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

LUẬN ÁN TIẾN SĨ KỸ THUẬT

HÀ NỘI – 2026

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

LỜI CẢM ƠN

Trong quá trình học tập, nghiên cứu để hoàn thành đề tài luận án "Phát hiện tin bài độc hại tiếng Việt sử dụng kết hợp đặc trưng văn bản và hình ảnh", tôi đã nhận được sự hỗ trợ to lớn từ nhiều cá nhân và tổ chức. Ngoài nỗ lực của bản thân, sự giúp đỡ từ quý thầy cô hướng dẫn, nhà trường, các khoa, đơn vị công tác cùng gia đình đã tạo điều kiện thuận lợi nhất. Tôi xin bày tỏ lòng biết ơn sâu sắc đối với tất cả sự hỗ trợ quý báu đó.

Lời tri ân đầu tiên xin được gửi đến PGS.TS. Hoàng Xuân Dậu và TS. Đinh Trường Duy, đã tận tâm định hướng, chỉ bảo tận tình và luôn đồng hành cùng tôi xuyên suốt quá trình thực hiện nghiên cứu cũng như hoàn thành luận án này.

Tôi cũng xin bày tỏ lòng biết ơn chân thành đến Khoa Đào tạo Sau Đại học cùng Lãnh đạo Học viện Công nghệ Bưu chính Viễn thông. Sự hỗ trợ từ phía Học viện đã kiến tạo một môi trường học thuật thuận lợi, giúp tôi hoàn thành nghiên cứu này. Đồng thời, tôi xin cảm ơn Ban Lãnh đạo và toàn thể cán bộ, giảng viên Khoa An toàn thông tin, Học viện Công nghệ Bưu chính Viễn thông đã luôn hỗ trợ, động viên tôi trong suốt quá trình học tập và nghiên cứu.

Tôi cũng xin trân trọng gửi lời cảm ơn tới Lãnh đạo Trường Đại học Mở Hà Nội, Lãnh đạo Khoa CNTT Trường Đại học Mở Hà Nội, FITHOULAB, cùng toàn bộ thầy cô, đồng nghiệp trong khoa CNTT nơi tôi đang công tác, đã luôn hỗ trợ, động viên tôi trong suốt quá trình học tập và nghiên cứu.

Cuối cùng, tôi xin gửi lời cảm ơn vô hạn tới gia đình, i3 Network Systems đã luôn ở bên cạnh, chia sẻ, động viên tôi những lúc khó khăn, hỗ trợ cả về vật chất lẫn tinh thần trong suốt quá trình nghiên cứu.

Hà Nội, ngày .... tháng .... năm 2026

Nghiên cứu sinh

Lê Ngọc An

MỤC LỤC

LỜI CẢM ƠN i
DANH MỤC BẢNG BIỂU v
DANH MỤC HÌNH VẼ vii
DANH MỤC CÁC TỪ VIẾT TẮT viii
PHẦN MỞ ĐẦU 1
1. TÍNH CẤP THIẾT CỦA LUẬN ÁN 1
2. MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU CỦA LUẬN ÁN 4
3. CÁC ĐÓNG GÓP CỦA LUẬN ÁN 5
4. BỐ CỤC CỦA LUẬN ÁN 5
CHƯƠNG 1: TỔNG QUAN VỀ PHÁT HIỆN TIN BÀI ĐỘC HẠI TIẾNG VIỆT 7
1.1. KHÁI QUÁT VỀ TIN BÀI ĐỘC HẠI TIẾNG VIỆT 7
1.1.1. Giới thiệu tin bài độc hại và sự lan truyền trên không gian mạng 7
1.1.2. Tiêu chí nhận diện tin bài độc hại 8
1.1.3. Đặc điểm ngôn ngữ và cấu trúc của tin bài độc hại dạng văn bản 11
1.1.4. Đặc điểm và vai trò của hình ảnh trong tin bài độc hại 13
1.1.5. Sự cần thiết của phát hiện và ngăn chặn tin bài độc hại 16
1.1.6. Các thách thức trong phát hiện tin bài độc hại tiếng Việt 17
1.2. CÁC PHƯƠNG PHÁP PHÁT HIỆN TIN BÀI ĐỘC HẠI 18
1.2.1. Giới thiệu 18
1.2.2. Phát hiện tin bài độc hại dựa trên học có giám sát 19
1.2.2.1. Các nghiên cứu trong nước 19
1.2.2.2. Các nghiên cứu quốc tế 22
1.2.3. Phát hiện tin bài độc hại dựa trên học bán giám sát 30
1.2.3.1. Các nghiên cứu trong nước 30
1.2.3.2. Các nghiên cứu quốc tế 33
1.2.4. Ưu điểm và hạn chế của các đề xuất đã có 35
1.3. CÁC MÔ HÌNH HỌC MÁY SỬ DỤNG TRONG LUẬN ÁN 37
1.3.1. Các mô hình xử lý văn bản 37
1.3.1.1. BERT 37
1.3.1.2. RoBERTa 38
1.3.1.3. PhoBERT 38
1.3.2. Các mô hình xử lý ảnh 39
1.3.2.1. Xception 39
1.3.2.2. VGG19 40
1.3.2.3. Swin Transformer V2 41
1.3.3. Học bán giám sát trong phân loại dữ liệu 43
1.3.4. Các mô hình hình kết hợp cho xử lý văn bản và ảnh 44
1.3.5. Các độ đo đánh giá 47

1.4. THU THẬP TẬP DỮ LIỆU TIN BÀI TIẾNG VIỆT 48
1.4.1. Các khó khăn trong thu thập dữ liệu tin bài độc hại tiếng Việt 48
1.4.2. Thu thập, làm sạch và gán nhãn tin bài tiếng Việt 49
1.5. HƯỚNG NGHIÊN CỨU CỦA LUẬN ÁN 54
1.5.1. Các vấn đề giải quyết trong luận án 55
1.5.2. Kiến trúc mô hình tổng thể cho các hướng nghiên cứu của luận án 55
1.6. KẾT LUẬN CHƯƠNG 1 57
CHƯƠNG 2: PHÁT HIỆN TIN BÀI ĐỘC HẠI TIẾNG VIỆT SỬ DỤNG KẾT HỢP ĐẶC TRƯNG VĂN BẢN VÀ HÌNH ẢNH 59
2.1. ĐẶT VẤN ĐỀ 59
2.2. MÔ HÌNH PHÁT HIỆN SỬ DỤNG KẾT HỢP DỰA TRÊN ĐẶC TRƯNG VĂN BẢN VÀ HÌNH ẢNH 60
2.2.1. Giới thiệu mô hình phát hiện 61
2.2.2. Tiền xử lý dữ liệu 62
2.2.3. Huấn luyện và phát hiện 63
2.2.4. Cài đặt 69
2.3. THỬ NGHIỆM VÀ ĐÁNH GIÁ 69
2.3.1. Tập dữ liệu thử nghiệm 70
2.3.1.1. Bộ dữ liệu cho chế độ phát hiện nhị phân 70
2.3.1.2. Bộ dữ liệu cho chế độ phát hiện đa lớp 70
2.3.2. Các kịch bản thử nghiệm và kết quả 71
2.3.2.1. Kết quả thử nghiệm phát hiện nhị phân 71
2.3.2.2. Kết quả thử nghiệm phát hiện đa lớp 72
2.3.3. Nhận xét và đánh giá 74
2.3.3.1. Hiệu suất của mô hình trong chế độ phát hiện nhị phân 74
2.3.3.2. Hiệu suất của mô hình trong chế độ phát hiện đa lớp 75
2.4. KẾT LUẬN CHƯƠNG 76
CHƯƠNG 3: PHÁT HIỆN TIN BÀI ĐỘC HẠI TIẾNG VIỆT DỰA TRÊN HỌC BÁN GIÁM SÁT 78
3.1. ĐẶT VẤN ĐỀ 78
3.2. MÔ HÌNH PHÁT HIỆN DỰA TRÊN HỌC BÁN GIÁM SÁT 80
3.2.1. Giới thiệu mô hình phát hiện dựa trên học bán giám sát 80
3.2.2. Tiền xử lý dữ liệu 82
3.2.3. Huấn luyện và phát hiện 82
3.2.4. Cài đặt 86
3.3. THỬ NGHIỆM VÀ ĐÁNH GIÁ 86
3.3.1. Tập dữ liệu thử nghiệm 86
3.3.2. Các kịch bản thử nghiệm và kết quả 86
3.3.3. Nhận xét và đánh giá 95

3.4. KẾT LUẬN CHƯƠNG 98
KẾT LUẬN 99
DANH MỤC CÁC CÔNG TRÌNH CÔNG BỐ 102
TÀI LIỆU THAM KHẢO 103

DANH MỤC BẢNG BIỂU

Bảng 1.1 Các tham số TP, FP, FN và TN trong ma trận nhầm lẫn..... 47
Bảng 2.1. Kiến trúc Mạng Nơ-ron cho mô hình phát hiện đa lớp 65
Bảng 2.2. Hiệu suất phát hiện của mô hình đề xuất trong chế độ nhị phân và các mô hình đơn phương thức 71
Bảng 2.3. Ma trận nhầm lẫn của mô hình đề xuất trong chế độ phát hiện nhị phân cho 2 nhóm tin bài: tin bài thường, tin bài độc hại 72
Bảng 2.4. So sánh hiệu suất của mô hình đề xuất trong chế độ phát hiện nhị phân với nghiên cứu liên quan..... 72
Bảng 2.5. Hiệu suất phát hiện của mô hình đề xuất trong chế độ đa lớp và các mô hình kết hợp trên dữ liệu đánh giá..... 73
Bảng 2.6. Tỉ lệ phát hiện (DR) của mô hình đề xuất trong chế độ đa lớp và các mô hình kết hợp trên dữ liệu kiểm tra..... 73
Bảng 2.7. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình đề xuất trong chế độ đa lớp 74
Bảng 3.1. Hiệu suất phát hiện của mô hình đề xuất theo các ngưỡng tin cậy trong học bán giám sát với 50% dữ liệu huấn luyện có gán nhãn..... 87
Bảng 3.2. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 5% dữ liệu huấn luyện có gán nhãn..... 88
Bảng 3.3. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 10% dữ liệu huấn luyện có gán nhãn..... 88
Bảng 3.4. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 15% dữ liệu huấn luyện có gán nhãn..... 89
Bảng 3.5. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 20% dữ liệu huấn luyện có gán nhãn..... 89
Bảng 3.6. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 30% dữ liệu huấn luyện có gán nhãn..... 90
Bảng 3.7. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 50% dữ liệu huấn luyện có gán nhãn..... 90
Bảng 3.8. Hiệu suất phát hiện tổng thể của mô hình bán giám sát đề xuất dựa trên PhoBERT + Swin Transformer V2 theo tỷ lệ dữ liệu huấn luyện có nhãn..... 91
Bảng 3.9. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 5% dữ liệu huấn luyện có gán nhãn 92
Bảng 3.10. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 10% dữ liệu huấn luyện có gán nhãn 92
Bảng 3.11. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 15% dữ liệu huấn luyện có gán nhãn 93

Bảng 3.12. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 20% dữ liệu huấn luyện có gán nhãn 93
Bảng 3.13. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 30% dữ liệu huấn luyện có gán nhãn 93
Bảng 3.14. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 50% dữ liệu huấn luyện có gán nhãn 94
Bảng 3.15. Số lượng và tỷ lệ phát hiện đúng (DR) cho từng nhãn tin bài của mô hình bán giám sát đề xuất theo tỷ lệ dữ liệu huấn luyện có nhãn..... 94
Bảng 3.16. Hiệu suất phát hiện của mô hình đề xuất và các mô hình bán giám sát trước đó trên 30% dữ liệu huấn luyện có nhãn 95
Bảng 3.17. So sánh hiệu suất phát hiện của mô hình đề xuất sử dụng 30% dữ liệu huấn luyện có nhãn với mô hình kết hợp dựa trên học có giám sát 95

DANH MỤC HÌNH VẼ

Hình 1.1. Tin bài độc hại dạng hình ảnh..... 15
Hình 1.2. Tin bài độc hại vu khống bôi nhọ dạng hình ảnh..... 52
Hình 1.3. Tin bài độc hại kêu gọi chống đối dạng hình ảnh..... 53
Hình 1.4. Tin bài độc hại thúc đẩy tư tưởng cực đoan dạng hình ảnh..... 54
Hình 1.5. Kiến trúc tổng thể cho phát hiện tin bài độc hại..... 56
Hình 2.1. Kiến trúc của mô hình kết hợp cho phân loại đa lớp..... 62
Hình 3.1. Kiến trúc của mô hình phát hiện đa lớp dựa trên học bán giám sát..... 81
Hình 3.2. Độ đo F1 của mô hình đề xuất theo tỷ lệ dữ liệu huấn luyện có nhãn..... 92

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Từ gốc	Tiếng Việt
1	ACC	Accuracy	Độ chính xác tổng thể
2	ANN	Artificial Neural Network	Mạng nơ-ron nhân tạo
3	BERT	Bidirectional Encoder Representations from Transformers	Biểu diễn mã hóa hai chiều từ Transformer
4	BiLSTM	Bidirectional Long Short-Term Memory	Bố nhớ dài-ngắn hạn hai chiều
5	CE	Cross-Entropy	Hàm mất mát entropy chéo
6	CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
7	DL	Deep Learning	Học sâu
8	FNR	False Negative Rate	Tỷ lệ âm tính giả
9	FPR	False Positive Rate	Tỷ lệ dương tính giả
10	GPU	Graphics Processing Unit	Bộ xử lý đồ họa
11	LLM	Large Language Model	Mô hình ngôn ngữ lớn
12	LSTM	Long Short-Term Memory	Bố nhớ dài-ngắn hạn
13	ML	Machine Learning	Học máy
14	NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
15	PhoBERT	Vietnamese BERT model	Mô hình BERT cho tiếng Việt
16	PL	Pseudo-Labeling	Gán nhãn giả
17	PPV	Positive Predictive Value	Độ chính xác dương tính
18	ReLU	Rectified Linear Unit	Hàm kích hoạt ReLU
19	RNN	Recurrent Neural Network	Mạng nơ-ron hồi tiếp
20	RoBERTa	Robustly Optimized BERT Pretraining Approach	Phương pháp tiền huấn luyện BERT tối ưu
21	SSL	Semi-Supervised Learning	Học bán giám sát
22	SVM	Support Vector Machine	Máy véc tơ hỗ trợ
23	SwinV2	Swin Transformer V2	Mô hình Swin Transformer V2
24	TPR	True Positive Rate	Độ bao phủ
25	VGG	Visual Geometry Group	Mạng nơ-ron VGG

PHẦN MỞ ĐẦU

1. TÍNH CẤP THIẾT CỦA LUẬN ÁN

Không gian mạng đã trở thành một bộ phận quan trọng của đời sống xã hội hiện đại. Sự phát triển của Internet và các nền tảng mạng xã hội giúp hàng triệu người dùng tại Việt Nam hằng ngày chia sẻ, tìm kiếm và lan truyền thông tin. Theo báo cáo Digital 2023, Việt Nam hiện có hơn 77 triệu người dùng Internet và trên 70 triệu người dùng mạng xã hội, chiếm hơn 70% dân số [1]. Cùng với số lượng người dùng Internet lớn, lượng nội dung được tạo ra mỗi ngày là rất lớn, bao phủ nhiều lĩnh vực từ chính trị, kinh tế, văn hóa, xã hội đến đời sống thường nhật. Quá trình này góp phần thúc đẩy kết nối xã hội và mở rộng khả năng tiếp cận thông tin. Tuy nhiên, cùng với những lợi ích mang lại, không gian mạng cũng chứng kiến sự gia tăng và lan truyền các loại tin bài độc hại với tốc độ chóng mặt.

Tin bài độc hại ở đây được hiểu là những nội dung mang tính chất giả mạo, sai lệch, xuyên tạc, vu khống, kích động hoặc gây hoang mang trong cộng đồng [2]. Các loại tin này thường được dàn dựng có chủ ý, lợi dụng đặc thù của môi trường mạng để lan tỏa nhanh chóng, vượt ngoài khả năng kiểm soát của các cơ quan chức năng cũng như của chính người dùng. Chúng không chỉ làm nhiễu loạn dòng chảy thông tin, khiến công chúng khó phân biệt đúng sai, mà còn gây ra những hệ lụy nghiêm trọng đối với nhận thức, hành vi của cá nhân và cộng đồng, tác động trực tiếp đến sự ổn định chính trị xã hội và an ninh quốc gia [3].

Thực tế cho thấy, sự lan truyền của tin bài độc hại đã và đang để lại những hậu quả nặng nề. Một số tin giả, tin xuyên tạc về y tế từng gây hoang mang dư luận trong thời kỳ dịch bệnh [4]; các tin sai lệch về kinh tế xã hội làm giảm niềm tin của người dân vào chính sách, trong khi đó những nội dung độc hại, chống phá về chính trị lại được khai thác như công cụ để chia rẽ khối đại đoàn kết dân tộc. Đặc biệt, nhiều tin bài độc hại hiện nay không còn đơn thuần chỉ là văn bản, mà được kết hợp chặt chẽ với hình ảnh, video và kỹ thuật truyền thông số, nhằm gia tăng sức thuyết phục và độ lan tỏa [5]. Điều này đặt ra thách thức mới trong việc phát hiện và xử lý,

khi các hệ thống giám sát chủ yếu dựa vào phân tích nội dung văn bản thuần túy để phát hiện tin bài độc hại bộc lộ nhiều hạn chế.

Một minh chứng điển hình cho mức độ nguy hiểm của các tin bài độc hại là vụ tấn công khủng bố xảy ra tại hai trụ sở xã Ea Tiêu và Ea Ktur, huyện Cư Kuin, tỉnh Đắk Lắk vào sáng ngày 11 tháng 6 năm 2023. Theo thông tin chính thức từ Bộ Công an, hơn 90 đối tượng có vũ trang, được các thế lực phản động trong và ngoài nước xúi giục, đã tổ chức tấn công bất ngờ vào trụ sở chính quyền, sử dụng súng, dao, lựu đạn và bom xăng, gây hậu quả đặc biệt nghiêm trọng, khiến nhiều cán bộ, chiến sĩ và người dân thiệt mạng, tài sản bị hủy hoại và tình hình an ninh trật tự địa phương bị xáo trộn. Đáng chú ý là trước, trong và sau vụ việc, trên mạng xã hội đã xuất hiện hàng loạt tin bài, hình ảnh và video bị cắt ghép, xuyên tạc nhằm kích động tâm lý bất mãn, lôi kéo cộng đồng tham gia lan truyền thông tin sai lệch [4]. Những tin bài này không chỉ bóp méo sự thật, mà còn tiếp tay cho hoạt động khủng bố, gây hoang mang dư luận trong cả nước [6]. Trường hợp này cho thấy sự gắn kết chặt chẽ giữa thông tin độc hại trên mạng với hành vi bạo lực ngoài đời thực, đồng thời khẳng định tầm quan trọng và sự cấp bách của việc nghiên cứu các giải pháp hiệu quả cho phát hiện, ngăn chặn kịp thời.

Mặc dù đã có nhiều nghiên cứu về phát hiện tin giả hoặc nội dung độc hại, nhưng thực tiễn cho thấy bài toán này vẫn đối mặt với nhiều thách thức lớn:

Thách thức thứ nhất Dữ liệu đa phương thức phức tạp: phần lớn các nghiên cứu ban đầu về phát hiện tin giả tập trung vào dữ liệu đơn phương thức, trong khi các nghiên cứu gần đây ngày càng quan tâm đến việc khai thác đồng thời nhiều phương thức như văn bản, hình ảnh, video và âm thanh [5]. Đặc biệt, việc kết hợp thông tin văn bản và hình ảnh cho phép mô hình khai thác các đặc trưng bổ trợ giữa hai phương thức, qua đó nâng cao hiệu quả nhận diện so với việc chỉ sử dụng một nguồn thông tin [7]. Các mô hình xử lý đơn phương thức vì vậy có thể gặp hạn chế khi phải xử lý các tin bài có nội dung được thể hiện đồng thời qua nhiều dạng thông tin.

Thách thức thứ hai Thiếu dữ liệu gán nhãn: tin bài độc hại thường xuất hiện trong nhiều ngữ cảnh khác nhau, được ngụy trang tinh vi và khó thu thập, dẫn đến rất khó thực hiện việc xây dựng tập dữ liệu tin bài đủ lớn và được gán nhãn đầy đủ.

Điều này làm giảm hiệu quả của các mô hình học máy giám sát truyền thống do hoàn toàn phụ thuộc vào dữ liệu đã gán nhãn [8].

Thách thức thứ ba Đặc thù tiếng Việt và tính chất đa lớp: tiếng Việt chứa nhiều biến thể ngôn ngữ, cấu trúc diễn đạt linh hoạt và không chuẩn hóa, gây khó khăn cho xử lý ngôn ngữ tự nhiên. Đồng thời, các tin bài độc hại thường gồm nhiều lớp với nhiều mục đích khác nhau, do đó bài toán phát hiện tin bài độc hại cần được giải quyết theo hướng phân loại đa lớp.

Trong bối cảnh đó, sự phổ biến của các kỹ thuật học sâu và học bán giám sát đã mở ra hướng đi mới, có tiềm năng giải quyết những thách thức nêu trên. Trong khi các phương pháp học sâu cho phép xây dựng những mô hình có khả năng tự động trích xuất đặc trưng phức tạp từ dữ liệu văn bản và hình ảnh, các kỹ thuật học bán giám sát giúp tận dụng nguồn dữ liệu chưa được gán nhãn vốn rất phong phú trên mạng xã hội, nhằm nâng cao hiệu quả huấn luyện mà không phụ thuộc quá nhiều vào dữ liệu được gán nhãn thủ công [9]. Việc kết hợp hai hướng tiếp cận này không chỉ phù hợp với điều kiện thực tiễn của Việt Nam, nơi dữ liệu được gán nhãn còn hạn chế, mà còn có ý nghĩa học thuật quan trọng trong việc phát triển các mô hình kết hợp cho bài toán phân loại đa lớp [10] [11].

Đặc biệt cho đến nay, chưa có nghiên cứu chuyên biệt tập trung vào bài toán phát hiện tin bài độc hại tiếng Việt với sự kết hợp giữa đặc trưng văn bản và hình ảnh, đồng thời ứng dụng phương pháp học bán giám sát [5] [12]. Đây là khoảng trống nghiên cứu cả về mặt lý luận và thực tiễn. Khoảng trống này càng trở nên đáng chú ý khi xét đến nhu cầu cấp thiết trong công tác đảm bảo sự ổn định chính trị và định hướng dư luận xã hội. Việc nghiên cứu các mô hình phát hiện tin bài độc hại tiếng Việt theo hướng đa phương thức và đa lớp, có khả năng hoạt động hiệu quả trong điều kiện thiếu dữ liệu có nhãn, không chỉ đáp ứng yêu cầu cấp bách về mặt ứng dụng, mà còn đóng góp vào kho tàng tri thức khoa học trong lĩnh vực trí tuệ nhân tạo và xử lý ngôn ngữ tự nhiên tiếng Việt.

Xuất phát từ những lý do nêu trên, đề tài "Phát hiện tin bài độc hại tiếng Việt sử dụng kết hợp đặc trưng văn bản và hình ảnh" được lựa chọn nghiên cứu trong khuôn khổ luận án tiến sĩ này. Định hướng nghiên cứu của đề tài là phát triển các mô hình nhận diện tin bài độc hại có độ tin cậy cao, phù hợp với đặc thù ngôn ngữ

tiếng Việt, có khả năng vận hành trên dữ liệu đa phương thức gồm văn bản và hình ảnh, đồng thời thích ứng với điều kiện hạn chế về dữ liệu đã được gán nhãn. Đây là hướng nghiên cứu mang tính cấp thiết, vừa giải quyết những thách thức cấp bách của xã hội, vừa có ý nghĩa khoa học trong tiến trình phát triển các công nghệ thông minh phục vụ an ninh mạng và truyền thông lành mạnh tại Việt Nam.

2. MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU CỦA LUẬN ÁN

Mục tiêu tổng quát của luận án là xây dựng các mô hình dựa trên trí tuệ nhân tạo có khả năng phát hiện các tin bài độc hại tiếng Việt sử dụng kết hợp đặc trưng văn bản và hình ảnh. Các mô hình đề xuất hướng đến việc khắc phục những hạn chế của các phương pháp truyền thống vốn chỉ tập trung vào xử lý đơn phương thức văn bản, đồng thời tận dụng hiệu quả dữ liệu chưa gán nhãn thông qua học bán giám sát. Với mục tiêu này, luận án không chỉ nhằm đóng góp vào cơ sở khoa học trong lĩnh vực xử lý ngôn ngữ tự nhiên và thị giác máy tính cho tiếng Việt, mà còn hướng đến giá trị thực tiễn trong công tác bảo đảm an ninh mạng, định hướng dư luận xã hội và hỗ trợ cơ quan đảm bảo an ninh thông tin trong kỷ nguyên số.

Để đạt được mục tiêu trên, luận án tập trung vào một số mục tiêu cụ thể. Trước hết, luận án tiến hành phân tích, hệ thống hóa các nghiên cứu liên quan trên thế giới và trong nước về phát hiện tin giả, tin độc hại, từ đó làm rõ khoảng trống nghiên cứu đối với bài toán phát hiện tin giả, tin độc hại tiếng Việt. Tiếp theo, luận án đề xuất một mô hình kết hợp cho phát hiện tin bài độc hại, trong đó sử dụng PhoBERT để trích xuất đặc trưng ngôn ngữ từ nội dung văn bản và Swin Transformer V2 để trích xuất đặc trưng hình ảnh của tin bài. Các đặc trưng này sau đó được hợp nhất trong một không gian chung nhằm phục vụ bài toán phân loại đa lớp, cho phép nhận diện và phân loại nhiều dạng tin bài độc hại vào các nhóm khác nhau như vu khống bôi nhọ, kêu gọi chống đối, thúc đẩy tư tưởng cực đoan. Tiếp theo, luận án áp dụng phương pháp học bán giám sát theo hướng gán nhãn giả (pseudo-labeling) [13] để khai thác nguồn dữ liệu chưa gán nhãn, qua đó nâng cao hiệu quả huấn luyện mô hình trong điều kiện hạn chế về dữ liệu có nhãn.

Phạm vi nghiên cứu của luận án được xác định nhằm đảm bảo tính khả thi và tính ứng dụng. Về loại hình dữ liệu, luận án tập trung vào hai phương thức chính là các tin bài dưới dạng văn bản và hình ảnh, không mở rộng sang các tin bài dưới

dạng âm thanh hoặc video. Về ngôn ngữ, luận án chỉ giới hạn trong tiếng Việt, phù hợp với bối cảnh nghiên cứu và nhu cầu cấp thiết của xã hội Việt Nam. Về dạng độc hại, luận án ưu tiên xử lý các loại tin bài mang tính độc hại liên quan đến chính trị như xuyên tạc, vu khống, kích động và gây hoang mang dư luận; đây là những dạng tin bài độc hại có nguy cơ cao đối với an ninh xã hội và trật tự công cộng. Về dữ liệu huấn luyện, luận án khai thác tập hợp dữ liệu có gán nhãn kết hợp với dữ liệu chưa gán nhãn, phản ánh đúng thực tiễn khi dữ liệu tin bài có nhãn tiếng Việt còn hạn chế.

Như vậy, mục tiêu và phạm vi nghiên cứu của luận án vừa đảm bảo tính bao quát về mặt khoa học, vừa hướng đến khả năng ứng dụng cao trong thực tiễn. Sự tập trung vào dữ liệu tiếng Việt, tin bài độc hại dạng văn bản và hình ảnh, cùng với cách tiếp cận đa lớp và bán giám sát, sẽ góp phần tạo nên một mô hình phát hiện tin độc hại có tính mới, có giá trị ứng dụng trong việc bảo vệ không gian thông tin và an ninh mạng tại Việt Nam.

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 được bố cục thành ba chương với nội dung như sau:

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

Chương này trình bày những cơ sở lý luận và thực tiễn liên quan đến bài toán phát hiện tin bài độc hại. Nội dung bao gồm: khái quát về tin bài độc hại, đặc điểm ngôn ngữ và hình ảnh trong các tin bài, các thách thức trong phát hiện, cũng như sự cần thiết của việc nghiên cứu và ngăn chặn. Chương cũng tổng hợp, phân tích các phương pháp phát hiện tin bài độc hại trong và ngoài nước, từ học có giám sát đến học bán giám sát, đồng thời chỉ ra ưu nhược điểm của các mô hình hiện có. Ngoài ra, chương giới thiệu một số mô hình học sâu điển hình được sử dụng trong luận án (PhoBERT, Swin Transformer V2…), các độ đo đánh giá và phương pháp thu thập dữ liệu. Trên cơ sở đó, chương 1 xác định rõ khoảng trống nghiên cứu mà luận án sẽ tập trung giải quyết.

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

Chương này đề xuất mô hình phát hiện tin bài độc hại tiếng Việt đa phương thức. Nội dung chương bao gồm: cơ sở hình thành hướng tiếp cận đa phương thức cho bài toán phát hiện tin bài trên không gian mạng, quy trình tiền xử lý dữ liệu, thiết kế kiến trúc mô hình phát tin bài độc hại đa lớp kết hợp đặc trưng văn bản và hình ảnh, đồng thời xây dựng các kịch bản thử nghiệm và phân tích kết quả đạt được. Thông qua các thực nghiệm và so sánh, chương 2 làm rõ hiệu quả của mô hình đề xuất trong việc khai thác đồng thời thông tin ngôn ngữ và hình ảnh.

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

Chương này mở rộng nghiên cứu bằng việc áp dụng học bán giám sát, nhằm tận dụng nguồn dữ liệu chưa gán nhãn phong phú trên không gian mạng. Luận án đề xuất các mô hình phát hiện đa lớp kết hợp với gán nhãn giả, đồng thời mô tả chi tiết quá trình tiền xử lý, huấn luyện và phát hiện. Các thực nghiệm được thực hiện để đánh giá hiệu quả của cách tiếp cận này, trong đó chú trọng đến khả năng thích ứng của mô hình khi tỷ lệ dữ liệu nhãn giảm xuống. Chương 3 khẳng định vai trò quan trọng của học bán giám sát trong việc nâng cao hiệu quả phát hiện tin độc hại tiếng Việt, đặc biệt trong điều kiện thiếu dữ liệu nhãn.

Phần kết luận tổng hợp lại các kết quả chính của luận án, khẳng định những đóng góp lý luận và thực tiễn, đồng thời nêu lên các hướng nghiên cứu tiếp theo.

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

Chương này tập trung trình bày một cách hệ thống cơ sở lý luận và thực tiễn liên quan đến phát hiện tin bài độc hại tiếng Việt. Nội dung chính gồm: (i) khái quát về tin bài độc hại, tiêu chí nhận diện và đặc điểm ngôn ngữ hình ảnh, (ii) tổng hợp các phương pháp phát hiện, từ học có giám sát đến học bán giám sát, cả trong nước và quốc tế, (iii) giới thiệu các mô hình học sâu được sử dụng trong luận án, (iv) trình bày phương pháp thu thập tập dữ liệu tin bài tiếng Việt. Trên cơ sở đó, chương 1 chỉ ra các khoảng trống nghiên cứu, làm nền tảng cho việc đề xuất mô hình ở các chương tiếp theo.

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 được hiểu là những nội dung thông tin có tính sai lệch, bóp méo, thiếu kiểm chứng hoặc mang yếu tố tiêu cực, có khả năng gây tác động xấu đến nhận thức, cảm xúc và hành vi của người tiếp nhận. Các tin bài này có thể bao gồm: tin giả, tin sai sự thật, thông tin giật gân, nội dung vu khống, xuyên tạc hoặc tin tức mang tính kích động. Điểm chung của tin bài độc hại là khả năng gây nhiễu loạn dòng chảy thông tin, làm suy giảm niềm tin vào các kênh truyền thông chính thống và tiềm ẩn nguy cơ ảnh hưởng tiêu cực đến ổn định xã hội.

Sự lan truyền của tin bài độc hại trở nên đặc biệt mạnh mẽ trong bối cảnh các nền tảng mạng xã hội phát triển. Các thông tin thường ưu tiên nội dung gây chú ý hoặc có tính lan tỏa cao, khiến cho những tin bài độc hại vốn thường được trình bày dưới dạng giật tít, sử dụng ngôn từ cực đoan hoặc kết hợp hình ảnh gây sốc trở nên dễ dàng tiếp cận và thu hút lượng lớn người đọc. Điều này làm gia tăng tốc độ phát tán và mức độ ảnh hưởng của tin bài độc hại so với các tin bài từ nguồn thông tin chính thống.

Thực tế đã ghi nhận nhiều sự kiện cho thấy mức độ nguy hại của việc phát tán thông tin độc hại. Minh chứng điển hình là sau vụ việc xảy ra tại Đắk Lắk năm 2023, nhiều tin bài sai lệch và xuyên tạc đã nhanh chóng được lan truyền trên mạng

xã hội, tạo nên sự hoang mang và những tranh luận thiếu căn cứ trong cộng đồng [4]. Trường hợp này cho thấy tin bài độc hại không chỉ tồn tại ở mức tin giả thông thường mà còn có khả năng tác động tiêu cực đến dư luận xã hội nếu không được phát hiện và xử lý kịp thời.

Trong luận án này, "tin bài độc hại" được sử dụng như khái niệm bao trùm đối với các tin bài có nội dung sai lệch, xuyên tạc, vu khống, kích động hoặc thúc đẩy tư tưởng cực đoan. Các khái niệm như tin giả, thông tin sai sự thật và ngôn từ thù ghét có phạm vi và đặc trưng riêng, do đó không được sử dụng đồng nhất với khái niệm tin bài độc hại. Luận án tập trung vào nhóm tin bài độc hại liên quan đến chính trị, được thể hiện dưới dạng văn bản kết hợp hình ảnh, và được phân loại thành ba nhóm độc hại: "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 được tham chiếu từ các quy định pháp luật hiện hành có liên quan đến an ninh mạng, hoạt động báo chí và phòng, chống tin giả, tin sai sự thật. Cụ thể, luận án cập nhật và tham chiếu Luật An ninh mạng số 116/2025/QH15, có hiệu lực từ ngày 01/07/2026 [14]; Luật Báo chí số 126/2025/QH15, có hiệu lực từ ngày 01/07/2026 [15]; và Nghị định số 328/2026/NĐ-CP của Chính phủ quy định phòng, chống tin giả, tin sai sự thật, có hiệu lực từ ngày 05/10/2026 [16].

Các văn bản trên được sử dụng làm cơ sở tham chiếu để xác định phạm vi và đặc trưng nội dung của dữ liệu nghiên cứu, không đồng nhất các nhãn dữ liệu của luận án với các khái niệm hoặc chế tài pháp lý. Trên cơ sở đó, luận án duy trì ba nhóm tin bài độc hại phục vụ bài toán phân loại đa lớp gồm: "Vu khống, bôi nhọ", "Kêu gọi chống đối" và "Thúc đẩy tư tưởng cực đoan". Việc phân nhóm được xây dựng dựa trên nội dung dữ liệu, mục tiêu nghiên cứu và các quy định pháp luật có liên quan.

Cụ thể, Công văn số 7169/VPCP-NC ngày 12/9/2012 của Văn phòng Chính phủ đã nhấn mạnh yêu cầu xử lý nghiêm các hành vi đăng tải, phát tán thông tin xuyên tạc, vu khống nhằm bôi nhọ bộ máy lãnh đạo, kích động chống phá Đảng và

Nhà nước. Công văn đồng thời yêu cầu các cơ quan chức năng tăng cường phối hợp trong việc phát hiện, ngăn chặn các trang mạng đưa tin sai sự thật, cũng như chủ động cung cấp thông tin chính xác để phản bác và định hướng dư luận [17]. Văn bản này cho thấy rõ vai trò của việc nhận diện sớm và xử lý thông tin độc hại trong bảo đảm an ninh và ổn định xã hội.

Ngoài ra, các bài viết trên Tạp chí Cộng sản cũng nhấn mạnh tầm quan trọng của việc "nhận diện các quan điểm sai trái, thù địch" nhằm phục vụ công tác đấu tranh phản bác, bởi những nội dung này có tác động trực tiếp đến nhận thức, niềm tin và sự ổn định tư tưởng trong xã hội [18]. Mặc dù tiếp cận dưới góc nhìn tư tưởng chính trị, các phân tích này cung cấp thêm cơ sở tham khảo quan trọng cho việc nhận diện thông tin sai lệch và độc hại trong môi trường truyền thông số.

Dưới góc độ luật pháp, Điều 13, Luật An ninh mạng 2025 đã quy định cụ thể các hành vi bị nghiêm cấm liên quan đến thông tin chống Nhà nước Cộng hòa xã hội chủ nghĩa Việt Nam, bao gồm: (1) tuyên truyền xuyên tạc, phỉ báng chính quyền nhân dân, tuyên truyền chống Nhà nước, kích động bạo loạn, chia rẽ dân tộc tôn giáo, (2) bịa đặt, gây hoang mang trong Nhân dân, (3) kích động biểu tình, bạo loạn, gây rối an ninh, trật tự, (4) làm nhục, vu khống tổ chức, cá nhân, xâm phạm danh dự, uy tín của tổ chức, hoặc gây thiệt hại đến lợi ích hợp pháp và (5) đăng tải thông tin sai sự thật xâm phạm lợi ích của Nhà nước, quyền và lợi ích hợp pháp của tổ chức, cá nhân, phá hoại chính sách kinh tế xã hội.

