BỘ KHOA HỌC VÀ CÔNG NGHỆ

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



LÊ NGỌC AN

PHÁT HIỆN TIN BÀI ĐỘC HẠI TIẾNG VIỆT SỬ DỤNG KẾT HỢP ĐẶC TRƯNG VĂN BẢN VÀ HÌNH ẢNH

CHUYÊN NGÀNH : Hệ THỐNG THÔNG TIN

MÃ SỐ: 9.48.01.04

LUẬN ÁN TIẾN SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC

1. PGS.TS. HOÀNG XUÂN DẬU

2. TS. ĐINH TRƯỜNG DUY

HÀ NỘI 2026

Công trình được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học:

1. PGS. TS. Hoàng Xuân Dậu
2. TS. Đinh Trường Duy

PHẦN MỞ ĐẦU

1. TÍNH CẤP THIẾT CỦA LUẬN ÁN

Từ bối cảnh thực tiễn với hơn 77 triệu người dùng Internet tại Việt Nam theo báo cáo Digital 2023, không gian mạng bùng nổ thông tin nhưng cũng đối mặt với sự lan truyền chóng mặt của các loại tin bài độc hại mang tính xuyên tạc, vu khống hay kích động. Những nội dung này đe dọa trực tiếp đến trật tự xã hội và an ninh quốc gia, mà điển hình là vụ việc kích động truyền thông xung quanh sự cố Đắk Lắk năm 2023.

Bài toán phát hiện tin bài độc hại hiện nay đang phải đối mặt với nhiều thách thức cốt lõi. Đầu tiên là sự xuất hiện của dữ liệu đa phương thức phức tạp khi thông tin độc hại được kết hợp tinh vi giữa văn bản và hình ảnh hay video, vượt quá khả năng xử lý của các hệ thống giám sát văn bản thuần túy. Bên cạnh đó, việc thiếu dữ liệu gán nhãn cũng tạo ra rào cản lớn do quy trình thu thập và dán nhãn thủ công cho các tập dữ liệu lớn đòi hỏi rất nhiều thời gian cùng chi phí. Ngoài ra, đặc thù tiếng Việt với cấu trúc đa dạng, linh hoạt kết hợp với tính chất phân loại đa lớp phức tạp của tin độc hại khiến bài toán càng trở nên khắt khe.

Trước thực trạng đó, việc kết hợp giữa học sâu nhằm trích xuất đặc trưng đa phương thức và học bán giám sát để tận dụng nguồn dữ liệu chưa gán nhãn phong phú được xem là một hướng đi vô cùng tiềm năng. Do hiện chưa có nghiên cứu chuyên biệt nào áp dụng đồng thời hai kỹ thuật này cho bài toán phát hiện tin độc hại tiếng Việt, việc triển khai đề tài mang ý nghĩa đặc biệt cấp thiết trên cả phương diện lý luận lẫn thực tiễn.

2. MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU

Mục tiêu tổng quát của luận án là xây dựng mô hình trí tuệ nhân tạo phát hiện tin bài độc hại tiếng Việt dựa trên sự kết hợp đặc trưng văn bản và hình ảnh, đồng thời ứng dụng học bán giám sát nhằm hỗ trợ công tác đảm bảo an ninh thông tin. Để đạt được mục tiêu này, luận án tiến hành hệ thống hóa lý thuyết liên quan, đề xuất mô hình tích hợp PhoBERT cho văn bản và Swin Transformer V2 cho hình ảnh phục vụ bài toán phân loại đa lớp, đồng thời áp dụng kỹ thuật gán nhãn giả nhằm khai thác hiệu quả nguồn dữ liệu chưa gán nhãn. Phạm vi nghiên cứu được giới hạn trong phương thức văn bản kết hợp hình ảnh bằng tiếng Việt, tập trung xử lý các

dạng tin bài mang tính chất chính trị như xuyên tạc, kích động hay vu khống, trên cơ sở kết hợp nguồn dữ liệu có nhãn và chưa gán nhãn thu thập từ các mạng xã hội.

3. CÁC ĐÓNG GÓP CỦA LUẬN ÁN

Đóng góp thứ nhất của luận án là đề xuất mô hình phát hiện tin bài độc hại tiếng Việt dựa trên sự kết hợp giữa các đặc trưng văn bản và hình ảnh. Mô hình đề xuất tích hợp PhoBERT cho xử lý dữ liệu văn bản và Swin Transformer V2 cho xử lý dữ liệu hình ảnh của tin bài, qua đó tận dụng sức mạnh của các mô hình học chuyển giao và học sâu trong việc khai thác thông tin ngữ nghĩa và thị giác.

Đóng góp thứ hai của luận án là đề xuất mô hình phát hiện tin bài độc hại tiếng Việt dựa trên phương pháp học bán giám sát. Phương pháp học bán giám sát được sử dụng để khắc phục vấn đề thiếu dữ liệu tin bài độc hại được gán nhãn. Thông qua cơ chế gán nhãn giả, mô hình có thể sử dụng một phần dữ liệu có nhãn và phần lớn dữ liệu chưa có nhãn thu thập từ nhiều nguồn trên mạng xã hội, từ đó cải thiện chất lượng huấn luyện.

4. BỐ CỤC CỦA LUẬN ÁN

Luận án gồm phần Mở đầu, Kết luận và 03 chương chính:

Chương 1: Tổng quan về phát hiện tin bài độc hại tiếng Việt.

Chương 2: Phát hiện tin bài độc hại tiếng Việt sử dụng kết hợp đặc trưng văn bản và hình ảnh.

Chương 3: Phát hiện tin bài độc hại tiếng Việt dựa trên học bán giám sát.

CHƯƠNG 1: TỔNG QUAN VỀ PHÁT HIỆN TIN BÀI ĐỘC HẠI TIẾNG VIỆT

1.1. KHÁI QUÁT VỀ TIN BÀI ĐỘC HẠI TIẾNG VIỆT

1.1.1. Giới thiệu tin bài độc hại và sự lan truyền trên không gian mạng

Tin bài độc hại là các nội dung sai lệch, xuyên tạc, vu khống, kích động hoặc thúc đẩy tư tưởng cực đoan, gây nhiễu loạn thông tin và ảnh hưởng tiêu cực đến nhận thức cộng đồng lẫn ổn định xã hội. Sự bùng nổ của mạng xã hội cùng các kỹ thuật giật tít, hình ảnh gây sốc đã khiến loại tin này lan truyền nhanh chóng, điển hình như các thông tin xuyên tạc sau sự cố tại Đắk Lắk năm 2023. Trong luận án này, "tin bài độc hại" được xác định là khái niệm bao trùm, phân biệt rõ với tin giả hay ngôn từ thù ghét đơn thuần. Nghiên cứu tập trung vào nhóm tin bài chính trị dạng văn bản kết hợp hình ảnh, được phân loại đa lớp thành 3 nhóm: "Vu khống, bôi nhọ", "Kêu gọi chống đối" và "Thúc đẩy tư tưởng cực đoan".

1.1.2. Tiêu chí nhận diện tin bài độc hại

Việc xác định phạm vi tin bài độc hại trong luận án dựa trên sự kết hợp giữa khung pháp lý và đặc trưng nội dung dữ liệu. Cơ sở pháp lý tham chiếu gồm Luật An ninh mạng 2025 (Điều 13) [14], Luật Báo chí 2025 [15], Nghị định 328/2026/NĐ-CP [16], Công văn số 7169/VPCP-NC về xử lý thông tin xuyên tạc, bôi nhọ lãnh đạo [17] và các định hướng đấu tranh phản bác trên Tạp chí Cộng sản [18]. Trên cơ sở đó, luận án xác định ba nhóm tiêu chí nhận diện chính cho bài toán phân loại đa lớp:

(1) Vu khống, bôi nhọ: Xúc phạm nghiêm trọng danh dự, uy tín cá nhân, tổ chức (đặc biệt là cán bộ, cơ quan Nhà nước); bịa đặt, cắt xén sự thật hoặc kết hợp hình ảnh/tài liệu không rõ nguồn gốc nhằm dẫn dắt sai lệch.

(2) Kêu gọi chống đối: Kích động, xúi giục, lôi kéo tụ tập đông người, biểu tình, gây rối trật tự hoặc cổ súy các hành vi phản kháng, chống lại chính quyền và người thi hành công vụ.

