KHOA KỸ THUẬT VÀ CÔNG NGHỆ
**BỘ MÔN CÔNG NGHỆ THÔNG TIN**

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

# ĐỀ CƯƠNG CHI TIẾT
# KHÓA LUẬN TỐT NGHIỆP NGÀNH CÔNG NGHỆ THÔNG TIN

Họ tên sinh viên: Dương Văn Hiệp MSSV: 110121209

Lớp: DA21TTB Khóa: 2021

Tên đề tài: Nhận diện cảm xúc từ văn bản tiếng Việt bằng mô hình NLP

1. Mục tiêu của đề tài:

Đề tài "Nhận diện cảm xúc từ văn bản tiếng Việt bằng mô hình NLP" hướng đến việc nghiên cứu, xây dựng và đánh giá một hệ thống có khả năng tự động nhận diện cảm xúc từ các đoạn văn bản tiếng Việt bằng cách ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) kết hợp với các mô hình học máy hiện đại.

**Về mục tiêu nghiên cứu – học tập:**

Làm quen và thực hành triển khai các mô hình học máy (Machine Learning) và học sâu (Deep Learning), bao gồm cả các kiến trúc mạng nơ-ron như RNN, LSTM, BERT, hay Transformer trong bài toán phân tích cảm xúc.

Nâng cao kiến thức và kỹ năng chuyên môn trong lĩnh vực xử lý ngôn ngữ tự nhiên, đặc biệt là đối với tiếng Việt – một ngôn ngữ có đặc thù riêng về cú pháp và ngữ nghĩa. Tìm hiểu và áp dụng các kỹ thuật tiền xử lý văn bản, vector hóa, trích xuất đặc trưng và phân loại cảm xúc trên văn bản tự nhiên.

Rèn luyện kỹ năng xây dựng quy trình nghiên cứu khoa học: từ việc thu thập dữ liệu, xử lý dữ liệu, huấn luyện mô hình, đánh giá kết quả cho đến việc tối ưu và trình bày hệ thống.

**Về mục tiêu ứng dụng,** đề tài hướng đến việc xây dựng một mô hình hoặc hệ thống có khả năng phân loại cảm xúc trong văn bản tiếng Việt với độ chính xác cao, từ đó có thể được áp dụng vào các lĩnh vực thực tiễn như:

Phân tích ý kiến người dùng trong các bài đánh giá sản phẩm, dịch vụ trên mạng xã hội, diễn đàn hoặc sàn thương mại điện tử.

Hỗ trợ giám sát dư luận xã hội, phát hiện xu hướng hoặc phản ứng tiêu cực tích cực từ cộng đồng đối với một sự kiện, cá nhân hoặc tổ chức.

Tăng cường khả năng tương tác của chatbot, trợ lý ảo bằng cách giúp hệ thống hiểu được cảm xúc của người dùng trong quá trình giao tiếp.

Làm nền tảng cho các nghiên cứu mở rộng về phân tích ngữ nghĩa, tóm tắt văn bản cảm xúc, hoặc tạo ra các hệ thống phản hồi thông minh, có cảm xúc trong tương lai.

**Tóm lại,** đề tài không chỉ phục vụ cho mục tiêu học thuật mà còn hướng đến việc tạo ra giá trị thực tiễn, góp phần thúc đẩy các nghiên cứu và ứng dụng công nghệ AI – NLP trong ngôn ngữ tiếng Việt.

2. Nội dung thực hiện:

**Tìm hiểu khái niệm và các hướng tiếp cận trong phân tích cảm xúc (Sentiment Analysis):** Đề tài sẽ tiến hành nghiên cứu các khái niệm nền tảng liên quan đến bài toán phân tích cảm xúc, bao gồm định nghĩa cảm xúc trong ngữ cảnh văn bản, các mức độ phân loại cảm xúc phổ biến (như tích cực, tiêu cực, trung lập hoặc các nhãn chi tiết hơn như vui, buồn, tức giận, sợ hãi,...), cùng với các cấp độ phân tích (mức độ câu, đoạn văn, tài liệu). Ngoài ra, đề tài sẽ khảo sát các hướng tiếp cận trong lĩnh vực này, bao gồm phương pháp dựa trên từ điển cảm xúc, cũng như các phương pháp hiện đại dựa trên mô hình học máy và học sâu. Việc nắm vững các hướng tiếp cận sẽ là cơ sở để xây dựng chiến lược phù hợp cho tiếng Việt – một ngôn ngữ có đặc điểm khác biệt so với tiếng Anh.

