

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG



LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC
NGÀNH CÔNG NGHỆ THÔNG TIN

Đề tài

TRANSFORMER
ĐỂ PHÂN LOẠI TIN GIẢ

TRANSFORMER
IN FAKE NEWS CLASSIFICATION

Sinh viên: Nguyễn Hồng Tuấn Phát
Mã số: B2111894
Khóa: 47

Cần Thơ, 04/2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN



LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC
NGÀNH CÔNG NGHỆ THÔNG TIN

Đề tài

**TRANSFORMER
ĐỂ PHÂN LOẠI TIN GIẢ**

**TRANSFORMER
IN FAKE NEWS CLASSIFICATION**

**Người hướng dẫn
ThS. Sử Kim Anh**

**Sinh viên thực hiện
Nguyễn Hồng Tuấn Phát
Mã số: B2111894
Khóa: 47**

Cần Thơ, 04/2025

LỜI CẢM ƠN

Trong suốt quá trình thực hiện và hoàn thành luận văn tốt nghiệp với đề tài “Transformer Đẻ Phân Loại Tin Giả”, em đã may mắn nhận được sự quan tâm, hướng dẫn và hỗ trợ tận tâm từ quý thầy cô, gia đình và bạn bè.

Trước hết, em xin gửi lời tri ân sâu sắc nhất đến cô Sứ Kim Anh. Cô đã tận tình hướng dẫn, định hướng nghiên cứu, chỉ bảo và đóng góp những ý kiến khoa học vô cùng giá trị, đồng thời cũng động viên, khích lệ em trong suốt hơn 4 tháng thực hiện đề tài, giúp em vượt qua những khó khăn và thử thách. Sự dùi dắt của cô là yếu tố then chốt để em có thể hoàn thành luận văn này.

Em cũng xin bày tỏ lòng biết ơn chân thành đến quý thầy cô Trường Công nghệ Thông tin và Truyền thông cùng quý thầy cô Khoa Phát triển Nông thôn – Trường Đại học Cần Thơ. Những kiến thức và kinh nghiệm quý báu mà thầy cô đã tận tâm truyền đạt trong suốt quá trình học tập là nền tảng vững chắc giúp em thực hiện nghiên cứu này.

Bên cạnh đó, em không thể không nhắc đến sự ủng hộ lớn lao từ gia đình và bạn bè. Mọi người đã luôn ở bên cạnh, động viên tinh thần và tạo mọi điều kiện thuận lợi nhất để em có thể tập trung vào việc học tập và nghiên cứu.

Mặc dù đã rất nỗ lực, song do sự hiểu biết và thời gian nghiên cứu còn giới hạn, luận văn chắc chắn không tránh khỏi những thiếu sót. Em kính mong nhận được những ý kiến đóng góp quý báu từ quý thầy cô và các bạn để luận văn được hoàn thiện hơn.

Cuối cùng, em xin kính chúc quý thầy cô dồi dào sức khỏe, hạnh phúc và gặt hái thêm nhiều thành công trong sự nghiệp giáo dục cao quý của mình.

Em xin chân thành cảm ơn!

Cần Thơ, ngày tháng 04 năm 2025

Sinh viên thực hiện

Nguyễn Hồng Tuấn Phát

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

Cần Thơ, ngày tháng 04 năm 2025

Giảng viên hướng dẫn

CAM KẾT

Luận văn này trình bày kết quả quá trình nghiên cứu độc lập của tôi, dựa trên việc tìm hiểu cơ sở lý thuyết và khảo sát các vấn đề thực tiễn liên quan. Quá trình này được thực hiện dưới sự hướng dẫn khoa học tận tình của ThS. Sử Kim Anh. Tôi xin cam đoan rằng các số liệu, phân tích và giải pháp được trình bày trong luận văn là trung thực, là thành quả lao động của riêng tôi và chưa từng được công bố dưới bất kỳ hình thức nào. Mọi nguồn tài liệu tham khảo sử dụng trong luận văn đều đã được trích dẫn đầy đủ và rõ ràng.

Tôi xin bày tỏ lòng biết ơn chân thành đối với những sự giúp đỡ quý báu đã góp phần tạo điều kiện cho tôi hoàn thành nghiên cứu này.