(3) Thúc đẩy tư tưởng cực đoan: Tuyên truyền xuyên tạc, phỉ báng chính quyền, lan truyền luận điệu chiến tranh tâm lý, gây chia rẽ dân tộc và thổi phồng tiêu cực nhằm tạo tâm lý hoang mang, khủng hoảng xã hội.

1.1.3. Đặc điểm ngôn ngữ và cấu trúc của tin bài độc hại dạng văn bản

Trong môi trường truyền thông số, văn bản độc hại chính trị đóng vai trò chủ đạo trong việc định hướng nhận thức và thúc đẩy sự lan truyền tiêu cực thông qua các đặc điểm ngôn ngữ, cấu trúc riêng biệt. Loại tin bài này thường khai thác lối diễn đạt giật gân, thiên lệch, kết hợp tiêu đề gây hiểu nhầm [19] [20] và lập luận thiếu logic dựa trên nguồn tin mơ hồ hoặc kỹ thuật "nửa thật nửa giả" [21] [22]. Về mặt kết cấu, chúng thường lặp lại thông điệp tuyên truyền có chủ đích [19] [20], đồng thời tận dụng tính đa nghĩa, thủ pháp tu từ hoặc biến tấu chính tả của tiếng Việt nhằm qua mặt kiểm duyệt [19]. Tất cả các đặc trưng này là cơ sở quan trọng giúp định hướng nghiên cứu và xây dựng các mô hình học máy, học sâu để phát hiện, phân loại tự động thông tin độc hại.

1.1.4. Đặc điểm và vai trò của hình ảnh trong tin bài độc hại

Bên cạnh văn bản, hình ảnh là thành tố trực quan quan trọng tạo nên sức lan tỏa và mức độ thuyết phục của tin bài độc hại trong môi trường truyền thông số. Loại nội dung này thường sử dụng kỹ thuật cắt ghép, chỉnh sửa hoặc lấy ảnh từ sự kiện khác gắn chú thích sai lệch [23] [24] nhằm đóng vai trò "chứng cứ trực quan" bổ trợ cho văn bản [25], từ đó kích thích mạnh mẽ cảm xúc phẫn nộ hay sợ hãi của người tiếp nhận [26]. Nhờ cơ chế ưu tiên hiển thị của thuật toán mạng xã hội, tin độc hại kèm hình ảnh dễ dàng gia tăng độ phủ sóng và tốc độ lan truyền [25]. Sự kết hợp đa phương thức này tạo ra thách thức lớn cho việc kiểm chứng, đặt ra yêu cầu tất yếu phải nghiên cứu đồng thời đặc trưng hình ảnh và văn bản để xây dựng các mô hình phát hiện tự động [27].

1.1.5. Sự cần thiết của phát hiện và ngăn chặn tin bài độc hại

Sự bùng nổ của tin bài độc hại trên Internet và mạng xã hội đang gây rối loạn môi trường thông tin, tạo sự hoang mang và đe dọa niềm tin xã hội. Do khối lượng dữ liệu khổng lồ và thủ đoạn phát tán ngày càng tinh vi, các biện pháp kiểm duyệt thủ công hay công cụ dựa trên từ khóa truyền thống đã không còn hiệu quả. Thực tế này đòi hỏi sự can thiệp của trí tuệ nhân tạo, đặc biệt là các mô hình học sâu và học bán giám sát có khả năng tự động xử lý dữ liệu đa phương thức phức tạp. Tuy nhiên, các nghiên cứu trong nước về phát hiện tin độc hại đa phương thức bằng tiếng Việt vẫn còn

khá khiêm tốn, tạo ra khoảng trống cấp thiết cần được giải quyết nhằm bảo đảm an toàn thông tin và nâng cao hiệu quả quản lý xã hội.

1.1.6. Các thách thức trong phát hiện tin bài độc hại tiếng Việt

Bài toán phát hiện tin bài độc hại tiếng Việt đối mặt với nhiều thách thức đặc thù, đến từ cả yếu tố dữ liệu, đặc điểm ngôn ngữ và sự ngụy trang nội dung ngày càng tinh vi. Rào cản lớn nhất hiện nay là sự khan hiếm của các tập dữ liệu đa dạng có nhãn đáng tin cậy, cùng với sự phức tạp của tiếng Việt về ngữ nghĩa, từ lóng và cách viết biến tấu trên môi trường mạng. Bên cạnh đó, sự kết hợp đa phương thức giữa văn bản và hình ảnh, biểu đồ hay video khiến các phương pháp xử lý đơn phương thức truyền thống trở nên kém hiệu quả. Sự thay đổi và biến động liên tục của các thủ đoạn phát tán đòi hỏi phải có những hướng tiếp cận mới như học sâu, học bán giám sát và các mô hình đa phương thức để nâng cao hiệu quả nhận diện thực tế.

1.2. CÁC PHƯƠNG PHÁP PHÁT HIỆN TIN BÀI ĐỘC HẠI

1.2.1. Giới thiệu

Các phương pháp phát hiện tin bài độc hại có thể chia thành ba nhóm: dựa trên từ khóa/danh sách, dựa trên luật và dựa trên học máy. Các phương pháp truyền thống tuy đơn giản, dễ triển khai nhưng phụ thuộc vào mẫu cố định nên khó thích ứng với ngôn ngữ biến đổi trên mạng xã hội. Ngược lại, các phương pháp học máy cho phép tự động học đặc trưng từ dữ liệu, trong đó học có giám sát đạt hiệu quả cao khi có đủ nhãn, còn học bán giám sát giúp tận dụng dữ liệu chưa gán nhãn để cải thiện khả năng phát hiện. Xu hướng hiện nay là chuyển sang các mô hình học máy có khả năng thích ứng và tổng quát hóa tốt hơn.

1.2.2. Phát hiện tin bài độc hại dựa trên học có giám sát

1.2.2.1. Các nghiên cứu trong nước

Nhiều nghiên cứu đã tiếp cận bài toán phát hiện nội dung độc hại và tin giả tiếng Việt bằng đa dạng phương pháp, từ học máy cổ điển như LGBM, TF-IDF [34] đến các mô hình học sâu, mô hình lai kết hợp đặc trưng (PhoBERT-CNN, BiLSTM, vELECTRA) và kỹ thuật kỹ nghệ đặc trưng, tăng cường dữ liệu [28], [29], [30], [31], [32], [33]. Trong đó, các hướng tiếp cận sử dụng mô hình ngôn ngữ tiền huấn luyện như PhoBERT [30],

[31], [32] thể hiện ưu thế vượt trội nhờ khả năng biểu diễn ngữ nghĩa mạnh mẽ. Tuy nhiên, các công trình hiện tại vẫn chủ yếu tập trung vào dữ liệu văn bản thuần túy hoặc siêu dữ liệu kèm theo mà chưa khai thác triệt để thông tin hình ảnh. Đồng thời, việc phụ thuộc lớn vào học có giám sát trên tập dữ liệu gán nhãn tốn kém cùng sự thiếu vắng các nghiên cứu chuyên sâu về bài toán phân loại đa lớp đối với nội dung nhạy cảm chính trị vẫn là những hạn chế lớn. Đây chính là khoảng trống tri thức then chốt mà luận án tập trung giải quyết.

1.2.2.2. Các nghiên cứu quốc tế

Các nghiên cứu quốc tế về phát hiện tin giả và nội dung độc hại đã có sự chuyển dịch rõ rệt từ các phương pháp học máy truyền thống (TF-IDF, SVM, XGBoost) [35], [57] sang các mô hình học sâu [38], [39], [40], [41], [43] và đặc biệt là các mô hình đa phương thức kết hợp văn bản - hình ảnh [36], [42], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56]. Nhiều công trình đã ghi nhận hiệu quả vượt trội nhờ áp dụng các kỹ thuật tiên tiến như cơ chế đồng chú ý (co-attention), học tương phản, căn chỉnh ngữ nghĩa qua CLIP hay hợp nhất động [47], [49], [51], [53]. Trong đó, các hướng tiếp cận đa phương thức của Uppada và cộng sự [55] hay Zhou và cộng sự [49] là những nghiên cứu gần gũi nhất với định hướng của luận án. Tuy nhiên, phần lớn các mô hình hiện nay chỉ được phát triển, đánh giá trên dữ liệu tiếng Anh hoặc tiếng Trung quy mô lớn, đồng thời phụ thuộc nặng nề vào học có giám sát trên tập dữ liệu gán nhãn đầy đủ. Đây là cơ sở thực tiễn để luận án tập trung phát triển mô hình phát hiện tin bài độc hại tiếng Việt dựa trên sự kết hợp giữa xử lý đa phương thức và học bán giám sát.

