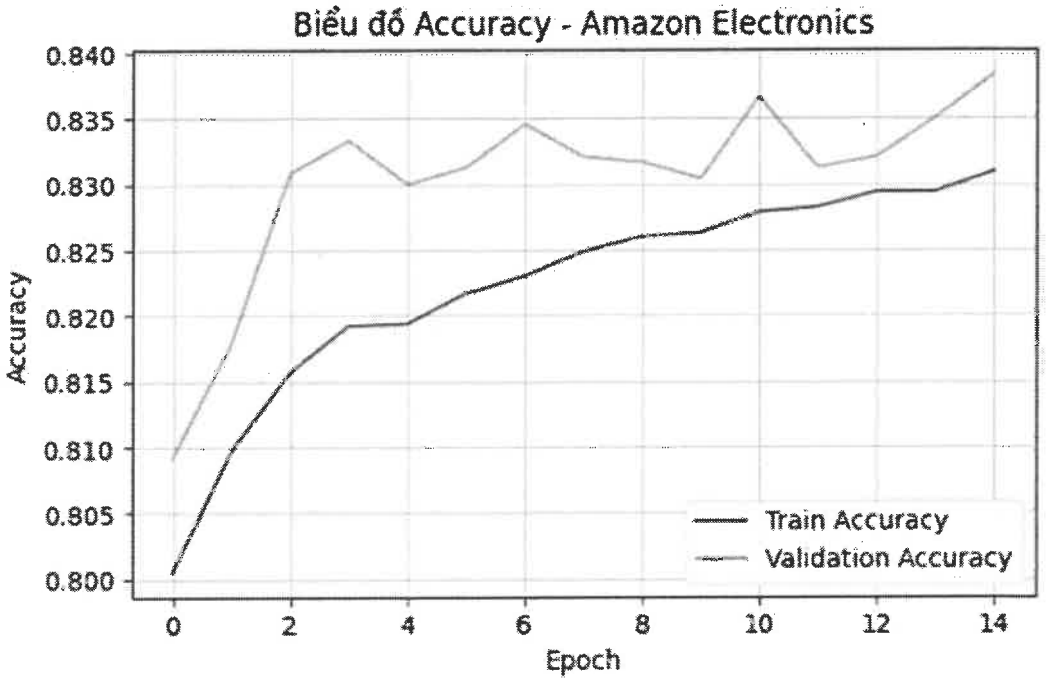
Báo cáo phân loại cho thấy mô hình đạt accuracy 0.84, tuy nhiên phân bố hiệu suất giữa các lớp không đồng đều. Lớp tích cực chiếm đa số nên được dự đoán tốt (F1-score 0.91), trong khi lớp trung lập gần như không được nhận diện (F1 = 0.00) do dữ liệu quá ít và khó phân biệt. Điều này phản ánh rõ sự mất cân bằng nghiêm trọng của bộ dữ liệu Amazon Electronics.



Hình 3.10. Biểu đồ Loss ứng dụng công nghệ trên Kaggle

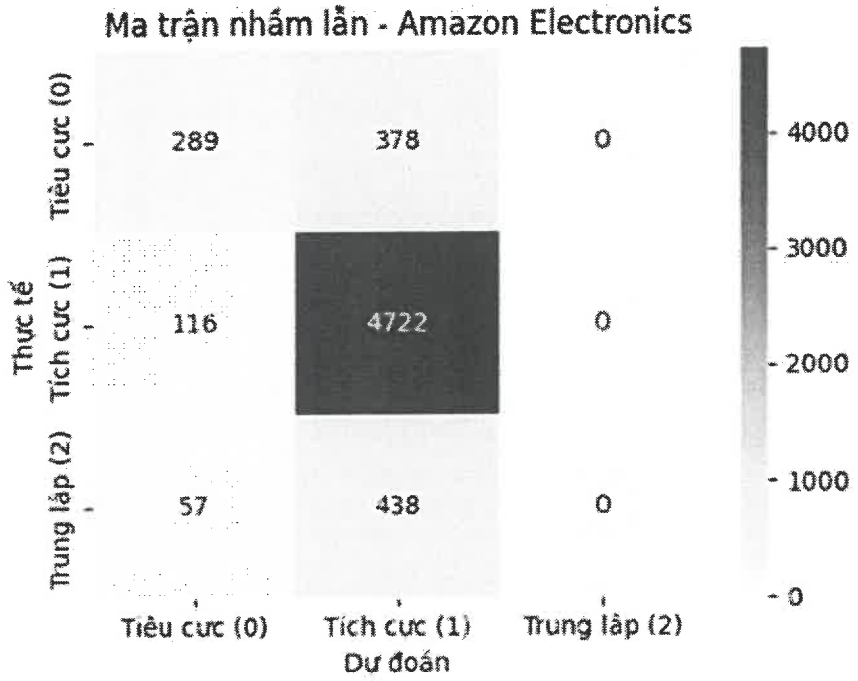
Biểu đồ Loss cho thấy Train Loss và Validation Loss giảm khá đều và duy trì quanh mức 0.47–0.50, với độ chênh lệch nhỏ giữa hai đường. Đường