Những quy định pháp luật có liên quan là cơ sở tham chiếu để luận án xác định phạm vi và xây dựng các tiêu chí phân loại dữ liệu tin bài độc hại phục vụ mục tiêu nghiên cứu. Từ cơ sở pháp lý trên, luận án xác định ba nhóm tiêu chí lớn tương ứng với ba nhãn độc hại được sử dụng trong bài toán phân loại tin bài đa lớp, gồm: "Vu khống, bôi nhọ", "Kêu gọi chống đối", và "Thúc đẩy tư tưởng cực đoan". Nội dung tiếp theo là mô tả chi tiết về các nhãn tin bài độc hại nêu trên:

(1) Nhóm tiêu chí "Vu khống, bôi nhọ"

Nhóm tiêu chí "Vu khống, bôi nhọ" được xây dựng trên cơ sở tham chiếu các quy định pháp luật hiện hành có liên quan đến hành vi đưa thông tin sai sự thật, xúc phạm danh dự, uy tín, nhân phẩm của tổ chức, cá nhân, đồng thời kết hợp với đặc

điểm nội dung của dữ liệu nghiên cứu. Cụ thể, tin bài được xếp vào nhóm "Vu khống, bôi nhọ" khi có một hoặc nhiều đặc điểm sau:

- Xúc phạm nghiêm trọng danh dự, uy tín, nhân phẩm của cá nhân hoặc tổ chức, đặc biệt là các cơ quan, cán bộ, lãnh đạo cơ quan Nhà nước.
- Đưa ra thông tin bịa đặt, sai sự thật hoặc không có căn cứ xác minh, gây tổn hại đến uy tín, danh dự hoặc quyền lợi hợp pháp của tổ chức, cá nhân.
- Cắt xén, bóp méo sự kiện, phát ngôn hoặc hình ảnh nhằm dẫn dắt người đọc đến nhận thức sai lệch.
- Kết hợp văn bản với hình ảnh, tài liệu hoặc biểu tượng không rõ nguồn gốc để tăng tính thuyết phục cho nội dung vu khống.

Nhóm tiêu chí này phản ánh đúng bản chất pháp lý của hành vi "làm nhục, vu khống" và "bịa đặt, sai sự thật" được quy định tại Điều 13.

(2) Nhóm tiêu chí "Kêu gọi chống đối"

Nhóm tiêu chí "Kêu gọi chống đối" được xây dựng trên cơ sở tham chiếu các quy định pháp luật hiện hành liên quan đến hành vi kích động, xúi giục, lôi kéo hoặc cổ súy các hành động gây ảnh hưởng đến an ninh, trật tự xã hội, kết hợp với đặc điểm nội dung của dữ liệu nghiên cứu. Tin bài thuộc nhóm này thường có các đặc điểm sau:

- Kêu gọi, vận động, xúi giục hoặc ám chỉ việc chống lại chính quyền nhân dân, cơ quan Nhà nước hoặc người thi hành công vụ.
- Lôi kéo, kích động tụ tập đông người, biểu tình, gây rối trật tự công cộng hoặc cản trở hoạt động của cơ quan, tổ chức.
- Sử dụng ngôn ngữ mang tính kích động, đe dọa, đối đầu, làm gia tăng tâm lý bất mãn hoặc hành vi phản kháng trong cộng đồng.
- Truyền tải nội dung cổ súy các hành động tập thể mang tính chống đối, kể cả dưới dạng gián tiếp.

Nhóm tiêu chí này phản ánh rõ hành vi kích động, xúi giục và gây mất ổn định an ninh, trật tự theo quy định của Điều 13.

(3) Nhóm tiêu chí "Thúc đẩy tư tưởng cực đoan"

Nhóm tiêu chí "Thúc đẩy tư tưởng cực đoan" được xây dựng trên cơ sở tham chiếu các quy định pháp luật hiện hành liên quan đến hành vi tuyên truyền, xuyên tạc, kích động thù hận, chia rẽ hoặc lan truyền thông tin sai lệch gây ảnh hưởng tiêu cực đến nhận thức xã hội, kết hợp với đặc điểm dữ liệu nghiên cứu. Tin bài thuộc nhóm này thường có các đặc điểm:

- Tuyên truyền xuyên tạc, phỉ báng chính quyền nhân dân hoặc phủ định vai trò của Nhà nước, pháp luật.
- Lan truyền các luận điệu mang tính chiến tranh tâm lý, kích động thù hận, chia rẽ xã hội hoặc gây nghi ngờ đối với thể chế.
- Thổi phồng các thông tin tiêu cực, suy diễn thiếu căn cứ nhằm tạo cảm giác khủng hoảng, mất ổn định hoặc sụp đổ.
- Kết hợp văn bản với hình ảnh, biểu tượng hoặc ngôn ngữ ẩn dụ để củng cố hệ tư tưởng cực đoan và tăng khả năng lan truyền.

Nhóm tiêu chí này phản ánh hành vi tuyên truyền chống Nhà nước và phát tán thông tin sai sự thật gây hoang mang trong 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 vẫn là phương tiện chủ đạo để truyền tải thông tin và định hướng nhận thức người tiếp nhận. Đối với tin bài độc hại, đặc biệt là nhóm tin mang nội dung độc hại về chính trị đã được xác định ở mục 1.1.2, các đặc điểm ngôn ngữ và cấu trúc văn bản đóng vai trò then chốt trong việc hình thành tác động tiêu cực và thúc đẩy quá trình lan truyền. Việc phân tích các đặc trưng này không chỉ có ý nghĩa về mặt ngôn ngữ học, mà còn là cơ sở quan trọng cho việc xây dựng các mô hình phát hiện tự động trong luận án.

Thứ nhất, về ngôn ngữ sử dụng: Tin bài độc hại thường khai thác lối diễn đạt mang tính giật gân, cực đoan, nhằm gây ấn tượng mạnh với người tiếp nhận. Các từ ngữ cường điệu, mang tính cảm xúc như "kinh hoàng", "gây chấn động", "nguy hiểm", "sụp đổ" xuất hiện với tần suất cao [19] [20]. Việc lựa chọn từ ngữ như vậy khiến văn bản dễ tạo hiệu ứng tâm lý sợ hãi, phẫn nộ hoặc tò mò, từ đó thúc đẩy

hành vi lan truyền mà không qua kiểm chứng. Đồng thời, tin độc hại thường thể hiện tính thiên lệch rõ rệt trong diễn đạt: thay vì trình bày đa chiều, khách quan, chúng thường khẳng định một chiều, nhấn mạnh yếu tố tiêu cực và bỏ qua bối cảnh đầy đủ của sự kiện.

Thứ hai, về cấu trúc tiêu đề: Tiêu đề đóng vai trò quan trọng trong việc định hướng sự chú ý của người đọc, và trong tin độc hại, đây là nơi tập trung nhiều thủ pháp gây hiểu nhầm. Các tiêu đề thường được xây dựng ngắn gọn, trực diện nhưng mang tính khơi gợi hoặc cường điệu, chẳng hạn sử dụng câu hỏi giả định ("Điều gì đang thực sự xảy ra sau sự kiện này?"), tạo cảm giác cấp bách ("Cần biết ngay trước khi quá muộn") hoặc cắt bỏ thông tin bối cảnh, khiến người đọc dễ bị dẫn dắt theo hướng suy đoán [20]. Điều này khiến khoảng cách giữa tiêu đề và nội dung thường rất lớn, làm gia tăng nguy cơ hiểu sai thông tin.

Thứ ba, về tính logic trong lập luận: Tin bài độc hại thường không tuân thủ các nguyên tắc lập luận chặt chẽ. Thay vì dựa trên bằng chứng cụ thể, chúng thiên về việc sử dụng suy đoán, ngụy biện và cảm tính. Các luận điểm thường được củng cố bằng những nguồn dẫn không rõ ràng ("theo một số nguồn tin", "một nhân vật giấu tên tiết lộ"), hoặc bằng cách lặp lại một quan điểm nhiều lần để tạo cảm giác thuyết phục. Kỹ thuật "nửa thật nửa giả" cũng phổ biến: thông tin có thật được sử dụng làm nền tảng, sau đó lồng ghép thêm chi tiết sai lệch hoặc suy diễn thiếu căn cứ nhằm hướng người đọc đến một kết luận nhất định [21] [22].

Thứ tư, về cấu trúc văn bản: Nhiều tin độc hại có kết cấu được thiết kế theo hướng tuyên truyền: mở đầu bằng tiêu đề gây sốc, phần thân bài triển khai thông tin theo hướng cảm xúc, nhấn mạnh yếu tố tiêu cực, và kết thúc bằng lời kêu gọi hành động hoặc chia sẻ. Sự lặp lại có chủ đích về từ ngữ và ý tưởng khiến cho thông điệp tiêu cực được khắc sâu dần trong nhận thức người tiếp nhận [19] [20].

Thứ năm, về đặc thù ngôn ngữ tiếng Việt: Trong bối cảnh tiếng Việt, các tin độc hại thường tận dụng tính đa nghĩa, giàu sắc thái tu từ để tăng cường sức tác động. Việc sử dụng ẩn dụ, phóng đại, châm biếm hoặc tiếng lóng phổ biến trong giới trẻ là thủ pháp quen thuộc. Bên cạnh đó, nhiều tin còn cố tình viết sai chính tả, thay đổi ký tự, hoặc sử dụng biểu tượng cảm xúc (emoji) nhằm qua mặt hệ thống

kiểm duyệt tự động. Đặc trưng này vừa gây khó khăn cho người đọc trong việc tiếp nhận khách quan, vừa tạo ra thách thức cho các hệ thống phát hiện tự động [19].

Ví dụ minh họa cho một tin bài độc hại dạng văn bản:

Tin độc hại: "Theo một số nguồn tin giấu tên, Việt Nam đang đứng trên bờ vực sụp đổ kinh tế vì nợ công tăng không kiểm soát. Chính quyền đã che giấu con số thực sự và tiếp tục vắt kiệt sức dân thông qua các loại thuế vô lý. Người dân cần tỉnh táo, không thể tiếp tục im lặng trước sự thật bị che đậy này!"

(Chú thích: Đây là nội dung văn bản độc hại hại về chính trị vi phạm điều 13 luật an ninh mạng 2025, chỉ phục vụ cho nghiên cứu này không được sử dụng với mục đích khác)

Tóm lại, tin bài độc hại dạng văn bản có thể được nhận diện thông qua một loạt đặc điểm về ngôn ngữ và cấu trúc: xu hướng giật gân, thiên lệch, thiếu logic, sử dụng nguồn tin mơ hồ, tiêu đề gây hiểu nhầm, lặp lại thông điệp và vận dụng các thủ pháp ngôn ngữ đặc thù. Những đặc trưng này chính là nền tảng để định hướng nghiên cứu và thiết kế các mô hình học máy, học sâu trong việc phát hiện, phân loại và ngăn chặn sự lan truyền của thông tin độc hại trong không gian mạng.

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 quan trọng tạo nên sức lan tỏa và mức độ thuyết phục của tin bài độc hại. Trong bối cảnh truyền thông hiện nay, khi người dùng thường lướt nhanh qua các dòng tin trên mạng xã hội, hình ảnh trở thành công cụ trực quan, gây ấn tượng tức thời và góp phần định hướng cách tiếp nhận thông tin. Việc nghiên cứu đặc điểm ngôn ngữ hình ảnh trong tin bài độc hại vì vậy là một yêu cầu cần thiết để hiểu đúng cơ chế lan truyền và mức độ nguy hại của loại thông tin này.

Thứ nhất, về đặc điểm hình thức: Hình ảnh trong tin độc hại thường được lựa chọn hoặc xử lý nhằm làm nổi bật tính tiêu cực, gây sốc hoặc gợi cảm xúc mạnh. Các kỹ thuật chỉnh sửa cơ bản như cắt ghép, thêm hiệu ứng màu sắc, phóng đại chi tiết, hoặc lựa chọn góc chụp đặc biệt được sử dụng để tạo cảm giác sự việc nghiêm trọng hơn thực tế [23]. Nhiều trường hợp, hình ảnh được lấy từ sự kiện khác nhưng

được chú thích sai lệch để đánh lừa người xem [24]. Chính sự "ngụy trang" này làm cho thông tin trở nên khó kiểm chứng và dễ tạo ra sự nhầm lẫn trong dư luận.

Thứ hai, về vai trò bổ trợ cho văn bản: Trong nhiều tin bài độc hại, văn bản và hình ảnh không tồn tại độc lập mà bổ sung cho nhau nhằm gia tăng tính thuyết phục. Văn bản có thể đưa ra một luận điểm chưa đủ cơ sở, nhưng hình ảnh đi kèm đóng vai trò như một "chứng cứ trực quan" khiến người đọc dễ dàng tin tưởng [25]. Ở chiều ngược lại, hình ảnh không rõ nguồn gốc hoặc không liên quan, khi được gắn kèm với lời dẫn giật gân, có thể tạo ra một mối liên tưởng sai lầm, khiến thông tin độc hại lan tỏa nhanh hơn.

Thứ ba, về vai trò trong kích thích cảm xúc: Hình ảnh có khả năng tác động trực tiếp đến thị giác và cảm xúc, từ đó ảnh hưởng mạnh đến thái độ và hành vi của người tiếp nhận [26]. Một bức ảnh gợi sự thương cảm, phẫn nộ hay sợ hãi có sức thuyết phục lớn hơn nhiều so với văn bản thuần túy. Tin bài độc hại thường khai thác khía cạnh này, sử dụng hình ảnh trẻ em, người cao tuổi, nạn nhân tai nạn hoặc các sự kiện bạo lực để gây sốc, đánh thẳng vào cảm xúc của cộng đồng mạng. Đây là yếu tố giải thích vì sao nhiều tin giả hoặc tin sai lệch có thể lan truyền với tốc độ nhanh chóng chỉ nhờ vào một hình ảnh ấn tượng.

Thứ tư, về vai trò trong cơ chế lan truyền: Trong môi trường mạng xã hội, các thuật toán thường ưu tiên hiển thị nội dung có hình ảnh hoặc video, do chúng thu hút nhiều lượt tương tác hơn so với văn bản thuần [25]. Điều này vô hình trung khiến tin bài độc hại có sử dụng hình ảnh dễ đạt được mức độ phủ sóng rộng. Người dùng có xu hướng chia sẻ hình ảnh mà ít khi kiểm chứng tính xác thực của chúng, từ đó làm gia tăng vòng xoáy lan truyền của thông tin độc hại.

Thứ năm, về thách thức trong phát hiện: Sự kết hợp giữa văn bản và hình ảnh đặt ra thách thức lớn cho việc phát hiện tin độc hại. Trong nhiều trường hợp, văn bản và hình ảnh riêng lẻ có thể không mang tính độc hại, nhưng khi kết hợp lại sẽ tạo ra thông tin sai lệch [27]. Ví dụ, một bức ảnh cũ được gắn với chú thích về sự kiện mới, hoặc một hình ảnh trung tính nhưng đi kèm lời dẫn sai lệch, đều có thể trở thành tin độc hại. Điều này cho thấy cần thiết phải nghiên cứu các phương pháp phân tích đa phương thức, trong đó đặc trưng hình ảnh và văn bản được xử lý đồng thời để nhận diện chính xác bản chất nội dung.

Ví dụ minh họa cho một tin bài độc hại dạng hình ảnh:



Hình 1.1. Tin bài độc hại dạng hình ảnh

(Chú thích: Đây là nội dung hình ảnh độc hại hại về chính trị vi phạm điều 13 luật an ninh mạng 2025, chỉ phục vụ cho nghiên cứu này không được sử dụng với mục đích khác)

Tóm lại, hình ảnh trong tin bài độc hại vừa là yếu tố tạo ra sức hấp dẫn, vừa là công cụ ngụy trang nguy hiểm, khiến thông tin trở nên khó kiểm chứng và dễ gây tác động tiêu cực đến dư luận xã hội. Vai trò của hình ảnh không chỉ dừng lại ở chức năng minh họa mà còn tham gia trực tiếp vào quá trình định hình nhận thức, kích thích cảm xúc và thúc đẩy lan truyền. Chính vì vậy, việc nghiên cứu và xử lý đặc trưng hình ảnh trong mô hình phát hiện tin độc hại là yêu cầu tất yếu, đặc biệt trong bối cảnh phát triển mạnh mẽ của các nền tảng truyền thông đa phương tiện hiện nay.

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

Trong những năm gần đây, cùng với sự phát triển của Internet và mạng xã hội, tin bài độc hại xuất hiện ngày càng nhiều và có tốc độ lan truyền nhanh chóng. Các nội dung sai sự thật, bóp méo hoặc giật gân không chỉ làm rối loạn môi trường thông tin mà còn gây tác động tiêu cực đến nhận thức và tâm lý của người dân. Nếu không được kiểm soát, tình trạng này có thể làm giảm niềm tin của xã hội vào các kênh thông tin chính thống, tạo ra sự hoang mang và tiềm ẩn nguy cơ bất ổn trong dư luận. Chính vì vậy, việc phát hiện và ngăn chặn tin bài độc hại là nhiệm vụ cấp thiết trong bối cảnh hiện nay.

Về phía cơ quan quản lý, nhiều chính sách và văn bản pháp luật đã được ban hành nhằm kiểm soát và xử lý nguồn tin độc hại. Tuy nhiên, do lượng thông tin phát sinh hằng ngày quá lớn và tốc độ lan truyền quá nhanh, việc chỉ dựa vào biện pháp hành chính hoặc kiểm duyệt thủ công là không khả thi. Cần có sự hỗ trợ của công nghệ, đặc biệt là các hệ thống trí tuệ nhân tạo, để phát hiện và xử lý tin bài độc hại một cách tự động, nhanh chóng và chính xác hơn.

Ở góc độ công nghệ, sự tinh vi trong cách thức phát tán tin độc hại đang là thách thức lớn. Nhiều nội dung được tạo ra bằng cách pha trộn thông tin thật giả, lồng ghép hình ảnh minh họa để tăng sức thuyết phục, hoặc sử dụng cách diễn đạt mơ hồ nhằm vượt qua các công cụ kiểm duyệt. Các phương pháp truyền thống dựa trên từ khóa hay bộ quy tắc vì vậy ngày càng bộc lộ hạn chế. Giải pháp phù hợp là nghiên cứu và áp dụng các mô hình học sâu, học bán giám sát có khả năng xử lý dữ liệu đa dạng và khai thác hiệu quả cả những tập dữ liệu chưa được gán nhãn đầy đủ.

Ở phạm vi quốc tế, nhiều quốc gia đã đầu tư nghiên cứu, triển khai hệ thống phát hiện tin giả và tin độc hại, song phần lớn tập trung vào tiếng Anh hoặc các ngôn ngữ phổ biến khác. Theo khảo sát của NCS, đến nay vẫn chưa có nhiều công trình chuyên biệt nghiên cứu ở trong nước về phát hiện tin bài độc hại tiếng Việt, nhất là tin bài ở dạng đa phương thức, trong khi đặc điểm ngôn ngữ và văn hóa sử dụng mạng lại có những nét riêng. Khoảng trống này càng nhấn mạnh sự cần thiết của việc xây dựng các mô hình phát hiện tin độc hại dành riêng cho tiếng Việt, đặc biệt là tin bài ở dạng đa phương thức, đáp ứng nhu cầu thực tiễn trong nước.

Từ những lý do trên có thể thấy rằng, phát hiện và ngăn chặn tin bài độc hại không chỉ là vấn đề kỹ thuật, mà còn liên quan trực tiếp đến việc giữ gìn môi trường thông tin trong sạch, bảo đảm an toàn cho cộng đồng và nâng cao hiệu quả quản lý xã hội. Đây là nhiệm vụ vừa mang ý nghĩa khoa học, vừa có giá trị ứng dụng thực tiễn cao.

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 không đơn thuần chỉ là áp dụng các mô hình có sẵn từ ngôn ngữ khác, mà chứa đựng nhiều thách thức đặc thù. Những khó khăn này đến từ cả yếu tố dữ liệu, đặc điểm ngôn ngữ, sự đa dạng trong cách thức thể hiện nội dung cũng như sự thay đổi liên tục của các hình thức lan truyền và ngụy trang thông tin độc hại.

Trước hết, vấn đề dữ liệu là một rào cản lớn. Tin bài độc hại tiếng Việt xuất hiện ở nhiều định dạng, từ văn bản ngắn gọn trên mạng xã hội cho đến các bài viết dài kèm hình ảnh minh họa hoặc đồ họa chế tác. Việc thu thập đủ dữ liệu đại diện cho các dạng thức này đã khó, trong khi dữ liệu có nhãn đáng tin cậy lại càng khan hiếm. Hầu hết nguồn dữ liệu tin bài độc hại tiếng Việt hiện nay chưa được gán nhãn đầy đủ hoặc thống nhất, gây khó khăn cho quá trình huấn luyện các mô hình học máy.

Thứ hai, ngôn ngữ tiếng Việt vốn có cấu trúc phức tạp, với nhiều từ đa nghĩa, đồng âm và cách viết không nhất quán (có dấu, không dấu, sử dụng ký tự thay thế). Trên môi trường mạng, người dùng còn thường xuyên viết tắt, dùng tiếng lóng hoặc kết hợp ký hiệu cảm xúc, khiến việc phân tích ngữ nghĩa trở nên kém chính xác. Đặc biệt, các đối tượng phát tán tin độc hại thường lợi dụng chính sự phức tạp này để ngụy trang nội dung, gây khó khăn cho hệ thống trong việc nhận diện.

Một khó khăn khác đến từ sự kết hợp đa phương thức. Nếu trước đây tin độc hại chủ yếu xuất hiện dưới dạng văn bản, thì hiện nay hình ảnh, biểu đồ, ảnh chế hay video ngắn thường xuyên được sử dụng để tăng mức độ thuyết phục. Việc tin độc hại thường kết hợp đồng thời giữa văn bản và hình ảnh khiến cho các phương pháp xử lý đơn phương thức trở nên kém hiệu quả, đòi hỏi hệ thống phải có khả năng phân tích song song nhiều dạng dữ liệu khác nhau.

Ngoài ra, tính biến động nhanh chóng của nội dung độc hại cũng đặt ra yêu cầu cao. Các chủ đề, hình thức thể hiện và kỹ thuật phát tán thường xuyên thay đổi nhằm tránh sự kiểm soát của các công cụ phát hiện. Điều này khiến cho những mô hình không được cập nhật liên tục sẽ nhanh chóng lạc hậu, giảm hiệu quả ứng dụng trong thực tế.

Từ những phân tích trên có thể thấy, phát hiện tin bài độc hại tiếng Việt có nhiều thách thức, vừa do yếu tố khách quan của ngôn ngữ, vừa do sự tinh vi ngày càng tăng của các kỹ thuật tạo và phát tán tin bài. Điều này đòi hỏi các nghiên cứu phải đi theo hướng tiếp cận mới, tận dụng sức mạnh của học sâu, học bán giám sát và các mô hình đa phương thức để đạt được hiệu quả cao hơn trong thực tiễn.

1.2. CÁC PHƯƠNG PHÁP PHÁT HIỆN TIN BÀI ĐỘC HẠI

1.2.1. Giới thiệu

Phát hiện tin bài độc hại trên không gian mạng là một bài toán đã được tiếp cận theo nhiều hướng khác nhau, gắn liền với sự phát triển của các kỹ thuật xử lý thông tin và trí tuệ nhân tạo. Về tổng thể, các phương pháp phát hiện tin bài độc hại có thể được phân thành một số nhóm: phương pháp dựa trên danh sách và từ khóa, phương pháp dựa trên luật, và phương pháp dựa trên học máy.

Các phương pháp dựa trên danh sách đen hoặc từ khóa hoạt động bằng cách so khớp nội dung tin bài với các tập từ, cụm từ hoặc nguồn được xác định trước là độc hại. Nhóm phương pháp này có ưu điểm là đơn giản, dễ triển khai và chi phí tính toán thấp. Tuy nhiên, do phụ thuộc vào các mẫu cố định, nên chúng khó xử lý các biến thể ngôn ngữ và cách diễn đạt của tiếng Việt, do vậy hiệu quả phát hiện còn hạn chế trong môi trường thông tin đa dạng, biến động.

Các phương pháp dựa trên luật sử dụng tập hợp các quy tắc được xây dựng thủ công dựa trên tri thức chuyên gia, nhằm mô tả các đặc trưng ngôn ngữ, cấu trúc hoặc ngữ cảnh của tin bài độc hại. Mặc dù có khả năng giải thích rõ ràng, nhóm phương pháp này phụ thuộc lớn vào kinh nghiệm chủ quan, khó mở rộng và thiếu khả năng thích ứng khi xuất hiện dạng nội dung mới.

Trong bối cảnh hạn chế của các phương pháp dựa trên danh sách và luật, các phương pháp học máy được nghiên cứu và áp dụng nhằm tự động khai thác đặc

trưng từ dữ liệu tin bài. Khác với các cách tiếp cận thủ công, phát hiện dựa trên học máy cho phép mô hình hiểu được mối quan hệ phức tạp giữa nội dung và nhãn độc hại dựa trên dữ liệu huấn luyện. Theo cách tiếp cận này, các phương pháp phát hiện tin bài độc hại có thể được phân loại thành hai nhóm chính: học có giám sát, sử dụng dữ liệu đã gán nhãn, và học bán giám sát, kết hợp dữ liệu có nhãn và chưa gán nhãn để cải thiện hiệu quả phát hiện.

Học có giám sát đạt hiệu quả cao khi có sẵn tập dữ liệu gán nhãn đầy đủ và đáng tin cậy. Tuy nhiên, trong thực tế, việc xây dựng các bộ dữ liệu lớn cho tin bài độc hại, đặc biệt đối với tiếng Việt, gặp nhiều khó khăn do chi phí gán nhãn cao và yêu cầu sâu về chuyên môn. Điều này dẫn đến sự quan tâm ngày càng lớn đối với học bán giám sát, do học bán giám sát có khả năng khai thác hiệu quả nguồn dữ liệu chưa gán nhãn phong phú trên không gian mạng để hỗ trợ quá trình huấn luyện.

Tóm lại, các phương pháp phát hiện tin bài độc hại đã chuyển dịch từ những cách tiếp cận dựa trên từ khóa và luật sang các phương pháp học máy có khả năng thích ứng tốt hơn với tính phức tạp của dữ liệu thực tế.

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

Trong lĩnh vực xử lý ngôn ngữ tự nhiên và phân tích dữ liệu, nhiều nghiên cứu đã khai thác phương pháp học có giám sát để giải quyết các bài toán phân loại khác nhau, từ nhận diện tin giả, phân tích quan điểm cho đến kiểm soát lan truyền thông tin trên mạng xã hội. Mặc dù hướng tiếp cận và mục tiêu cụ thể của từng công trình có sự khác biệt, điểm chung nổi bật là việc sử dụng dữ liệu đã gán nhãn làm cơ sở để huấn luyện mô hình. Các kết quả này vừa khẳng định vai trò trung tâm của học có giám sát trong nhiều tác vụ liên quan đến thông tin trên không gian mạng, vừa tạo nền tảng quan trọng để tham khảo và kế thừa trong nghiên cứu về phát hiện và phân loại tin bài độc hại tiếng Việt.

1.2.2.1. Các nghiên cứu trong nước

Nguyễn Hữu Phát và cộng sự (2021) [28] đã đề xuất một giải pháp ứng dụng thực tiễn để giải quyết vấn đề bình luận nhạy cảm, phản cảm bằng cách xây dựng một hệ thống chatbot tự động. Nghiên cứu nhấn mạnh những thách thức đặc thù của tiếng Việt, nơi người dùng thường sử dụng các hình thức "lách luật" tinh vi như từ

đồng nghĩa, nói lái, viết tắt để biểu đạt nội dung tiêu cực. Để đối phó, nhóm tác giả đã thiết kế một kiến trúc chatbot gồm ba khối chức năng: (1) Khối tiền xử lý để chuẩn hóa và làm sạch văn bản, (2) Khối xác định mức độ phản cảm để phân loại nội dung, và (3) Khối trả lời để đưa ra phản hồi tự động. Hệ thống đã được huấn luyện và kiểm thử trên một bộ dữ liệu quy mô lớn do nhóm tự thu thập, gồm 100.000 bình luận từ Facebook, và đạt được độ chính xác 75%. Nghiên cứu còn đặt ra mục tiêu phát triển chatbot thành một thư viện có thể tích hợp dễ dàng vào nhiều nền tảng khác nhau.

Nguyễn Thị Thanh Thủy và cộng sự (2023) [29] đã tập trung vào bài toán phát hiện phát ngôn tiêu cực trong các văn bản tiếng Việt phi chính thức trên mạng xã hội. Nhận thấy rằng các lỗi thường gặp như sai chính tả, viết tắt, và dùng từ lóng làm giảm hiệu quả của các mô hình ngôn ngữ, một đóng góp quan trọng của nghiên cứu là việc tích hợp một mô-đun sửa lỗi chính tả vào giai đoạn tiền xử lý. Về mặt kiến trúc, nhóm tác giả đã đề xuất một phương pháp lai (hybrid approach) mạnh mẽ, kết hợp nhiều loại đặc trưng ngữ nghĩa: đặc trưng có ngữ cảnh sâu từ BERT, đặc trưng hiệu quả với từ hiếm từ FastText, và đặc trưng cấp độ ký tự từ charCNN. Các véc-tơ đặc trưng này sau đó được hợp nhất và đưa vào một mạng nơ-ron hồi quy hai chiều (BiLSTM) để thực hiện nhiệm vụ phân loại. Kết quả thực nghiệm trên bộ dữ liệu ViHSD đã khẳng định hiệu quả của cả việc làm sạch dữ liệu đầu vào và chiến lược kết hợp đa đặc trưng.

Khanh Q. Tran và cộng sự (2022) [30] đã đề xuất một hệ thống toàn diện và hiệu năng cao để phát hiện phát ngôn thù địch và phản cảm bằng tiếng Việt. Nghiên cứu giới thiệu một mô hình lai mới, kết hợp sức mạnh của mô hình ngôn ngữ PhoBERT trong việc nắm bắt ngữ cảnh và Mạng nơ-ron tích chập trong việc trích xuất các mẫu cục bộ (PhoBERT-CNN). Để giải quyết vấn đề mất cân bằng dữ liệu thường gặp, nghiên cứu đã áp dụng kỹ thuật tăng cường dữ liệu EDA (Easy Data Augmentation). Thử nghiệm trên hai bộ dữ liệu chuẩn là ViHSD và HSD-VLSP cho thấy mô hình đề xuất đạt hiệu năng vượt trội so với các phương pháp tiên tiến nhất, với điểm F1-score ấn tượng là 67,46% và 98,45%. Ngoài ra, nghiên cứu còn chứng minh tính thực tiễn bằng việc xây dựng một ứng dụng có khả năng xử lý dữ liệu theo luồng (streaming) trong thời gian thực.

Luan Thanh Nguyen [31] và cộng sự (2021) đã mang lại một đóng góp kép quan trọng cho cộng đồng xử lý ngôn ngữ tự nhiên tiếng Việt. Đóng góp thứ nhất là việc xây dựng và công bố bộ dữ liệu UIT-ViCTSD, gồm 10.000 bình luận được gán nhãn thủ công cho hai nhiệm vụ song song là phát hiện phát ngôn mang tính xây dựng và phát ngôn độc hại. Đóng góp thứ hai là việc đề xuất một hệ thống phân loại hiệu quả dựa trên mô hình học chuyển giao PhoBERT. Kết quả thực nghiệm cho thấy mô hình này vượt trội so với các mô hình cơ sở, đạt điểm F1-score là 78,59% cho nhiệm vụ nhận diện phát ngôn xây dựng. Mặc dù điểm F1-score cho nhiệm vụ phát hiện độc hại (59,40%) còn hạn chế do sự mất cân bằng dữ liệu, nghiên cứu đã mở ra một hướng đi quan trọng và cung cấp một tài nguyên giá trị cho các nghiên cứu sau này.

Anh-Tuan Huynh và Phuoc Tran (2025) [32] đã thực hiện một nghiên cứu chuyên sâu về phát hiện tin giả tiếng Việt, với đóng góp chính là xây dựng một bộ dữ liệu mới từ Facebook và các trang tin tức, đồng thời áp dụng các kỹ thuật xử lý mất cân bằng dữ liệu hiệu quả như bổ sung từ bộ dữ liệu VFND và lấy mẫu ngẫu nhiên quá mức. Về phương pháp, nghiên cứu tập trung khai thác các mô hình Transformer, đặc biệt là PhoBERT, và khám phá sức mạnh tổng hợp khi kết hợp PhoBERT với các kỹ thuật nhúng từ truyền thống như TF-IDF và Word2Vec. Kết quả thực nghiệm đã chứng minh rằng mô hình kết hợp PhoBERT + Word2Vec mang lại hiệu suất cao nhất, với độ chính xác 90,4% và điểm F1-score 90,5%, khẳng định rằng việc bổ sung thông tin ngữ nghĩa từ các véc-tơ từ được huấn luyện trước giúp mô hình phân loại tin giả vượt trội hơn.

Khoa Pham Dang và cộng sự (2023) [33] đã giới thiệu một phương pháp lai tiên tiến để phát hiện tin giả tiếng Việt, với một kỹ thuật kỹ nghệ đặc trưng (feature engineering) độc đáo. Cốt lõi của phương pháp này là "diễn đạt thành lời" (verbalization) các siêu dữ liệu như số lượt thích, bình luận, chia sẻ và thông tin thời gian thành các câu văn tự nhiên, sau đó nối chúng vào nội dung gốc của bài đăng. Cách tiếp cận này cho phép các mô hình Transformer, vốn được thiết kế để xử lý văn bản, có thể hiểu và tận dụng được các thông tin phi văn bản này. Khi áp dụng trên bộ dữ liệu ReINTEL, mô hình dựa trên vELECTRA của nhóm đã thiết lập một hiệu năng đỉnh cao mới (state-of-the-art) với điểm AUC là 0.9575, qua đó khẳng

định rằng các chỉ số tương tác xã hội là những yếu tố dự báo mạnh mẽ về độ tin cậy của tin tức.

Thuan Ngoc Le và cộng sự (2024) [34] đã thực hiện một nghiên cứu khẳng định tính khả thi và hiệu quả của việc sử dụng các mô hình học máy cổ điển trong bài toán phát hiện tin giả tiếng Việt. Sử dụng bộ dữ liệu từ cuộc thi ReINTEL, nghiên cứu tập trung vào cách tiếp cận truyền thống: đầu tiên, các đặc trưng văn bản được trích xuất bằng kỹ thuật TF-IDF, sau đó các đặc trưng này được dùng để huấn luyện một loạt các mô hình học máy phổ biến như Logistic Regression, SVM, Random Forest, XGBoost và Light Gradient Boosting Machine (LGBM). Kết quả thực nghiệm cho thấy mô hình LGBM đạt hiệu năng tốt nhất với độ chính xác 88,21% và điểm F1-score 79,67%. Công trình này là một minh chứng cho thấy các thuật toán học máy cổ điển, khi kết hợp với kỹ thuật xử lý ngôn ngữ cơ bản, vẫn là một lựa chọn hiệu quả và tiết kiệm tài nguyên tính toán.

Bài toán phát hiện nội dung độc hại và tin giả tiếng Việt đã nhận được sự quan tâm đáng kể với nhiều hướng tiếp cận khác nhau, từ các phương pháp học máy truyền thống đến các mô hình học sâu và học chuyển giao hiện đại. Trong đó, các nghiên cứu sử dụng các mô hình ngôn ngữ tiền huấn luyện như PhoBERT [26], [27], [28] là những hướng tiếp cận gần nhất với luận án, khi tận dụng khả năng biểu diễn ngữ nghĩa mạnh mẽ của các mô hình Transformer để nâng cao hiệu quả phân loại.

Tuy nhiên, các nghiên cứu hiện tại vẫn tồn tại một số hạn chế. Thứ nhất, phần lớn các phương pháp chủ yếu tập trung vào dữ liệu văn bản hoặc kết hợp thêm các đặc trưng như siêu dữ liệu, nhưng chưa khai thác thông tin hình ảnh. Thứ hai, các phương pháp hiện nay chủ yếu dựa trên học có giám sát, do đó phụ thuộc lớn vào dữ liệu gán nhãn, trong khi việc thu thập và gán nhãn dữ liệu tiếng Việt thường tốn kém và khó mở rộng. Thứ ba, các nghiên cứu chưa tập trung rõ ràng vào bài toán phát hiện tin bài độc hại đa lớp với các nhóm nội dung có tính chất phức tạp, đặc biệt là các nội dung nhạy cảm về chính trị.

1.2.2.2. Các nghiên cứu quốc tế

Z. Khanam và cộng sự (2021) [35] đã thực hiện một công trình mang tính ứng dụng, tập trung vào việc triển khai các mô hình học máy truyền thống cho bài toán

phát hiện tin giả. Cách tiếp cận của họ chủ yếu dựa vào trích xuất đặc trưng, trong đó các đặc trưng từ vựng được trích xuất bằng TF-IDF. Điểm nổi bật của phương pháp này là việc bổ sung các đặc trưng ngôn ngữ định lượng được rút ra từ việc phân tích thành phần câu, chẳng hạn như số lượng danh từ, động từ, tính từ. Các thuật toán cổ điển như Naïve Bayes, SVM, Random Forest, và XGBoost đã được thử nghiệm trên bộ dữ liệu LIAR. Kết quả cho thấy mô hình XGBoost đạt hiệu năng cao nhất với độ chính xác khoảng 75%, chứng minh sự hiệu quả của việc kết hợp các đặc trưng ngôn ngữ được chế tác thủ công với các mô hình học máy cổ điển.

Abdulhameed Al Obaid và cộng sự (2022) [36] đã đề xuất một phương pháp học sâu đa phương thức kết hợp, có tên là DEFD (Diverse Ensemble of Fake-news Detectors), nhằm giải quyết những hạn chế của các mô hình kết hợp truyền thống như số lượng tham số lớn và nguy cơ quá vừa (overfitting). Điểm đột phá của nghiên cứu nằm ở kiến trúc độc đáo, trong đó nhiều 'learner' được xây dựng trên một bộ trích xuất đặc trưng sâu dùng chung và chỉ khác biệt ở các mô-đun chú ý riêng lẻ. Để đảm bảo sự đa dạng, nhóm tác giả đã giới thiệu một hàm mất mát lai mới, không chỉ tối ưu hóa độ chính xác mà còn 'ép buộc' mỗi learner phải tập trung vào các phần khác nhau của nội dung. Về kỹ thuật, mô hình sử dụng XLNet để mã hóa văn bản, VGG19 cho xử lý hình ảnh, và tích hợp Focal Loss để xử lý mất cân bằng dữ liệu. Kết quả thử nghiệm trên các tập dữ liệu PolitiFact và Gossip cho thấy DEFD vượt trội đáng kể so với các phương pháp tiên tiến khác.

Arkadipta De và cộng sự (2021) [37] đã giải quyết thách thức phát hiện phát ngôn thù địch trong tiếng Hindi, một ngôn ngữ thiếu tài nguyên bằng cách định hình bài toán theo một cấu trúc phân loại hai cấp: thô (hateful/non-hateful) và tinh (fake, hate, offensive, etc.). Nền tảng của phương pháp là việc tận dụng các embedding đa ngôn ngữ được huấn luyện trước từ kiến trúc Transformer như mBERT và XLM-Roberta. Đóng góp quan trọng nhất của nghiên cứu nằm ở chiến lược phân loại tinh, trong đó cách tiếp cận One-vs-Rest, xây dựng các bộ phân loại nhị phân riêng biệt cho từng loại hình thù địch, đã chứng tỏ hiệu quả vượt trội so với mô hình phân loại đa nhãn trực tiếp. Cách tiếp cận OvR này đã giúp mô hình đạt được hiệu năng đỉnh cao, với điểm F1-score lên tới 92.60% cho việc phát hiện thù địch ban đầu.

