TRƯỜNG ĐẠI HỌC KỸ THUẬT – CÔNG NGHỆ CẦN THƠ

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

Ngành: Khoa học máy tính

MÔ HÌNH PHÂN LOẠI CHỦ ĐỀ TỰ ĐỘNG CHO BẢN TIN THỜI SỰ TRUYỀN HÌNH

TỪ THÁI BẢO

Cần Thơ, năm 2023

TRƯỜNG ĐẠI HỌC KỸ THUẬT – CÔNG NGHỆ CẦN THƠ

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

Ngành: Khoa học máy tính

MÔ HÌNH PHÂN LOẠI CHỦ ĐỀ TỰ ĐỘNG CHO BẢN TIN THỜI SỰ TRUYỀN HÌNH

Cán bộ hướng dẫn:

ThS: Nguyễn Tấn Phú

Sinh viên thực hiện:

Từ Thái Bảo

MSSV: 1900222

Cần Thơ, năm 2023

XÁC NHẬN ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

Ngành: Khoa học máy tính

MÔ HÌNH PHÂN LOẠI CHỦ ĐỀ TỰ ĐỘNG CHO BẢN TIN THỜI SỰ TRUYỀN HÌNH

CÁN BỘ HƯỚNG DẪN	**SINH VIÊN THỰC HIỆN**
(Ký và ghi rõ họ tên)	*(Ký và ghi rõ họ tên)*

Ngày bảo vệ:……../…..…/……

TRƯỞNG BAN	**CÁN BỘ PHẢN BIỆN**	**THƯ KÝ**
(Ký và ghi rõ họ tên)	*(Ký và ghi rõ họ tên)*	*(Ký và ghi rõ họ tên)*

LỜI CAM ĐOAN

Tôi tên: Từ Thái Bảo, MSSV: 1900222, Lớp: Khoa học máy tính 0119.

Tôi xin cam đoan rằng đây là công trình nghiên cứu của chính bản thân tôi và được sự hướng dẫn thực hiện của ThS Nguyễn Tấn Phú.

Các số liệu, kết quả thực nghiệm là trung thực. Mọi thứ được dựa trên sự cố gắng cũng như sự nỗ lực của bản thân. Tài liệu tham khảo có trong đồ án tốt nghiệp đã được liệt kê và nêu rõ ra tại danh mục tài liệu tham khảo.

Tôi xin cam đoan những điều được trình bày ở trên là đúng sự thật, nếu sai sót xin hoàn toàn chịu trách nhiệm.

Cần Thơ, ngày 19 tháng 11 năm 2023

Sinh viên thực hiện

Từ Thái Bảo

LỜI CẢM ƠN

Trước tiên với tình cảm sâu sắc và chân thành nhất, cho phép tôi được bày tỏ lòng biết ơn đến tất cả các cá nhân và tổ chức đã tạo điều kiện hỗ trợ, giúp đỡ tôi trong suốt quá trình học tập và thực hiện đề tài này.

Tôi chân thành gửi lời cảm ơn đến Thầy Nguyễn Tấn Phú, người đã tận tâm hướng dẫn và trực tiếp giúp đỡ tôi trong suốt quá trình nghiên cứu đồ án. Với sự chỉ bảo nhiệt tình của Thầy, tôi đã có định hướng tốt trong việc triển khai và thực hiện các yêu cầu trong quá trình làm đồ án.

Tôi xin chân thành gửi lời cảm ơn đến gia đình. Những người đã luôn giành cho tôi những tình cảm yêu thương nhất, những người đã luôn hỗ trợ theo dõi những bước đi trong suốt cuộc hành trình vừa qua. Đã gửi những lời động viên trong suốt quá trình thực hiện nghiên cứu.

Tôi xin chân thành gửi lời cảm ơn đến tất cả các Thầy Cô khoa Công nghệ thông tin đã tạo điều kiện cho tôi trong suốt quá trình học tập. Những người đã dìu dắt tôi tận tình, đã truyền đạt cho tôi những kiến thức và những bài học quý giá trong suốt thời gian tôi theo học tại trường.

Và cuối cùng cảm ơn tất cả bạn bè, những người đã sát cánh cùng tôi qua những niềm vui, cùng chia sẻ những khó khăn, thử thách đã làm cho hành trình này trở nên ý nghĩa hơn bao giờ hết. Cảm ơn các bạn vì đã luôn tin tưởng và ủng hộ tôi.

Xin chân thành cảm ơn!

TÓM TẮT

Trong nghiên cứu này, giới thiệu về các phương pháp tiếp cận để phân loại các chủ đề cho bản tin thời sự truyền hình với độ chính xác cao. Sử dụng phương pháp biểu diễn văn bản bằng đặc trưng TF-IDF kết hợp với các mô hình học máy SVM, KNN, CNN và PhoBERT để so sánh độ chính xác của các mô hình này. Dữ liệu nghiên cứu là các video về bản tin thời sự truyền hình tại Việt Nam, được thu thập từ YouTube và thực hiện chuyển đổi thành văn bản. Tổng cộng có 14.503 mẫu, chia thành 11 chủ đề khác nhau như: Chính trị Xã hội, Dự báo thời tiết, Kinh tế, Môi trường, Nông nghiệp, Pháp luật, Sức khỏe, Thế giới, Thể thao, Văn hóa và Giáo dục. Kết quả thực nghiệm cho thấy phương pháp sử dụng mô hình PhoBERT đạt độ chính xác lên tới 98%, vượt trội so với các mô hình khác được thử nghiệm như SVM, KNN và CNN. Nghiên cứu này mở ra triển vọng mới trong ứng dụng học máy cho việc phân loại chủ đề bản tin thời sự truyền hình, đặc biệt là sự hiệu quả của mô hình PhoBERT.

Từ khoá: *phân loại chủ đề, bản tin, thời sự, TF-IDF, SVM, KNN CNN, PhoBERT.*

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Tên đầy đủ	Dịch nghĩa
AI	Artificial Intelligence	Trí tuệ nhân tạo
CNN	Convolutional Neural Network	Mạng neural tích chập
KNN	K-Nearest Neighbors	Thuật toán học máy
LDA	Latent Dirichlet Allocation	Phân tích chủ đề
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
NLU	Natural Language Understanding	Hiểu ngôn ngữ tự nhiên
NLG	Natural Language Generation	Tạo ngôn ngữ tự nhiên
ML	Machine Learning	Học máy
PhoBERT	Pho Bidirectional Encoder Representations from Transformers	Mô hình ngôn ngữ
SVM	Support Vector Machine	Thuật toán học máy
TF-IDF	Term Frequency-Inverse Document Frequency	Rút trích đặc trưng văn bản
SVD	Singular Value Decomposition	Phân tách giá trị

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

CHƯƠNG I: GIỚI THIỆU

1.1. Đặt vấn đề

Bản tin thời sự truyền hình là một trong những nguồn thông tin quan trọng và phổ biến trong cuộc sống hàng ngày của mọi người. Để cung cấp thông tin đáng tin cậy và phản ánh chính xác về các sự kiện xảy ra trên khắp thế giới, việc phân loại chủ đề tự động cho bản tin truyền hình trở nên vô cùng cấp thiết. Công nghệ hiện đại, đặc biệt là trong lĩnh vực trí tuệ nhân tạo (AI) và xử lý ngôn ngữ tự nhiên (NLP), đã phát triển đáng kể và có khả năng giúp tự động hóa quá trình này.

Trong bối cảnh ngày càng phát triển của truyền thông và truyền hình. Sự gia tăng không ngừng về lượng thông tin được sản xuất và truyền tải qua các phương tiện truyền thông đòi hỏi một cách tiếp cận thông tin hiệu quả. Mô hình phân loại chủ đề tự động có vai trò quan trọng trong việc tổ chức, phân loại và tóm tắt các bản tin thời sự truyền hình, từ đó giúp đảm bảo rằng người xem có thể dễ dàng tiếp cận thông tin một cách nhanh chóng và hiệu quả.

Việc áp dụng trí tuệ nhân tạo vào phân loại nội dung truyền thông là một yếu tố quan trọng đối với sự tiến bộ của ngành truyền thông và truyền hình tại Việt Nam. Trong bối cảnh số hóa và xu hướng tiêu thụ nội dung trực tuyến ngày càng gia tăng, việc tự động phân loại chủ đề trong các bản tin thời sự truyền hình giúp tăng cường khả năng tìm kiếm và phân loại nội dung trên các nền tảng truyền thông kỹ thuật số. Điều này sẽ không chỉ giúp giảm thời gian và công sức trong việc tạo ra nội dung mà còn nâng cao trải nghiệm người xem, đặc biệt là trong một thời đại mà thông tin đang trở nên ngày càng phong phú và đa dạng.

Nghiên cứu và phát triển mô hình phân loại chủ đề tự động trong ngữ cảnh của truyền hình Việt Nam cũng sẽ đóng một vai trò quan trọng trong việc cung cấp giải pháp để theo kịp các xu hướng truyền thông quốc tế. Điều này có thể giúp tạo ra các dịch vụ truyền thông thông minh hơn, như tạo ra gợi ý nội dung cá nhân hóa dựa trên sở thích của người xem và tạo ra các hệ thống quản lý nội dung truyền thông hiệu quả hơn.

Trên thế giới, các nghiên cứu và ứng dụng trong lĩnh vực này đã đạt được sự chú ý đặc biệt. Mô hình phân loại chủ đề tự động không chỉ giúp tối ưu hóa quá trình sản xuất nội dung mà còn cung cấp cơ hội phát triển các dịch vụ truyền thông thông minh, như tạo ra gợi ý nội dung dựa trên sở thích của người xem và cải thiện hệ thống quản lý

nội dung truyền thông. Các nghiên cứu quốc tế cũng đã đánh dấu sự tiến bộ trong việc áp dụng học máy và trí tuệ nhân tạo để xử lý dữ liệu truyền hình và đạt được kết quả đáng kể trong việc phân loại chủ đề tự động.

Do đó, đề tài này không chỉ thúc đẩy sự phát triển trong lĩnh vực nghiên cứu mà còn có tiềm năng ứng dụng rộng rãi trong ngành truyền thông Việt Nam, mà còn đóng góp tích cực vào việc nâng cao chất lượng và hiệu quả của sản phẩm truyền thông trong thời đại số hóa, đồng thời giúp Việt Nam không chỉ tiếp cận một nguồn thông tin đa dạng mà còn có khả năng sản xuất nội dung truyền thông đáng tin cậy và phù hợp với nhu cầu của người xem.

1.2. Công trình nghiên cứu có liên quan

Trong quá trình nghiên cứu, nhận thấy rằng đã tồn tại nhiều nghiên cứu trên thế giới về vấn đề phân loại bản tin truyền hình, với hai hướng chính trong việc giải quyết vấn đề này. Hai hướng chính này bao gồm việc sử dụng lĩnh vực xử lý hình ảnh (Image processing) và xử lý ngôn ngữ tự nhiên (NLP). Trong các nghiên cứu đã tồn tài nhận thấy rằng phần lớn các thực nghiệm và nghiên cứu được tiến hành trên các bản tin thời sự có tính chất quốc tế. Một số nghiên cứu và thực nghiệm đã công bố có liên quan đến vấn đề như:

Nghiên cứu của Li và cộng sự [1] đã đề xuất một phương pháp để phân loại tin tức bằng việc sử dụng phân bổ Latent Dirichlet Allocation (LDA) cùng với thuật toán Softmax Regression. Đề xuất này bắt đầu bằng việc áp dụng phương pháp LDA để phân tích và xác định các chủ đề trong bộ văn bản nghiên cứu. Tiếp theo, quá trình tiến hành xử lý văn bản và trích xuất đặc trưng được thực hiện. Dữ liệu được huấn luyện bằng thuật toán Softmax Regression để phân loại tin tức vào 3 chủ đề dựa trên phân bổ xác suất. Kết quả thử nghiệm trên tập dữ liệu đã thể hiện rằng phương pháp này đạt được hiệu suất phân loại tin tức rất tốt, với chỉ số F1-Score đạt 88%.

Mô hình phát hiện tin tức giả đã được nghiên cứu và xây dựng bởi tác giả Ahmed và cộng sự [2]. Nghiên cứu này kết hợp các phương pháp và mô hình học máy để xác định tính đáng tin cậy của các bài báo và thông tin trên mạng. Cụ thể, phương pháp rút trích đặc trưng TF-IDF kết hợp với các mô hình học máy như Naïve Bayes, Passive Aggressive và SVM đã được áp dụng trong quá trình xây dựng mô hình. Nghiên cứu đã

tiến hành thực nghiệm trên nhiều tập dữ liệu về tin tức giả để đánh giá hiệu suất của mô hình. Kết quả của thực nghiệm đã cho thấy mức độ chính xác đạt 93%.

Nghiên cứu của tác giả Manzato và Goularte [3] tập trung vào việc áp dụng thuật toán Genetic Algorithm (GA). Dữ liệu thực nghiệm trong nghiên cứu này được lấy từ các danh mục video được xác định trước thông qua việc sử dụng công cụ GraphEdit từ Microsoft DirectX SDK2 để phân tách và thu thập phụ đề. Sau khi thu thập dữ liệu, quá trình tiền xử lý được thực hiện. Phương pháp trích xuất đặc trưng Term-Document Matrix được áp dụng, sau đó áp dụng phân rã ma trận Singular Value Decomposition (SVD) nhằm loại bỏ thông tin không quan trọng trong dữ liệu văn bản. Tiếp theo, dữ liệu được huấn luyện trên thuật toán GA. Kết quả thử nghiệm trên tập dữ liệu đã cho thấy phương pháp này đạt được hiệu suất phân loại rất tốt. Chỉ số đánh giá Precision và Recall lần lượt đạt khoảng 89% và 80%.

Nghiên cứu gần đây của tác giả Gao [4], tập trung vào việc phân loại chủ đề bản tin thời sự trong các video sử dụng mô hình ResNet-v2. Phương pháp nghiên cứu dựa trên cơ sở xử lý hình ảnh. Bắt đầu với việc xử lý dữ liệu video bằng cách tách từng khung hình ra từ các đoạn video về các chủ đề thời sự và sau đó đưa chúng qua lớp tích chập của mạng CNN để trích xuất đặc trưng. Sau khi các đặc trưng được trích xuất, mô hình ResNet-v2 được huấn luyện để nhận diện và phân loại chủ đề của bản tin thời sự. Kết quả thực nghiệm trên tập dữ liệu cho thấy mô hình đạt độ chính sát lên đến 91.47%.

Trong nghiên cứu được mô tả trong [5] tập trung vào khám phá mô hình XLNet để thực hiện nhiệm vụ phân loại chương trình truyền hình. Tập dữ liệu bao gồm 1000 chương trình, được phân chia thành 5 chủ đề chính: Ca nhạc, Phim truyện, Thể thao, Thời sự và Tổng hợp. Kết quả thực nghiệm đã chỉ ra rằng mô hình trong nghiên cứu [5] đạt được độ chính xác chủ yếu trên 90% đối với các chủ đề như Phim truyện và Tổng hợp. Tuy nhiên, tác giả cũng nhấn mạnh rằng việc giữ cân bằng giữa các loại dữ liệu huấn luyện đã có ảnh hưởng đáng kể đến độ chính xác của mô hình.

1.3. Mục đích nghiên cứu

Nghiên cứu tổng quan về ngôn ngữ và xử lý ngôn ngữ tự nhiên và một số mô hình phân loại chủ đề tự động cho bản tin thời sự truyền hình tại Việt Nam. Mô hình này sẽ có khả năng tự động phân loại và gắn nhãn các chủ đề quan trọng trong các bản tin thời

sự truyền hình, từ đó giúp tối ưu hóa quá trình sản xuất nội dung và nâng cao hiệu suất của các dịch vụ truyền thông.

Xây dựng bộ dữ liệu thực nghiệm về các chủ đề bản tin thời sự truyền hình Việt Nam phục vụ cho nghiên cứu. Đánh giá hiệu suất của các mô hình phân loại chủ đề qua việc so sánh kết quả phân loại giữa các mô hình. Ứng dụng mô hình phân loại chủ đề thời sự truyền hình vào thực tế.

1.4. Đối tượng và phạm vi nghiên cứu

1.4.1. Đối tượng nghiên cứu

Đối tượng của nghiên cứu chính là bản tin thời sự truyền hình, đặc biệt là các bản tin thời sự truyền hình sản xuất và phát sóng tại Việt Nam. Cụ thể là các đoạn video chứa thông tin về các sự kiện và chủ đề đa dạng trong bản tin thời sự.

1.4.2. Phạm vi nghiên cứu

Nghiên cứu tập trung vào việc phân loại chủ đề tự động cho các bản tin thời sự truyền hình. Nghiên cứu sẽ sử dụng dữ liệu từ các nguồn truyền hình Việt Nam trong một khoảng thời gian cụ thể và áp dụng các kỹ thuật học máy, trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên để phát triển mô hình phân loại chủ đề. Phạm vi địa lý của nghiên cứu giới hạn trong lãnh thổ Việt Nam. Mục tiêu là tạo ra một mô hình phân loại chủ đề tự động hiệu quả và có thể áp dụng rộng rãi trong ngành truyền thông tại Việt Nam, góp phần nâng cao chất lượng và hiệu suất sản phẩm truyền thông trong thời đại số hóa.

1.5. Phương pháp nghiên cứu

Phương pháp nghiên cứu trong đề tài này sẽ tuân theo một chuỗi quy trình khoa học và được thực hiện theo các giai đoạn cụ thể. Đầu tiên sẽ tiến hành thu thập một bộ dữ liệu lớn là các video chứa các bản tin thời sự truyền hình tại Việt Nam trong khoảng thời gian nhất định. Dữ liệu này sẽ bao gồm nhiều nguồn truyền hình và các chủ đề khác nhau, để đảm bảo tính đa dạng và đầy đủ trong quá trình phát triển mô hình. Sau đó sẽ chuyển đổi dữ liệu từ dạng video sang dạng văn bản.

Tiếp theo, sẽ thực hiện quá trình tiền xử lý dữ liệu để làm sạch và chuẩn hóa thông tin từ các bản tin. Quá trình này bao gồm loại bỏ nhiễu, tách từ và câu để sử dụng trong các mô hình học máy.

Sau khi đã có dữ liệu sạch và chuẩn hóa, sẽ tiến hành trích xuất đặc trưng cho dữ liệu. Sau đó sẽ đưa các đặc trưng vào phát triển mô hình phân loại chủ đề tự động bằng cách sử dụng các thuật toán học máy, mô hình học sâu và mô hình ngôn ngữ.

Các mô hình sẽ được đào tạo và kiểm định bằng cách sử dụng các phương pháp đánh giá hiệu suất chính xác và hiệu quả. Tối ưu hóa các tham số của mô hình để đảm bảo tính khả thi và độ tin cậy trong việc phân loại chủ đề.

Cuối cùng, sẽ áp dụng mô hình đã phát triển vào dữ liệu thời sự truyền hình thực tế và đánh giá hiệu suất của nó trong việc phân loại chủ đề tự động. Kết quả sẽ được phân tích và trình bày để đưa ra kết luận và những đề xuất cho ứng dụng thực tiễn trong lĩnh vực truyền thông tại Việt Nam.

1.6. Cấu trúc của đồ án

Đồ án tốt nghiệp được cấu trúc thành 5 phần chính, tập trung vào việc nghiên cứu về cơ sở lý thuyết của NLP, phương pháp trích xuất đặc trưng TF-IDF, cũng như các thuật toán học máy như SVM, KNN, mô hình học sâu như CNN và PhoBERT vào việc phân loại bản tin thời sự trên truyền hình. Đồ án tập trung trình bày về phương pháp thực hiện để giải quyết vấn đề, kết quả của các thực nghiệm thực tế và cuối cùng là việc trình bày kết luận của nghiên cứu.

Bố cục của đồ án cụ thể bao gồm:

Chương I: Giới thiệu

Đề cập đến tình hình hiện tại của các hệ thống phân loại bản tin thời sự truyền hình, cùng việc xem xét các nghiên cứu trước đây về chủ đề này. Cuối cùng đi vào mục tiêu và nội dung chính của đề tài.

Chương II: Cơ sở lý thuyết

Phần thứ nhất của chương, sẽ cung cấp một cái nhìn tổng quan về cơ sở lý thuyết của ngôn ngữ, nhấn mạnh vào chức năng và phạm vi ứng dụng của ngôn ngữ. Tiếp theo, tập trung vào việc tìm hiểu về xử lý ngôn ngữ tự nhiên và quá trình phân loại văn bản. Nội dung này giúp hiểu rõ hơn về lý thuyết cơ bản và quá trình áp dụng công nghệ vào việc xử lý ngôn ngữ trong các ứng dụng thực tế.