1.2.3. Phát hiện tin bài độc hại dựa trên học bán giám sát

1.2.3.1. Các nghiên cứu trong nước

Các nghiên cứu trong nước đã bắt đầu ứng dụng các phương pháp học bán giám sát để giải quyết rào cản khan hiếm dữ liệu gán nhãn trong nhiều bài toán xử lý ngôn ngữ tự nhiên tiếng Việt, tiêu biểu như phân loại tin tức [58], nhận dạng thực thể có tên [59], [60] và trích rút quan hệ [61]. Nhiều kỹ thuật từ tự động gán nhãn dựa trên tri thức chuyên ngành, thuật toán lan truyền nhãn (Label Propagation) cho đến học lặp bootstrapping (MIRA) đã chứng minh được tính hiệu quả và vượt trội so với học có giám sát khi tập

huấn luyện bị hạn chế [58], [59], [60], [61]. Trong đó, hướng tiếp cận của Nguyễn Minh Nhật và Trần Huỳnh Minh Tân [58] là gần nhất với luận án khi kết hợp thành công học sâu (viBERT) với cơ chế tự gán nhãn. Mặc dù vậy, các công trình trong nước chủ yếu vẫn giới hạn trên dữ liệu văn bản thuần túy, phụ thuộc kỹ thuật bán giám sát truyền thống và chưa giải quyết bài toán phát hiện tin bài độc hại đa phương thức hay phân loại đa lớp các nội dung nhạy cảm chính trị. Đây chính là động lực quan trọng để luận án tập trung xây dựng mô hình phát hiện tin bài độc hại tiếng Việt kết hợp xử lý đa phương thức và các kỹ thuật học bán giám sát hiện đại.

1.2.3.2. Các nghiên cứu quốc tế

Các nghiên cứu quốc tế đã phát triển nhiều phương pháp tiên tiến nhằm giải quyết rào cản khan hiếm dữ liệu gán nhãn trong phát hiện tin giả và tin độc hại, nổi bật là các hướng tiếp cận bán giám sát, học ít mẫu và học đa phương thức [62], [63], [64], [65], [66]. Các công trình tiêu biểu bao gồm mô hình SEMI-FND kết hợp ensemble văn bản với NasNet Mobile [62], MetaFEND sử dụng học siêu dữ liệu [63], các phương pháp lan truyền nhãn giả trên đồ thị (LOSS-GAT, GCN) [64], [65] và kỹ thuật tự huấn luyện chọn lọc độ tin cậy [66]. Trong đó, mô hình SEMI-FND [62] và MetaFEND [63] có sự tương đồng nhất với luận án khi khai thác dữ liệu đa phương thức trong điều kiện nhãn gán hạn chế. Tuy nhiên, các phương pháp này vẫn còn hạn chế do chủ yếu tập trung vào văn bản hoặc đồ thị mà chưa tích hợp hiệu quả hình ảnh, phụ thuộc nhiều vào lượng nhãn ban đầu hoặc chưa giải thích rõ vai trò đóng góp của từng phương thức. Đặc biệt, các mô hình hiện tại đa số xây dựng trên tiếng Anh hay tiếng Trung, tạo nên khoảng trống tri thức lớn đối với tiếng Việt mà luận án tập trung giải quyết thông qua hướng tiếp cận đa phương thức kết hợp học bán giám sát.

1.2.4. Ưu điểm và hạn chế của các đề xuất đã có

Các nghiên cứu gần đây, đặc biệt là mô hình học sâu và Transformer, đã cải thiện đáng kể hiệu quả phát hiện tin độc hại; một số hướng đa phương thức còn khai thác thêm hình ảnh để nâng cao độ chính xác. Tuy nhiên, các nghiên cứu trong nước chủ yếu vẫn xử lý văn bản, thiếu đa phương thức; đồng thời phần lớn phương pháp phụ thuộc học có giám sát, đòi hỏi dữ liệu nhãn lớn. Ngoài ra, nhiều công trình chỉ dừng ở phân loại nhị phân, chưa phản ánh đầy đủ tính đa lớp của tin độc hại.

Do đó, cần một mô hình tích hợp cho tiếng Việt có khả năng xử lý đa phương thức, tận dụng dữ liệu chưa gán nhãn (bán giám sát) và hỗ trợ phân loại đa lớp. Từ đó, luận án đặt ra hai câu hỏi nghiên cứu:

RQ1. Đâu là mô hình đa phương thức có khả năng xử lý hiệu quả cả dữ liệu văn bản và hình ảnh cho phát hiện tin bài độc hại tiếng Việt?

RQ2. Đâu là mô hình dựa trên học bán giám sát có khả năng phát hiện tin bài độc hại tiếng Việt hiệu quả mà chỉ sử dụng một phần dữ liệu có gán nhãn?

1.3. CÁC MÔ HÌNH HỌC MÁY SỬ DỤNG TRONG LUẬN ÁN

1.3.1. Các mô hình xử lý văn bản

Luận án tập trung vào các mô hình ngôn ngữ dựa trên kiến trúc của Transformer, cho phép biểu diễn ngữ nghĩa toàn diện thay thế cho các phương pháp truyền thống.

1.3.1.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) là mô hình ngôn ngữ hai chiều do Google phát triển, sử dụng cơ chế Masked Language Modeling để hiểu ngữ cảnh sâu.

Ưu điểm: Khả năng hiểu ngữ cảnh toàn diện, biểu diễn ngữ nghĩa giàu thông tin, linh hoạt trong tinh chỉnh.

Nhược điểm: Tiền huấn luyện chủ yếu trên tiếng Anh nên cần nhiều nguồn lực khi áp dụng cho tiếng Việt; yêu cầu tài nguyên tính toán cao.

1.3.1.2. RoBERTa

RoBERTa là phiên bản cải tiến của BERT do Facebook AI phát triển, tối ưu hóa quá trình tiền huấn luyện bằng cách loại bỏ nhiệm vụ NSP và sử dụng kỹ thuật masking động.

Ưu điểm: Hiệu suất cao và ổn định hơn BERT khi xử lý các tập dữ liệu lớn và phức tạp.

Nhược điểm: Vẫn gặp rào cản ngôn ngữ với tiếng Việt gốc; kích thước mô hình lớn gây khó khăn trong triển khai thực tế.

1.3.1.3. PhoBERT

PhoBERT là mô hình ngôn ngữ tiền huấn luyện chuyên biệt cho tiếng Việt, xây dựng trên kiến trúc RoBERTa và kho ngữ liệu tiếng Việt quy mô lớn.

Ưu điểm: Hiệu suất vượt trội trên dữ liệu tiếng Việt; nắm bắt tốt đặc thù hình thái và ngữ pháp phức tạp mà không cần chuyển ngữ.

Nhược điểm: Đòi hỏi tài nguyên tính toán lớn; cần kết hợp thêm các kỹ thuật phụ trợ để mô hình hóa tốt hơn mối quan hệ giữa các nhãn.

1.3.2. Các mô hình xử lý ảnh

Để nhận diện tin bài độc hại đa phương thức, luận án xem xét các mô hình từ mạng nơ-ron tích chập (CNN) đến kiến trúc Transformer cho thị giác máy tính.

1.3.2.1. Xception

Xception sử dụng tích chập tách biệt theo chiều sâu (Depthwise Separable Convolution), kế thừa và mở rộng dòng Inception.

Ưu điểm: Giảm thiểu số lượng tham số, hiệu quả trong học đặc trưng cục bộ và xử lý nhanh.

Nhược điểm: Khả năng khai thác ngữ nghĩa liên vùng hạn chế, khó nhận diện các nội dung ngụy trang tinh vi.

1.3.2.2. VGG19

VGG19 là mạng CNN kinh điển gồm 19 lớp với các bộ lọc 3×3 đồng nhất, thường được dùng làm mô hình trích xuất đặc trưng nền tảng.