**Khảo sát và lựa chọn tập dữ liệu tiếng Việt phù hợp:** Dữ liệu là yếu tố then chốt trong việc huấn luyện và đánh giá mô hình phân tích cảm xúc. Do đó, đề tài sẽ thực hiện khảo sát, thu thập và đánh giá các bộ dữ liệu tiếng Việt có sẵn phục vụ cho bài toán nhận diện cảm xúc. Một số nguồn dữ liệu có thể được khai thác bao gồm các diễn đàn, trang đánh giá sản phẩm, bình luận mạng xã hội như Facebook, Shopee, Tiki, hoặc dữ liệu từ các bài báo, blog. Trong trường hợp dữ liệu không đủ hoặc chưa được gán nhãn đầy đủ, đề tài có thể thực hiện

bước tiền xử lý và gán nhãn thủ công hoặc bán tự động. Dữ liệu sau khi thu thập sẽ được làm sạch, chuẩn hóa và phân chia thành tập huấn luyện, tập kiểm thử và tập kiểm tra.

**Tìm hiểu và ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) đối với tiếng Việt:** Văn bản tiếng Việt có nhiều thách thức trong NLP như ngôn ngữ không dấu, từ đa nghĩa, cú pháp linh hoạt. Do đó, đề tài sẽ nghiên cứu các kỹ thuật tiền xử lý ngôn ngữ phù hợp như tách từ, chuẩn hóa văn bản, loại bỏ từ dừng, xử lý từ viết tắt, biểu tượng cảm xúc,... Ngoài ra, đề tài cũng sẽ xem xét các phương pháp biểu diễn văn bản (text representation) như Word2Vec từ các mô hình đã được huấn luyện như PhoBERT – một mô hình BERT dành riêng cho tiếng Việt.

**Nghiên cứu và triển khai các thuật toán học máy (Machine Learning):** Sau khi hoàn thành bước xử lý dữ liệu, đề tài sẽ thử nghiệm các mô hình học máy cổ điển như Naive Bayes, SVM (Support Vector Machine), Decision Tree, KNN,... Các mô hình này sẽ được huấn luyện với dữ liệu đã xử lý để làm cơ sở đánh giá và so sánh hiệu năng ban đầu. Các tham số mô hình sẽ được tinh chỉnh (tuning) để cải thiện độ chính xác, đồng thời đánh giá độ hiệu quả qua các chỉ số như Accuracy, Precision, Recall, F1-score.

**Nghiên cứu và áp dụng các mô hình học sâu (Deep Learning):** Trong giai đoạn tiếp theo, đề tài sẽ nghiên cứu các kiến trúc học sâu hiện đại nhằm nâng cao độ chính xác và khả năng biểu diễn ngữ nghĩa sâu hơn. Các mô hình như CNN, RNN, LSTM, BiLSTM và đặc biệt là các mô hình transformer như BERT, PhoBERT sẽ được xem xét triển khai. Đề tài cũng sẽ đánh giá hiệu quả của các mô hình pre-trained được huấn luyện trên dữ liệu tiếng Việt để tận dụng sức mạnh học biểu diễn ngữ nghĩa từ mô hình lớn. Việc huấn luyện và tinh chỉnh các mô hình này đòi hỏi xử lý tốt các yếu tố kỹ thuật như điều chỉnh learning rate, batch size, epoch, dropout,... và sử dụng GPU để tối ưu thời gian huấn luyện.