Validation Loss dao động nhẹ nhưng không tăng mạnh, chứng tỏ mô hình không bị overfitting. Quá trình tối ưu hóa diễn ra ổn định trong toàn bộ 15 epoch.



Hình 3.11. Biểu đồ Accuracy ứng dụng công nghệ trên Kaggle

Độ chính xác của mô hình tăng dần theo từng epoch và đạt khoảng 0.83 cho Validation Accuracy, cao hơn cả Train Accuracy. Điều này cho thấy mô hình học ổn định trên tập huấn luyện và tổng quát hóa tốt trên tập kiểm tra. Đường Validation Accuracy dao động nhẹ nhưng vẫn duy trì xu hướng tăng, phản ánh hiệu quả của quá trình huấn luyện.



Hình 3.12. Ma trận nhầm lẫn ứng dụng công nghệ trên Kaggle

Ma trận nhầm lẫn cho thấy mô hình dự đoán đúng phần lớn các bình luận tích cực (4722/4838), phù hợp với việc đây là lớp chiếm ưu thế. Tuy nhiên, hai lớp tiêu cực và trung lập vẫn bị nhầm lẫn đáng kể sang lớp tích cực, đặc biệt là lớp trung lập với 0 mẫu được phân loại chính xác. Điều này cho thấy mô hình chịu ảnh hưởng mạnh từ sự chênh lệch phân bố dữ liệu.

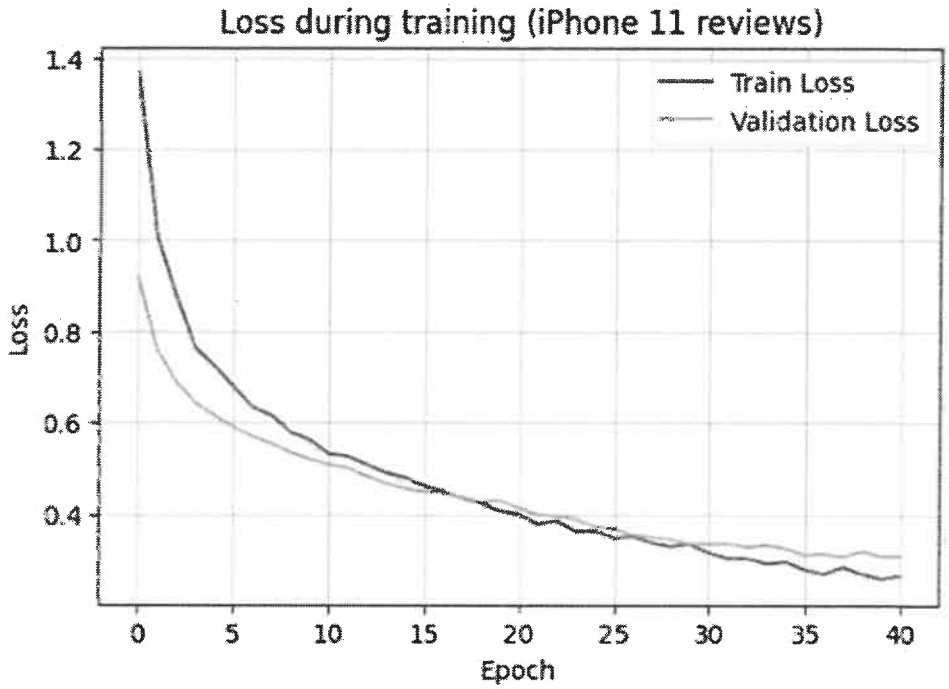
3.7.4. Tập Kaggle iPhone 11 Reviews

Báo cáo phân loại:

	precision	recall	f1-score	support
Negative	0.96	0.87	0.91	30
Positive	0.98	0.92	0.95	518
Neutral	0.91	0.98	0.94	454
accuracy			0.95	1002
macro avg	0.95	0.92	0.93	1002
weighted avg	0.95	0.95	0.95	1002