Ưu điểm: Cấu trúc đơn giản, ổn định, khai thác tốt các đặc trưng thị giác cơ bản như đường nét, màu sắc

Nhược điểm: Số lượng tham số rất lớn; thiếu cơ chế chú ý khiến việc học quan hệ không gian toàn cục kém hiệu quả.

1.3.2.3. Swin Transformer V2

Swin Transformer V2 ứng dụng cơ chế Attention theo cửa sổ trượt phân cấp, giải quyết các hạn chế về độ ổn định và khả năng mở rộng của kiến trúc Transformer cho hình ảnh.

Ưu điểm: Học tốt ngữ nghĩa toàn cục và mối tương quan giữa các vùng ảnh; hiệu quả với các hình ảnh ẩn dụ, bố cục phức tạp.

Nhược điểm: Cấu trúc phức tạp, đòi hỏi quy trình huấn luyện khắt khe để đạt được độ hội tụ tối ưu.

1.3.3. Học bán giám sát trong phân loại dữ liệu

Học bán giám sát kết hợp dữ liệu có nhãn và chưa gán nhãn để cải thiện hiệu quả mô hình, đặc biệt phù hợp khi dữ liệu gán nhãn khan hiếm. Các hướng chính gồm pseudo-labeling, consistency regularization và lan truyền nhãn, trong đó pseudo-labeling được ưu tiên nhờ đơn giản, dễ triển khai và phù hợp với dữ liệu tiếng Việt. Phương pháp này giúp mở rộng tập huấn luyện và tăng khả năng tổng quát hóa, nhưng cần kiểm soát nhiễu để tránh lan truyền sai. Trong luận án, học bán giám sát là thành phần quan trọng nhằm nâng cao hiệu quả phát hiện tin bài độc hại trong điều kiện dữ liệu hạn chế.

1.3.4. Các mô hình hình kết hợp cho xử lý văn bản và ảnh

Các mô hình đa phương thức gồm ba hướng: kết hợp sớm, kết hợp muộn và kết hợp lai. Trong đó, kết hợp sớm đơn giản, hiệu quả và phù hợp với dữ liệu hạn chế, trong khi kết hợp muộn và lai có khả năng khai thác sâu hơn nhưng phức tạp và dễ quá khớp. Luận án lựa chọn kết hợp sớm bằng cách ghép đặc trưng từ PhoBERT (văn bản) và Swin Transformer V2 (ảnh), giúp cân bằng hiệu quả, tính ổn định và khả năng triển khai trong bối cảnh dữ liệu tiếng Việt còn hạn chế.

1.3.5. Các độ đo đánh giá

Để đánh giá được hiệu suất mô hình phân loại, các độ đo được sử dụng bao gồm: Accuracy, Precision, Recall, F1-score và Detection Rate (DR).

Độ đo Accuracy là độ chính xác tổng thể, được tính theo công thức:

$$Accuracy = \frac{TP+TN}{tổng\ số\ mẫu} \quad (1.1)$$

Độ đo Precision là độ chính xác, được tính theo công thức:

$$Precision = \frac{TP}{TP+FP} \quad (1.2)$$

Độ đo Recall là độ nhạy hay độ bao phủ, được tính theo công thức:

$$Recall = \frac{TP}{TP+FN} \quad (1.3)$$

F1-score là độ đo biểu diễn trung bình điều hòa giữa 2 độ đo Precision và Recall, được tính theo công thức:

$$F1 = \frac{2(precision*recall)}{precision+recall} \quad (1.4)$$

Tỉ lệ phát hiện trên dữ liệu test được tính trên công thức:

$$\text{DR} = \frac{Số\ lượng\ mẫu\ dự\ đoán\ đúng}{Tổng\ số\ mẫu\ thử\ nghiệm} \quad (1.5)$$

Các tham số TP, FP, FN và TN được biểu diễn trong ma trận nhầm lẫn trên Bảng 1.1

Bảng 1.1 Các tham số TP, FP, FN và TN trong ma trận nhầm lẫn

		Lớp thực tế	
		Độc hại	Bình thường
Lớp dự đoán	Độc hại	TP (True Positive)	FP (False Positive)
	Bình thường	FN (False Negative)	TN (True Negative)

1.4. THU THẬP TẬP DỮ LIỆU TIN BÀI TIẾNG VIỆT

1.4.1. Các khó khăn trong thu thập dữ liệu tin bài độc hại tiếng Việt

Việc thu thập dữ liệu tin bài độc hại tiếng Việt gặp nhiều khó khăn do nội dung đa dạng, biến đổi nhanh và thường được ngụy trang tinh vi dưới dạng văn bản kết hợp hình ảnh. Quá trình gán nhãn phức tạp và thiếu nhất quán do phụ thuộc vào đánh giá chủ quan. Ngoài ra, còn tồn tại các ràng buộc về pháp lý và đạo đức trong thu thập dữ liệu. Bên cạnh đó, tiếng Việt thiếu các bộ dữ liệu chuẩn hóa quy mô lớn và dữ liệu thường mất cân bằng (ít mẫu độc hại). Những hạn chế này đặt ra nhu cầu áp dụng các phương pháp như học bán giám sát và mô hình đa phương thức.

1.4.2. Thu thập, làm sạch và gán nhãn tin bài tiếng Việt

Luận án xây dựng tập dữ liệu tin bài tiếng Việt do chưa có bộ dữ liệu chuẩn hóa cho bài toán phát hiện nội dung độc hại chính trị. Quy trình gán nhãn dựa trên Điều 13 Luật An ninh mạng 2025, gồm ba nhãn độc hại: vu khống, bôi nhọ; kêu gọi chống đối; thúc đẩy tư tưởng cực đoan, cùng nhãn tin bài thường.

Dữ liệu được gán nhãn độc lập bởi hai nhóm và đánh giá độ nhất quán bằng hệ số Krippendorff's Alpha ($\alpha = 0,84$), cho thấy độ tin cậy cao; các mẫu chưa thống nhất được rà soát bởi nhóm thứ ba. Quy trình nhiều bước giúp giảm sai lệch chủ quan và nâng cao tính nhất quán.

Kết quả, hai bộ dữ liệu được xây dựng: (i) 12.000 mẫu cho học có giám sát (cân bằng giữa các lớp), (ii) 18.000 mẫu cho học bán giám sát. Bộ dữ liệu bao gồm cả văn bản và hình ảnh, phản ánh tính đa dạng, làm nền tảng thực nghiệm cho các mô hình đa phương thức và học bán giám sát.

1.5. HƯỚNG NGHIÊN CỨU CỦA LUẬN ÁN

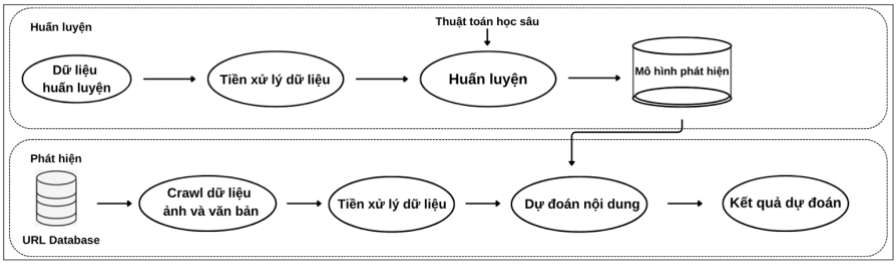
1.5.1. Các vấn đề giải quyết trong luận án

Bài toán 1: Phát hiện tin bài độc hại tiếng Việt dựa trên mô hình học sâu đa phương thức. Dữ liệu đầu vào gồm văn bản và hình ảnh của tin bài; đầu ra là nhãn phân loại. Lời giải của bài toán 1 là câu trả lời cho Câu hỏi nghiên cứu RQ1.

Bài toán 2: Phát hiện tin bài độc hại tiếng Việt dựa trên học bán giám sát. Dữ liệu đầu vào gồm một phần nhỏ dữ liệu có nhãn và một lượng lớn dữ liệu chưa gán nhãn; đầu ra là mô hình có khả năng phân loại hiệu quả trong điều kiện dữ liệu hạn chế. Lời giải của bài toán 2 là câu trả lời cho Câu hỏi nghiên cứu RQ2.

1.5.2. Kiến trúc mô hình tổng thể cho các hướng nghiên cứu của luận án

Luận án đề xuất kiến trúc hệ thống phát hiện tin bài độc hại tiếng Việt theo hướng đa phương thức, gồm hai giai đoạn chính: huấn luyện và phát hiện.