**Đánh giá, so sánh mô hình và tổng hợp kết quả:** Sau khi huấn luyện các mô hình học máy và học sâu, đề tài sẽ tiến hành đánh giá và so sánh kết quả theo các tiêu chí định lượng. Việc đánh giá không chỉ dựa trên độ chính xác, mà còn xét đến tính tổng quát, khả năng mở rộng và thời gian xử lý. Qua đó, lựa chọn mô hình tối ưu để làm mô hình cuối cùng.

**Xây dựng hệ thống thử nghiệm và trình bày kết quả:** Sau khi hoàn thiện quá trình nghiên cứu và huấn luyện mô hình nhận diện cảm xúc, đề tài sẽ tiến hành xây dựng một ứng dụng thử nghiệm có khả năng phân tích và phân loại cảm xúc của người dùng từ các phản hồi, bình luận trên các trang thương mại điện tử. Hệ thống đóng vai trò như một minh chứng thực tế cho hiệu quả của mô hình đã xây dựng, đồng thời thể hiện tiềm năng ứng dụng trong các bài toán công nghiệp.

3. Phương pháp thực hiện:

Để đạt được mục tiêu của đề tài, quá trình thực hiện sẽ được chia thành các bước cụ thể, tuần tự theo quy trình xử lý dữ liệu và phát triển mô hình học máy/học sâu trong lĩnh vực xử lý ngôn ngữ tự nhiên. Cụ thể, các phương pháp được áp dụng bao gồm:

**Khảo sát lý thuyết và tổng quan nghiên cứu:** Tìm hiểu các khái niệm liên quan đến phân tích cảm xúc (sentiment analysis), các cấp độ cảm xúc (câu, đoạn, tài liệu) và các loại cảm xúc phổ biến. Nghiên cứu các phương pháp dựa trên học máy và học sâu. Khảo sát các công trình nghiên cứu trước đó liên quan đến phân tích cảm xúc trong tiếng Việt và các ngôn ngữ khác.

**Thu thập và xử lý dữ liệu:** Tìm kiếm và lựa chọn các bộ dữ liệu có sẵn chứa văn bản tiếng Việt được gán nhãn cảm xúc, thực hiện các bước tiền xử lý ngôn ngữ tự nhiên, bao gồm làm sạch văn bản (loại bỏ ký tự đặc biệt, liên kết, html tag,...), tách từ bằng thư viện, chuẩn hóa văn bản, mã hóa văn bản.

**Xây dựng mô hình học máy và học sâu:** Thử nghiệm các mô hình học máy cổ điển để làm cơ sở so sánh, triển khai các mô hình học sâu phổ biến cho xử lý chuỗi văn bản, tinh chỉnh siêu tham số và tối ưu mô hình.

**Huấn luyện và đánh giá mô hình:** Chia dữ liệu thành các tập huấn luyện, kiểm thử và kiểm tra theo tỷ lệ hợp lý, sử dụng các thước đo đánh giá như **Accuracy**, **Precision**, **Recall**, **F1-score**, **Confusion Matrix** để đo lường hiệu quả mô hình, so sánh kết quả giữa các mô hình học máy, học sâu và mô hình tiền huấn luyện để chọn ra phương án tối ưu.

**Triển khai ứng dụng thử nghiệm:** Phát triển một ứng dụng web đơn giản cho phép người dùng nhập văn bản phản hồi hoặc bình luận sản phẩm từ các sàn thương mại điện tử, và nhận kết quả phân tích cảm xúc ngay lập tức. Ứng dụng sử dụng mô hình đã huấn luyện như một REST API phía backend (có thể dùng Flask, FastAPI), frontend đơn giản (có thể dùng React). Có thể bổ sung tính năng phân tích hàng loạt phản hồi, biểu đồ thống kê cảm xúc, lọc theo sản phẩm.

**Tổng hợp kết quả và đề xuất hướng phát triển:** Đưa ra nhận xét, phân tích những ưu – nhược điểm của các mô hình đã thử nghiệm. So sánh hiệu quả mô hình với các kết quả trong các công trình nghiên cứu khác. Đề xuất hướng phát triển tiếp theo như mở rộng phạm vi cảm xúc, áp dụng cho các lĩnh vực khác như y tế, giáo dục, chăm sóc khách hàng,...