Phần thứ hai của chương, sẽ đề cập đến những phương pháp tiếp cận giải quyết vấn đề phân loại bản tin thời sự trên truyền hình. Trong phần này, sẽ khám phá kỹ thuật

trích xuất đặc trưng TF-IDF cùng việc tìm hiểu về các mô hình và thuật toán thông dụng trong NLP cho việc phân loại văn bản.

Phần cuối cùng của chương, sẽ tìm hiểu một số thư viện và các công cụ hỗ trợ được áp dụng trong quá trình thực nghiệm. Các công cụ này bao gồm các thư viện mã nguồn mở, các nền tảng phân tích dữ liệu và các ứng dụng hỗ trợ quá trình huấn luyện mô hình và đánh giá hiệu suất của nó trong việc phân loại bản tin.

Chương III: Phương pháp thực hiện

Trong chương này, trình bày về mô hình nghiên cứu được áp dụng để giải quyết bài toán cụ thể. Đầu tiên, trình bày về quy trình thu thập và xử lý dữ liệu, đồng thời triển khai các phương pháp tiếp cận bài toán đã được nghiên cứu chi tiết trong Chương II. Phần kết của chương sẽ tập trung vào việc đánh giá chất lượng của mô hình thông qua việc áp dụng các tiêu chí đánh giá cho bài toán phân loại.

Chương IV: Kết quả thực nghiệm

Chương này tập trung trình bày kết quả thuật nghiệm của các mô hình đã đề xuất, tập trung vào việc so sánh và đánh giá chất lượng của từng mô hình.

Chương V: Kết luận

Chương này nhằm tổng kết các nhiệm vụ chính đã được hoàn thành và kết quả đạt được. Ngoài việc trình bày kết quả, phần này cũng sẽ chỉ ra các hạn chế còn tồn tại và đề xuất các hướng nghiên cứu và phát triển tiềm năng trong tương lai.

CHƯƠNG II: CƠ SỞ LÝ THUYẾT

2.1. Tổng quan

2.1.1. Ngôn ngữ

2.1.1.1. Khái niệm

Theo nghiên cứu của tác giả Dũng và Hùng [6], họ đã đề xuất khái niệm về ngôn ngữ là một hệ thống dấu hiệu đặc biệt được sử dụng như một phương tiện giao tiếp quan trọng nhất và cũng là phương tiện tư duy của con người.

Trong khi đó, nhận định của tác giả Sáng [7] về khái niệm ngôn ngữ là mô tả những âm thanh được tạo ra bởi con người. Đây là một hệ thống âm thanh liên quan đến ý thức và chứa đựng thông tin. Ngôn ngữ bao gồm các yếu tố như âm thanh, từ vựng và các quy tắc kết hợp chúng mà những người trong cùng một cộng đồng sử dụng như một phương tiện để giao tiếp với nhau.

Qua các nhận định trên có thể hiểu ngôn ngữ được hiểu là bất cứ ngôn ngữ nào được phát sinh mà không trải qua bất cứ một suy nghĩ nào trước đó trong não bộ của con người. Ngôn ngữ tồn tại dưới nhiều trạng thái trong cuộc sống. Đây được coi là ngôn ngữ mà bất kỳ ai cũng có thể tiếp thu và học tập thông qua ngôn ngữ nói để hình thành kiến thức cho bản thân. Việc không tuân thủ theo bất cứ một sự định hướng cũng như hướng dẫn chỉ định từ đầu đã tạo nên những nét riêng biệt khiến ngôn ngữ tự nhiên khác với những ngôn ngữ thông thường.

Ngôn ngữ không chỉ đóng vai trò giao tiếp, mà còn là một phương tiện mạnh mẽ giúp con người tạo ra, truyền đạt và lưu giữ tri thức, văn hóa và lịch sử. Mỗi ngôn ngữ đều mang trong mình một phần không gian tư duy và cách nhìn thế giới độc đáo. Như vậy ngôn ngữ không chỉ giúp trao đổi thông tin một cách hiệu quả mà còn thể hiện sự đa dạng và phong phú của những quan điểm, giá trị và suy nghĩ trong cộng đồng. Điều này tạo ra một sự kết nối sâu sắc giữa con người và môi trường xung quanh.

Việc nghiên cứu và ứng dụng ngôn ngữ trong các lĩnh vực khoa học ngày càng trở nên quan trọng và phổ biến hơn. Ngôn ngữ tự nhiên không chỉ đóng góp vào việc hiểu rõ hơn về tư duy con người mà còn mở ra cánh cửa cho sự phát triển trong các lĩnh vực như trí tuệ nhân tạo, xử lý ngôn ngữ tự nhiên và khoa học dữ liệu.

Thông qua việc nghiên cứu và khai thác ngôn ngữ tự nhiên có thể giải quyết các vấn đề phức tạp, từ dự báo thời tiết, phân tích dữ liệu y tế đến dự đoán xu hướng xã hội và thậm chí là khám phá những khía cạnh mới trong vũ trụ. Các công cụ và kỹ thuật ngôn ngữ học đã và đang định hình cách tương tác với công nghệ và cách hiểu về thế giới xung quanh.

2.1.1.2. Chức năng

Ngôn ngữ [6] được định nghĩa như một hệ thống dấu hiệu đặc biệt với những đặc điểm và chức năng quan trọng.

a) Ngôn ngữ là phương tiện giao tiếp quan trọng nhất của con người

Giao tiếp là hoạt động truyền đạt và trao đổi thông tin. Có nhiều loại phương tiện giao tiếp khác nhau bao gồm giao tiếp ngôn ngữ và giao tiếp phi ngôn ngữ. Tuy con người sử dụng nhiều phương tiện giao tiếp khác nhau như đèn giao thông, cử chỉ và tiếng chuông báo nhưng không có phương tiện nào quan trọng bằng ngôn ngữ bởi vì:

- Phương tiện giao tiếp phổ biến nhất: Ngôn ngữ là cần thiết cho tất cả mọi người và có phạm vi sử dụng không giới hạn.
- Khả năng thể hiện đầy đủ và chính xác: Ngôn ngữ có khả năng truyền đạt tất cả tư tưởng, tình cảm và cảm xúc của con người một cách rõ ràng và chính xác. Các phương tiện giao tiếp khác có thể biểu đạt những tình cảm này một cách độc đáo, sâu sắc và tinh tế, nhưng thường hạn chế về phạm vi sử dụng.

Chức năng giao tiếp của ngôn ngữ bao gồm truyền thông tin, yêu cầu hành động, bộc lộ cảm xúc, xác lập và duy trì quan hệ xã hội. Nó không chỉ áp dụng trong cùng một thế hệ mà còn giữ vai trò quan trọng trong truyền tải thông điệp qua các thế hệ khác nhau. Ngôn ngữ đóng vai trò quan trọng trong việc con người truyền đạt thông điệp cho các thế hệ tương lai.

b) Ngôn ngữ là phương tiện tư duy

Ngôn ngữ không chỉ đóng vai trò là phương tiện giao tiếp mà còn là một phương tiện tư duy quan trọng. Điều này có nghĩa rằng ngôn ngữ cho phép con người thực hiện các hoạt động tư duy. Ngôn ngữ không chỉ được sử dụng khi cần truyền đạt ý tưởng, tình cảm và cảm xúc cho người khác trong quá trình giao tiếp. Ngôn ngữ còn xuất hiện khi con người nói một mình hoặc thậm chí trong quá trình suy nghĩ riêng tư, mà không

cần phải thể hiện ra ngoài bằng từ ngữ. Khái niệm, phán đoán và suy luận, tức là những hình thức cơ bản của tư duy, thường tồn tại dưới dạng biểu đạt bằng ngôn ngữ. Một cách ngược lại, nếu thiếu tư duy ngôn ngữ trở thành những âm thanh không mang ý nghĩa, vô nghĩa. Ngôn ngữ và tư duy có sự thống nhất mạnh mẽ không thể tách rời lẫn nhau.

Mặc dù ngôn ngữ và tư duy thường làm việc cùng nhau nhưng chúng không đồng nhất. Ngôn ngữ là phương tiện biểu đạt, trong khi tư duy là nội dung được biểu đạt. Mỗi ngôn ngữ mang theo những đặc trưng riêng không thể thấy ở các ngôn ngữ khác. Trong khi đó tư duy ở mức cơ bản không thể hiện sự khác biệt quá lớn giữa các dân tộc.

Ngôn ngữ không chỉ được sử dụng làm phương tiện giao tiếp vì nó không chỉ đơn giản là một loạt âm thanh mà còn là một loạt âm thanh biểu thị tư tưởng của con người, là kết quả của hoạt động tư duy. Vì vậy, chức năng của ngôn ngữ trong giao tiếp mật thám liên quan chặt chẽ đến chức năng của nó trong việc thể hiện hoạt động tư duy.

2.1.1.3. Các lĩnh vực nghiên cứu

Trong nghiên cứu [6] cho thấy rằng lĩnh vực nghiên cứu của ngôn ngữ gồm 3 khía cạnh chính bao gồm: ngữ âm học, ngữ pháp học và ngữ nghĩa học.

a) Ngữ âm học

Ngữ âm học là lĩnh vực nghiên cứu chuyên sâu về các khía cạnh âm thanh trong ngôn ngữ. Nó không chỉ tập trung vào việc phân tích và mô tả các âm thanh ngôn ngữ mà còn khám phá mối quan hệ phức tạp giữa ngữ viết và âm thanh của ngôn ngữ.

Khi nghiên cứu về ngữ âm trong lĩnh vực ngôn ngữ tự nhiên, nghiên cứu này thường quan tâm đến cấu trúc âm thanh và các yếu tố âm học, xem xét ngữ âm như một khía cạnh vật lý hoặc chất liệu trong ngôn ngữ. Tuy nhiên, không nên xem xét ngữ âm chỉ từ góc độ vật lý mà còn phải nhìn vào khía cạnh chức năng quan trọng của chúng, bởi vì ngữ âm mang theo ý nghĩa và tham gia vào việc truyền đạt thông tin và giao tiếp trong xã hội con người.

Khía cạnh xã hội của ngữ âm có thể thấy qua các điểm sau:

- Mỗi xã hội và mỗi dân tộc sử dụng một ngôn ngữ cụ thể với hệ thống ngữ âm riêng biệt. Một số ngữ âm có thể trở nên quen thuộc và tự nhiên đối với một xã hội nhưng chúng có thể gây khó khăn và không rõ ràng đối với xã hội khác.

- Mỗi khoa học ngôn ngữ có phương pháp tiếp cận và xử lý âm thanh theo hướng riêng biệt. Ví dụ trong lĩnh vực ngôn ngữ học, khi quan sát ngôn ngữ tiếng Việt, thấy được sự phân biệt giữa hai âm thanh o và ô, trong khi một số ngôn ngữ khác không thể hiện sự tách biệt này. Đồng thời, tiếng Việt cũng đặt ra sự phân biệt giữa âm thanh t và th, trong khi tiếng Anh không có sự phân biệt này.

b) *Ngữ pháp học*

Ngữ pháp học là một lĩnh vực nghiên cứu về ngôn ngữ, tập trung vào việc nghiên cứu sâu về hình thái của từ, cùng với các quy tắc và nguyên tắc cấu tạo từ và câu. Hệ thống này bao gồm những quy tắc cụ thể về cách từ được hình thành, biến hình và cách câu được cấu tạo. Ngữ pháp học gồm có hai phân ngành hẹp hơn là: hình thái học và cú pháp học.

- Hình thái học tập trung vào việc nghiên cứu cấu trúc và tính hình thái của từ, bao gồm sự phân tích và hiểu về cách từ được hình thành, cấu trúc từ và việc xác định từ loại. Đặc biệt, trong các ngôn ngữ điển hình như tiếng Anh, tiếng Pháp, tiếng Nga,… việc nghiên cứu hình thái từ đóng một vai trò quan trọng. Tuy nhiên, trong trường hợp của các ngôn ngữ không điển hình như tiếng Việt, tiếng Hán và nhiều ngôn ngữ khác, lĩnh vực này không bao gồm nghiên cứu về hình thái từ và được thường được gọi là từ pháp học. Vấn đề cấu tạo từ cũng là một phần của nghiên cứu từ vựng học và có thể coi như là một phạm trù trung gian giữa từ vựng và ngữ pháp. Từ loại mặc dù có thể được định nghĩa bởi môn từ loại học, tồn tại độc lập và không thay đổi dựa trên hình thái và cú pháp.

- Cú pháp học tập trung vào nghiên cứu cấu trúc của câu, bao gồm quy tắc cấu trúc câu và ngữ đoạn. Sự phân biệt giữa hình thái học và cú pháp học thường có tính chất ước định và phụ thuộc vào loại hình ngôn ngữ. Trong ngôn ngữ biến hình sự phân biệt này có thể rõ ràng hơn, trong khi trong ngôn ngữ đơn lập như tiếng Việt và tiếng Hán sự phân biệt giữa hình thái học và cú pháp học không được thực hiện và có thể coi môn học này là ngữ pháp học chung. Ngay cả trong các ngôn ngữ biến hình, hình thái học và cú pháp học có thể có những sự liên quan với nhau.

c) Ngữ nghĩa học

Ngữ nghĩa học là lĩnh vực nghiên cứu chuyên sâu về ý nghĩa của các biểu thức ngôn ngữ, bất kể liệu chúng đứng một mình hoặc được liên kết với bối cảnh cụ thể. Một cách tổng quan, ý nghĩa của một biểu thức ngôn ngữ đại diện cho nội dung tinh thần mà nó truyền đạt.

Nghĩa của từ hoặc biểu thức ngôn ngữ có thể được xem xét từ nhiều khía cạnh khác nhau. Ngôn ngữ đóng vai trò quan trọng trong việc giao tiếp giữa con người, trong việc miêu tả cả những sự vật cụ thể và những khái niệm trừu tượng, cũng như trong việc thể hiện suy nghĩ và tình cảm cá nhân. Ngôn ngữ được coi là một hệ thống dấu hiệu mà trong đó ý nghĩa của từ hay biểu thức ngôn ngữ tồn tại trong ngữ cảnh và quan hệ với bối cảnh giao tiếp. Khía cạnh này liên quan đến nội dung thông tin truyền đạt.

Không thể tách rời ý nghĩa từ ngữ cảnh cụ thể trong đó chúng được sử dụng. Một câu có thể hiểu khác nhau tùy thuộc vào ngữ cảnh.

Ngoài ra, ý nghĩa của từ hay biểu thức ngôn ngữ là một hiện tượng tinh thần, tức là nó tồn tại trong trí óc của người nghe hoặc người nói. Ý nghĩa khi xem xét từ góc độ này trở thành một vấn đề về tri nhận và quan điểm cá nhân.

Cuối cùng, có thể khám phá ý nghĩa của từ hoặc câu trong ngữ cảnh của chúng trong mối quan hệ với nhau. Từ góc độ này, có thể xác định mối quan hệ giữa các từ và biểu thức chẳng hạn như việc nói rằng hai từ là trái nghĩa hoặc đồng nghĩa.

2.1.2. Xử lý ngôn ngữ tự nhiên

2.1.2.1. Khái niệm

Theo nghiên cứu của tác giả Eisenstein [8]. Khái niệm này có thể được hiểu như việc tổng hợp các phương pháp để đưa ngôn ngữ con người vào khả năng truy cập của máy tính.

Bên cạnh đó, khái niệm này hoàn toàn tương đồng với kết luận của tác giả Geitgey [9] cho rằng xử lý ngôn ngữ tự nhiên đóng vai trò quan trọng trong lĩnh vực trí tuệ nhân tạo, trong đó máy tính được tập trung vào việc hiểu và xử lý ngôn ngữ con người.

Từ các quan điểm trên, có thể thấy rõ mục tiêu chính của NLP là phát triển các giải pháp và công nghệ để hỗ trợ máy tính trong việc hiểu và thực hiện hiệu quả các nhiệm vụ liên quan đến ngôn ngữ con người.

2.1.2.2. Lĩnh vực

Theo nghiên cứu của tác giả Jurafsky và Martin [10] đã đưa ra nhận định rằng NLP có thể được phân chia thành hai lĩnh vực chính mặc dù chúng không hoàn toàn độc lập. Hai lĩnh vực chính trong NLP được nhận định là: xử lý tiếng nói và xử lý văn bản.

a) Xử lý tiếng nói

Tập trung vào việc nghiên cứu và phát triển các thuật toán và chương trình máy tính nhằm xử lý và hiểu ngôn ngữ con người khi được diễn đạt dưới dạng tiếng nói. Lĩnh vực này tập trung vào việc nghiên cứu các phương pháp chuyển đổi tín hiệu âm thanh thành ngôn ngữ, cũng như phát triển các công nghệ nhằm nhận diện và hiểu ngữ cảnh, ngữ điệu và ý nghĩa được truyền đạt thông qua tiếng nói.

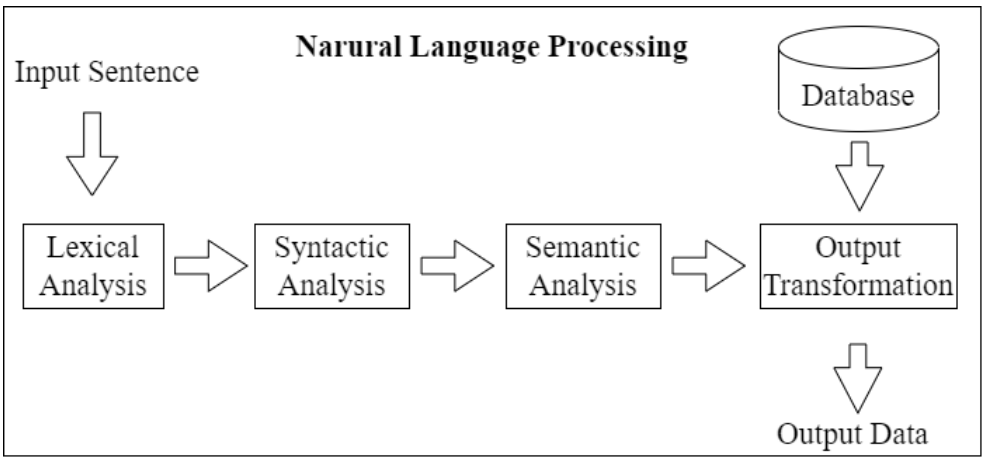
b) Xử lý văn bản

Tập trung vào việc phân tích và nghiên cứu phát triển các thuật toán và chương trình máy tính để hiểu và xử lý dữ liệu văn bản. Lĩnh vực này tập trung vào việc áp dụng các phương pháp phân tích ngôn ngữ tự nhiên, xử lý ngôn ngữ tự nhiên và học máy để hiểu và rút trích thông tin từ các tài liệu văn bản, bao gồm cả việc phân tích cú pháp, nhận diện thực thể có tên và tóm tắt nội dung.

Hai lĩnh vực này cùng hướng tới mục tiêu là hiểu và xử lý ngôn ngữ con người, mặc dù từ góc độ khác nhau. Qua việc tập trung vào những khía cạnh khác nhau của ngôn ngữ, NLP đã liên tục phát triển các phương pháp và công nghệ để hiểu và tương tác với ngôn ngữ con người một cách hiệu quả và tự nhiên hơn.

2.1.2.3. Quy trình

Tapsai và các cộng sự [11] đã đưa ra nhận định rằng lĩnh vực NLP đã trải qua hơn 40 năm phát triển, được khám phá trong nhiều nghiên cứu nhằm tạo ra nền tảng cho việc tương tác với máy tính thông qua sử dụng nhiều phương pháp và kỹ thuật khác nhau. Các quy trình chính trong lĩnh vực NLP thường được phân thành 4 bước cơ bản được minh họa chi tiết trong Hình 2.1.



Hình 2. 1: Quy trình của xử lý ngôn ngữ tự nhiên [11]

Quy trình của xử lý ngôn ngữ tự nhiên bao gồm các bước cụ thể bao gồm:

a) Phân tích từ vựng (Lexical Analysis)

Giai đoạn này tập trung vào quá trình phân tách một câu thành các từ hoặc các đơn vị nhỏ, các từ được tách thường được gọi là các *token*, nhằm xác định ý nghĩa của từ hoặc đơn vị này và quan hệ của nó với toàn bộ câu.

b) Phân tích cú pháp (Syntactic Analysis)

Bước này tập trung vào việc xác định mối quan hệ giữa các từ và cụm từ khác nhau trong một câu, chuẩn hóa cấu trúc của chúng và biểu thị các mối quan hệ theo cấu trúc phân cấp để hiểu cú pháp của câu.

c) Phân tích ngữ nghĩa (Semantic Analysis)

Giai đoạn này tập trung vào việc kết nối cấu trúc cú pháp từ cấp độ của các cụm từ, mệnh đề, câu và đoạn văn đến cấp độ của toàn bộ văn bản với ý nghĩa độc lập với ngôn ngữ cụ thể của chúng.

d) Chuyển đổi đầu ra (Output Transformation)

Giai đoạn này tạo ra một đầu ra dựa trên phân tích ngữ nghĩa của văn bản hoặc giọng nói, phù hợp với mục tiêu của ứng dụng. Điều này có thể là bản dịch, việc sửa lỗi ngữ pháp hoặc tạo ra phản hồi dựa trên quy tắc hoặc dữ liệu huấn luyện tùy thuộc vào mục đích cụ thể của ứng dụng NLP.