Hệ thống thực hiện thu thập dữ liệu từ các nguồn trực tuyến, tiền xử lý và trích xuất đặc trưng văn bản-hình ảnh, sau đó huấn luyện mô hình học

sâu để xây dựng bộ phân loại. Trong giai đoạn vận hành, dữ liệu mới được đưa vào mô hình để phát hiện và phân loại tin bài, kết quả được lưu trữ và hiển thị phục vụ tra cứu và cảnh báo.

1.6. KẾT LUẬN CHƯƠNG 1

Chương 1 trình bày cơ sở lý luận và tổng quan nghiên cứu về phát hiện tin bài độc hại tiếng Việt, chỉ ra đặc trưng đa phương thức và các hạn chế của phương pháp hiện tại. Luận án lựa chọn hướng tiếp cận kết hợp PhoBERT, Swin Transformer V2 và học bán giám sát. Đồng thời, các thách thức về dữ liệu và tính biến đổi nội dung được xác định. Từ đó, định hướng xây dựng mô hình đa phương thức bán giám sát cho các chương tiếp theo.

CHƯƠNG 2: PHÁT HIỆN TIN BÀI ĐỘC HẠI TIẾNG VIỆT SỬ DỤNG KẾT HỢP ĐẶC TRƯNG VĂN BẢN VÀ HÌNH ẢNH

2.1. ĐẶT VẤN ĐỀ

Các nghiên cứu trong nước hiện chủ yếu xử lý văn bản đơn lẻ nên dễ bỏ sót các nội dung độc hại ngụy trang tinh vi qua hình ảnh, trong khi các giải pháp đa phương thức quốc tế lại khó áp dụng trực tiếp do rào cản ngôn ngữ và sự thiếu hụt dữ liệu tiếng Việt. Để khắc phục hạn chế này, Chương 2 đề xuất một mô hình đa phương thức chuyên biệt cho tiếng Việt nhằm khai thác mối tương quan giữa văn bản và hình ảnh, giúp tối ưu độ chính xác và khả năng nhận diện trong các kịch bản thực tế.

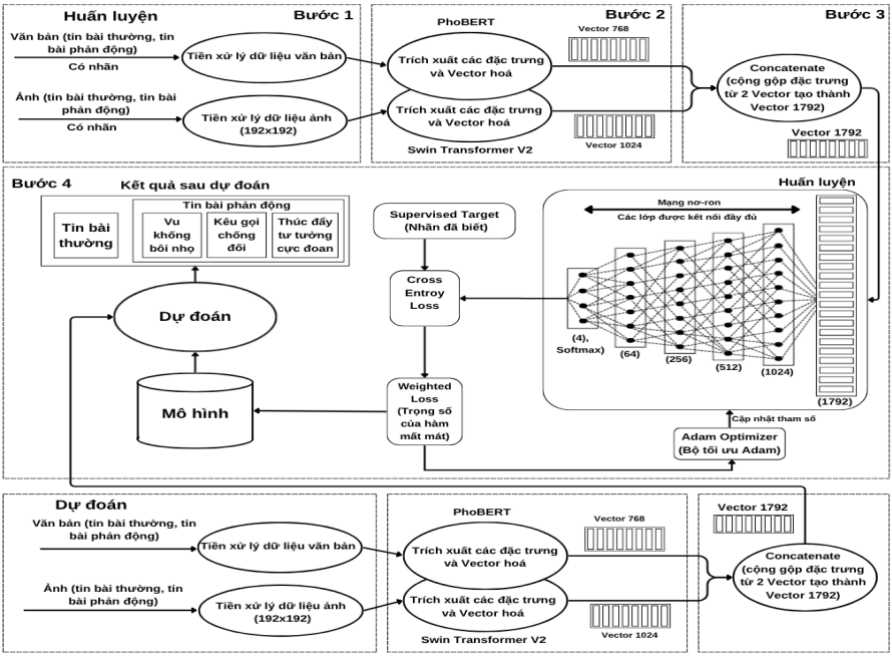
2.2. MÔ HÌNH PHÁT HIỆN SỬ DỤNG KẾT HỢP DỰA TRÊN ĐẶC TRƯNG VĂN BẢN VÀ HÌNH ẢNH

2.2.1. Giới thiệu mô hình phát hiện

Bài toán phân loại đa lớp được đặt ra nhằm phản ánh đúng tính đa dạng của các loại tin bài độc hại, từ đó hỗ trợ lựa chọn phương thức xử lý phù hợp cho từng nhóm nội dung. So với phân loại nhị phân, cách tiếp cận này cho phép nhận diện chi tiết hơn các dạng độc hại như “vu khống, bôi nhọ”, “kêu gọi chống đối” hay “thúc đẩy tư tưởng cực đoan”.

Mô hình đề xuất sử dụng PhoBERT để trích xuất đặc trưng văn bản và Swin Transformer V2 để khai thác đặc trưng hình ảnh, sau đó chuẩn hóa và hợp nhất trong một kiến trúc đa phương thức. Lớp phân loại sử dụng hàm

softmax cùng hàm mất mát Cross-Entropy nhằm học phân bố xác suất trên các nhãn và tối ưu khả năng phân biệt giữa các lớp.



Hình 2.1. Kiến trúc của mô hình kết hợp cho phân loại đa lớp

Cách tiếp cận này giúp mô hình khai thác tốt mối quan hệ giữa các nhãn, phản ánh rõ mức độ và loại hình độc hại trong từng tin bài, đồng thời vẫn đảm bảo khả năng mở rộng khi bổ sung các nhãn mới

2.2.2. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu nhằm chuẩn hóa văn bản và hình ảnh trước khi đưa vào mô hình. Với văn bản, dữ liệu được làm sạch (chuyển chữ thường, loại bỏ ký tự đặc biệt, URL, từ viết tắt), tách từ tiếng Việt, đệm độ dài và mã hóa thành token ID phù hợp với PhoBERT. Với hình ảnh, dữ liệu được chuẩn hóa bằng cách resize về kích thước 192×192, chuẩn hóa giá trị điểm ảnh và chuyển sang dạng tensor để phù hợp với Swin Transformer V2.

2.2.3. Huấn luyện và phát hiện

Quá trình huấn luyện bao gồm các bước sau:

Bước 1 (B1): Các tin bài đầu vào, bao gồm cả văn bản và hình ảnh, được tiền xử lý theo quy trình thống nhất như trình bày tại Mục 2.2.2. Văn bản được chuẩn hóa, tách từ, đệm chuỗi và mã hóa thành token ID. Hình ảnh được resize về kích thước 192×192 pixel và chuẩn hóa giá trị điểm ảnh.

Bước 2 (B2): Văn bản sau tiền xử lý được đưa vào mô hình PhoBERT để sinh vector đặc trưng ngôn ngữ có kích thước 768 chiều. Đồng thời, hình ảnh được xử lý qua Swin Transformer V2 để sinh vector đặc trưng thị giác 1024 chiều.

Bước 3 (B3): Hai vector đặc trưng đầu ra từ PhoBERT và Swin Transformer V2 được kết hợp bằng phép nối (concatenate), tạo thành một vector hợp nhất có kích thước 1792.

Bước 4 (B4): Sau khi có được vector đặc trưng hợp nhất x' (kích thước 1792), bước tiếp theo là huấn luyện mô hình dự đoán $f_\theta(x')$ để phân loại tin bài thành tin bài độc hại hoặc tin bài thường. Mô hình này là một mạng nơ-ron được thiết kế để học cách ánh xạ các đặc trưng phức tạp sang các quyết định phân loại.

Kiến trúc Mạng Nơ-ron: Mạng nơ-ron $f_\theta(x')$ là một mạng Feed-Forward với các lớp kết nối đầy đủ. Kiến trúc của mạng được thiết kế để trích xuất các đặc trưng bậc cao và đưa ra dự đoán xác suất.

Quá trình dự đoán bao gồm các bước sau:

Bước 1 (B1): Các tin bài đầu vào, gồm cả văn bản và hình ảnh, được xử lý tương tự như giai đoạn huấn luyện. Văn bản được chuẩn hóa, tách từ, đệm chuỗi và mã hóa thành token IDs, hình ảnh được thay đổi kích thước về 192×192 điểm ảnh và chuẩn hóa giá trị điểm ảnh để đưa về định dạng tensor.

