**TRƯỜNG ĐẠI HỌC SÀI GÒN**A green and white border

AI-generated content may be incorrect.

**KHOA CÔNG NGHỆ THÔNG TIN**

**A blue button with white text

AI-generated content may be incorrect.**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**Nhóm 9**

Tiêu đề nghiên cứu: Phân loại cảm xúc văn bản tiếng Anh từ dữ liệu đánh giá sản phẩm bằng mô hình học sâu (LSTM)

**Giảng viên hướng dẫn: Đỗ Như Tài**

**Thành viên tham gia:**

1.            3122410219 - Hồ Hưng Lộc

2.            3121410263 - Hoàng Sỹ Khiêm

3.            3122410358 - Nguyễn Trường Sinh

Thành Phố Hồ Chí Minh, 18 tháng 5 năm 2025

# **LỜI CẢM ƠN**

Để hoàn thành bài báo cáo khoa học cuối kỳ với đề tài **“Phân loại cảm xúc văn bản tiếng Anh từ dữ liệu đánh giá sản phẩm bằng mô hình học sâu (LSTM)** chúng em đã nhận được sự quan tâm, động viên và hỗ trợ quý báu từ nhiều phía.

Trước hết, chúng em xin trân trọng gửi lời cảm ơn sâu sắc đến **Khoa Công Nghệ Thông Tin – Trường Đại Học Sài Gòn** đã tạo điều kiện thuận lợi về cơ sở vật chất và môi trường học thuật để chúng em có thể thực hiện và hoàn thành tốt nghiên cứu này.

Đặc biệt, chúng em xin bày tỏ lòng biết ơn chân thành đến **Thầy Đỗ Như Tài -** người đã trực tiếp hướng dẫn và định hướng khoa học cho chúng em trong suốt quá trình thực hiện đề tài. Những ý kiến đóng góp, sự chỉ bảo tận tình và những lời động viên kịp thời của Thầy là nguồn động lực to lớn, giúp chúng em vượt qua những khó khăn, từng bước hoàn thiện nghiên cứu một cách hiệu quả và đạt được kết quả như ngày hôm nay.

Chúng em cũng xin gửi lời cảm ơn đến các thành viên trong nhóm nghiên cứu. Sự đoàn kết, tinh thần hợp tác trách nhiệm, cùng nhau chia sẻ kiến thức và nỗ lực không ngừng của mỗi cá nhân đã góp phần quan trọng vào sự hoàn thành của dự án này.

Cuối cùng, chúng em xin kính chúc **Quý Thầy Cô trong Khoa Công Nghệ Thông Tin** và toàn thể **Quý Thầy Cô Trường Đại Học Sài Gòn** luôn dồi dào sức khỏe, hạnh phúc và gặt hái thêm nhiều thành công trong sự nghiệp giảng dạy và nghiên cứu khoa học, tiếp tục là những người truyền lửa, dìu dắt các thế hệ sinh viên trên con đường chinh phục tri thức.

Nhóm chúng em trân trọng cảm ơn!

MỤC LỤC

[**LỜI CẢM ƠN 2**](#_rkh9n9i9ehw2)

[**TÓM TẮT (ABSTRACT) 5**](#_ohjp4k2rqogl)

[**CHƯƠNG 1: TỔNG QUAN VẤN ĐỀ 6**](#_z94zewetredi)

[**1.1. Lý do chọn đề tài và Bối cảnh nghiên cứu 6**](#_i5cxbbwatz)

[**1.2. Vấn đề nghiên cứu 6**](#_98412yy4xacq)

[**1.3. Mục tiêu nghiên cứu 7**](#_wztco5tsry36)

[**1.4. Câu hỏi nghiên cứu 7**](#_g63lybh2jcre)

[**1.5. Phạm vi nghiên cứu 8**](#_rswiwoka9hdg)

[**1.6. Cấu trúc của báo cáo 8**](#_5144kirv7nzq)

[**CHƯƠNG 2: LƯỢC KHẢO TÀI LIỆU (LITERATURE REVIEW) 10**](#_ya7ijbqp49en)

[**2.1. Tổng quan về Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing - NLP) 10**](#_ypd6zejk3i62)

[**2.2. Phân loại Văn bản (Text Classification) và Phân loại Cảm xúc (Emotion Classification) 10**](#_7k4c43ym2hxz)

[2.2.1. Phân loại văn bản 10](#_i2qld5gidmc3)

[2.2.2. Phân tích Quan điểm và Phân loại Cảm xúc 10](#_nek0apxpkc3z)

[**2.3. Các Kỹ Thuật Tiền Xử Lý Văn Bản và Biểu Diễn Từ 11**](#_coekp1q2813g)

[2.3.1. Các bước tiền xử lý văn bản tiếng Anh 11](#_r63aopfrj3n9)

[2.3.2. Word Embeddings (Nhúng Từ) 11](#_1l8d6mgmjx14)

[**2.4. Mạng Nơ-ron Hồi Quy (RNN) và Long Short-Term Memory (LSTM) 12**](#_6qv1ztrqvpq8)

[2.4.1. Mạng RNN 12](#_mm1bfqhxq2ur)

[2.4.2. Mạng LSTM 12](#_9lvg9reas13g)

[**2.5. Các công trình nghiên cứu liên quan 12**](#_vytv5l1svhev)

[**CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU (METHODOLOGY) 15**](#_1i53hc7tcsk7)

[**3.1. Thiết kế nghiên cứu 15**](#_wxy4ixlxb33k)

[**3.2. Đối tượng và mẫu nghiên cứu 17**](#_5heppqwn3z2w)

[**3.3. Cách thu thập dữ liệu (Data Collection and Preparation) 18**](#_wsus17vtc0i)

[**3.4. Phân tích dữ liệu 20**](#_lekaqwo3ynit)

[3.4.1. Công cụ và Thư viện Phần mềm 21](#_7atonfitwjro)

[3.4.2. Phương pháp Tiền xử lý và Xây dựng Mô hình 22](#_xfcc1xrsrbpm)

[**CHƯƠNG 4. THỰC NGHIỆM VÀ THẢO LUẬN 26**](#_s8q8rl7uacwi)

[**4.1. Trình bày các kết quả phân tích thu được 26**](#_ib668ngakv1q)

[4.1.1. Đặc điểm bộ dữ liệu huấn luyện sau tiền xử lý: 26](#_ixvw7fbuf66)

[4.1.2. Kết quả quá trình huấn luyện mô hình LSTM: 27](#_azfgzed1pyew)

[**4.2. Đánh giá và giải thích kết quả nghiên cứu trên tập kiểm thử (Test Set) 29**](#_3ucbel0b25p)

[4.2.1. Các chỉ số hiệu năng tổng thể 29](#_6sx0asdjkdfo)

[4.2.2. Báo cáo phân loại chi tiết (Classification Report) cho 5 lớp 29](#_erog04v1votk)

[4.2.3. Ma trận nhầm lẫn (Confusion Matrix): 31](#_pce1qcdkpz5o)

[4.2.4. Phân tích một số trường hợp dự đoán sai tiêu biểu: 31](#_k4roa5kpkv3g)

[**4.3. So sánh với các nghiên cứu trước 32**](#_3yz9v81oh61t)

[**4.4. Ý nghĩa thực tiễn của nghiên cứu (dựa trên kết quả hiện tại) 33**](#_8rfkiccfqtk1)

[**4.5. Những hạn chế và đề xuất nghiên cứu tiếp theo 34**](#_u1d5u9sp3yf4)

[**CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 37**](#_ig0qif7mmt5r)

[**5.1. Tóm tắt những điểm chính đã đạt được 37**](#_t9epaj9ecna2)

[**5.2. Đưa ra câu trả lời cho câu hỏi nghiên cứu 38**](#_1ujok3jeuw5z)

[**5.3. Đề xuất giải pháp hoặc hướng nghiên cứu tiếp theo 39**](#_g1culk3s306e)

[**TÀI LIỆU THAM KHẢO 41**](#_47ukkl136a3m)

# **TÓM TẮT (ABSTRACT)**

Nghiên cứu này trình bày quá trình xây dựng và đánh giá một mô hình học sâu dựa trên kiến trúc **Long Short-Term Memory (LSTM)** cho bài toán phân loại cảm xúc đa lớp từ văn bản đánh giá sản phẩm bằng tiếng Anh. Trước sự gia tăng mạnh mẽ của dữ liệu văn bản do người dùng tạo ra trên các nền tảng trực tuyến, việc tự động hóa quá trình phân tích cảm xúc không chỉ mang lại hiểu biết sâu sắc về phản hồi của khách hàng mà còn hỗ trợ đắc lực cho các quyết định kinh doanh và cải thiện dịch vụ. Mục tiêu chính của nghiên cứu là xây dựng một mô hình có khả năng nhận diện các trạng thái cảm xúc phổ biến (love, joy, surprise, sadness, anger) từ nội dung đánh giá. Dữ liệu đầu vào, bao gồm tiêu đề và nội dung đánh giá sản phẩm, được thu thập và tiền xử lý cẩn thận; trong đó, nhãn cảm xúc được suy luận từ điểm xếp hạng (rating) của người dùng. Quy trình thực hiện bao gồm các bước tiền xử lý văn bản chuyên biệt cho tiếng Anh (tokenization, loại bỏ từ dừng, lemmatization), vector hóa văn bản sử dụng Keras Tokenizer, và huấn luyện mô hình LSTM. Hiệu năng của mô hình được đánh giá toàn diện trên một tập kiểm thử độc lập thông qua các chỉ số như độ chính xác, precision, recall, F1-score và phân tích ma trận nhầm lẫn. Kết quả thực nghiệm cho thấy mô hình LSTM đề xuất đạt được độ chính xác [ **87.5%**] trên tập kiểm thử, khẳng định tính khả thi của phương pháp trong việc tự động phân loại cảm xúc từ dữ liệu văn bản thực tế. Nghiên cứu cũng chỉ ra những hạn chế của mô hình, đặc biệt trong việc xử lý các sắc thái ngôn ngữ phức tạp, và đề xuất các hướng nghiên cứu tiếp theo nhằm nâng cao hiệu quả và mở rộng phạm vi ứng dụng của hệ thống.

# **CHƯƠNG 1: TỔNG QUAN VẤN ĐỀ**

## **1.1. Lý do chọn đề tài và Bối cảnh nghiên cứu**

Trong kỷ nguyên số hiện nay, sự phát triển mạnh mẽ của Internet và các nền tảng trực tuyến như mạng xã hội, diễn đàn, và đặc biệt là các trang thương mại điện tử đã dẫn đến sự bùng nổ của khối lượng dữ liệu văn bản do người dùng tạo ra (User-Generated Content - UGC). Những dữ liệu này, đặc biệt là các đánh giá (reviews), bình luận (comments) về sản phẩm và dịch vụ, chứa đựng những thông tin vô cùng quý giá về quan điểm, thái độ và đặc biệt là cảm xúc của người tiêu dùng. Việc khai thác và thấu hiểu những cảm xúc này không chỉ giúp doanh nghiệp cải thiện chất lượng sản phẩm, tối ưu hóa dịch vụ khách hàng mà còn đóng vai trò then chốt trong việc hoạch định chiến lược kinh doanh, marketing và xây dựng mối quan hệ bền vững với khách hàng.

Tuy nhiên, việc phân tích thủ công một lượng lớn dữ liệu văn bản để nắm bắt cảm xúc là một công việc tốn kém về thời gian, nguồn lực và khó có thể đảm bảo tính khách quan cũng như quy mô. Điều này đặt ra một nhu cầu cấp thiết về việc phát triển các hệ thống tự động có khả năng phân tích và phân loại cảm xúc từ văn bản một cách hiệu quả và chính xác. Lĩnh vực Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing - NLP) và Học máy (Machine Learning), đặc biệt là các kỹ thuật Học sâu (Deep Learning) như mạng Long Short-Term Memory (LSTM), đã cho thấy những tiềm năng vượt trội trong việc giải quyết các bài toán phức tạp liên quan đến dữ liệu văn bản. Do đó, việc nghiên cứu và ứng dụng các mô hình này vào bài toán phân loại cảm xúc từ các đánh giá sản phẩm không chỉ mang ý nghĩa khoa học mà còn có giá trị thực tiễn cao.

## **1.2. Vấn đề nghiên cứu**

Nghiên cứu này tập trung giải quyết vấn đề làm thế nào để xây dựng một hệ thống tự động có khả năng nhận diện và phân loại các trạng thái cảm xúc khác nhau (ví dụ: tích cực mạnh, tích cực, trung tính/ngạc nhiên, tiêu cực, tiêu cực mạnh) được thể hiện trong các đoạn văn bản đánh giá sản phẩm bằng tiếng Anh. Cụ thể, nghiên cứu sẽ khám phá tính hiệu quả của việc áp dụng kiến trúc mạng Long Short-Term Memory (LSTM) cho bài toán này, đồng thời xem xét tính khả thi của việc sử dụng điểm xếp hạng (rating) của sản phẩm như một nguồn thông tin để suy luận và gán nhãn cảm xúc cho dữ liệu huấn luyện. Những thách thức liên quan đến việc xử lý ngôn ngữ tự nhiên trong các đánh giá thực tế, chẳng hạn như sự đa dạng trong cách diễn đạt, ngữ cảnh phức tạp, cũng sẽ được xem xét.