Nishant Rai và cộng sự (2022) [38] đã đề xuất một phương pháp học sâu lai để giải quyết bài toán phát hiện tin giả, tập trung vào việc khai thác các đặc trưng ngôn ngữ từ tiêu đề của các bài báo. Kiến trúc này kết hợp sức mạnh của mô hình BERT để tạo ra các biểu diễn từ được ngữ cảnh hóa sâu sắc, và Mạng Nơ-ron Hồi quy LSTM được xếp chồng lên trên để nắm bắt các phụ thuộc tuần tự và các mẫu dài hạn. Ý tưởng cốt lõi là tận dụng khả năng hiểu ngữ nghĩa của BERT để cung cấp một đầu vào chất lượng cao cho LSTM, từ đó nâng cao khả năng của mô hình trong việc nhận diện các tín hiệu tinh vi của tin giả. Kết quả thực nghiệm trên hai bộ dữ liệu PolitiFact và GossipCop cho thấy mô hình kết hợp BERT+LSTM đã cải thiện độ chính xác lần lượt là 2,50% và 1,10% so với việc chỉ sử dụng một mô hình BERT đơn lẻ.

Pawan Kumar Verma và cộng sự (2021) [39] đã giới thiệu MCred, một phương pháp lai tiên tiến được thiết kế để tận dụng đồng thời cả ngữ nghĩa cục bộ và toàn cục của văn bản. Kiến trúc của MCred bao gồm hai luồng xử lý song song: một luồng sử dụng mô hình BERT để nắm bắt các mối quan hệ ngữ nghĩa toàn cục, luồng còn lại sử dụng Mạng Nơ-ron Tích chập với nhiều kích thước kernel khác nhau để trích xuất các đặc trưng n-gram cục bộ. Các biểu diễn đặc trưng từ hai luồng này sau đó được hợp nhất và đưa vào một mạng nơ-ron dày đặc để đưa ra dự đoán cuối cùng. Kết quả thực nghiệm trên nhiều bộ dữ liệu cho thấy MCred không chỉ vượt trội so với các mô hình học máy truyền thống mà còn cải thiện độ chính xác so với mô hình tiên tiến nhất tại thời điểm đó (FakeBERT).

Subhranil Das và cộng sự (2025) [40] đã đề xuất một mô hình lai kết hợp Mạng Nơ-ron Tích chập và Mạng Nơ-ron Hồi quy để giải quyết bài toán phát hiện tin giả. Phương pháp này được thiết kế để tận dụng các thế mạnh bổ trợ của cả hai kiến trúc: CNN được sử dụng để trích xuất các mẫu văn bản cục bộ và các đặc trưng n-gram, trong khi RNN có nhiệm vụ mô hình hóa các phụ thuộc tuần tự và ngữ cảnh dài hạn trong văn bản. Dữ liệu được tiền xử lý và sử dụng các embedding từ được huấn luyện trước GloVe để khởi tạo. Kết quả thực nghiệm trên bộ dữ liệu SherLockFakenewsNet cho thấy mô hình lai đạt được hiệu năng ấn tượng với độ chính xác kiểm thử là 91,80% và điểm F1-score là 0.91, chứng minh rằng việc kết hợp hai kiến trúc này tạo ra một mô hình mạnh mẽ và hiệu quả.

Aqeel Sahi và cộng sự (2025) [41] đã giải quyết những hạn chế của các mô hình hiện tại trong việc nắm bắt các đặc điểm động và cấu trúc của mạng lưới lan truyền tin giả. Họ đã giới thiệu một kiến trúc lai mới có tên là SGDM-GRU, kết hợp giữa học sâu đồ thị quang phổ và Mạng GRU. Trong phương pháp này, mạng lưới lan truyền tin tức được biểu diễn dưới dạng đồ thị, sau đó một mô hình học sâu đồ thị quang phổ được sử dụng để học các biểu diễn đặc trưng, nắm bắt các thông tin quan hệ và cấu trúc ẩn. Các đặc trưng dựa trên đồ thị này sau đó được đưa vào một mạng GRU để mô hình hóa các động lực học theo thời gian. Kết quả thực nghiệm trên bốn bộ dữ liệu chuẩn cho thấy mô hình SGDM-GRU đạt được hiệu suất vượt trội, với độ chính xác lên tới 98%.

Ayetiran và Özgöbek (2024) [42] đã đề xuất một phương pháp đột phá nhằm giải quyết "khoảng trống ngữ nghĩa" và "khoảng trống không đồng nhất" trong nội dung đa phương thức. Đóng góp cốt lõi là khái niệm "phương thức hợp nhất", trong đó các phương thức phi văn bản được chuyển đổi thành dạng văn bản thông qua chú thích hình ảnh và nhận dạng ký tự quang học (OCR). Trên nền tảng này, một kiến trúc học sâu kết hợp BiLSTM và CNN được xây dựng, với điểm sáng tạo là cơ chế "chú ý liên phương thức" cho phép mô hình học sự tương quan giữa các biểu diễn văn bản của các phương thức khác nhau. Khi áp dụng trên ba bộ dữ liệu chuẩn, phương pháp đề xuất đã cho thấy hiệu năng vượt trội so với các công trình tiên tiến khác.

Deepali Goyal Dev và cộng sự (2024) [43] đã giới thiệu LSTMCNN, một phương pháp học sâu lai được thiết kế để phát hiện tin giả, kết hợp các thế mạnh bổ trợ của CNN và LSTM. Trong kiến trúc này, CNN có nhiệm vụ trích xuất các đặc trưng cục bộ và các mẫu n-gram quan trọng từ văn bản, trong khi LSTM được sử dụng để nắm bắt các phụ thuộc tuần tự và ngữ cảnh dài hạn. Phương pháp này sử dụng các embedding từ được huấn luyện trước GloVe để chuyển đổi văn bản thành các biểu diễn véc-tơ. Kết quả thực nghiệm cho thấy mô hình LSTMCNN đạt được hiệu suất vượt trội với độ chính xác lên tới 98%, vượt qua đáng kể so với các mô hình học máy truyền thống và các kiến trúc học sâu đơn lẻ khác được thử nghiệm.

Visweswaran và các cộng sự [44] đề xuất một phương pháp phát hiện tin giả đa phương thức dựa trên cơ chế kết hợp chú ý giữa ba nguồn dữ liệu: văn bản, hình

ảnh và siêu dữ liệu. Thay vì chỉ đơn thuần nối đặc trưng, nhóm tác giả triển khai một mạng attention đồng bộ, cho phép mô hình tự động học ra mối liên kết quan trọng giữa từng phần nội dung văn bản và vùng đặc trưng trong ảnh. Hệ thống được huấn luyện trên các tập dữ liệu lớn như Twitter và Weibo, cho kết quả vượt trội so với các phương pháp baseline và có khả năng học được sự mâu thuẫn giữa văn bản và hình ảnh.

Guo và Song (2022) [45] đã đề xuất một phương pháp phát hiện tin giả đa phương thức mới, giải quyết hạn chế của các mô hình hiện tại vốn phụ thuộc vào các đặc trưng hình ảnh ở tầng cuối. Nghiên cứu giới thiệu một kiến trúc dựa trên "khối" (block-based), khai thác "kiến thức trung gian" từ cả hai phương thức. Trong kiến trúc này, luồng thời gian được nắm bắt bởi RNN xử lý văn bản, trong khi luồng không gian được lấy từ cả các tầng trung gian và tầng cuối của CNN xử lý hình ảnh. Một cơ chế "khối chú ý" sau đó được sử dụng để hợp nhất hai luồng thông tin này, cho phép mô hình học được mối tương quan chi tiết giữa các phần của văn bản với các cấp độ đặc trưng hình ảnh khác nhau. Kết quả trên bộ dữ liệu Weibo cho thấy phương pháp đề xuất vượt trội đáng kể so với các mô hình nền tảng.

Junxiao Xue và cộng sự (2021) [46] đã giới thiệu MCNN (Multimodal Consistency Neural Network), một phương pháp phát hiện tin giả đa phương thức tiên tiến được thiết kế để nắm bắt sự bất nhất giữa nội dung văn bản-hình ảnh và các dấu hiệu giả mạo vật lý. Kiến trúc này sử dụng một mô-đun chuyên biệt dựa trên thuật toán Phân tích mức độ lỗi để nhận diện các dấu hiệu chỉnh sửa ảnh, và một mô-đun đo lường sự tương đồng để định lượng mức độ nhất quán giữa các biểu diễn ngữ nghĩa của văn bản và hình ảnh. Toàn bộ kiến trúc được huấn luyện end-to-end bằng một hàm mất mát lai, tối ưu hóa đồng thời cả độ chính xác phân loại và sự tương đồng liên phương thức. Kết quả thực nghiệm trên bốn bộ dữ liệu đã chứng minh hiệu quả vượt trội của phương pháp MCNN.

Yang Wu và cộng sự (2021) [47] đã giải quyết hạn chế trong việc mô hình hóa các mối quan hệ tương quan phức tạp giữa các phương thức, vốn thường chỉ được kết hợp thông qua phép nối véc-tơ đơn giản. Nghiên cứu đã giới thiệu MCAN (Multimodal Co-Attention Networks), một kiến trúc học sâu end-to-end được thiết kế để nắm bắt các phụ thuộc liên phương thức. Phương pháp này trích xuất đặc

trưng văn bản từ BERT, đặc trưng hình ảnh trong miền không gian từ VGG19, và một đóng góp độc đáo là đặc trưng hình ảnh trong miền tần số để phát hiện giả mạo. Điểm sáng tạo cốt lõi của MCAN nằm ở cơ chế hợp nhất dựa trên việc xếp chồng nhiều tầng "đồng chú ý", cho phép mô hình hóa tường minh các tương tác giữa các phương thức.

Long Ying và cộng sự (2021) [48] đã giới thiệu MMCN (Multi-level Multi-modal Cross-attention Network), một kiến trúc tiên tiến để giải quyết các hạn chế trong việc hợp nhất thông tin đa phương thức và khai thác ngữ nghĩa văn bản đa cấp. Đóng góp đột phá đầu tiên của MMCN nằm ở mạng chú ý chéo đa phương thức, hợp nhất các biểu diễn từ cấp độ chi tiết bằng cách đưa các véc-tơ từ của BERT và các véc-tơ vùng ảnh của ResNet vào một khối Transformer chung. Đóng góp quan trọng thứ hai là mạng mã hóa đa cấp, khai thác đồng thời cả các đặc trưng cú pháp cấp thấp và ngữ nghĩa cấp cao từ các tầng khác nhau của BERT. Bằng cách kết hợp các biểu diễn đa cấp và được làm giàu chéo này, MMCN đã đạt được hiệu suất vượt trội trên hai bộ dữ liệu công khai.

Yangming Zhou và cộng sự (2022) [49] đã giới thiệu FND-CLIP, một phương pháp phát hiện tin giả đa phương thức mới, tận dụng mô hình được huấn luyện trước CLIP không chỉ để trích xuất đặc trưng mà còn như một cơ chế định hướng nhằm đo lường tường minh sự nhất quán ngữ nghĩa giữa văn bản và hình ảnh. Điểm sáng tạo cốt lõi của phương pháp này là việc sử dụng "điểm tương đồng liên phương thức" tính toán từ CLIP để điều chỉnh một cách linh hoạt mức độ đóng góp của các đặc trưng hợp nhất, trong khi một cơ chế chú ý theo phương thức sẽ cân bằng các thông tin từ đơn phương thức và đa phương thức. Đạt được hiệu năng đỉnh cao trên ba bộ dữ liệu chuẩn, nghiên cứu đã chứng minh hiệu quả của việc sử dụng các mô hình đa phương thức lớn để định hướng tường minh cho quá trình hợp nhất.

Yangming Zhou và cộng sự (2023) [50] đã giới thiệu MMFN (Multi-grained Multi-modal Fusion Network), một phương pháp được thiết kế để bắt chước quy trình đánh giá của con người bằng cách tích hợp thông tin ở nhiều mức độ chi tiết và giải quyết "vấn đề bất đối xứng". Kiến trúc này xử lý các đặc trưng ở hai cấp độ: một cấp độ chi tiết, nơi các đặc trưng văn bản và hình ảnh được hợp nhất qua cơ chế chú ý chéo và một cấp độ tổng thể, nơi mô hình CLIP cung cấp sự hiểu biết ngữ

nghĩa toàn cục. Quan trọng hơn, mô hình sử dụng điểm tương đồng từ CLIP để điều chỉnh linh hoạt sự đóng góp của các đặc trưng hợp nhất, cho phép nó giảm bớt sự phụ thuộc vào thông tin đa phương thức khi chúng không nhất quán.

Longzheng Wang và cộng sự (2023) [51] đã giới thiệu COOLANT, một phương pháp học tương phản liên phương thức được thiết kế tinh vi để giải quyết bài toán phát hiện tin giả. Thay vì sử dụng các nhãn tương phản "cứng", phương pháp này làm mềm mục tiêu học thông qua một nhiệm vụ nhất quán phụ trợ, cho phép căn chỉnh các đặc trưng một cách tinh vi hơn. Sau đó, một cơ chế chú ý có định hướng được sử dụng để tổng hợp một cách thông minh các đặc trưng đơn phương thức và đa phương thức dựa trên mức độ bất đối xứng được định lượng giữa chúng. Đạt được hiệu năng đỉnh cao mới, COOLANT đã chứng minh rằng việc điều chỉnh cẩn thận quá trình học tương phản và kết hợp nó với một cơ chế hợp nhất có thể diễn giải là một chiến lược hiệu quả cao.

Jing Jing và cộng sự (2022) [52] đã giải quyết hạn chế của các phương pháp hiện tại về sự phụ thuộc quá mức vào các đặc trưng ngữ nghĩa ở tầng sâu. Nghiên cứu đã giới thiệu MPFN (Multimodal Progressive Fusion Network), một kiến trúc được thiết kế để khai thác và hợp nhất thông tin đa phương thức một cách có hệ thống qua nhiều cấp độ chi tiết. Thay vì thực hiện một lần hợp nhất duy nhất ở tầng cuối, MPFN thực hiện một chiến lược hợp nhất lũy tiến (progressive fusion), trong đó các đặc trưng ở các tầng nông của các phương thức khác nhau được hợp nhất trước, và kết quả này lại tiếp tục được hợp nhất với các đặc trưng ở các tầng sâu hơn. Nghiên cứu này khẳng định rằng việc hợp nhất thông tin một cách lũy tiến và có hệ thống giúp mô hình nắm bắt được một biểu diễn toàn diện hơn.

Hongzhen Lv và cộng sự (2024) [53] đã giới thiệu Multimodal Dynamic Fusion, một phương pháp đột phá giải quyết những thách thức về tính không chắc chắn và trọng số hợp nhất cố định. Thay vì hợp nhất một cách tĩnh tại, kiến trúc này triển khai một chiến lược hợp nhất động. Đầu tiên, một Mô-đun Ước tính Bất định (UEM) mô hình hóa sự không chắc chắn bên trong mỗi phương thức. Sau đó, một Mạng Hợp nhất Động (DFN) sử dụng Lý thuyết bằng chứng Dempster-Shafer để tính toán các điểm tin cậy cho từng phương thức và quyết định một cách linh hoạt xem nên tin tưởng vào phương thức nào. Nghiên cứu đã chứng minh rằng việc mô

hình hóa tường minh sự không chắc chắn và áp dụng một cơ chế ra quyết định hợp nhất động là một chiến lược hiệu quả cao.

Cao và cộng sự [54] là một trong những nhóm nghiên cứu đầu tiên nhấn mạnh vai trò của nội dung trực quan trong phát hiện tin giả. Nghiên cứu cho thấy các bài viết có kèm hình ảnh/video có khả năng lan truyền nhanh hơn và tác động sâu hơn tới người đọc, đặc biệt là khi nội dung được thiết kế để gây hiệu ứng cảm xúc. Đây là cơ sở quan trọng cho thấy việc phát hiện tin bài độc hại không thể chỉ dừng ở văn bản, mà cần xử lý cả phương tiện trực quan.

Uppada và cộng sự [55] đề xuất mô hình phát hiện tin giả đa phương thức trên tập Fakeddit, một trong những tập dữ liệu lớn nhất hiện nay. Họ kết hợp các mô hình ngôn ngữ như BERT, RoBERTa với các mô hình thị giác như ResNet50, VGG19 và Xception. Trong đó, kết hợp BERT và Xception cho F1-score cao nhất (93%). Tuy nhiên, nghiên cứu này chỉ triển khai trên dữ liệu tiếng Anh và không xử lý tốt nội dung có tính chính trị phức tạp.

Armin cộng sự [56] tiếp cận bài toán bằng mô hình kết hợp dữ liệu từ Văn bản + Ảnh, bình luận và siêu dữ liệu. Họ áp dụng nhiều kỹ thuật trộn đặc trưng như concatenate, sum và maximum pooling để tối ưu hóa đầu ra. Mô hình đạt độ chính xác lên tới 88%, cho thấy tiềm năng của các hệ thống xử lý thông tin đa nguồn, tuy nhiên yêu cầu dữ liệu đồng bộ và không đánh giá trên tiếng Việt là một hạn chế.

Nguyen và Gokhale [57] tập trung phát hiện quan điểm chống chính phủ trên Twitter. Dữ liệu được trích đặc trưng bằng n-grams, TF-IDF và đưa vào các mô hình học máy như SVM, RF, Logistic Regression và cả DistilBERT. Kết quả cho Accuracy ~85%, F1 ~82%, tuy nhiên mô hình này không xử lý được ảnh và phụ thuộc vào ngữ cảnh xã hội cụ thể, khó khái quát cho tiếng Việt.

Từ các nghiên cứu quốc tế có thể nhận thấy rằng, các hướng tiếp cận hiện đại trong phát hiện tin giả và nội dung độc hại đã có sự chuyển dịch rõ rệt từ các phương pháp học máy truyền thống sang các mô hình học sâu, đặc biệt là các mô hình đa phương thức. Trong đó, các nghiên cứu như Uppada và cộng sự [51] và Zhou và cộng sự [45] là những hướng tiếp cận gần nhất với luận án khi cùng khai

thác đồng thời đặc trưng văn bản và hình ảnh, cũng như sử dụng các mô hình học sâu tiên tiến để nâng cao hiệu quả phát hiện.

Tuy nhiên, các nghiên cứu này vẫn tồn tại một số hạn chế. Thứ nhất, các mô hình đa phương thức chủ yếu được xây dựng và đánh giá trên dữ liệu tiếng Anh với quy mô lớn, do đó chưa phản ánh đầy đủ các đặc thù của tiếng Việt. Thứ hai, phần lớn các phương pháp đều dựa trên học có giám sát, phụ thuộc nhiều vào dữ liệu gán nhãn, trong khi bài toán thực tế thường thiếu hụt dữ liệu có nhãn.

Từ những phân tích trên, có thể thấy rằng việc xây dựng một mô hình phát hiện tin bài độc hại tiếng Việt theo hướng kết hợp đa phương thức và học bán giám sát, là một hướng nghiên cứu cần thiết. Đây cũng chính là định hướng mà luận án tập trung giải quyết.

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

Học bán giám sát (Semi-Supervised Learning SSL) nổi lên như một hướng tiếp cận đầy hứa hẹn để giải quyết một trong những thách thức lớn nhất của học có giám sát: sự phụ thuộc vào một lượng lớn dữ liệu đã được gán nhãn thủ công. Việc gán nhãn cho tin tức là một quá trình tốn kém, mất thời gian và đòi hỏi chuyên môn cao. Học bán giám sát tìm cách giảm bớt gánh nặng này bằng cách tận dụng một lượng lớn dữ liệu không nhãn có sẵn trên không gian mạng, kết hợp với một lượng nhỏ dữ liệu có nhãn. Các kỹ thuật phổ biến trong SSL bao gồm tự gán nhãn giả, các phương pháp dựa trên đồ thị (graph-based methods), và học nhất quán.

1.2.3.1. Các nghiên cứu trong nước

Nguyễn Minh Nhật và Trần Huỳnh Minh Tân (2024) [58] đã giới thiệu một phương pháp bán giám sát ứng dụng học sâu, nhằm giải quyết thách thức cốt lõi là tự động phân loại tin tức chứng khoán tiếng Việt và giảm thiểu chi phí gán nhãn thủ công. Tên gọi của phương pháp ngụ ý mạnh mẽ về một cách tiếp cận bán giám sát, được thiết kế để tự động tạo ra một bộ dữ liệu chất lượng cao cho việc huấn luyện mô hình mà không cần sự can thiệp của con người. Về mặt kiến trúc, phương pháp này đề xuất một quy trình tự động gán nhãn dựa trên kiến thức chuyên ngành, sử dụng các từ khóa chuyên biệt như mã chứng khoán, tên công ty niêm yết và thuật ngữ tài chính để phân loại sơ bộ các bài báo. Dựa trên bộ dữ liệu được tạo tự động

này, nghiên cứu đã tiến hành huấn luyện một bộ phân loại học sâu sử dụng kiến trúc BERT, cụ thể là mô hình viBERT cho tiếng Việt. Kết quả thực nghiệm trên bộ dữ liệu hơn 40.000 bài báo cho thấy mô hình đạt độ chính xác trên 96% khi được kiểm chứng với dữ liệu do chuyên gia gán nhãn, chứng tỏ hiệu quả vượt trội trong việc tự động lọc và xác định tin tức liên quan đến thị trường

Lê Hương Thanh và cộng sự (2013) [59] đã giới thiệu một phương pháp Nhận dạng thực thể có tên dựa trên thuật toán Lan truyền nhãn (Label Propagation), nhằm giải quyết thách thức cốt lõi là giảm thiểu sự phụ thuộc vào dữ liệu gán nhãn thủ công quy mô lớn. Tên gọi của phương pháp ngụ ý mạnh mẽ về một cách tiếp cận bán giám sát, được thiết kế để hoạt động hiệu quả ngay cả khi kho dữ liệu gán nhãn có kích thước hạn chế. Về mặt kỹ thuật, phương pháp này đề xuất một độ đo tương đồng từ vựng mới dựa trên ngữ cảnh và cây ngữ nghĩa tiếng Việt để xây dựng đồ thị lan truyền. Song song đó, nghiên cứu đã thực hiện một cải tiến quan trọng đối với thuật toán gốc bằng cách đưa ra một cơ chế điều chỉnh trọng số nhằm giảm thiểu ảnh hưởng của các nhãn có tần suất cao. Kết quả thực nghiệm trên bộ dữ liệu báo chí tiếng Việt cho thấy phương pháp đề xuất đạt độ chính xác vượt trội hơn thuật toán lan truyền nhãn gốc, và hiệu quả hơn mô hình Trường ngẫu nhiên có điều kiện (CRF) có giám sát khi lượng dữ liệu gán nhãn là rất nhỏ.

Phạm Quang H. và cộng sự (2015) [60] đề xuất một phương pháp bán giám sát kết hợp kỹ thuật bootstrapping với mô hình CRF được huấn luyện online. Phương pháp này đặc biệt hữu ích khi dữ liệu gán nhãn cho bài toán Nhận dạng thực thể có tên (NER) trong tiếng Việt khan hiếm trong khi nguồn dữ liệu chưa có nhãn lại rất dồi dào. Mô hình xây dựng một bộ phân loại ban đầu từ một lượng nhỏ dữ liệu gán nhãn, sau đó sử dụng thuật toán học online MIRA để lặp lại quy trình gán nhãn cho các câu chưa được chú thích và tự động thêm các mẫu có độ tin cậy cao nhất vào tập huấn luyện. Dù mô hình đạt F1-score đến 89.16% trên tiếng Việt, nghiên cứu cũng chỉ ra rằng hiệu quả có thể bị suy giảm do sự lan truyền lỗi, khi một dự đoán sai trong một vòng lặp có thể ảnh hưởng tiêu cực đến các mô hình ở các vòng lặp sau.

Lê Thanh Hương và cộng sự (2014) [61] đề xuất một phương pháp bán giám sát sử dụng Lan truyền nhãn để giải quyết bài toán trích rút quan hệ trong tiếng

Việt. Phương pháp này đặc biệt hữu ích trong bối cảnh dữ liệu gán nhãn khan hiếm trong khi kho văn bản chưa gán nhãn lại rất dồi dào. Mô hình xây dựng một đồ thị kết nối các câu chứa cặp thực thể và sử dụng thuật toán lan truyền để gán nhãn cho các mẫu chưa được chú thích. Để cải thiện hiệu quả, nghiên cứu giới thiệu một phương pháp đo lường độ tương đồng câu mới và một cơ chế để giảm thiểu ảnh hưởng của các nhãn có tần suất cao. Thử nghiệm cho thấy phương pháp này không chỉ cho kết quả tốt hơn đáng kể so với thuật toán lan truyền nhãn gốc mà còn vượt trội hơn phương pháp học có giám sát SVM khi lượng dữ liệu huấn luyện bị hạn chế, cho thấy tiềm năng lớn trong việc xử lý các ngôn ngữ ít tài nguyên.

Từ các nghiên cứu trong nước có thể nhận thấy rằng, các hướng tiếp cận bán giám sát đã được quan tâm và áp dụng trong nhiều bài toán xử lý ngôn ngữ tiếng Việt như phân loại văn bản, nhận dạng thực thể có tên và trích rút quan hệ. Trong đó, nghiên cứu của Nguyễn Minh Nhật và Trần Huỳnh Minh Tân [54] là hướng tiếp cận gần nhất với luận án khi kết hợp học sâu với cơ chế tự động gán nhãn nhằm giảm chi phí xây dựng dữ liệu huấn luyện.

Tuy nhiên, các nghiên cứu hiện tại vẫn tồn tại một số hạn chế. Thứ nhất, phần lớn các phương pháp chủ yếu tập trung vào dữ liệu văn bản, chưa xem xét hoặc khai thác thông tin từ các phương thức khác như hình ảnh, trong khi đây là đặc trưng phổ biến của tin bài trên mạng xã hội. Thứ hai, các phương pháp bán giám sát chủ yếu dựa trên các kỹ thuật truyền thống như lan truyền nhãn hoặc bootstrapping, chưa tận dụng đầy đủ khả năng của các mô hình học sâu hiện đại trong việc học biểu diễn dữ liệu. Thứ ba, các nghiên cứu trong nước chưa tập trung trực tiếp vào bài toán phát hiện tin bài độc hại, đặc biệt là các nội dung nhạy cảm về chính trị, cũng như chưa giải quyết bài toán phân loại đa lớp trong bối cảnh dữ liệu đa phương thức.

Từ những phân tích trên, có thể thấy rằng việc xây dựng một mô hình phát hiện tin bài độc hại tiếng Việt theo hướng kết hợp đa phương thức và học bán giám sát, đồng thời khai thác hiệu quả các mô hình học sâu hiện đại, là một hướng nghiên cứu cần thiết. Đây cũng chính là định hướng mà luận án tập trung giải quyết nhằm khắc phục những hạn chế của các nghiên cứu hiện có.

1.2.3.2. Các nghiên cứu quốc tế

Singh và cộng sự (2022) [62] đã giới thiệu SEMI-FND, một phương pháp đa phương thức mới dựa trên kiến trúc ensemble xếp chồng nhằm giải quyết đồng thời hai thách thức cốt lõi là nâng cao độ chính xác và tối ưu hóa tốc độ xử lý. Tên gọi SEMI-FND ngụ ý mạnh mẽ về một cách tiếp cận bán giám sát, được thiết kế để hoạt động hiệu quả ngay cả khi dữ liệu gán nhãn bị hạn chế. Về mặt kiến trúc, đối với phương thức văn bản, phương pháp này áp dụng cơ chế xếp chồng kết hợp hai mô hình ngôn ngữ là BERT và ELECTRA. Đối với hình ảnh, nghiên cứu đã thực hiện một phân tích so sánh sâu rộng để xác định NasNet Mobile là lựa chọn tối ưu. Kết quả thực nghiệm trên hai bộ dữ liệu Twitter MediaEval và Weibo cho thấy SEMI-FND đạt độ chính xác vượt trội, đồng thời giảm đáng kể số lượng tham số huấn luyện.

Wang và cộng sự (2021) [63] đã giải quyết thách thức cốt lõi của việc phát hiện tin giả trên các sự kiện mới nổi, nơi mà dữ liệu gán nhãn cực kỳ khan hiếm. Để giải quyết bài toán học ít mẫu, nghiên cứu đã giới thiệu MetaFEND, một phương pháp đột phá kết hợp giữa học siêu dữ liệu và các tiến trình nơ ron. Điểm sáng tạo của MetaFEND nằm ở việc thiết kế một "nhiệm vụ học mô phỏng" cho phép mô hình học được một bộ tham số khởi tạo tối ưu có khả năng thích ứng nhanh. Sau đó, mô hình sử dụng hiệu quả một vài mẫu tin đã được xác minh làm ngữ cảnh để đưa ra dự đoán. Mô hình còn được tăng cường bởi một cơ chế "chú ý cứng" (hard attention) dựa trên Gumbel-Softmax, cho phép nó chọn lọc và tập trung vào mẫu tin tức có liên quan nhất trong tập hỗ trợ.

Lakzaei và cộng sự [64] đề xuất LOSS-GAT, một mô hình kết hợp one-class learning với Mạng chú ý đồ thị (GAT) để truyền lan pseudo-label. Phương pháp này đặc biệt hữu ích khi dữ liệu của một lớp (ví dụ: tin thật) dồi dào trong khi dữ liệu của lớp kia (tin giả) thì khan hiếm. Mô hình xây dựng một đồ thị kết nối các mẫu dữ liệu và sau đó sử dụng kiến trúc GAT để lan truyền thông tin nhãn từ một số ít các mẫu đã được gán nhãn sang một lượng lớn các mẫu chưa có nhãn. Dù mô hình đạt độ chính xác 83.6% trên tiếng Anh, nghiên cứu cũng chỉ ra rằng hiệu quả bị suy giảm khi áp dụng sang ngôn ngữ khác, cho thấy thách thức khi chuyển đổi mô hình giữa các ngôn ngữ và dữ liệu không đồng nhất.

Meel và cộng sự [65] sử dụng Mạng tích chập đồ thị (GCN) kết hợp với khoảng cách Word Mover's trong một khuôn khổ học bán giám sát. Trong phương pháp này, các bài báo được xem như các nút trong một đồ thị, và các cạnh được tạo ra giữa các bài báo tương tự nhau về mặt ngữ nghĩa, đo bằng khoảng cách Word Mover's. Sau đó, mô hình GCN được áp dụng trên cấu trúc đồ thị này để học các biểu diễn đặc trưng và lan truyền thông tin nhãn từ các nút đã được gán nhãn sang các nút chưa được gán nhãn. Mô hình cho kết quả tốt trong tiếng Anh nhưng không đề cập đến xử lý ảnh, do đó thiếu khả năng tương thích với dữ liệu đa phương thức.

Li và cộng sự [66] đề xuất một mô hình phát hiện bán giám sát sử dụng word2vec và Bi-LSTM, cùng một lớp đánh giá độ tin cậy cho các nhãn giả. Đây là một dạng của phương pháp tự huấn luyện, trong đó mô hình được huấn luyện ban đầu trên một lượng nhỏ dữ liệu có nhãn, sau đó được dùng để dự đoán nhãn cho dữ liệu không nhãn. Lớp đánh giá độ tin cậy đóng vai trò then chốt, chỉ chọn những dự đoán có độ chắc chắn cao nhất để đưa các cặp (dữ liệu, nhãn giả) này vào lại tập huấn luyện. Mô hình đạt độ chính xác tới 90%, nhưng nghiên cứu cũng lưu ý rằng nó vẫn phụ thuộc đến 80% vào dữ liệu gán nhãn ban đầu, gây khó khăn nếu áp dụng vào bối cảnh dữ liệu có nhãn cực kỳ khan hiếm.

Từ các nghiên cứu quốc tế có thể nhận thấy rằng, các phương pháp phát hiện tin giả/tin độc hại đã đạt được nhiều kết quả đáng chú ý theo các hướng tiếp cận khác nhau như học đa phương thức, học bán giám sát và học siêu dữ liệu. Trong đó, các mô hình như SEMI-FND [58] và MetaFEND [59] có sự tương đồng nhất với hướng nghiên cứu của luận án khi cùng khai thác dữ liệu đa phương thức và giải quyết bài toán trong điều kiện dữ liệu gán nhãn hạn chế.

Tuy nhiên, vẫn tồn tại một số hạn chế đáng chú ý. Thứ nhất, các phương pháp đa phương thức như SEMI-FND chủ yếu tập trung vào việc cải thiện hiệu năng thông qua kiến trúc ensemble, nhưng chưa làm rõ vai trò đóng góp của từng nguồn dữ liệu (văn bản và hình ảnh) trong quá trình ra quyết định. Thứ hai, các phương pháp học bán giám sát và học ít mẫu như MetaFEND, LOSS-GAT [60] hay các phương pháp dựa trên GCN [61] chủ yếu khai thác cấu trúc đồ thị hoặc dữ liệu văn bản, chưa tích hợp hiệu quả thông tin hình ảnh trong bài toán phát hiện tin bài đa phương thức. Thứ ba, một số phương pháp tự huấn luyện như [62] vẫn phụ thuộc

đáng kể vào dữ liệu gán nhãn ban đầu, làm hạn chế khả năng mở rộng trong bối cảnh dữ liệu thực tế.

Bên cạnh đó, phần lớn các nghiên cứu hiện tại được xây dựng và đánh giá trên dữ liệu tiếng Anh hoặc tiếng Trung, trong khi bài toán phát hiện tin bài độc hại tiếng Việt vẫn chưa được khai thác đầy đủ, đặc biệt trong bối cảnh dữ liệu đa phương thức và hạn chế về nhãn.

Từ những phân tích trên, có thể thấy rằng việc xây dựng một mô hình phát hiện tin bài độc hại tiếng Việt theo hướng kết hợp đa phương thức và học bán giám sát, đồng thời làm rõ vai trò của từng nguồn dữ liệu trong quá trình phân loại, là một hướng nghiên cứu cần thiết và có ý nghĩa. Đây cũng chính là định hướng chính mà luận án tập trung giải quyết.

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

Trong những năm gần đây, nghiên cứu về phát hiện tin bài độc hại đã đạt được một số kết quả đáng ghi nhận. Các mô hình học sâu, đặc biệt là các biến thể dựa trên Transformer, đã chứng minh khả năng nắm bắt ngữ nghĩa và bối cảnh ngôn ngữ tốt hơn nhiều so với các phương pháp truyền thống. Một số công trình quốc tế còn mở rộng theo hướng đa phương thức, kết hợp văn bản với hình ảnh để khai thác thông tin toàn diện hơn. Đối với dữ liệu đa phương thức, các nghiên cứu [45], [46], [49], [51], [52] và [58] cho thấy việc kết hợp thông tin văn bản và hình ảnh có khả năng khai thác các tín hiệu bổ sung mà phương pháp đơn phương thức khó biểu diễn đầy đủ. Đặc biệt, FND-CLIP [45] và MMFN [46] chú trọng mô hình hóa quan hệ và mức độ nhất quán giữa hai phương thức, trong khi MDF [49] hướng tới hợp nhất động dựa trên độ tin cậy của từng phương thức. Các kết quả này cho thấy xử lý đồng thời văn bản và hình ảnh là một hướng tiếp cận có cơ sở cho bài toán phát hiện thông tin sai lệch và nội dung độc hại.

Tuy nhiên, khi đặt trong mối tương quan so sánh giữa trong nước và quốc tế, có thể nhận thấy một số hạn chế nổi bật. Trước hết, các nghiên cứu tại Việt Nam vẫn chủ yếu dựa vào phương pháp đơn phương thức, tập trung gần như hoàn toàn vào phân tích nội dung văn bản của tin bài. Trong khi đó, thực tế cho thấy phần lớn tin bài độc hại thường kết hợp giữa văn bản và hình ảnh, tận dụng cả ngôn từ lẫn

yếu tố trực quan để gia tăng mức độ thuyết phục. Việc thiếu vắng cơ chế xử lý đa phương thức đồng nghĩa với việc chưa khai thác hết các tín hiệu quan trọng trong dữ liệu, khiến khả năng phát hiện còn hạn chế so với tình huống thực tế.

Một điểm yếu khác là sự phụ thuộc lớn vào học có giám sát. Cả trong nước lẫn quốc tế, đa số các hệ thống hiện tại đều yêu cầu bộ dữ liệu gán nhãn đầy đủ để huấn luyện. Điều này kéo theo chi phí tốn kém, khó mở rộng và thiếu tính thích ứng trước các sự kiện hoặc loại hình tin độc hại mới xuất hiện. Với tiếng Việt, vấn đề này còn nghiêm trọng hơn do thiếu hụt các kho dữ liệu chuẩn hóa quy mô lớn. Trong bối cảnh đó, các phương pháp học bán giám sát có khả năng tận dụng dữ liệu chưa gán nhãn để cải thiện mô hình nhưng đến nay vẫn chưa được nghiên cứu một cách hệ thống. Đây là khoảng trống cần được tập trung khai thác.

Ngoài ra, nhiều công trình hiện nay có xu hướng đơn giản hóa bài toán phân loại tin bài thành phân loại nhị phân, chẳng hạn chia thành "tin bài thường/ tin bài độc hại". Cách tiếp cận này tuy dễ triển khai và đánh giá, nhưng lại chưa phản ánh đúng sự phức tạp của hiện tượng. Trong thực tế, một bài viết có thể đồng thời chứa nhiều yếu tố độc hại, vừa xuyên tạc, vừa kích động, hoặc vừa bôi nhọ cá nhân lại vừa gieo rắc hoang mang. Việc giới hạn ở nhị phân vô hình trung làm giảm khả năng phân tích chi tiết và chưa đáp ứng được yêu cầu phân loại đa lớp trong bối cảnh ngày càng phức tạp.