Bước 2 (B2): Văn bản đầu vào được đưa qua mô hình PhoBERT để tạo vector đặc trưng 768 chiều, trong khi hình ảnh được xử lý bằng Swin Transformer V2 để sinh vector đặc trưng 1024 chiều.

Bước 3 (B3): Hai vector đặc trưng nói trên được kết hợp bằng phép nối (con-catenate) thành một vector hợp nhất có kích thước 1792 chiều, tương tự như trong giai đoạn huấn luyện.

Bước 4 (B4): Vector hợp nhất được đưa vào mô hình mạng nơ-ron FFNN đã được huấn luyện trước đó. Mô hình sẽ tính toán và đưa ra một giá

trị xác suất đầu ra, nằm trong khoảng (0, 1), thông qua hàm softmax tại lớp cuối cùng.

Bước 5 (B5): Nhãn dự đoán cuối cùng được xác định bằng cách chọn lớp có xác suất softmax cao nhất. Nếu xác suất cao nhất thuộc về nhãn "tin bài thường", mẫu được phân loại là tin bài thường và tương tự với nhóm tin độc hại. Cách phân loại này đảm bảo mô hình đưa ra quyết định dựa trên phân phối xác suất đầy đủ thay vì một ngưỡng cố định.

2.2.4. Cài đặt

Môi trường thực nghiệm được thiết lập trên một cấu hình phần cứng chuyên biệt nhằm đáp ứng yêu cầu tính toán chuyên sâu của các mô hình học sâu. Cụ thể, (CPU) Intel Xeon E5-2680 v3, xử lý Đồ họa (GPU) NVIDIA Tesla P40 với 24 GB bộ nhớ. Hệ thống được cấu hình để hỗ trợ phiên bản CUDA 12.6, quá trình thực nghiệm được kiểm soát thông qua các siêu tham số như batch size: 64, số chu kỳ huấn luyện epoch: 30, tốc độ học learning rate: 0,001, và số lớp nhãn của bài toán.

2.3. THỬ NGHIỆM VÀ ĐÁNH GIÁ

2.3.1. Tập dữ liệu thử nghiệm

2.3.1.1. Bộ dữ liệu cho chế độ phát hiện nhị phân

Bộ dữ liệu cho chế độ phát hiện nhị phân được thu thập gồm 4000 tin bài tiếng Việt, bao gồm:

- 1000 tin bài độc hại dạng văn bản và 1000 tin bài độc hại dạng ảnh từ các nguồn như diễn đàn hải ngoại, fanpage mạng xã hội có nội dung xuyên tạc, độc hại. Các tin bài thuộc nhóm này được gán nhãn 'Tin bài độc hại'.

- 1000 tin bài thường dạng văn bản và 1000 tin bài thường dạng ảnh từ các trang tin tức chính thống và diễn đàn phổ biến trong nước. Các tin bài thuộc nhóm này được gán nhãn 'Tin bài thường'.

2.3.1.2. Bộ dữ liệu cho chế độ phát hiện đa lớp

Bộ dữ liệu cho chế độ phát hiện đa lớp được thu thập gồm 12.000 tin bài tiếng Việt sử dụng để huấn luyện và kiểm tra mô hình đa phương pháp đề xuất, bao gồm:

- Các tin bài có nội dung độc hại: gồm 3.000 bài viết dưới dạng văn bản và 3.000 bài dưới dạng ảnh thuộc 3 nhóm nhãn (mỗi nhãn 1.000 bài) từ nhiều nguồn khác nhau, như trên các diễn đàn hải ngoại, các fanpage của

các mạng xã hội. Các tin bài thuộc 3 nhóm này được dán nhãn lần lượt là ‘Vu khống, bôi nhọ’, ‘Kêu gọi chống đối’, ‘Thúc đẩy tư tưởng cực đoan’.

- Các tin bài có nội dung tin bài thường: 3000 bài viết dưới dạng văn bản và 3000 bài dưới dạng ảnh thu thập từ các trang tin tức và các diễn đàn. Các tin bài thuộc nhóm này được dán nhãn ‘Tin bài thường’.

2.3.2. Các kịch bản thử nghiệm và kết quả

2.3.2.1. Kết quả thử nghiệm phát hiện nhị phân

Thử nghiệm phát hiện nhị phân được thiết kế chuẩn mực trên tập dữ liệu 4.000 tin bài tiếng Việt với tỷ lệ phân chia tối ưu (70% huấn luyện, 15% đánh giá và 15% kiểm tra), bảo đảm tính khách quan và tin cậy cho kết quả phân loại. Mô hình kết hợp đa phương thức đề xuất PhoBERT + Swin Transformer V2 thể hiện tính vượt trội tuyệt đối khi đạt độ chính xác Accuracy 97,00% và F1-score 97,00%, vượt trội hơn so với các mô hình đơn phương thức. Việc cải thiện 0,5% so với mô hình văn bản đơn lẻ tốt nhất (PhoBERT với 96,50%) khẳng định giá trị quan trọng của việc bổ sung đặc trưng thị giác nhằm xử lý hiệu quả các trường hợp thông tin nhiễu hoặc nhập nhằng. So với các nghiên cứu liên quan sử dụng BERT + Xception (F1-score đạt khoảng 93,25% - 93,29%), mô hình đề xuất thể hiện ưu thế rõ rệt nhờ việc tối ưu hóa đặc trưng ngôn ngữ tiếng Việt kết hợp trích xuất ngữ nghĩa hình ảnh hiện đại. Phân tích ma trận nhầm lẫn cho thấy số lượng bỏ sót tin độc hại rất thấp (chỉ 8/400 mẫu), chứng minh mô hình có độ tin cậy cao và hoàn toàn đáp ứng yêu cầu của các hệ thống giám sát an ninh mạng thực tế

2.3.2.2. Kết quả thử nghiệm phát hiện đa lớp

2.3.3. Nhận xét và đánh giá

Mô hình đa phương thức đề xuất PhoBERT + Swin Transformer V2 thể hiện tính vượt trội tuyệt đối trong kịch bản phát hiện đa lớp với độ chính xác Accuracy đạt 95,67% và F1-score đạt 95,68%, vượt xa các mô hình kết hợp khác như BERT + Swin Transformer V2 (92,50%) hay RoBERTa + Swin Transformer V2 (66,45%). Mô hình duy trì tỷ lệ phát hiện (DR) rất cao và đồng đều trên các nhóm nội dung, cụ thể đạt 97,78% với 'Tin bài thường', 94,67% với 'Vu khống, bôi nhọ', 97,33% với 'Kêu gọi chống đối' và 96,00% với 'Thúc đẩy tư tưởng cực đoan'. Việc phân tích ma trận nhầm lẫn cho thấy mức độ sai lệch giữa các nhóm nhãn độc hại là rất thấp, khẳng

định khả năng trích xuất và học các đặc trưng phân biệt riêng biệt của mô hình đối với từng loại hành vi độc hại. Sự kết hợp giữa PhoBERT và Swin Transformer V2 giúp khắc phục triệt để hạn chế của các phương pháp xử lý đơn phương thức nhờ khả năng nhận diện chính xác các văn bản nhúng trong hình ảnh cũng như các thông tin ngụy trang tinh vi. Kết quả thực nghiệm chứng minh mô hình không chỉ đạt hiệu suất tổng thể cao mà còn đáp ứng tốt yêu cầu phân loại chi tiết trong bài toán giám sát và phòng chống tin bài độc hại thực tế.

2.3.3.1. Hiệu suất của mô hình trong chế độ phát hiện nhị phân

Trong kịch bản phát hiện nhị phân (Bảng 2.2), mô hình kết hợp PhoBERT và Swin Transformer V2 đạt hiệu suất cao nhất với độ chính xác 97%, nhỉnh hơn so với mô hình văn bản đơn lẻ và vượt trội so với các mô hình chỉ sử dụng ảnh. Mức cải thiện tuy không lớn về mặt tuyệt đối nhưng cho thấy việc bổ sung đặc trưng hình ảnh đã giúp mô hình xử lý tốt hơn các trường hợp khó, nơi thông tin văn bản chưa đủ rõ ràng.