2.1.2.4. Nhiệm vụ

Trong nghiên cứu của tác giả Jain và Chadha [12] đã nhấn mạnh rằng trong lĩnh vực NLP có nhiều vấn đề thực tế cần được giải quyết. Dưới đây là một số vấn đề chính đã được nêu ra:

a) Phân tích cảm xúc (Sentiment Analysis)

Phân tích cảm xúc hay còn gọi là khai phá quan điểm, nhằm khám phá thái độ của người nói, người viết hoặc văn bản đối với một chủ đề cụ thể, sản phẩm hoặc tổng thể về ngữ cảnh của một tài liệu.

Việc phân tích quan điểm có thể thực hiện ở ba mức độ khác nhau: mức độ tài liệu, mức độ câu, mức độ thực thể và khía cạnh. Phương pháp này đang được sử dụng rộng rãi trong nhiều lĩnh vực như quản lý thương hiệu, phân tích sản phẩm và việc hiểu rõ tâm trạng của khách hàng.

b) Tóm tắt văn bản (Text Summarization)

Tóm tắt văn bản là quá trình tạo ra một tóm tắt ngắn gọn từ văn bản dài, vẫn bảo toàn thông tin chính và ý nghĩa tổng thể. Quá trình này chia thành hai loại chính là: tóm tắt trích xuất và tóm tắt sáng tạo.

- Tóm tắt trích xuất liên quan đến việc rút trích các cụm từ và câu quan trọng từ văn bản gốc để tạo ra một bản tóm tắt, tóm gọn chỉ một phần nhỏ của nội dung ban đầu.
- Tóm tắt sáng tạo, ngược lại tạo ra một bản tóm tắt mới, có thể sử dụng từ và cụm từ không có trong văn bản gốc, tương tự như cách con người tóm tắt một tài liệu.

c) Phân loại văn bản (Text Classification)

Phân loại văn bản là quá trình phân loại các văn bản vào các danh mục đã cho hoặc phát hiện các chủ đề mới trong văn bản. Bằng việc phân tích dữ liệu văn bản và dự đoán lớp của nó, việc phân loại hiệu quả và hiệu suất có thể được đạt được giúp thông tin trở nên dễ quản lý và dễ hiểu hơn.

d) Hỏi đáp (Question Answering)

Hỏi đáp là một hệ thống có khả năng trả lời một câu hỏi được đặt ra bởi con người một cách chính xác. Nhiệm vụ này không chỉ giải quyết các câu hỏi đơn giản đến những câu hỏi phức tạp yêu cầu quá trình suy luận và hiểu biết bối cảnh.

Mục tiêu của một hệ thống hỏi đáp là cung cấp những câu trả lời chính xác, ngắn gọn và liên quan đến các truy vấn của người dùng. Phát triển các hệ thống này đòi hỏi sự hiểu biết sâu sắc về cả ngôn ngữ tự nhiên và quá trình tạo ra ngôn ngữ tự nhiên, tạo nên một nhiệm vụ khó khăn nhưng có tác động to lớn trong lĩnh vực NLP.

e) Nhận dạng thực thể có tên (Named Entity Recognition)

Việc xác định thực thể là một bước quan trọng trong quá trình trích xuất thông tin từ văn bản, mục tiêu là xác định và phân loại các thực thể được đặt tên vào các danh mục định trước. Các danh mục này bao gồm tên của cá nhân, tổ chức, địa điểm, sự kiện, số lượng,... Các thực thể thường được xác định là từ hoặc chuỗi từ liên tục thường xuất hiện để chỉ đến cùng một loại thực thể. Mỗi thực thể được xác định và phân loại vào một danh mục được xác định trước. Nhiệm vụ này thường được thực hiện thông qua việc phân tích ngữ cảnh của từ và từ láng giềng trong văn bản.

f) Phân tích cú pháp (Dependency Parsing)

Phân tích cú pháp là một nhiệm vụ nhằm trích xuất cấu trúc ngữ pháp của câu, xác định mối quan hệ giữa các từ trọng tâm và các từ sửa đổi những từ trọng tâm đó. Đây thực chất là quá trình xem xét các phụ thuộc giữa các cụm từ trong một câu để xác định cấu trúc ngữ pháp của nó.

Mỗi câu được phân tích thành các phần dựa trên sự tương phụ thuộc giữa các đơn vị ngôn ngữ trong câu. Quá trình này dựa trên giả thuyết về mối quan hệ trực tiếp giữa mỗi đơn vị trong câu và các mối quan hệ này được gọi là phụ thuộc. Phân tích cú pháp bao gồm việc xác định các mối quan hệ phụ thuộc của mỗi từ, xác định từ nào phụ thuộc vào từ nào.

2.1.2.5. Thành phần

Nghiên cứu của tác giả Kaur và Singh [13] đã đưa ra kết luận rằng NLP (Natural Language Processing) được hình thành từ 2 thành phần là NLU (Natural Language Understanding) và NLG (Natural Language Generation) .

- NLU đảm nhận trách nhiệm về việc hiểu dữ liệu đầu vào từ người dùng dựa trên cú pháp ngôn ngữ, nội dung được xác định dựa trên ý định và thực thể mà dữ liệu được diễn đạt. Nó chịu trách nhiệm cho việc phân tích ngôn ngữ từ các khía cạnh khác nhau và ánh xạ dữ liệu đầu vào cho các ngôn ngữ tự nhiên.

- Một khía cạnh khác, NLG thao tác với việc tạo ra văn bản từ dữ liệu có cấu trúc. NLG chỉ phân tích dữ liệu và diễn đạt nó thành ngôn ngữ giao tiếp. NLG không chỉ giới hạn ở việc phân tích dữ liệu mà còn tập trung vào quá trình tạo ra văn bản một cách logic và mạch lạc từ dữ liệu đã được xử lý. Mục tiêu chính của NLG là biến dữ liệu thành ngôn ngữ tự nhiên, tạo ra văn bản trôi chảy, dễ hiểu và có tính ứng dụng cao.

2.1.2.6. Thánh thức

Lĩnh vực NLP phải đối mặt với nhiều thách thức lớn do tính phức tạp và không nhất quán của ngôn ngữ con người, được nhận định bởi các tác giả Dilmegani [14] và Roldós [15]. Những thách thức quan trọng này làm cho phát triển và triển khai các hệ thống NLP trở nên khó khăn và đầy thách thức. Một số thách thức đáng kể mà NLP phải đối mặt, bao gồm:

a) Ngôn ngữ khiếm nhã

Ngôn ngữ khiếm nhã tạo ra thách thức đối với các mô hình học máy vì chúng thường sử dụng các từ và cụm từ có ý nghĩa tích cực hoặc tiêu cực theo định nghĩa, trong khi thực tế có thể mang ý nghĩa ngược lại, điều này đòi hỏi khả năng phân biệt ngôn ngữ nghệ thuật và ngôn ngữ tường thuật.

b) Từ đa nghĩa

Một từ hoặc cụm từ có thể có nhiều ý nghĩa khác nhau tùy thuộc vào bối cảnh của câu. Điều này gây khó khăn trong việc hiểu đúng ngữ cảnh và ý nghĩa của từng từ.

c) Tiếng lóng hoặc tiếng địa phương

Tiếng lóng hoặc tiếng địa phương thường không tuân theo các quy tắc ngữ pháp chuẩn và thay đổi liên tục. Điều này khiến việc hiểu và xử lý các biến thể ngôn ngữ trở nên phức tạp.

d) Ngôn ngữ chuyên ngành

Ngôn ngữ chuyên ngành sử dụng các thuật ngữ và cách diễn đạt riêng biệt chỉ được hiểu bởi những người trong cùng lĩnh vực. Điều này đòi hỏi khả năng xử lý và hiểu sâu về kiến thức chuyên ngành.

e) Thiên vị trong dữ liệu huấn luyện

Sự thiên vị trong dữ liệu huấn luyện có thể dẫn đến các mô hình học máy có kết quả thiên vị hoặc không chính xác, khiến cho việc tổng quát hóa đối với các tình huống thực tế trở nên khó khăn.

2.1.3. Phân loại văn bản

2.1.3.1. Khái niệm

Theo như kết quả nghiên cứu của tác giả Hồng [16] đã đưa ra nhận định rằng phân loại văn bản là một nhiệm vụ quan trọng trong lĩnh vực NLP. Mục tiêu là sắp xếp các tài liệu văn bản vào một hoặc nhiều phân loại hoặc lớp đã xác định trước. Mục đích chính của quá trình này là xác định xem một văn bản cụ thể thuộc về lớp ngữ nghĩa cụ thể nào mà đã được định rõ trước đó.

Nhiệm vụ của phân loại văn bản là xếp các tài liệu văn bản vào các phân loại với mục tiêu thu thập thông tin nhanh hơn và cung cấp lĩnh vực cụ thể để nghiên cứu sâu hơn các tài liệu tương tự. Trong quá khứ, các hệ thống thu thập thông tin thường sử dụng biểu đồ phân loại truyền thống. Tuy nhiên, hầu hết các giải thuật phân nhóm ngày nay sử dụng mô hình không gian vector để biểu diễn các tài liệu văn bản.

Theo nghiên cứu của tác giả Huy [17] đã đưa ra nhận định rằng mục tiêu của bài toán phân loại chủ đề văn bản có thể được hình dung dưới góc nhìn toán học như công thức (2.1).

$$f(DC): \mathrm{D} \rightarrow \mathrm{C} \qquad (2.1)$$

Trong đó:

- D là tập hợp các văn bản, gồm n văn bản: D = {d1, d2, ..., dn}
- C là tập hợp các chủ đề hoặc các nhãn chủ đề, gồm m chủ đề: C = {c1, c2,...,cm}
- *f(DC)* là hàm ánh xạ từ tập văn bản D vào tập chủ đề C, đại diện cho quá trình phân loại hoặc nhận dạng chủ đề của các văn bản

Mục tiêu của bài toán nhận dạng văn bản là xây dựng và tối ưu hàm *f(DC)* sao cho khi đưa vào một văn bản mới, hàm này có khả năng dự đoán và gán một chủ đề tương ứng từ tập C cho văn bản đó. Để làm được điều này, hàm *f(DC)* cần được học từ dữ liệu huấn luyện, trong đó mỗi văn bản đã được gán nhãn chủ đề đúng.

2.1.3.2. Lĩnh vực

Trong nghiên cứu [17] đã cho thấy rằng phân loại văn bản thường được thực hiện dưới 2 lĩnh vực chính: phân loại theo chủ đề và phân loại theo ngữ nghĩa.

a) Phân loại theo chủ đề

Dựa vào chủ đề mà văn bản có thể thuộc vào. Tập văn bản được phân thành các chủ đề khác nhau như: Giáo dục, Thể Thao, Du Lịch, Sức Khoẻ và nhiều chủ đề khác.

b) Phân loại theo ngữ nghĩa

Dựa vào ngữ nghĩa của văn bản để phân loại chúng. Ví dụ về các ứng dụng như: Phân tích cảm xúc, Xác định spam hoặc không spam, Đề xuất hoặc không đề xuất và nhiều ứng dụng khác.

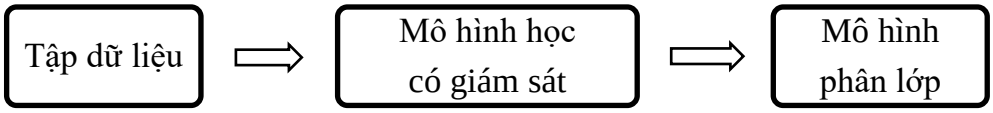
2.1.3.3. Mô hình tổng quát

Theo kết quả nghiên cứu của tác giả Hòa và Lăng [18] từng đề cập rằng mô hình phân loại văn bản dựa trên phương pháp thống kê và học có giám sát được mô tả bao gồm hai giai đoạn quan trọng:

a) Giai đoạn huấn luyện

Trong giai đoạn này, quá trình bắt đầu với việc sử dụng một tập dữ liệu huấn luyện, trong đó mỗi mục trong tập này được gán vào một hoặc nhiều lớp dựa trên các đặc điểm ngữ nghĩa. Mục tiêu chính là biểu diễn mỗi mục trong tập huấn luyện bằng một mô hình số hóa thường là một vector, để mô tả văn bản tương ứng trong tập huấn luyện. Mô hình số hóa này giúp biểu diễn ngữ nghĩa và đặc điểm của các văn bản một cách số học để máy tính có thể hiểu và xử lý chúng.

Giai đoạn huấn luyện trong nhận dạng văn bản được tổng quát hóa như Hình 2.2.



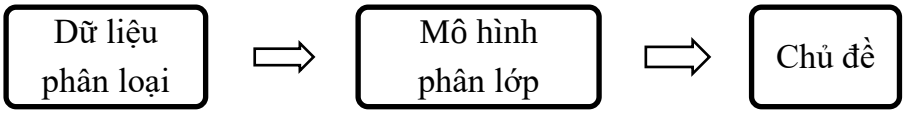
Hình 2. 2: Sơ đồ giai đoạn huấn luyện phân loại văn bản [18]

- **Đầu vào**: Dữ liệu huấn luyện gồm các văn bản và mô hình học có giám sát
- **Đầu ra**: Mô hình phân lớp được tạo ra trong quá trình huấn luyện

b) Giai đoạn phân loại

Sau khi hoàn thành giai đoạn huấn luyện và có mô hình phân lớp đã được xây dựng, giai đoạn phân loại sẽ được tiến hành để phân loại các văn bản mới dựa trên kiến thức đã học từ tập huấn luyện.

Giai đoạn phân loại bao gồm các bước như Hình 2.3.



Hình 2. 3: Sơ đồ giai đoạn phân loại văn bản [18]

- **Đầu vào**: Dữ liệu cần phân loại và mô hình phân lớp đã được tạo trong giai đoạn huấn luyện
- **Đầu ra**: Kết quả là chủ đề hoặc lớp mà các văn bản mới thuộc về sau khi được xử lý bởi mô hình phân loại

Thông qua quá trình này, mô hình phân lớp đã được học từ tập huấn luyện được áp dụng để dự đoán và gán chủ đề cho các văn bản mới, đóng vai trò quan trọng trong việc tự động phân loại và quản lý thông tin trong nhiều ứng dụng thực tế.

2.2. Các phương pháp tiếp cận bài toán

2.2.1. Đặc trưng TF-IDF

Dựa trên nghiên cứu của tác giả Nam [19] đã đưa ra nhận định rằng phương pháp trích đặc trưng TF-IDF là một phương pháp thống kê được biết đến để xác định mức độ quan trọng của một từ trong một văn bản so với một tập hợp các văn bản khác nhau.

TF-IDF được sử dụng để xác định trọng số của một từ trong văn bản, thể hiện mức độ quan trọng của một từ trong bản thân văn bản và so với các văn bản trong tập dữ liệu. Giá trị TF-IDF càng cao cho thấy độ quan trọng cao của từ trong văn bản, tuy nhiên, giá trị này cũng phụ thuộc vào số lần từ xuất hiện trong văn bản cùng với tần suất xuất hiện của từ trong toàn bộ tập dữ liệu.

a) TF (Term Frequency)

Đây là tần suất xuất hiện của một từ trong văn bản. Vì độ dài của các văn bản có thể khác nhau, nên việc tính TF thường được chuẩn hóa bằng cách chia tần suất xuất hiện của từ cho tổng số từ trong văn bản. Giá trị TF được xác định trong công thức (2.2).

$$TF(t,d) = \frac{f(t,d)}{\text{sum}\{f(\text{w,d}) \text{ w}\in\text{d}\}} \quad (2.2)$$

Trong đó:

- *TF(t,d):* là tần suất xuất hiện của từ t trong văn bản d
- *f(t,d):* là số lần xuất hiện của từ t trong văn bản d
- sum{*f*(w,d): w ∈ d}: Tổng số từ xuất hiện trong văn bản d

b) IDF (Inverse Document Frequency)

Đây là nghịch đảo của tần suất xuất hiện của một từ trong các văn bản của tập dữ liệu. IDF giúp đánh giá tầm quan trọng của từ. Một số từ có thể xuất hiện nhiều lần trong các văn bản nhưng không quan trọng để thể hiện ý nghĩa của văn bản, do đó cần giảm mức độ quan trọng của những từ này bằng cách sử dụng IDF bằng công thức (2.3).

$$IDF(t,D) = log\frac{|D|}{|\{d \in D: t \in d\}|} \quad (2.3)$$

Trong đó:

- $IDF(t,D)$: là giá trị IDF của từ t trong tập văn bản D
- $|D|$: Tổng số văn bản trong tập D
- $|\{d \in D: t \in d\}|$: Số tài liệu chứa từ t trong tập d

c) TF-IDF (Term Frequency-Inverse Document Frequency)

Công thức tính giá trị TF-IDF của một từ trong một văn bản được xác định như công thức (2.4).

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D) \quad (2.4)$$

Trong đó:

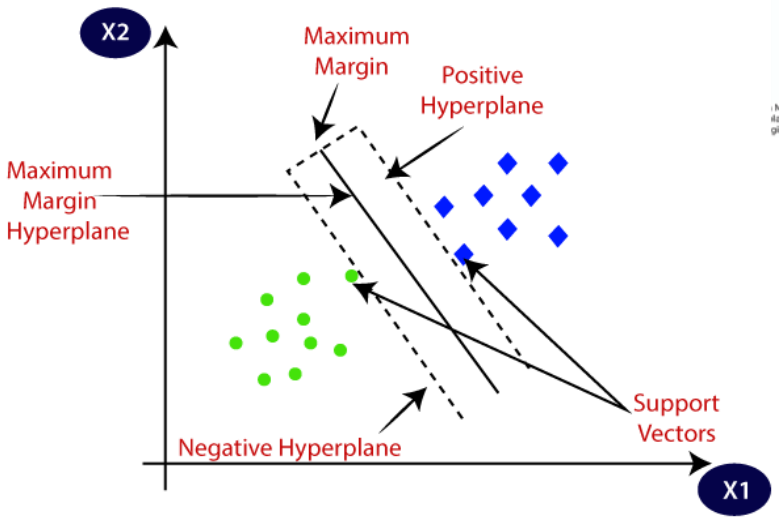
- $TF(t,d)$: là tần suất xuất hiện của một từ trong văn bản
- $IDF(t,D)$: là nghịch đảo của tần suất xuất hiện của một từ trong các văn bản của tập dữ liệu.

Các từ trong văn bản không chỉ có mức độ quan trọng khác nhau đối với văn bản mà còn quan trọng trong việc phân loại văn bản. Các từ có giá trị TF-IDF cao thường là những từ xuất hiện nhiều trong văn bản đó và xuất hiện ít trong các văn bản khác. Phương pháp TF-IDF giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao, chính là các từ khóa quan trọng của văn bản đó.

2.2.2. Thuật toán Support Vector Machine

Thuật toán Support Vector Machine (SVM) là một phương pháp học có giám sát dựa trên lý thuyết học thống kê do Vapnik [20] phát triển được sử dụng phổ biến trong nhiều lĩnh vực, đặc biệt là lĩnh vực phân loại và nhận dạng.

Mục tiêu chính của SVM là xây dựng một siêu phẳng (hyperplane) tốt nhất để phân tách các lớp dữ liệu khác nhau trong không gian đặc trưng. Siêu phẳng này được xác định bằng cách tối ưu hóa khoảng cách lớn nhất (margin) giữa siêu phẳng và các điểm dữ liệu gần nhất của các lớp khác nhau. Khoảng cách này thường được gọi là lề (margin). Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các vector hỗ trợ (Support Vector).



Hình 2. 4: Thuật toán SVM [21]

SVM đặt ra một nhiệm vụ quan trọng là tạo ra một ranh giới quyết định sao cho khoảng cách từ các điểm dữ liệu đến siêu phẳng quyết định là lớn nhất có thể. Điều này đảm bảo tính phân loại chính xác cho các dữ liệu mới mà chưa từng được thấy trong quá trình huấn luyện.