Từ những hạn chế nêu trên có thể thấy, khoảng trống nghiên cứu hiện nay chủ yếu nằm ở việc thiếu một mô hình tích hợp phù hợp cho tiếng Việt, vừa khai thác được các đặc trưng đa phương thức, vừa thích ứng với điều kiện hạn chế dữ liệu gán nhãn. Mô hình đó cần đồng thời giải quyết được ba vấn đề then chốt: khả năng xử lý đa phương thức (kết hợp văn bản và hình ảnh), khả năng tận dụng dữ liệu chưa gán nhãn thông qua học bán giám sát, và năng lực phân loại đa lớp thay vì chỉ dừng ở nhị phân. Trên thế giới, từng hướng tiếp cận này đã có nghiên cứu chuyên sâu, song việc kết hợp cả ba thành phần trong một khung thống nhất, lại ứng dụng cho tiếng Việt với những đặc thù riêng về ngôn ngữ, vẫn là một thách thức lớn. Đây cũng chính là hướng đi mà luận án lựa chọn nhằm đóng góp cả về mặt lý luận và thực tiễn. Cụ thể, luận án này sẽ trả lời hai câu hỏi nghiên cứu chính sau:

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

Trong lĩnh vực xử lý ngôn ngữ tự nhiên, sự xuất hiện của các mô hình ngôn ngữ lớn dựa trên kiến trúc Transformer đã tạo ra bước tiến đột phá. Các mô hình này không chỉ thay thế những phương pháp truyền thống dựa trên n-gram hoặc RNN, mà còn mở ra khả năng biểu diễn ngữ nghĩa toàn diện hơn, phục vụ hiệu quả cho nhiều tác vụ khác nhau, trong đó có phát hiện tin bài độc hại. Trong luận án này, ba mô hình được xem xét và sử dụng bao gồm BERT [67], RoBERTa [25] và PhoBERT [11].

1.3.1.1. BERT

BERT [67] (Bidirectional Encoder Representations from Transformers) là một mô hình ngôn ngữ sâu được phát triển bởi Google, nổi bật nhờ khả năng học biểu diễn ngữ nghĩa theo chiều hai chiều (bidirectional) thông qua kiến trúc Transformer.

Mô hình này được huấn luyện với hai nhiệm vụ chính: Masked Language Modeling (MLM) và Next Sentence Prediction (NSP), giúp BERT hiểu rõ hơn về ngữ cảnh của từ và mối quan hệ giữa các câu trong văn bản.

Ưu điểm của BERT là có khả năng biểu diễn ngữ nghĩa giàu thông tin và nắm bắt ngữ cảnh toàn diện, đặc biệt phù hợp cho các tác vụ phân tích ngôn ngữ tự nhiên như phân loại văn bản, trích chọn thông tin hoặc phát hiện các biểu hiện ngữ nghĩa phức tạp. Ngoài ra, kiến trúc Transformer còn giúp mô hình xử lý tốt các văn bản dài và duy trì tính linh hoạt trong nhiều ứng dụng.

Tuy nhiên, nhược điểm của BERT là mô hình gốc được huấn luyện chủ yếu trên dữ liệu tiếng Anh, do đó khi áp dụng cho các ngôn ngữ khác, bao gồm tiếng Việt, có thể gặp sai lệch về ngữ nghĩa nếu không được tinh chỉnh (fine-tuning) hoặc

xây dựng phiên bản tiền huấn luyện phù hợp. Bên cạnh đó, kích thước lớn và yêu cầu tài nguyên tính toán cao cũng là thách thức đáng kể khi triển khai trong thực tế.

1.3.1.2. RoBERTa

RoBERTa [25] (Robustly Optimized BERT Pretraining Approach) là một phiên bản cải tiến của BERT [67] do Facebook AI phát triển. Mô hình này loại bỏ nhiệm vụ Next Sentence Prediction, thay đổi kỹ thuật masking theo hướng động, đồng thời được huấn luyện trên tập dữ liệu lớn hơn với batch size lớn hơn, nhờ đó cải thiện đáng kể hiệu suất so với BERT. RoBERTa có thể được sử dụng như một bộ mã hóa ngôn ngữ mạnh mẽ cho hệ thống phát hiện nội dung độc hại, đặc biệt hiệu quả khi kết hợp với các kiến trúc phân loại đa lớp.

Ưu điểm của RoBERTa là đạt hiệu quả cao hơn BERT trong hầu hết các tác vụ xử lý ngôn ngữ tự nhiên nhờ tối ưu hóa quá trình tiền huấn luyện. Mô hình này cũng mạnh mẽ và ổn định hơn khi huấn luyện trên tập dữ liệu lớn, đồng thời dễ thích nghi cho các tác vụ đặc thù như phân loại văn bản độc hại. Tuy nhiên, nhược điểm của RoBERTa là tương tự như BERT khi mô hình gốc được huấn luyện chủ yếu trên tiếng Anh, do đó cần được tinh chỉnh hoặc sử dụng phiên bản tiền huấn luyện cho tiếng Việt. Ngoài ra, dung lượng lớn và yêu cầu tài nguyên tính toán cao cũng là thách thức đáng kể khi triển khai trong môi trường có hạn chế về phần cứng.

1.3.1.3. PhoBERT

PhoBERT [12] là mô hình ngôn ngữ tiền huấn luyện dành riêng cho tiếng Việt, được xây dựng dựa trên kiến trúc RoBERTa [68] và huấn luyện trên kho ngữ liệu tiếng Việt quy mô lớn. Nhờ kế thừa sức mạnh của RoBERTa và được điều chỉnh phù hợp với đặc thù ngôn ngữ, PhoBERT đã đạt hiệu quả vượt trội trong nhiều tác vụ xử lý ngôn ngữ tự nhiên tiếng Việt như phân loại văn bản, phân tích cảm xúc hay nhận dạng thực thể. Với đặc điểm hình thái và ngữ pháp phức tạp của tiếng Việt, PhoBERT giúp cải thiện đáng kể khả năng biểu diễn ngữ nghĩa và ngữ cảnh.

Trong phạm vi nghiên cứu này, PhoBERT được sử dụng để xử lý dữ liệu văn bản trong bài toán phát hiện tin độc hại. Mặc dù luận án tập trung vào phát hiện tin độc hại tiếng Việt nói chung, tập dữ liệu huấn luyện và đánh giá mô hình chủ yếu

phản ánh nhóm tin độc hại về chính trị, vốn là loại hình phổ biến và có tác động xã hội nghiêm trọng. Do đó, việc ứng dụng PhoBERT trong bối cảnh này không chỉ chứng minh khả năng nắm bắt ngữ nghĩa phức tạp của tiếng Việt mà còn khẳng định hiệu quả khi triển khai cho những trường hợp thực tiễn có độ nhạy cảm cao.

Ưu điểm của PhoBERT là mang lại hiệu suất vượt trội trên dữ liệu tiếng Việt, không cần chuyển ngữ hay điều chỉnh đặc biệt như BERT hoặc RoBERTa gốc. Mô hình đặc biệt phù hợp cho các bài toán phân loại văn bản tiếng Việt, bao gồm cả kịch bản đa lớp khi được kết hợp với kiến trúc học sâu thích hợp. Nhược điểm của PhoBERT là vẫn yêu cầu tài nguyên tính toán lớn và chưa phù hợp với các ứng dụng thời gian thực nếu không được rút gọn. Ngoài ra, PhoBERT chưa mô hình hóa rõ mối quan hệ giữa các nhãn, vì vậy hiệu quả có thể được cải thiện khi kết hợp với các kỹ thuật attention giữa nhãn hoặc kiến trúc phân nhánh.

1.3.2. Các mô hình xử lý ảnh

Trong bối cảnh tin bài độc hại không chỉ tồn tại dưới dạng văn bản mà còn gắn liền với hình ảnh, việc lựa chọn các mô hình học sâu để trích xuất đặc trưng từ dữ liệu ảnh là một yêu cầu quan trọng. Các mô hình xử lý ảnh dựa trên mạng nơ-ron tích chập và Transformer đã chứng minh hiệu quả vượt trội trong việc nhận diện, phân loại và biểu diễn ngữ nghĩa trực quan. Trong phạm vi luận án này, ba mô hình được xem xét bao gồm Xception, VGG19 và Swin Transformer V2.

1.3.2.1. Xception

Xception [69] (Extreme Inception) là một kiến trúc mạng nơ-ron tích chập sâu được phát triển như một mở rộng của dòng Inception, với trọng tâm là việc sử dụng tích chập tách biệt theo chiều sâu. Cơ chế này giúp giảm thiểu số lượng tham số và tăng hiệu quả huấn luyện mà vẫn giữ lại khả năng học các đặc trưng thị giác cần thiết. Nhờ đó, Xception đã cho thấy hiệu quả trong một số bài toán phân loại hình ảnh có cấu trúc phức tạp hoặc chứa yếu tố biểu tượng.

Trong bài toán phát hiện nội dung độc hại dưới dạng ảnh, Xception thường được sử dụng như một baseline mạnh, đặc biệt trong các nghiên cứu so sánh mô hình xử lý ảnh. Với khả năng học tốt các đặc trưng cục bộ, Xception có thể nhận diện các biểu hiện hình ảnh có tính trực quan cao như meme, ảnh ghép, hoặc ảnh

chứa biểu tượng chính trị rõ ràng. Tuy nhiên, khi đối mặt với các hình ảnh có bố cục ngụy trang tinh vi, biểu tượng nhỏ, hoặc mức độ trừu tượng cao, thường gặp trong các chiến dịch truyền thông độc hại hiện đại, thì khả năng khai thác ngữ nghĩa liên vùng của Xception vẫn còn hạn chế.

So với các kiến trúc CNN truyền thống hoặc Xception, mô hình Swin Transformer V2, vốn sử dụng cơ chế attention trong cửa sổ trượt, thể hiện ưu thế vượt trội về khả năng học ngữ nghĩa không gian toàn cục, đồng thời mở rộng linh hoạt cho các kiến trúc phân cấp. Điều này cho phép Swin V2 xử lý tốt hơn các loại ảnh độc hại có yếu tố ẩn dụ, ngôn ngữ hình ảnh phân tán hoặc bố cục phức hợp, vốn là điểm yếu của hầu hết các kiến trúc dựa trên tích chập.

Tóm lại, Xception là một lựa chọn hiệu quả trong các hệ thống yêu cầu xử lý ảnh nhanh và tương đối đơn giản về mặt cấu trúc, nhưng vẫn chưa đủ linh hoạt khi cần phát hiện các hình ảnh độc hại có tính ngữ nghĩa phức tạp. Do đó, trong các hệ thống phát hiện nội dung độc hại quy mô lớn hoặc đa dạng phương thức, Xception có thể được dùng làm đối sánh hoặc thành phần bổ trợ, trong khi các mô hình tiên tiến hơn như Swin Transformer V2 nên được ưu tiên làm kiến trúc chính.

1.3.2.2. VGG19

VGG19 [70] là một trong những kiến trúc mạng nơ-ron tích chập nền tảng, được phát triển bởi nhóm Visual Geometry Group tại Đại học Oxford. Mô hình này gồm 19 lớp trọng số, trong đó có 16 lớp tích chập và 3 lớp kết nối đầy đủ. Tất cả các lớp tích chập sử dụng bộ lọc kích thước 3×3, kết hợp với các lớp gộp cực đại (max pooling) kích thước 2×2 để giảm dần kích thước không gian của đặc trưng. Cấu trúc đồng nhất và tuần tự này giúp VGG19 dễ triển khai, dễ huấn luyện và dễ áp dụng cho nhiều bài toán thị giác máy tính khác nhau.

Trong nghiên cứu này, VGG19 được sử dụng dưới dạng mô hình học sâu có sẵn (pretrained model), đã được huấn luyện trước trên tập dữ liệu ImageNet. Thay vì huấn luyện lại từ đầu, mô hình được dùng chủ yếu cho mục đích trích xuất đặc trưng hình ảnh, một bước quan trọng trong hệ thống phân loại đa phương thức.

Cụ thể, ảnh đầu vào sẽ được đưa qua VGG19 để sinh ra vector đặc trưng dạng nhúng (embedding vector), sau đó kết hợp với đặc trưng văn bản (trích xuất bởi

PhoBERT) để phục vụ phân loại nội dung độc hại theo hướng đa lớp. Việc sử dụng VGG19 theo cách này có nhiều lợi thế trong thực nghiệm ban đầu: đơn giản trong triển khai, không yêu cầu tài nguyên huấn luyện lớn, và có khả năng khai thác tốt các đặc trưng trực quan như đường nét, màu sắc, bố cục vốn thường xuất hiện trong các loại hình ảnh độc hại như: biểu ngữ chống đối, ảnh ghép xuyên tạc, hoặc các hình ảnh có yếu tố thị giác nhạy cảm. VGG19 hoạt động ổn định với ảnh có bố cục rõ ràng và nội dung dễ phân vùng, giúp hệ thống phát hiện nhanh các tín hiệu thị giác nổi bật mà không cần xử lý ngữ nghĩa sâu.

Tuy nhiên, VGG19 vẫn tồn tại một số điểm hạn chế. Do không tích hợp các cơ chế attention hay transformer, mô hình khó học được mối quan hệ ngữ nghĩa toàn cục giữa các vùng ảnh. Điều này khiến VGG19 kém hiệu quả hơn khi gặp các ảnh độc hại có cấu trúc bố cục phức tạp, biểu tượng lồng ghép ngụy trang hoặc các chi tiết phân tán. Ngoài ra, số lượng tham số lớn khiến mô hình ngốn bộ nhớ và không phù hợp cho các hệ thống yêu cầu thời gian thực hoặc có tài nguyên hạn chế.

Tổng thể, VGG19 được sử dụng trong nghiên cứu như một mô hình nền truyền thống để so sánh và đối chiếu. Mặc dù không phải là kiến trúc hiện đại nhất, nhưng VGG19 vẫn cung cấp một baseline hữu ích và đáng tin cậy trong việc đánh giá hiệu quả của các mô hình tiên tiến hơn, đặc biệt là Swin Transformer V2, vốn được lựa chọn làm thành phần chính trong hệ thống đề xuất.

1.3.2.3. Swin Transformer V2

Swin Transformer V2 [71] là một phiên bản nâng cấp của kiến trúc Swin Transformer, mô hình thị giác đầu tiên ứng dụng cơ chế attention theo cửa sổ trượt trong mạng thị giác dựa trên Transformer. Được giới thiệu bởi Microsoft Research vào năm 2022, Swin Transformer V2 giải quyết nhiều hạn chế của mô hình ban đầu, đặc biệt về khả năng tổng quát hóa, độ ổn định khi huấn luyện với batch size lớn, và khả năng mở rộng cho các bài toán thị giác phức tạp.

Khác với mạng CNN truyền thống vốn chỉ học được đặc trưng cục bộ thông qua tích chập, Swin Transformer V2 sử dụng multi-head self-attention trong từng cửa sổ nhỏ của ảnh (window-based attention), cho phép mô hình học mối quan hệ không gian trong phạm vi hẹp một cách hiệu quả. Sau đó, nhờ cơ chế "cửa sổ trượt", mô hình tiếp tục học mối tương quan giữa các vùng ảnh khác nhau, từ đó mở

rộng khả năng biểu diễn ngữ nghĩa toàn cục mà không phải đánh đổi hiệu suất tính toán.

Trong bài toán phát hiện nội dung độc hại dưới dạng hình ảnh, Swin Transformer V2 thể hiện ưu thế rõ rệt ở khả năng nhận diện biểu tượng ngụy trang, ảnh ghép châm biếm, hoặc bố cục phi tuyến tính vốn thường xuyên được sử dụng trong các chiến dịch truyền thông độc hại. Thay vì chỉ tập trung vào vùng đối tượng nổi bật, Swin Transformer V2 có thể học được các biểu hiện rải rác khắp ảnh, chẳng hạn như dòng chữ nhỏ gắn trên áo, hình biểu tượng nằm trong góc ảnh, hoặc sự kết hợp giữa nền ảnh và chi tiết ám chỉ chính trị.

Ngoài ra, Swin Transformer V2 có kiến trúc phân cấp tương tự như CNN, cho phép trích xuất đặc trưng ở nhiều độ phân giải khác nhau. Điều này giúp mô hình đặc biệt phù hợp với các ảnh có độ phân giải cao hoặc chứa nhiều chi tiết nhỏ, mà vẫn đảm bảo tốc độ xử lý và hiệu suất bộ nhớ, vốn là yêu cầu quan trọng khi triển khai trong môi trường thực tế.

Không chỉ vượt trội về mặt kỹ thuật, Swin Transformer V2 còn cho phép kết hợp dễ dàng với các mô hình xử lý văn bản như BERT hoặc PhoBERT để xây dựng hệ thống đa phương thức. Trong mô hình đề xuất của luận án, Swin Transformer V2 đóng vai trò là nhánh xử lý hình ảnh, tạo ra embedding ngữ nghĩa ảnh để kết hợp với đặc trưng văn bản từ PhoBERT, qua đó phục vụ phân loại nội dung độc hại đa lớp.

Thực nghiệm trong nghiên cứu cũng cho thấy Swin Transformer V2 đạt kết quả tốt hơn các mô hình CNN truyền thống như VGG19 hay Xception trong việc phát hiện hình ảnh độc hại có tính ẩn dụ hoặc lồng ghép phức tạp. Cụ thể, Swin Transformer V2 có thể phát hiện tốt hơn các trường hợp hình ảnh chứa biểu tượng bị biến thể, khẩu hiệu bị mã hóa, hoặc thiết kế gián tiếp hướng đến các chủ đề chính trị nhạy cảm, điều mà các mô hình dựa hoàn toàn vào tích chập khó nhận diện.

Tóm lại, Swin Transformer V2 là lựa chọn phù hợp cho các hệ thống phát hiện nội dung độc hại hiện đại. Với cơ chế attention hiệu quả, kiến trúc phân cấp, khả năng mở rộng tốt và tính linh hoạt khi tích hợp vào mô hình đa phương thức, Swin

Transformer V2 không chỉ đáp ứng tốt yêu cầu kỹ thuật mà còn thể hiện rõ tiềm năng ứng dụng trong thực tiễn giám sát không gian mạng hiện nay.

1.3.3. Học bán giám sát trong phân loại dữ liệu

Học bán giám sát là phương pháp kết hợp giữa dữ liệu có nhãn và dữ liệu chưa có nhãn nhằm nâng cao hiệu quả huấn luyện mô hình. Trong bối cảnh thực tế, việc xây dựng các tập dữ liệu gán nhãn đầy đủ cho các tác vụ phức tạp như phát hiện tin bài độc hại thường gặp nhiều khó khăn do chi phí, thời gian và yêu cầu chuyên môn cao. Ngược lại, dữ liệu chưa gán nhãn lại rất dồi dào, đặc biệt trên không gian mạng với lượng lớn tin bài được đăng tải hàng ngày. Do đó, học bán giám sát trở thành một hướng tiếp cận tiềm năng, giúp khai thác tối đa nguồn dữ liệu sẵn có.

Các kỹ thuật học bán giám sát thường xoay quanh ba nguyên lý cốt lõi:

- Giả định tính liên tục (continuity assumption): các điểm dữ liệu gần nhau trong không gian đặc trưng thường có cùng nhãn.
- Giả định cụm (cluster assumption): dữ liệu có xu hướng hình thành cụm, và các điểm trong cùng cụm nhiều khả năng chia sẻ cùng một nhãn.
- Giả định đa tạp (manifold assumption): dữ liệu phân bố trên một đa tạp có chiều thấp hơn so với không gian quan sát, do đó có thể tận dụng cấu trúc này để cải thiện phân loại.

Một số hướng tiếp cận cho học bán giám sát tiêu biểu như: (i) gán nhãn giả (pseudo-labeling), theo đó mô hình được huấn luyện ban đầu trên dữ liệu có nhãn, sau đó sử dụng để dự đoán và gán nhãn tạm thời cho dữ liệu chưa gán nhãn, (ii) ràng buộc tính nhất quán (consistency regularization), trong đó buộc mô hình duy trì tính nhất quán của dự đoán khi đầu vào bị biến đổi nhẹ và (iii) các phương pháp dựa trên đồ thị hoặc lan truyền nhãn (label propagation), trong đó khai thác cấu trúc lân cận trong không gian đặc trưng.

Các phương pháp dựa trên ràng buộc tính nhất quán (như Mean Teacher hoặc FixMatch) cho thấy hiệu quả cao trong các bài toán thị giác máy tính với lượng dữ liệu lớn và có thể áp dụng các phép biến đổi mạnh lên dữ liệu. Tuy nhiên, trong bài toán phân loại tin bài đa phương thức tiếng Việt, việc xác định các phép biến đổi phù hợp cho cả văn bản và hình ảnh là không đơn giản, đặc biệt với văn bản tiếng

Việt có nhiều biến thể ngôn ngữ, dễ làm thay đổi ngữ nghĩa. Trong khi đó, các phương pháp lan truyền nhãn dựa trên đồ thị đòi hỏi chi phí tính toán lớn và khó mở rộng khi lượng dữ liệu tăng, đồng thời phụ thuộc nhiều vào chất lượng biểu diễn không gian đặc trưng.

So với các phương pháp trên, pseudo-labeling có ưu điểm nổi bật về tính đơn giản, dễ triển khai và khả năng tận dụng trực tiếp mô hình đã huấn luyện. Phương pháp này không yêu cầu thiết kế các hàm ràng buộc phức tạp hoặc cấu trúc đồ thị bổ sung, do đó phù hợp với điều kiện dữ liệu tiếng Việt còn hạn chế cả về quy mô và độ tin cậy của nhãn. Cách tiếp cận này giúp mở rộng tập huấn luyện, giảm hiện tượng quá khớp khi dữ liệu có nhãn hạn chế, đồng thời nâng cao khả năng tổng quát hóa.

Một lợi thế quan trọng của học bán giám sát là khả năng thích ứng với sự thay đổi liên tục của dữ liệu trên mạng xã hội. Các loại tin bài độc hại thường biến đổi nhanh theo sự kiện và xu hướng xã hội, khiến dữ liệu gán nhãn sẵn dễ nhanh chóng lỗi thời. Bằng việc bổ sung thêm dữ liệu mới chưa gán nhãn, các mô hình bán giám sát có thể duy trì tính cập nhật và hiệu quả phát hiện.

Tuy nhiên, học bán giám sát cũng đối mặt với một số thách thức. Nếu dữ liệu chưa gán nhãn chứa nhiều nhiễu hoặc sai lệch, việc gán nhãn giả có thể dẫn đến hiệu ứng lan truyền sai. Bên cạnh đó, các phương pháp bán giám sát thường đòi hỏi quá trình tinh chỉnh cẩn thận để cân bằng giữa dữ liệu có nhãn và chưa có nhãn, tránh làm suy giảm hiệu quả của mô hình.

Trong khuôn khổ luận án, học bán giám sát đóng vai trò then chốt trong việc xử lý tình huống thiếu hụt dữ liệu gán nhãn tiếng Việt, đồng thời bổ trợ cho các mô hình đa phương thức nhằm nâng cao hiệu quả phát hiện tin bài độc hại trong môi trường thực tế.

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

Trong bối cảnh thông tin trên không gian mạng được trình bày dưới dạng kết hợp giữa văn bản và hình ảnh nhiều hơn, việc chỉ tập trung vào một phương thức đơn lẻ không còn đủ để phản ánh đầy đủ nội dung và ý đồ truyền tải. Nhiều tin bài độc hại không chỉ sử dụng ngôn ngữ mang tính xuyên tạc, kích động mà còn khai

thác hình ảnh minh họa, biểu tượng để gia tăng mức độ thuyết phục và tác động tâm lý đối với người đọc. Do đó, các mô hình học máy kết hợp cả văn bản và ảnh trở thành một hướng tiếp cận quan trọng trong phát hiện và phân loại tin bài độc hại.

Các mô hình kết hợp (multimodal models) thường được xây dựng theo ba hướng tiếp cận chính:

- Kết hợp sớm (early fusion): đặc trưng của văn bản và ảnh được trích xuất riêng biệt sau đó ghép nối (concatenate) thành một vector đặc trưng chung. Cách tiếp cận này đơn giản và hiệu quả trong các trường hợp dữ liệu không quá phức tạp, nhưng đôi khi chưa khai thác triệt để được mối quan hệ tương tác giữa các phương thức.
- Kết hợp muộn (late fusion): đặc trưng của từng phương thức được xử lý qua các mô hình riêng biệt, sau đó kết quả dự đoán được tích hợp thông qua các cơ chế như voting, averaging hoặc các tầng kết hợp có trọng số. Phương pháp này giúp tận dụng thế mạnh riêng của từng mô hình, nhưng có thể bỏ qua sự tương tác tinh vi giữa văn bản và hình ảnh.
- Kết hợp lai (hybrid fusion): là sự kết hợp linh hoạt giữa hai hướng tiếp cận trên, cho phép mô hình vừa học đặc trưng chung vừa giữ được thông tin chuyên biệt của từng phương thức.

Phương pháp kết hợp sớm tiến hành trích xuất đặc trưng từ văn bản và hình ảnh thông qua các mô hình học sâu chuyên biệt, sau đó ghép nối các đặc trưng này thành một vector chung trước khi đưa vào bộ phân loại. Ưu điểm của phương pháp này nằm ở tính đơn giản, chi phí tính toán thấp và khả năng huấn luyện hiệu quả ngay cả khi tài nguyên hạn chế. Tuy nhiên, do việc hợp nhất đặc trưng diễn ra ở giai đoạn đầu, mô hình có thể chưa khai thác đầy đủ các tương tác phức tạp giữa nội dung ngôn ngữ và tín hiệu thị giác, đặc biệt trong các trường hợp ngụy trang tinh vi hoặc nội dung đa tầng.

Trái ngược với đó, phương pháp kết hợp muộn xử lý văn bản và hình ảnh bằng các mô hình tách biệt và chỉ thực hiện tích hợp ở giai đoạn cuối thông qua các cơ chế như voting, averaging hoặc các lớp hợp nhất có trọng số. Cách tiếp cận này cho phép tận dụng tối đa sức mạnh chuyên biệt của từng mô hình, góp phần nâng cao độ

ổn định của kết quả phân loại và khả năng mở rộng kiến trúc hệ thống. Tuy nhiên, do tách rời hai phương thức trong quá trình học đặc trưng, phương pháp kết hợp muộn có thể bỏ qua các mối quan hệ phụ thuộc tinh vi giữa văn bản và hình ảnh, dẫn tới hạn chế trong các tình huống yêu cầu hiểu ngữ cảnh sâu.

Phương pháp kết hợp lai được đề xuất nhằm khắc phục các hạn chế của hai hướng tiếp cận trên, thông qua việc kết hợp linh hoạt giữa hợp nhất đặc trưng ở mức sớm và tích hợp kết quả ở mức muộn. Đây là chiến lược có tiềm năng mang lại hiệu quả cao nhất vì vừa duy trì khả năng học đặc trưng chung, vừa bảo toàn thông tin chuyên biệt của từng phương thức. Tuy nhiên, kiến trúc kết hợp lai thường phức tạp và yêu cầu tài nguyên tính toán lớn, dẫn đến chi phí huấn luyện cao và khó triển khai trong các hệ thống thực tế quy mô lớn.

Trong phạm vi bài toán của luận án, dữ liệu đa phương thức tiếng Việt có quy mô hạn chế và không đồng đều giữa các lớp. Trong bối cảnh này, các phương pháp fusion phức tạp dễ dẫn đến hiện tượng quá khớp (overfitting) hoặc học các tương quan giả (spurious correlations), do số lượng tham số lớn trong khi dữ liệu huấn luyện không đủ để ước lượng ổn định. Ngược lại, phương pháp concatenation cho phép bảo toàn thông tin riêng biệt của từng miền (văn bản và hình ảnh) mà không áp đặt giả định mạnh về mối quan hệ giữa chúng.

Ngoài ra, lựa chọn concatenation còn phù hợp với thiết kế tổng thể của hệ thống học bán giám sát trong luận án. Cụ thể, khi áp dụng gán nhãn giả (pseudo-labeling), việc sử dụng một mô hình có cấu trúc đơn giản giúp giảm lan truyền sai số từ nhãn giả và tăng tính ổn định trong quá trình huấn luyện lặp. Trong khi đó, các mô hình fusion phức tạp có xu hướng khuếch đại sai số nếu nhãn giả không đủ chính xác.

Trong khuôn khổ luận án này, hướng tiếp cận kết hợp sớm được lựa chọn nhờ khả năng cân bằng giữa tính hiệu quả và tính khả thi khi xử lý dữ liệu thực tế tiếng Việt. Cụ thể, đặc trưng văn bản được trích xuất bằng mô hình ngôn ngữ PhoBERT, trong khi đặc trưng hình ảnh được mã hóa bằng Swin Transformer V2. Hai nguồn đặc trưng này được hợp nhất và được đưa vào bộ phân loại cho các bài toán phân loại đa lớp. Kết quả thực nghiệm cho thấy phương pháp kết hợp sớm đạt hiệu năng cao hơn các mô hình đơn phương thức được đánh giá trong cùng kịch bản

thực nghiệm, đồng thời cho thấy tính khả thi của việc kết hợp đặc trưng văn bản và hình ảnh.

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:

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

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:

$$F_1 = \frac{2(Precision \times Recall)}{Precision+Recall} \quad (1.4)$$

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

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

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 xây dựng tập dữ liệu là bước khởi đầu và cũng là một trong những khâu quan trọng trong nghiên cứu phát hiện tin bài độc hại. Tuy nhiên, quá trình thu thập và xử lý dữ liệu tiếng Việt cho bài toán này gặp phải nhiều khó khăn đặc thù, ảnh hưởng trực tiếp đến chất lượng của mô hình và độ tin cậy của kết quả.

Trước hết, sự đa dạng và biến đổi không ngừng của tin bài độc hại tạo ra thách thức lớn trong việc thu thập dữ liệu. Tin bài độc hại có thể tồn tại dưới nhiều hình thức: từ các bài viết dạng tin tức, bài đăng trên mạng xã hội, đến hình ảnh minh họa, đồ họa chế, hoặc các đoạn văn bản ngắn gọn kèm theo chú thích. Nội dung của chúng thường xuyên thay đổi theo sự kiện xã hội, chính trị, kinh tế, hoặc theo các chủ đề nóng đang thu hút sự quan tâm của cộng đồng. Do đó, một tập dữ liệu được xây dựng tại một thời điểm cụ thể có thể nhanh chóng trở nên lỗi thời, không còn phản ánh đầy đủ các hình thức và xu hướng mới của tin độc hại.

Tiếp đến, tính ngụy trang và mức độ tinh vi của tin độc hại khiến việc thu thập và gán nhãn trở nên phức tạp. Nhiều tin bài được viết với ngôn từ mơ hồ, hàm ý, sử dụng lối viết ẩn dụ, châm biếm hoặc cắt ghép hình ảnh nhằm tạo ra ấn tượng sai lệch mà khó nhận diện bằng mắt thường. Một số trường hợp, nội dung văn bản nếu tách rời khỏi ngữ cảnh thì hoàn toàn bình thường, nhưng khi kết hợp với hình ảnh hoặc tiêu đề lại mang ý nghĩa xuyên tạc. Chính đặc điểm này làm cho quá trình xác định tin bài độc hại không chỉ phụ thuộc vào phân tích câu chữ mà còn đòi hỏi đánh giá toàn bộ bối cảnh và cách thức trình bày thông tin.

Bên cạnh đó, khó khăn trong gán nhãn dữ liệu là vấn đề nổi bật. Quá trình gán nhãn đòi hỏi người thực hiện phải có sự am hiểu về ngôn ngữ, văn hóa, và thậm chí cả bối cảnh xã hội để nhận diện đúng mức độ và loại hình độc hại. Tuy nhiên, việc gán nhãn thường dựa vào nhận định chủ quan, dễ dẫn đến sự thiếu thống nhất giữa các chuyên gia. Một bài viết có thể được một chuyên gia đánh giá là "vu khống bôi nhọ", trong khi người khác lại xếp vào nhóm "xuyên tạc sự thật" hoặc "thúc đẩy tư tưởng cực đoan". Sự khác biệt trong cách nhìn nhận này dẫn đến tình trạng thiếu thống nhất trong dữ liệu gán nhãn, từ đó làm suy giảm hiệu quả của mô hình học máy, vốn đòi hỏi dữ liệu đầu vào phải có mức độ nhất quán cao.

Ngoài ra, khó khăn về pháp lý và đạo đức trong thu thập dữ liệu cũng là một vấn đề không thể bỏ qua. Tin bài độc hại thường xuất hiện trên các nền tảng mạng xã hội hoặc các trang web không chính thống. Việc thu thập và sử dụng những dữ liệu này cần được cân nhắc kỹ lưỡng để tuân thủ quy định về quyền riêng tư, bản quyền và pháp luật Việt Nam. Đồng thời, trong bối cảnh nghiên cứu khoa học, việc công bố dữ liệu cũng phải đảm bảo không tiếp tay cho sự lan truyền của các nội dung độc hại, tránh gây ảnh hưởng tiêu cực tới cộng đồng.

Một khó khăn khác là thiếu hụt các tập dữ liệu công khai chuẩn hóa cho tiếng Việt. Trong khi tiếng Anh và một số ngôn ngữ phổ biến khác đã có nhiều kho dữ liệu lớn, được cộng đồng nghiên cứu chia sẻ và chuẩn hóa, thì tiếng Việt hiện nay vẫn thiếu vắng các tập dữ liệu quy mô lớn và được gán nhãn đầy đủ cho bài toán phát hiện tin độc hại. Phần lớn các nghiên cứu phải tự xây dựng tập dữ liệu riêng, dẫn đến sự phân mảnh và khó khăn trong việc so sánh, đánh giá khách quan giữa các mô hình.

Cuối cùng, cần nhấn mạnh đến tính bất cân bằng trong dữ liệu. Số lượng tin độc hại chiếm tỷ lệ nhỏ so với khối lượng thông tin bài thường trên mạng, dẫn đến tập dữ liệu thường bị lệch về phía dữ liệu "tin bài thường". Điều này gây khó khăn cho mô hình trong việc học và nhận diện các đặc trưng của tin độc hại, làm tăng nguy cơ bỏ sót hoặc nhận diện sai. Việc xử lý mất cân bằng dữ liệu đòi hỏi áp dụng thêm các kỹ thuật đặc thù như tái mẫu, gán trọng số hoặc sử dụng học bán giám sát để tận dụng dữ liệu chưa nhãn.

Tóm lại, việc thu thập dữ liệu tin bài độc hại tiếng Việt gặp nhiều thách thức từ tính đa dạng, tinh vi của nội dung cho đến khó khăn trong gán nhãn, vấn đề pháp lý đạo đức, sự thiếu hụt kho dữ liệu chuẩn hóa và tình trạng mất cân bằng trong phân bố dữ liệu. Đây là những yếu tố cần được tính đến khi thiết kế mô hình phát hiện, đồng thời cũng là động lực thúc đẩy việc nghiên cứu các phương pháp học bán giám sát và kết hợp đa phương thức để khắc phục hạn chế về dữ liệu.

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

Trong nghiên cứu phát hiện tin bài độc hại tiếng Việt, việc xây dựng một tập dữ liệu có chất lượng đóng vai trò quyết định đối với hiệu quả mô hình. Hiện tại,

chưa tồn tại bộ dữ liệu tiêu chuẩn hóa hay công khai nào phục vụ riêng cho bài toán nhận diện nội dung độc hại mang tính chính trị, do đó luận án tiến hành xây dựng tập dữ liệu mới nhằm phản ánh đúng thực tế và đảm bảo độ bao phủ cần thiết về phạm vi và mức độ đa dạng của nội dung.

Quy trình gán nhãn trong nghiên cứu được xây dựng dựa trên Điều 13 Luật An ninh mạng năm 2025. Nhóm tác giả tiến hành trích chọn, đưa ra bộ tiêu chí nhận diện trên cơ sở đã nêu ở mục 1.1.2, ba nhóm nhãn độc hại chính bao gồm: "Vu khống, bôi nhọ", "Kêu gọi chống đối", và "Thúc đẩy tư tưởng cực đoan". Các tin bài không biểu hiện đặc điểm thuộc các khoản nêu trong Điều 13 được gán nhãn "tin bài thường".

Để đảm bảo tính khách quan và độ tin cậy của dữ liệu, quá trình gán nhãn được tổ chức theo cơ chế đánh giá độc lập. Bộ mẫu dữ liệu được cung cấp cho hai nhóm gán nhãn độc lập.

Sau bước gán nhãn độc lập, mức độ nhất quán giữa hai nhóm gán nhãn được đánh giá thông qua hệ số đồng thuận Krippendorff's Alpha [72]. Chỉ số này cho phép định lượng mức độ thống nhất trong quá trình gán nhãn, qua đó phản ánh độ tin cậy của bộ tiêu chí đã xây dựng. Việc lựa chọn Krippendorff's Alpha thay vì các độ đo khác xuất phát từ một số lý do sau:

- Thứ nhất, Krippendorff's Alpha phù hợp với bài toán phân loại đa lớp, cho phép xử lý linh hoạt mà không bị giới hạn bởi số lượng lớp.
- Thứ hai, Krippendorff's Alpha cho phép xử lý trường hợp dữ liệu thiếu (missing labels) hoặc không đầy đủ, điều này phù hợp với thực tế khi một số mẫu có thể chưa đạt được sự đồng thuận ngay từ đầu.
- Thứ ba, chỉ số này đánh giá mức độ đồng thuận sau khi đã loại bỏ yếu tố trùng hợp ngẫu nhiên, do đó phản ánh chính xác hơn độ tin cậy thực sự của quá trình gán nhãn.

Trên cơ sở lựa chọn chỉ số đánh giá phù hợp, dữ liệu gán nhãn được tổ chức lại để phục vụ cho quá trình tính toán. Cụ thể, từ bảng kết quả gán nhãn, một ma trận đồng xuất hiện (coincidence matrix) được xây dựng nhằm thống kê tần suất xuất hiện của các cặp nhãn trên cùng một mẫu dữ liệu. Trong ma trận này, mỗi phần

tử n_{ij} biểu diễn số lần một mẫu được gán nhãn i bởi nhóm thứ nhất và nhãn j bởi nhóm thứ hai. Ma trận này phản ánh trực tiếp mối quan hệ giữa các nhãn và là cơ sở để tính toán mức độ sai khác.

Hệ số Alpha được xác định thông qua tỷ lệ giữa sai khác quan sát và sai khác kỳ vọng. Hệ số Alpha được sử dụng theo công thức sau:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Trong đó:

- α : Hệ số Krippendorff's Alpha
- D_o: Độ sai khác quan sát (được tính bằng cách tổng hợp tất cả các trường hợp không đồng thuận giữa 2 nhóm và chuẩn hóa theo tổng số quan sát).
- D_e: Độ sai khác kỳ vọng (được tính dựa trên phân bố tổng thể của các nhãn trong toàn bộ tập dữ liệu).