2.3.3.2. Hiệu suất của mô hình trong chế độ phát hiện đa lớp

Trong kịch bản phát hiện đa lớp (Bảng 2.5), mô hình đề xuất tiếp tục duy trì hiệu năng cao và ổn định trên các chỉ số đánh giá (Accuracy, Precision, Recall, F1-score đều xấp xỉ 95-96%). Kết quả này khẳng định hiệu quả của cách tiếp cận đa phương thức trong việc phân biệt các nhóm nhãn độc hại khác nhau, đồng thời cho thấy mô hình có khả năng tổng quát hóa tốt trong bối cảnh dữ liệu đa dạng và phức tạp.

2.4. KẾT LUẬN CHƯƠNG

Chương này đã đề xuất mô hình phát hiện tin bài độc hại tiếng Việt dựa trên kết hợp đặc trưng văn bản và hình ảnh, khắc phục hạn chế của các phương pháp đơn phương thức. Kết quả thực nghiệm cho thấy mô hình đạt hiệu năng cao và ổn định, đặc biệt trong bài toán đa lớp với F1-score trên 95%.

Những kết quả này khẳng định hiệu quả của hướng tiếp cận đa phương thức và đóng góp về khả năng phân loại đa lớp trong điều kiện dữ liệu thực tế. Trên cơ sở đó, chương tiếp theo sẽ phát triển theo hướng học bán giám sát nhằm tận dụng dữ liệu chưa gán nhãn và nâng cao hơn nữa hiệu quả mô hình.

CHƯƠNG 3: PHÁT HIỆN TIN BÀI ĐỘC HẠI TIẾNG VIỆT DỰA TRÊN HỌC BÁN GIÁM SÁT

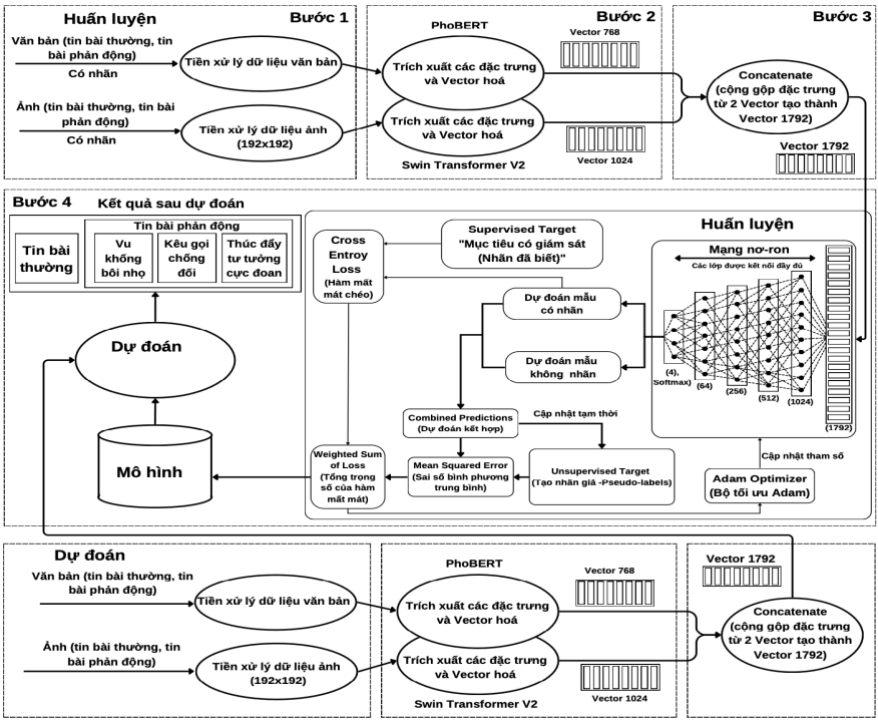
3.1. ĐẶT VẤN ĐỀ

Trong bối cảnh dữ liệu gán nhãn cho bài toán phát hiện tin bài độc hại tiếng Việt còn hạn chế, trong khi dữ liệu chưa gán nhãn lại rất phong phú, việc áp dụng học bán giám sát trở thành hướng tiếp cận quan trọng nhằm tận dụng tối đa nguồn dữ liệu sẵn có và nâng cao hiệu quả mô hình.

3.2. MÔ HÌNH PHÁT HIỆN DỰA TRÊN HỌC BÁN GIÁM SÁT

Mục này mô tả và phân tích mô hình đa phương thức dựa trên học bán giám sát cho bài toán phân loại đa lớp, trong đó nhãn độc hại được tách thành 3 nhãn con tương tự như đã trình bày tại mục 2.2.

3.2.1. Giới thiệu mô hình phát hiện dựa trên học bán giám sát



Hình 3.1. Kiến trúc mô hình phát hiện đa lớp dựa trên học bán giám sát

Mô hình học bán giám sát trong luận án được xây dựng nhằm khai thác đồng thời dữ liệu có nhãn và chưa gán nhãn cho bài toán phát hiện tin bài

độc hại đa lớp tiếng Việt. Đặc trưng văn bản được trích xuất bằng PhoBERT và đặc trưng hình ảnh được học bằng Swin Transformer V2, sau đó được hợp nhất thành vector chung đưa vào mạng phân loại.

Trong quá trình huấn luyện, mô hình áp dụng cơ chế gán nhãn giả (pseudo-labeling), trong đó dữ liệu chưa gán nhãn được dự đoán nhãn và lựa chọn theo ngưỡng tin cậy để bổ sung vào tập huấn luyện. Cách tiếp cận này giúp mở rộng dữ liệu hiệu quả và cải thiện khả năng khái quát hóa của mô hình trong điều kiện dữ liệu thực tế hạn chế.

3.2.2. Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu được thực hiện theo quy trình đã mô tả ở mục 2.2.2.

3.2.3. Huấn luyện và phát hiện

Dựa trên Giải thuật 3.1, có thể thấy các bước từ (B1) đến (B4) trong quá trình huấn luyện tuân theo quy trình tương tự như đã mô tả tại mục 2.2.3. Tuy nhiên, tại bước (B4), dữ liệu sau khi đi qua lớp Softmax sẽ tiếp tục được đưa vào huấn luyện theo phương thức bán giám sát.

Để tận dụng dữ liệu chưa gán nhãn, chúng tôi sử dụng phương pháp tạo nhãn giả. Mô hình gán nhãn cho các mẫu chưa có nhãn dựa trên độ tin cậy của dự đoán: nhãn phải đạt xác xuất 0.6 mới được gán nhãn. Ngưỡng tin cậy này giúp giảm thiểu nhãn sai, cải thiện chất lượng huấn luyện.

Trong quá trình huấn luyện, mô hình được tiếp tục hoàn thiện bằng cách "huấn luyện lại" trên một tập dữ liệu kết hợp. Tập dữ liệu này bao gồm cả dữ liệu gốc đã được gán nhãn thủ công và dữ liệu chưa được gán nhãn, nhưng đã được gán "nhãn giả" nhờ vào kết quả dự đoán ban đầu của mô hình. Quá trình huấn luyện lại này sử dụng cả hai thành phần mất mát: mất mát giám sát (tính trên dữ liệu có nhãn gốc) và mất mát không giám sát (tính trên dữ liệu gán nhãn giả).

3.2.4. Cài đặt

Quá trình cài đặt và triển khai mô hình được thực hiện bằng Python 3.12.6 với nền tảng PyTorch được thực hiện theo quy trình đã mô tả ở mục 2.2.4. Số chu kỳ huấn luyện epoch: 30, tốc độ học learning rate: 0,001, số lớp nhãn của bài toán là 4 và ngưỡng độ tin cậy cho gán nhãn giả là 0,6.

3.3. THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.3.1. Tập dữ liệu thử nghiệm

Bộ dữ liệu cho mô hình phát hiện đa lớp kết hợp học bán giám sát gồm 18000 tin bài tiếng Việt sử dụng để huấn luyện và kiểm tra mô hình đa phương pháp kết hợp học bán giám sát cho phát hiện tin bài phản động tiếng Việt, bao gồm:

Các tin bài có nội dung độc hại: gồm 4500 bài viết dưới dạng văn bản và 4500 bài dưới dạng ảnh từ nhiều nguồn khác nhau, như trên các diễn đàn hải ngoại, các fanpage của các mạng xã hội. Các tin bài thuộc 3 nhóm nhãn (mỗi nhãn 1500 tin bài dạng văn bản, 1500 tin bài dạng ảnh) được gán nhãn lần lượt là 'Vu khống, bôi nhọ', 'Kêu gọi chống đối', 'Thúc đẩy tư tưởng cực đoan'.