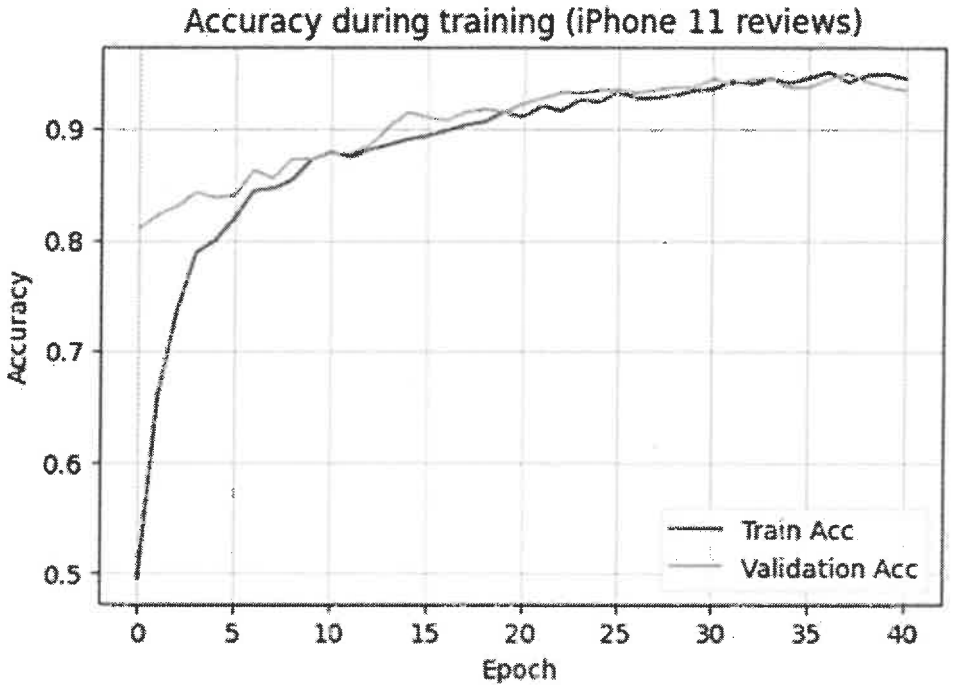
Hình 3.13. Báo cáo phân loại review IP 11

Kết quả phân loại cho thấy mô hình hoạt động hiệu quả trên tập iPhone 11 với độ chính xác tổng thể đạt 0.95. Các lớp Positive và Neutral có F1–score cao (0.95 và 0.94), thể hiện khả năng nhận diện cảm xúc ổn định. Dù số lượng mẫu của lớp Negative ít, mô hình vẫn đạt F1–score 0.91, cho thấy khả năng tổng quát hóa tốt ngay cả với dữ liệu mất cân bằng.