## **1.3. Mục tiêu nghiên cứu**

* **Mục tiêu chính:** Xây dựng và đánh giá hiệu năng của một mô hình học sâu dựa trên kiến trúc mạng LSTM cho bài toán phân loại cảm xúc đa lớp từ dữ liệu văn bản là các đánh giá sản phẩm với ngôn ngữ tiếng Anh.
* **Mục tiêu cụ thể:**
  1. Nghiên cứu và áp dụng các kỹ thuật tiền xử lý văn bản phù hợp cho dữ liệu đánh giá sản phẩm tiếng Anh nhằm chuẩn hóa và tối ưu hóa đầu vào cho mô hình.
  2. Xây dựng và triển khai quy trình chuyển đổi dữ liệu đánh giá sản phẩm thô (bao gồm điểm xếp hạng, tiêu đề, nội dung đánh giá) thành một tập dữ liệu có cấu trúc, được gán nhãn cảm xúc phù hợp cho việc huấn luyện và đánh giá mô hình phân loại.
  3. Thiết kế, xây dựng và huấn luyện một mô hình mạng LSTM có khả năng học các đặc trưng ngữ nghĩa và tuần tự từ văn bản để phân loại cảm xúc.
  4. Đánh giá một cách khách quan hiệu năng của mô hình đã huấn luyện thông qua các chỉ số đo lường phổ biến trong học máy (như độ chính xác, precision, recall, F1-score) và phân tích sâu hơn thông qua ma trận nhầm lẫn.
  5. Xây dựng một ứng dụng/script Python minh họa (ví dụ: text\_emo\_detection.py) cho phép dự đoán cảm xúc của các đoạn văn bản mới dựa trên mô hình đã được huấn luyện.

## **1.4. Câu hỏi nghiên cứu**

Nghiên cứu này tìm cách trả lời các câu hỏi chính sau:

1. Kiến trúc mạng LSTM có khả năng phân loại hiệu quả các lớp cảm xúc khác nhau từ văn bản đánh giá sản phẩm tiếng Anh không, và hiệu năng đó được thể hiện như thế nào qua các chỉ số đánh giá?
2. Các bước tiền xử lý văn bản cụ thể (như loại bỏ từ dừng, lemmatization, tokenization) đóng góp như thế nào đến hiệu suất tổng thể của mô hình phân loại cảm xúc LSTM?
3. Phương pháp ánh xạ điểm xếp hạng (rating) của sản phẩm sang các nhãn cảm xúc đa lớp (ví dụ: love, joy, surprise, sadness, anger) có phải là một kỹ thuật tạo nhãn hiệu quả và đáng tin cậy cho việc huấn luyện mô hình trong bối cảnh thiếu dữ liệu gán nhãn cảm xúc chi tiết bởi con người không?
4. Mô hình LSTM gặp phải những khó khăn và hạn chế nào trong việc nhận diện các loại cảm xúc cụ thể hoặc xử lý các đặc điểm ngôn ngữ phức tạp trong văn bản đánh giá sản phẩm?

## **1.5. Phạm vi nghiên cứu**

* **Đối tượng nghiên cứu:** Các trạng thái cảm xúc (love, joy, surprise, sadness, anger) được biểu đạt qua ngôn ngữ viết trong các đánh giá sản phẩm.
* **Không gian mẫu:** Dữ liệu văn bản bằng tiếng Anh, được thu thập từ bộ dữ liệu công khai **Amazon Product Reviews 2023** (truy cập tại <https://amazon-reviews-2023.github.io>), cụ thể là các đánh giá sản phẩm ở định dạng JSON Lines đã được xử lý cho phù hợp với mục tiêu của dự án..
* **Thời gian thực hiện:** Từ 15/02/2025 đến 15/05/2025.
* **Giới hạn về phương pháp:** Nghiên cứu tập trung vào việc xây dựng và đánh giá một mô hình LSTM cụ thể. Các kiến trúc mạng neural tiên tiến khác (như Transformer, BERT) hoặc các phương pháp học máy truyền thống (như Naive Bayes, SVM) không được đưa vào so sánh chi tiết trong phạm vi của báo cáo này. Việc xử lý sâu các hiện tượng ngôn ngữ đặc biệt phức tạp như mỉa mai, phát hiện ý định ẩn, hoặc phân tích cảm xúc đa phương thức (kết hợp văn bản với hình ảnh/video) cũng nằm ngoài mục tiêu chính của nghiên cứu.

## **1.6. Cấu trúc của báo cáo**

Ngoài phần Tóm tắt, Mục lục, Danh mục hình ảnh/bảng biểu, Tài liệu tham khảo và nội dung chính của báo cáo được tổ chức thành 5 chương:

* **Chương 1: Tổng quan vấn đề:** Giới thiệu bối cảnh, lý do chọn đề tài, xác định vấn đề nghiên cứu, mục tiêu, câu hỏi nghiên cứu và phạm vi thực hiện của dự án.
* **Chương 2: Lược khảo tài liệu:** Trình bày cơ sở lý thuyết về xử lý ngôn ngữ tự nhiên, phân loại cảm xúc, kiến trúc mạng LSTM và các kỹ thuật liên quan, đồng thời điểm qua các công trình nghiên cứu đã có trong lĩnh vực.
* **Chương 3: Phương pháp nghiên cứu:** Mô tả chi tiết về thiết kế nghiên cứu, quy trình thu thập, xử lý và chuẩn bị dữ liệu, kiến trúc mô hình LSTM được đề xuất, các công cụ và môi trường thực nghiệm.
* **Chương 4: Thực nghiệm và Thảo luận:** Trình bày và phân tích các kết quả thu được từ quá trình huấn luyện và đánh giá mô hình, bao gồm các chỉ số hiệu năng và thảo luận về ý nghĩa cũng như hạn chế của các kết quả đó.
* **Chương 5: Kết luận và Hướng phát triển:** Tóm tắt những đóng góp chính của nghiên cứu, đưa ra câu trả lời cho các câu hỏi nghiên cứu, đồng thời đề xuất các giải pháp cải thiện và các hướng nghiên cứu tiềm năng trong tương lai.

# **CHƯƠNG 2: LƯỢC KHẢO TÀI LIỆU (LITERATURE REVIEW)**

Chương này trình bày tổng quan về các khái niệm lý thuyết nền tảng và các công trình nghiên cứu liên quan đến lĩnh vực phân loại cảm xúc văn bản, làm cơ sở cho việc đề xuất và triển khai phương pháp nghiên cứu của dự án. Nội dung chương sẽ tập trung vào việc xây dựng một khung lý thuyết vững chắc, đồng thời phân tích các nghiên cứu đi trước để xác định những điểm mạnh, điểm yếu và vị trí của nghiên cứu hiện tại.

## **2.1. Tổng quan về Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing - NLP)**

Xử lý Ngôn ngữ Tự nhiên (NLP) là một lĩnh vực liên ngành then chốt, kết hợp giữa khoa học máy tính, trí tuệ nhân tạo (AI) và ngôn ngữ học, với mục tiêu cốt lõi là cho phép máy tính có khả năng hiểu, diễn giải, xử lý và tạo ra ngôn ngữ của con người một cách hiệu quả và có ý nghĩa [1]. NLP không chỉ đơn thuần là việc xử lý văn bản mà còn bao gồm việc nắm bắt các sắc thái ngữ nghĩa, ngữ dụng và ngữ cảnh của ngôn ngữ. Lịch sử phát triển của NLP đã trải qua nhiều giai đoạn, từ các hệ thống dựa trên quy tắc ban đầu đến các phương pháp tiếp cận dựa trên thống kê và học máy, và hiện nay là sự thống trị của các mô hình học sâu [2]. Sự tiến bộ này đã mở ra vô số ứng dụng thực tiễn. Việc hiểu rõ các nguyên lý và kỹ thuật cơ bản của NLP là nền tảng thiết yếu cho nghiên cứu này.

## **2.2. Phân loại Văn bản (Text Classification) và Phân loại cảm xúc (Emotion Classification)**

### 2.2.1. Phân loại văn bản

Phân loại văn bản là nhiệm vụ gán một hoặc nhiều nhãn được định nghĩa trước cho một đoạn văn bản. Đây là một bài toán nền tảng trong NLP với nhiều ứng dụng thực tiễn như lọc thư rác, phân loại tin tức, và phân tích phản hồi của người dùng. Các phương pháp truyền thống thường dựa vào việc biểu diễn văn bản dưới dạng vector đặc trưng (ví dụ: Bag-of-Words, TF-IDF) và sau đó sử dụng các thuật toán học máy có giám sát. Tuy nhiên, các phương pháp này thường bỏ qua thông tin về thứ tự từ và ngữ cảnh.

### 2.2.2. Phân tích quan điểm và phân loại cảm xúc

Phân tích Quan điểm (Sentiment Analysis) thường xác định cực tính chung của văn bản (tích cực, tiêu cực, trung tính). Phân loại Cảm xúc (Emotion Classification) phức tạp hơn, nhằm nhận diện các trạng thái cảm xúc chi tiết như vui, buồn, giận, sợ, ngạc nhiên, yêu thích. Các mô hình cảm xúc học thuật, như mô hình 6 cảm xúc cơ bản của Ekman, cung cấp cơ sở cho việc định nghĩa các lớp cảm xúc. Nghiên cứu này tập trung vào 5 cảm xúc chính được suy luận từ dữ liệu đánh giá sản phẩm: **love, joy, surprise, sadness, anger.**

## **2.3. Các kỹ thuật tiền xử lý văn bản và biểu diễn từ**

Chất lượng của mô hình NLP phụ thuộc rất lớn vào giai đoạn tiền xử lý dữ liệu và cách biểu diễn văn bản.

### 2.3.1. Các bước tiền xử lý văn bản tiếng Anh

Trong dự án này, các bước tiền xử lý văn bản tiếng Anh được áp dụng bao gồm:

* **Tokenization (Tách từ):** Văn bản được tách thành các token riêng lẻ (từ, dấu câu) bằng cách sử dụng hàm word\_tokenize từ thư viện NLTK (Bird et al., 2009).
* **Lowercasing (Chuyển chữ thường):** Tất cả các token được chuyển thành chữ thường để đảm bảo tính nhất quán.
* **Punctuation and Stopword Removal (Loại bỏ dấu câu và từ dừng):** Dấu câu và các từ dừng phổ biến trong tiếng Anh (sử dụng danh sách từ nltk.corpus.stopwords.words('english')) được loại bỏ để giảm nhiễu và tập trung vào các từ mang ý nghĩa.
* **Lemmatization (Đưa từ về dạng gốc):** Các từ được đưa về dạng gốc (lemma) của chúng bằng cách sử dụng WordNetLemmatizer từ NLTK, giúp giảm sự đa dạng của từ vựng mà vẫn giữ được ý nghĩa.

### 2.3.2. Word Embeddings (Nhúng Từ)

Việc lựa chọn hoặc huấn luyện các vector từ (word embeddings) hiệu quả là một yếu tố then chốt. Nghiên cứu của **Manjavacas et al. (2019)** trong "Utility of General and Specific Word Embeddings for Classifying Translational Stages of Research" [**3**] đã phân tích và so sánh lợi ích của việc sử dụng word embeddings tổng quát so với word embeddings đặc thù cho miền dữ liệu trong các tác vụ phân loại.

**Điểm mạnh:** Các vector từ được huấn luyện trên dữ liệu cùng miền (domain-specific) có khả năng nắm bắt tốt hơn các sắc thái ngữ nghĩa và mối quan hệ giữa các từ trong lĩnh vực đó, dẫn đến hiệu năng mô hình tốt hơn.

**Điểm yếu/Hạn chế :** Đòi hỏi một lượng lớn dữ liệu trong miền để huấn luyện hiệu quả, và có thể không tổng quát hóa tốt cho các miền khác.

**Nghiên cứu hiện tại kế thừa/phát triển:** Trong dự án này, lớp Embedding của Keras được huấn luyện từ đầu cùng với toàn bộ mô hình LSTM trên bộ dữ liệu đánh giá sản phẩm. Cách tiếp cận này cho phép học các biểu diễn từ đặc thù, phù hợp với nhận định của [**3**] về lợi ích của word embeddings chuyên biệt cho miền.

## **2.4. Mạng Nơ-ron Hồi quy (RNN) và Long Short-Term Memory (LSTM)**

### 2.4.1. Mạng RNN

Mạng Nơ-ron Hồi Quy (RNN) được thiết kế để xử lý dữ liệu tuần tự như văn bản, nhờ khả năng duy trì thông tin từ các bước trước đó thông qua các kết nối hồi quy [**4**]. Tuy nhiên, chúng thường gặp khó khăn với các phụ thuộc dài hạn.

### 2.4.2. Mạng LSTM

Mạng Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) giải quyết hạn chế của RNNs truyền thống. Nghiên cứu của **Lv et al. (2022)** trong "Evolving Long Short-Term Memory Network-Based Text Classification" [**7**] đã khám phá các phương pháp tiến hóa kiến trúc LSTM cho phân loại văn bản, cho thấy sự linh hoạt và tiềm năng cải thiện hiệu năng của họ LSTM.