Các tin bài có nội dung tin bài thường: 4500 bài viết dưới dạng văn bản và 4500 bài dưới dạng ảnh thu thập từ các trang tin tức và các diễn đàn. Các tin bài thuộc nhóm này được dán nhãn 'tin bài thường'.

3.3.2. Các kịch bản thử nghiệm và kết quả

Thực nghiệm được tiến hành trên tập dữ liệu gồm 18.000 tin bài (9.000 văn bản và 9.000 hình ảnh) được chia ngẫu nhiên theo tỷ lệ 70% huấn luyện, 15% đánh giá và 15% kiểm tra. Các mô hình kết hợp khác nhau giữa các bộ mã hóa văn bản (BERT, RoBERTa, PhoBERT) và bộ mã hóa hình ảnh (VGG19, Xception, Swin Transformer V2) đã được huấn luyện và đánh giá độc lập qua 6 kịch bản với tỷ lệ dữ liệu có gán nhãn giảm dần từ 50%, 30%, 20%, 15%, 10% đến 5%. Khảo sát trên ngưỡng tin cậy gán nhãn giả cho thấy mức ngưỡng 0,6 mang lại sự cân bằng tối ưu giữa chất lượng và số lượng nhãn giả, giúp mô hình đạt hiệu suất cao nhất. Mô hình kết hợp đề xuất PhoBERT + Swin Transformer V2 đạt hiệu năng vượt trội so với các mô hình kết hợp khác ở tất cả kịch bản, duy trì F1-score trên 93% khi tỷ lệ dữ liệu có nhãn từ 30% trở lên và đạt 85,78% ngay cả khi dữ liệu có nhãn chỉ ở mức 5%. Kết quả đánh giá chi tiết theo từng nhãn cho thấy tỷ lệ phát hiện đúng (DR) của mô hình đề xuất đạt trên 85% cho 3/4 nhóm tin bài khi dùng 30% dữ liệu gán nhãn, khẳng định hiệu quả vượt trội so với các nghiên cứu bán giám sát liên quan như LOSS-GAT

3.3.3. Nhận xét và đánh giá

Các kịch bản thử nghiệm đã được thiết kế bài bản và toàn diện, giúp đánh giá chính xác khả năng thích ứng của mô hình khi tỷ lệ dữ liệu gán nhãn giảm dần từ 50% xuống 5%. Việc lựa chọn ngưỡng tin cậy 0,6 cho kỹ thuật gán nhãn giả thể hiện sự tính toán hợp lý, đảm bảo sự cân bằng giữa số lượng và độ tin cậy của các mẫu dữ liệu được bổ sung. Kết quả thực nghiệm chứng minh tính ưu việt vượt trội của mô hình kết hợp PhoBERT + Swin Transformer V2 so với các mô hình đối sánh và các nghiên cứu bán giám sát trước đó. Sự ổn định về hiệu năng (với F1-score đạt trên 93% chỉ với 30% dữ liệu có nhãn) đã khẳng định mạnh mẽ tính hiệu quả và khả năng khái quát hóa của phương pháp đề xuất. Những phân tích chi tiết theo từng nhóm nhãn và ma trận nhầm lẫn không chỉ làm rõ thế mạnh nhận diện mà còn mang lại giá trị thực tiễn cao cho việc ứng dụng mô hình trong các hệ thống giám sát thực tế.

3.4. KẾT LUẬN CHƯƠNG

Chương 3 trình bày mô hình phát hiện tin bài độc hại tiếng Việt bằng học bán giám sát, kết hợp đặc trưng văn bản từ PhoBERT và hình ảnh từ Swin Transformer V2 nhằm khai thác hiệu quả nguồn dữ liệu chưa gán nhãn dồi dào qua cơ chế gán nhãn giả. Kết quả thực nghiệm vượt trội so với các mô hình đối sánh, đồng thời duy trì hiệu năng cao ngay cả khi tỷ lệ dữ liệu gán nhãn chỉ đạt từ 5–20%. Bên cạnh khả năng phân biệt tin thường và tin độc hại, mô hình còn nhận diện tương đối chính xác từng nhóm nội dung độc hại với độ chính xác và F1-score tổng thể ổn định trên 90% khi sử dụng từ 10% dữ liệu gán nhãn trở lên. Những đóng góp này khẳng định tính khả thi vượt trội của học bán giám sát trong bài toán thực tế, đồng thời đã được công bố tại Tạp chí Học viện Công nghệ Bưu chính Viễn thông và Hội thảo quốc tế VCRIS 2025.

KẾT LUẬN

Trước thách thức nghiêm trọng từ tin bài độc hại tiếng Việt (đặc biệt là nhóm nội dung chính trị), luận án đã phát triển các mô hình đa phương thức kết hợp văn bản (PhoBERT) và hình ảnh (Swin Transformer V2) dựa trên hai hướng tiếp cận học có giám sát và học bán giám sát. Nghiên cứu đã đề xuất thành công mô hình phân loại đa lớp (4 nhóm nội dung) và mô hình

học bán giám sát sử dụng kỹ thuật gán nhãn giả giúp duy trì hiệu năng cao (F1-score trên 93%) ngay cả khi dữ liệu có nhãn bị hạn chế. Để phục vụ thực nghiệm, luận án đã tự thu thập và xây dựng bộ dữ liệu đa phương thức tiếng Việt với quy mô 18.000 mẫu, tạo nền tảng vững chắc cho tính tin cậy của các kết quả đạt được. Mặc dù đạt nhiều thành tựu, nghiên cứu vẫn tồn tại một số hạn chế như phạm vi dữ liệu mới chỉ dừng ở văn bản – hình ảnh, tập trung vào mảng chính trị, giới hạn trong tiếng Việt, cũng như nguy cơ tích lũy lỗi từ gán nhãn giả và sự nhầm lẫn giữa các nhãn tương đồng. Từ đó, luận án đề xuất các hướng phát triển tiếp theo bao gồm mở rộng dữ liệu đa dạng, ứng dụng kỹ thuật học sâu tiên tiến (như self-supervised hay few-shot learning), tích hợp thêm phương thức âm thanh/video và hướng tới xây dựng hệ thống giám sát thời gian thực.

DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ

TẠP CHÍ KHOA HỌC

[CT1] Lê Ngọc An, Hoàng Xuân Dậu, Dương Trần Đức, Đinh Tuấn Long, Một mô hình đa phương pháp cho phát hiện tin bài phản động tiếng Việt, Tạp chí Khoa học Công nghệ thông tin và Truyền thông, Số 01 (CS.01) 2025. Link: https://tapchi.ptit.edu.vn/jstic-ptit/index.php/jstic/article/view/1662/505.

[CT2] Ngoc An Le, Xuan Dau Hoang, Xuan Hanh Vu, Thi Thu Trang Ninh, A novel multimodal model for detecting Vietnamese toxic news using PhoBERT and Swin Transformer V2, IJEECS 2025, ISSN: 2502-4752, e-ISSN: 2502-4760, https://ijeecs.iaescore.com/index.php/IJEECS/article/view/43186, Scopus Q4.

[CT3] Lê Ngọc An, Hoàng Xuân Dậu, Đinh Trường Duy, Phát hiện tin bài độc hại tiếng việt đa lớp dựa trên học bán giám sát, Tạp chí Khoa học Công nghệ thông tin và Truyền thông, Tập 11 - Số 01 (CS.01), Trang 10-19, 2026.

HỘI NGHỊ KHOA HỌC

[CT4] Ngoc An Le, Xuan Dau Hoang, Thi Thu Trang Ninh, Huu Dung Le, Detecting Vietnamese Multi-label Reactionary Articles Using the Combination of Text and Image, FAIR 2025, Hanoi, Vietnam. Link: https://ebook365.vn/proceedings-of-the-18th-national-conference-on-fundamental-and-applied-information-technology-research-fair-2025-b1457.html.

[CT5] Ngoc An Le, Xuan Dau Hoang and Thi Thu Trang Ninh, A Multimodal Model for Detecting Vietnamese Toxic News Using Semi-supervised Learning, VCRIS 2025, Link: https://ieeexplore.ieee.org/document/11250571.