SVM có khả năng giải quyết cả các bài toán phân loại tuyến tính và phi tuyến bằng cách ánh xạ dữ liệu vào một không gian chiều cao hơn, sử dụng hàm nhân (Kernel). Một số hàm nhân phổ biến được sử dụng để biến đổi dữ liệu được thể hiện trong Bảng 2.1

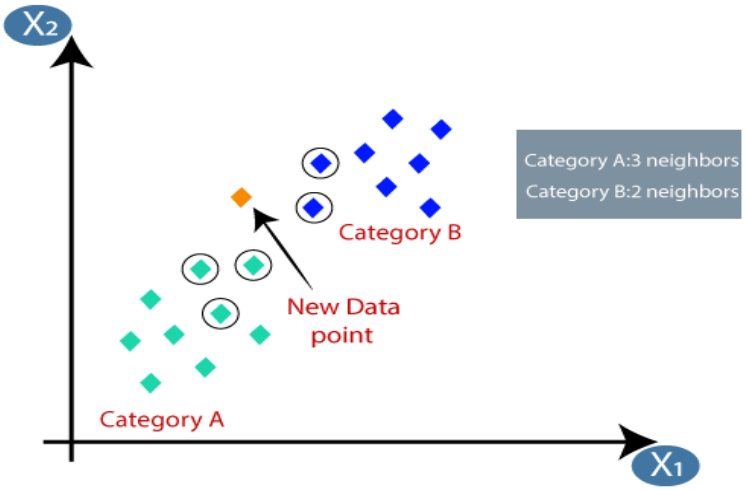
Bảng 2. 1: Một số hàm nhân được sử dụng trong SVM

Kernel	**Công thức**
Linear	$K(x, y) = x * y$
Polynomial	$K(x, y) = (\alpha * x * y + c)^d$
Radial Basis Function (RBF)	$K(x, y) = \exp(-\gamma * \|x - y\|^2)$
Sigmoid	$K(x, y) = \tanh(\alpha * x * y + c)$

Trong nghiên cứu này, không đi sâu vào cơ sở lý thuyết toán học chi tiết của SVM. Mục tiêu chính của nghiên cứu là áp dụng SVM để tập trung vào việc thực hiện các thực nghiệm. Cơ sở lý thuyết toán học được thể hiện qua nghiên cứu [20].

2.2.3. Thuật toán K-Nearest Neighbors

Theo nhận định của tác giả Tiệp [22], thuật toán K-nearest neighbor (KNN) là một thuật toán trong lĩnh vực học máy có giám sát. Trong quá trình huấn luyện, thuật toán KNN không thực hiện bất kỳ quá trình học nào từ dữ liệu huấn luyện, thay vào đó nó lưu trữ toàn bộ dữ liệu huấn luyện và hoạt động dựa trên dữ liệu này khi cần dự đoán đầu ra cho các mẫu dữ liệu mới. Điều này dẫn đến việc KNN thuộc loại thuật toán học đơn giản nghĩa là không có bất kỳ tính toán nào được thực hiện trước khi cần dự đoán đầu ra cho dữ liệu mới.



Hình 2. 5: Thuật toán KNN [23]

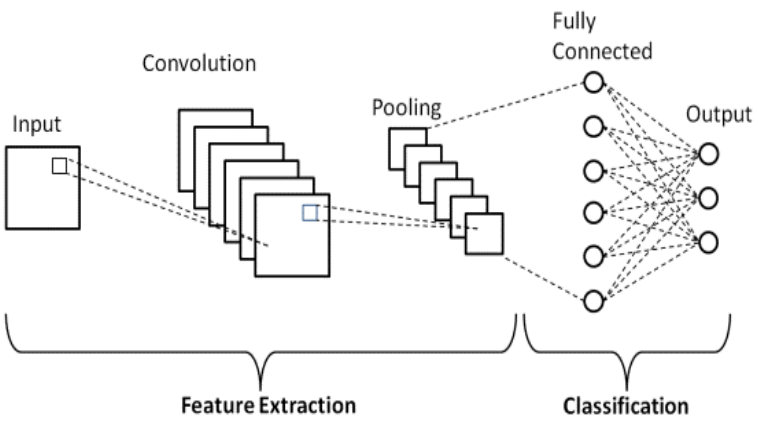
Khi có một điểm dữ liệu mới cần dự đoán đầu ra. KNN sẽ thực hiện quá trình tìm kiếm trong bộ dữ liệu huấn luyện để tìm ra K điểm gần nhất với điểm dữ liệu mới này, với K là một tham số mà người dùng xác định trước. Sau khi tìm được các điểm gần nhất, KNN sẽ sử dụng chúng để đưa ra dự đoán cho điểm dữ liệu mới dựa trên các trọng số về khoảng cánh của các lớp của các điểm gần nhất này. Cách tiếp cận này dựa trên giả định rằng các điểm dữ liệu tương tự về mặt không gian sẽ thuộc cùng một lớp.

Mặc dù KNN đơn giản và dễ triển khai, nhưng nó có một số hạn chế, bao gồm khả năng ảnh hưởng lớn của việc lựa chọn K và cần phải lưu trữ toàn bộ dữ liệu huấn luyện. Tuy nhiên, KNN vẫn được sử dụng rộng rãi trong các ứng dụng thực tế và thường được sử dụng như một công cụ đầu tiên để thử nghiệm các bài toán phân loại dữ liệu. Một cách tổng quan, KNN là một thuật toán dự đoán đầu ra của một điểm dữ liệu mới bằng cách sử dụng thông tin từ K điểm dữ liệu gần nhất trong tập huấn luyện.

2.2.4. Mạng Convolutional Neural Network

Theo nhận định của tác giả Wood [24], Mạng Convolutional Neural Network (CNN) đại diện cho một loại kiến trúc mạng học sâu được tạo ra để xử lý các dữ liệu có cấu trúc, đặc biệt là hình ảnh. CNN đã tạo ra một sự cách mạng trong lĩnh vực thị giác máy tính và trở thành tiêu chuẩn tiên phong trong nhiều ứng dụng liên quan đến xử lý hình ảnh. Nó không chỉ giúp cải thiện đáng kể hiệu suất trong các nhiệm vụ như phân loại hình ảnh mà còn đã chứng tỏ khả năng thành công trong việc xử lý ngôn ngữ tự nhiên để phân loại và xử lý văn bản.

Bên cạnh đó, nghiên cứu của tác giả Tuấn [25] từng đề cập rằng kiến trúc mạng CNN có thể khác nhau tùy vào mục đích sử dụng và loại dữ liệu được sử dụng, một số lớp cơ bản được sử dụng trong hầu hết các kiến trúc của mạng CNN.



Hình 2. 6: Mạng CNN [26]

a) Lớp Convolutional

Đây là lớp đầu tiên trong quá trình trích xuất đặc trưng từ dữ liệu đầu vào. Lớp Convolutional sử dụng các bộ lọc (filters) để quét qua từng phần của dữ liệu đầu vào, nhằm phát hiện các đặc trưng cục bộ của dữ liệu và tạo ra một tập hợp các ma trận đặc trưng (feature maps). Mục đích chính của lớp tích chập là xác định sự xuất hiện của các mẫu tại các vị trí cụ thể trong dữ liệu.

b) Lớp Pooling

Lớp này giảm kích thước của các ma trận đặc trưng (feature maps) bằng cách áp dụng các hàm lấy mẫu trên các vùng không trùng lặp của các ma trận đặc trưng (feature maps). Hai loại phổ biến của hàm lấy mẫu trong CNN là Max Pooling và Average Pooling. Chúng cho phép lựa chọn giá trị lớn nhất hoặc giá trị trung bình từ một vùng của ma trận đặc trưng, từ đó làm nổi bật các đặc trưng quan trọng trong ma trận đặc trưng và làm giảm kích thước của chúng.

c) Lớp Fully Connected

Lớp này được sử dụng để kết nối tất cả các neuron của lớp trước với tất cả các neuron của lớp sau, tạo thành một mạng neuron truyền thẳng đầy đủ. Nó thường được đặt ở cuối của kiến trúc CNN trước khi tạo đầu ra của mô hình. Lớp này giúp mô hình học và biểu diễn các đặc trưng phức tạp của dữ liệu và sử dụng các hàm kích hoạt để đưa ra phân phối xác suất cho các lớp đầu ra trong tác vụ phân loại.

2.2.5. Mô hình ngôn ngữ PhoBERT

Theo kết quả nghiên cứu của tác giả Quốc và Tuấn [27] đã đề xuất mô hình PhoBERT (Pho Bidirectional Encoder Representations from Transformers) là một mô hình mã hóa hai chiều dữ liệu sử dụng nhiều khối Transformer. Mô hình này được đào tạo trên một tập dữ liệu văn bản tiếng Việt. Thực nghiệm tại Viện nghiên cứu Trí tuệ nhân tạo VinAI.

Mô hình PhoBERT đã trải qua quá trình huấn luyện trên một tập dữ liệu văn bản lớn với tổng dung lượng khoảng 20GB. Tập dữ liệu này bao gồm một phần dữ liệu từ Wikipedia tiếng Việt với dung lượng khoảng 1GB, cùng với phần còn lại là dữ liệu thu thập từ các nguồn tin tức tiếng Việt. Quá trình huấn luyện này đã đóng góp quan trọng

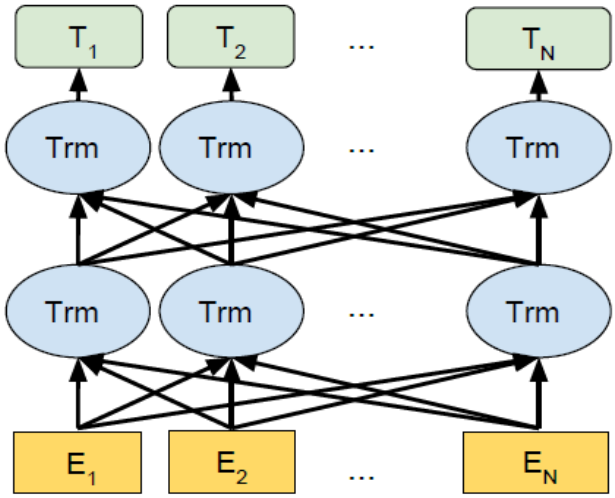
vào việc tạo ra một mô hình ngôn ngữ tiếng Việt có hiệu suất cao và khả năng đa dạng trong việc hiểu và phân tích ngôn ngữ.

Và mô hình này sử dụng lớp *RDRSegmenter* của thư viện *VncoreNLP* để trích xuất các từ trong dữ liệu đầu vào trước khi vào lớp mã hoá. Bảng 2.2 mô tả thông tin cơ bản hình thành nên kiến trúc mô hình PhoBERT.

Bảng 2. 2: Các kiến trúc của mô hình PhoBERT [18]

Mô hình	**Tham số**	**Transformers**	**Hidden Layer**
PhoBERT base	135M	12	768
PhoBERT large	370M	24	1024

PhoBERT là một phương pháp huấn luyện trước các biểu diễn ngôn ngữ, mà chúng dựa trên kiến trúc mô hình mạng mô phỏng theo hệ thống neuron thần kinh con người. Mô hình này giúp phân biệt rõ hơn ngữ cảnh bằng cách xem xét toàn bộ các từ trong một câu. Điều này khác biệt hoàn toàn so với phương pháp truyền thống dựa trên thứ tự xuất hiện của từng từ trong câu. PhoBERT cho phép mô hình ngôn ngữ hiểu về ngữ cảnh của một từ vựng dựa trên sự liên kết với các từ xung quanh nó, thay vì chỉ dựa vào từ trước hoặc sau nó.



Hình 2. 7: Kiến trúc của mô hình PhoBERT [18]

Mô hình PhoBERT sử dụng khả năng tiếp nhận thông tin từ cả hai hướng của một từ để cải thiện sự hiểu biết và khả năng đại diện ngôn ngữ. Điều này đánh dấu một sự tiến bộ quan trọng trong lĩnh vực NLP, đặc biệt là trong việc làm cho các mô hình ngôn

ngữ trở nên thông minh và linh hoạt hơn trong việc hiểu ngôn ngữ (NLU) và sinh ngôn ngữ (NLG).

Với một nguồn tri thức khổng lồ từ mô hình PhoBERT. Việc ứng dụng lại cần phải tinh chỉnh sao cho phù hợp thành một tác vụ cụ thể. Việc tận dụng lại tri thức từ mô hình ngôn ngữ lớn PhoBERT sẽ giúp các tác vụ được tối ưu với dữ liệu mới vì mô hình đã được học trên cả hai nguồn dữ liệu tri thức đó là dữ liệu huấn luyện và dữ liệu khổng lồ đã được học tập trước đó.

2.3. Các nền tảng công nghệ sử dụng

2.3.1. Pytube

Pytube [28] là một thư viện Python mã nguồn mở được sử dụng để trích xuất và xử lý video từ YouTube một cách thuận tiện. Nó cung cấp một giao diện dễ sử dụng để tải xuống video từ YouTube, cho phép người sử dụng tải về video theo yêu cầu và thực hiện xử lý dữ liệu với video này.

Thư viện này hỗ trợ cách tiếp cận thông qua API YouTube cho phép truy cập thông tin về video như độ phân giải, thời lượng, tác giả và các yếu tố dữ liệu khác. Thư viện cung cấp khả năng tải video ở nhiều định dạng và chất lượng khác nhau.

Trong môi trường nghiên cứu và phát triển ứng dụng, *Pytube* được sử dụng để thu thập dữ liệu, nghiên cứu về video trên mạng xã hội và cung cấp nguồn tài nguyên phong phú cho các dự án liên quan đến video và âm thanh trực tuyến.

2.3.2. Pydub

Pydub [29] được thiết kế để đơn giản hóa quy trình xử lý âm thanh, cho phép người sử dụng thao tác với các định dạng âm thanh khác nhau một cách linh hoạt. Việc tích hợp và sử dụng *Pydub* không đòi hỏi người lập trình có kiến thức chuyên sâu về xử lý âm thanh, giúp nâng cao khả năng tiếp cận của cộng đồng phát triển.

Pydub hỗ trợ nhiều định dạng file âm thanh phổ biến như MP3, WAV, FLAC và nhiều định dạng khác. Điều này giúp các nhà nghiên cứu và nhà phát triển dễ dàng làm việc với dữ liệu âm thanh từ nhiều nguồn khác nhau mà không phải lo lắng về định dạng.

Pydub cung cấp nhiều chức năng xử lý âm thanh, từ cắt, ghép đến làm mịn sóng âm. Các chức năng này không chỉ giúp tiết kiệm thời gian mà còn mang lại hiệu suất cao cho quá trình xử lý dữ liệu âm thanh.

Với *Pydub*, người sử dụng có thể dễ dàng điều chỉnh các thông số âm thanh như âm lượng, tần số và độ phủ sóng. Điều này mở ra nhiều cơ hội trong việc tinh chỉnh và tối ưu hóa âm thanh cho các ứng dụng cụ thể.

2.3.3. Speech_recognition

Speech_recognition [30] là một thư viện Python mã nguồn mở chuyên dụng trong việc nhận dạng giọng nói từ các nguồn âm thanh, mở ra nhiều ứng dụng từ ghi âm, điều khiển bằng giọng nói, cho đến hệ thống hỗ trợ người dùng trong nhiều lĩnh vực khác nhau.

Theo kết quả nghiên cứu [31] cho thấy rằng thư viện *speech_recognition* có độ chính xác tương đương với các thư viện khác, nhưng có tốc độ xử lý chậm hơn. Cụ thể, thư viện *speech_recognition* có độ chính xác trung bình là 91,5%, trong khi các thư viện khác có độ chính xác trung bình từ 91,6% đến 92,5%. Thư viện *CMU Sphinx* có độ chính xác cao nhất, nhưng có tốc độ xử lý chậm nhất. Thư viện *Kaldi* có tốc độ xử lý nhanh nhất, nhưng độ chính xác thấp hơn so với thư viện *CMU Sphinx*. Thư viện *Google Cloud Speech-to-Text* có độ chính xác và tốc độ xử lý tương đương với thư viện *CMU Sphinx*. Kết quả nghiên cứu [31] cho thấy rằng hiệu năng xử lý của thư viện *speech_recognition* phụ thuộc vào nhiều yếu tố, bao gồm tập dữ liệu được sử dụng để đào tạo mô hình, yêu cầu cụ thể của ứng dụng, và phần cứng được sử dụng.

Với tính linh hoạt và khả năng tích hợp cao, thư viện *speech_recognition* trở thành một công cụ quan trọng trong lĩnh vực NLP. Sự tiện ích và khả năng mở rộng của nó mở ra nhiều cánh cửa cho ứng dụng và nghiên cứu trong việc tận dụng thông tin từ ngôn ngữ nói.

2.3.4. Concurrent.futures

Trong việc nghiên cứu và phát triển phần mềm, việc xử lý các tác vụ đồng thời là một phần quan trọng để tối ưu hóa hiệu suất của ứng dụng. Thư viện *concurrent.futures* [32] trong ngôn ngữ lập trình Python là một công cụ mạnh mẽ để thực hiện xử lý song song và quản lý nhiều luồng công việc cùng một lúc. Tính năng của thư viện này đã đem lại những cơ hội đáng kể cho việc tối ưu hóa thời gian xử lý và cải thiện hiệu suất cho các ứng dụng Python.

Thư viện *concurrent.futures* cung cấp hai lớp chính là *ThreadPoolExecutor* và *ProcessPoolExecutor*, cho phép thực thi các tác vụ đồng thời sử dụng các luồng hoặc tiến trình.

- *ThreadPoolExecutor* cho phép xử lý đa luồng thông qua việc tạo và quản lý một nhóm các luồng để thực hiện các tác vụ một cách song song. Điều này giúp tận dụng các luồng có sẵn trong hệ thống để thực hiện các tác vụ mà không ảnh hưởng đến hiệu suất chung.

- *ProcessPoolExecutor* sử dụng một nhóm các tiến trình để thực hiện các tác vụ. Mỗi tiến trình là một bản sao hoàn chỉnh của quá trình gốc, có khả năng thực hiện các tác vụ độc lập với các tiến trình khác. Điều này cho phép tận dụng tài nguyên hệ thống, đặc biệt là trên các hệ thống có nhiều tài nguyên CPU.

2.3.5. Pyvi

Pyvi [33] là một công cụ mã nguồn mở dành cho NLP. *Pyvi* cung cấp các công cụ mạnh mẽ để phân tích và xử lý văn bản tiếng Việt. Nó bao gồm các chức năng chính như phân đoạn từ, tách từ, phân loại từ loại và các công cụ hỗ trợ khác cho việc nghiên cứu và phát triển trong lĩnh vực NLP.

Pyvi ra đời với mục tiêu đơn giản hóa quá trình xử lý văn bản tiếng Việt bằng cách cung cấp các công cụ dễ sử dụng và linh hoạt. Bằng việc kết hợp các thuật toán và tài nguyên ngôn ngữ, *Pyvi* cung cấp khả năng xử lý văn bản hiệu quả, từ việc tách từ đơn giản đến việc phân tích ngữ nghĩa phức tạp. Điều này mang lại sự linh hoạt và hiệu quả trong việc xử lý ngôn ngữ tự nhiên cho cộng đồng phát triển và nghiên cứu. Công cụ này đã thu hút sự quan tâm của cộng đồng NLP và các nhà nghiên cứu trong lĩnh vực NLP cho tiếng Việt, giúp họ tiết kiệm thời gian và nỗ lực trong việc xây dựng các ứng dụng và dịch vụ với ngôn ngữ tiếng Việt.

2.3.6. Scikit-learn

Thư viện *scikit-learn* là một công cụ mạnh mẽ và linh hoạt trong lĩnh vực học máy và khoa học dữ liệu. *Scikit-learn* cung cấp một loạt các công cụ và thuật toán để thực hiện nhiều tác vụ khác nhau trong việc xử lý dữ liệu, phân tích dữ liệu và dự đoán.

Theo nghiên cứu [34], thư viện *scikit-learn* được xây dựng trên cơ sở của ngôn ngữ lập trình Python và tập trung vào việc tiện lợi, hiệu suất và có tính ứng dụng cao.

Thư viện này cung cấp các công cụ mạnh mẽ cho việc tiền xử lý dữ liệu, bao gồm các công cụ chuẩn hóa, mã hóa và trích xuất đặc trưng. Ngoài ra, nó cung cấp các thuật toán học máy phổ biến như hồi quy, phân loại, gom cụm và rất nhiều thuật toán khác.

Thư viện này cũng đi kèm với các công cụ hữu ích để đánh giá mô hình và tinh chỉnh siêu tham số, giúp người dùng xác định mô hình tốt nhất cho dữ liệu của họ. Điều này đảm bảo rằng người dùng có thể thực hiện các thí nghiệm một cách hiệu quả và chính xác để đạt được kết quả tối ưu. Và nhận được sự hỗ trợ từ cộng đồng người dùng.

2.3.7. TensorFlow

TensorFlow [35], phát triển bởi Google Brain Team, đã trở thành một trong những công cụ phổ biến hàng đầu trong lĩnh vực học máy và trí tuệ nhân tạo. *TensorFlow* đã thu hút sự chú ý của cộng đồng do các tính năng mạnh mẽ và tích hợp linh hoạt.

TensorFlow cung cấp một cách tiếp cận đồng nhất và linh hoạt để xây dựng, huấn luyện và triển khai các mô hình học máy. Với cú pháp dễ sử dụng và khả năng tương tác cao, *TensorFlow* cho phép các nhà nghiên cứu và nhà phát triển triển khai các mô hình từ đơn giản đến phức tạp một cách dễ dàng.