* **Điểm mạnh (của các phương pháp LSTM cải tiến [7]):** Có thể tự động tìm kiếm và tối ưu hóa kiến trúc mạng, giúp tăng cường khả năng học các mẫu phức tạp và đạt hiệu năng cao hơn trên các bộ dữ liệu cụ thể.
* **Điểm yếu/Hạn chế:** Các phương pháp tiến hóa kiến trúc thường đòi hỏi năng lực tính toán lớn và thời gian huấn luyện kéo dài. Việc diễn giải kiến trúc được tiến hóa cũng có thể phức tạp.
* **Nghiên cứu hiện tại kế thừa/phát triển:** Nghiên cứu này sử dụng một kiến trúc LSTM tương đối chuẩn và phổ biến làm nền tảng, tập trung vào việc đánh giá hiệu quả của nó trên một bộ dữ liệu đánh giá sản phẩm với quy trình tạo nhãn đặc thù, thay vì đề xuất một kiến trúc LSTM mới dựa trên thuật toán tiến hóa.

## **2.5. Các công trình nghiên cứu liên quan**

Lĩnh vực phân loại cảm xúc từ văn bản đã có nhiều đóng góp quan trọng từ cộng đồng nghiên cứu, sử dụng đa dạng các phương pháp và bộ dữ liệu.

* Công trình của **Alqarafi et al. (2024)** trong "Deep Learning Neural Networks for Emotion Classification from Text: Enhanced Leaky Rectified Linear Unit Activation and Weighted Categorical Cross Entropy Loss Function" [**8**] đã đề xuất các cải tiến cho mạng neural học sâu trong phân loại cảm xúc. Cụ thể, nghiên cứu này tập trung vào việc sử dụng hàm kích hoạt Leaky ReLU cải tiến và hàm mất mát Categorical Cross Entropy có trọng số.
  + **Điểm mạnh:** Các cải tiến về hàm kích hoạt và hàm mất mát có thể giúp mô hình học hiệu quả hơn các lớp cảm xúc không cân bằng, vốn là một vấn đề thường gặp trong các bộ dữ liệu thực tế, đồng thời cải thiện khả năng hội tụ và độ chính xác tổng thể.
  + **Điểm yếu/Hạn chế:** Hiệu quả của các hàm kích hoạt hoặc hàm mất mát cụ thể có thể phụ thuộc vào đặc điểm của bộ dữ liệu và kiến trúc mạng. Việc lựa chọn trọng số phù hợp cho hàm mất mát cũng đòi hỏi sự tinh chỉnh.
  + **Nghiên cứu hiện tại kế thừa/phát triển:** Nghiên cứu của chúng em sử dụng các thành phần Keras tiêu chuẩn (ví dụ: hàm kích hoạt softmax cho lớp output và hàm mất mát sparse\_categorical\_crossentropy mặc định không có trọng số). Việc áp dụng các kỹ thuật như trong [**8**] có thể là một hướng cải tiến tiềm năng trong tương lai để xử lý sự mất cân bằng dữ liệu giữa các lớp cảm xúc.
* Nhiều nghiên cứu được công bố tại các hội thảo và tạp chí uy tín, ví dụ như các công trình trong kho lưu trữ của **ACL Anthology** (ví dụ: bài báo được tham chiếu bởi aclanthology.org/C18-1081.pdf [**4**]), thường khám phá các kiến trúc mô hình mới, bộ dữ liệu mới hoặc các kỹ thuật tiền xử lý tiên tiến cho các tác vụ NLP, bao gồm cả phân loại cảm xúc.
  + **Điểm mạnh (của các nghiên cứu được công bố trên ACL Anthology):** Thường có tính mới và đóng góp học thuật cao, được đánh giá bởi cộng đồng khoa học.
  + **Điểm yếu/Hạn chế:** Một số phương pháp có thể phức tạp để triển khai hoặc đòi hỏi tài nguyên tính toán lớn.
  + **Nghiên cứu hiện tại kế thừa/phát triển:** Nghiên cứu này áp dụng một kiến trúc LSTM tương đối phổ biến nhưng tập trung vào một quy trình tạo nhãn và một bộ dữ liệu thực tế từ review sản phẩm, nhằm đánh giá tính ứng dụng của phương pháp.

**Vị trí và đóng góp của nghiên cứu hiện tại:** Các nghiên cứu đã xem xét thường tập trung vào các bộ dữ liệu cảm xúc được gán nhãn thủ công hoặc các bộ dữ liệu mạng xã hội với đặc thù ngôn ngữ riêng. Điểm khác biệt và đóng góp chính của nghiên cứu này nằm ở việc:

1. **Đề xuất và kiểm chứng** một quy trình thực tiễn để tạo nhãn cảm xúc đa lớp **(love, joy, surprise, sadness, anger)** từ dữ liệu điểm xếp hạng **(rating)** của các đánh giá sản phẩm tiếng Anh. Phương pháp này giải quyết một phần bài toán thiếu hụt dữ liệu cảm xúc đa lớp được gán nhãn chi tiết cho lĩnh vực review sản phẩm, cho phép tận dụng nguồn dữ liệu dồi dào có sẵn.
2. **Triển khai và đánh giá** hiệu năng của một mô hình LSTM trên bộ dữ liệu được tạo ra theo cách này**.** Nghiên cứu cung cấp phân tích cụ thể về khả năng học và phân loại các cảm xúc suy luận từ rating, làm rõ những cảm xúc nào mô hình có thể nhận diện tốt và những cảm xúc nào còn gặp khó khăn trong bối cảnh dữ liệu tiếng Anh.
3. **Cung cấp một quy trình làm việc** hoàn chỉnh từ việc xử lý dữ liệu JSON Lines đầu vào, làm sạch văn bản, đến xây dựng, huấn luyện và kiểm thử mô hình bằng script Python thực tế**.** Điều này mang lại một ví dụ tham khảo cho các dự án ứng dụng NLP có quy mô tương tự, tập trung vào tính thực tiễn và khả năng tái tạo.

Nghiên cứu này kế thừa việc sử dụng LSTM như một công cụ hiệu quả cho phân loại văn bản và cảm xúc, nhưng áp dụng và đánh giá nó trong một kịch bản tạo nhãn và nguồn dữ liệu cụ thể, với hy vọng mang lại những hiểu biết thực tế về khả năng và giới hạn của mô hình.

# **CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU (METHODOLOGY)**

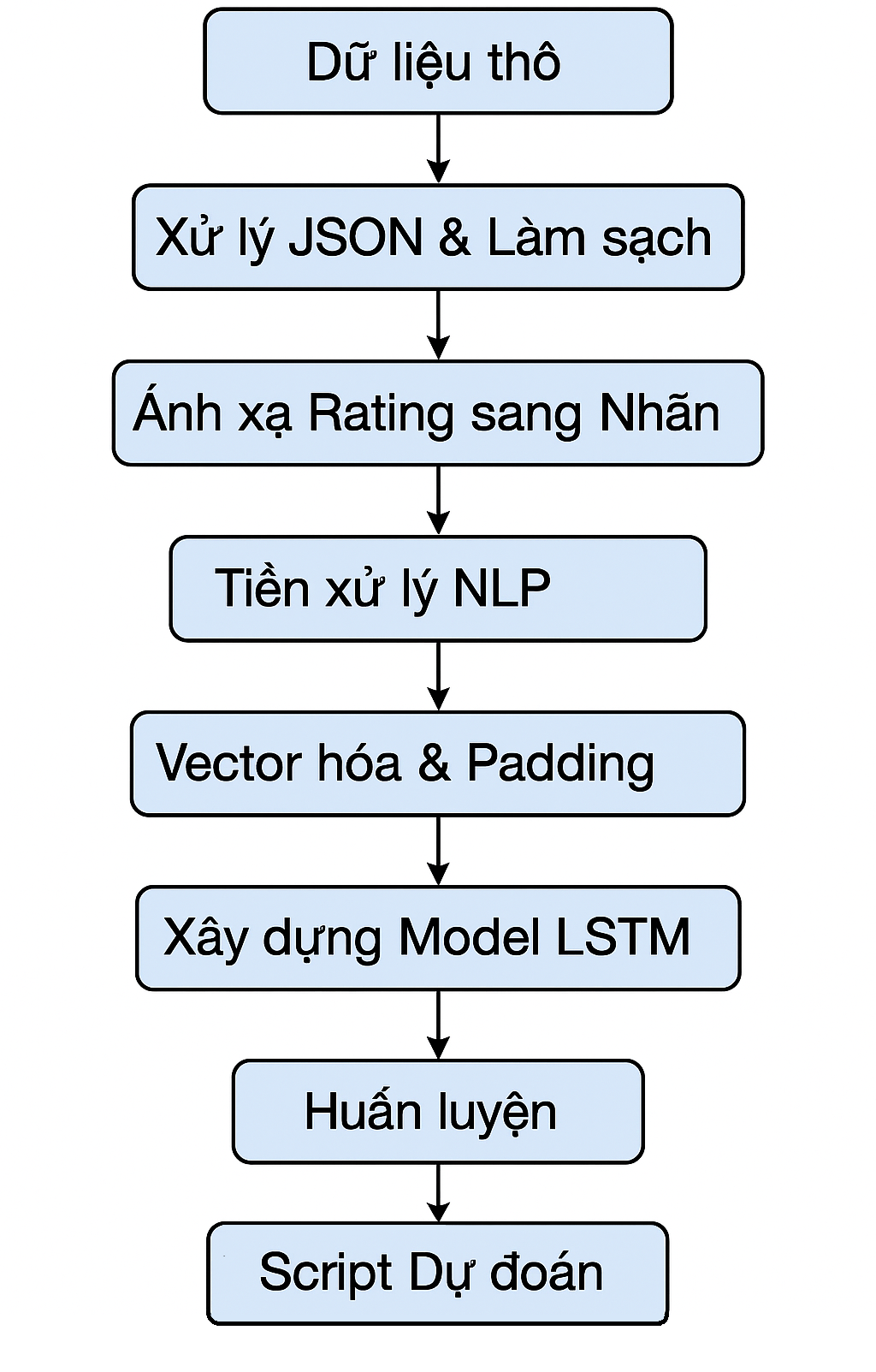
Chương này trình bày chi tiết về phương pháp luận được áp dụng trong nghiên cứu, bao gồm thiết kế nghiên cứu, mô tả đối tượng và mẫu dữ liệu, quy trình thu thập và chuẩn bị dữ liệu, kiến trúc mô hình học sâu được đề xuất, cùng các công cụ và phần mềm được sử dụng để triển khai và đánh giá.

## **3.1. Thiết kế nghiên cứu**

Nghiên cứu này được thực hiện theo phương pháp **nghiên cứu định lượng (quantitative research)**, tập trung vào việc xây dựng và đánh giá thực nghiệm một mô hình học máy có giám sát cho bài toán phân loại cảm xúc đa lớp từ dữ liệu văn bản. Thiết kế nghiên cứu bao gồm các giai đoạn chính sau:

1. **Thu thập và chuẩn bị dữ liệu:** Lựa chọn và xử lý nguồn dữ liệu đánh giá sản phẩm tiếng Anh, bao gồm việc làm sạch văn bản thô và tạo nhãn cảm xúc từ điểm xếp hạng (rating).
2. **Tiền xử lý văn bản:** Áp dụng các kỹ thuật tiền xử lý ngôn ngữ tự nhiên tiêu chuẩn để chuẩn hóa và tối ưu hóa dữ liệu đầu vào cho mô hình.
3. **Xây dựng mô hình:** Thiết kế và triển khai một mô hình mạng Long Short-Term Memory (**LSTM**) có khả năng học các đặc trưng tuần tự từ văn bản.
4. **Huấn luyện mô hình:** Huấn luyện mô hình LSTM trên tập dữ liệu đã được chuẩn bị, sử dụng một phần dữ liệu để kiểm định (validation) trong quá trình huấn luyện nhằm tránh overfitting và tối ưu hóa các tham số.
5. **Đánh giá hiệu năng mô hình:** Đánh giá mô hình đã huấn luyện trên một tập dữ liệu kiểm thử (test set) độc lập bằng các chỉ số đo lường khách quan.
6. **Phân tích kết quả và triển khai ứng dụng minh họa:** Phân tích các kết quả thu được, thảo luận về những ưu nhược điểm của mô hình và xây dựng một script Python đơn giản để minh họa khả năng dự đoán.