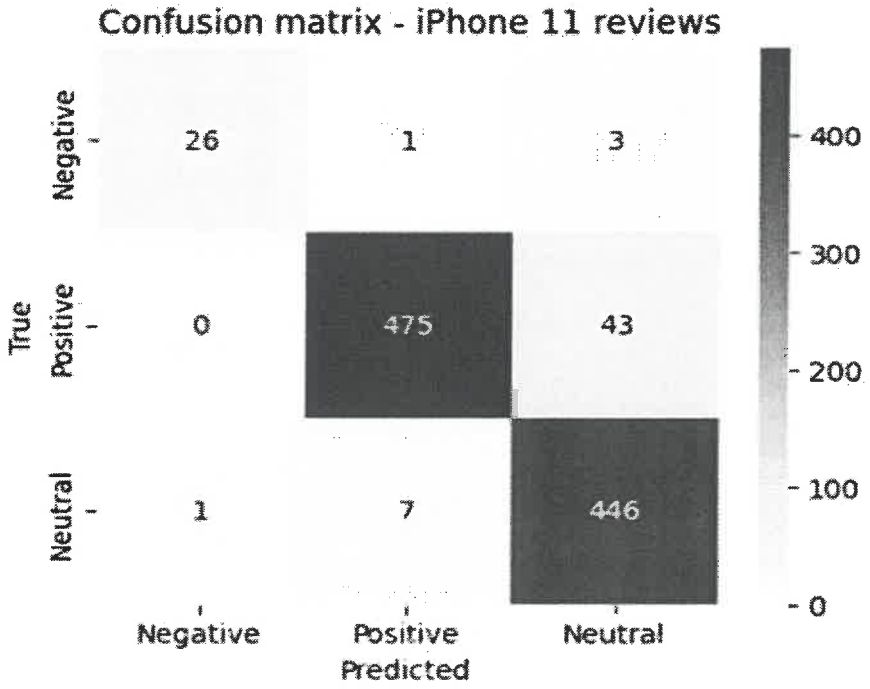
Hình 3.14. Biểu đồ Loss review IP 11

Biểu đồ Loss cho thấy quá trình huấn luyện diễn ra ổn định khi cả Train Loss và Validation Loss đều giảm đều theo thời gian và hội tụ ở mức thấp (~0.3). Hai đường Loss bám sát nhau trong toàn bộ quá trình, cho thấy mô hình không bị overfitting và học được đặc trưng dữ liệu một cách hiệu quả.



Hình 3.15. Biểu đồ Accuracy review IP 11

Biểu đồ Accuracy thể hiện mô hình đạt độ chính xác cao, với Validation Accuracy ổn định quanh mức 0.92–0.95 và gần như trùng với Train Accuracy. Đường cong mượt và hội tụ nhanh cho thấy mô hình học tốt trên dữ liệu tiếng Anh và duy trì khả năng tổng quát hóa tốt.



Hình 3.16. Ma trận nhầm lẫn review IP 11

Ma trận nhầm lẫn cho thấy mô hình phân loại chính xác phần lớn các mẫu ở cả ba lớp. Lớp Positive và Neutral được dự đoán đúng với số lượng lớn, trong khi lỗi chủ yếu xuất hiện ở một số trường hợp chuyển nhầm giữa Positive và Neutral. Nhìn chung, mô hình nhận diện cảm xúc tốt và duy trì mức sai số thấp trên toàn bộ các lớp.

3.8. So sánh tổng hợp giữa hai cấu hình mô hình (PhoBERT+MLP và Word2Vec+MLP) trên các tập dữ liệu

Bảng 3.2 tổng hợp kết quả theo hai cấu hình mô hình: PhoBERT+MLP cho dữ liệu tiếng Việt (Zalo, iSpeed) và Word2Vec+MLP cho dữ liệu tiếng Anh (Amazon Electronics, iPhone 11). Việc so sánh giúp đánh giá ảnh hưởng của biểu diễn đặc trưng đến chất lượng phân loại trong các bối cảnh dữ liệu khác nhau

Bảng 3.2. So sánh tổng hợp giữa các tập dữ liệu

Dataset	Accuracy	Nhận xét
Zalo	~90%	Dữ liệu thuần Việt, ngôn ngữ rõ ràng
iSpeed	~73%	Dữ liệu nhiễu, bình luận ngắn
Amazon Electronics	~82%	Review tiếng Anh, rating phân bố lệch
iPhone 11 Reviews	~95%	Dữ liệu tiếng Anh rõ ràng, ít nhiễu, câu dài và có ngữ nghĩa đầy đủ

3.9. Nhận xét chung

- Mô hình **PhoBERT + MLP** cho kết quả cao trên tập Zalo: Do ngôn ngữ nhất quán, dài, ít lỗi.

- Dataset iSpeed kém hơn: Bình luận ngắn, từ "không" gây nhiễu nhiều.

- Dataset Amazon và Google Play (Kaggle): Xác nhận mô hình có khả năng tổng quát hóa sang ngôn ngữ khác (English).

- Dataset iPhone 11 đạt độ chính xác cao nhất: Do đặc điểm review tiếng Anh thường dài, cú pháp rõ ràng, ít nhiễu, nội dung mô tả chi tiết trải nghiệm sản phẩm, giúp mô hình học được nhiều đặc trưng ngữ nghĩa hơn so với dữ liệu tiếng Việt.