TensorFlow cũng hỗ trợ nhiều loại mô hình học máy, bao gồm mạng neural, mô hình học sâu, học tăng cường và nhiều loại mô hình khác. Điều này làm cho nó trở thành một công cụ linh hoạt không chỉ cho việc nghiên cứu mà còn cho các ứng dụng thương mại.

TensorFlow không chỉ là một thư viện mã nguồn mở mà còn là một cộng đồng rộng lớn, với hàng ngàn người dùng và nhà phát triển trên khắp thế giới, đóng góp vào việc phát triển và cải tiến liên tục của nó. Sự kết hợp giữa tính linh hoạt, hiệu suất và cộng đồng sáng tạo đã tạo nên sức hút lớn của *TensorFlow* trong cộng đồng ML và AI.

2.3.8. PyTorch

PyTorch [36] là một thư viện mã nguồn mở được sử dụng rộng rãi trong lĩnh vực ML và AI. Nó cung cấp một cách linh hoạt và mạnh mẽ để xây dựng và huấn luyện mô hình mạng neural và thực hiện các tính toán số học trên dữ liệu. *PyTorch* được phát triển bởi Facebook's AI Research Lab (FAIR) và được phát hành lần đầu vào năm 2016. Tính đến thời điểm này, *PyTorch* đã trở thành một trong những công cụ quan trọng nhất cho việc nghiên cứu và triển khai các ứng dụng trí tuệ nhân tạo.

PyTorch được xây dựng dựa trên việc sử dụng các cấu trúc dữ liệu linh hoạt gọi là *tensors*, giúp người dùng thực hiện các phép toán số học một cách hiệu quả trên GPU. Điều này cho phép việc tính toán song song, làm tăng tốc quá trình huấn luyện mô hình so với việc sử dụng CPU.

Một trong những điểm mạnh của *PyTorch* là cách tiếp cận dễ dàng và linh hoạt trong việc xây dựng mô hình. Người dùng có thể tận dụng các tính năng tự động lan truyền ngược để tự động tính đạo hàm, giúp đơn giản hóa quá trình huấn luyện mô hình. Điều này cung cấp sự thuận tiện cho việc thử nghiệm và điều chỉnh mô hình một cách linh hoạt. Ngoài ra, cộng đồng người dùng *PyTorch* rất lớn và năng động, cung cấp nhiều tài liệu học tập, ví dụ và mã nguồn mở. Công cụ này cũng được hỗ trợ bởi nhiều công ty hàng đầu trong ngành công nghiệp, từ việc nghiên cứu đến triển khai ứng dụng thực tế.

2.3.1. Streamlit

Thư viện *Streamlit* [37] là một công cụ mạnh mẽ trong lĩnh vực phân tích dữ liệu và xử lý dữ liệu tương tác. Nó cho phép người dùng tạo giao diện người dùng (GUI) cho các ứng dụng dữ liệu một cách nhanh chóng và dễ dàng thông qua việc sử dụng ngôn ngữ lập trình Python. *Streamlit* đã nhanh chóng trở thành một công cụ phổ biến trong cộng đồng khoa học dữ liệu do tính linh hoạt và khả năng tương tác cao.

Streamlit được thiết kế để giúp người dùng tạo các ứng dụng web tương tác một cách dễ dàng. Nó cung cấp các công cụ và thư viện giúp người dùng xây dựng giao diện người dùng trực quan và hiệu quả một cách nhanh chóng, đồng thời hỗ trợ việc hiển thị dữ liệu phức tạp một cách rõ ràng.

Streamlit có một loạt các tính năng mạnh mẽ bao gồm khả năng tạo ra các biểu đồ, bảng và đồ thị một cách linh hoạt, cho phép người dùng tương tác với dữ liệu một cách trực quan. Điều này cung cấp một phương tiện hiệu quả để trình bày và truy cập thông tin dữ liệu một cách dễ dàng hơn.

Trong lĩnh vực nghiên cứu và phát triển ứng dụng dữ liệu, *Streamlit* đóng vai trò quan trọng bằng cách cung cấp một cơ chế linh hoạt, nhanh chóng để tạo và chia sẻ ứng dụng tương tác dựa trên dữ liệu, giúp nghiên cứu sinh, nhà phân tích dữ liệu và nhà phát triển tận dụng dữ liệu một cách hiệu quả.

CHƯƠNG III: PHƯƠNG PHÁP THỰC HIỆN

3.1. Mô hình nghiên cứu tổng quan

Trong nghiên cứu này, trình bày quá trình tổng hợp và cụ thể từ việc thu thập dữ liệu đến xây dựng và đào tạo mô hình phân loại chủ đề tự động cho bản tin thời sự truyền hình. Mục tiêu là cung cấp một cái nhìn chi tiết về cách thực hiện nghiên cứu này, từng bước một nhằm đảm bảo tính khả thi và đáng tin cậy của phương pháp được sử dụng.

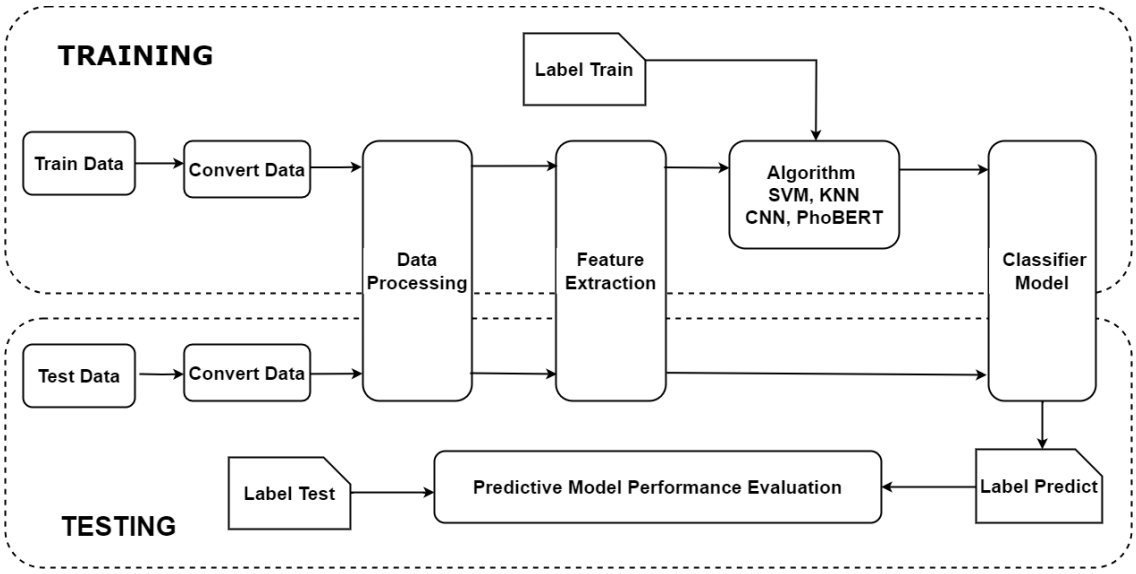
Bước đầu tiên trong quá trình nghiên cứu là thu thập dữ liệu video từ các kênh YouTube của các đài truyền hình đang hoạt động trong khu vực Việt Nam. Dữ liệu video này sau đó sẽ được chuyển đổi thành dạng văn bản bằng cách sử dụng thư viện xử lý âm thanh để nhận dạng. Dữ liệu văn bản thô thu được sẽ tiếp tục được tiền xử lý để loại bỏ các yếu tố không cần thiết và chuẩn bị cho các bước tiếp theo.

Tiếp theo, quá trình trích xuất đặc trưng cho dữ liệu được thực hiện. Dữ liệu sau khi được tiền xử lý sẽ được biểu diễn dưới dạng các đặc trưng có thể được sử dụng trong việc đào tạo và đánh giá mô hình.

Dữ liệu thực nghiệm trong nghiên cứu này được chia thành hai tập: tập dữ liệu huấn luyện và tập dữ liệu kiểm tra.

- Tập dữ liệu huấn luyện được sử dụng để xây dựng và đào tạo các mô hình học máy, học sâu và mô hình ngôn ngữ.
- Tập dữ liệu kiểm tra được sử dụng để đánh giá hiệu suất của các mô hình đã được huấn luyện.

Quá trình nghiên cứu này bao gồm việc tiến hành thu thập, chuyển đổi, tiền xử lý, trích xuất đặc trưng và chia thành các tập dữ liệu khác nhau để huấn luyện các mô hình. Đánh giá các mô hình dựa trên các tiêu chí đánh giá áp dụng cho bài toán phân loại. Mục tiêu cuối cùng là tạo ra một mô hình phân loại chủ đề tự động có khả năng dự đoán và phân loại các bản tin thời sự truyền hình một cách hiệu quả. Hình 3.1 trình bày sơ đồ tổng quan về quá trình nghiên cứu đã thực hiện.



Hình 3. 1: Mô hình tổng quát của hệ thống thực nghiệm

3.2. Thu thập dữ liệu

Trong giai đoạn thu thập dữ liệu, sẽ sử dụng các công cụ và thư viện để thu thập dữ liệu video từ YouTube một cách hiệu quả. Thư viện *pytube* được sử dụng để tương tác và làm việc với nội dung video từ YouTube, trong khi thư viện *concurrent.futures* hỗ trợ việc thực thi các tác vụ bất đồng bộ trên video để tối ưu hóa thời gian.

Thư viện *pytube* cho phép thực hiện các tác vụ quan trọng như tải xuống video, trích xuất thông tin về video và lấy ra Playlist. Sử dụng nó để tạo kết nối và tải về video từ các đài truyền hình trên YouTube. Điều này giúp truy cập một lượng lớn nội dung video một cách tự động.

Thư viện *concurrent.futures* cung cấp lớp *ThreadPoolExecutor* hỗ trợ thực thi nhiều luồng đồng thời tại một thời điểm, giúp tận dụng tối đa tài nguyên của máy tính. Sử dụng nó để đồng thời tải xuống nhiều video cùng một lúc, làm tăng tốc quá trình thu thập dữ liệu và tiết kiệm thời gian.

Trong quá trình thu thập dữ liệu. Áp dụng một điều kiện lọc quan trọng. Chỉ những video có thời lượng nhỏ hơn hoặc bằng 5 phút mới được tải xuống. Điều này giúp quản lý dung lượng lưu trữ và đảm bảo rằng nội dung video có độ dài ngắn hơn để thuận tiện cho việc xử lý và huấn luyện sau này.

Mỗi video được gán nhãn tương ứng dựa trên Playlist của các đài truyền hình trên YouTube. Nhãn này đại diện cho chủ đề hoặc danh mục của video, giúp phân biệt và phân loại dữ liệu một cách chính xác. Việc này quan trọng để đảm bảo dữ liệu thu thập

đáng tin cậy và chất lượng. Bảng 3.1 trình bày các chủ đề cùng với các từ khóa phổ biến tương ứng. Các từ khóa được xác định dựa trên tần suất xuất hiện cao nhất liên quan đến mỗi chủ đề.

Bảng 3. 1: Mô tả các chủ đề cùng với các từ khóa phổ biến tương ứng

Chủ đề	**Từ khóa**
Chính trị Xã hội	ủy_ban, nhân_dân, địa_phương, thành_phố, hội_đồng, chủ_tịch, tổ_chức, thường_vụ, mặt_trận, tổ_quốc,…
Dự báo thời tiết	nhiệt_độ, độ_ẩm, khu_vực, dự_báo, ngày_nắng, ngày_mưa, mưa_rào, mây, thay_đổi, huyện, tỉnh,…
Kinh tế	đô_la, tỷ_đồng, doanh_nghiệp, thị_trường, ngân_hàng, đầu_tư, phát_triển, sản_suất, hàng_hóa, xuất_khẩu,…
Môi trường	cháy_rừng, ô_nhiễm, lũ_lụt, thiên_tai, hạn_hán, đất, nước, sạc_lỡ, thiện_hại, tuyến_đường, đám_cháy,…
Nông nghiệp	hợp_tác xã, thu hoạch, lúa_gạo, trái_cây, phân_bón, nông_sản, nông_dân, doanh_nghiệp, sâu_bệnh, ,…
Pháp luật	đối_tượng, truy_nã, công_an, cảnh_sát, điều_tra, pháp_luật, xét_xử, phát_hiện, ma_túy, cá_độ, bắt giữ,…
Sức khỏe	bệnh_nhân, thuốc, bệnh, bác_sĩ, cấp cứu, dịch_bệnh, covid, viêm_gan, bệnh_viện, triệu_chứng, y_tế,…
Thế giới	trung_quốc, mỹ, nga, triều_tiên, bắc_kinh, ấn_độ, tổng_thống, chính_phủ, liên_hợp quốc, thế_giới, khu_vực, châu_âu, quốc_gia,…
Thể thao	thi_đấu, cầu_thủ, trận_đấu, đội_tuyển, vận_động_ viên, câu lạc_bộ, mùa_giải, huyến_luyện_viên, bóng_đá,…
Văn hóa	lễ_hội, du_lịch, truyền_thống, du_lịch, nghệ_thuật, tác_phẩm, di sản, tranh, biểu_diễn, tinh_thần,…
Giáo dục	trường, trung_học, phổ_thông, quốc_gia, học_sinh, kỳ thi, đại_học, đào_tạo, nhà_trường, sinh_viên,…

3.3. Chuyển đổi dữ liệu

Sau khi đã thu thập dữ liệu dưới dạng video từ YouTube, tiến hành chuyển đổi dữ liệu này thành dạng văn bản để có thể thực hiện các phân tích và xử lý dữ liệu tiếp theo. Quá trình chuyển đổi này được thực hiện qua các bước.

Thư viện *pydub* là một công cụ mạnh mẽ để xử lý âm thanh. Sử dụng lớp *AudioSegment* từ thư viện *pydub* để chuyển đổi video thành định dạng âm thanh. Điều này giúp loại bỏ phần hình ảnh và chỉ giữ lại âm thanh trong video.

Sau khi đã có âm thanh từ video, sử dụng thư viện *speech_recognition* để thực hiện nhận dạng giọng nói và chuyển đổi nó thành dạng văn bản. Thư viện này hỗ trợ nhận dạng tiếng nói từ nhiều nguồn âm thanh khác nhau, giúp trích xuất thông tin từ video một cách tự động.

Việc chuyển đổi dữ liệu từ video thành văn bản có thể tốn nhiều thời gian, đặc biệt khi có nhiều video cần xử lý. Để tối ưu hóa quá trình chuyển đổi và tiết kiệm thời gian, sử dụng lớp *ThreadPoolExecutor* từ thư viện *concurrent.futures*. Lớp này cho phép thực hiện chuyển đổi đồng thời trên nhiều video cùng một lúc, làm tăng tốc độ xử lý dữ liệu.

Kết quả của quá trình chuyển đổi là các tệp tin văn bản, mỗi tệp tương ứng với nội dung âm thanh từ một video. Các tệp tin này được lưu trữ để phục vụ cho các quá trình xử lý và huấn luyện sau này.

3.4. Tiền xử lý dữ liệu

Dữ liệu thu thập từ video ban đầu có thể không hoàn hảo và cần phải trải qua một loạt bước xử lý để làm sạch và chuẩn hóa nó trước khi tiếp tục vào quá trình phân tích và huấn luyện. Các bước xử lý dữ liệu đảm bảo tính đáng tin cậy và hiệu quả của dữ liệu văn bản.

Dữ liệu thu thập có thể chứa các đoạn văn bản rỗng hoặc dữ liệu với lỗi chính tả. Điều này có thể ảnh hưởng đến kết quả của quá trình phân tích và huấn luyện. Sử dụng thư viện *pyvi*, ta có thể sử dụng lớp *ViUtils* để sửa lỗi chính tả cho Tiếng Việt và loại bỏ các đoạn văn bản rỗng. Nếu dữ liệu văn bản chứa các thẻ HTML hoặc các kí tự kéo dài không cần thiết, cần xóa chúng để đảm bảo dữ liệu là dạng văn bản thuần túy và không bị nhiễu.

Sử dụng lớp *ViTokenizer* từ thư viện *pyvi*, có thể thực hiện tách từ cho văn bản Tiếng Việt một cách hiệu quả. Khi sử dụng lớp *ViTokenizer* các từ có nghĩa bổ trợ nhau sẽ được gom thành một cụm từ. Ngoài ra, việc loại bỏ các từ không mang nhiều ý nghĩa cho văn bản (stopwords) là quan trọng để giảm kích thước của dữ liệu và tập trung vào các từ quan trọng.

Cuối cùng, quá trình chuẩn hóa văn bản sẽ được thực hiện để đưa văn bản về một dạng chuẩn nhất, nhằm tạo sự đồng nhất trong các đoạn văn bản.

3.5. Trích xuất đặc trưng

Máy tính không thể hiểu ngôn ngữ tự nhiên một cách trực tiếp mà chỉ có thể xử lý thông tin ngôn ngữ khi được biểu diễn dưới dạng không gian vector. Quá trình trích xuất đặc trưng từ dữ liệu văn bản trong nghiên cứu bao gồm một chuỗi các bước quan trọng.

Trước hết, phương pháp N-Gram được áp dụng để biểu diễn thông tin văn bản dưới dạng không gian vector. Phương pháp này tách văn bản thành các từ đơn lẻ, mỗi từ được coi là một đơn vị độc lập. Trong nghiên cứu, giá trị N được thiết lập là 1, tương đương với việc sử dụng unigram, trong đó mỗi từ được biểu diễn độc lập và không xem xét đến thứ tự hay mối quan hệ tuần tự giữa chúng.

Sau đó, phương pháp TF-IDF được sử dụng để trích xuất đặc trưng, dựa trên công thức (2.4). Một từ điển từ vựng được xác định trước dựa trên phương pháp N-Gram. Mỗi văn bản sau đó được biểu diễn dưới dạng một vector với chiều dài tương đương số từ trong từ điển và các giá trị trong vector phản ánh tần suất xuất hiện của từ trong văn bản. Phương pháp này không quan tâm đến ngữ pháp hay thứ tự của từ. Trong quá trình triển khai, lớp *TfidfVectorizer* từ thư viện *Sklearn* được sử dụng để thực hiện trích xuất đặc trưng theo phương pháp TF-IDF.

Tuy nhiên, do kích thước của không gian vector sau khi áp dụng TF-IDF là rất lớn, điều này có thể gây nhiễu và ảnh hưởng đến khả năng phân loại. Vì vậy, nghiên cứu áp dụng phương pháp phân rã ma trận SVD để giảm số chiều không gian đặc trưng. Quá trình này loại bỏ các thuộc tính không nổi bật hoặc có tần suất xuất hiện thấp, giúp giảm kích thước vector mà không mất đi quan hệ tuyến tính giữa các phần tử. Phương pháp SVD chuyển đổi vector ban đầu thành một vector mới có kích thước nhỏ hơn, nhưng vẫn giữ lại những thông tin quan trọng nhất từ vector gốc. Để thực hiện việc giảm số chiều của thuộc tính, nghiên cứu sử dụng lớp *TruncatedSVD* từ thư viện *Sklearn*.

3.6. Xây dựng mô hình

Sau khi đã thu được đặc trưng của của tập dữ liệu. Tiến hành xây dựng 4 mô hình khác nhau nhằm nghiên cứu và phân tích sự hiệu quả của các mô hình. Các mô hình được lựa chọn bao gồm thuật toán SVM và KNN, mạng CNN và mô hình PhoBERT.

3.6.1. SVM

Trong quá trình quá trình đào tạo thuật toán SVM với mục tiêu phân loại dữ liệu, đã thực hiện trích xuất đặc trưng bằng phương pháp TF-IDF và tiến hành áp dụng phép giảm chiều SVD. Để lựa chọn các siêu tham số tối ưu cho SVM, sẽ áp dụng phương pháp tìm kiếm siêu tham số sử dụng lớp *GridSearchCV* có sẵn trong thư viện *Sklearn*. Phương pháp này hoạt động bằng cách thực hiện việc duyệt qua toàn bộ không gian siêu tham số đã xác định trước và đánh giá hiệu suất của mô hình cho từng tổ hợp siêu tham số này bằng kỹ thuật *Cross validation*. Quá trình này cho phép tìm ra các tổ hợp siêu tham số tối ưu nhất cho mô hình SVM.

Các siêu tham số được điều chỉnh để tối ưu hóa mô hình đã được thiết lập dựa trên Bảng 3.2. Quá trình này giúp xác định các giá trị tối ưu cho siêu tham số như C, Gamma và Kernel của SVM, đảm bảo rằng mô hình được cấu hình để đạt được hiệu suất cao nhất trong việc phân loại văn bản.