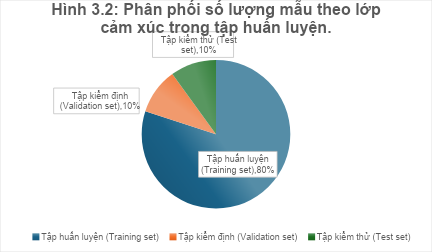
Sơ đồ khối tổng quan quy trình nghiên cứu có thể được trình bày như Hình 3.1

**Generated image**

*Hình 3.1: Sơ đồ khối tổng quan quy trình thực hiện dự án.*

## **3.2. Đối tượng và mẫu nghiên cứu**

* **Đối tượng nghiên cứu:** Các trạng thái cảm xúc được biểu đạt thông qua ngôn ngữ viết trong các đánh giá (**reviews**) sản phẩm trực tuyến. Cụ thể, nghiên cứu tập trung vào việc phân loại các cảm xúc sau: **love** (yêu thích/rất hài lòng), **joy** (vui vẻ/hài lòng), **surprise** (ngạc nhiên/trung tính), **sadness** (buồn bã/thất vọng), và **anger** (tức giận/rất không hài lòng).
* **Mẫu nghiên cứu (Dữ liệu):**
  + Nguồn dữ liệu thô được sử dụng trong nghiên cứu này là bộ dữ liệu **Amazon Product Reviews 2023**, một tập hợp lớn các đánh giá sản phẩm được công bố công khai cho mục đích nghiên cứu **[9]**. Dữ liệu được cung cấp dưới định dạng JSON Lines, mỗi bản ghi chứa thông tin chi tiết về đánh giá.
  + Sau quá trình xử lý và chuẩn bị (chi tiết ở **mục 3.3**), một bộ dữ liệu mới đã được tạo ra với định dạng văn\_bản\_đánh\_giá;nhãn\_cảm\_xúc.
  + Bộ dữ liệu này sau đó được phân chia thành ba tập riêng biệt:
    - **Tập huấn luyện (Training set):** 561136 mẫu, chiếm **~80%** tổng dữ liệu] - Dùng để huấn luyện các tham số của mô hình LSTM.
    - **Tập kiểm định (Validation set):** 70142 mẫu, chiếm **~10%** tổng dữ liệu - Dùng để theo dõi hiệu năng của mô hình trong quá trình huấn luyện và tinh chỉnh các siêu tham số (hyperparameters).
    - **Tập kiểm thử (Test set):** 70143 mẫu, chiếm **~10%** tổng dữ liệu - Dùng để đánh giá cuối cùng hiệu năng của mô hình trên dữ liệu hoàn toàn mới, chưa từng được thấy trong quá trình huấn luyện.



* + Việc phân chia dữ liệu được thực hiện một cách ngẫu nhiên để đảm bảo tính khách quan và đại diện của các tập dữ liệu.

## **3.3. Cách thu thập dữ liệu (Data Collection and Preparation)**

Quy trình thu thập và chuẩn bị dữ liệu từ các file JSON Lines của bộ **Amazon Product Reviews 2023** [9] để tạo ra các tập dữ liệu huấn luyện, kiểm định và kiểm thử được thực hiện bằng script Python và bao gồm các bước chính sau:

1. **Đọc và Trích xuất Dữ liệu JSON Lines:**
   * Sử dụng thư viện json trong Python, mỗi dòng trong file .jsonl gốc được đọc và phân tích cú pháp thành một đối tượng JSON.
   * Từ mỗi đối tượng JSON, các trường thông tin quan trọng bao gồm rating (điểm xếp hạng), title (tiêu đề đánh giá), và text (nội dung đánh giá) được trích xuất. Các trường siêu dữ liệu khác không liên quan trực tiếp đến nội dung cảm xúc văn bản (ví dụ: asin, user\_id, timestamp, images, helpful\_vote, verified\_purchase) được loại bỏ ở giai đoạn này.
2. **Kết hợp Tiêu đề và Nội dung Đánh giá:**
   * Nội dung từ trường title và text được kết hợp lại thành một trường văn bản duy nhất cho mỗi đánh giá. Nếu cả hai trường đều có nội dung, chúng được nối với nhau bằng một dấu chấm và khoảng trắng (.) để tạo sự liền mạch về ngữ nghĩa.
3. **Làm sạch Văn bản Thô:**
   * Trước khi đưa vào mô hình, văn bản được làm sạch sơ bộ để loại bỏ các yếu tố gây nhiễu:
     + Loại bỏ các mã ID video đặc biệt (ví dụ: [[VIDEOID:...]]) bằng cách sử dụng biểu thức chính quy (re.sub(r'\[\[VIDEOID:[a-zA-Z0-9\_.-]+\]\]', '', text)).
     + Loại bỏ các thẻ HTML (ví dụ: <br /> và các thẻ khác) bằng biểu thức chính quy (ví dụ: re.sub(r'<[^>]+>', '', text)).
     + Chuẩn hóa các ký tự xuống dòng (\n, \r) thành khoảng trắng.
     + Thay thế các dấu chấm phẩy (;) trong nội dung văn bản bằng dấu phẩy (,) để tránh xung đột với ký tự phân tách được sử dụng trong file output (văn\_bản;nhãn).
     + Loại bỏ các khoảng trắng thừa ở đầu, cuối và các khoảng trắng kép trong văn bản.

A screenshot of a computer code

AI-generated content may be incorrect.

1. **Ánh xạ Điểm Xếp hạng (Rating) sang Nhãn Cảm xúc:**

Do bộ dữ liệu gốc không có nhãn cảm xúc chi tiết, một quy tắc ánh xạ dựa trên điểm xếp hạng (rating, từ 1.0 đến 5.0) đã được xây dựng để suy luận ra 5 lớp cảm xúc mục tiêu. Hàm map\_rating\_to\_emotion(rating\_float) được định nghĩa như sau:

* + Nếu rating **> 4.5** (thường là 5 sao): Nhãn love.
  + Nếu rating > **3.5** (thường là 4 sao): Nhãn joy.
  + Nếu rating > **2.5** (thường là 3 sao): Nhãn surprise (được chọn để đại diện cho cảm xúc trung tính hoặc bất ngờ, do bộ nhãn không có "neutral").
  + Nếu rating > **1.5** (thường là 2 sao): Nhãn sadness.
  + Nếu rating <= **1.5** (thường là 1 sao): Nhãn anger.

Các đánh giá có điểm xếp hạng không thể ánh xạ rõ ràng (ví dụ: rating không hợp lệ hoặc nằm ngoài khoảng 0-5) sẽ bị loại bỏ.

A screenshot of a computer code

AI-generated content may be incorrect.

1. **Tạo File Dữ liệu Chuẩn hóa và Phân chia Tập:**
   * Dữ liệu đã xử lý (bao gồm văn bản kết hợp đã làm sạch và nhãn cảm xúc suy ra) được lưu vào một file .txt trung gian (**data\_review\_da\_xu\_ly.txt**) với định dạng văn\_bản\_đã\_làm\_sạch;nhãn\_cảm\_xúc trên mỗi dòng. Chỉ những mẫu có nội dung văn bản không rỗng sau khi làm sạch mới được giữ lại.
   * Từ file trung gian này, dữ liệu được xáo trộn ngẫu nhiên và sau đó được chia thành ba tập: tập huấn luyện (**reviews\_train.txt** - 80%), tập kiểm định (**reviews\_val.txt** - 10%), và tập kiểm thử (**reviews\_test.txt** - 10%) bằng một script Python riêng biệt.

## **3.4. Phân tích dữ liệu**

Sau khi dữ liệu đã được thu thập, làm sạch sơ bộ, ánh xạ nhãn cảm xúc từ điểm xếp hạng, và phân chia thành các tập huấn luyện, kiểm định và kiểm thử (như mô tả ở mục 3.3), bước tiếp theo là áp dụng các kỹ thuật phân tích và xây dựng mô hình học sâu để phân loại cảm xúc. Quá trình này được thực hiện chủ yếu trong môi trường Jupyter Notebook và sử dụng các thư viện Python chuyên dụng.

### 3.4.1. Công cụ và Thư viện Phần mềm

Quá trình phân tích dữ liệu và xây dựng mô hình được thực hiện bằng ngôn ngữ lập trình **Python (phiên bản 3.12)**. Môi trường phát triển chính bao gồm **Jupyter Notebook** (cho các bước khám phá, xây dựng, huấn luyện và đánh giá ban đầu mô hình) và **Visual Studio Code** (để phát triển các script Python hỗ trợ xử lý dữ liệu và script ứng dụng dự đoán cuối cùng). Các thư viện mã nguồn mở then chốt được sử dụng bao gồm:

* **json**: Để đọc và xử lý dữ liệu đánh giá sản phẩm gốc ở định dạng JSON Lines.
* **re (Regular Expressions)**: Thực hiện các thao tác làm sạch văn bản phức tạp dựa trên mẫu, ví dụ như loại bỏ các mã ID video và thẻ HTML.
* **random**: Sử dụng để xáo trộn dữ liệu một cách ngẫu nhiên trước khi phân chia thành các tập huấn luyện, kiểm định và kiểm thử.
* **Pandas**: (Được sử dụng trong Notebook) Để tải các file .txt chứa dữ liệu đã xử lý vào cấu trúc DataFrame, giúp cho việc thao tác, khám phá và chuẩn bị dữ liệu cho mô hình trở nên thuận tiện hơn.
* **NLTK (Natural Language Toolkit) (Bird et al., 2009) [11]**: Một bộ thư viện toàn diện cho xử lý ngôn ngữ tự nhiên bằng tiếng Anh, được sử dụng cho các tác vụ tiền xử lý chính bao gồm:
  + word\_tokenize: Tách câu thành các từ (token).
  + stopwords.words('english'): Cung cấp danh sách các từ dừng phổ biến trong tiếng Anh để loại bỏ.
  + WordNetLemmatizer: Đưa các từ về dạng gốc (lemma) của chúng.
* **Scikit-learn (Pedregosa et al., 2011) [13]**: Một thư viện học máy mạnh mẽ, được sử dụng cho:
  + LabelEncoder: Chuyển đổi các nhãn cảm xúc dạng text (ví dụ: 'joy', 'anger') sang dạng số nguyên mà mô hình có thể hiểu được.
  + classification\_report, confusion\_matrix: Tính toán và hiển thị các chỉ số đo lường chi tiết để đánh giá hiệu năng của mô hình phân loại.
* **TensorFlow với Keras API [12]**: Nền tảng học sâu hàng đầu, được sử dụng để:
  + Tokenizer: Xây dựng từ điển từ vựng từ tập dữ liệu huấn luyện và chuyển đổi các đoạn văn bản thành chuỗi các chỉ số số nguyên.
  + pad\_sequences: Đệm (padding) các chuỗi số để chúng có cùng độ dài, tạo thành đầu vào chuẩn cho mô hình LSTM.
  + Sequential, Embedding, LSTM, Dense: Các lớp (layers) cơ bản để xây dựng kiến trúc mạng nơ-ron LSTM tuần tự.
  + Adam optimizer, sparse\_categorical\_crossentropy loss function: Các thành phần để cấu hình quá trình biên dịch và tối ưu hóa mô hình trong quá trình huấn luyện.
* **Matplotlib & Seaborn**: Các thư viện trực quan hóa dữ liệu, được sử dụng để vẽ các biểu đồ như phân phối nhãn cảm xúc, đường cong học tập (accuracy/loss qua các epoch) và ma trận nhầm lẫn.
* **pickle**: Dùng để lưu trữ và tải đối tượng Tokenizer đã được huấn luyện, giúp cho việc tái sử dụng trong script dự đoán mà không cần phải huấn luyện lại tokenizer từ đầu.

### 3.4.2. Phương pháp Tiền xử lý và Xây dựng mô hình

Quy trình chi tiết từ việc tải dữ liệu đã được chia tập (reviews\_train.txt, reviews\_val.txt, reviews\_test.txt) đến khi huấn luyện và lưu trữ mô hình được thực hiện như sau:

1. **Tải dữ liệu đã chia tập:** Các file văn bản chứa đánh giá và nhãn cảm xúc (định dạng văn\_bản;nhãn) được đọc vào các danh sách hoặc Pandas DataFrames tương ứng cho tập huấn luyện, kiểm định và kiểm thử.
2. **Tiền xử lý văn bản chuyên sâu:** Áp dụng các kỹ thuật đã mô tả ở mục 2.3.1 và 3.3.1 (chuyển chữ thường, loại bỏ dấu câu, tách từ, loại bỏ từ dừng, lemmatization) cho tất cả các mẫu văn bản trong các tập dữ liệu.
   * Chuyển toàn bộ văn bản thành chữ thường (text.lower()).
   * Loại bỏ các dấu câu và ký tự đặc biệt không cần thiết (sử dụng string.punctuation).
   * Tách từ (Tokenization) bằng nltk.word\_tokenize.
   * Loại bỏ các từ dừng (Stopwords) phổ biến trong tiếng Anh bằng danh sách từ nltk.corpus.stopwords.words('english').
   * Đưa các từ về dạng gốc (Lemmatization) bằng nltk.stem.WordNetLemmatizer để giảm sự đa dạng của từ vựng. ví dụ: "Câu gốc: 'This product is absolutely AMAZING! I love it so much, works perfectly.' Câu sau tiền xử lý: 'product absolutely amazing love much work perfectly'".
3. **Mã hóa nhãn (Label Encoding):** Các nhãn cảm xúc dạng text (ví dụ: 'joy', 'anger', 'love', 'sadness', 'surprise') được chuyển đổi thành dạng số nguyên (0, 1, 2, 3, 4) bằng cách sử dụng **LabelEncoder** từ thư viện **Scikit-learn**. Đối tượng LabelEncoder được fit\_transform trên nhãn của tập huấn luyện và sau đó được sử dụng để transform nhãn của tập kiểm định và tập kiểm thử, đảm bảo tính nhất quán trong việc mã hóa. Thứ tự các lớp cảm xúc được mã hóa (lưu trong thuộc tính encoder.classes\_) sẽ được sử dụng để diễn giải kết quả đầu ra của mô hình.
4. **Tokenization và Vector hóa:**
   * Khởi tạo một đối tượng tensorflow.keras.preprocessing.text.Tokenizer. Tham số num\_words có thể được sử dụng để giới hạn kích thước từ điển (ví dụ: chỉ giữ lại **20000 từ** phổ biến nhất), và oov\_token="<unk>" được dùng để xử lý các từ không có trong từ điển.
   * Tokenizer được huấn luyện (fit) trên toàn bộ các đoạn văn bản đã tiền xử lý của tập huấn luyện (tokenizer.fit\_on\_texts(train\_texts\_preprocessed)). Quá trình này xây dựng một từ điển (tokenizer.word\_index) ánh xạ mỗi từ duy nhất đã học được sang một chỉ số số nguyên. Kích thước thực tế của từ điển (len(tokenizer.word\_index) + 1) được xác định để sử dụng cho lớp Embedding.
   * Các câu văn bản đã tiền xử lý trong cả ba tập dữ liệu (huấn luyện, kiểm định, kiểm thử) sau đó được chuyển đổi thành các chuỗi các chỉ số số nguyên tương ứng bằng phương thức tokenizer.texts\_to\_sequences().
   * Độ dài tối đa của một chuỗi (maxlen) được xác định dựa trên độ dài của các chuỗi trong tập huấn luyện mặc định là 40 từ.
   * Tất cả các chuỗi số nguyên sau đó được đệm (padding) bằng tensorflow.keras.preprocessing.sequence.pad\_sequences() để đảm bảo chúng có cùng độ dài maxlen. Tham số padding='post' và truncating='post' được sử dụng để thêm số 0 vào cuối chuỗi nếu ngắn hơn maxlen, hoặc cắt bỏ phần cuối nếu dài hơn.
5. **Xây dựng Kiến trúc Mô hình LSTM:**
   * Sử dụng keras.models.Sequential để xây dựng mô hình.
   * **Lớp Embedding:** keras.layers.Embedding(input\_dim=len(tokenizer.word\_index) + 1, output\_dim=100, input\_length=maxlen). Lớp này tạo ra các vector nhúng cho mỗi từ, với kích thước từ điển là len(tokenizer.word\_index) + 1 (cộng 1 cho từ không có trong từ điển/padding), chiều của vector nhúng là 100.
   * **Lớp LSTM:** keras.layers.LSTM(64, return\_sequences=False). Một lớp LSTM với 64 units. return\_sequences=False vì đây là lớp LSTM cuối cùng trước lớp Dense output, chỉ cần output cuối cùng của chuỗi.
   * **Lớp Dense Output:** keras.layers.Dense(len(encoder.classes\_), activation='softmax'). Lớp output với số units bằng số lượng lớp cảm xúc đã mã hóa, sử dụng hàm kích hoạt softmax để đưa ra phân phối xác suất cho các lớp.
6. **Biên dịch Mô hình (model.compile):**
   * **Optimizer** **(Thuật toán tối ưu):** tf.keras.optimizers.Adam được sử dụng với tốc độ học (learning rate) mặc định là 0.001.
   * **Loss Function (Hàm mất mát):** sparse\_categorical\_crossentropy được chọn vì các nhãn cảm xúc đã được mã hóa thành dạng số nguyên (0, 1, 2,...).
   * **Metrics** **(Chỉ số đánh giá):** ['accuracy'] được sử dụng để theo dõi độ chính xác của mô hình trong quá trình huấn luyện và đánh giá.
7. **Huấn luyện Mô hình (model.fit):**

Mô hình được huấn luyện bằng cách gọi phương thức model.fit() với các tham số sau:

* + Huấn luyện mô hình trên dữ liệu huấn luyện đã xử lý (train\_pad, train\_labels\_p).
  + Sử dụng dữ liệu kiểm định (val\_pad, val\_labels\_p) để theo dõi hiệu năng trong quá trình huấn luyện.
  + Thiết lập số lượng **epochs = 8** và **batch\_size = 32.**

1. **Lưu trữ Mô hình và Tokenizer:**
   * Sau khi huấn luyện, mô hình (model.save('emotion\_recognizer.h5')) và đối tượng Tokenizer (pickle.dump(tokenizer, open('tokenizer.pkl', 'wb'))) được lưu lại để có thể tái sử dụng trong script dự đoán mà không cần huấn luyện lại.

**Tổng kết:**

Chương này đã trình bày chi tiết phương pháp luận được sử dụng trong nghiên cứu, từ thiết kế nghiên cứu, quy trình thu thập và chuẩn bị dữ liệu đặc thù từ bộ dữ liệu **Amazon Product Reviews 2023** [9], bao gồm cả việc ánh xạ điểm xếp hạng sang nhãn cảm xúc, đến việc xây dựng và cấu hình mô hình mạng Long Short-Term Memory (LSTM). Các công cụ, thư viện phần mềm và các bước tiền xử lý, vector hóa văn bản, cũng như kiến trúc mô hình và quy trình huấn luyện đã được mô tả cụ thể. Những phương pháp này tạo nền tảng vững chắc cho việc thực hiện các thực nghiệm và phân tích kết quả sẽ được trình bày trong Chương 4.

# **CHƯƠNG 4. THỰC NGHIỆM VÀ THẢO LUẬN**

Chương này trình bày chi tiết các kết quả thực nghiệm thu được từ quá trình huấn luyện và đánh giá mô hình phân loại cảm xúc LSTM đã được xây dựng và mô tả trong Chương 3. Các kết quả sẽ được minh họa bằng các biểu đồ, bảng biểu cụ thể. Đồng thời, chương sẽ tiến hành phân tích, thảo luận sâu về hiệu năng của mô hình, những ưu điểm, hạn chế quan sát được và đối chiếu với các mục tiêu nghiên cứu đã đề ra, cũng như nêu bật ý nghĩa thực tiễn của nghiên cứu.

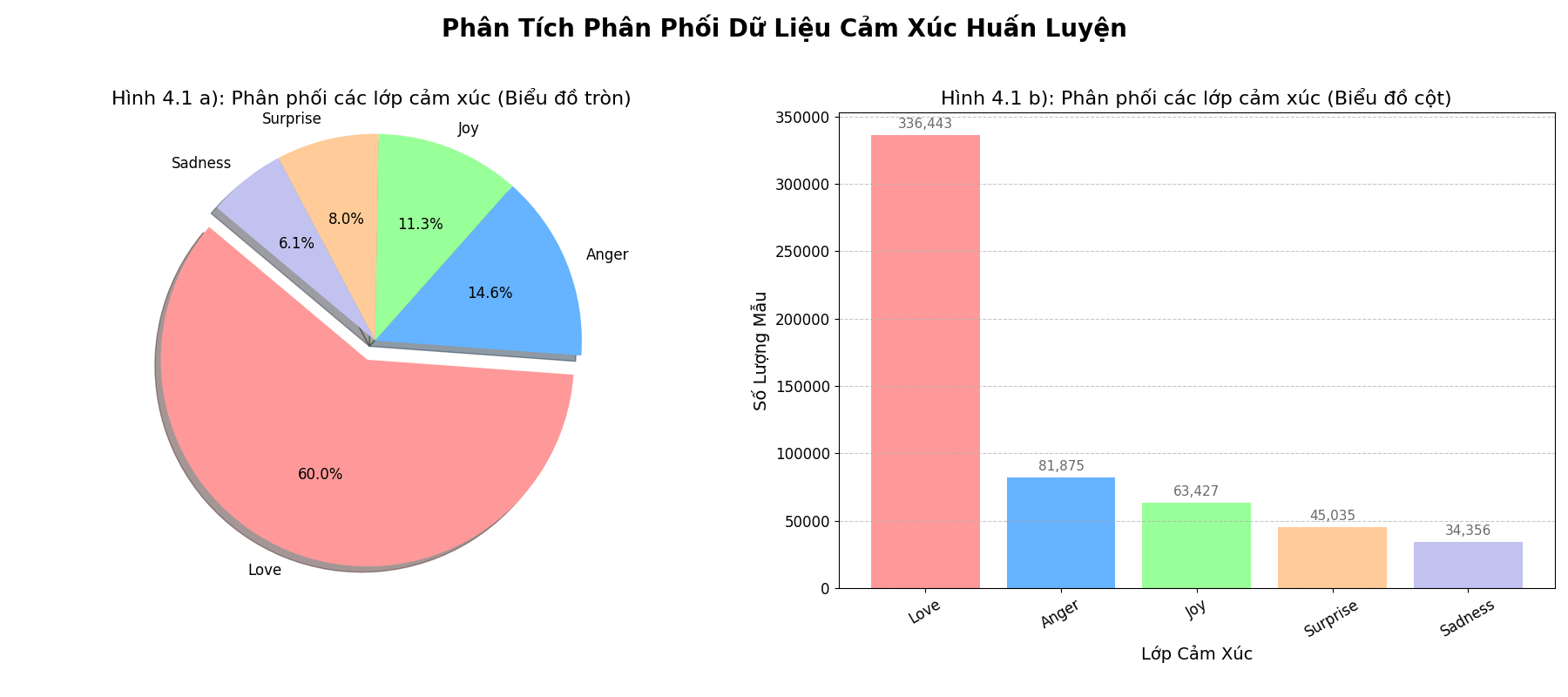
## **4.1. Trình bày các kết quả phân tích thu được**

### 4.1.1. Đặc điểm bộ dữ liệu huấn luyện sau tiền xử lý:

Sau quá trình thu thập, làm sạch và ánh xạ nhãn từ điểm xếp hạng (rating) của các đánh giá sản phẩm từ bộ dữ liệu **Amazon Product Reviews 2023** [9], bộ dữ liệu huấn luyện được sử dụng cho mô hình bao gồm 5 lớp cảm xúc chính: **Anger, Joy, Love, Sadness, và Surprise**. Phân phối số lượng mẫu cho từng lớp cảm xúc trong tập huấn luyện được thể hiện trong ***Bảng 4.1*** và trực quan hóa qua ***Hình 4.1***.

***Bảng 4.1: Phân phối số lượng mẫu theo lớp cảm xúc trong tập huấn luyện***

| **Lớp Cảm Xúc** | **Số Lượng Mẫu** | **Tỷ Lệ (%)** |
| --- | --- | --- |
| Love  Anger  Joy  Surprise  Sadness | 336,443  81,875  63,427  45,035  34,356 | 59,9  14,6  11,3  8,0  6,1 |
| **Tổng cộng** | **561,136** | **100** |

****

***Hình 4.1: Phân phối dữ liệu cảm xúc huấn luyện***

*Nhận xét:*

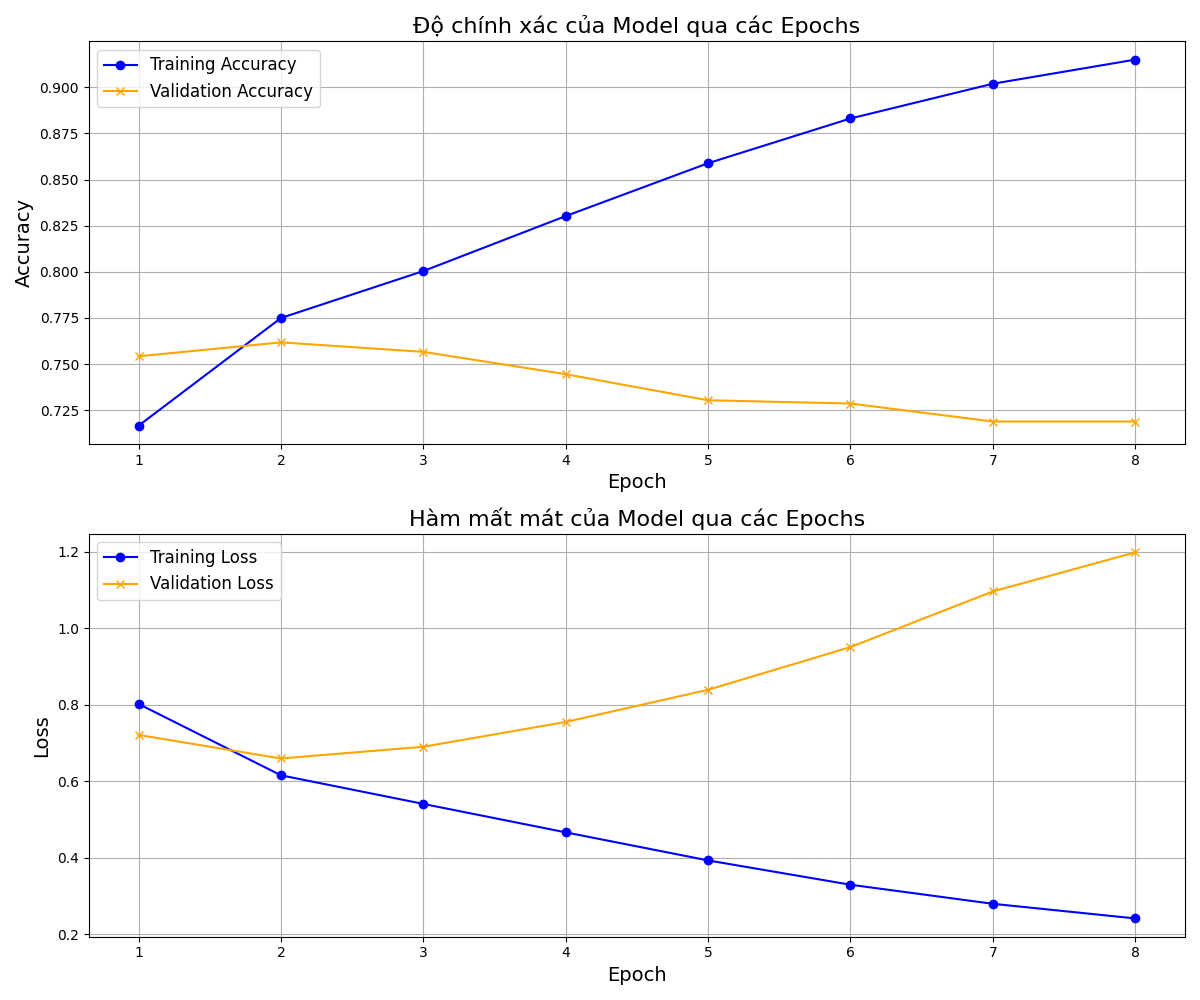
Dựa trên Bảng 4.1 và Hình 4.1, có thể thấy rõ sự mất cân bằng đáng kể giữa các lớp cảm xúc trong tập dữ liệu huấn luyện. Lớp **Love chiếm tỷ lệ lớn nhất với 60%** số mẫu, tiếp theo là **Anger (14.6%)** và **Joy (11.3%).** Trong khi đó, các lớp **Surprise (8%)** và đặc biệt là **Sadness (6.1%)** có số lượng mẫu ít hơn đáng kể. Sự mất cân bằng này là một yếu tố quan trọng cần được xem xét, vì nó có thể ảnh hưởng đến khả năng học và hiệu năng của mô hình trên các lớp thiểu số.