Giá trị hệ số Krippendorff's Alpha thu được trong nghiên cứu đạt mức

$\alpha = 0.84$, vượt ngưỡng 0,8 theo các tiêu chuẩn đánh giá, cho thấy mức độ đồng thuận cao giữa các nhóm và đảm bảo độ tin cậy của dữ liệu gán nhãn.

Mặc dù đạt độ nhất quán cao, đối với các trường hợp vẫn còn tồn tại sự không thống nhất về nhãn, dữ liệu được chuyển sang bước đánh giá bổ sung. Cụ thể, một nhóm thứ ba tiến hành rà soát và gán nhãn lại dựa trên bộ tiêu chí đã thiết lập, nhằm đảm bảo tính chính xác và nhất quán của dữ liệu.

Quy trình gán nhãn nhiều bước không chỉ giúp giảm thiểu sai lệch chủ quan, mà còn nâng cao độ tin cậy và khả năng tái lập của bộ dữ liệu, từ đó góp phần đảm bảo tính chính xác và giá trị khoa học của các kết quả thực nghiệm trong luận án.

Kết quả, hai bộ dữ liệu hoàn chỉnh đã được xây dựng để phục vụ thử nghiệm các mô hình trong luận án:

- Đối với mô hình học có giám sát, bộ dữ liệu gồm 1.2000 tin bài, 6.000 tin bài độc hại được phân đều cho ba nhãn con, và 6.000 tin bài thường.

- Trong bối cảnh học bán giám sát, bộ dữ liệu đa lớp đạt 18.000 tin bài (9.000 tin bài độc hại, 9.000 tin bài thường), bảo đảm quy mô đủ lớn để khai thác hiệu quả các kỹ thuật bán giám sát.

Ví dụ minh họa cho các nhãn độc hại trong tập dữ liệu:

- **Tin độc hại về vu khống bôi nhọ**

  Dạng văn bản: "Đảng Cộng sản tiếp tục kìm hãm sự phát triển của đất nước, khiến Việt Nam lạc hậu so với các nước Hàn Quốc, Nhật Bản, cũng như các nước phương Tây. Đồng thời, chính sách độc tài của đảng tước đoạt quyền cơ bản của người dân, vi phạm nhân quyền và đàn áp tôn giáo."

  (Chú thích: Đây là nội dung văn bản độc hại hại về chính trị vi phạm điều 13 luật an ninh mạng 2025, chỉ phục vụ cho nghiên cứu này không được sử dụng với mục đích khác)

  Dạng hình ảnh:



Hình 1.2. Tin bài độc hại vu khống bôi nhọ dạng hình ảnh

(Chú thích: Đây là nội dung hình ảnh độc hại hại về chính trị vi phạm điều 13 luật an ninh mạng 2025, chỉ phục vụ cho nghiên cứu này không được sử dụng với mục đích khác)

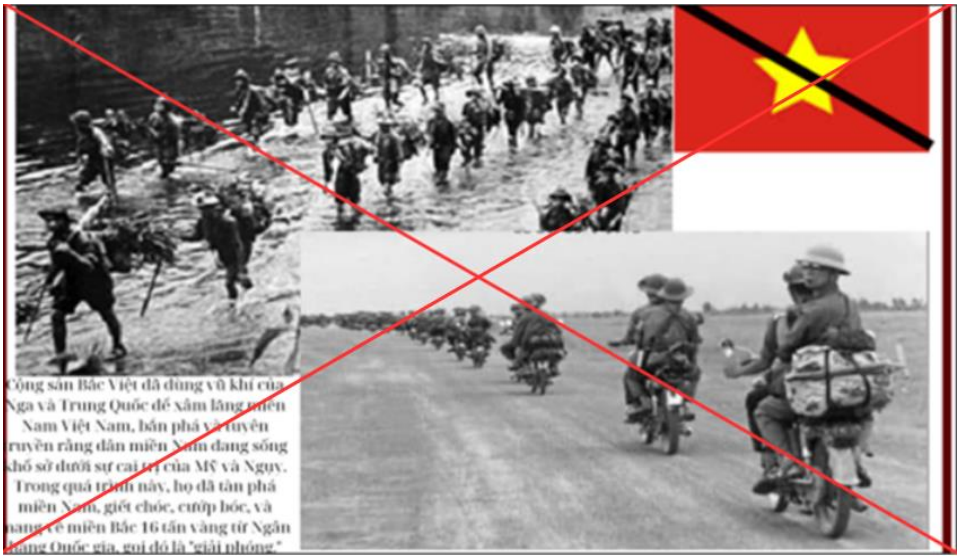
- **Tin độc hại kêu gọi chống đối**

  Dạng văn bản: Người dân miền Nam Việt Nam không cần sự 'giải phóng' giả dối mà chỉ là một chiêu trò để che giấu tham vọng của Việt Cộng. Việt

Nam và chính quyền tay sai quá lộng hành rồi, sai lầm chính là không hành động nên hãy đứng lên đi hỡi nhân dân cả nước 'Xóa sổ chính quyền'.

(Chú thích: Đây là nội dung văn bản độc hại hại về chính trị vi phạm điều 13 luật an ninh mạng 2025, chỉ phục vụ cho nghiên cứu này không được sử dụng với mục đích khác)

Dạng hình ảnh:



Hình 1.3. Tin bài độc hại kêu gọi chống đối dạng hình ảnh

(Chú thích: Đây là nội dung hình ảnh độc hại hại về chính trị vi phạm điều 13 luật an ninh mạng 2025, chỉ phục vụ cho nghiên cứu này không được sử dụng với mục đích khác)

- **Tin độc hại thúc đẩy tư tưởng cực đoan**

  Dạng văn bản: "Trong các chế độ độc tài, đặc biệt là cộng sản, đấu đá quyền lực giữa các phe nhóm thường rất gay gắt nhưng thường được kiểm soát để tránh bất ổn. Tuy nhiên, hiện nay, xung đột trong lãnh đạo Đảng Cộng sản Việt Nam đã vượt quá mức thỏa hiệp và không còn kiểm soát được. Tin đồn về các cuộc triệt hạ nội bộ tràn ngập mạng xã hội, khiến Đảng mất khả năng phản ứng kịp thời."

  (Chú thích: Đây là nội dung văn bản độc hại hại về chính trị vi phạm điều 13 luật an ninh mạng 2025, chỉ phục vụ cho nghiên cứu này không được sử dụng với mục đích khác)

Dạng hình ảnh:



Hình 1.4. Tin bài độc hại thúc đẩy tư tưởng cực đoan dạng hình ảnh

(Chú thích: Đây là nội dung hình ảnh độc hại hại về chính trị vi phạm điều 13 luật an ninh mạng 2025, chỉ phục vụ cho nghiên cứu này không được sử dụng với mục đích khác)

Qua đó, bộ dữ liệu không chỉ phản ánh đặc trưng đa dạng và phức tạp của tin bài độc hại tiếng Việt mà còn đặt nền tảng thực nghiệm quan trọng cho các mô hình phát hiện kết hợp và học bán giám sát, góp phần giải quyết một trong những dạng điển hình và nguy hại nhất của tin độc hại trên không gian mạng hiện nay.

1.5. HƯỚNG NGHIÊN CỨU CỦA LUẬN ÁN

Từ việc phân tích các ưu điểm và hạn chế của các nghiên cứu liên quan trong mục 1.2.4, có thể nhận thấy rằng các phương pháp hiện tại vẫn còn tồn tại một số hạn chế, đặc biệt trong việc khai thác hiệu quả thông tin đa phương thức và xử lý bài toán trong điều kiện dữ liệu gán nhãn hạn chế. Do đó, việc xây dựng một mô hình phát hiện tin bài độc hại tiếng Việt có khả năng kết hợp hiệu quả giữa đặc trưng văn bản và hình ảnh, đồng thời tận dụng được nguồn dữ liệu chưa gán nhãn, vẫn là một vấn đề có ý nghĩa nghiên cứu cả về mặt lý thuyết và thực tiễn. Phần tiếp theo trong mục này nêu rõ các vấn đề sẽ được giải quyết trong luận án nhằm trả lời cho 2 câu hỏi nghiên cứu nêu trong mục 1.2.4.

1.5.1. Các vấn đề giải quyết trong luận án

Hướng nghiên cứu của luận án là phát hiện tin bài độc hại tiếng Việt dựa trên học sâu, kết hợp đa phương thức và học bán giám sát. Trên cơ sở đó, luận án tập trung giải quyết các vấn đề sau:

(1) Xây dựng mô hình phát hiện tin bài độc hại đa phương thức: Khai thác đồng thời đặc trưng văn bản và hình ảnh nhằm nâng cao khả năng nhận diện các nội dung độc hại có biểu hiện phức tạp. Việc kết hợp hai nguồn dữ liệu giúp mô hình không chỉ dựa vào ngôn ngữ mà còn tận dụng thông tin trực quan.

(2) Xây dựng mô hình học bán giám sát cho bài toán phát hiện tin độc hại: Giảm phụ thuộc vào dữ liệu gán nhãn bằng cách tận dụng dữ liệu chưa gán nhãn thông qua kỹ thuật pseudo-labeling, từ đó mở rộng khả năng áp dụng trong thực tế.

(3) Giải quyết bài toán phân loại đa lớp: Cho phép phân loại tin bài thuộc dạng độc hại khác nhau, phản ánh đúng bản chất của dữ liệu thực tế thay vì giới hạn ở phân loại nhị phân.

Từ các các vấn đề cần giải quyết nêu trên, luận án xác định hai bài toán chính cần giải quyết:

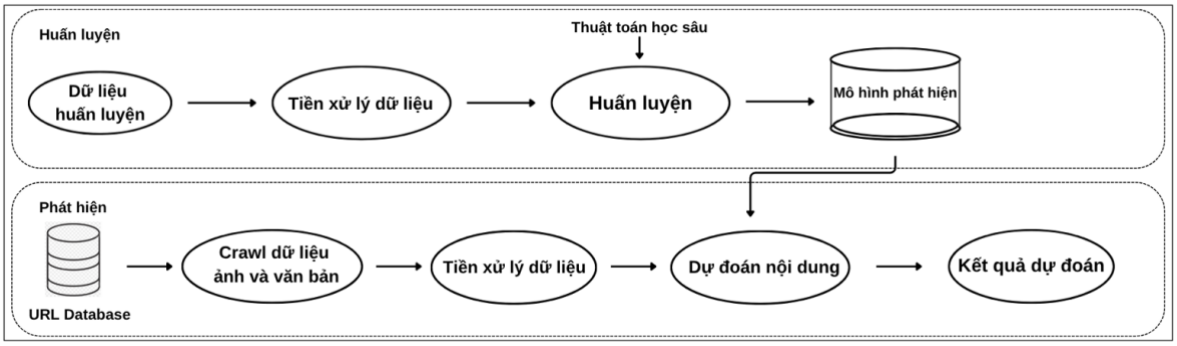
Bài toán 1: Phát hiện tin bài độc hại tiếng Việt dựa trên mô hình học sâu đa phương thức. Dữ liệu đầu vào gồm văn bản và hình ảnh của tin bài; đầu ra là nhãn phân loại. Lời giải của bài toán 1 là câu trả lời cho Câu hỏi nghiên cứu RQ1.

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 trình bày kiến trúc tổng thể của hệ thống phát hiện tin bài độc hại tiếng Việt dựa trên kết hợp đa phương thức. Kiến trúc hệ thống được xây dựng nhằm giải quyết các vấn đề đã đặt ra, bao gồm khai thác đồng thời đặc trưng văn bản và hình ảnh, cũng như nâng cao hiệu quả phát hiện trong điều kiện dữ liệu gán

nhãn hạn chế. Hệ thống được tổ chức thành hai giai đoạn chính: giai đoạn huấn luyện và giai đoạn phát hiện, như minh họa trong Hình 1.5.



Hình 1.5. Kiến trúc tổng thể cho phát hiện tin bài độc hại

Chức năng chính của hệ thống là:

Xây dựng mô hình phát hiện tin bài độc hại Tiếng Việt từ dữ liệu huấn luyện.

Chức năng này thực hiện tiền xử lý dữ liệu, trích xuất đặc trưng văn bản và hình ảnh, huấn luyện mô hình phát hiện bằng thuật toán học sâu.

Đầu vào: Tập dữ liệu huấn luyện tiếng Việt bao gồm văn bản, hình ảnh và nhãn tương ứng.

Xử lý: Tiền xử lý dữ liệu, trích xuất đặc trưng văn bản và hình ảnh, huấn luyện mô hình phân loại và kiểm tra.

Đầu ra: File lưu mô hình phát hiện tin bài độc hại và độ chính xác kiểm tra.

Quản lý nguồn dữ liệu (URL, trang mạng xã hội).

Chức năng này quản lý danh sách các nguồn dữ liệu trực tuyến phục vụ thu thập tin bài, bao gồm các tính năng xem danh sách, thêm mới, sửa, xóa và cập nhật.

Đầu vào: Thông tin các nguồn dữ liệu được giám sát.

Xử lý: Thực hiện các chức năng quản lý như thêm mới, chỉnh sửa, xoá và cập nhật danh sách nguồn dữ liệu, lưu trữ thông tin trong CSDL.

Đầu ra: Thông tin các nguồn dữ liệu được được giám sát và lưu trong CSDL.

Thu thập dữ liệu văn bản và hình ảnh.

Chức năng này thu thập dữ liệu tin bài từ các nguồn đã được quản lý, bao gồm trích xuất văn bản và hình ảnh từ nguồn dữ liệu.

Đầu vào: Danh sách thông tin nguồn dữ liệu.

Xử lý: Thực hiện crawl dữ liệu từ các nguồn, lấy nội dung văn bản, hình ảnh và lưu trữ dữ liệu vào CSDL.

Đầu ra: Tập dữ liệu tin bài gồm văn bản và hình ảnh được lưu trong CSDL.

Phát hiện tin bài độc hại

Chức năng này xử lý trích xuất đặc trưng, phân loại và phát hiện tin bài độc hại dựa trên mô hình đã huấn luyện.

Đầu vào: Dữ liệu thu thập từ nguồn dữ liệu (văn bản và hình ảnh).

Xử lý: Tiền xử lý văn bản, chuyển đổi dữ liệu về định dạng phù hợp với mô hình học sâu, phát hiện.

Đầu ra: Kết quả phân loại tin bài.

Lưu trữ kết quả phân loại và hiển thị.

Chức năng này lưu trữ toàn bộ kết quả phát hiện tin bài để phục vụ tra cứu, phân tích và sinh cảnh báo đối với các tin bài độc hại.

Đầu vào: Kết quả phân loại tin bài.

Xử lý: Kết nối đến CSDL, lưu toàn bộ kết quả dự đoán vào CSDL, hiển thị kết quả trên giao diện người dùng

Đầu ra: Các thông tin được lưu trữ đầy đủ trong CSDL và kết quả hiển thị cho người dùng.

1.6. KẾT LUẬN CHƯƠNG 1

Chương 1 đã hệ thống hóa các khía cạnh lý luận và thực tiễn cơ bản liên quan đến bài toán phát hiện tin bài độc hại tiếng Việt. Từ việc làm rõ khái niệm, tiêu chí nhận diện đến phân tích đặc trưng ngôn ngữ và vai trò của hình ảnh, chương đã chỉ ra rằng tin bài độc hại trên không gian mạng thường biểu hiện dưới dạng đa phương thức, tinh vi về ngôn ngữ và dễ lợi dụng yếu tố trực quan để tăng tính thuyết phục. Phân tích các nghiên cứu trong và ngoài nước cho thấy đã có nhiều tiến bộ về mặt mô hình và kỹ thuật, tuy nhiên phần lớn các công trình trong nước vẫn chủ yếu tập

trung vào phương thức văn bản đơn lẻ, đồng thời còn dựa nặng vào dữ liệu gán nhãn và thường giải bài toán ở mức phân loại nhị phân.

Trên cơ sở đánh giá các mô hình học máy được sử dụng phổ biến, chương đã giới thiệu các lựa chọn kỹ thuật phù hợp cho luận án: mô hình ngôn ngữ chuyên biệt cho tiếng Việt (PhoBERT) nhằm trích xuất đặc trưng ngữ nghĩa, mô hình thị giác hiện đại (Swin Transformer V2) để trích xuất đặc trưng trực quan, và phương pháp học bán giám sát (gán nhãn giả) để giảm phụ thuộc vào dữ liệu gán nhãn. Đồng thời, chương cũng nêu rõ các thách thức chính gồm: khó khăn trong thu thập và gán nhãn dữ liệu tiếng Việt, tính đa dạng và biến đổi nhanh của thủ pháp phát tán tin độc hại, vấn đề bất cân bằng nhãn và yêu cầu tài nguyên tính toán khi xử lý đa phương thức.

Từ những phân tích trên, có thể rút ra hai luận điểm nền tảng làm đầu mối cho các chương tiếp theo: (i) cần thiết phải phát triển các giải pháp đa phương thức, kết hợp đồng thời văn bản và hình ảnh để nắm bắt đầy đủ các đặc trưng của tin độc hại, (ii) việc tích hợp học bán giám sát và mô hình chuyên biệt cho tiếng Việt là hướng khả thi để cải thiện tính ứng dụng trong điều kiện dữ liệu nhãn hạn chế. Các nhận định này sẽ là cơ sở khoa học để xây dựng kiến trúc mô hình, chiến lược thu thập dữ liệu và thiết kế thử nghiệm trong Chương 2 và Chương 3 của luận án.

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

Chương này tập trung xây dựng, thử nghiệm và đánh giá mô hình đa phương thức cho phát hiện tin bài độc hại tiếng Việt dựa trên kết hợp đặc trưng văn bản và hình ảnh. Nội dung chính của chương gồm: (i) đặt vấn đề và cơ sở hình thành hướng tiếp cận đa phương thức, (ii) giới thiệu chi tiết kiến trúc mô hình phát hiện đa lớp, (iii) mô tả quy trình tiền xử lý dữ liệu, huấn luyện và triển khai, (iv) trình bày kết quả thử nghiệm, phân tích và so sánh với một số phương pháp liên quan. Trên cơ sở đó, chương 2 khẳng định tính khả thi của hướng tiếp cận đa phương thức, đồng thời tạo nền tảng cho việc mở rộng sang học bán giám sát được trình bày ở Chương 3.

2.1. ĐẶT VẤN ĐỀ

Chương 1 đã trình bày tổng quan về tin bài độc hại và các phương pháp phát hiện hiện có, qua đó cho thấy phần lớn các nghiên cứu trong nước mới dừng lại ở xử lý nội dung văn bản đơn lẻ, trong khi hình ảnh thường được coi là thành phần minh họa, ít khi được khai thác một cách hệ thống. Điều này dẫn đến hạn chế trong khả năng nhận diện, bởi trong thực tế, tin bài độc hại thường không chỉ dựa trên ngôn ngữ mà còn được tăng cường sức thuyết phục thông qua các yếu tố trực quan. Ngoài ra, nhiều công trình vẫn tiếp cận bài toán dưới dạng phân loại nhị phân, khiến cho mô hình khó thích ứng khi đối mặt với dữ liệu thực tế có tính đa dạng cao và thay đổi liên tục.

Trong bối cảnh truyền thông số hiện nay, tin bài độc hại thường được xây dựng dưới dạng đa phương thức, kết hợp đồng thời văn bản và hình ảnh để tạo hiệu ứng tác động mạnh mẽ hơn. Văn bản có thể sử dụng ngôn ngữ mơ hồ, ám chỉ, hoặc mang tính cường điệu, trong khi hình ảnh đi kèm có thể bị chỉnh sửa, cắt ghép hoặc sử dụng sai ngữ cảnh nhằm dẫn dắt nhận thức người xem. Chính sự phối hợp này khiến cho việc phát hiện nội dung độc hại trở nên phức tạp, bởi nếu chỉ phân tích một thành phần riêng lẻ, nhiều thông tin quan trọng có thể bị bỏ sót, làm giảm độ chính xác của hệ thống.

Các nghiên cứu quốc tế gần đây đã bắt đầu chú trọng đến hướng tiếp cận đa phương thức, kết hợp văn bản và hình ảnh trong cùng một kiến trúc, nhờ đó đạt được kết quả khả quan hơn. Tuy nhiên, việc áp dụng trực tiếp cho tiếng Việt gặp nhiều thách thức do đặc thù ngôn ngữ, sự hạn chế về dữ liệu gán nhãn và sự khác biệt trong bối cảnh xã hội. Điều này đòi hỏi cần có một mô hình được thiết kế chuyên biệt, vừa tận dụng các thành tựu mới trong học sâu, vừa phù hợp với đặc điểm ngữ nghĩa và thực tế dữ liệu tin bài độc hại tiếng Việt.

Xuất phát từ yêu cầu đó, Chương 2 tập trung xây dựng và thử nghiệm mô hình phát hiện tin bài độc hại tiếng Việt dựa trên kết hợp đặc trưng văn bản và hình ảnh. Mục tiêu của chương là khai thác hiệu quả mối quan hệ bổ sung giữa hai nguồn dữ liệu để nâng cao khả năng nhận diện, đặc biệt trong những trường hợp phức tạp mà nội dung độc hại được ngụy trang tinh vi.

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

Mục này mô tả và phân tích mô hình đa phương thức cho phát hiện tin bài độc hại về chính trị tiếng Việt dựa trên kết hợp các đặc trưng văn bản và hình ảnh cho bài toán phân loại đa lớp, trong đó nhãn độc hại được tách thành 3 nhãn con: “Vu khống, bôi nhọ”, “Kêu gọi chống đối” và “Thúc đẩy tư tưởng cực đoan”.

Trong mô hình đa phương thức đề xuất, PhoBERT được lựa chọn làm bộ mã hóa văn bản nhờ khả năng biểu diễn ngữ nghĩa vượt trội cho tiếng Việt. PhoBERT kế thừa kiến trúc RoBERTa nhưng được tiền huấn luyện trên một kho ngữ liệu tiếng Việt quy mô lớn, qua đó khắc phục những hạn chế của các mô hình BERT và RoBERTa vốn huấn luyện chủ yếu trên dữ liệu tiếng Anh. Các nghiên cứu trước đây đã chứng minh PhoBERT đạt hiệu quả cao hơn hẳn trong các tác vụ phân loại văn bản, phân tích cảm xúc và nhận dạng thực thể tiếng Việt. Do đó, PhoBERT được xem là lựa chọn tối ưu để xử lý ngữ cảnh và trích xuất đặc trưng ngôn ngữ từ các tin bài tiếng Việt.

Đối với dữ liệu hình ảnh, luận án sử dụng Swin Transformer V2. Đây là một trong những mô hình thị giác hiện đại, có khả năng học biểu diễn đa cấp độ thông qua cơ chế “shifted windows”, cho phép nắm bắt cả chi tiết cục bộ và quan hệ toàn

cục trong ảnh. So với các CNN truyền thống, Swin Transformer V2 thể hiện ưu thế rõ rệt trong các tác vụ phân loại ảnh, nhận dạng vật thể và đặc biệt là các bài toán cần kết hợp thị giác với ngôn ngữ. Hơn nữa, Swin Transformer V2 có khả năng mở rộng tốt, phù hợp để xử lý các bộ dữ liệu ảnh đa dạng, phức tạp và thường xuyên bị thao túng trong tin bài độc hại, như ảnh cắt ghép, chỉnh sửa, gán ghép sai ngữ cảnh.

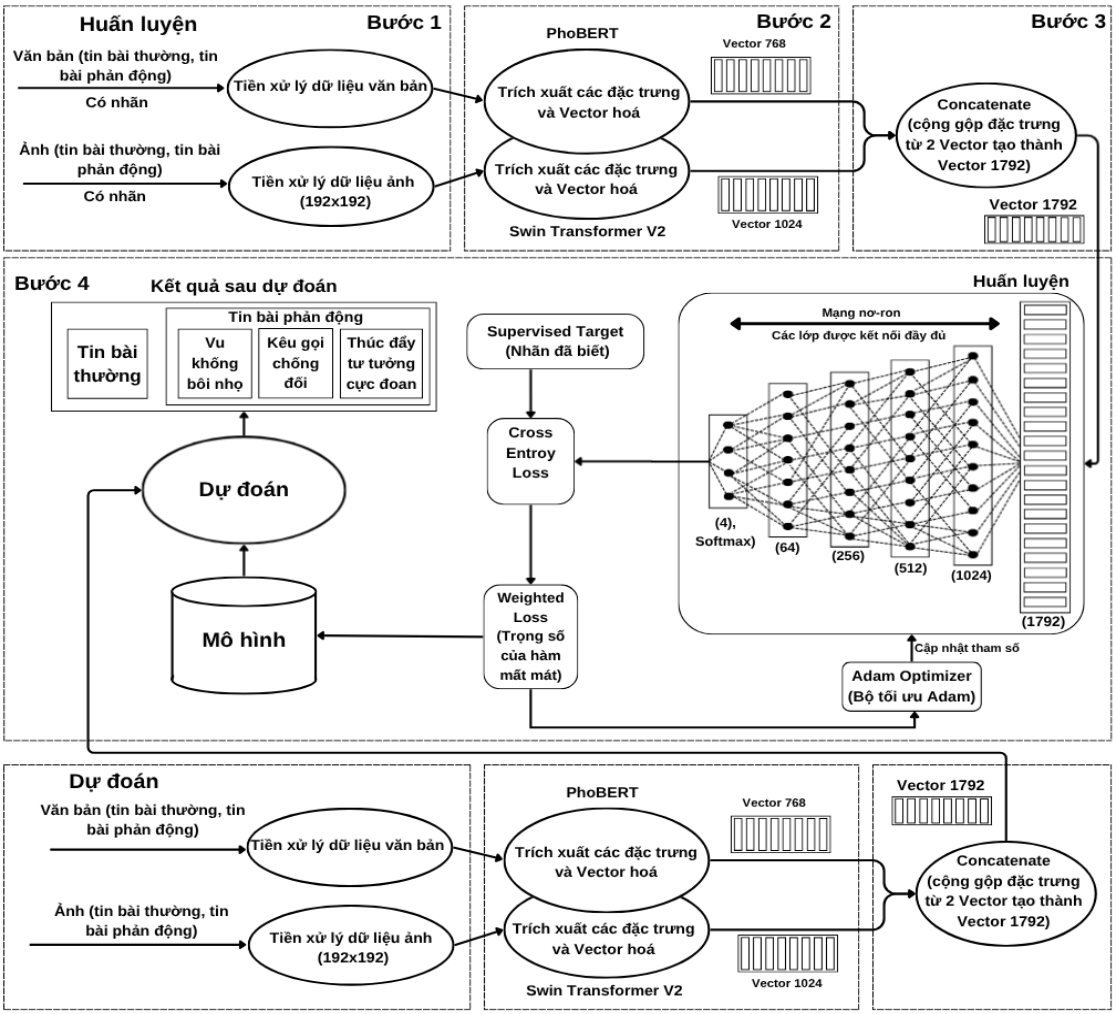
Việc kết hợp PhoBERT và Swin Transformer V2 trong mô hình đa phương thức đề xuất vì vậy mang tính chiến lược: PhoBERT đảm nhiệm việc nắm bắt đặc trưng ngôn ngữ, trong khi Swin Transformer V2 khai thác các tín hiệu trực quan từ hình ảnh. Hai dòng đặc trưng được sinh bởi PhoBERT và Swin Transformer V2 sau khi được chuẩn hóa và hợp nhất sẽ bổ trợ lẫn nhau, giúp mô hình phản ánh đầy đủ hơn tính chất phức tạp của tin bài độc hại về chính trị trên không gian mạng.

2.2.1. Giới thiệu mô hình phát hiện

Bài toán phân loại đa lớp được đặt ra trong bối cảnh các tin bài độc hại có mục đích và hướng đến tập người xem khác nhau nên cần được phát hiện chính xác để có phương thức xử lý, ngăn chặn phù hợp. Chẳng hạn, các tin bài thuộc nhóm "Vu khống, bôi nhọ" cần được nhận diện chính xác và có hình thức xử lý phù hợp như tổ chức các chiến dịch tuyên truyền, giáo dục, cung cấp thông tin chính thức, đầy đủ và quảng bá rộng rãi đến từng người dân. Vì vậy, mô hình phân loại đa lớp phản ánh sát thực tế hơn bản chất và phương thức xử lý tin bài hiện nay.

Hình 2.1 mô tả kiến trúc của mô hình phát hiện tin bài độc hại dựa trên kết hợp các đặc trưng văn bản và hình ảnh cho bài toán phân loại đa lớp. Theo đó, đặc trưng ngữ nghĩa của văn bản được trích xuất bằng PhoBERT, trong khi đặc trưng hình ảnh được khai thác từ Swin Transformer V2. Các đặc trưng này sau đó được chuẩn hóa và đưa vào tầng hợp nhất đa phương thức, nhằm học tối ưu mối quan hệ tương quan giữa văn bản và hình ảnh.

Nghiên cứu áp dụng hàm kích hoạt softmax tại lớp phân loại đa lớp. Cách tiếp cận này cho phép mô hình học phân bố xác suất trên toàn bộ tập nhãn, từ đó xác định lớp có xác suất cao nhất tương ứng với dạng độc hại chính trong tin bài. Hàm mất mát được sử dụng là Cross-Entropy điều chỉnh cho kịch bản đa lớp, nhằm tối ưu khả năng phân biệt giữa các lớp nhãn khác nhau.



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

Mô hình đa phương thức cho bài toán phân loại đa lớp có các ưu điểm sau: Thứ nhất, mô hình có khả năng học tốt hơn mối quan hệ phụ thuộc giữa các nhãn, điều thường bị bỏ qua trong các thiết kế đa lớp truyền thống. Thứ hai, cách tiếp cận này hỗ trợ phân tích rõ ràng hơn sự phân bố nguy cơ của từng loại độc hại trong một tin bài. Thứ ba, kiến trúc vẫn đảm bảo khả năng mở rộng, cho phép bổ sung nhãn mới mà không làm thay đổi cấu trúc cốt lõi.

2.2.2. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một bước không thể thiếu nhằm chuẩn hóa và định dạng lại dữ liệu trước khi đưa vào mô hình huấn luyện. Trong nghiên cứu này, dữ liệu đầu vào bao gồm cả văn bản và hình ảnh, do đó quá trình tiền xử lý được thực hiện tương ứng cho từng loại. Với dữ liệu văn bản, đặc biệt là văn bản được thu thập tự động từ Internet, thường tồn tại dưới dạng không đồng nhất, chứa nhiều ký tự lạ, biểu tượng cảm xúc, liên kết hoặc các từ viết tắt khó xử lý. Do đó, bước đầu tiên là

chuẩn hóa định dạng văn bản bằng cách chuyển toàn bộ nội dung về chữ thường, loại bỏ các ký tự đặc biệt và thông tin dư thừa như URL hay biểu tượng không cần thiết. Sau đó, các từ viết tắt ít phổ biến cũng được loại bỏ nhằm giảm nhiễu cho mô hình. Tiếp theo, văn bản được phân tách từ bằng công cụ tách từ tiếng Việt, nhằm đảm bảo rằng chuỗi văn bản đầu vào phản ánh đúng các đơn vị ngữ nghĩa độc lập, giúp mô hình học sâu dễ dàng xử lý hơn. Sau khi đã phân tách từ, văn bản sẽ được đệm để đạt độ dài chuỗi cố định, phù hợp với định dạng đầu vào mà PhoBERT yêu cầu. Cuối cùng, toàn bộ văn bản đã chuẩn hóa được mã hóa thành chuỗi các mã token ID để mô hình PhoBERT có thể tiếp nhận và xử lý trực tiếp.

Với dữ liệu hình ảnh, quá trình tiền xử lý nhằm đưa ảnh về định dạng chuẩn phù hợp với mô hình Swin Transformer V2. Cụ thể, tất cả các ảnh được thay đổi kích thước về 192×192 điểm ảnh để đảm bảo tính nhất quán trong đầu vào. Tiếp theo, ảnh được chuẩn hóa giá trị điểm ảnh nhằm đưa các kênh màu về cùng một thang đo chuẩn. Cuối cùng, ảnh được chuyển đổi sang dạng tensor, định dạng dữ liệu phù hợp cho đầu vào của các mô hình học sâu xử lý ảnh như Swin Transformer V2.

Thông qua các bước tiền xử lý trên, dữ liệu đầu vào cho mô hình được chuẩn hóa và làm sạch, tạo tiền đề quan trọng cho việc huấn luyện mô hình đa phương thức nhằm phát hiện tin bài độc hại tiếng Việt một cách hiệu quả.

2.2.3. Huấn luyện và phát hiện

Giải thuật 2.1 mô hình hoá quá trình huấn luyện được biểu diễn trên Hình 2.1.

Giải thuật 2.1: Quá trình huấn luyện

Input:

x_{i_t}: Văn bản đầu vào (training sample).

x_{i_i}: Ảnh đầu vào (training sample).

y_i: Nhãn lớp của mẫu xi, yi ∈ {0,1,2,3}.

S: Tập hợp các mẫu huấn luyện có nhãn.

Output:

θ *:* Tham số của mô hình sau khi huấn luyện.

Process:

PhoBERT(x): Trích xuất đặc trưng văn bản.

SwinTransformerV2(x): Trích xuất đặc trưng ảnh.

$f_\theta(\mathbf{x}')$ Mạng nơ-ron FFNN (lớp cuối softmax) với tham số có thể huấn luyện θ.

for t in [1, num epochs] do

for each minibatch B do

B1: Tiền xử lý

B2: Trích xuất đặc trưng

$\mathbf{x}'_{t\in B} \leftarrow \text{PhoBERT}(\mathbf{x}_{t\in B})$ ▷ Trích xuất đặc trưng văn bản

$\mathbf{x}'_{i\in B} \leftarrow \text{SwinTransformerV2}(\mathbf{x}_{i\in B})$ ▷ Trích xuất đặc trưng ảnh

B3: Kết hợp đặc trưng

$\mathbf{x}' \in B \leftarrow \text{Concatenate}(\mathbf{x}'_t, \mathbf{x}'_i)$ ▷ Kết hợp đặc trưng văn bản và ảnh

B4: Huấn luyện mô hình

$z_{i\in B} \leftarrow f_\theta(\mathbf{x}')$ Đầu ra của mạng nơ-ron (sau softmax để tạo ra xác suất dự đoán)

$$L_s^{i\in B} \leftarrow \frac{1}{|B|}\sum_{i\in(B\cap S)}[y_i \log(z_i) + (1-y_i)\log(1-z_i)]$$

$L_s = -\frac{1}{|B|}\sum_{i\in B}\sum_{c=1}^{C} y_{ic}\log(z_{ic})$ ▷ Thành phần mất mát giám sát (Multiclass Cross-Entropy)

Cập nhật θ ▷ Tối ưu hóa các tham số của mạng nơ-ron

end for (kết thúc vòng lặp minibatch)

end for (kết thúc vòng lặp epoch)

return $\boldsymbol{\theta}$

Quá trình huấn luyện mô hình phát hiện tin bài độc hại đa lớp đề xuất trong luận án được thực hiện theo kiến trúc học sâu đa phương thức, kết hợp giữa đặc trưng ngôn ngữ và thị giác. Tổng thể hệ thống được triển khai theo 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 chiều theo công thức sau:

$$\mathrm{x}' \in \mathbb{R}^{d_t+d_i} = \text{Concatenate}([\mathrm{x}'_t, \mathrm{x}'_i]) \qquad (2.1)$$

Trong đó:

- x' là vector đặc trưng hợp nhất thuộc không gian vector thực R với số chiều bằng $d_t + d_i$.
- d_t là số chiều của vector đặc trưng văn bản x_t'.
- d_i là số chiều của vector đặc trưng ảnh x_i'.
- $Concatenate([x'_t, x'_i])$ là hàm nối (concatenate) vector x_t' và x_i' theo chiều cuối cùng.
- Cụ thể: $d_t = 768, \quad d_i = 1024$ và $d_t + d_i = 1792$

Phép ghép nối hai vector đặc trưng sau bước trích xuất được xem là chiến lược kết hợp sớm (early fusion) ở mức đặc trưng. Phép nối này giúp giữ nguyên toàn bộ thông tin từ cả văn bản và hình ảnh, thay vì hòa trộn hoặc rút gọn như các phương pháp cộng hoặc trung bình, vốn dễ làm mất thông tin chi tiết.

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 đa phương thức 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.

Bảng 2.1. Kiến trúc Mạng Nơ-ron cho mô hình phát hiện đa lớp

Mạng nơ-ron	**Diễn giải**
Lớp kết nối đầy đủ thứ nhất (Dense)	Lớp này có 1024 nơ-ron và sử dụng hàm kích hoạt ReLU để tạo ra tính phi tuyến tính.
Lớp Dropout thứ nhất	Lớp này được sử dụng để giảm thiểu overfitting bằng cách ngẫu nhiên bỏ qua một số nơ-ron trong quá trình huấn luyện. Tỷ lệ dropout được đặt là 0.2.
Lớp kết nối đầy đủ thứ hai (Dense):	Lớp này có 512 nơ-ron và sử dụng hàm kích hoạt ReLU.

Lớp Dropout thứ hai	Lớp này được sử dụng để giảm thiểu overfitting với tỷ lệ dropout là 0.2.
Lớp kết nối đầy đủ thứ ba (Dense):	Lớp này có 256 nơ-ron và sử dụng hàm kích hoạt ReLU.
Lớp Dropout thứ ba	Lớp này được sử dụng để giảm thiểu overfitting với tỷ lệ dropout là 0.2.
Lớp kết nối đầy đủ thứ tư (Dense): Lớp kết nối đầy đủ thứ năm (Dense):	Lớp này có 64 nơ-ron và sử dụng hàm kích hoạt ReLU. Lớp này có 4 nơ-ron, sử dụng hàm kích hoạt softmax để đưa ra xác suất dự đoán

Mô hình phân loại tin tức sử dụng mạng nơ-ron tự thiết kế với kiến trúc nhiều lớp. Lớp đầu vào nhận vector đặc trưng hợp nhất có kích thước 1792 chiều, biểu diễn các đặc trưng số học trích xuất từ cả văn bản và hình ảnh của tin tức. Tiếp theo là một chuỗi các lớp ẩn, được thiết kế để tự động trích xuất các đặc trưng bậc cao và phức tạp từ dữ liệu. Trong kiến trúc này, số lượng lớp ẩn và số lượng nơ-ron trong mỗi lớp được điều chỉnh để tối ưu hóa hiệu suất. Ví dụ, mô hình có thể sử dụng kiến trúc với các lớp ẩn giảm dần số lượng nơ-ron, chẳng hạn 1024, 512, 256 và 64 nơ-ron. Hàm kích hoạt ReLU được sử dụng giữa các lớp ẩn để tăng khả năng học các mối quan hệ phi tuyến tính. Công thức của một tầng tuyến tính như sau:

$$\mathrm{a} = \mathrm{x}'\mathrm{w} + \mathrm{b} \quad \textit{(phép nhân ma trận và cộng bias)} \qquad (2.2)$$

Trong đó:

- x' là dữ liệu đầu vào (đã chuyển vị).
- w là trọng số của lớp.
- b là độ lệch (bias) của lớp.
- a là đầu ra của lớp (chưa qua hàm kích hoạt).