Bảng 3. 2: Trạng thái các tham số trong SVM khi sử dụng GridSearchCV

GridSearchCV	**Siêu tham số**			
C	0.1	1	10	100
Gamma	1	0.1	0.01	0.001
Kernel	rbf			

3.6.2. KNN

Trong giai đoạn đào tạo thuật toán KNN với mục tiêu phân loại dữ liệu, đã tiến hành trích xuất đặc trưng bằng phương pháp TF-IDF và sau đó sử dụng phép giảm chiều SVD. Để xác định các siêu tham số tối ưu cho mô hình KNN, phương pháp tìm kiếm tham số *GridSearchCV* đã sử dụng. Phương pháp này hoạt động bằng cách duyệt qua toàn bộ không gian siêu tham số đã xác định trước và đánh giá hiệu suất của mô hình cho từng tổ hợp siêu tham số này bằng kỹ thuật *Cross validation*. Quá trình này giúp tìm ra các tổ hợp siêu tham số tối ưu nhất cho mô hình KNN.

Các siêu tham số được điều chỉnh để tối ưu hóa mô hình đã được thiết lập dựa trên Bảng 3.3. Quá trình này giúp xác định các giá trị tối ưu cho siêu tham số như K, Weights và P của mô hình KNN, đảm bảo rằng mô hình được cấu hình để đạt được hiệu suất cao nhất trong việc phân loại văn bản.

Bảng 3. 3: Trạng thái các tham số trong KNN khi sử dụng GridSearchCV

GridSearchCV	**Siêu tham số**	
K	1, 3, 5…25,27,29	
Weights	uniform	distance
P	1	2

3.6.3. CNN

Trong quá trình huấn luyện mô hình CNN với mục tiêu phân loại dữ liệu, đã thực hiện trích xuất đặc trưng bằng phương pháp TF-IDF và sau đó sử dụng phép giảm chiều dữ liệu thông qua SVD. Mô hình tóm tắt của mạng CNN được thể hiện qua Bảng 3.4. Các lớp của mạng được chọn bằng phương pháp thử và sai. Thực hiện thực nghiệm bằng việc sử dụng thư viện *Tensorflow* để xây dựng mô hình CNN.

Bảng 3. 4: Tóm tắt kiến trúc CNN thực nghiệm

Layer	**Output shape**	**Param #**
InputLayer	[(None, 300)]	0
Reshape	(None, 10, 30)	0
Bidirectional	(None, 10, 256)	122.880
Conv1D	(None, 8, 100)	76.900
Flatten	(None, 800)	0
Dense_1	(None, 512)	410.112
Dropout_1	(None, 512)	0
Dense_2	(None, 256)	131.328
Dropout_2	(None, 256)	0
Dense_3	(None, 128)	32.896
Dense_4	(None, 64)	8.256
Dense_5	(None, 11)	715
Total params: 783.087		
Trainable params: 783.08		
Non-trainable params: 0		

Mô hình mạng CNN thực nghiệm với kiến trúc mạng trong Bảng 3.4**.** Đặc trưng văn bản đầu vào có kích thước (None, 300) sẽ được định dạng lại thành kích thước (None, 10, 30) được đưa vào lớp Bidirectional.

Lớp Bidirectional nhận đầu vào có kích thước đầu vào với độ dài 10 và mỗi phần tử có kích thước là 256 được xử lý theo cả hai hướng. Cho phép mô hình học thông tin từ cả hai phía của chuỗi dữ liệu, cải thiện khả năng dự đoán và hiểu ngữ cảnh.

Lớp Conv1D có 100 bộ lọc, mỗi bộ lọc có kích thước Kernel là 3 và sử dụng hàm kích hoạt ReLU.

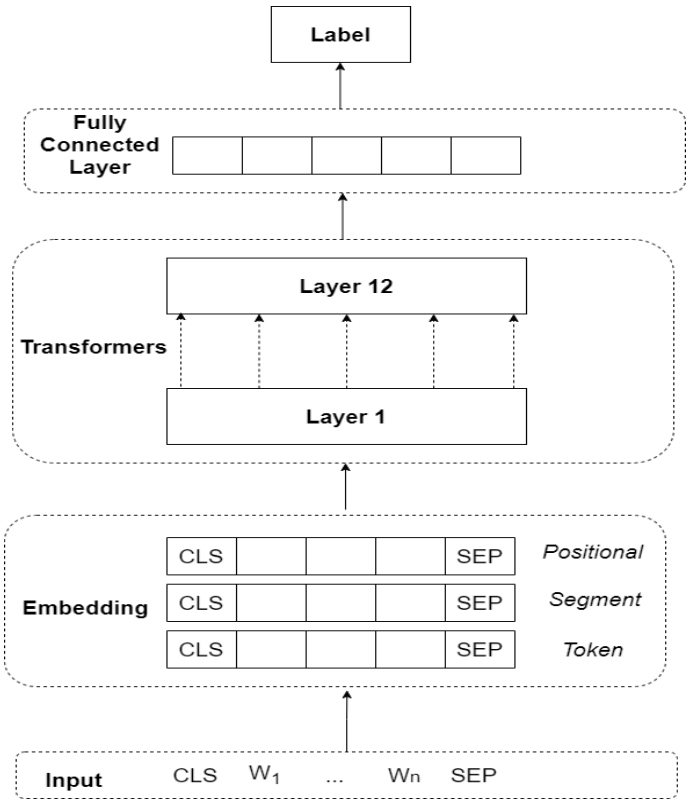
Lớp Flatten được sử dụng để chuyển đổi dữ liệu từ lớp Conv1D là ma trận hai chiều thành vector.

Lớp Dense_1 và Dense_2 với hàm kích hoạt ReLU và Dropout 20% để thu được đầu ra cuối cùng sẽ được xử lý cho các lớp tiếp theo.

Lớp Dense_3, Dense_4, Dense_5 nhận đầu ra của lớp Dense 2 với hàm kích hoạt ReLU, lớp này kết nối sử dụng hàm kích hoạt Softmax để phân phối xác suất cho từng chủ đề.

3.6.4. PhoBERT

Trong quá trình huấn luyện PhoBERT. Để điều chỉnh mô hình PhoBERT cho phù hợp với bài toán phân loại đa lớp, cần phải điều chỉnh các tham số trong mô hình. Sau đó, mô hình sẽ được huấn luyện lại trên tập dữ liệu tiền xử lý để đảm bảo rằng nó thực hiện nhiệm vụ phân loại một cách hiệu quả. Kiến trúc của mô hình PhoBERT cho tác vụ phân loại được thể hiện qua Hình 3.2**.** Thực hiện thử nghiệm bằng việc sử dụng thư viện *PyTorch* để xây dựng mô hình tinh chỉnh PhoBERT.



Hình 3. 2: Kiến trúc PhoBERT cho tác vụ phân loại

a) Khối Input

Mỗi mẫu văn bản đầu vào được biến đổi thành một chuỗi các vector token và sau đó được bổ sung bằng hai token [CLS] và [SEP]. Kích thước của các vector này không vượt quá 256 token.

- Token [CLS] đại diện cho toàn bộ câu trong nhiệm vụ phân loại.
- Token [SEP] đánh dấu sự kết thúc của câu.

b) Khối Embedding

Các vector biểu diễn của các token trong chuỗi sau đó được chuyển qua một chuỗi ba lớp liên tiếp có kích thước tương đồng, bao gồm các lớp chi tiết như sau:

- Lớp nhúng từ (Token embedding)
- Lớp nhúng phân đoạn (Segmentation embedding)
- Lớp nhúng vị trí (Position embedding)

c) Khối Transformer

Trong quá trình xử lý dữ liệu văn bản sử dụng kiến trúc mạng Transformer, bắt đầu bằng việc thực hiện việc kết hợp các vector nhúng cho mỗi token, vector nhúng cho mỗi đoạn văn bản và vector nhúng cho vị trí của mỗi token. Quá trình này bao gồm việc gộp các vector nhúng này để tạo ra đầu vào cho mạng Transformer. Trong thực nghiệm này, sử dụng mạng Transformer với tổng cộng 12 khối, được xây dựng dựa trên phiên bản PhoBERTbase. Quá trình này cung cấp cơ sở thông tin vững chắc cho mạng Transformer để tiến hành quá trình học và trích xuất thông tin từ các đầu vào văn bản.

d) Khối Fully Connected Layer

Trong giai đoạn này, thông tin từ các lớp trước đó sau khi đã trải qua quá trình trích xuất và trích chọn đặc trưng, sẽ được kết hợp toàn bộ các neuron được biểu diễn dưới dạng một Fully Connected Layer. Sau khi thông tin được lan truyền qua các lớp trước, một bộ phân loại cuối cùng được sử dụng để tính toán xác suất phân phối của mẫu dữ liệu đó vào mỗi lớp hay nhãn. Để thực hiện điều này, hàm kích hoạt Softmax được áp dụng. Hàm kích hoạt Softmax giúp chuẩn hóa giá trị đầu ra, biến chúng thành xác suất để dự đoán mẫu dữ liệu thuộc về từng lớp. Dự đoán cuối cùng của mô hình sẽ dựa trên xác suất cao nhất thu được từ hàm kích hoạt Softmax, xác định lớp phân loại mà mẫu dữ liệu được dự đoán thuộc về với xác suất cao nhất.

3.7. Các tiêu chí đánh giá mô hình

Với bài toán phân loại, việc đánh giá hiệu suất của mô hình là một phần quan trọng để đảm bảo tính chính xác và đáng tin cậy của quá trình phân loại. Để thực hiện đánh giá hiệu suất mô hình, một công cụ quan trọng được sử dụng là ma trận nhầm lẫn (confusion matrix). Ma trận nhầm lẫn là một biểu đồ đa chiều cung cấp thông tin về sự phân phối của dự đoán của mô hình so với thực tế. Ma trận nhầm lẫn điển hình được thể hiện trong Bảng 3.5.

Bảng 3. 5: Ma trận nhầm lẫn cho bài toán phân loại nhị phân

Confusion Matrix		**Predict Class**	
		Positive (P)	**Negative (N)**
Actual Class	**Positive (P)**	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

Ma trận này bao gồm các thành phần chính: True Positives (TP), True Negatives (TN), False Positives (FP) và False Negatives (FN) trong đó:

- **True Positive (TP)**: Mô hình dự đoán đúng một mẫu dương tính. Nói cách khác, mô hình cho rằng mẫu thuộc lớp dương tính và thực tế mẫu thuộc lớp dương tính.
- **False Positive (FP)**: Mô hình dự đoán sai một mẫu âm tính thành dương tính. Nói cách khác, mô hình cho rằng mẫu thuộc lớp dương tính, nhưng thực tế mẫu lại thuộc lớp âm tính.
- **True Negative (TN)**: Mô hình dự đoán đúng một mẫu âm tính. Nói cách khác, mô hình cho rằng mẫu thuộc lớp âm tính và thực tế mẫu thuộc lớp âm tính.
- **False Negative (FN)**: Mô hình dự đoán sai một mẫu dương tính thành âm tính. Nói cách khác, mô hình cho rằng mẫu thuộc lớp âm tính, nhưng thực tế mẫu lại thuộc lớp dương tính.

Sau khi có các chỉ số đánh giá có thể xác định các tiêu chí đánh giá để đo lường hiệu suất của bộ phân loại bao gồm:

- **Accuracy:** Độ bao phủ đo lường khả năng của mô hình xác định đúng các trường hợp dương tính (Positive) trên tổng số trường hợp thực sự là dương tính.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3.1}$$

- **Precision:** Precision đo lường khả năng của mô hình xác định đúng các trường hợp dương tính (Positive) trên tổng số trường hợp được dự đoán là dương tính.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

- **Recall:** Độ bao phủ đo lường khả năng của mô hình xác định đúng các trường hợp dương tính (Positive) trên tổng số trường hợp thực sự là dương tính.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

- **F1 Score**: Độ đo F1 là một số liệu tổng hợp của Precision và Recall, thường được sử dụng khi cần cân nhắc cả hai khía cạnh của hiệu suất mô hình.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.4}$$

Một mô hình phân loại với các chỉ số cao sẽ có hiệu suất chính xác và đáng tin cậy hơn trong việc dự đoán và phân loại dữ liệu. Các chỉ số này càng cao thì mô hình càng chính xác và đáng tin cậy hơn, giúp nâng cao chất lượng và ứng dụng của mô hình trong thực tế.

CHƯƠNG IV: KẾT QUẢ THỰC NGHIỆM

4.1. Dữ liệu thực nghiệm

Trên bộ dữ liệu được thu thập và tiền xử lý với tổng số 14.503 được gán 11 nhãn: Chính trị Xã hội, Dự báo thời tiết, Kinh tế, Môi trường, Nông nghiệp, Pháp luật, Sức khỏe, Thế giới, Thể thao, Văn hóa và Giáo dục được mô tả chi tiết trong Bảng 4.1.

Bảng 4. 1: Chủ đề và số lượng mẫu dữ liệu dùng trong thực nghiệm

Chủ đề	**Mẫu dữ liệu**			
	Train	**Test**	**Tổng**	**Tỷ lệ(%)**
Chính trị Xã hội	1.165	292	**1.457**	10.05
Dự báo thời tiết	1.055	264	**1.319**	9.09
Kinh tế	1.046	262	**1.308**	9.02
Môi trường	933	233	**1.166**	8.04
Nông nghiệp	1.123	281	**1.404**	9.68
Pháp luật	1.094	273	**1.367**	9.43
Sức khỏe	1.050	262	**1.312**	9.05
Thế giới	1.084	271	**1.355**	9.34
Thể thao	1.050	263	**1.313**	9.05
Văn hóa	1.097	274	**1.371**	9.45
Giáo dục	905	226	**1.131**	7.80
Tổng cộng	**11.602**	**2.901**	**14.503**	**100**

Trong thực nghiệm này, tập trung vào hai vấn đề quan trọng trong quá trình xây dựng mô hình, đó là hiện tượng *Overfitting* và *Underfitting*. Cả hai vấn đề này có khả năng gây ra sự suy giảm hiệu suất của mô hình và tạo ra sự không ổn định trong việc dự đoán dữ liệu mới.

Với mô hình SVM và KNN. Sử dụng kỹ thuật *K-Fold* nhằm giải quyết hai vấn đề trên. Trong thực nghiệm này đã lựa chọn giá trị K = 5 để thực hiện. Với mô hình CNN và PhoBERT sẽ phân tách 10% dữ liệu huấn luyện thành dữ liệu đánh giá.

4.2. Môi trường thực nghiệm

Môi trường thực nghiệm đã được thiết lập với mục đích nghiên cứu và đánh giá hiệu suất của mô hình được đề xuất trong nghiên cứu này. Môi trường này bao gồm cả môi trường huấn luyện và môi trường triển khai, mỗi môi trường có những đặc điểm riêng như Bảng 4.2.

Bảng 4. 2: Môi trường huấn luyện và triển khai

Phần cứng	**Môi trường huấn luyện**	**Môi trường triển khai**
Hệ điều hành	Ubuntu 18.04 LTS	Windows 11
Bộ xử lý	Intel Xeon CPU @ 2.20 GHz	Intel(R) Core(TM) i3-1005G1 CPU @ 1.20GHz
GPU	Tesla T4 12GB	Không sử dụng GPU
RAM	16 GB	8 GB
Ổ cứng	256 GB SSD	512 GB SSD

4.3. Kết quả thực nghiệm

4.3.1. SVM

Mô hình SVM sẽ được thực nghiệm trên tập dữ liệu kiểm tra. Thông tin về kết quả phân loại từ mô hình bao gồm cả các siêu tham số tối ưu tốt nhất, được xác định thông qua quá trình tìm kiếm siêu tham số tốt nhất sử dụng phương pháp *GridSearchCV*. Các thông tin này đã được ghi chép trong Bảng 4.3.

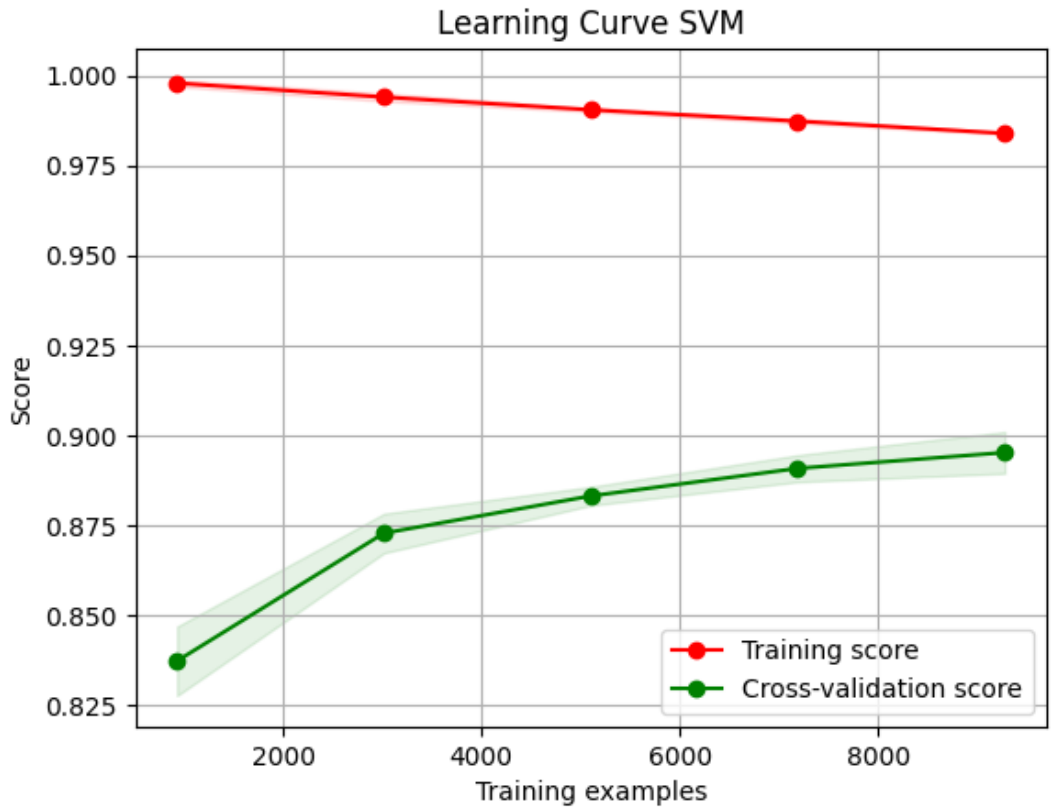
Bảng 4. 3: Bảng tổng hợp so sánh kết quả thực nghiệm với SVM

Trường hợp	**Kernel**	**C**	**Gamma**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
1	**RBF**	**10**	**1**	**91**	**91**	**91**	**91**
2	RBF	10	0.1	89	89	89	89
3	RBF	1	1	90	90	90	90

Kết quả cao nhất của thực nghiệm của mô hình SVM từ Bảng 4.3, khi sử dụng Kernel = RBF với các tham số C = 10 và Gamma = 1, có thể được giải thích vì Kernel = RBF được chọn vì khả năng linh hoạt trong xử lý cả dữ liệu tuyến tính và phi tuyến tính thông qua ánh xạ dữ liệu vào không gian nhiều chiều, từ đó tạo ra đường biên phân loại phức tạp hơn. Tham số C đóng vai trò quan trọng trong việc kiểm soát sự phức tạp

của đường biên quyết định, trong khi tham số Gamma điều chỉnh phạm vi ảnh hưởng của mỗi điểm dữ liệu. Sự kết hợp đã góp phần tạo ra hiệu suất tối ưu cho mô hình SVM trong việc phân loại dữ liệu.

Biểu đồ ở Hình 4.1 thể hiện độ chính xác trên tập dữ liệu huấn luyện và tập dữ liệu đánh giá của mô hình SVM với số lượng mẫu huấn luyện khác nhau. Độ chính xác của mô hình tăng lên khi số lượng mẫu huấn luyện tăng lên. Điều này cho thấy rằng mô hình đã đạt tới khả năng học tập tối đa của mình với dữ liệu huấn luyện hiện có.



Hình 4. 1: Đồ thị học tập của mô hình SVM

Kết quả kiểm tra của mô hình SVM ứng với tập dữ liệu kiểm tra được thể hiện chi tiết trong Bảng 4.4.

Bảng 4. 4: Kết quả kiểm chứng bộ phân lớp bằng thuật toán SVM

Chủ đề	SVM			
	Precision (%)	**Recall (%)**	**F1 Score (%)**	**Accuracy (%)**
Chính trị Xã hội	92	89	91	
Dự báo thời tiết	99	100	99	
Kinh tế	90	91	91	
Môi trường	87	90	89	
Nông nghiệp	94	92	93	
Pháp luật	86	90	88	91
Sức khỏe	90	88	89	
Thế giới	92	88	90	
Thể thao	96	97	96	
Văn hóa	83	82	83	
Giáo dục	89	90	89	
Tổng cộng	**91**	**91**	**91**	**91**

4.3.2. KNN

Mô hình KNN sẽ được thực nghiệm trên tập dữ liệu kiểm tra. Kết quả về phân loại từ mô hình bao gồm việc xác định các siêu tham số tối ưu nhất, được tìm kiếm thông qua quá trình *GridSearchCV* để tối ưu hóa hiệu suất mô hình. Các thông tin chi tiết về kết quả này đã được ghi nhận và ghi lại trong Bảng 4.5.