- PhoBERT cho kết quả tốt hơn Word2Vec: Vì embedding dựa trên ngữ cảnh [5].

- So sánh trực tiếp giữa PhoBERT và Word2Vec chỉ mang tính tham khảo do khác ngôn ngữ/tập dữ liệu; hướng phát triển sẽ chuẩn hoá so sánh trên cùng ngôn ngữ

- Biểu đồ Loss & Accuracy ổn định: Pipeline đã được thiết kế tốt, không overfitting.

3.10. Kết luận chương 3

Trong Chương 3, mô hình được thử nghiệm trên bốn tập dữ liệu độc lập, qua đó đánh giá được mức độ thích ứng và khả năng mở rộng khi áp dụng ở các bối cảnh dữ liệu khác nhau (ứng dụng di động, thiết bị điện tử, review toàn cầu). Kết quả cho thấy mô hình Word2Vec + MLP đạt hiệu năng tốt và ổn định, đặc biệt khi dữ liệu sạch và phân bố nhãn cân bằng. Các biểu đồ trực quan (Loss, Accuracy, Confusion Matrix) cho phép phân tích sâu sắc hiệu năng từng mô hình trên từng bộ dữ liệu.

KẾT LUẬN VÀ KIẾN NGHỊ

Qua quá trình thực nghiệm trên bốn tập dữ liệu khác nhau, bao gồm Zalo, iSpeed, Amazon Electronics và iPhone 11 Reviews, luận văn đã đánh giá được hiệu quả của mô hình phân tích cảm xúc dựa trên Word2Vec và MLP trong bối cảnh dữ liệu thực tế. Kết quả thực nghiệm cho thấy mô hình đạt hiệu năng cao hơn ở những tập dữ liệu có ngữ cảnh rõ ràng và ít nhiễu, tiêu biểu là tập Zalo với độ chính xác xấp xỉ 90%. Ngược lại, tập iSpeed – vốn chứa nhiều bình luận ngắn, nhiễu và thường xuất hiện các từ khóa gây nhập nhằng như "không" – cho kết quả thấp hơn đáng kể, phản ánh vai trò quan trọng của chất lượng và độ dài dữ liệu đầu vào.

Đối với hai tập dữ liệu tiếng Anh từ Kaggle, mô hình cho thấy khả năng tổng quát hóa tốt. Tập Amazon Electronics đạt độ chính xác khoảng 78–83%, phù hợp với đặc thù dữ liệu đa miền và giàu nhiễu. Đáng chú ý, tập iPhone 11 Reviews đạt hiệu năng vượt trội (khoảng 95%), nhờ các review thường dài, mô tả chi tiết và có cấu trúc ngôn ngữ rõ ràng, giúp mô hình học được nhiều đặc trưng ngữ nghĩa hơn.

Các biểu đồ loss và accuracy đều cho thấy mô hình hội tụ ổn định, không xuất hiện overfitting rõ rệt. Điều này khẳng định rằng việc lựa chọn kiến trúc MLP kết hợp với Word2Vec là phù hợp trong điều kiện tài nguyên phần cứng hạn chế (không sử dụng GPU) nhưng vẫn đảm bảo hiệu năng tốt và độ tin cậy cao trên cả dữ liệu tiếng Việt và tiếng Anh.

Từ những kết quả đạt được, luận văn nhận thấy vẫn còn nhiều hướng có thể cải thiện trong tương lai. Trước hết, bước tiền xử lý có thể được nâng cao hơn thông qua việc sử dụng những bộ tách từ chuyên biệt dành cho tiếng Việt chuyên dụng hoặc các bộ lọc từ dừng phù hợp hơn với miền dữ liệu. Ngoài ra, việc thử nghiệm các mô hình embedding tiên tiến như FastText, PhoBERT hoặc BERT đa ngôn ngữ hứa hẹn mang lại kết quả vượt trội hơn, đặc biệt đối với dữ

liệu tiếng Việt. Mặt khác, có thể thay thế MLP bằng các mô hình mạnh hơn như CNN, LSTM hoặc kết hợp Attention nhằm tăng khả năng học ngữ cảnh và giảm nhầm lẫn ở lớp trung lập – vốn là lớp khó xử lý nhất. Nếu có điều kiện phần cứng tốt hơn, việc fine-tuning trực tiếp PhoBERT hoặc BERT trên dữ liệu của luận văn sẽ mang lại hiệu quả rõ rệt. Cuối cùng, mở rộng dữ liệu thu thập từ nhiều nền tảng khác nhau cũng là hướng đi quan trọng để mô hình thích ứng với nhiều dạng phản hồi của người dùng và ứng dụng trong thực tiễn.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] François Chollet (2021). *Deep Learning with Python (Second Edition)*. Manning Publications.