4. Bố cục đề tài:

**Chương 1: Giới thiệu đề tài:** Lý do chọn đề tài, mục tiêu đề tài, đối tượng và phạm vi nghiên cứu, phương pháp nghiên cứu, cấu trúc báo cáo.

**Chương 2: Cơ sở lý thuyết:** Tổng quan về xử lý ngôn ngữ tự nhiên (NLP), khái niệm và vai trò của phân tích cảm xúc (Sentiment Analysis), các kỹ thuật biểu diễn văn bản, tổng quan về học máy (Machine Learning) và học sâu (Deep Learning), các mô hình phổ biến dùng trong phân tích cảm xúc.

**Chương 3: Phân tích và thiết kế hệ thống:** Phân tích bài toán nhận diện cảm xúc trong văn bản tiếng Việt, yêu cầu hệ thống (chức năng và phi chức năng), thiết kế quy trình xử lý dữ liệu, thiết kế kiến trúc hệ thống (mô hình tổng thể, các thành phần chính), mô hình luồng dữ liệu và xử lý, lựa chọn công nghệ.

**Chương 4: Thực nghiệm và triển khai:** Thu thập và xử lý dữ liệu văn bản tiếng Việt, tiền xử lý, vector hóa văn bản, huấn luyện các mô hình học máy (ML) và học sâu (DL), đánh giá mô hình bằng các thước đo, so sánh kết quả giữa các mô hình, tinh chỉnh và lựa chọn mô hình tối ưu.

**Chương 5: Xây dựng ứng dụng thử nghiệm:** Mục tiêu và chức năng ứng dụng, thiết kế giao diện người dùng, kết nối mô hình với hệ thống phân tích phản hồi từ thương mại điện tử, mô phỏng quy trình, thử nghiệm kết quả ứng dụng thực tế.

**Chương 6: Đánh giá và kết luận:** Tổng kết kết quả đạt được, đánh giá hiệu quả mô hình và hệ thống, khó khăn, hạn chế trong quá trình thực hiện, hướng phát triển và mở rộng trong tương lai.

**Tài liệu tham khảo, Phụ lục.**

5. Tài liệu tham khảo:

Các tài liệu tham khảo sử dụng trong đề tài được lựa chọn từ nhiều nguồn uy tín, bao gồm bài báo khoa học quốc tế, tài liệu nghiên cứu trong nước, sách chuyên ngành và các nguồn dữ liệu, công cụ thực tế. Những tài liệu này đóng vai trò quan trọng trong việc xây dựng cơ sở lý thuyết, lựa chọn mô hình và định hướng triển khai hệ thống. Một số tài liệu tham khảo hiện tại:

[1] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media, 2009.

[2] F. Chollet, *Deep Learning with Python*, 2nd ed. Shelter Island, NY: Manning Publications, 2021.

[3] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained Language Models for Vietnamese," 2020. [Online]. Available: https://arxiv.org/abs/2003.00744

[4] T. M. Triết, Giáo trình Xử lý ngôn ngữ tự nhiên. TP. Hồ Chí Minh: NXB Đại học Quốc gia TP.HCM, 2017.

[5] Nguyễn Thanh Tuấn, Deep Learning cơ bản, 2021. [Online]. Available: https://drive.google.com/file/d/1lNjzISABdoc7SRq8tg-xkCRRZRABPCKi/view