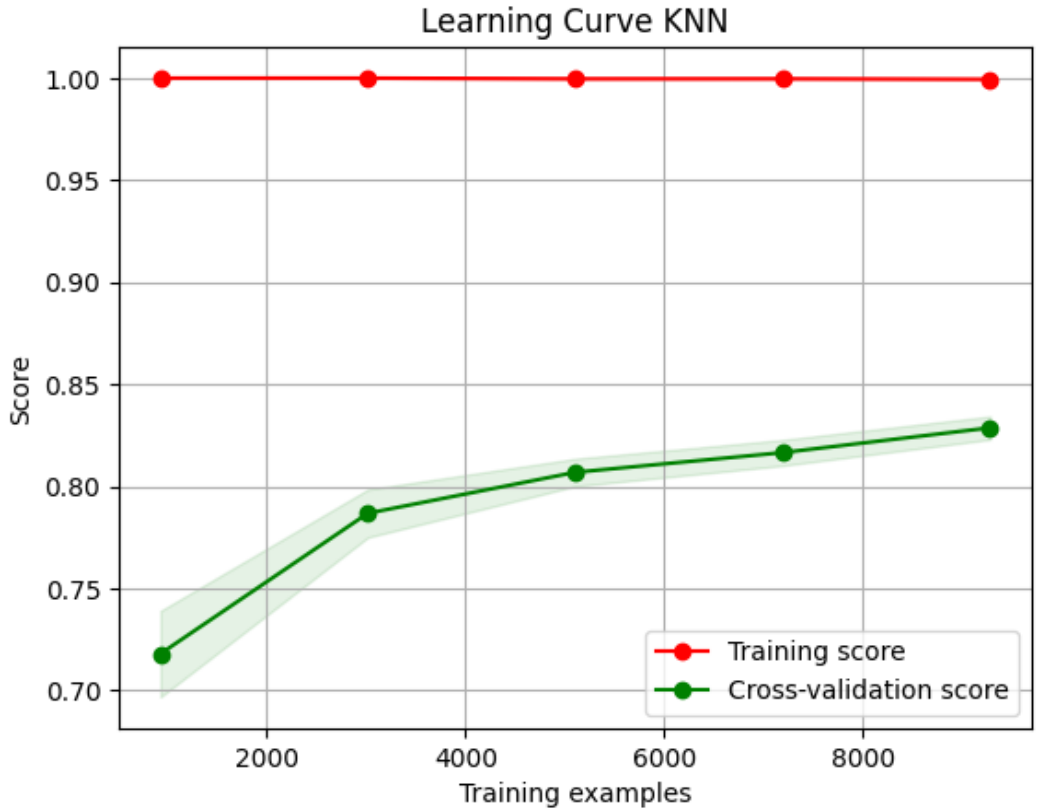
Bảng 4. 5: Bảng tổng hợp so sánh kết quả thực nghiệm với KNN

Trường hợp	K	Weights	P	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	**5**	**distance**	**1**	**83**	**83**	**83**	**83**
2	5	distance	2	82	82	82	82
3	7	distance	1	82	82	82	82

Kết quả thực nghiệm cho thấy mô hình KNN từ Bảng 4.4 khi sử dụng các tham số K = 5, Weights = distance và P = 1 đạt được kết quả cao nhất. Việc chọn tham số K = 5 có thể cân bằng giữa việc giảm nhiễu và giữ lại tính tổng quát của mô hình. Tham số Weights dựa trên khoảng cách giữa điểm dữ liệu đang xét và các điểm lân cận. Các điểm

gần sẽ có ảnh hưởng lớn hơn đối với quyết định phân loại. Tham số P = 1 ảnh hưởng đến cách tính toán khoảng cách giữa các điểm dữ liệu. Sự kết hợp đã góp phần tạo ra hiệu suất tối ưu cho mô hình KNN trong việc phân loại dữ liệu.

Biểu đồ trong Hình 4.2 biểu thị sự chính xác trên tập dữ liệu huấn luyện và tập dữ liệu xác thực của mô hình KNN với biến thiên về số lượng mẫu huấn luyện. Theo dõi đường biểu diễn, có thể nhận thấy rằng độ chính xác của mô hình gia tăng đồng đều khi số lượng mẫu huấn luyện tăng. Kết quả này chỉ ra rằng mô hình đã đạt đến khả năng học tập tối ưu của nó khi đối mặt với lượng dữ liệu huấn luyện hiện thời.



Hình 4. 2: Đồ thị học tập của mô hình KNN

Kết quả kiểm tra của mô hình KNN với tập dữ liệu kiểm tra được thể hiện trong Bảng 4.6.

Bảng 4. 6: Kết quả kiểm chứng bộ phân lớp bằng thuật toán KNN

Chủ đề	KNN			
	Precision (%)	**Recall (%)**	**F1 Score (%)**	**Accuracy (%)**
Chính trị Xã hội	86	72	78	
Dự báo thời tiết	98	99	98	
Kinh tế	87	85	86	
Môi trường	82	78	80	
Nông nghiệp	89	84	86	
Pháp luật	81	81	81	83
Sức khỏe	87	81	84	
Thế giới	78	81	79	
Thể thao	85	93	89	
Văn hóa	61	73	66	
Giáo dục	79	85	82	
Tổng cộng	**83**	**83**	**83**	**83**

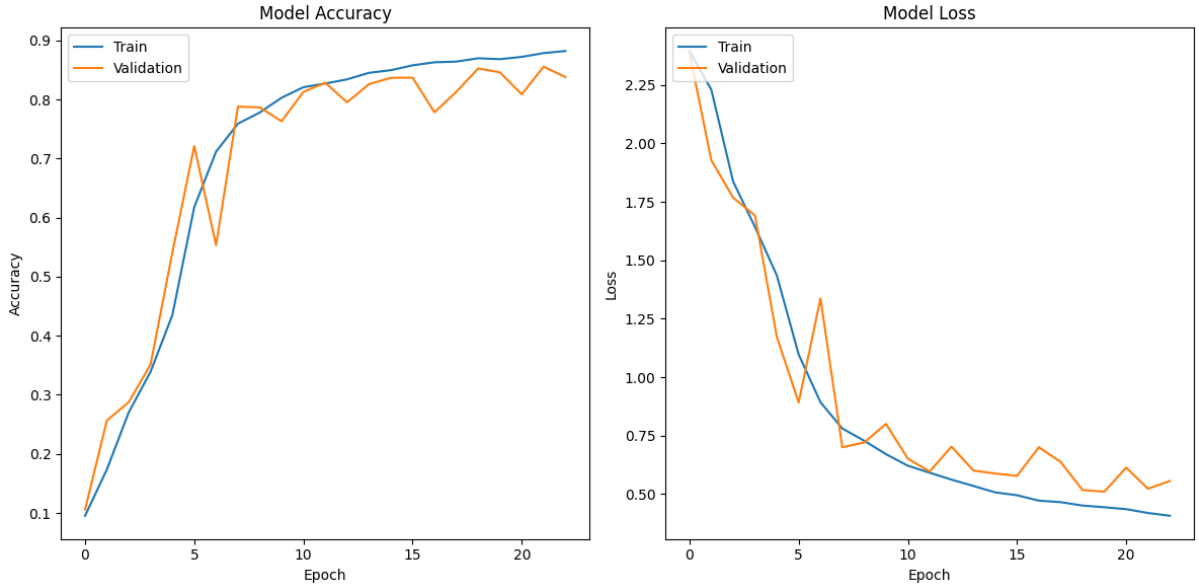
4.3.3. CNN

Trong quá trình huấn luyện mô hình CNN, giá trị Batchsize được lựa chọn là 32 để định lượng số lượng mẫu dữ liệu được sử dụng mỗi lần cập nhật trọng số mạng. Epochs được điều chỉnh tự động thông qua kỹ thuật *Early Stopping*, dừng quá trình huấn luyện khi hiệu suất tỷ lệ lỗi trên tập dữ liệu đánh giá không còn giảm. Điều chỉnh các tham số trong mô hình đã dẫn đến sự thay đổi đáng kể về tốc độ học, độ chính xác và thời gian huấn luyện của mô hình. Bảng 4.7 mô tả kết quả thực khi thay đổi Epoch qua các trường hợp.

Bảng 4. 7: Bảng tổng hợp so sánh kết quả thực nghiệm với CNN

Trường hợp	**Batchsize**	**Optimizer**	**Epoch**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
1	32	RMSprop	12	86	86	86	86
2	32	Adam	12	88	88	88	88
3	32	Nadam	7	88	88	88	88
4	**32**	**SGD**	**23**	**89**	**89**	**89**	**89**

Kết quả thực nghiệm mô hình CNN từ Bảng 4.7 đạt kết quả cao nhất 89% với các tham số Batchsize = 32, Optimizer = SGD và Epoch = 23. Biểu đồ ở Hình 4.3 thể hiện tỷ lệ lỗi và độ chính xác trên tập dữ liệu huấn luyện và tập dữ liệu xác thực đánh giá hiệu suất phân loại với số Epoch = 23 cho thấy các thay đổi trong quá trình thực nghiệm.



Hình 4. 3: Tỉ lệ lỗi và độ chính xác của mô hình mạng CNN với Epoch = 23

Kết quả kiểm tra của mô hình CNN với tập dữ liệu kiểm tra được thể hiện trong Bảng 4.8.

Bảng 4. 8: Kết quả kiểm chứng bộ phân lớp bằng mạng CNN

Chủ đề	CNN			
	Precision (%)	**Recall (%)**	**F1 Score (%)**	**Accuracy (%)**
Chính trị Xã hội	91	88	90	
Dự báo thời tiết	99	100	99	
Kinh tế	88	91	90	
Môi trường	84	89	87	
Nông nghiệp	90	92	91	
Pháp luật	84	92	88	89
Sức khỏe	85	85	85	
Thế giới	90	86	88	
Thể thao	92	96	94	
Văn hóa	88	72	79	
Giáo dục	88	88	88	
Tổng cộng	**89**	**89**	**89**	**89**

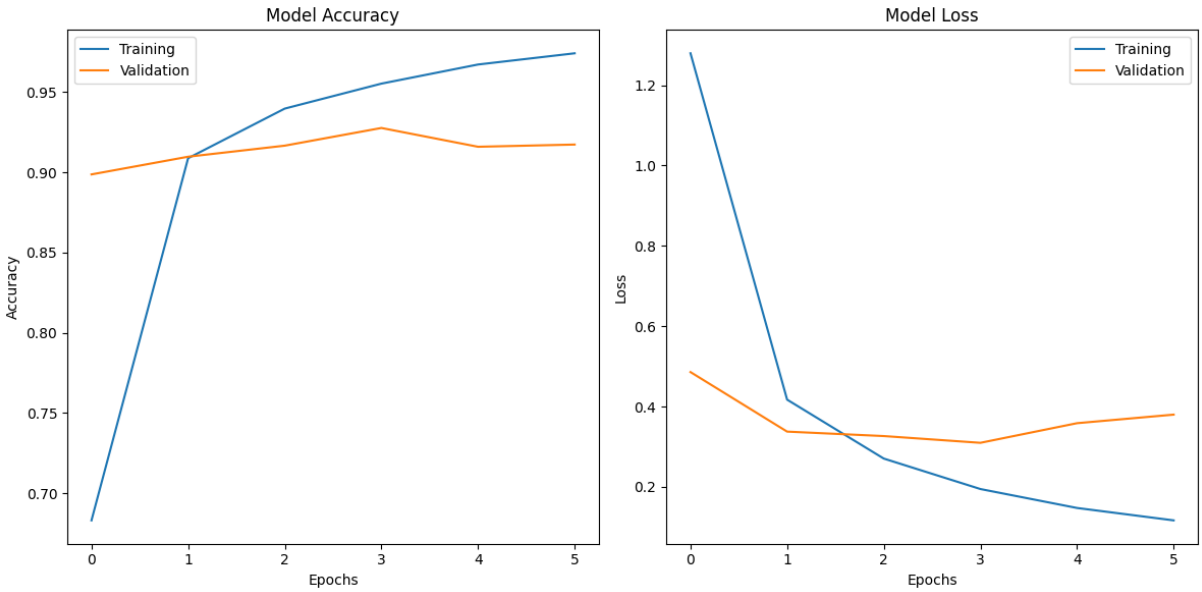
4.3.4. PhoBERT

Với mô hình PhoBERT thực nghiệm với tập dữ liệu kiểm tra. Sử dụng thuật toán tối ưu hóa AdamW được đề xuất bởi các nhà nghiên cứu [27]. Số Epochs được xác định bằng kỹ thuật *Early Stopping*, dừng quá trình huấn luyện khi tỷ lệ lỗi trên tập dữ liệu đánh giá không còn giảm. Với các siêu tham số được cấu hình trong quá trình huấn luyện. Điều chỉnh các tham số trong mô hình đã dẫn đến sự thay đổi đáng kể về tốc độ học, độ chính xác và thời gian huấn luyện của mô hình. Bảng 4.9 mô tả kết quả thực nghiệm khi thay đổi các tham số qua các trường hợp.

Bảng 4. 9: Bảng tổng hợp so sánh kết quả thực nghiệm với PhoBERT

Trường hợp	Learning Rate	Maxlength	Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1	**5e-5**	**256**	**5**	**98**	**98**	**98**	**98**
2	2e-5	128	7	92	92	92	92
3	3e-5	256	6	94	94	94	94

Kết quả thực nghiệm mô hình PhoBERT từ Bảng 4.9 đạt kết quả cao nhất 98% với các tham số Learning Rate = 5e-5, Maxlength = 256 và Epoch = 5. Biểu đồ ở Hình 4.4 thể hiện tỉ lệ lỗi và độ chính xác trên tập dữ liệu huấn luyện và tập dữ liệu xác thực đánh giá hiệu suất phân loại với số Epoch = 5 cho thấy các thay đổi trong quá trình thực nghiệm.



Hình 4. 4: Tỉ lệ lỗi và độ chính xác của mô hình PhoBERT với Epoch = 5

Kết quả kiểm tra của mô hình PhoBERT với tập dữ liệu kiểm tra được thể hiện trong Bảng 4.10.

Bảng 4. 10: Kết quả kiểm chứng bộ phân lớp bằng mô hình PhoBERT

Chủ đề	PhoBERT			
	Precision (%)	**Recall (%)**	**F1 Score (%)**	**Accuracy (%)**
Chính trị Xã hội	99	97	98	
Dự báo thời tiết	100	100	100	
Kinh tế	98	98	98	
Môi trường	97	96	97	
Nông nghiệp	99	99	99	
Pháp luật	96	95	96	98
Sức khỏe	98	98	98	
Thế giới	96	98	97	
Thể thao	100	99	99	
Văn hóa	95	96	95	
Giáo dục	98	98	98	
Tổng cộng	**98**	**98**	**98**	**98**

4.4. So sánh kết quả

Dựa trên các kết quả thực nghiệm từ các mô hình đề xuất và kết hợp với các kỹ thuật rút trích TF-IDF, giảm chiều dữ liệu SVD, tìm kiếm siêu tham số *GirdSearchCV* và kỹ thuật *Early Stopping* dừng quá trình huấn luyện khi hiệu suất không cải thiện. Kết quả thực nghiệm thể hiện qua các tiêu chí đánh giá *Accuracy, Precision, Recall, F1-Score* được thể hiện qua Bảng 4.11.

Bảng 4. 11: Kết quả thực nghiệm tốt nhất của các mô hình

Mô hình	Tiêu chí			
	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
SVM	91	91	91	91
KNN	93	83	83	83
CNN	89	89	89	89
PhoBERT	**98**	**98**	**98**	**98**

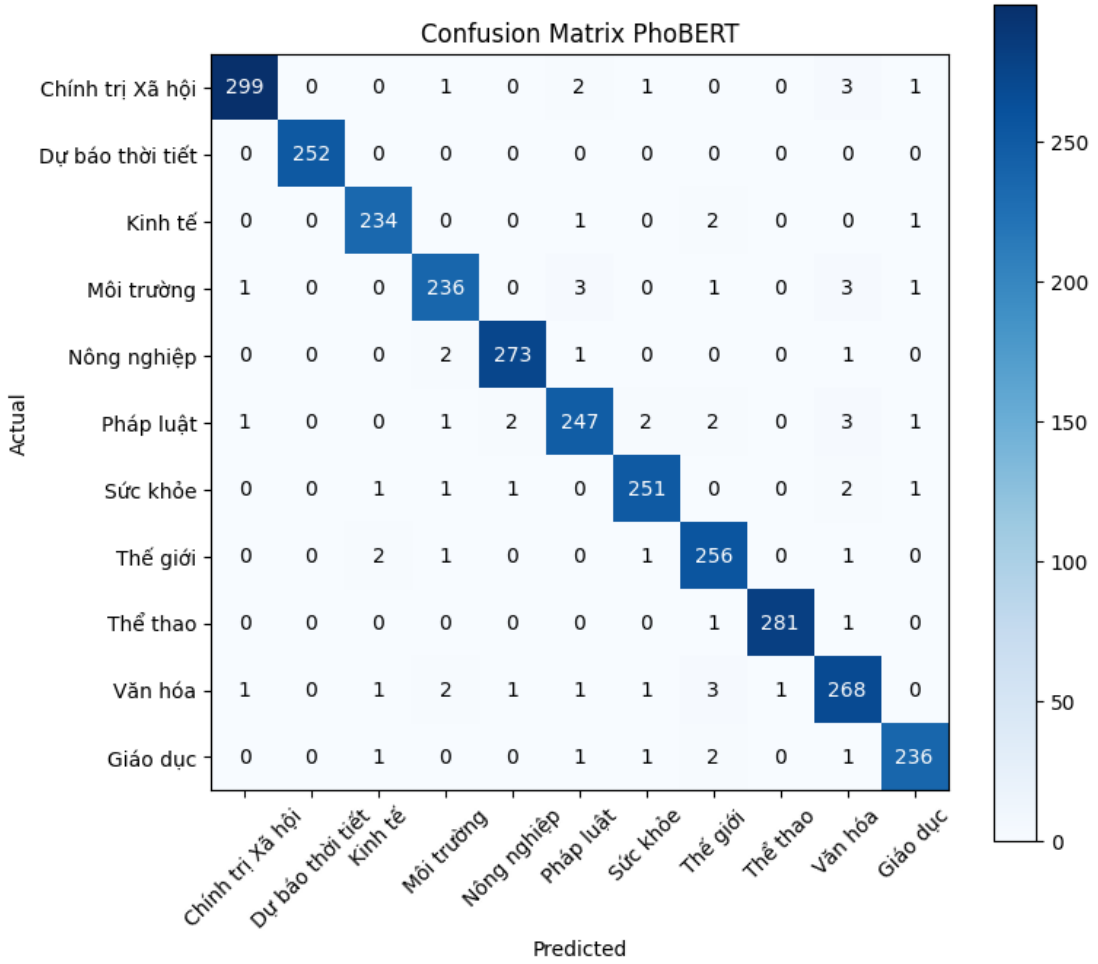
Kết quả thực nghiệm trong nghiên cứu này đã cung cấp một cái nhìn rõ ràng về hiệu suất của 4 phương pháp phân loại khác nhau. Kết quả chứng minh rằng mô hình PhoBERT có hiệu suất xuất sắc hơn 98% so với các phương pháp truyền thống SVM, KNN và CNN trong việc phân loại dữ liệu.

Mô hình ngôn ngữ PhoBERT đã đạt được tỷ lệ phân loại tốt nhất với tỷ lệ đúng 98%. Sự ưu việt này có thể được giải thích bằng việc nó đã được huấn luyện trên một lượng lớn dữ liệu văn bản, bao gồm nguồn tri thức văn bản khổng lồ. Quá trình đào tạo trên dữ liệu lớn này đã giúp mô hình PhoBERT nắm bắt được ngữ cảnh và ngữ nghĩa của các từ và câu một cách chi tiết và tổng quan hơn. Điều này đã dẫn đến khả năng phân loại vượt trội và hiệu suất tốt hơn.

Mặt khác, mô hình SVM và CNN đã đạt được tỷ lệ phân loại xấp xỉ nhau lần lượt là 91% và 89%. Sự khác biệt trong hiệu suất này có thể được giải thích bằng việc hai mô hình này đã được huấn luyện trên một tập dữ liệu nhỏ hơn và chưa có sự hỗ trợ từ một nguồn tri thức văn bản đa dạng. Các đặc trưng được học tập bởi SVM và CNN cũng có tính cục bộ hơn, chỉ phản ánh thông tin trong tập dữ liệu huấn luyện cụ thể. Điều này có nghĩa rằng khi gặp các đặc trưng mới hoặc dữ liệu phức tạp hơn, khả năng biểu diễn của hai mô hình này có thể bị hạn chế và dẫn đến sự giảm sút trong khả năng phân loại.