Quá trình xử lý dữ liệu qua mỗi lớp ẩn bao gồm việc tính toán tích vô hướng của vector trước với ma trận trọng số (w) của lớp đó, sau đó cộng với vector bias (b). Ví dụ, đầu ra 'a' của lớp đầu tiên được tính bằng công thức $a = x'w + b$, trong đó x' là vector đặc trưng hợp nhất, w là ma trận trọng số, và b là vector bias. Kết quả này sau đó được truyền qua hàm kích hoạt ReLU trước khi được đưa vào lớp tiếp theo.

Cuối cùng, lớp đầu ra gồm 4 nơ-ron, tương ứng với bốn nhóm tin bài: tin bài thường, vu khống bôi nhọ, kêu gọi chống đối, và thúc đẩy tư tưởng cực đoan. Các giá trị logit này được đưa qua hàm softmax để tạo ra phân phối xác suất trên 4 nhãn, đảm bảo tổng xác suất bằng 1. Hàm softmax được định nghĩa như sau:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \qquad j \in \{1,2,3,4\} \text{ (số lượng lớp đầu ra).} \quad (2.3)$$

Trong đó:

- z_i là giá trị logit tại đầu ra của nơ-ron thứ i.
- e^{z_i} e^{z_i} là giá trị hàm mũ của logit đó.
- $\sum_{j=1}^{n} e^{z_j}$ là tổng của tất cả giá trị hàm mũ từ các logit.

Như vậy, softmax chuyển đổi các giá trị logit thành xác suất dự đoán. Mỗi nhánh softmax dự đoán xác suất cho một nhóm hành vi, cho phép mô hình gán đồng thời nhiều nhãn.

Trong quá trình huấn luyện, mạng nơ-ron được tối ưu hóa bằng thuật toán Adam với tốc độ học khởi đầu là α=1×10^{-4}. Mô hình được huấn luyện theo từng minibatch có kích thước 32 mẫu. Toàn bộ trọng số của các mô hình nền tảng (PhoBERT và Swin Transformer V2) được giữ nguyên (đóng băng), tức là chỉ huấn luyện phần mạng phân loại phía sau nhằm tận dụng triệt để đặc trưng đã được tiền huấn luyện.

Để đo lường sai số trong quá trình huấn luyện và cập nhật tham số mô hình, nghiên cứu sử dụng hàm mất mát Cross-Entropy (CE) áp dụng riêng biệt cho từng nhánh softmax trong tầng phân loại. Hàm này phản ánh mức độ sai lệch giữa xác suất dự đoán và phân phối nhãn thực tế trong từng nhóm hành vi độc hại. Công thức của hàm mất mát Cross-Entropy như sau:

$$L_s^{i\in B} = \frac{1}{|B|}\sum_{i\in B}[y_i \log(z_i) + (1 - y_i)\log(1 - z_i)] \quad (2.4)$$

Trong đó:

- L_s là giá trị mất mát (loss) cho cả batch B.
- $|B|$ là kích thước của batch B (số lượng mẫu trong batch).

- $\sum_{i \in B}$ là tổng trên tất cả các mẫu i trong batch B.
- y_i là nhãn thực tế của mẫu thứ i (0 hoặc 1).
- z_i là xác suất dự đoán của mẫu thứ i (giá trị sau sigmoid).
- log( ) là logarit tự nhiên (ln).

Hàm mất mát tổng thể được tính dựa trênhàm cross-entropy tại lớp phân loại. Cách tiếp cận này phù hợp với cấu trúc phân loại đa lớp, giúp mô hình học cách phân biệt hiệu quả giữa các lớp nhãn khác nhau và lựa chọn lớp đặc trưng nhất cho mỗi tin bài.

Cuối cùng tập hợp các tham số của mô hình là θ được huấn luyện hoàn chỉnh, quá trình dự đoán được thực hiện nhằm phân loại các tin bài đầu vào. Quy trình dự đoán của hệ thống được thể hiện trực quan trong Hình 2.1.

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 (concatenate) 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

Quá trình cài đặt và triển khai mô hình được thực hiện bằng Python 3.12.6 với nền tảng PyTorch. Các thư viện chính bao gồm:

- Xử lý dữ liệu và tính toán: numpy, pandas, scikit-learn.
- Mạng nơ-ron và tối ưu: torch, torch.nn, torch.optim.
- Xử lý văn bản: transformers (sử dụng PhoBERT).
- Xử lý hình ảnh: PIL và timm (sử dụng Swin Transformer V2).
- Đánh giá mô hình: scikit-learn (accuracy, precision, recall, F1-score, confusion matrix).
- Quản lý dữ liệu: torch.utils.data (Dataset, DataLoader).

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, hệ thống tận dụng Đơn vị Xử lý Trung tâm (CPU) Intel Xeon E5-2680 v3 để quản lý luồng dữ liệu và các tác vụ tiền xử lý, trong khi việc tăng tốc huấn luyện được đảm nhiệm bởi Đơn vị 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. Các thiết lập và cấu hình đảm bảo khả năng mở rộng linh hoạt cho các thử nghiệm khác nhau trên dữ liệu tiếng Việt.

2.3. THỬ NGHIỆM VÀ ĐÁNH GIÁ

Để đánh giá mô hình đa phương thức cho phát hiện tin bài độc hại đa lớp tiếng Việt, luận án thực hiện thử nghiệm mô hình đề xuất ở 2 chế độ: phát hiện nhị phân và phát hiện đa lớp. Mục đích của thử nghiệm mô hình đề xuất ở chế độ phát hiện nhị phân là đánh giá hiệu suất phát hiện tổng thể và so sánh với các mô hình đơn phương thức, cũng như so sánh với các mô hình phát hiện nhị phân trong các nghiên cứu liên quan. Trong chế độ phát hiện nhị phân, mô hình được thử nghiệm

để phân loại tin bài thành 2 nhóm nhãn: ‘Tin bài thường’ và ‘Tin bài độc hại’. Với chế độ phát hiện đa lớp, mục đích của thử nghiệm mô hình đề xuất là đánh giá hiệu suất phát hiện tổng thể cũng như chi tiết cho từng nhãn và so sánh với các mô hình phát hiện đa lớp, đa phương thức khác. Trong chế độ phát hiện đa lớp, mô hình được thử nghiệm để phân loại tin bài thành 1 trong 4 nhóm nhãn: ‘Tin bài thường’, ‘Vu khống, bôi nhọ’, ‘Kêu gọi chống đối’ và ‘Thúc đẩy tư tưởng cực đoan’.

2.3.1. Tập dữ liệu thử nghiệm

Luận án triển khai thực nghiệm bộ dữ liệu được xây dựng và sử dụng chủ yếu tập trung vào nhóm tin độc hại mang nội dung độc hại về chính trị. Đây là loại hình tin độc hại phổ biến, có tác động xã hội nghiêm trọng và thường xuyên xuất hiện trên không gian mạng. Việc lựa chọn nhóm dữ liệu này giúp mô hình được huấn luyện trên một tập thông tin có tính đặc thù cao, đồng thời phản ánh đúng những thách thức trong nhận diện nội dung độc hại nhạy cảm và phức tạp. Hai bộ dữ liệu thử nghiệm được sử dụng, bao gồm một bộ cho chế độ nhị phân và một bộ cho chế độ đa lớp.

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 4.000 tin bài tiếng Việt, bao gồm:

- 1.000 tin bài độc hại dạng văn bản và 1.000 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’.

- 1.000 tin bài thường dạng văn bản và 1.000 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: 3.000 bài viết dưới dạng văn bản và 3.000 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

Tập dữ liệu 4000 tin bài được chia thành 70% cho huấn luyện, 15% cho đánh giá và 15% cho kiểm tra sử dụng để đánh giá mô hình trong chế độ phát hiện nhị phân. Bảng 2.2 cung cấp hiệu suất phát hiện của mô hình kết hợp đa phương thức đề xuất trong chế độ phát hiện nhị phân và các mô hình đơn phương thức. Bảng 2.3 cung cấp chi tiết ma trận nhầm lẫn của mô hình đề xuất trong chế độ phát hiện nhị phân cho 2 nhóm tin bài: tin bài thường, tin bài độc hại. Có thể thấy mô hình kết hợp đa phương thức đề xuất dựa trên PhoBERT+Swin Transformer V2 cho hiệu suất phát hiện chung cao hơn đáng kể so với các mô hình đơn phương thức,

Bảng 2.4 cung cấp số liệu so sánh hiệu suất phát hiện chung của mô hình kết hợp đa phương thức đề xuất với các mô hình kết hợp đa phương thức trong các nghiên cứu liên quan trong chế độ phát hiện nhị phân. Cũng có thể thấy, mô hình kết hợp đa phương thức đề xuất dựa trên PhoBERT+Swin Transformer V2 cho hiệu suất phát hiện chung vượt trội so với các mô hình kết hợp đa phương thức dựa trên BERT+Xception (Concatenate) và BERT+Xception (Maximum) [51].

Bảng 2.2. Hiệu suất phát hiện của mô hình đề xuất trong chế độ nhị phân và các mô hình đơn phương thức

Mô hình	Đặc trưng	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Xception	Ảnh	82,25	83,68	79,90	81,75
Swin Transformer V2	Ảnh	85,00	84,24	85,93	85,07
BERT	Văn bản	91,25	91,26	91,23	91,24
PhoBERT	Văn bản	96,50	96,49	96,60	96,50

PhoBERT+Swin Transformer V2	**Văn bản+Ảnh**	**97,00**	**97,00**	**97,10**	**97,00**

Bảng 2.3. Ma trận nhầm lẫn của mô hình đề xuất trong chế độ phát hiện nhị phân cho 2 nhóm tin bài: tin bài thường, tin bài độc hại

Lớp	Tin bài thường (400 tin)	Tin bài độc hại (400 tin)
Tin bài thường	382	18
Tin bài độc hại	8	392

Bảng 2.4. So sánh hiệu suất của mô hình đề xuất trong chế độ phát hiện nhị phân với nghiên cứu liên quan

Mô hình	Đặc trưng	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BERT+Xception (Concatenate) [55]	Văn bản+Ảnh	91,70	93,29	93,22	93,25
BERT+Xception (Maximum) [55]	Văn bản+Ảnh	91,68	93,76	92,83	93,29
PhoBERT+Swin Transformer V2	**Văn bản+Ảnh**	**97,00**	**97,00**	**97,10**	**97,00**

2.3.2.2. Kết quả thử nghiệm phát hiện đa lớp

Tập dữ liệu 12.000 tin bài được chia thành 70% cho huấn luyện, 15% cho đánh giá và 15% cho kiểm tra. Mô hình kết hợp đề xuất dựa trên PhoBERT và Swin Transformer V2 với các mô hình kết hợp khác được thử nghiệm với cùng tập dữ liệu để đo hiệu suất chung (gồm các độ đo Accuracy, Precision, Recall, F1-score) và tỷ lệ phát hiện (DR – Detection rate) cho từng nhãn tin bài, gồm 'tin bài thường', 'Vu khống bôi nhọ', 'Kêu gọi chống đối' và 'Thúc đẩy tư tưởng cực đoan'. Kịch bản này được sử dụng để đánh giá và so sánh khả năng phát hiện đa lớp của các mô hình kết hợp.

Bảng 2.5 và 2.6 lần lượt cung cấp hiệu suất chung và tỷ lệ phát hiện cho từng nhãn tin bài của mô hình trong chế độ đa lớp đề xuất và các mô hình kết hợp xử lý

dữ liệu văn bản và hình ảnh. Bảng 2.7 cung cấp chi tiết ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình đề xuất trong chế độ đa lớp. Có thể thấy mô hình kết hợp đề xuất dựa trên PhoBERT và Swin Transformer V2 cho hiệu suất phát hiện vượt trội so với các mô hình kết hợp khác.

Bảng 2.5. Hiệu suất phát hiện của mô hình đề xuất trong chế độ đa lớp và các mô hình kết hợp trên dữ liệu đánh giá

Mô hình	Đặc trưng	Accuracy %	Precision %	Recall %	F1-score %
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**95,67**	**95,71**	**95,67**	**95,68**
PhoBERT + Xception	Văn bản + Ảnh	95,00	94,07	95,81	94,93
PhoBERT + VGG19	Văn bản + Ảnh	77,33	78,28	76,07	77,16
BERT + Swin Transformer V2	Văn bản + Ảnh	92,44	92,60	92,40	92,50
BERT + Xception	Văn bản + Ảnh	91,44	91,86	91,24	91,55
BERT + VGG19	Văn bản + Ảnh	69,78	68,50	64,63	66,51
RoBERTa + VGG19	Văn bản + Ảnh	72,67	70,19	71,28	70,73
RoBERTa + Swin Transformer V2	Văn bản + Ảnh	64,89	70,32	62,98	66,45
RoBERTa + Xception	Văn bản + Ảnh	68,67	67,30	68,10	67,70

Bảng 2.6. Tỉ lệ phát hiện (DR) của mô hình đề xuất trong chế độ đa lớp và các mô hình kết hợp trên dữ liệu kiểm tra

Mô hình	Đặc trưng	Tin bài thường (%)	Vu khống, bôi nhọ (%)	Kêu gọi chống đối (%)	Thúc đẩy tư tưởng cực đoan (%)
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**97,78**	**94,67**	**97,33**	**96,00**
PhoBERT + Xception	Văn bản + Ảnh	95,33	90,00	90,00	94,00
PhoBERT + VGG19	Văn bản + Ảnh	99,78	53,33	90,67	63,33
BERT + Swin	Văn bản + Ảnh	96,44	91,33	96,67	96,33

Transformer V2					
BERT + Xception	Văn bản + Ảnh	98,67	95,67	90,67	91,33
BERT + VGG19	Văn bản + Ảnh	59,56	80,67	67,33	30,67
RoBERTa + VGG19	Văn bản + Ảnh	78,89	74,00	96,00	66,67
RoBERTa + Swin Transformer V2	Văn bản + Ảnh	88,22	90,67	66,67	22,00
RoBERTa + Xception	Văn bản + Ảnh	87,33	88,00	82,67	91,33

Quy ước: Lớp 0 – Tin bài thường, lớp 1Vu khống bôi nhọ, lớp 2 Kêu gọi chống đối, lớp 3 Thúc đẩy tư tưởng cực đoan.

Bảng 2.7. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình đề xuất trong chế độ đa lớp

Lớp	Tin bài thường (450 tin)	Vu khống, bôi nhọ (150 tin)	Kêu gọi chống đối (150 tin)	Thúc đẩy tư tưởng cực đoan (150 tin)
0	437	7	0	6
1	7	138	2	3
2	3	3	142	2
3	1	4	1	144

2.3.3. Nhận xét và đánh giá

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

Trong kịch bản nhị phân, mô hình kết hợp PhoBERT và Swin Transformer V2 đạt độ chính xác 97%, cao hơn so với PhoBERT đơn lẻ (96,5%) và vượt trội so với các mô hình xử lý ảnh như Xception (82,25%) hoặc Swin Transformer V2 đơn lẻ (85%). Khi mô hình đơn phương thức (PhoBERT) đã đạt hiệu suất cao (96.5%), việc tiệm cận ngưỡng tối ưu khiến biên độ cải thiện trở nên hẹp dần. Với mức tăng 0,5% của mô hình đa phương phức cho thấy khả năng trong việc khai thác các đặc trưng bổ trợ từ hình ảnh để giải quyết những trường hợp nhiễu hoặc nhập nhằng mà văn bản đơn thuần không thể xử lý.

Đặc biệt, khi so sánh với các nghiên cứu trước đó sử dụng BERT kết hợp Xception, mô hình đề xuất cũng thể hiện sự cải thiện rõ rệt, với mức F1-score đạt

97,0% so với khoảng 93% của các phương pháp cũ. Như vậy, sự lựa chọn PhoBERT, vốn tối ưu cho tiếng Việt, kết hợp với một kiến trúc thị giác hiện đại như Swin Transformer V2 đã mang lại sự cộng hưởng tốt, khắc phục được hạn chế khi chỉ dựa vào đặc trưng ngôn ngữ hoặc chỉ dựa vào đặc trưng hình ảnh.

Phân tích ma trận nhầm lẫn trong bài toán nhị phân (Bảng 2.3) cũng củng cố nhận định này. Mô hình dự đoán đúng 382/400 tin bài thường và 392/400 tin độc hại. Số lượng nhầm lẫn còn lại (18 tin bài thường bị gán thành độc hại và 8 tin độc hại bị gán thành tin bài thường) chiếm tỉ lệ nhỏ, cho thấy khả năng phân biệt của mô hình là khá chắc chắn. Trong thực tiễn giám sát thông tin, việc giảm thiểu nhầm lẫn loại II (tin độc hại bị bỏ sót) có ý nghĩa quan trọng, và mô hình đề xuất đã đạt được mức sai sót rất thấp ở khía cạnh này.

Khi so sánh với các mô hình khác, mô hình đề xuất thể hiện ưu thế rõ rệt: Theo Bảng 2.4 hiệu suất của PhoBERT + Swin Transformer V2 (F1: 97,00%) cao hơn hẳn so với các mô hình kết hợp như BERT + Xception (Concatenate) đạt F1-score từ 93,25%, BERT + Xception (Maximum) đạt F1-score 93,29%) [55].

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

Mô hình đề xuất dựa trên sự kết hợp giữa PhoBERT và Swin Transformer V2, đã mang lại kết quả vượt trội trên các độ đo đánh giá trong chế độ phát hiện đa lớp. Cụ thể, mô hình đạt độ chính xác Accuracy (ACC) là 95,67%, độ chính xác Precision (PPV) là 95,71%, độ bao phủ Recall (TPR) là 95,67% và điểm F1-score đạt 95,68%. Kết quả này khẳng định rằng việc tận dụng đồng thời đặc trưng văn bản và hình ảnh giúp mô hình đạt được hiệu suất phát hiện cao nhất trong số các cấu hình thử nghiệm.

Mô hình không chỉ đạt hiệu suất tổng thể cao mà còn duy trì tỉ lệ phát hiện (DR) rất ổn định và đồng đều giữa các nhóm nhãn khác nhau. Nhãn "Tin bài thường" đạt tỉ lệ phát hiện cao nhất với 97,78%, điều này chứng tỏ mô hình có khả năng phân loại chính xác các tin tức chính thống, giúp giảm thiểu tối đa hiện tượng báo động giả trong thực tế giám sát. Các nhãn độc hại có tỉ lệ phát hiện đối với các nhóm nhãn phức tạp đều ở mức ấn tượng, "Vu khống, bôi nhọ" đạt 94,67%, "Kêu gọi chống đối" đạt 97,33% và "Thúc đẩy tư tưởng cực đoan" đạt 96%. Ma trận nhầm lẫn (Bảng 2.7) cho thấy sự nhầm lẫn giữa các nhóm nhãn độc hại là rất thấp,

minh chứng cho khả năng học được các đặc trưng phân biệt riêng biệt của mô hình đối với từng loại hành vi độc hại dù chúng có những điểm tương đồng về mặt biểu đạt.

Hiệu suất ưu việt của mô hình PhoBERT + Swin Transformer V2 so với nghiên cứu trước đó được lý giải bởi hai yếu tố chính: (i) PhoBERT hiệu quả hơn đáng kể so với BERT trong việc xử lý văn bản tiếng Việt, đồng thời Swin Transformer V2 có khả năng xử lý các nội dung văn bản nhúng trong hình ảnh tốt hơn so với Xception, giúp sự kết hợp này đặc biệt phù hợp cho bài toán hiện tại; và (ii) việc kết hợp PhoBERT với Swin Transformer V2 cho phép mô hình khai thác các thế mạnh bổ trợ lẫn nhau từ cả hai kiến trúc, từ đó trích xuất toàn diện hơn các đặc trưng phân biệt từ cả văn bản và hình ảnh, giúp nâng cao khả năng phân loại giữa tin bài độc hại và tin bài thường.

Ngoài những ưu điểm nêu trên, mô hình đề xuất có một số hạn chế, gồm: (i) Chưa khai thác triệt để tương tác liên phương thức, do sử dụng kỹ thuật ghép nối các vector đặc trưng ở giai đoạn đầu, mô hình có thể chưa học được những mối quan hệ tương tác phức tạp và tinh vi giữa nội dung ngôn ngữ và tín hiệu thị giác. Điều này đặc biệt ảnh hưởng đến khả năng nhận diện các nội dung có tính chất ngụy trang đa tầng. (ii) Tính linh hoạt của kiến trúc: Mặc dù phương pháp kết hợp sớm có chi phí tính toán thấp hơn, nhưng nó thiếu sự linh hoạt trong việc bảo toàn các thông tin chuyên biệt sâu của từng phương thức đơn lẻ so với các kiến trúc phức tạp hơn.

2.4. KẾT LUẬN CHƯƠNG

Chương này đã trình bày mô hình phát hiện tin bài độc hại tiếng Việt dựa trên kết hợp đặc trưng văn bản và hình ảnh. Mô hình được thiết kế nhằm khắc phục những hạn chế của các phương pháp đơn phương thức, đồng thời phản ánh đúng hơn đặc trưng đa phương thức của tin bài trên không gian mạng.

Kết quả thử nghiệm cho thấy mô hình đề xuất đạt hiệu quả cao và ổn định hơn so với các mô hình riêng lẻ và một số nghiên cứu liên quan. Trong kịch bản đa lớp, mô hình cũng chứng minh khả năng phân loại chính xác ở mức cao (F1-score trên 95%), đồng thời cho thấy ưu thế rõ rệt trong việc xử lý các trường hợp tin bài phức hợp.

Đóng góp của luận án không nằm ở việc đề xuất mới riêng lẻ các mô hình PhoBERT, Swin Transformer V2 hay phép kết hợp vector, mà ở việc xây dựng và đánh giá một khung phát hiện tin bài độc hại tiếng Việt đa lớp dựa trên kết hợp văn bản–hình ảnh; đồng thời mở rộng khung này sang điều kiện thiếu dữ liệu gán nhãn bằng học bán giám sát.

Các kết quả ở chương này đã được đăng trên Tạp chí Khoa học Công nghệ thông tin và Truyền thông (2025) [CT1], được đăng trên Tạp chí quốc tế The Indonesian Journal of Electrical Engineering and Computer Science (2025, Scopus Q4) [CT2], được đăng ở Hội thảo quốc gia FAIR 2025 [CT4].

Trên cơ sở những kết quả đạt được ở chương 2, chương tiếp theo sẽ tập trung khai thác hướng tiếp cận học bán giám sát, nhằm tiếp tục nâng cao hiệu quả của hệ thống phát hiện tin bài độc hại tiếng Việt trong điều kiện hạn chế về dữ liệu đã được gán nhãn.

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

Chương 3 kế thừa mô hình đa phương thức được xây dựng ở Chương 2. Trong đó, PhoBERT và Swin Transformer V2 tiếp tục được sử dụng để trích xuất đặc trưng văn bản và hình ảnh. Chương này tập trung xây dựng, thử nghiệm và đánh giá mô hình đa phương thức có bổ sung cơ chế học bán giám sát và pseudo-labeling nhằm giảm phụ thuộc vào dữ liệu có nhãn cho phát hiện tin bài độc hại tiếng Việt. Nội dung gồm: (i) đặt vấn đề và chỉ ra những hạn chế của các phương pháp học có giám sát trong bối cảnh dữ liệu tin bài tiếng Việt được gán nhãn còn hạn chế, (ii) giới thiệu chi tiết kiến trúc mô hình bán giám sát cho bài toán phân loại đa lớp, (iii) mô tả quy trình tiền xử lý dữ liệu, cơ chế gán nhãn giả và huấn luyện, (iv) trình bày kết quả thử nghiệm, so sánh và đánh giá hiệu quả của mô hình. Trên cơ sở đó, chương 3 khẳng định tính khả thi của việc áp dụng học bán giám sát, đồng thời cho thấy tiềm năng mở rộng và ứng dụng thực tiễn của mô hình phát hiện tin bài độc hại tiếng Việt trong điều kiện dữ liệu gán nhãn còn khan hiếm.

3.1. ĐẶT VẤN ĐỀ

Trong nghiên cứu phát hiện tin bài độc hại tiếng Việt, một trong những thách thức lớn nhất xuất phát từ sự khan hiếm và thiếu ổn định của dữ liệu gán nhãn. Khác với dữ liệu tin tức chính thống vốn được lưu trữ lâu dài và có cấu trúc nhất quán, tin bài độc hại đặc biệt là nhóm có nội dung độc hại chính trị thường tồn tại trong thời gian ngắn và có xu hướng bị gỡ bỏ nhanh chóng bởi các hệ thống kiểm duyệt trên mạng xã hội. Những đặc điểm này khiến việc thu thập một lượng dữ liệu đủ lớn, cân bằng và đại diện cho nhiều dạng thức độc hại trở nên khó khăn. Bên cạnh đó, tin độc hại thường được thể hiện với ngôn từ biến hóa, giật gân, nội dung xuyên tạc hoặc ngụy tạo, khiến quá trình nhận diện thủ công trở nên phức tạp.

Thêm vào đó, việc gán nhãn cho tin bài độc hại là một quá trình đòi hỏi chuyên môn cao, đặc biệt với các nội dung mang tính chính trị nhạy cảm, phải phân biệt được ranh giới giữa phản biện xã hội thông thường và thông tin có tính chất vu khống, kêu gọi chống đối hoặc thúc đẩy tư tưởng cực đoan. Do vậy, việc gán nhãn

hoàn toàn thủ công vừa tốn kém thời gian, vừa khó mở rộng trên quy mô lớn. Kết quả là tập dữ liệu có nhãn thu được thường nhỏ, mất cân bằng giữa các nhóm nhãn và không phản ánh đầy đủ sự đa dạng của tin bài độc hại trong thực tế.

Ngược lại, dữ liệu chưa gán nhãn trên không gian mạng lại xuất hiện với số lượng rất lớn và liên tục tăng theo thời gian. Các bài đăng, bình luận, hình ảnh và nội dung đa phương thức được tạo ra mỗi ngày chứa nhiều thông tin tiềm ẩn yếu tố độc hại nhưng chưa được phân loại. Đây là nguồn dữ liệu phong phú và mang tính đại diện cao cho môi trường thông tin thực tiễn. Tuy nhiên, các phương pháp học có giám sát truyền thống không thể khai thác hiệu quả nguồn dữ liệu này, dẫn tới việc lãng phí một lượng lớn thông tin hữu ích.

Trong bối cảnh đó, học bán giám sát là một hướng tiếp cận phù hợp để khai thác đồng thời dữ liệu có nhãn và dữ liệu chưa gán nhãn. Khác với học có giám sát vốn chỉ dựa vào dữ liệu đã được gán nhãn, học bán giám sát cho phép mô hình tận dụng đồng thời dữ liệu có nhãn và dữ liệu chưa có nhãn để cải thiện hiệu quả huấn luyện. Cơ chế gán nhãn giả (pseudo-labeling) đóng vai trò quan trọng trong quy trình này: mô hình ban đầu được huấn luyện trên tập dữ liệu nhỏ có nhãn, sau đó được sử dụng để gán nhãn tự động cho dữ liệu chưa gán nhãn; phần dữ liệu mở rộng này tiếp tục được đưa vào quá trình huấn luyện, giúp mô hình thích ứng dần với dữ liệu thực tế và nâng cao khả năng khái quát hóa.

Cách tiếp cận bán giám sát không chỉ khắc phục hạn chế về quy mô dữ liệu có nhãn, mà còn giúp mô hình thích ứng tốt hơn trước sự thay đổi liên tục của các dạng tin độc hại mới. Việc kết hợp thêm dữ liệu chưa gán nhãn làm tăng độ đa dạng của tập huấn luyện, giảm nguy cơ mô hình bị thiên lệch vào các mẫu phổ biến và tăng độ bền vững trong môi trường thực tế. Đồng thời, phương pháp này giúp giảm đáng kể chi phí và công sức cho quá trình gán nhãn thủ công, từ đó phù hợp hơn với bối cảnh nghiên cứu tiếng Việt, nơi nguồn lực gán nhãn chuyên sâu còn hạn chế.

Từ những phân tích nêu trên, chương 3 của luận án tập trung nghiên cứu và phát triển mô hình phát hiện tin bài độc hại tiếng Việt dựa trên học bán giám sát kết hợp đặc trưng văn bản và hình ảnh. Mục tiêu của chương nhằm tận dụng tối đa nguồn dữ liệu chưa nhãn từ không gian mạng, khắc phục vấn đề thiếu hụt dữ liệu gán nhãn và nâng cao khả năng phát hiện các dạng nội dung độc hại trong điều kiện

thực tiễn. Đây không chỉ là hướng tiếp cận có ý nghĩa khoa học đối với lĩnh vực xử lý ngôn ngữ tự nhiên và phân tích nội dung đa phương thức tiếng Việt, mà còn mang giá trị tham khảo trong việc xây dựng các hệ thống hỗ trợ giám sát và phát hiện nội dung độc hại trên môi trường mạng.

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

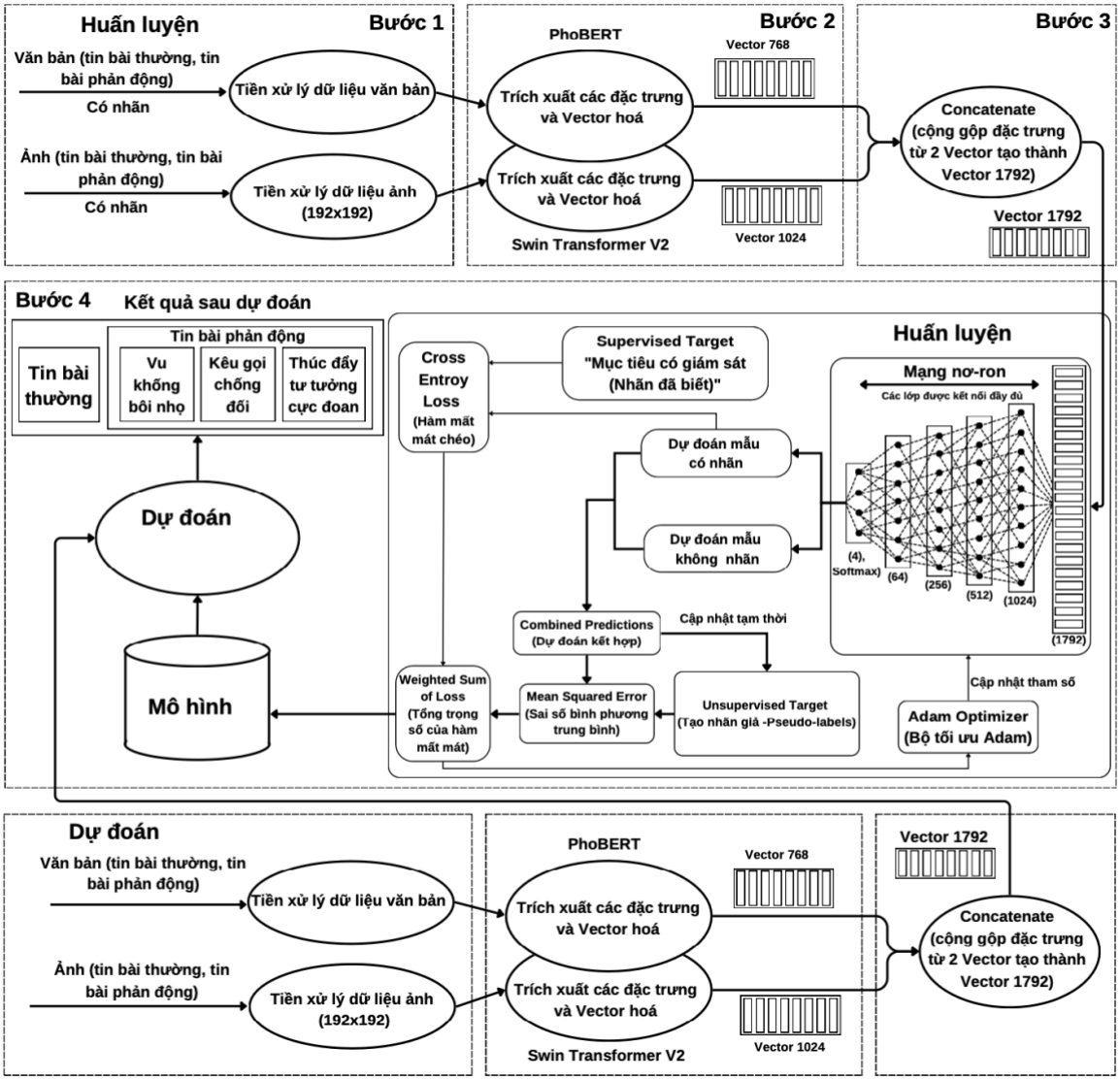
Trong bối cảnh tin bài độc hại thường đồng thời mang nhiều đặc điểm như xuyên tạc, kêu gọi chống đối hoặc thúc đẩy tư tưởng cực đoan, việc áp dụng phân loại đa lớp là cần thiết nhằm phản ánh đầy đủ tính chất phức tạp của dữ liệu. Mô hình kết hợp dựa trên học bán giám sát trong luận án được thiết kế nhằm tận dụng đồng thời thông tin ngữ nghĩa từ văn bản và đặc trưng trực quan từ hình ảnh, đồng thời khai thác hiệu quả lượng dữ liệu chưa gán nhãn thông qua cơ chế gán nhãn giả, như mô tả trên Hình 3.1.

Ở bước đầu, đặc trưng ngữ nghĩa được trích xuất từ văn bản bằng PhoBERT, trong khi đặc trưng hình ảnh được học thông qua Swin Transformer V2. Các đặc trưng này sau đó được chuẩn hóa và cộng gộp, hình thành một vector đặc trưng hợp nhất. Vector này được đưa vào mạng nơ-ron giảm chiều nhiều lớp tự thiết kế, đóng vai trò như tầng phân loại chính.

Trong mô hình dựa trên bán giám sát, đầu ra được xây dựng dựa trên hàm kích hoạt softmax với số chiều tương ứng với số lớp. Điều này cho phép hệ thống phân bổ xác suất cho từng loại độc hại, qua đó xác định lớp có xác suất cao nhất tương ứng với mỗi tin bài. Hàm mất mát sử dụng trong huấn luyện là cross-entropy, phù hợp cho bài toán đa lớp và giúp tối ưu hóa khả năng phân biệt giữa các lớp khác nhau.

Để vượt qua thách thức thiếu dữ liệu gán nhãn, luận án tích hợp kỹ thuật gán nhãn giả trong quá trình huấn luyện. Cụ thể, mô hình ban đầu được huấn luyện trên tập dữ liệu gán nhãn, sau đó sử dụng để gán nhãn giả cho dữ liệu chưa gán nhãn.

Các mẫu dữ liệu này, sau khi được lựa chọn dựa trên ngưỡng tin cậy, được bổ sung vào tập huấn luyện nhằm tái huấn luyện mô hình. Quá trình lặp lại này giúp hệ thống khai thác được khối lượng lớn dữ liệu chưa gán nhãn, đồng thời mở rộng lượng dữ liệu được sử dụng trong quá trình huấn luyện, qua đó duy trì hiệu năng của mô hình trong điều kiện tỷ lệ dữ liệu có nhãn bị hạn chế.



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

Nhờ cấu trúc này, mô hình không chỉ phát hiện chính xác các tin bài độc hại chứa nhiều yếu tố khác nhau mà còn cho thấy khả năng duy trì hiệu năng trong các kịch bản có tỷ lệ dữ liệu gán nhãn khác nhau. Đây là điểm mới quan trọng của luận án, góp phần khẳng định hiệu quả của việc kết hợp học bán giám sát và đa phương thức trong xử lý tin bài độc hại tiếng Việt.

3.2.2. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng nhằm chuẩn bị đầu vào phù hợp cho mô hình đa phương thức kết hợp học bán giám sát. Quy trình này được thực hiện cho cả hai dạng dữ liệu văn bản và hình ảnh. Quá trình tiền xử lý dữ liệu được thực hiện theo quy trình đã mô tả ở mục 2.2.2.

Ngoài ra trong học bán giám sát, dữ liệu chưa gán nhãn được mô hình gán nhãn giả. Các mẫu có xác suất dự đoán vượt ngưỡng tin cậy sẽ được chọn lọc và xử lý tương tự dữ liệu đã gán nhãn, đảm bảo tính thống nhất của đầu vào. Cách tiếp cận này cho phép mở rộng tập huấn luyện qua nhiều vòng lặp gán nhãn giả, giúp cải thiện khả năng khái quát hóa của mô hình.

Như vậy, bước tiền xử lý dữ liệu vừa đảm bảo tính đồng nhất, vừa tăng cường khả năng học của mô hình, đóng vai trò quan trọng trong việc nâng cao hiệu quả phát hiện tin bài độc hại trong bối cảnh học bán giám sát.

3.2.3. Huấn luyện và phát hiện

Giải thuật 3.1 mô hình hoá quá trình huấn luyện biểu diễn trên Hình 3.1.

Giải thuật 3.1: Quá trình huấn luyện

Input:

x_{i_t}: Văn bản đầu vào (training sample).

x_{i_i}: Ảnh đầu vào (training sample).

y_i: Nhãn lớp của mẫu xi, yi ∈ {0,1,2,3}.

S: Tập hợp các mẫu huấn luyện có nhãn.

Output:

$\boldsymbol{\theta}$ *:* Tham số của mô hình sau khi huấn luyện.

Process:

α: Hệ số điều chỉnh mức độ ảnh hưởng của các dự đoán trước đó, $0 \leq \alpha \leq 1$.

w(t): Tăng trọng số không giám sát.

PhoBERT(x): Trích xuất đặc trưng văn bản.

SwinTransformerV2(x): Trích xuất đặc trưng ảnh.

$\boldsymbol{f_\theta(x')}$*:* Mạng nơ-ron FFNN (lớp cuối softmax) với tham số có thể huấn luyện θ.

$Z \leftarrow 0_{[N]}$ ⊳ Khởi tạo dự đoán tổng hợp (ensemble predictions) (N: số lượng mẫu).