6. Kế hoạch thực hiện đề tài

| Tuần | Từ ngày - đến ngày | Công việc thực hiện | Ghi chú |
|---|---|---|---|
| 1 | Từ ngày 08/4/2025 đến 13/4/2025 | - Nghiên cứu, tìm hiểu tổng quan về lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và bài toán nhận diện cảm xúc tiếng Việt.<br>- Xác định mục tiêu nghiên cứu và phạm vi thực hiện.<br>- Tìm kiếm, tổng hợp tài liệu tham khảo: sách, bài báo, tài nguyên học thuật bằng tiếng Anh và tiếng Việt.<br>- Viết đề cương chi tiết và xây dựng kế hoạch thực hiện đề tài. | |
| 2 | Từ ngày 14/4/2025 đến 20/4/2025 | - Nghiên cứu sâu hơn về các khái niệm liên quan: phân tích cảm xúc (Sentiment Analysis), các loại cảm xúc và các hướng tiếp cận phổ biến.<br>- Tìm hiểu các đặc trưng của văn bản tiếng Việt trong xử lý ngôn ngữ tự nhiên.<br>- Nghiên cứu về các kỹ thuật tiền xử lý văn bản: tách từ, loại bỏ stop words, chuẩn hóa văn bản tiếng Việt. | |
| 3 | Từ ngày 21/4/2025 đến 27/4/2025 | - Tìm kiếm, đánh giá và lựa chọn bộ dữ liệu phục vụ bài toán (hoặc crawl dữ liệu từ phản hồi khách hàng trên các sàn thương mại điện tử). | |

| | | | |
|---|---|---|---|
| | | - Tiến hành tiền xử lý dữ liệu: làm sạch, chuẩn hóa, gán nhãn cảm xúc nếu bộ dữ liệu chưa được gán nhãn.<br>- Phân tích và trực quan hóa dữ liệu để hiểu rõ hơn về phân bố cảm xúc và cấu trúc dữ liệu. | |
| 4 | Từ ngày 28/4/2025 đến 04/5/2025 | - Tìm hiểu các thuật toán học máy (Machine Learning) và học sâu (Deep Learning) phổ biến cho bài toán phân tích cảm xúc<br>- Cài đặt và thử nghiệm các mô hình cơ bản với tập dữ liệu đã xử lý. | |
| 5 | Từ ngày 05/5/2025 đến 11/5/2025 | - Huấn luyện các mô hình học máy và học sâu đã chọn.<br>- Tiến hành điều chỉnh siêu tham số, cải tiến hiệu suất và so sánh kết quả giữa các mô hình.<br>- Ghi nhận các khó khăn trong quá trình huấn luyện và chuẩn bị giải pháp khắc phục. | |
| 6 | Từ ngày 12/5/2025 đến 18/5/2025 | | |
| 7 | Từ ngày 19/5/2025 đến 25/5/2025 | - Xây dựng hệ thống thử nghiệm: thiết kế giao diện ứng dụng website cho phép nhập văn bản đầu vào và hiển thị kết quả phân tích cảm xúc.<br>- Tích hợp mô hình phân tích cảm xúc vào hệ thống, đảm bảo hoạt động trơn tru trên dữ liệu đầu vào thực tế. | |
| 8 | Từ ngày 26/5/2025 đến 01/6/2025 | - Kiểm thử ứng dụng với dữ liệu thực tế: phản hồi khách hàng, bài viết mạng xã hội, hoặc nhận xét từ trang thương mại điện tử. | |

| | | | |
|---|---|---|---|
| | | - Cải thiện mô hình và tối ưu hiệu suất dựa trên kết quả thực nghiệm. | |
| 9 | Từ ngày 02/6/2025 đến 08/6/2025 | - Tổng hợp kết quả thử nghiệm mô hình.<br>- Viết thực nghiệm và kết quả vào báo cáo: phân tích chi tiết độ chính xác, ưu nhược điểm của các mô hình, và đánh giá hiệu quả hệ thống đã xây dựng. | |
| 10 | Từ ngày 09/6/2025 đến 15/6/2025 | - Hoàn thiện các nội dung còn lại của quyển báo cáo.<br>- Rà soát lỗi chính tả, định dạng, trích dẫn tài liệu tham khảo.<br>- Nộp báo cáo tốt nghiệp hoàn chỉnh đúng hạn. | |

*Trà Vinh, ngày 17 tháng 4 năm 2025*

**GIẢNG VIÊN HƯỚNG DẪN** | **SINH VIÊN THỰC HIỆN**

**Nguyễn Nhứt Lam** | **Dương Văn Hiệp**