Mô hình KNN trong nghiên cứu này đạt tỷ lệ phân loại thấp nhất, chỉ 83%. Hiệu suất kém này có thể giải thích bằng việc KNN xác định lớp của một điểm dữ liệu thông qua việc sử dụng các điểm gần kề trong không gian đặc trưng, dự đoán phụ thuộc vào các điểm dữ liệu lân cận đòi hỏi tính toán khoảng cách cao, dễ bị ảnh hưởng bởi nhiễu, không tự động học đặc trưng quan trọng và cần sự lựa chọn cẩn thận về giá trị K. Điều này dẫn đến khả năng phân loại kém khi gặp phải dữ liệu mới hoặc dữ liệu phức tạp.

Để chứng minh hiệu suất của mô hình PhoBERT so với các mô hình còn lại, tiến hành phân tích ma trận nhầm lẫn như được biểu diễn trong Hình 4.3. Ma trận nhầm lẫn cung cấp chi tiết về sự nhầm lẫn giữa các chủ đề thực tế và các chủ đề được mô hình PhoBERT dự đoán trên tập dữ liệu kiểm tra. Kết quả cho thấy tỉ lệ dự đoán sai sót của mô hình PhoBERT là rất thấp, chứng tỏ khả năng hiệu suất xuất sắc của mô hình này.



Hình 4. 5: Ma trận nhầm lẫn ứng với mô hình PhoBERT

Có thể giải thích sự nhầm lẫn trong dự đoán bằng việc xem xét sự nhiễu trong quá trình gán nhãn trên tập dữ liệu. Trong nhiều trường hợp, tập dữ liệu huấn luyện có thể chứa các thông tin không chính xác hoặc mâu thuẫn về chủ đề, dẫn đến việc gán nhãn không chính xác. Điều này có thể dẫn đến sự nhầm lẫn khi mô hình được áp dụng cho các dữ liệu kiểm tra mới. Tuy nhiên, mô hình PhoBERT vẫn thể hiện khả năng chính xác cao trong việc phân loại chủ đề, cho thấy tính ổn định và đáng tin cậy của nó trong ứng dụng thực tế.

CHƯƠNG V: KẾT LUẬN

Sau quá trình tìm hiểu và thực hiện đề tài đã trình bày những tổng kết của quá trình thực nghiệm, bao gồm nhìn nhận kết quả các công việc đã làm được, các hạn chế của nghiên cứu này và đề xuất hướng phát triển tiềm năng cho tương lai.

5.1. Kết quả đạt được

Nghiên cứu tổng quan về vấn đề và xây dựng một sơ đồ hệ thống tổng quát để giải quyết bài toán phân loại chủ đề truyền hình thời sự.

Thực hiện một nghiên cứu chi tiết về kiến trúc và thành phần của các phương pháp tiếp cận bài toán phân loại, bao gồm cả đặc trưng TF-IDF, thuật toán SVM và KNN, mạng CNN và mô hình ngôn ngữ PhoBERT.

Xây dựng hoàn thành bộ dữ liệu thử nghiệm để giải quyết bài toán phân loại chủ đề truyền hình thời sự gồm 11 chủ đề thuộc quốc gia Việt Nam.

Giải quyết bài toán phân loại bằng cách áp dụng nhiều phương pháp khác nhau và sau đó so sánh chất lượng của các mô hình theo các tiêu chí đánh giá.

Lựa chọn mô hình tốt nhất dựa trên các tiêu chí đánh giá và triển khai hệ thống phân loại tự động trên nền tảng trang web để tương tác với người dùng.

Kết quả cuối cùng của nghiên cứu đã chứng minh rằng mô hình phân loại tự động đã phát triển có khả năng phân loại các bài viết trên bảng tin thời sự truyền hình với độ chính xác cao. Điều này đã thể hiện sự tiềm năng của mô hình trong việc tự động hóa quá trình phân loại tin tức.

5.2. Hạn chế

Trong quá trình thử nghiệm hệ thống thì nhận thấy được rằng vẫn còn một số hạn chế còn tồn động bao gồm:

Trong tập dữ liệu được sử dụng, đã ghi nhận sự xuất hiện của nhiễu dữ liệu, chủ yếu là do mâu thuẫn về chủ đề hoặc các thông tin không liên quan đến nội dung bảng tin thời sự truyền hình.

Một số chính sách bảo mật và vấn đề liên quan đến phân quyền truy cập đã làm hạn chế quyền truy cập và sử dụng dữ liệu từ Youtube. Điều này có thể đã ảnh hưởng đến tính đa dạng của dữ liệu và khả năng phát triển mô hình.

Thời gian xử lý dữ liệu trong quá trình thử nghiệm phụ thuộc vào tốc độ đường truyền internet và độ phức tạp của dữ liệu. Việc xử lý video và văn bản truyền hình có thể đòi hỏi nguồn tài nguyên tính toán lớn và thời gian xử lý khá dài, điều này làm giảm hiệu suất hệ thống.

5.3. Hướng phát triển

Để cải thiện độ chính xác và khả năng tổng quát hóa tập trung bổ sung dữ liệu chất lượng. Điều này đồng nghĩa với việc thu thập dữ liệu từ nhiều nguồn và nguồn tài liệu đáng tin cậy để đảm bảo tính đúng đắn và đa dạng khả năng tổng quát hóa.

Xử lý trực tiếp dữ liệu video mà không cần qua quá trình chuyển đổi. Quá trình chuyển đổi có thể dẫn đến mất mát thông tin và làm giảm độ chính xác. Việc phát triển các thuật toán và mô hình có khả năng làm việc trực tiếp với dữ liệu video gốc có thể là một bước quan trọng để cải thiện hiệu suất.

Nghiên cứu cũng có thể mở rộng để xử lý thông tin truyền hình theo thời gian thực. Ví dụ như phân loại sự kiện và tin tức trực tiếp khi chúng xảy ra. Điều này đòi hỏi phải phát triển các phương pháp xử lý và nhận dạng thời gian thực phức tạp hơn, có khả năng hoạt động nhanh chóng và hiệu quả trong thời gian thực.

Khai thác tối đa sức mạnh của các mô hình ngôn ngữ lớn. Các mô hình này có khả năng hiểu và tự động trích xuất thông tin từ văn bản, giúp tăng cường khả năng xử lý dữ liệu văn bản kèm theo trong các video.

Tích hợp hệ thống đa nền tảng có thể hỗ trợ khai thác khả năng phân loại của mô hình cung cấp một cái nhìn toàn diện hơn về nội dung video và cải thiện khả năng phát hiện và phân loại sự kiện, tin tức, hoặc nội dung khác một cách chính xác

TÀI LIỆU THAM KHẢO

[1] Z. Li, W. Shang and M. Yan, "News text classification model based on topic model," *IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS),* pp. 1-5, 2016.

[2] S. Ahmed, K. Hinkelmann and F. Corradini, "Development of Fake News Model Using Machine Learning through Natural Language Processing," *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering,* 2020.

[3] M. Manzato and R. Goularte, "Video news classification for automatic content personalization: A genetic algorithm based approach," 2008.

[4] Y. Gao, "News Video Classification Model Based on ResNet-2 and Transfer Learning," *Security and Communication Networks,* 2021.

[5] V. Q. Long, N. N. H. Anh, T. M. Sơn và T. T. Hà, "Phân Loại Tên Chương Trình Truyền Hình Theo Chủ Đề Phát sóng Sử Dụng Mô Hình XLNet," *JOURNAL OF TECHNICALEDUCATION SCIENCE,* pp. 8-16, 2022.

[6] H. Dũng và B. M. Hùng, Giáo trình Dẫn luận ngôn ngữ học, Hà Nội: Nhà xuất bản Đại học Sư phạm, 2007.

[7] T. V. Sáng, Dẫn luận ngôn ngữ học, Đà Nẵng: Trường Đại Học Sư Phạm - Đại học Đà Nẵng, 2019.

[8] J. Eisenstein, Natural Language Processing, 2018.

[9] A. Geitgey, "Natural Language Processing is Fun!," 2018. [Online]. Available: https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e. [Đã truy cập 15 10 2023].

[10] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition, Prentice-Hall 2nd edition, 2009.

[11] C. Tapsai, P. Meesad and C. Haruechaiyasak, "LS-ART: Thai Language Segmentation by AutomaticRanking Trie," *Conference: The 9th International Conference Autonomous Systems,* 2016.

[12] V. Jain and A. Chadha, "Neural Nets," *Distilled Notes for Stanford CS224n: Natural Language Processing with Deep Learning,* p. https://aman.ai/, 2021.

[13] G. Kaur and G. Singh, "Importance of Natural Language Processing, Its Features, Components and Applications," *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE),* vol. 6, no. 10, 2019.

[14] C. Dilmegani, "Complete Guide to NLP in 2023: How It Works & Top Use Cases," 12 10 2023. [Online]. Available: https://research.aimultiple.com/nlp/. [Accessed 31 10 2023].

[15] I. Roldós, "Major Challenges of Natural Language Processing (NLP)," 22 12 2020. [Online]. Available: https://monkeylearn.com/blog/natural-language-processing-challenges/. [Accessed 31 10 2023].

[16] L. T. M. Hồng, Phân loại văn bản bằng phương pháp Support vector machine, Hà Nội: Trường Đại học Bách Khoa Hà Nội, 2006.

[17] L. T. Huy, Nghiên cứu cải tiến một số phương pháp phân loại văn bản tự động và áp dụng trong xử lý văn bản tiếng Việt, Hà Nội: Trường Đại học Công nghệ - Hà Nội, 2008.

[18] N. L. M. Hòa và T. V. Lăng, "GOM CỤM BÀI BÁO THEO CHỦ ĐỀ," *HUFLIT Journal of Science,* p. https://hjs.huflit.edu.vn/index.php/hjs/article/view/147, 2023.

[19] Đ. V. Nam, "Nghiên cứu mô hình học máy Naïve Bayes trong phân lớp văn bản; Ứng dụng phân lớp cho tập dữ liệu các nhận xét trên Twitter," *HỘI NGHỊ TOÀN QUỐC KHOA HỌC TRÁI ĐẤT VÀ TÀI NGUYÊN VỚI PHÁT TRIỂN BỀN VỮNG (ERSD 2020),* p. https://qlkh.humg.edu.vn/CongBo/Download/4413?FileName=ERSD2020_DangVanNam_NLP.pdf, 2020.

[20] V. Vapnik, The Nature of Statistical Learning Theory, New York, 1999.

[21] S. Bandgar, "https://medium.com/," Support Vector Machine., 13 11 2021. [Online]. Available: https://medium.com/nerd-for-tech/support-vector-machine-92fa3c57d33b.

[22] V. H. Tiệp, Machine Learning cơ bản, Nhà Xuất Bản Khoa Học Và Kỹ Thuật, 2018.

[23] G. Singh, "linkedin.com," k-Nearest Neighbors, 10 6 2020. [Online]. Available: https://www.linkedin.com/pulse/k-nearest-neighbors-gauransh-singh/.

[24] T. Wood, "DeepAI," 15 10 2023. [Trực tuyến]. Available: https://deepai.org/machine-learning-glossary-and-terms/convolutional-neural-network.

[25] N. T. Tuấn, Deep Learning cơ bản, 2020.

[26] D. j. Sharma, S. Dutta and D. D. j. Bora, "REGA: Real-Time Emotion , Gender , Age Detection Using CNN – A Review," *Conference: The International Conference on Research in Management & Technovation,* 2020.

[27] N. D. Quoc and N. A. Tuan, "PhoBERT: Pre-trained language models for Vietnames," *arxiv,* 2020.

[28] N. Ficano, "pytube Documentation," 20 5 2023. [Online]. Available: https://pytube.io/_/downloads/en/latest/pdf/.

[29] J. Robert, "pydub.com," 2021. [Trực tuyến]. Available: https://github.com/jiaaro/pydub/blob/master/API.markdown.

[30] D. Amos, "The Ultimate Guide To Speech Recognition With Python," 21 3 2018. [Trực tuyến]. Available: https://realpython.com/python-speech-recognition/.

[31] M. Inman, "A Comparison of Speech Recognition Libraries for Python," IEEE Signal Processing Letters, vol. 23, no. 11, pp. 1589-1592, 2016.

[32] B. Slatkin, Effective Python: 90 Specific Ways to Write Better Python, 2015.

[33] V. T. Trần, “Python Vietnamese Core NLP Toolkit,” 2017. [Trực tuyến]. Available: https://github.com/trungtv/pyvi.d

[34] G. V. A. G. V. M. B. T. O. G. M. B. P. P. R. W. V. D. J. V. A. P. D. C. M. B. M. P. François Pedregosa, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research 12*, pp. 2825-2830, 2011.

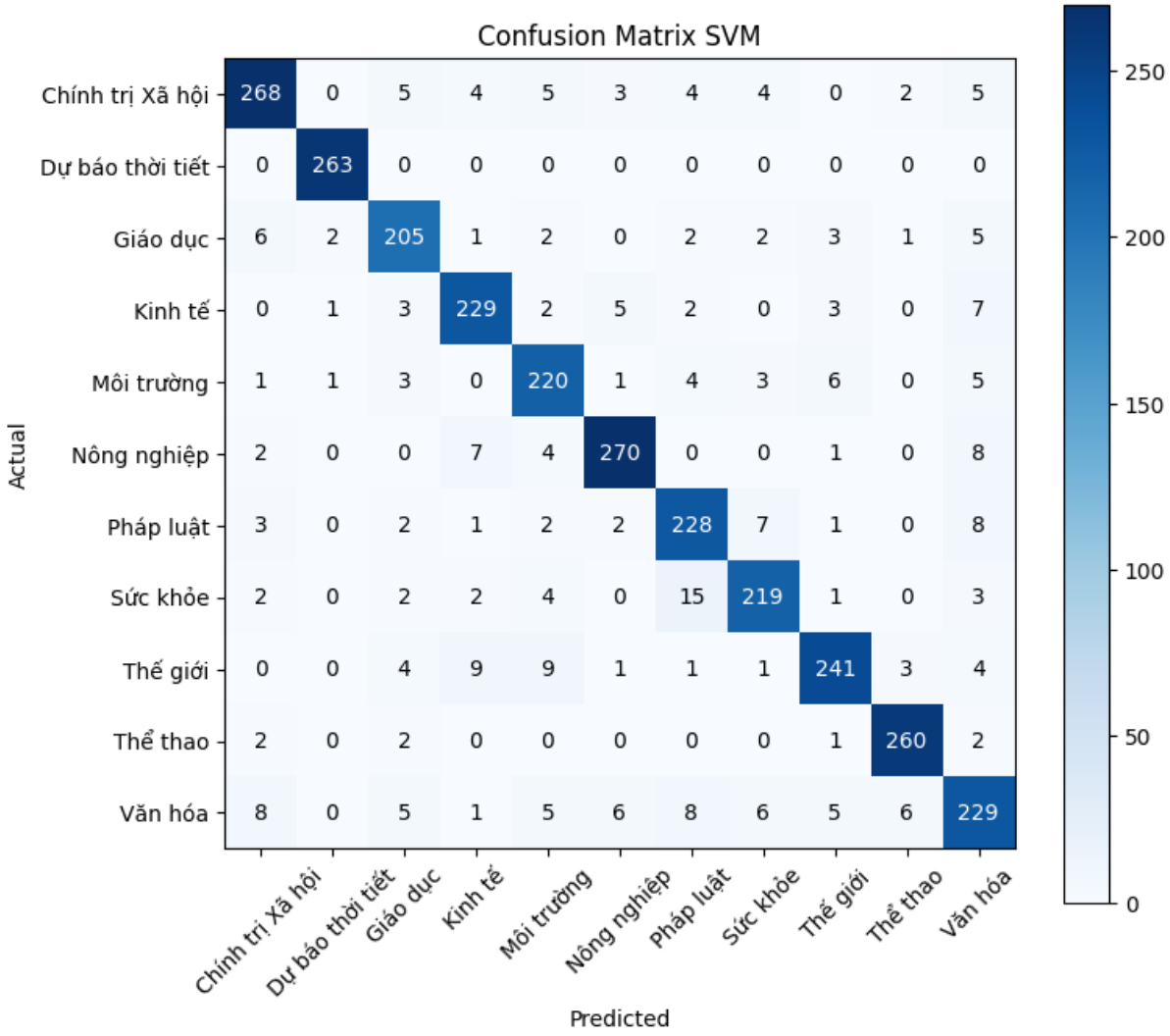
[35] B. Pang, E. Nijkamp and Y. N. Wu, "Deep Learning With TensorFlow: A Review," *Journal of Educational and Behavioral Statistics*, pp. 227-248, 2020.

[36] E. Stevens, L. Antiga and T. Viehmann, Deep Learning with PyTorch, Manning Publications, 2020.

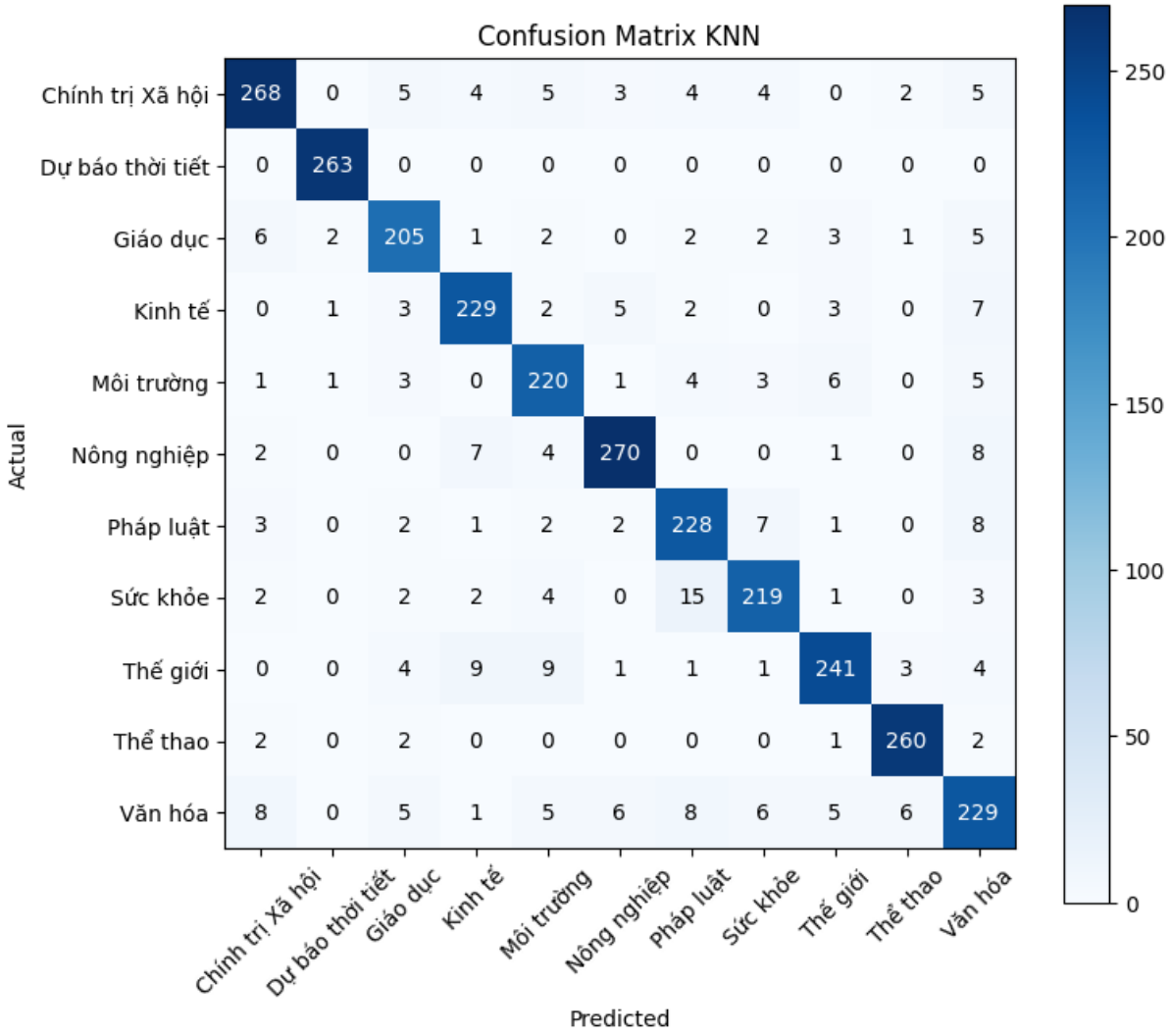
[37] M. Khorasani, M. Abdou and J. H. Fernández, Web Application Development with Streamlit, Berkeley, 2022.

PHỤ LỤC

Phụ lục 1: Ma trận nhầm lẫn ứng với mô hình SVM



Phụ lục 2: Ma trận nhầm lẫn ứng với mô hình KNN



Phụ lục 3: Ma trận nhầm lẫn ứng với mô hình CNN