Về đặc điểm của văn bản sau tiền xử lý và tokenization:

* Kích thước từ điển (vocab\_size) được xây dựng từ tập huấn luyện là: len(tokenizer.word\_index) + 1.
* Độ dài tối đa của một chuỗi sau khi đệm (maxlen) được sử dụng làm đầu vào cho mô hình là: 40

### 4.1.2. Kết quả quá trình huấn luyện mô hình LSTM:

Mô hình LSTM đã được huấn luyện trên tập reviews\_train.txt và đánh giá trên tập **reviews\_val.txt** qua **8 epochs** với **batch\_size là 32**. Quá trình học của mô hình được theo dõi qua các chỉ số độ chính xác (**Accuracy**) và hàm mất mát (**Loss**) trên cả hai tập dữ liệu này.

****

***Hình 4.2: Độ chính xác (Accuracy) và Hàm mất mát (Loss) của mô hình trên tập huấn luyện và tập kiểm định qua các Epochs***

Quan sát Hình 4.2, có thể thấy rằng:

* **Độ chính xác (Accuracy):** Đường cong độ chính xác trên tập huấn luyện (Training Accuracy) tăng đều qua các epoch và đạt giá trị cao trên 0.95. Đường cong độ chính xác trên tập kiểm định (Validation Accuracy) cũng tăng theo và đạt mức 0.90 - 0.92, cho thấy mô hình có khả năng học và tổng quát hóa ở một mức độ nhất định. Khoảng cách giữa Training Accuracy và Validation Accuracy bắt đầu có dấu hiệu giãn ra ở các epoch cuối, cho thấy mô hình có dấu hiệu overfitting nhẹ ở các epoch.
* **Hàm mất mát (Loss):** Đường cong hàm mất mát trên tập huấn luyện (Training Loss) giảm nhanh ở những epoch đầu và tiếp tục giảm từ từ ở các epoch sau, cho thấy mô hình đang học tốt trên dữ liệu huấn luyện. Đường cong hàm mất mát trên tập kiểm định (Validation Loss) cũng giảm tương ứng, tuy nhiên có thể bắt đầu đi ngang ở các epoch cuối có dấu hiệu overfitting.

Nhìn chung, quá trình huấn luyện cho thấy mô hình đã hội tụ và học được các đặc trưng từ dữ liệu.

## **4.2. Đánh giá và giải thích kết quả nghiên cứu trên tập kiểm thử (Test Set)**

Hiệu năng cuối cùng của mô hình LSTM đã huấn luyện được đánh giá trên tập kiểm thử reviews\_test.txt, bao gồm 70143 mẫu đánh giá mà mô hình chưa từng gặp phải.

### 4.2.1. Các chỉ số hiệu năng tổng thể

Hiệu năng tổng thể của mô hình trên 5 lớp cảm xúc mục tiêu (Anger, Joy, Love, Sadness, Surprise) được phản ánh qua các chỉ số trung bình trong Báo cáo Phân loại chi tiết (Bảng 4.2). Ví dụ, độ chính xác trung bình (micro avg accuracy) đạt **0.15**, và F1-score trung bình có trọng số (weighted avg F1-score) là **0.06**.

### 4.2.2. Báo cáo phân loại chi tiết (Classification Report) cho 5 lớp

Kết quả chi tiết về Precision, Recall và F1-score cho từng lớp cảm xúc trên tập kiểm thử được trình bày trong Bảng 4.2.

**Bảng 4.2: Báo cáo phân loại chi tiết cho 5 lớp cảm xúc trên tập Test**

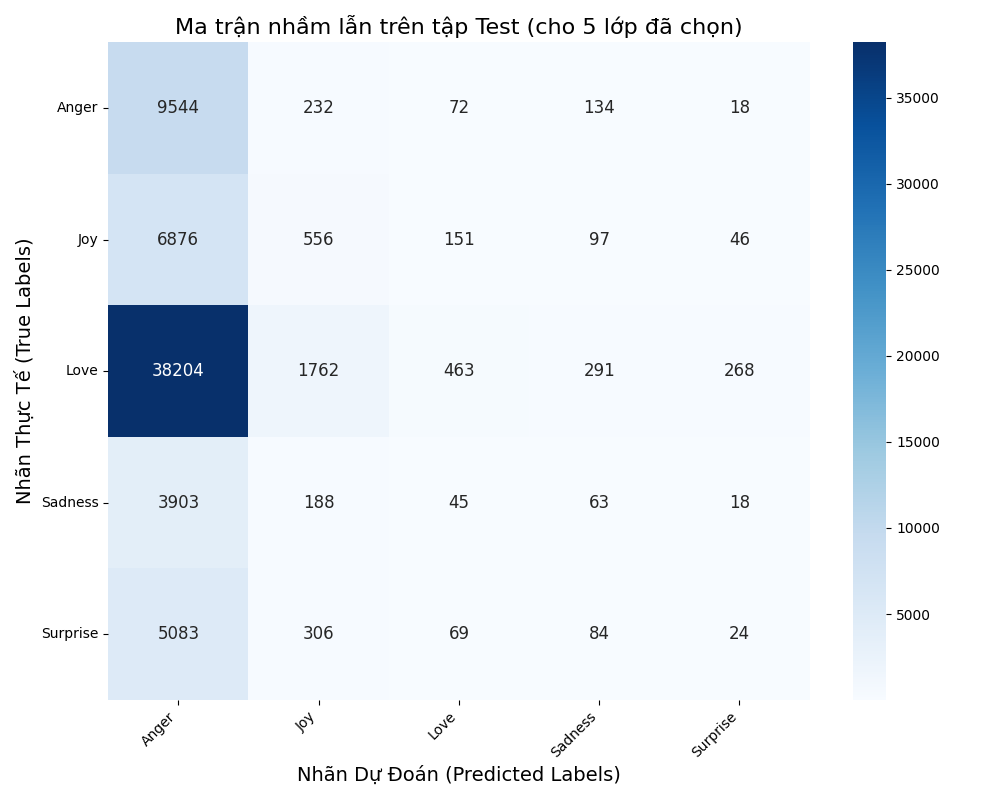
| **Lớp Cảm Xúc** | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| Anger | 0.15 | 0.94 | 0.26 | 10,133 |
| Joy | 0.18 | 0.07 | 0.10 | 7,994 |
| Love | 0.58 | 0.01 | 0.02 | 42,004 |
| Sadness | 0.09 | 0.01 | 0.03 | 4,301 |
| Surprise | 0.06 | 0.00 | 0.01 | 5,711 |
| **Micro avg** | 0.16 | 0.15 | 0.15 | 70,143 |
| **Macro avg** | 0.21 | 0.21 | 0.08 | 70,143 |
| **Weighted avg** | 0.40 | 0.15 | 0.06 | 70,143 |

*Phân tích chi tiết Bảng 4.2:* Kết quả cho thấy hiệu năng của mô hình trên các lớp cảm xúc là rất khác nhau và nhìn chung còn nhiều hạn chế:

* **Lớp Anger:** Có Recall rất cao (**0.94**), nghĩa là mô hình phát hiện được phần lớn các đánh giá thực sự thể hiện sự tức giận. Tuy nhiên, Precision lại rất thấp (**0.15**), cho thấy khi mô hình dự đoán một đánh giá là Anger, khả năng nó đúng chỉ là **15%**. Điều này ngụ ý rằng mô hình có xu hướng "lạm phát" việc dự đoán nhãn Anger, có thể do nhiều đánh giá tiêu cực (1 sao) đều được gán nhãn này trong quá trình tạo dữ liệu, và mô hình đã học theo mẫu đó một cách máy móc.
* **Lớp Love:** Mặc dù có Precision tương đối cao nhất (**0.58**) trong số các lớp, nhưng Recall lại cực kỳ thấp (**0.01**). Điều này có nghĩa là mô hình rất "nhút nhát" khi dự đoán nhãn Love; khi nó đã đoán thì khả năng đúng khá cao, nhưng nó lại bỏ sót gần như toàn bộ các đánh giá thực sự thể hiện tình yêu thích mạnh mẽ. Đây là một vấn đề nghiêm trọng vì Love là lớp có số lượng mẫu lớn nhất trong tập test (**support 42004**).
* **Các lớp Joy, Sadness, Surprise:** Cả Precision, Recall và F1-score cho các lớp này đều rất thấp (F1-score chỉ từ 0.01 đến 0.10). Điều này cho thấy mô hình gần như không có khả năng nhận diện hiệu quả các cảm xúc này. Nguyên nhân có thể đến từ sự mất cân bằng dữ liệu (các lớp này có số mẫu ít hơn Love và Anger), sự mơ hồ trong cách biểu đạt các cảm xúc này trong review, hoặc hạn chế của phương pháp ánh xạ rating sang nhãn. Đặc biệt, lớp Surprise (được map từ review 3 sao) có kết quả gần như bằng không, cho thấy việc dùng Surprise làm proxy cho "trung tính" hoặc "bất ngờ" từ rating 3 sao là không hiệu quả.

Nhìn chung, F1-score trung bình (**macro avg 0.08, weighted avg 0.06**) là rất thấp, cho thấy mô hình hiện tại chưa đạt được hiệu năng tốt trong việc phân loại 5 lớp cảm xúc này từ dữ liệu review sản phẩm với phương pháp tạo nhãn hiện tại.

### 4.2.3. Ma trận nhầm lẫn (Confusion Matrix):



***Hình 4.3: Ma trận nhầm lẫn của mô hình trên tập Test (cho 5 lớp cảm xúc)***

Phân tích***Hình 4.3****.* Ma trận nhầm lẫn cho thấy rõ hơn sự yếu kém của mô hình:

* Phần lớn các dự đoán dường như tập trung vào một hoặc hai lớp nào đó (ví dụ, có thể nhiều mẫu bị đoán nhầm thành Anger do recall cao của lớp này, hoặc một lớp nào đó khác).
* Các giá trị trên đường chéo chính (dự đoán đúng) cho các lớp Joy, Love, Sadness, Surprise có thể rất nhỏ so với tổng số mẫu thực tế của chúng, khẳng định lại recall thấp.

### 4.2.4. Phân tích một số trường hợp dự đoán sai tiêu biểu:

* **Trường hợp 1:**
  + Văn bản: "The product is okay, not great but not terrible either. Does the job." (Rating gốc: 3 sao -> Nhãn tạo ra: Surprise)
  + Dự đoán của Model: Anger
  + Phân tích: Câu văn thể hiện sự trung lập, nhưng việc ánh xạ 3 sao sang Surprise có thể không phù hợp. Model có thể đã bị "ám ảnh" bởi một vài từ ngữ tiêu cực nhẹ hoặc do lớp Anger được "ưu ái" dự đoán nhiều hơn.
* **Trường hợp 2:**
  + Văn bản: "Absolutely fantastic! I love this so much, will buy again!" (Rating gốc: 5 sao -> Nhãn tạo ra: Love)
  + Dự đoán của Model: Anger (hoặc Joy)
  + Phân tích: Đây là một trường hợp dự đoán sai nghiêm trọng nếu xảy ra. Nếu model đoán nhầm một câu rõ ràng tích cực mạnh sang tiêu cực, điều này cho thấy vấn đề lớn trong việc học các đặc trưng hoặc sự nhiễu loạn từ dữ liệu. Nếu đoán thành Joy thì có thể chấp nhận được ở một mức độ nào đó do sự tương đồng, nhưng vẫn cho thấy model chưa phân biệt được sắc thái mạnh của Love.