$z \leftarrow 0_{[N]}$ ⊳ Khởi tạo vector mục tiêu không giám sát (unsupervised target vectors).

for t in [1, num epochs] do

  for each minibatch B do

B1: Tiền xử lý: (Các tin bài được tiền xử lý trước khi đưa vào PhoBERT và Swin Transformer V2, như mô tả tại Mục 3.2.2)

B2: Trích xuất đặc trưng

$\mathbf{x'}_{t\in B} \leftarrow \text{PhoBERT}(\mathbf{x}_{t\in B})$ ▷ Trích xuất đặc trưng văn bản

$\mathbf{x'}_{i\in B} \leftarrow \text{SwinTransformerV2}(\mathbf{x}_{i\in B})$ ▷ Trích xuất đặc trưng ảnh

B3: Kết hợp đặc trưng

$\mathbf{x'} \in B \leftarrow \text{Concatenate}(\mathbf{x'}_t, \mathbf{x'}_i)$ ▷ Kết hợp đặc trưng văn bản và ảnh

B4: Huấn luyện mô hình

$z_{i\in B} \leftarrow f_\theta(\mathbf{x'})$Đầu ra của mạng nơ-ron (sau softmax để tạo ra xác suất dự đoán)

$L_s^{i\in B} \leftarrow \frac{1}{|B|}\sum_{i\in(B\cap S)}[y_i \log(z_i) + (1-y_i)\log(1-z_i)]$ ▷ Thành phần mất mát giám sát (Multiclass Cross-Entropy)

$\tilde{z}_{i\in B} \leftarrow \text{Mean}(z_{i\in B})$ ▷ Tính trung bình các dự đoán trong minibatch

$L_u^{i\in B} \leftarrow \frac{C_1}{|B|}\sum_{i\in B}(z_i - \tilde{z}_i)^2$ ▷ Thành phần mất mát không giám sát (Mean Squared Error)

Loss $\leftarrow L_s^{i\in B} + w(t)L_u^{i\in B}$ ▷ Tổng mất mát (có trọng số cho mất mát không giám sát)

Cập nhật $\boldsymbol{\theta}$ ▷ Tối ưu hóa các tham số của mạng nơ-ron

end for (kết thúc vòng lặp minibatch)

$Z \leftarrow \alpha Z + (1\ \alpha)z$ ▷ Cập nhật dự đoán tổng hợp (Temporal Update)

$z \leftarrow \frac{Z}{1-\alpha^t}$ ▷ Hiệu chỉnh sai lệch cho vector mục tiêu không giám sát

end for (kết thúc vòng lặp epoch)

return $\boldsymbol{\theta}$

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).

$$L_u^i = \left(\frac{C_1}{|B|}\right) \sum_{i \in B} (z_i - \tilde{z}_i)^2 \quad (3.1)$$

Trong đó:

- L_u^i là mất mát không giám sát.
- C_1 là siêu tham số điều chỉnh mức độ ảnh hưởng của L^u_i.
- $|B|$ là kích thước minibatch.
- z_i là xác suất dự đoán của mô hình cho mẫu i (sau softmax).
- $\tilde{z}_i$ là nhãn giả gán cho mẫu i.
- $(z_i - \tilde{z}_i)^2$ là bình phương sai số giữa dự đoán và nhãn giả.

Thành phần mất mát không giám sát sử dụng Mean Squared Error (MSE) để đo lường sự khác biệt giữa xác suất dự đoán (z_i) của mô hình và nhãn giả ($\tilde{z}_i$) gán cho dữ liệu chưa được gán nhãn. Công thức tính mất mát không giám sát (L_u^i) là $L_u^i = \left(\frac{C_1}{|B|}\right) \sum_{i \in B} (z_i - \tilde{z}_i)^2$, trong đó $|B|$ là kích thước của minibatch, và C_1 là một siêu tham số được sử dụng để điều chỉnh mức độ ảnh hưởng của thành phần mất mát không giám sát này đến quá trình tối ưu hóa tổng thể của mô hình.

Việc kết hợp cả mất mát giám sát và không giám sát giúp mô hình tận dụng được tối đa thông tin từ cả dữ liệu đã gán nhãn và dữ liệu chưa gán nhãn, từ đó cải thiện khả năng phân loại. Hàm mất mát tổng thể như sau:

$$\text{Loss} = L_s^i + w(t) L_u^i \quad (3.2)$$

Để tối ưu hóa mô hình chúng tôi tính toán một hàm mất mát tổng thể bằng cách kết hợp mất mát giám sát (L_s^i) và mất mát không giám sát (L_u^i) thông qua tổng trọng số: $\text{Loss} = L_s^i + w(t) L_u^i$. Trong công thức này, $w(t)$ là một hàm tăng trọng số không giám sát, cho phép chúng tôi điều chỉnh tầm quan trọng của thành phần mất mát không giám sát trong quá trình huấn luyện, thường là tăng dần theo thời gian (epoch).

Điều này giúp mô hình ban đầu tập trung vào dữ liệu đã được gán nhãn chính xác, và sau đó dần dần khai thác thông tin từ dữ liệu đã được gán nhãn giả.

$$Z \leftarrow \alpha Z + (1\ \alpha)z \tag{3.3}$$

Công thức (3.3) thực hiện việc cập nhật vector dự đoán tổng hợp Z bằng cách kết hợp các dự đoán trước đó (được lưu trữ trong Z) với các dự đoán mới nhất z từ epoch hiện tại. Tham số α (hệ số điều chỉnh mức độ ảnh hưởng của các dự đoán trước đó) điều chỉnh mức độ ảnh hưởng của các dự đoán trước đó

Trong quá trình huấn luyện bán giám sát, một thành phần quan trọng là vector mục tiêu không giám sát z, được sử dụng để tính toán hàm mất mát giữa các dự đoán hiện tại và nhãn giả. Tuy nhiên, ở giai đoạn đầu khi mô hình chưa được huấn luyện đầy đủ, các dự đoán tổng hợp tích lũy trong vector Z có thể chứa nhiều sai lệch do tính không ổn định của các dự đoán ban đầu.

Để khắc phục điều này, mô hình áp dụng bước hiệu chỉnh sai lệch bằng công thức:

$$z \leftarrow \frac{Z}{1-\alpha^t} \tag{3.4}$$

Công thức (3.4) giúp hiệu chỉnh sai lệch vector Z bằng cách chia cho hệ số suy giảm theo thời gian $(1 - \alpha^t)$, từ đó làm giảm ảnh hưởng của các sai số tích lũy trong giai đoạn sớm của huấn luyện. Nhờ đó, các nhãn giả được sinh ra từ z trở nên ổn định hơn và phản ánh chính xác hơn kỳ vọng của mô hình ở thời điểm hiện tại.

Việc hiệu chỉnh này không chỉ tăng độ tin cậy cho nhãn giả, mà còn góp phần quan trọng trong việc ổn định hàm mất mát không giám sát, đảm bảo mô hình học được các đặc trưng hữu ích từ dữ liệu chưa gán nhãn mà không bị chi phối bởi nhiễu ban đầu. Đây là một bước thiết yếu để nâng cao chất lượng huấn luyện và cải thiện khả năng tổng quát hóa của mô hình.

Các thuật toán tối ưu hóa như Adam được sử dụng lặp đi lặp lại trên các minibatches khác nhau, nhằm tìm ra bộ tham số tối ưu giúp mô hình đạt được độ chính xác cao nhất trên dữ liệu mới.

Sau khi hoàn tất quá trình huấn luyện tập hợp các tham số $\boldsymbol{\theta}$, mô hình được sử dụng để dự đoán các nhãn độc hại có thể đồng thời xuất hiện trong một tin bài đầu vào. Trong quá trình dự đoán tuân theo quy trình tương tự như đã mô tả tại mục 2.2.3.

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 18.000 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 4.500 bài viết dưới dạng văn bản và 4.500 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 1.500 tin bài dạng văn bản, 1.500 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: 4.500 bài viết dưới dạng văn bản và 4.500 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

Tập dữ liệu thu thập được chia ngẫu nhiên với 70% dữ liệu cho huấn luyện, 15% cho đánh giá và 15% dữ liệu cho kiểm tra, được sử dụng để huấn luyện và đánh giá hiệu suất phát hiện của mô hình kết hợp đề xuất dựa trên PhoBERT và Swin Transformer V2 và các mô hình kết hợp khác.

Cụ thể lần lượt sử dụng các mô hình kết hợp BERT và Swin Transformer V2, RoBERTa và Swin Transformer V2, BERT và VGG19, PhoBERT và VGG19,

RoBERTa và VGG19, BERT và Xception, RoBERTa và Xception, PhoBERT và Xception, PhoBERT và Swin Transformer V2 để xử lý tập dữ liệu gồm 9.000 bài viết dưới dạng văn bản và 9.000 bài dưới dạng ảnh (tập dữ liệu gồm 4.500 tin bài thường và 4.500 tin bài độc hại thuộc 3 nhóm nhãn dưới dạng văn bản, và 4.500 tin bài thường và 4.500 tin bài độc hại thuộc 3 nhóm nhãn dưới dạng ảnh). Trong đó, dữ liệu huấn luyện được thử nghiệm 6 lần độc lập bằng cách thay đổi tỷ lệ phần trăm dữ liệu được gán nhãn bắt đầu từ 5%, 10%, 15%, 20%, 30% và 50%.

Bảng từ 3.1 cung cấp hiệu suất phát hiện của mô hình đề xuất theo các ngưỡng tin cậy trong học bán giám sát với 50% dữ liệu huấn luyện có gán nhãn. Các Bảng từ 3.2-3.7 cung cấp hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng lần lượt 5%, 10%, 15%, 20%, 30% và 50% dữ liệu huấn luyện có gán nhãn. Có thể thấy với 5% dữ liệu huấn luyện có gán nhãn, mô hình đề xuất đã cho các độ đo Accuracy và F1 đều trên 85%. Khi tăng tỷ lệ dữ liệu huấn luyện có gán nhãn, hiệu suất phát hiện tăng. Và ổn định ở khoảng 30% dữ liệu huấn luyện có gán nhãn.

Bảng 3.8 và Hình 3.2 cung cấp hiệu suất phát hiện tổng thể của mô hình bán giám sát đề xuất dựa trên PhoBERT + Swin Transformer V2 theo tỷ lệ dữ liệu huấn luyện có nhãn. Thông tin từ Bảng 3.8 và Hình 3.2 khẳng định khi tỷ lệ dữ liệu có nhãn giảm từ 50% xuống 30%, mô hình vẫn duy trì F1 trên 93%; ở mức 20%, F1 đạt 92,18%; và ngay cả với 5% dữ liệu có nhãn, F1 vẫn đạt 85,78%. Kết quả cho thấy mô hình có khả năng khai thác dữ liệu chưa gán nhãn và giảm sự phụ thuộc vào dữ liệu được gán nhãn thủ công.

Bảng 3.1. Hiệu suất phát hiện của mô hình đề xuất theo các ngưỡng tin cậy trong học bán giám sát với 50% dữ liệu huấn luyện có gán nhãn

Mô hình	Ngưỡng tin cậy	Accuracy %	Precision %	Recall %	F1-score %
PhoBERT + Swin Transformer V2	0,8	91,04	91,56	91,04	91,18
PhoBERT + Swin Transformer V2	0,7	90,59	90,85	90,59	90,71
PhoBERT + Swin	**0.6**	**93,48**	**93,43**	**93,48**	**93,45**

Transformer V2					
PhoBERT + Swin Transformer V2	0,5	91,78	91,70	91,78	91,73
PhoBERT + Swin Transformer V2	0,4	90,89	91,24	90,98	91,11

Bảng 3.2. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 5% dữ liệu huấn luyện có gán nhãn

Mô hình	Đặc trưng	Accuracy %	Precision %	Recall %	F1-score %
BERT + Swin Transformer V2	Văn bản + Ảnh	82,52	82,40	82,52	82,46
BERT + VGG19	Văn bản + Ảnh	79,26	78,62	79,26	78,94
BERT + Xception	Văn bản + Ảnh	76,67	77,47	76,67	77,07
RoBERTa + Swin Transformer V2	Văn bản + Ảnh	60,07	51,59	60,07	55,51
RoBERTa + Xception	Văn bản + Ảnh	64,74	61,62	64,74	63,14
RoBERTa + VGG19	Văn bản + Ảnh	73,85	71,95	73,85	72,89
PhoBERT + VGG19	Văn bản + Ảnh	74,15	79,95	74,15	76,94
PhoBERT + Xception	Văn bản + Ảnh	81,78	82,22	81,78	82,00
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**85,56**	**86,01**	**85,56**	**85,78**

Bảng 3.3. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 10% dữ liệu huấn luyện có gán nhãn

Mô hình	Đặc trưng	Accuracy %	Precision %	Recall %	F1-score %
BERT + Swin Transformer V2	Văn bản + Ảnh	85,85	85,58	85,85	85,71
BERT + VGG19	Văn bản + Ảnh	83,19	84,09	83,19	83,64
BERT + Xception	Văn bản + Ảnh	78,67	79,02	78,67	78,84
RoBERTa + Swin Transformer V2	Văn bản + Ảnh	66,22	62,67	63,94	63,30

RoBERTa + Xception	Văn bản + Ảnh	66,81	64,60	64,70	64,65
RoBERTa + VGG19	Văn bản + Ảnh	77,11	77,02	76,13	76,57
PhoBERT + VGG19	Văn bản + Ảnh	80,22	80,02	80,22	80,12
PhoBERT + Xception	Văn bản + Ảnh	86,74	86,66	86,74	86,66
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**90,37**	**90,24**	**90,37**	**90,30**

Bảng 3.4. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 15% dữ liệu huấn luyện có gán nhãn

Mô hình	Đặc trưng	Accuracy %	Precision %	Recall %	F1-score %
BERT + Swin Transformer V2	Văn bản + Ảnh	87,11	87,05	87,11	87,08
BERT + VGG19	Văn bản + Ảnh	85,48	85,16	85,48	85,32
BERT + Xception	Văn bản + Ảnh	82,07	82,36	82,07	82,21
RoBERTa + Swin Transformer V2	Văn bản + Ảnh	69,56	67,90	69,56	68,72
RoBERTa + Xception	Văn bản + Ảnh	69,93	65,40	69,93	67,59
RoBERTa + VGG19	Văn bản + Ảnh	78,07	78,05	78,07	78,06
PhoBERT + VGG19	Văn bản + Ảnh	84,67	84,55	84,67	84,61
PhoBERT + Xception	Văn bản + Ảnh	86,96	87,73	86,96	87,34
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**91,63**	**91,59**	**91,63**	**91,61**

Bảng 3.5. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 20% dữ liệu huấn luyện có gán nhãn

Mô hình	Đặc trưng	Accuracy %	Precision %	Recall %	F1-score %
BERT + Swin Transformer V2	Văn bản + Ảnh	89,41	89,42	89,41	89,41
BERT + VGG19	Văn bản + Ảnh	86,07	86,38	86,07	86,22
BERT + Xception	Văn bản + Ảnh	82,89	82,58	82,89	82,73

RoBERTa + Swin Transformer V2	Văn bản + Ảnh	71,70	70,76	70,90	70,83
RoBERTa + Xception	Văn bản + Ảnh	69,41	69,56	69,10	69,33
RoBERTa + VGG19	Văn bản + Ảnh	80,07	79,75	79,71	79,73
PhoBERT + VGG19	Văn bản + Ảnh	86,89	86,54	86,89	86,71
PhoBERT + Xception	Văn bản + Ảnh	89,41	89,46	89,41	89,43
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**92,15**	**92,22**	**92,15**	**92,18**

Bảng 3.6. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 30% dữ liệu huấn luyện có gán nhãn

Mô hình	Đặc trưng	Accuracy %	Precision %	Recall %	F1-score %
BERT + Swin Transformer V2	Văn bản + Ảnh	91,78	91,87	91,78	91,82
BERT + VGG19	Văn bản + Ảnh	87,93	88,31	87,93	88,12
BERT + Xception	Văn bản + Ảnh	85,26	85,31	85,26	85,28
RoBERTa + Swin Transformer V2	Văn bản + Ảnh	69,19	68,91	69,19	69,05
RoBERTa + Xception	Văn bản + Ảnh	72,67	71,34	72,67	72,00
RoBERTa + VGG19	Văn bản + Ảnh	82,81	81,48	82,81	82,14
PhoBERT + VGG19	Văn bản + Ảnh	88,52	88,41	88,52	88,46
PhoBERT + Xception	Văn bản + Ảnh	89,63	90,83	89,63	90,23
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**93,19**	**93,30**	**93,19**	**93,14**

Bảng 3.7. Hiệu suất phát hiện của mô hình đề xuất và các mô hình kết hợp khác dựa trên học bán giám sát sử dụng 50% dữ liệu huấn luyện có gán nhãn

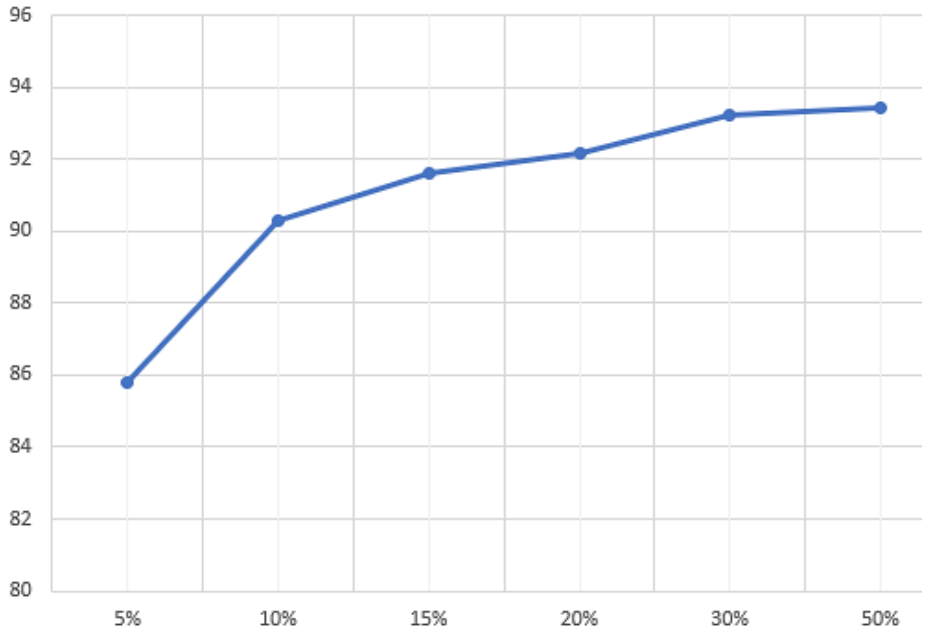
Mô hình	Đặc trưng	Accuracy %	Precision %	Recall %	F1-score %
BERT + Swin Transformer V2	Văn bản + Ảnh	92,37	92,44	92,37	92,40

BERT + Xception	Văn bản + Ảnh	87,19	87,52	87,19	87,35
BERT + VGG19	Văn bản + Ảnh	89,93	89,70	89,93	89,81
RoBERTa + Swin Transformer V2	Văn bản + Ảnh	74,52	74,14	74,12	74,13
RoBERTa + Xception	Văn bản + Ảnh	76,30	75,58	75,52	75,55
RoBERTa + VGG19	Văn bản + Ảnh	83,19	83,39	82,97	83,18
PhoBERT + VGG19	Văn bản + Ảnh	88,37	88,55	88,37	88,46
PhoBERT + Xception	Văn bản + Ảnh	93,41	93,36	93,41	93,38
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**93,48**	**93,43**	**93,48**	**93,45**

Bảng 3.8. Hiệu suất phát hiện tổng thể của mô hình bán giám sát đề xuất dựa trên PhoBERT + Swin Transformer V2 theo tỷ lệ dữ liệu huấn luyện có nhãn

Tỷ lệ dữ liệu gán nhãn	Accuracy %	Precision %	Recall %	F1-score %
5%	85,56	86,01	85,56	85,78
10%	90,37	90,24	90,37	90,30
15%	91,63	91,59	91,63	91,61
20%	92,15	92,22	92,15	92,18
30%	**93,19**	**93,30**	**93,19**	**93,24**
50%	93,48	93,43	93,48	93,45

Các Bảng từ 3.9-3.14 cung cấp ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng lần lượt 5%, 10%, 15%, 20%, 30%, 50% dữ liệu huấn luyện có gán nhãn. Ngoài ra, Bảng 3.15 cung cấp số lượng mẫu và tỷ lệ phát hiện đúng (DR) cho từng nhãn tin bài của mô hình bán giám sát đề xuất theo tỷ lệ dữ liệu huấn luyện có nhãn. Quy ước: Lớp 0 – tin bài thường, lớp 1 - tin độc hại, lớp 2 - tin độc hại 2, lớp 3 - tin độc hại 3.



Hình 3.2. Độ đo F1 của mô hình đề xuất theo tỷ lệ dữ liệu huấn luyện có nhãn

Bảng 3.9. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 5% dữ liệu huấn luyện có gán nhãn

Lớp	Tin bài thường (675 tin)	Tin độc hại 1 (225 tin)	Tin độc hại 2 (225 tin)	Tin độc hại 3 (225 tin)
0	636	21	11	7
1	50	133	28	14
2	8	9	204	4
3	28	10	56	131

Bảng 3.10. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 10% dữ liệu huấn luyện có gán nhãn

Lớp	Tin bài thường (675 tin)	Tin độc hại 1 (225 tin)	Tin độc hại 2 (225 tin)	Tin độc hại 3 (225 tin)
0	641	25	2	7
1	42	161	9	13
2	4	18	188	15
3	15	20	9	181

Bảng 3.11. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 15% dữ liệu huấn luyện có gán nhãn

Lớp	Tin bài thường (675 tin)	Tin độc hại 1 (225 tin)	Tin độc hại 2 (225 tin)	Tin độc hại 3 (225 tin)
0	651	18	2	4
1	45	173	3	4
2	5	37	174	9
3	15	31	3	176

Bảng 3.12. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 20% dữ liệu huấn luyện có gán nhãn

Lớp	Tin bài thường (675 tin)	Tin độc hại 1 (225 tin)	Tin độc hại 2 (225 tin)	Tin độc hại 3 (225 tin)
0	662	6	1	6
1	58	153	3	11
2	8	10	192	15
3	14	8	3	200

Bảng 3.13. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 30% dữ liệu huấn luyện có gán nhãn

Lớp	Tin bài thường (675 tin)	Tin độc hại 1 (225 tin)	Tin độc hại 2 (225 tin)	Tin độc hại 3 (225 tin)
0	634	20	1	20
1	21	161	4	39
2	2	5	192	26
3	1	0	1	223

Bảng 3.14. Ma trận nhầm lẫn cho 4 nhóm tin bài với tập dữ liệu đánh giá mô hình bán giám sát đề xuất sử dụng 50% dữ liệu huấn luyện có gán nhãn

Lớp	Tin bài thường (675 tin)	Tin độc hại 1 (225 tin)	Tin độc hại 2 (225 tin)	Tin độc hại 3 (225 tin)
0	655	16	1	3
1	26	188	5	6
2	2	7	212	4
3	9	7	3	206

Bảng 3.15. Số lượng và tỷ lệ phát hiện đúng (DR) cho từng nhãn tin bài của mô hình bán giám sát đề xuất theo tỷ lệ dữ liệu huấn luyện có nhãn

Tỷ lệ dữ liệu gán nhãn	Các nhãn tin bài							
	Tin bài thường		Vu khống bôi nhọ		Kêu gọi chống đối		Thúc đẩy tư tưởng cực đoan	
	Số lượng	DR (%)	Số lượng	DR (%)	Số lượng	DR (%)	Số lượng	DR (%)
5%	636	94,22	133	59,11	204	90,67	131	58,22
10%	641	94,96	161	71,56	188	83,56	181	80,44
15%	651	96,44	173	76,89	174	77,33	176	78,22
20%	662	98,07	153	68,00	192	85,33	200	88,89
30%	634	93,93	161	71,56	192	85,33	223	99,11
50%	655	97,04	188	83,56	212	94,22	206	91,56
Tổng	675	100	225	100	225	100	225	100

Bảng 3.16 cung cấp hiệu suất phát hiện của mô hình kết hợp đề xuất dựa trên PhoBERT và Swin Transformer V2 sử dụng học bán giám sát với các nghiên cứu liên quan. Có thể nhận thấy rằng phương pháp đề xuất vượt trội hơn hai mô hình trong [64] [65].

Bảng 3.16. Hiệu suất phát hiện của mô hình đề xuất và các mô hình bán giám sát trước đó trên 30% dữ liệu huấn luyện có nhãn

Mô hình	Đặc trưng	Accuracy %	Precision %	Recall %	F1-score %
LOSS-GAT [64]	Văn bản	83.62			83.61
GCN [65]	Văn bản	81.26	75.60	86.59	80.72
PhoBERT + Swin Transformer V2	**Văn bản + Ảnh**	**93,19**	**93,30**	**93,19**	**93,14**

Bảng 3.17. So sánh hiệu suất phát hiện của mô hình đề xuất sử dụng 30% dữ liệu huấn luyện có nhãn với mô hình kết hợp dựa trên học có giám sát

Mô hình kết hợp	Loại mô hình	Accuracy %	Precision %	Recall %	F1-score %
PhoBERT + Swin Transformer V2 [CT2]	Có giám sát	95,00	95,00	95,00	95,00
PhoBERT + Swin Transformer V2 [CT4]	Có giám sát	95,67	95,71	95,67	95,68
PhoBERT + Swin Transformer V2	**Bán giám sát**	**93,19**	**93,30**	**93,19**	**93,24**

3.3.3. Nhận xét và đánh giá

Các kết quả được đối sánh với các cấu hình cơ sở và mô hình kết hợp khác trong cùng điều kiện dữ liệu, qua đó đánh giá riêng ảnh hưởng của lựa chọn mô hình văn bản, mô hình hình ảnh và chiến lược học bán giám sát.

Từ các kết quả thử nghiệm cho trên các Bảng 3.1 – 3.17, có thể rút ra một số nhận xét:

- Tại Bảng 3.1, khi ngưỡng tin cậy tăng, các mẫu dữ liệu giả được lựa chọn có độ chính xác cao hơn, góp phần cải thiện độ ổn định của mô hình. Tuy nhiên, ngưỡng quá cao sẽ làm giảm số lượng mẫu dữ liệu được bổ sung, dẫn đến hạn chế khả năng mở rộng tập huấn luyện. Ngược lại, khi giảm ngưỡng tin cậy, số lượng mẫu dữ liệu giả tăng lên, giúp mô hình khai thác tốt hơn thông tin từ dữ liệu chưa gán nhãn, nhưng đồng thời cũng làm gia tăng nguy cơ nhiễu do các nhãn giả kém

chính xác. Kết quả thực nghiệm cho thấy ngưỡng tin cậy 0,6 đạt được sự cân bằng hợp lý giữa số lượng và chất lượng dữ liệu giả, từ đó giúp mô hình duy trì hiệu suất ổn định và khả năng tổng quát tốt.

- Các mô hình kết hợp sử dụng PhoBERT và VGG19, Xception, Swin Transformer V2 cho hiệu suất phát hiện tốt hơn đáng kể so với các mô hình kết hợp sử dụng BERT và VGG19, Xception, Swin Transformer V2 với tất cả các kịch bản thử nghiệm. Điều này là do PhoBERT được tiền huấn luyện trên tập dữ liệu tiếng Việt rất lớn nên có khả năng xử lý tin bài tiếng Việt hiệu quả hơn so với BERT.

- Các mô hình kết hợp sử dụng Swin Transformer V2 cho hiệu suất phát hiện tốt hơn đáng kể so với các mô hình kết hợp sử dụng VGG19 và Xception với tất cả các kịch bản thử nghiệm. Điều này là do mô hình Swin Transformer V2 có khả năng xử lý dữ liệu văn bản nhúng trong ảnh và các đặc trưng không gian tốt hơn so với mô hình VGG19 và Xception.

- Hầu hết các mô hình kết hợp dựa trên học bán giám sát đều cho hiệu suất phát hiện khá tốt. Chỉ với 10% dữ liệu huấn luyện có nhãn, hầu hết các mô hình kết hợp đều cho độ đo F1 đạt trên 80% (riêng mô hình đề xuất đạt 90,30%) và F1 tăng ổn định khi tỷ lệ dữ liệu huấn luyện có nhãn tăng. Đặc biệt, Bảng 3.8 và Hình 3.2 cho thấy mô hình đề xuất dựa trên học bán giám sát kết hợp với PhoBERT cho xử lý dữ liệu văn bản và Swin Transformer V2 cho xử lý dữ liệu ảnh đạt độ đo F1 trên 92% với 20% dữ liệu huấn luyện có nhãn và duy trì ổn định ở mức trên 93% khi lượng dữ liệu huấn luyện có nhãn được sử dụng đạt 30%.

Bảng 3.14 từ kết quả ma trận nhầm lẫn, có thể thấy lớp “Vu khống, bôi nhọ” (Tin độc hại 1) có số lượng dự đoán đúng là 188/225, thấp hơn so với các lớp khác như Tin độc hại 2 (212/225) và Tin độc hại 3 (206/225). Đồng thời, lớp này có hiện tượng bị nhầm sang lớp Tin bài thường (26 trường hợp) và một phần nhỏ sang các lớp độc hại khác. Điều này cho thấy đặc trưng nhận diện giữa lớp “Vu khống, bôi nhọ” và các nội bình thường trong dữ liệu chưa tách rõ ràng. Nguyên nhân chủ yếu là do mô hình gặp khó khăn trong việc phân biệt ranh giới giữa thông tin sai lệch và ý kiến mang tính cá nhân hoặc chưa được kiểm chứng. Nhiều trường hợp nội dung không chứa từ khóa rõ ràng mà phụ thuộc vào ngữ cảnh hoặc kiến thức ngoài mà mô hình chưa được học, nên khó nhận diện chính xác. Về tác động đến chỉ số đánh

giá, do macro-F1 được tính trung bình trên từng lớp, nên khi một lớp có độ chính xác và khả năng nhận diện thấp hơn, giá trị macro-F1 chung sẽ bị giảm theo, mặc dù các lớp còn lại vẫn đạt kết quả tương đối tốt. Kết quả này cũng phản ánh đúng đặc thù khó của lớp "Vu khống, bôi nhọ" trong dữ liệu thực tế.

Kết quả phát hiện cho từng nhãn cụ thể trên Bảng 3.15 cho thấy, mô hình đề xuất dựa trên học bán giám sát kết hợp với PhoBERT và Swin Transformer V2 có khả năng phát hiện đúng khá cao với hầu hết các nhóm tin bài. Với 30% dữ liệu huấn luyện có nhãn, mô hình đề xuất cho tỷ lệ phát hiện đúng (DR) trên 85% với 3/4 nhóm tin bài. Nhóm tin bài 'Vu khống bôi nhọ' có DR thấp nhất cũng đạt trên 71%.

Bảng 3.16 cung cấp so sánh hiệu suất phát hiện của mô hình đề xuất với nghiên cứu có liên quan cùng với 30% dữ liệu huấn luyện có nhãn. Có thể thấy, mô hình kết hợp đề xuất có hiệu suất phát hiện cao vượt trội so với 2 mô hình trong [64] [65](LOSS-GAT [64] và GCN [65]). Hiệu suất phát hiện của mô hình đề xuất tốt hơn so với các mô hình nghiên cứu liên quan bởi hai lý do sau: (i) mô hình PhoBERT cho xử lý dữ liệu văn bản và Swin Transformer V2 cho xử lý dữ liệu ảnh cung cấp nhiều thông tin hữu ích để phân biệt hiệu quả giữa tin bài thường và các dạng tin bài độc hại, và (ii) việc lựa chọn kết hợp mô hình PhoBERT với Swin Transformer V2 khai thác được điểm mạnh của cả 2 mô hình riêng lẻ, giúp tăng khả năng nhận diện các đặc trưng cả trên văn bản và trên ảnh, nhờ vậy có khả năng phân biệt tốt hơn giữa tin bài thường và tin bài độc hại.

Ngoài ra, khi so sánh hiệu suất phát hiện của mô hình đề xuất sử dụng 30% dữ liệu huấn luyện có nhãn với mô hình kết hợp dựa trên học có giám sát với 100% dữ liệu huấn luyện có nhãn tại [CT2][CT4], cùng sử dụng PhoBERT và Swin Transformer V2. Mặc dù chỉ sử dụng 30% dữ liệu huấn luyện có nhãn, nhưng mô hình đề xuất cho hiệu suất phát hiện tiệm cận với các mô hình đề xuất trong [CT2][CT4].

Ngoài những ưu điểm được nêu thì mô hình đề xuất vẫn còn một số hạn chế như: (i) Nguy cơ lan truyền lỗi: Nếu dữ liệu chưa gán nhãn thu thập từ mạng xã hội chứa quá nhiều nhiễu hoặc thông tin sai lệch, việc tự động gán nhãn giả có thể dẫn đến hiệu ứng lan truyền sai sót, làm ảnh hưởng tiêu cực đến chất lượng huấn luyện

của mô hình ở các vòng lặp sau. (ii) Khó khăn trong cân bằng dữ liệu: Phương pháp bán giám sát đòi hỏi quá trình tinh chỉnh các siêu tham số cực kỳ cẩn thận để đạt được sự cân bằng giữa dữ liệu có nhãn và dữ liệu chưa có nhãn, tránh làm suy giảm hiệu quả tổng thể của hệ thống. (iii) Nhầm lẫn giữa các nhãn tương đồng: Kết quả thực nghiệm cho thấy vẫn tồn tại sai sót khi phân loại giữa các nhóm nhãn có ngữ nghĩa gần nhau.

3.4. KẾT LUẬN CHƯƠNG

Chương 3 trình bày mô hình phát hiện tin bài độc hại tiếng Việt dựa trên học bán giám sát, với sự kết hợp đặc trưng văn bản từ PhoBERT và đặc trưng hình ảnh từ Swin Transformer V2. Trên cơ sở phân tích khó khăn trong việc thu thập dữ liệu gán nhãn, mô hình được thiết kế để khai thác hiệu quả dữ liệu chưa gán nhãn thông qua cơ chế gán nhãn giả, từ đó giảm phụ thuộc vào nguồn dữ liệu gán nhãn đầy đủ.

Kết quả thực nghiệm cho thấy mô hình dựa trên học bán giám sát kết hợp PhoBERT và Swin Transformer V2 đạt hiệu suất vượt trội so với các mô hình đơn lẻ cũng như nhiều mô hình kết hợp khác. Đặc biệt, ngay cả khi tỷ lệ dữ liệu gán nhãn chỉ ở mức thấp (5–20%), mô hình vẫn duy trì được độ chính xác và F1-score ở mức cao, chứng minh tính hiệu quả và khả năng khái quát hóa tốt của hướng tiếp cận bán giám sát.

Các phân tích chi tiết cũng chỉ ra rằng mô hình không chỉ mạnh trong việc phân biệt Tin bài thường với tin độc hại, mà còn có khả năng nhận diện tương đối chính xác các nhóm nội dung độc hại khác nhau. Mặc dù vẫn tồn tại sai sót giữa các nhóm nhãn có ngữ nghĩa gần nhau, hiệu năng tổng thể ổn định và duy trì ở mức trên 90% khi tỷ lệ dữ liệu gán nhãn đạt từ 10% trở lên.

Những kết quả này khẳng định tính khả thi của việc áp dụng học bán giám sát trong phát hiện tin bài độc hại tiếng Việt, đặc biệt trong bối cảnh dữ liệu gán nhãn hạn chế và dữ liệu chưa gán nhãn dồi dào.

Các kết quả ở chương này đã được đăng trên Tạp chí Khoa học Công nghệ Thông Tin và Truyền thông [CT3] và ở Hội thảo quốc tế VCRIS 2025 [CT5].

KẾT LUẬN

Tin bài độc hại tiếng Việt, đặc biệt là nhóm tin mang nội dung độc hại về chính trị, đã và đang trở thành một thách thức nghiêm trọng trên không gian mạng. Loại tin bài này thường có đặc điểm lan truyền nhanh, phương thức thể hiện tinh vi, đồng thời kết hợp cả văn bản và hình ảnh, khiến cho các phương pháp phát hiện truyền thống gặp nhiều hạn chế. Luận án này tập trung phát triển các mô hình đa phương thức dựa trên học có giám sát và học bán giám sát cho phát hiện các tin bài độc hại trên không gian mạng được thể hiện dưới dạng kết hợp văn bản và hình ảnh.

Các kết quả đạt được

Luận án đã thực hiện được các mục tiêu nghiên cứu đã đề ra thông qua hai hướng tiếp cận chính, với các kết quả cụ thể như sau:

(1) Đề xuất mô hình đa phương thức kết hợp đặc trưng văn bản (sử dụng PhoBERT) và đặc trưng hình ảnh (sử dụng Swin Transformer V2). Mô hình này cho phép khai thác đồng thời đặc trưng văn bản và hình ảnh giúp khai thác tối ưu mối quan hệ bổ trợ giữa ngôn ngữ và thị giác để cải thiện khả năng nhận diện các tin bài độc hại tiếng Việt. Hơn nữa, mô hình đề xuất có khả năng phát hiện tin bài đa lớp, cụ thể với 4 nhóm: tin bài thường, vu khống bôi nhọ, kêu gọi chống đối và thúc đẩy tư tưởng cực đoan.

(2) Đề xuất mô hình phát hiện tin bài độc hại tiếng Việt dựa trên phương pháp học bán giám sát, sử dụng kỹ thuật gán nhãn giả (pseudo-labeling) nhằm tận dụng hiệu quả nguồn dữ liệu chưa được gán nhãn. Mô hình phát hiện đề xuất dựa trên học bán giám sát sử dụng kế thừa mô hình đa phương thức kết hợp đặc trưng văn bản và đặc trưng hình ảnh từ đề xuất (1). Kết quả thực nghiệm cho thấy phương pháp học bán giám sát có khả năng khai thác dữ liệu chưa gán nhãn và duy trì hiệu năng phân loại ở mức cao khi lượng dữ liệu có nhãn được giảm đáng kể (tỷ lệ dữ liệu có nhẫn thử nghiệm từ 5% đến 50%), qua đó chứng minh tính khả thi của phương pháp trong bối cảnh lượng dữ liệu được gán nhãn hạn chế.