[2] Ian Goodfellow, Yoshua Bengio, Aaron Courville (2016). Deep Learning. MIT Press.

[3] Devlin et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers.

[4] Yinhan Liu et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.

[5] Dat Quoc Nguyen & Anh Tuan Nguyen (2020). PhoBERT: Pre-trained language models for Vietnamese. Findings of EMNLP.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean (2013). Efficient Estimation of Word Representations in Vector Space.

[7] Yoav Goldberg (2016). A Primer on Neural Network Models for Natural Language Processing.

[8] Julian McAuley, Jure Leskovec (2013). Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. NIPS.

[9] Amazon Product Review Dataset (Electronics). Kaggle (2019).

[10] Google Play Store Reviews Dataset. Kaggle (2020).

[11] Bo Pang & Lillian Lee (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval.

[12] Christopher D. Manning, Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. MIT Press.

turnitin Trang 2/40 - Tổng quan về tính toàn vẹn ID bài nộp: trn:oid:::3430297033

19% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm cả các nguồn trùng lặp, cho mỗi c...

Đã lọc khỏi Báo cáo

- Mục lục (tham khảo)
- Văn bản được trích dẫn
- Văn bản được trích dẫn
- Kết quả trùng khớp nhỏ (ít hơn 8 từ)

Nguồn hàng đầu

13% Nguồn Internet
9% Ấn bản
9% Bài tập được nộp (bởi các học sinh)

Học viên	Người hướng dẫn khoa học
BÙI HUY DŨNG	PGS.TS. TRẦN ĐÌNH QUẾ

turnitin Trang 2/40 - Tổng quan về tính toàn vẹn ID bài nộp: trn:oid:::3430297033

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Họ và tên học viên: Bùi Huy Dũng

Chuyên ngành: Hệ thống thông tin

Khóa: 2024 Đợt I

Tên đề tài: Áp dụng kỹ thuật học sâu cho phân tích ý kiến khách hàng về các sản phẩm công nghệ

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

Ngày bảo vệ: 24/01/2026

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Trình bày lại rõ mô hình thực thi trong đề án	Đã bổ sung và trình bày rõ mô hình thực thi của đề án tại mục 2.1.1 "Mô hình thực thi của đề án", mô tả 2 pha Offline (thu thập–tiền xử lý–gán nhãn–chia tập–huấn luyện–đánh giá) và Online/Batch (suy luận dự đoán cho bình luận mới và xuất kết quả). Đồng thời bổ sung mục 2.1.2 "Kiến trúc thực thi và các thành phần hệ thống" theo các tầng: dữ liệu – xử lý – mô hình – kết quả, giúp quy trình thực thi rõ ràng và dễ triển khai.
2	Thực thi và so sánh giải pháp đề xuất với một số mô hình khác	Đã mở rộng thực nghiệm và bổ sung phần so sánh bằng cách thiết kế 2 cấu hình mô hình theo ngôn ngữ/dữ liệu: PhoBERT (feature extraction) + MLP cho 2 tập tiếng Việt (Zalo, iSpeed) và Word2Vec + MLP cho 2 tập tiếng Anh (Kaggle). Luận văn đã mô tả rõ mục tiêu thực nghiệm, cách chuẩn bị dữ liệu, mô hình biểu diễn và mô hình phân loại để đảm bảo so sánh nhất quán; đồng thời bổ sung mục 3.8 "So sánh tổng hợp…" và Bảng 3.2 để tổng hợp kết quả giữa các tập dữ liệu/mô hình.

Hà Nội, ngày..09. tháng 02 năm 2016

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
TS. Nguyễn Duy Phương	TS. Đỗ Thị Liên	PGS.TS. Trần Đình Quế	Bùi Huy Dũng