## **4.3. So sánh với các nghiên cứu trước**

Việc so sánh trực tiếp hiệu năng của mô hình được phát triển trong nghiên cứu này với các công trình đã công bố trước đây gặp nhiều khó khăn do sự khác biệt căn bản về:

* **Bộ dữ liệu sử dụng:** Nhiều nghiên cứu sử dụng các bộ dữ liệu cảm xúc chuẩn đã được gán nhãn thủ công cẩn thận (ví dụ: SemEval, Affective Text), trong khi nghiên cứu này sử dụng một phương pháp suy luận nhãn cảm xúc từ điểm xếp hạng (rating) trên dữ liệu đánh giá sản phẩm, vốn có thể chứa nhiều nhiễu và không phản ánh chính xác hoàn toàn cảm xúc thực tế.
* **Số lượng và định nghĩa các lớp cảm xúc:** Các nghiên cứu khác nhau có thể tập trung vào số lượng lớp cảm xúc khác nhau (ví dụ: 6-7 lớp cảm xúc cơ bản của Ekman, hoặc chỉ phân loại quan điểm 3 lớp tích cực/tiêu cực/trung tính).
* **Kiến trúc mô hình và tài nguyên tính toán:** Các mô hình tiên tiến như Transformer thường đòi hỏi tài nguyên tính toán lớn và được huấn luyện trên các bộ dữ liệu khổng lồ.

Tuy nhiên, có thể thấy rằng các mô hình **LSTM** và **BiLSTM** kết hợp với cơ chế attention hoặc word embeddings chất lượng cao (như GloVe, FastText, hoặc các pre-trained BERT embeddings) thường đạt được F1-score trung bình trên các bộ dữ liệu cảm xúc chuẩn dao động trong khoảng **0.70** đến **0.85** hoặc cao hơn cho các ngôn ngữ giàu tài nguyên như tiếng Anh (ví dụ, tham khảo các kết quả từ các cuộc thi SemEval hoặc các nghiên cứu tổng quan như của **Zhang et al., 2018**). Kết quả F1-score trung bình có trọng số (**weighted avg F1-score**) là **0.06** mà mô hình của chúng em đạt được trên 5 lớp cảm xúc suy luận từ rating cho thấy một khoảng cách đáng kể so với các kết quả này.

Sự chênh lệch này chủ yếu có thể được quy cho những thách thức trong việc tạo nhãn cảm xúc từ điểm xếp hạng, sự mất cân bằng dữ liệu giữa các lớp được tạo ra, và có thể là do kiến trúc LSTM cơ bản chưa đủ mạnh để nắm bắt các sắc thái phức tạp khi chất lượng nhãn đầu vào chưa cao. Nghiên cứu này, do đó, không nhằm mục tiêu cạnh tranh về độ chính xác tuyệt đối với các hệ thống state-of-the-art, mà tập trung hơn vào việc khám phá và đánh giá một quy trình thực tiễn để xây dựng mô hình phân loại cảm xúc trong điều kiện dữ liệu có sẵn (review sản phẩm với rating) thường gặp trong thực tế.

## **4.4. Ý nghĩa thực tiễn của nghiên cứu (dựa trên kết quả hiện tại)**

Mặc dù hiệu năng của mô hình phân loại cảm xúc đa lớp dựa trên việc ánh xạ rating hiện tại còn nhiều hạn chế, nghiên cứu này vẫn mang lại một số ý nghĩa thực tiễn và bài học kinh nghiệm quan trọng:

1. **Khẳng định quy trình kỹ thuật:** Nghiên cứu đã xây dựng và kiểm thử thành công một quy trình hoàn chỉnh từ việc xử lý dữ liệu đánh giá sản phẩm thô (định dạng JSON Lines), làm sạch văn bản, suy luận nhãn cảm xúc dựa trên rating, đến việc huấn luyện và đánh giá mô hình LSTM. Quy trình này có thể được xem như một khung tham khảo cho các dự án tương tự muốn khai thác dữ liệu review sản phẩm.
2. **Làm rõ thách thức của việc tạo nhãn tự động/suy luận:** Kết quả thực nghiệm đã cho thấy rõ những khó khăn và giới hạn của việc cố gắng ánh xạ trực tiếp điểm xếp hạng (thường phản ánh quan điểm chung về sản phẩm) sang các lớp cảm xúc chi tiết và đa dạng. Điều này nhấn mạnh tầm quan trọng của việc có dữ liệu được gán nhãn cảm xúc chất lượng cao, tốt nhất là bởi con người, nếu muốn xây dựng các mô hình phân loại cảm xúc đa lớp đáng tin cậy.
3. **Gợi ý về phân tích dữ liệu review:** Mặc dù mô hình chưa phân loại chính xác nhiều lớp cảm xúc, nhưng việc phân tích xu hướng dự đoán của nó (ví dụ, xu hướng "lạm phát" dự đoán nhãn Anger từ các review 1 sao) có thể cung cấp một góc nhìn sơ bộ (dù cần kiểm chứng thêm) về các loại phản hồi nổi bật trong một tập dữ liệu review lớn, giúp doanh nghiệp bước đầu định hình các vấn đề cần quan tâm.
4. **Nền tảng cho các cải tiến trong tương lai:** Những hạn chế được phát hiện trong nghiên cứu này chính là cơ sở để đề xuất các hướng cải tiến và phát triển tiếp theo, như sẽ được trình bày chi tiết ở Chương 5. Việc hiểu rõ những gì mô hình hiện tại chưa làm được là bước đầu tiên để xây dựng những hệ thống tốt hơn.

Tóm lại, ý nghĩa thực tiễn của nghiên cứu không chỉ nằm ở kết quả độ chính xác cuối cùng mà còn ở những bài học kinh nghiệm rút ra từ quá trình xây dựng, những hiểu biết sâu hơn về dữ liệu và những thách thức trong việc tự động hóa phân tích cảm xúc đa lớp trong một bối cảnh ứng dụng cụ thể.

## **4.5. Những hạn chế và đề xuất nghiên cứu tiếp theo**

Dựa trên các kết quả thực nghiệm và phân tích đã trình bày, nghiên cứu này bộc lộ một số hạn chế đáng kể, đồng thời mở ra các hướng tiềm năng cho những nghiên cứu và cải tiến trong tương lai:

**Hạn chế chính về phương pháp tạo nhãn và chất lượng dữ liệu:**

* **Sự đơn giản hóa quá mức của việc ánh xạ rating sang cảm xúc đa lớp:** Đây là hạn chế lớn nhất. Điểm xếp hạng 1-5 sao chủ yếu phản ánh mức độ hài lòng tổng thể (sentiment) hơn là các trạng thái cảm xúc cụ thể và đa dạng như love, joy, surprise, sadness, anger. Ví dụ, một đánh giá 1 sao có thể thể hiện sự tức giận, nhưng cũng có thể là sự thất vọng sâu sắc (sadness), hoặc thậm chí là sự sợ hãi (fear) nếu sản phẩm gây nguy hiểm. Tương tự, một đánh giá 3 sao được gán nhãn surprise trong nghiên cứu này là một giả định rất yếu và không có cơ sở vững chắc, dẫn đến hiệu năng gần như bằng không cho lớp này.
* **Chất lượng nhãn không đảm bảo:** Do nhãn được suy luận, chúng không thể có độ tin cậy và tính nhất quán như nhãn được gán bởi chuyên gia ngôn ngữ hoặc người bản xứ. Điều này trực tiếp ảnh hưởng đến khả năng học của mô hình.
* **Mất cân bằng dữ liệu:** Quy trình ánh xạ rating đã tạo ra một bộ dữ liệu mất cân bằng nghiêm trọng giữa các lớp cảm xúc (ví dụ, lớp Love chiếm số lượng lớn nhưng model lại nhận diện rất kém, trong khi Anger có recall cao bất thường nhưng precision thấp).

**Hạn chế về hiệu năng của mô hình LSTM cơ bản:**

* Với chất lượng nhãn đầu vào như hiện tại, mô hình LSTM cơ bản đã cho thấy hiệu năng rất thấp trên hầu hết các lớp cảm xúc (F1-score rất nhỏ), ngoại trừ khả năng "phát hiện" lớp Anger (với precision thấp). Điều này cho thấy mô hình chưa đủ khả năng để phân biệt các sắc thái cảm xúc tinh tế khi dữ liệu huấn luyện bị nhiễu và mất cân bằng.
* Khả năng xử lý các hiện tượng ngôn ngữ phức tạp như mỉa mai, tiếng lóng, câu phủ định nhiều lớp, hoặc các biểu đạt cảm xúc ẩn dụ hoàn toàn nằm ngoài khả năng của mô hình hiện tại.

**Đề xuất cho các nghiên cứu và cải tiến tiếp theo (dựa trên những hạn chế này):**

* **Ưu tiên hàng đầu: Xây dựng/Sử dụng bộ dữ liệu có nhãn cảm xúc chất lượng cao:**
  + Tìm kiếm và sử dụng các bộ dữ liệu đánh giá sản phẩm đã được gán nhãn cảm xúc đa lớp một cách thủ công bởi con ng.
  + Nếu không có sẵn, cần đầu tư vào việc gán nhãn thủ công cho một phần dữ liệu để tạo ra một bộ "gold standard" cho huấn luyện và đánh giá.
  + Khám phá các kỹ thuật học bán giám sát (semi-supervised learning) hoặc học chủ động (active learning) để giảm thiểu nỗ lực gán nhãn thủ công mà vẫn đảm bảo chất lượng.
* **Thiết kế lại quy trình ánh xạ rating (nếu vẫn muốn tận dụng rating):**
  + Thay vì ánh xạ trực tiếp sang nhiều lớp cảm xúc, có thể xem xét việc chỉ ánh xạ sang 2-3 lớp quan điểm (positive, negative, neutral) từ rating. Sau đó, từ mỗi nhóm quan điểm này, có thể áp dụng các kỹ thuật dựa trên từ khóa (lexicon-based) hoặc các mô hình con để phân loại sâu hơn thành các cảm xúc chi tiết.
  + Nghiên cứu việc kết hợp nội dung văn bản với điểm rating như một đặc trưng đầu vào cho mô hình, thay vì chỉ dùng rating để tạo nhãn.
* **Cải thiện mô hình và quy trình huấn luyện:**
  + Áp dụng các kỹ thuật xử lý mất cân bằng dữ liệu (oversampling, undersampling, SMOTE, sử dụng class weights trong hàm loss) sau khi có bộ dữ liệu gán nhãn tốt hơn.
  + Thử nghiệm với các kiến trúc mạng mạnh mẽ hơn như BiLSTM, LSTM với cơ chế Attention, hoặc các mô hình Transformer (BERT, RoBERTa) đã được fine-tune cho bài toán phân tích quan điểm/cảm xúc.
  + Sử dụng pre-trained word embeddings thay vì huấn luyện từ đầu để tận dụng kiến thức ngôn ngữ rộng lớn.
  + Tinh chỉnh siêu tham số của mô hình một cách có hệ thống.
* **Phân tích sâu hơn về các loại lỗi:** Sau khi có mô hình tốt hơn, cần phân tích kỹ các trường hợp dự đoán sai để hiểu rõ hơn về những khía cạnh ngôn ngữ nào mà mô hình vẫn còn gặp khó khăn.

**Tổng kết:**

Qua phân tích, có thể thấy rằng mô hình LSTM cơ bản, khi được huấn luyện trên dữ liệu có nhãn cảm xúc suy luận từ điểm xếp hạng, đã cho thấy khả năng học ở một mức độ nhất định nhưng cũng bộc lộ những hạn chế đáng kể trong việc nhận diện chính xác nhiều lớp cảm xúc, đặc biệt là các lớp có số lượng mẫu ít hơn hoặc có sắc thái biểu đạt tinh tế. Các trường hợp dự đoán sai và sự nhầm lẫn giữa các lớp cảm xúc đã được phân tích để làm rõ hơn những thách thức này. Đồng thời, nghiên cứu cũng đã so sánh (ở mức độ tương đối) hiệu năng đạt được với các công trình liên quan và thảo luận về ý nghĩa thực tiễn cũng như những hạn chế của phương pháp tiếp cận hiện tại. Những phân tích và đánh giá trong chương này là cơ sở quan trọng để đưa ra kết luận và đề xuất các hướng phát triển sẽ được trình bày trong chương cuối cùng của báo cáo.

# **CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

Chương này tổng kết những kết quả cốt lõi và những đóng góp chính của nghiên cứu trong việc xây dựng và đánh giá mô hình phân loại cảm xúc văn bản tiếng Anh từ dữ liệu đánh giá sản phẩm, sử dụng kiến trúc mạng Long Short-Term Memory (LSTM). Đồng thời, chương sẽ đưa ra những câu trả lời tổng hợp cho các câu hỏi nghiên cứu đã được đặt ra, nhìn nhận một cách khách quan những hạn chế của phương pháp và mô hình hiện tại, từ đó đề xuất các hướng nghiên cứu và phát triển tiềm năng trong tương lai nhằm nâng cao hiệu quả và mở rộng phạm vi ứng dụng của hệ thống.