Bên cạnh đó, để phục vụ cho quá trình huấn luyện và đánh giá mô hình, luận án đã tiến hành thu thập và xây dựng bộ dữ liệu tin bài tiếng Việt với quy mô

18.000 mẫu, bao gồm cả văn bản và hình ảnh. Bộ dữ liệu này đóng vai trò là nền tảng thực nghiệm quan trọng, góp phần đảm bảo tính khách quan và độ tin cậy của các kết quả nghiên cứu.

Các tồn tại, hạn chế

Bên cạnh những kết quả đạt được, luận án cũng ghi nhận một số hạn chế cần được tiếp tục nghiên cứu: (i) Phạm vi phương thức dữ liệu: Luận án hiện mới chỉ tập trung xử lý hai phương thức là văn bản và hình ảnh, chưa mở rộng sang các định dạng phức tạp khác như âm thanh hoặc video. (ii) Giới hạn về loại hình nội dung: Dữ liệu nghiên cứu chủ yếu tập trung vào nhóm tin bài độc hại về chính trị, chưa bao phủ các loại tin bài độc hại khác trên không gian mạng. (iii) Ngôn ngữ đơn lẻ: Nghiên cứu chỉ giới hạn trong phạm vi tiếng Việt, chưa có sự thử nghiệm hoặc so sánh đa ngôn ngữ. (iv) Vấn đề phát sinh từ học bán giám sát: Việc gán nhãn giả vẫn tồn tại nguy cơ lan truyền lỗi. Nếu các dự đoán ban đầu sai lệch, những sai sót này có thể tích lũy và làm giảm chất lượng huấn luyện của mô hình ở các giai đoạn sau. (v) Sai sót giữa các nhãn tương đồng: Thực nghiệm cho thấy mô hình đôi khi vẫn nhầm lẫn giữa các nhóm nhãn có ngữ nghĩa gần nhau hoặc các trường hợp ngụy trang tinh vi.

Hướng phát triển

Để hoàn thiện và nâng cao tính ứng dụng của nghiên cứu, luận án đề xuất các hướng phát triển tiếp theo: (i) Mở rộng quy mô và đa dạng dữ liệu: Tiếp tục thu thập dữ liệu từ nhiều nguồn khác nhau trên mạng xã hội để tăng tính đa dạng và khả năng ứng dụng thực tế của mô hình. (ii) Nghiên cứu các kỹ thuật học sâu tiên tiến: Khám phá các phương pháp như học tự giám sát (self-supervised learning) hoặc học ít mẫu (few-shot learning) nhằm giảm thiểu tối đa sự phụ thuộc vào dữ liệu gán nhãn thủ công. (iii) Tích hợp thêm các phương thức mới: Nghiên cứu khả năng kết hợp thêm các yếu tố về video và âm thanh để có cái nhìn toàn diện hơn về nội dung độc hại đa phương tiện. (iv) Xây dựng hệ thống giám sát thời gian thực: Phát triển một hệ thống hoàn chỉnh có khả năng hoạt động trực tuyến, hỗ trợ các cơ quan chức năng giám sát và ngăn chặn tin bài độc hại theo thời gian thực trên không gian mạng.

Quan hệ giữa các Công trình đã công bố với các Đóng góp của luận án

Công trình	Nội dung	Đóng góp của luận án
CT1	Đề xuất mô hình đa phương thức cho phát hiện tin bài độc hại tiếng Việt dựa trên PhoBERT + Swin Transformer V2 (nhị phân)	Đóng góp (1)
CT2	Đề xuất mô hình đa phương thức cho phát hiện tin bài độc hại tiếng Việt dựa trên PhoBERT + Swin Transformer V2 (nhị phân, mở rộng)	Đóng góp (1)
CT3	Đề xuất mô hình đa phương thức cho phát hiện tin bài độc hại tiếng Việt dựa trên học bán giám sát (đa lớp, mở rộng)	Đóng góp (2)
CT4	Đề xuất mô hình đa phương thức cho phát hiện tin bài độc hại tiếng Việt dựa trên PhoBERT + Swin Transformer V2 (đa lớp, mở rộng)	Đóng góp (1)
CT5	Đề xuất mô hình đa phương thức cho phát hiện tin bài độc hại tiếng Việt dựa trên học bán giám sát (nhị phân)	Đóng góp (2)

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, Hanoi, 10/2025. Link: https://ieeexplore.ieee.org/document/11250571.

TÀI LIỆU THAM KHẢO

[1] Social, DataReportal and We Are, “Digital 2023: Vietnam,” pp. https://datareportal.com/reports/digital-2023-vietnam, 2023.

[2] C. W. Derakhshan and Hossein D., “Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making,” pp. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c, 2017.

[3] T. T. P. Nguyen, D. C. Nguyen, A. T. T. Nguyen, L. H. Nguyen, G. T. Vu, C. T. Nguyen, T. H. Nguyen, and H. T. Le, “Fake News Affecting the Adherence of National Response Measures During the COVID-19 Lockdown Period: The Experience of Vietnam,” Frontiers in Public Health, tập 8, số 10.3389/fpubh.2020.589872, p. https://pubmed.ncbi.nlm.nih.gov/33072716/, 2020.

[4] “Bộ Công an: 'Nhóm tấn công hai trụ sở xã ở Đăk Lăk là khủng bố',” pp. https://vnexpress.net/bo-cong-an-nhom-tan-cong-hai-tru-so-xa-o-dak-lak-la-khung-bo-4620619.html, 2023.

[5] M. M. S. Nasser, N. I. B. Arshad, A. Ali, H. S. Alhussian, F. Saeed, A. Da’u and I. Nafea, “A systematic review of multimodal fake news detection on social media using deep learning models,” Results in Engineering, tập 26, số 104752, p. doi: 10.1016/j.rineng.2025.104752., 2025.

[6] D. Q. Nguyen, A. T. Nguyen, and T. H. Vu, “VLSP Shared Task: Vietnamese News Classification and Fake News Detection,” Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing (VLSP 2020), pp. https://aclanthology.org/2020.vlsp-1.pdf, 2020.

[7] S.-Y. Lin, Y.-C. Chen, Y.-H. Chang, S.-H. Lo and K.-M. Chao, “Text–image multimodal fusion model for enhanced fake news detection,” Science Progress, tập 107, p. doi: 10.1177/00368504241292685, 2024.

[8] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, “Semi-supervised Learning with Deep Generative Models,” p. https://arxiv.org/abs/1406.5298, 2014.

[9] V. C. Pham and T. Al-Dala’in, “Natural Language Processing (NLP) for Hate Speech Detection in Vietnamese Language: Challenges and Implementation,” Proceedings of the Third International Conference on Advances in Computing Research (ACR’25), pp. https://link.springer.com/chapter/10.1007/978-3-031-87647-9_10, 2025.

[10] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” số 10.48550/arXiv.1312.6114, p. https://arxiv.org/abs/1312.6114, 2014.

[11] Q. Xie, Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le, “Unsupervised Data Augmentation for Consistency Training,” Advances in Neural Information

Processing Systems (NeurIPS), p. https://arxiv.org/abs/1904.12848, 2020.

[12] D. Q. Nguyen and A. T. Nguyen, “Pre-trained language models for Vietnamese,” Findings of the Association for Computational Linguistics: EMNLP 2020, số 10.18653/v1/2020.findings-emnlp.92, p. https://www.scinapse.io/papers/3098637735, 2020.

[13] D. H. Lee, “Pseudo-Label : The Simple and Efficient Semi-Supervised LearningMethod for Deep Neural Networks,” pp. https://www.researchgate.net/publication/280581078_Pseudo-Label_The_Simple_and_Efficient_Semi-Supervised_Learning_Method_for_Deep_Neural_Networks#pf6, 2013.

[14] Quốc hội (2025), “Luật An ninh mạng số 116/2025/QH15,” p. https://vanban.chinhphu.vn/?pageid=27160&docid=216499&classid=1&orggroupid=1, Ngày ban hành ngày 10/12/2025, có hiệu lực ngày 01/07/2026.

[15] Quốc hội (2025), “Luật Báo chí số 126/2025/QH15,” p. https://vanban.chinhphu.vn/?pageid=27160&docid=216507&classid=1&orggroupid=1, ban hành ngày 10/12/2025, có hiệu lực từ ngày 01/07/2026.

[16] Chính phủ (2026), “Nghị định số 328/2026/NĐ-CP quy định phòng, chống tin giả, tin sai sự thật,” p. https://vanban.chinhphu.vn/?pageid=27160&docid=219241, ban hành ngày 19/08/2026, có hiệu lực từ ngày 05/10/2026..

[17] “Công văn số 7169/VPCP-NC của Văn phòng Chính phủ: V/v xử lý thông tin có nội dung chống Đảng và Nhà nước,” p. https://vanban.chinhphu.vn/default.aspx?pageid=27160&docid=163586.

[18] T. N. Tran, “Nhận diện và đấu tranh phản bác các quan điểm sai trái, thù địch đòi xóa bỏ vai trò lãnh đạo, cầm quyền của Đảng Cộng sản Việt Nam,” Tạp chí Cộng sản (online), pp. https://www.tapchicongsan.org.vn/web/guest/dau-tranh-phan-bac-cac-luan-dieu-sai-trai-thu-dich/chi-tiet/-/asset_publisher/YqSB2JpnYto9/content/nhan-dien-va-dau-tranh-phan-bac-cac-quan-diem-sai-trai-thu-dich-doi-xoa-bo-vai-tro-lanh-dao-cam-quyen-cua-dang-co, 30-01-2024.

[19] M. Mahyoob, J. Algaraady, and M. Alrahaili, “Linguistic-Based Detection of Fake News in Social Media,” International Journal of English Linguistics, tập 11, số 1, p. https://www.ccsenet.org/journal/index.php/ijel/article/view/0/44204, 2020.

[20] R. Sousa-Silva, “Fighting the Fake: A Forensic Linguistic Analysis to Fake News Detection,” International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique, tập 35, số 6, p. https://pmc.ncbi.nlm.nih.gov/articles/PMC9047580, 2022.

[21] S. Aneja, N. Aneja, and P. Kumaraguru, “Predictive Linguistic Cues for Fake News: A Societal Artificial Intelligence Problem,” IAES International Journal of Artificial Intelligence (IJ-AI), tập 11, số 4, p. https://arxiv.org/abs/2211.14505, 2022.

[22] B. Lutz, M. Adam, S. Feuerriegel, N. Pröllochs, and D. Neumann, “Which Linguistic Cues Make People Fall for Fake News? A Comparison of Cognitive and Affective Processing,” p. https://arxiv.org/abs/2312.03751, 2023.

[23] S. Qian, C. Shen, and J. Zhang, “Fighting cheapfakes: using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation,” Journal of Computer-Mediated Communication, tập Volume 28, số Issue 1, p. https://academic.oup.com/jcmc/article/28/1/zmac024/6839987, January 2023.

[24] M. H., V. D. M. Toni, and V. Rens, “How persuasive are political cheapfakes disseminated via social media? The effects of out-of-context visual disinformation on message credibility and issue agreement(,” Information, Communication & Society, pp. https://www.pbs.org/newshour/science/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation, 2024.

[25] I. Segura-Bedmar and S. Alonso-Bartolome, “Multimodal Fake News Detection,” tập 13, số 6, p. doi: 10.3390/info13060284, 2022.

[26] “How Photos Fuel the Spread of Fake News,” pp. https://www.wired.com/2016/12/photos-fuel-spread-fake-news, 2016.

[27] X. Zhou, J. Wu, and R. Zafarani, “SAFE: Similarity-Aware Multi-Modal Fake News Detection,” tập 12085, số 10.1007/978-3-030-47436-2_27, p. https://arxiv.org/abs/2003.04981, 2020.

[28] H. P. Nguyễn, M. C. Đỗ, and V. Q. Hoàng, “Proposing algorithm based on chatbot to detect sensitive comments,” JSTIC – PTIT, 2021.

[29] T. T. T. Nguyen and N. D. Nguyen, “Detecting negative speech on social networks using deep learning models and spelling errors,” p. JSTIC – PTIT, 2024.

[30] K. Q. Tran, A. T. Nguyen, P. G. Hoang, C. D. Luu, T. H. Do, and K. V. Nguyen, “Vietnamese hate and offensive detection using phobert-cnn and social media streaming data,” Proceedings of the 2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), p. https://arxiv.org/abs/2206.00524, 2022.

[31] L. T. Nguyen, K. V. Nguyen, and N. L. T. Nguyen, “Constructive and toxic speech detection for open-domain social media comments in vietnamese,” tập cs.CL, số 10.1007/978-3-030-79457-6_49, p. https://arxiv.org/abs/2103.10069, 2021.

[32] A. T. Huynh and P. Tran, “Utilizing Transformer Models To Detect Vietnamese Fake News on Social Media Platforms,” pp. https://science.tdtu.edu.vn/publication/2025/utilizing-transformer-models-detect-vietnamese-fake-news-social-media-platforms, 2025.

[33] K. D. Pham, K. V. Nguyen, and N. L. T. Nguyen, “Improving Vietnamese Fake News Detection based on Contextual Language Model and Handcrafted Features,” Science and Technology Development Journal, tập 7, số 1, pp. https://scispace.com/papers/improving-vietnamese-fake-news-detection-based-on-

contextual-48qez8hf6p, 2023.

[34] T. N. Le, K. B. T. Nguyen, and T. V. K. Ngo, “An Application of Machine Learning Models in Detecting Vietnamese Fake News,” International Journal of Software Innovation (IJSI), tập 12, số 1, pp. https://www.igi-global.com/article/an-application-of-machine-learning-models-in-detecting-vietnamese-fake-news/362623, 2024.

[35] Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, “Fake News Detection Using Machine Learning Approaches,” IOP Conference Series: Materials Science and Engineering, tập 1099, số 1, pp. https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/meta, 2021.

[36] A. Al Obaid, H. Khotanlou, M. Mansoorizadeh, and D. Zabihzadeh, “Multimodal Fake-News Recognition Using Ensemble of Deep Learners,” tập 24, số 9 (doi: 10.3390/e24091242), p. https://pmc.ncbi.nlm.nih.gov/articles/PMC9497605/, 2022.

[37] A. De, V. E., K. K. Maurya, and M. S. Desarkar, “Coarse and Fine-Grained Hostility Detection in Hindi Posts using Fine Tuned Multilingual Embeddings,” số 10.48550/arXiv.2101.04998, p. https://arxiv.org/abs/2101.04998, 2021.

[38] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, “Fake News Classification using transformer based enhanced LSTM and BERT,” International Journal of Cognitive Computing in Engineering, tập 3, số 10.1016/j.ijcce.2022.03.003, p. https://www.sciencedirect.com/science/article/pii/S2666307422000092, 2022.

[39] P. K. Verma, P. Agrawal, V. Madaan, and R. Prodan, “MCred: multi-modal message credibility for fake news detection using BERT and CNN,” Journal of Ambient Intelligence and Humanized Computing, tập 14, số 10.1007/s12652-022-04085-2, pp. https://link.springer.com/article/10.1007/s12652-022-04338-2, 2022.

[40] S. Das, R. Kumari, and R. K. Singh, “Detection of Fake News by Convolutional Neural Networks and Recurrent Neural Networks,” Procedia Computer Science, tập 258, số 10.1016/j.procs.2025.10.123, p. https://www.sciencedirect.com/science/article/pii/S1877050925015844, 2025.

[41] A. Sahi, M. Albdair, M. Diykh, S. Abdulla, H. Alghayab, K. Aljebur, and S. K. D. Alkhafaji, “SGDM-GRU: Spectral graph deep learning based Gated Recurrent Unit model for accurate fake news detection,” Expert Systems with Applications, tập 271, số DOI: 10.1016/j.eswa.2025.127572, p. https://www.sciencedirect.com/science/article/pii/S0957417425011947, 2025.

[42] E. F. Ayetiran and O. Ozgöbek, “An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection,” Information Systems, tập 122, số 10.1016/j.is.2024.102378, p. https://www.sciencedirect.com/science/article/pii/S030643792400036X, 2024.

[43] D. G. Dev, V. Bhatnagar, B. S. Bhati, M. Gupta, and A. Nanthaamornphong, “LSTMCNN: A hybrid machine learning model to unmask fake news,” Heliyon, tập 10, số 3, p. https://www.sciencedirect.com/science/article/pii/S2405844024012751,

2024.

[44] S. Tahmasebi, S. Hakimov, R. Ewerth, and E. Müller-Budack, “Improving Generalization for Multimodal Fake News Detection,” Proceedings of the 2023 ACM International Conference on Multimedia Retrieval (ICMR), số 10.1145/3591106.3592230, p. https://arxiv.org/abs/2305.18599, 2023.

[45] Y. Guo and W. Song, “A temporal-and-spatial flow based multimodal fake news detection by pooling and attention blocks,” IEEE Access, tập 9, p. https://ieeexplore.ieee.org/document/9989380, 2022.

[46] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, “Detecting fake news by exploring the consistency of multimodal data,” Information Processing & Management, tập 58, số 5, p. https://www.sciencedirect.com/science/article/pii/S0306457321001060, 2021.

[47] Y. Wu, P. Zhan, Y. Zhou, and Z. Xu, “Multimodal Fusion with Co-Attention Networks for Fake News Detection,” Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021), số 10.18653/v1/2021.findings-acl.226, pp. https://www.semanticscholar.org/paper/Multimodal-Fusion-with-Co-Attention-Networks-for-Wu-Zhan/a852e1d3adae5179ca75abe1b93b4dcb410cc082, 2021.

[48] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, “Multi-Level Multi-Modal Cross-Attention Network for Fake News Detection,” IEEE Access, tập 9, số 10.1109/ACCESS.2021.3114153, p. https://ieeexplore.ieee.org/abstract/document/9541113, 2021.

[49] Y. Zhou, Q. Ying, Z. Qian, S. Li, and X. Zhang, “Multimodal Fake News Detection via CLIP-Guided Learning,” số 10.1145/3581783.3611833, p. https://arxiv.org/abs/2205.14304, 2022.

[50] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, “Multi-modal Fake News Detection on Social Media via Multi-grained Information Fusion,” Proceedings of the 2023 ACM Multimedia Conference (MM '23), số 10.1145/3581783.3611833, p. https://arxiv.org/abs/2304.00827, 2023.

[51] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, and S. Wang, “Cross-modal Contrastive Learning for Multimodal Fake News Detection,” tập cs.LG, số 10.1145/3581783.3613850, p. https://arxiv.org/abs/2302.14057, 2023.

[52] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, “Multimodal fake news detection via progressive fusion networks,” Information Processing & Management, tập 60, số 1, p. https://www.sciencedirect.com/science/article/pii/S0306457322002217, 2023.

[53] H. Lv, W. Yang, F. Wei, J. Peng, and H. Geng, “MDF: A Dynamic Fusion Model for Multi-modal Fake News Detection,” số 10.48550/arXiv.2406.19702, p. https://arxiv.org/abs/2406.19776, 2024.

[54] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, “Exploring the Role of Visual Content in Fake News Detection,” p. https://arxiv.org/pdf/2003.05096, 2020.

[55] S. K. Uppada, P. Patel, and S. B., “An image and text-based multimodal model for detecting fake news in OSN’s,” Journal of Intelligent Information Systems, tập 61, pp. https://link.springer.com/article/10.1007/s10844-022-00764-y, 2022.

[56] K. Armin, S. Djordje, and Z. Matthias, “Multimodal Detection of Information Disorder from Social Media,” Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), tập 10.1109/ICMEW53276.2021.9455964, p. https://arxiv.org/pdf/2105.15165, 2021.

[57] H. Nguyen and S. Gokhale, “An efficient approach to identifying anti-government sentiment on Twitter during Michigan protests,” PeerJ Computer Science, tập 8, số 10.7717/peerj-cs.1127, p. https://pmc.ncbi.nlm.nih.gov/articles/PMC9748841/, 2022.

[58] N. M. Nguyen and T. H. M. Tran, “A semi-supervised approach for Vietnamese stock news classification with deep learning,” ournal of Finance and Marketing, p. https://doi.org/10.52932/jfm.vi2.500, 2024.

[59] H. T. Le, R. C. Sam, H. C. Nguyen, and T. T. Nguyen, “Named Entity Recognition in Vietnamese Text Using Label Propagation,” số 10.1109/SOCPAR.2013.7054160, pp. ISBN: 978-1-4799-3400-3/13 (Kỷ yếu hội nghị SoCPaR 2013), 2013.

[60] Q. H. Pham, M. L. Nguyen, B. T. Nguyen, and V. C. Nguyen, “Semi-supervised Learning for Vietnamese Named Entity Recognition using Online Conditional Random Fields,” pp. Kỷ yếu hội nghị "Proceedings of the Fifth Named Entity Workshop, joint with 53rd ACL and the 7th IJCNLP", 2015.

[61] T. H. Lê, C. Sam, T. T. Nguyễn, T. L. Nguyễn, and M. D. Trịnh, “Trích rút quan hệ giữa các thực thể từ văn bản tiếng Việt sử dụng phương pháp lan truyền nhãn,” p. https://vjs.ac.vn/jcc/article/view/2566/pdf, 2014.

[62] S. Prabhav, S. Ridam, K. P. S. Rana, and V. Kumar, “SEMI-FND: Stacked Ensemble Based Multimodal Inference For Faster Fake News Detection,” p. https://arxiv.org/abs/2205.08159, 2022.

[63] Y. Wang, F. Ma, H. Wang, K. Jha, and J. Gao, “Multimodal Emergent Fake News Detection via Meta Neural Process Networks,” Proceedings of the 2021 Web Conference (WWW '21), số 10.1145/3442381.3449914, p. https://arxiv.org/abs/2106.13711, 2021.

[64] B. Lakzaei, M. H. Chehreghani, and A. Bagheri, “Label Propagation and One-Class Semi-Supervised Graph Attention Network for Fake News Detection,” tập cs.SI, số 10.48550/arXiv.2311.08272, p. https://arxiv.org/pdf/2402.08401, 2024.

[65] P. Meel and D. K. Vishwakarma, “Fake News Detection using Semi-Supervised Graph Convolutional Network,” Multimedia Tools and Applications, tập 80, số 10.1007/s11042-020-10211-1, p. https://www.semanticscholar.org/reader/a355abfab37a8daba7157434123ce8f17e175ee1, 2021.

[66] X. Li, P. Lu, L. Hu, X. Wang, and L. Lu, “A novel self-learning semi- supervised

deep learning network to detect fake news on social media," Multimedia Tools and Applications, tập 81, số 1182, pp. https://link.springer.com/article/10.1007/s11042-021-11065-x, 2021.

[67] A. Tam, "A Brief Introduction to BERT," Machine Learning Mastery, pp. https://machinelearningmastery.com/a-brief-introduction-to-bert.

[68] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, O. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," pp. arXiv preprint arXiv: 1907.11692, 2019..

[69] M. Visweswaran, J. Mohan, S. S. Kumar, and K. P. Soman, "Synergistic Detection of Multimodal Fake News Leveraging Attention Fusion," Procedia Computer Science, tập 235, số DOI: 10.1016/j.procs.2024.04.017, p. https://www.sciencedirect.com/science/article/pii/S1877050924006938.

[70] G. Boesch, "Analyzing Depthwise Separable Convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. https://viso.ai/deep-learning/xception-model, 2017.

[71] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin Transformer V2: Scaling Up Capacity and Resolution," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), số 10.1109/CVPR52688.2022.01170, p. https://huggingface.co/papers/2111.09883, 2022.

[72] K. Krippendorff, "Reliability in Content Analysis: Some Common Misconceptions and Recommendations," Human Communication Research, tập Volume 30, số Issue 3, pp. https://academic.oup.com/hcr/article-abstract/30/3/411/4331534, 2004.

[73] "AN TOÀN THÔNG TIN, AN NINH MẠNG VÀ BẢO VỆ QUYỀNRIÊNG TƯ - HỆ THỐNG QUẢN LÝ AN TOÀN THÔNG TIN - CÁC YÊU CẦU," p. https://drive.google.com/file/d/1yAkJvC4trdxI_RXJl4HLMNv3m8Cx0O4d/view?usp=sharing.

[74] "18:2024:TT-BTTTT," pp. https://drive.google.com/file/d/1Tg87bT6KPM2rL_GuQaYL8ARoc-dpa-21/view?usp=sharing.

[75] "Chỉ thị số 14:CT-TTg," pp. https://drive.google.com/file/d/1B9-9fczhqvp1GMwUki9ulUjgSqFS3jD2/view?usp=sharing.

[76] "Out-of-context photos are a powerful low-tech form of misinformation(20)," pp. https://www.pbs.org/newshour/science/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation, 2020.

[77] Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J. and Li, J., "Exploring the Role of Visual Content in Fake News Detection," Proceedings of the 2020 International Conference on Disinformation in Social Media, p. https://openreview.net/pdf?id=XPZIaotutsD.

[78] “Over 900 websites banned for advertising in Vietnam in 2024,” (MIC), Ministry of Information and Communications, pp. https://english.mst.gov.vn/over-900-websites-banned-for-advertising-in-vietnam-in-2024-19725010209494810 3.htm, January 2, 2025 (09:49 am GMT+7).

[79] (MST), Bộ Khoa học và Công Nghệ, “Tăng cường quản lý buôn bán và quảng cáo sai sự thật trên không gian mạng,” pp. https://mst.gov.vn/tang-cuong-quan-ly-buon-ban-va-quang-cao-sai-su-that-tren-khong-gian-mang-197241224213236396.htm, 14/12/2024.

[80] G. Perboli and E. Arabnezhad, “A Machine Learning-based DSS for mid and long-term company crisis prediction,” Expert Systems with Applications, tập 183, số 115421, p. https://www.sciencedirect.com/science/article/pii/S0957417421001998?via%3Dihub , 2021.

[81] M. C. Souza, B. M. Nogueira, R. G. Rossi, R. M. Marcacini, B. N. D. Santos, and S. O. Rezende, “A network-based positive and unlabeled learning approach for fake news detection,” Machine Learning, tập 110, số 115421, p. https://pubmed.ncbi.nlm.nih.gov/34815619/, 2021.

[82] D. Varshney and D. K. Vishwakarma, “A review on rumour prediction and veracity assessment in online social network,” Expert Systems with Applications, tập 168, số 114208, p. https://www.researchgate.net/publication/346870208_A_review_on_rumour_prediction_and_veracity_assessment_in_online_social_network, 2021.

[83] E. Miranda, M. Aryuni, Y. Fernando, and T. M. Kibtiah, “A study of radicalism contents detection in Twitter: insights from support vector machine technique,” 2020.

[84] A. Bondielli and F. Marcelloni, “A survey on fake news and rumour detection techniques,” Information Sciences, tập 497, số 10.1016/j.ins.2019.05.035, p. https://www.sciencedirect.com/science/article/abs/pii/S0020025519304372, 2019.

[85] D. Varshney and D. K. Vishwakarma, “A unified approach for detection of Clickbait videos on YouTube using cognitive,” Applied Intelligence, tập 51, số 7, p. https://pubmed.ncbi.nlm.nih.gov/34764575/, 2021.

[86] M. Paraschiv, N. Salamanos, C. Iordanou, N. Laoutaris, and M. Sirivianos, “A Unified Graph-Based Approach to Disinformation Detection using Contextual and Semantic Relations,” arXiv, p. https://arxiv.org/abs/2109.11781, 2021 , công bố hội nghị ICWSM 2022.

[87] R. A. Alzahrani and M. Aljabri, “AI-Based Techniques for Ad Click Fraud Detection and Prevention: Review and Research Directions.,” Journal of Sensor and Actuator Networks, tập 12, số 1, p. https://ouci.dntb.gov.ua/en/works/7qyGdee7/, 2022.

[88] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,” p. https://doi.org/10.48550/arXiv.1909.11942.

[89] O. Araque and C. A. Iglesias, “An approach for radicalization detection based on emotion signals and semantic,” IEEE Access, tập 8, p. https://www.researchgate.net/publication/338663740_An_Approach_for_Radicalization_Detection_Based_on_Emotion_Signals_and_Semantic_Similarity, 2020.

[90] T. H. Vo, K. C. Ninh, and A. K. Tran, “Automatic fake news detection: achievements and challenges,” pp. https://www.neliti.com/publications/453661/automatic-fake-news-detection-achievements-and-challenges, 2022.

[91] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” số 10.48550/arXiv.1810.04805, p. https://arxiv.org/abs/1810.04805, 2018.

[92] J. S. Brennen, F. M. Simon, and R. K. Nielsen, “Beyond (Mis)Representation: Visuals in COVID-19 Misinformation,” The International Journal of Press/Politics, tập 26, số 1, p. https://pmc.ncbi.nlm.nih.gov/articles/PMC7548543, 2021.

[93] K. Shu, A. Bhattacharjee, F. Alatawi, T. Nazer, K. Ding, M. Karami, and H. Liu, “Combating Disinformation in a Social Media Age,” tập cs.SI, số 10.48550/arXiv.2007.07388, p. https://arxiv.org/abs/2007.07388, 2020.

[94] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, “Combating fake news: A survey on identification and mitigation techniques,” ACM Transactions on Intelligent Systems and Technology (TIST), tập 10, số 3, p. https://dl.acm.org/doi/10.1145/3305260, 2019.

[95] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” Proceedings of the eleventh annual conference on Computational learning theory (COLT '98), số 10.1145/279943.279962, p. https://dl.acm.org/doi/10.1145/279943.279962, 1998.

[96] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep learning-based text classification: A comprehensive review,” CM Computing Surveys (CSUR), tập 54, số 3, p. https://dl.acm.org/doi/10.1145/3439726, 2021.

[97] D. E. Allen, M. McAleer, and A. K. Singh, “Daily market news sentiment and stock prices,” Applied Economics, tập 51, số 29, p. https://www.tandfonline.com/doi/full/10.1080/00036846.2018.1564115, 2019.

[98] A. I. A. Abd-Elaal, A. Z. Badr, and H. M. K. Mahdi, “Detecting Violent Radical Accounts on Twitter,” International Journal of Advanced Computer Science and Applications (IJSA), tập 11, số 8, p. https://thesai.org/Publications/ViewPaper?Volume=11&Issue=8&Code=IJACSA&SerialNo=65, 2020.

[99] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, “Detecting Offensive Language in Tweets Using Deep Learning,” Applied Intelligence, tập 48, số 12, p. https://arxiv.org/abs/1801.04433, 2018.

[100] B. Vidgen and T. Yasseri, “Detecting weak and strong Islamophobic hate speech on social media,” số 10.48550/arXiv.1812.10400, p. https://arxiv.org/abs/1812.10400, 2018.

[101] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, “Detection and classification of social media-based extremist affiliation using sentiment analysis techniques,” Human-centric Computing and Information Sciences, tập 9, số 10.1186/s13673-019-0185-6, pp. https://link.springer.com/article/10.1186/s13673-019-0185-6, 2019.

[102] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, “Detection and Resolution of Rumours in Social Media: A Survey,” ACM Computing Surveys (CSUR), tập 51, số 2, p. https://arxiv.org/abs/1704.00656, 2018.

[103] S. E. Reid, M. Valasik, A. Bagavati, “Examining the physical manifestation of alt-right gangs: from online trolling to street fighting,” 2020.

[104] D. Nguyen, T. T. Nguyen, and C. V. Nguyen, “Fake advertisements detection using automated multimodal learning: a case study for Vietnamese real estate data.,” Applied Intelligence, tập 55, pp. https://link.springer.com/article/10.1007/s10489-025-06238-2, 2025 (Published online: 23 January 2025)..

[105] K. Shu, A. Sliva, S. Wang, and J. Tang, “Fake News Detection on Social Media: A Data Mining Perspective,” ACM SIGKDD Explorations Newsletter, tập 19, số 1, p. https://www.researchgate.net/publication/318981549_Fake_News_Detection_on_Social_Media_A_Data_Mining_Perspective, 2017.

[106] Y. Wang, S. Qian, and C. Xu, “Fake News Detection via Knowledge-driven Multimodal Graph Convolutional Networks,” Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR), số 10.1145/3372278.3390698, pp. https://www.semanticscholar.org/paper/Fake-News-Detection-via-Knowledge-driven-Multimodal-Wang-Qian/fecf3e4c9587a68e1d523a0c0682571fad0a74a9, 2020.

[107] P. Meel and D. K. Vishwakarma, “Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities,” Expert Systems with Applications, tập 153, số 10.1016/j.eswa.2019.112986, p. https://www.sciencedirect.com/science/article/abs/pii/S0957417419307043?via%3Dihub, 2020.

[108] G. Aditi, L. Hemank, K. Ponnurangam, and J. Anupam, “Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy,” p. https://dblp.org/rec/conf/www/0003LKJ13.html, 2013.

[109] A. Klein, “From Twitter to Charlottesville: analyzing the fighting words between the alt-right and antifa,” p. https://ijoc.org/index.php/ijoc/article/viewFile/10076/2533,

2019.

[110] G. Grill, “Future protest made risky: examining social media based civil unrest prediction research and products,” 2021.

[111] B. Duivenvoorde, “Generative AI and the future of marketing: A consumer protection perspective,” p. https://www.sciencedirect.com/science/article/pii/S2212473X25000148, 2025 (Available online 2 May 2025).

[112] Y. J. Lu and C. T. Li, “GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), số 10.18653/v1/2020.acl-main.48, p. https://arxiv.org/abs/2004.11648, 2020.

[113] L. Y. Xiao, “Illegal loot box advertising on social media? An empirical study using the Meta and TikTok ad transparency repositories,” tập 56, số 106069, p. https://www.sciencedirect.com/science/article/pii/S0267364924001353, 2025 (Available online 9 November 2024).

[114] J. R. DeCook, “Memes and symbolic violence: #proudboys and the use of memes for propaganda and the construction of collective identity,” 2018.

[115] U. Kursuncu, M. Gaur, C. Castillo, A. Alambo, K. Thirunarayan, V. Shalin, D. Achilov, I. B. Arpinar, and A. Sheth, “Modeling Islamist Extremist Communications on Social Media using Contextual Dimensions: Religion, Ideology, and Hate,” Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), tập 10.1145/3341161.3342894, p. https://arxiv.org/abs/1908.06520, 2019.

[116] S. Mussiraliyeva, M. Bolatbek, B. Omarov, Z. Medetbek, G. Baispay, and R. Ospanov, “On detecting online radicalization and extremism using natural language processing,” 2020.

[117] M. P. S. Gôlo, M. C. de Souza, R. G. Rossi, S. O. Rezende, B. M. Nogueira, and R. M. Marcacini, “One-class learning for fake news detection through multimodal variational autoencoders,” Engineering Applications of Artificial Intelligence, tập 122, số 10.1016/j.engappai.2023.106088, p. https://www.sciencedirect.com/science/article/abs/pii/S0952197623002725, 2023.

[118] M. Del Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, “Polarization and Fake News: Early Warning of Potential Misinformation Targets,” tập 13, số 2 (Article No.: 10), p. https://dl.acm.org/doi/10.1145/3316809, 2019.

[119] A. Hajibagheri and G. Sukhthankar, “Political polarization over global warming: analyzing Twitter data on climate change (poster),” 2014.

[120] M. M. Vu, “Preventing misinformation spread on online social networks: a methodology(29),” Journal of Science and Technology on Information and Communications (JSTIC) – PTIT, 2023.

[121] Z. Wu, D. Pi, J. Chen, M. Xie, and J. Cao, “Rumor detection based on propagation graph neural network with attention mechanism,” tập 226, số 10.1016/j.knosys.2021.106978, p. https://pmc.ncbi.nlm.nih.gov/articles/PMC7274137/, 2020.

[122] J. R. Steven, M. Etienne, M. Youssef, J. Jerret, and G. Vaibhava, “Self-critical Sequence Training for Image Captioning.,” p. https://arxiv.org/pdf/1612.00563, 2017.

[123] D. Duong, T. Nguyen, and M. Dang, “Stock market prediction using financial news articles on Ho Chi Minh Stock Exchange,” Proceedings of the 8th International Conference on Knowledge and Systems Engineering (KSE), số 10.1109/KSE.2016.7758030, pp. https://www.researchgate.net/profile/L-Minh-Dang/publication/303370683_101145285754628576 19/links/5c34785aa6fdccd6b59b0895/101145-28575462857619.pdf, 2016.

[124] H. Allcott and M. Gentzkow, “Social Media and Fake News in the 2016 Election,” Journal of Economic Perspectives, tập 31, số 2 (doi: 10.1257/jep.31.2.211), p. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211, 2017.

[125] W. Khan, M. A. Ghazanfar, M. A. Azam, A. Karami, K. H. Alyoubi, and A. Alfakeeh, “Stock market prediction using machine learning classifiers and social media news,” Journal of Ambient Intelligence and Humanized Computing, tập 13, số 7 (doi: 10.1007/s12652-020-01839-w), p. https://www.researchgate.net/publication/339938321_Stock_market_prediction_using_machine_learning_classifiers_and_social_media_news, 2020.

[126] M. Makrehchi, S. Shah, and W. Liao, “Stock prediction using event-based sentiment analysis,” Proceedings of the 2013 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), tập 1, số DOI: 10.1109/WI-IAT.2013.48, p. https://www.academia.edu/6794685/Stock_Prediction_Using_Event_based_Sentiment_Analysis, 2013.

[127] M. Zheng, B. Liu, and L. Sun, “Study of Deep Learning-Based Legal Judgment Prediction in Internet of Things Era.,” Computational Intelligence and Neuroscience, tập 2022, số 10.1155/2022/8490760, p. https://onlinelibrary.wiley.com/doi/10.1155/2022/8490760, 2022..

[128] G. Caldarelli, R. De-Nicola, F. Vigna, M. Petrocchi, and F. Saracco, “The role of bot squads in political propaganda on Twitter,” 2020.

[129] X. Dong, U. Victor, and L. Qian, “Two-path deep semisupervised learning for timely fake news detection,” p. https://arxiv.org/abs/2002.00763, 2020.

[130] M. Nouh, J. R. C. Nurse, and M. Goldsmith, “Understanding the Radical Mind: Identifying Signals to Detect Extremist Content on Twitter,” Proceedings of the 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), số DOI: 10.1109/ISI.2019.8823380, p. https://arxiv.org/abs/1905.08067, 2019.

[131] L. A. Cotfas, C. Delcea, R. Gherai, and I. Roxin, “Unmasking people’s opinions behind mask-wearing during COVID-19 pandemic: a Twitter stance analysis,” 2021.

[132] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, “ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation.,” Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), p. https://arxiv.org/abs/2205.06457., 2022..

[133] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao, “Weak supervision for fake news detection via reinforcement learning,” Proceedings of the AAAI Conference on Artificial Intelligence, tập 34, số 01 (doi: 10.1609/aaai.v34i01.5389), p. https://arxiv.org/abs/1912.12520, 2020.