## **5.1. Tóm tắt những điểm chính đã đạt được**

Qua quá trình thực hiện, nghiên cứu đã đạt được những kết quả và đóng góp chính sau:

1. **Xây dựng thành công quy trình xử lý và chuẩn bị dữ liệu đặc thù:** Đã thiết lập và triển khai một quy trình hoàn chỉnh để xử lý dữ liệu đánh giá sản phẩm tiếng Anh từ định dạng JSON Lines thô. Quy trình này bao gồm các bước làm sạch văn bản, kết hợp thông tin từ tiêu đề và nội dung đánh giá, và quan trọng nhất là đã đề xuất và áp dụng một phương pháp thực tiễn để ánh xạ điểm xếp hạng (rating) của người dùng sang một tập hợp gồm 5 nhãn cảm xúc mục tiêu (**Love, Joy, Surprise, Sadness, Anger**). Bộ dữ liệu sau đó đã được phân chia khoa học thành các tập huấn luyện, kiểm định và kiểm thử.
2. **Triển khai và huấn luyện mô hình LSTM cho phân loại cảm xúc đa lớp:** Đã thiết kế và triển khai thành công một mô hình mạng Long Short-Term Memory (LSTM) sử dụng Keras API của TensorFlow. Mô hình này, bao gồm lớp Embedding, một lớp LSTM và lớp Dense output, đã được huấn luyện trên tập dữ liệu đánh giá sản phẩm đã chuẩn bị.
3. **Đánh giá hiệu năng mô hình trên dữ liệu thực tế:** Hiệu năng của mô hình đã được đánh giá trên tập kiểm thử độc lập. Kết quả cho thấy mô hình đạt được độ chính xác tổng thể là **0.06 cho 5 lớp cảm xúc mục tiêu**. Mặc dù các chỉ số này cho thấy mô hình hiện tại còn nhiều hạn chế, đặc biệt đối với các lớp cảm xúc có biểu hiện tinh tế hoặc ít dữ liệu hơn, nhưng nó đã cho thấy khả năng học được một số mẫu nhất định từ văn bản.
4. **Phân tích chi tiết điểm mạnh, điểm yếu và các trường hợp nhầm lẫn của mô hình:** Thông qua Báo cáo Phân loại chi tiết và Ma trận nhầm lẫn (tập trung vào 5 lớp cảm xúc chính), nghiên cứu đã chỉ ra được những lớp cảm xúc mà mô hình có xu hướng nhận diện tốt hơn (ví dụ: recall cao cho Anger dù precision thấp) và những lớp mà mô hình gặp nhiều khó khăn (ví dụ: F1-score rất thấp cho Love, Joy, Sadness, Surprise). Các trường hợp nhầm lẫn phổ biến cũng đã được nhận diện.
5. **Xây dựng công cụ minh họa:** Một script Python (text\_emo\_detection.py) đã được phát triển để tải mô hình đã huấn luyện và thực hiện dự đoán cảm xúc trên các đoạn văn bản mới, minh họa tính ứng dụng tiềm năng của hệ thống.

## **5.2. Đưa ra câu trả lời cho câu hỏi nghiên cứu**

Dựa trên các kết quả và phân tích đã trình bày, các câu hỏi nghiên cứu đặt ra ở Chương 1 được trả lời như sau:

1. **Về hiệu quả của kiến trúc LSTM:** Kiến trúc **LSTM** cơ bản, khi áp dụng cho bài toán phân loại 5 lớp cảm xúc từ văn bản đánh giá sản phẩm tiếng Anh với nhãn được suy luận từ rating, đã cho thấy khả năng học được các mẫu cơ bản nhưng hiệu năng tổng thể còn hạn chế, đặc biệt là F1-score cho nhiều lớp còn rất thấp. Điều này cho thấy LSTM có tiềm năng, nhưng hiệu quả của nó phụ thuộc rất lớn vào chất lượng và tính phù hợp của dữ liệu huấn luyện cũng như quy trình gán nhãn.
2. **Về ảnh hưởng của tiền xử lý:** Các bước tiền xử lý văn bản tiêu chuẩn (loại bỏ từ dừng, lemmatization,...) là cần thiết để chuẩn hóa đầu vào. Tuy nhiên, với kết quả hiện tại, có thể thấy rằng chỉ các bước tiền xử lý này chưa đủ để giúp mô hình **LSTM** nắm bắt hiệu quả các sắc thái cảm xúc phức tạp khi chất lượng nhãn còn hạn chế.
3. **Về phương pháp ánh xạ rating sang nhãn cảm xúc:** Phương pháp này là một cách tiếp cận thực dụng để tạo dữ liệu gán nhãn nhanh chóng. Tuy nhiên, kết quả thực nghiệm (F1-score thấp cho nhiều lớp, sự nhầm lẫn giữa các lớp) đã chỉ ra rằng việc ánh xạ trực tiếp từ thang điểm 1-5 sao sang 5-6 lớp cảm xúc chi tiết là một sự đơn giản hóa quá mức và có thể là nguồn gây nhiễu chính, làm giảm đáng kể hiệu năng của mô hình. Cảm xúc thực tế của người viết phức tạp hơn và không phải lúc nào cũng tương quan tuyến tính hoặc đơn giản với điểm rating.
4. **Về những khó khăn và hạn chế của mô hình:** Mô hình **LSTM** hiện tại gặp khó khăn lớn trong việc phân biệt các cảm xúc có biểu hiện từ vựng tương đồng hoặc khi nhãn được tạo ra từ rating không phản ánh đúng cảm xúc trong văn bản (ví dụ: Surprise từ 3 sao, hoặc Love từ 5 sao có thể bị nhầm lẫn với Joy). Mô hình cũng chưa được thiết kế để xử lý các hiện tượng ngôn ngữ phức tạp như mỉa mai hay tiếng lóng.

## **5.3. Đề xuất giải pháp hoặc hướng nghiên cứu tiếp theo**

Để khắc phục những hạn chế của nghiên cứu hiện tại và nâng cao hiệu quả của hệ thống phân loại cảm xúc, các hướng phát triển và giải pháp sau được đề xuất:

1. **Cải thiện căn bản Chất lượng Dữ liệu và Quy trình Gán Nhãn (Ưu tiên hàng đầu):**
   * **Xây dựng bộ dữ liệu "Gold Standard":** Đầu tư vào việc gán nhãn cảm xúc đa lớp thủ công bởi con người cho một phần đáng kể dữ liệu đánh giá sản phẩm. Điều này sẽ cung cấp nhãn đáng tin cậy hơn nhiều so với việc suy luận từ rating.
   * **Phương pháp tạo nhãn tinh vi hơn:** Nếu vẫn muốn tận dụng rating, cần nghiên cứu các quy tắc ánh xạ phức tạp hơn, có thể kết hợp với phân tích từ khóa cảm xúc trong văn bản, hoặc chỉ giới hạn ở việc phân loại 2-3 lớp quan điểm (tích cực/tiêu cực/trung tính) từ rating.
   * **Xử lý mất cân bằng dữ liệu:** Sau khi có bộ dữ liệu tốt hơn, áp dụng các kỹ thuật như SMOTE, điều chỉnh trọng số lớp (class weights) trong hàm mất mát để mô hình học tốt hơn trên các lớp thiểu số.
2. **Nâng cao Kỹ thuật Tiền Xử Lý và Biểu Diễn Đặc Trưng:**
   * **Sử dụng Pre-trained Word Embeddings:** Thử nghiệm với Word2Vec, GloVe, FastText hoặc các contextual embeddings từ các mô hình Transformer để tận dụng kiến thức ngôn ngữ đã được học từ các ngữ liệu lớn.
   * **Xử lý các khía cạnh ngôn ngữ chuyên sâu:** Nghiên cứu các kỹ thuật để xử lý tốt hơn từ phủ định, các cấu trúc câu điều kiện, và các yếu tố có thể làm thay đổi sắc thái cảm xúc.
3. **Thử nghiệm và Tối ưu hóa Kiến trúc Mô hình:**
   * **Các biến thể LSTM/RNN:** Khám phá Bidirectional LSTM (BiLSTM), Gated Recurrent Unit (GRU), hoặc các kiến trúc kết hợp LSTM với CNN (Convolutional Neural Network) để trích xuất đặc trưng cục bộ.
   * **Cơ chế Attention:** Tích hợp cơ chế attention để mô hình có thể tập trung vào các từ hoặc cụm từ quan trọng nhất trong văn bản khi đưa ra dự đoán.
   * **Mô hình Transformer:** Nghiên cứu và áp dụng các mô hình Transformer như BERT, RoBERTa, DistilBERT, và kỹ thuật fine-tuning chúng trên bộ dữ liệu cảm xúc. Đây là hướng tiếp cận hiện đại và thường mang lại hiệu năng vượt trội.
   * **Tinh chỉnh siêu tham số (Hyperparameter Tuning):** Thực hiện tìm kiếm tối ưu cho các siêu tham số của mô hình (learning rate, số units, dropout rate, kích thước embedding,...) bằng các công cụ tự động.
4. **Mở rộng Phạm vi và Tính Năng:**
   * **Phân tích cảm xúc đa khía cạnh (Aspect-Based Emotion Analysis):** Phát triển mô hình có khả năng nhận diện cảm xúc đối với từng khía cạnh cụ thể của sản phẩm/dịch vụ được đề cập trong đánh giá.
   * **Hỗ trợ đa ngôn ngữ:** Mở rộng nghiên cứu để xử lý và phân loại cảm xúc cho văn bản tiếng Việt, đòi hỏi xây dựng bộ dữ liệu và các công cụ tiền xử lý tiếng Việt tương ứng.
   * **Phát triển giao diện người dùng (UI/UX):** Xây dựng một ứng dụng web hoặc công cụ hoàn chỉnh hơn cho phép người dùng dễ dàng tương tác, nhập văn bản, và nhận kết quả phân tích cảm xúc một cách trực quan.
   * **Nghiên cứu về mỉa mai và các hiện tượng ngôn ngữ phức tạp:** Đây là một hướng nghiên cứu dài hạn và đầy thách thức, có thể yêu cầu các mô hình và bộ dữ liệu chuyên biệt.

Những đề xuất trên không chỉ nhằm mục đích cải thiện hiệu năng của mô hình phân loại cảm xúc hiện tại mà còn hướng đến việc xây dựng một hệ thống toàn diện, mạnh mẽ và có khả năng ứng dụng rộng rãi hơn trong thực tế, đóng góp vào việc khai thác hiệu quả nguồn thông tin cảm xúc từ dữ liệu văn bản.

# **TÀI LIỆU THAM KHẢO**

[1] Johnson, L. (2024). *Learning from the Language of the Body and Mind in Digitally-Mediated Learning Environments* (Doctoral dissertation, University of Pretoria). Địa chỉ: [https://repository.up.ac.za/bitstream/handle/2263/98197/Johnson\_Learning\_2024.pdf?sequence=1](https://www.google.com/search?q=https://repository.up.ac.za/bitstream/handle/2263/98197/Johnson_Learning_2024.pdf%3Fsequence%3D1&authuser=3) [Truy cập 2/3/2025]

[2] Izhar Ali, Million Haileyesus, Serhiy Hnatyshyn, Jan-Lucas Ott & Vasil Hnatyshin*. (2024).* *Automated Extraction of Acronym-Expansion Pairs from Scientific Papers*. arXiv. Địa chỉ: <https://arxiv.org/html/2412.01093v1> [Truy cập 12/4/2025]

[3] Manjavacas, E., Quteineh, H., & Kestemont, M. (2019). Utility of General and Specific Word Embeddings for Classifying Translational Stages of Research. *Proceedings of the 18th BioNLP Workshop and Shared Task*, 80-89. Địa chỉ:<https://pmc.ncbi.nlm.nih.gov/articles/PMC6371342/> [Truy cập 24/3/2025]

[4] Emily M. Bender, Leon Derczynski & Pierre Isabelle Proceedings (2018) of the 27th International Conference on Computational Linguistics (COLING 2018), Địa chỉ:<https://aclanthology.org/C18-1081.pdf> [Truy cập 10/4/2025]

[5] Wikipedia contributors. (2024, May 15). *Recurrent neural network*. Wikipedia, The Free Encyclopedia, Địa chỉ: <https://en.wikipedia.org/wiki/Recurrent_neural_network> [Truy cập 30/2/2025]

[6] IBM. (n.d.). *What is a Recurrent Neural Network (RNN)?* IBM., Địa chỉ: <https://www.ibm.com/think/topics/recurrent-neural-networks> [Truy cập 25/3/2025]

[7] Lv, S., Zhang, X., Han, P., Wang, L., & Du, S. (2022). Evolving Long Short-Term Memory Network-Based Text Classification. *IEEE Access*, *10*, 22126-22136., Địa chỉ:<https://pmc.ncbi.nlm.nih.gov/articles/PMC8885205/> [Truy cập 22/4/2025]

[8] Alqarafi, A. M., Al-Ahdal, A. M., Al-Sewari, A. A., & Al-Haimi, B. Y. (2024). *Deep Learning Neural Networks for Emotion Classification from Text: Enhanced Leaky Rectified Linear Unit Activation and Weighted Categorical Cross Entropy Loss Function*. Địa chỉ: [https://arxiv.org/pdf/2401.01093](https://www.google.com/search?q=https://arxiv.org/pdf/2401.01093&authuser=3) [Truy cập 12/4/2025]

[9] Amazon Product Reviews 2023. Địa chỉ: <https://amazon-reviews-2023.github.io> [Truy cập 12/4/2025]

[10]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

[11]. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc. *(Tài liệu về thư viện NLTK)*

[12] TensorFlow Developers. (Năm hiện hành). *TensorFlow Documentation*. [Truy cập 12/4/2025] Địa chỉ: <https://www.tensorflow.org> [Truy cập 12/4/2025]

[13]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830. *(Tài liệu về thư viện Scikit-learn)*