Cần Thơ, ngày tháng 04 năm 2025

Sinh viên thực hiện

Nguyễn Hồng Tuấn Phát

MỤC LỤC

DANH MỤC BIÊU BẢNG	i
DANH MỤC HÌNH ẢNH	ii
DANH MỤC TỪ VIẾT TẮT	iii
TÓM LUỢC	v
ABSTRACT	vi
PHẦN I: GIỚI THIỆU	1
1. Đặt vấn đề	1
2. Những liên cứu liên quan	1
3. Mục tiêu đề tài	5
4. Đối tượng và phạm vi nghiên cứu	5
4.1 Đối tượng nghiên cứu	5
4.2 Phạm vi nghiên cứu	5
5. Phương pháp nghiên cứu	5
6. Nội dung nghiên cứu	6
7. Bố cục của quyển luận văn	7
PHẦN II: NỘI DUNG	8
CHƯƠNG 1: ĐẶC TẢ YÊU CẦU	8
1.1 Giới thiệu	8
1.1.1 Mục đích	8
1.1.2 Phạm vi sản phẩm	8
1.1.3 Đối tượng sử dụng	9
1.2 Mô tả tổng quan	9
1.2.1 Bối cảnh và Mục tiêu ứng dụng	9
1.2.2 Mô tả chung về Hệ thống	9
1.2.3 Các chức năng chính	10
1.2.4 Đặc điểm người dùng	10
1.2.5 Các ràng buộc	10
1.2.6 Sơ đồ UseCase Tổng quát	11
1.3 Yêu cầu cụ thể	11
1.3.1 Các Tác nhân (Actors)	11
1.3.2 Chi tiết Use Cases	11

1.3.3	Yêu cầu Phi chức năng (Non – Functional Requirements – NFR)	18
1.3.4	Yêu cầu Dữ liệu (Data Requirements – DAT).....	19
1.3.5	Yêu cầu Giao diện (Interface Requirements).....	20
	CHƯƠNG 2: CƠ SỞ LÝ THUYẾT.....	21
2.1	Tin giả (Fake News)	21
2.2	Xử lý ngôn ngữ tự nhiên	22
2.3	Transformer.....	23
2.4	BERT	25
2.5	RoBERTa.....	27
2.6	Fine – tuning và Vai trò của Lớp Softmax trong Phân loại	29
	CHƯƠNG 3: PHƯƠNG PHÁP THỰC NGHIỆM	31
3.1	Tổng quan	31
3.2	Phân loại tin tức thật giả	32
3.2.1	Tập dữ liệu.....	32
3.2.2	Phân tích dữ liệu.....	34
3.2.3	Xây dựng phương pháp.....	48
	CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM	51
4.1	Thiết lập môi trường	51
4.2	Cơ sở đánh giá	51
4.3	Điều chỉnh siêu tham số và đào tạo mô hình	52
4.4	So sánh và Thảo luận kết quả	64
4.5	Phát triển ứng dụng	71
4.5.1	Đối tượng sử dụng.....	72
4.5.2	Chức năng của ứng dụng.....	72
4.5.3	Kết quả	72
	PHẦN III: KẾT LUẬN	86
1.	Kết quả đạt được.....	86
2.	Hướng phát triển.....	86
	TÀI LIỆU THAM KHẢO.....	88

DANH MỤC BIỂU BẢNG

Bảng 1. Số lượng chi tiết cho các tập huấn luyện (train) và kiểm thử (test) được sử dụng để nghiên cứu.....	33
Bảng 2. Kết quả phân loại chủ đề trên tập dữ liệu ISOT Fake News.....	34
Bảng 3. Kết quả phân loại chủ đề trên tập dữ liệu Fake News.....	39
Bảng 4. Kết quả phân loại chủ đề trên tập dữ liệu Fake or Real News	43
Bảng 5. Kết quả phân loại chủ đề trên tập dữ liệu Fake News Detection	45
Bảng 6. Kết quả đánh giá hiệu suất trên tập dữ liệu ISOT Fake News Dataset	53
Bảng 7. Kết quả đánh giá hiệu suất trên tập dữ liệu Fake News Dataset.....	56
Bảng 8. Kết quả đánh giá hiệu suất trên tập dữ liệu Fake or Real News Dataset	59
Bảng 9. Kết quả đánh giá hiệu suất trên tập dữ liệu Fake News Detection Dataset	62
Bảng 10. So sánh mô hình đề xuất với các nghiên cứu trước đó trên tập dữ liệu ISOT fake news	65
Bảng 11. So sánh mô hình đề xuất với các nghiên cứu trước đó trên tập dữ liệu Fake news	66
Bảng 12. So sánh mô hình đề xuất với các nghiên cứu trước đó trên tập dữ liệu Fake or Real news	67
Bảng 13. So sánh mô hình đề xuất với các nghiên cứu trước đó trên tập dữ liệu Fake news detection.....	68
Bảng 14. Kiểm thử chéo với các tập dữ liệu	69

DANH MỤC HÌNH ẢNH

Hình 1. UseCase Tổng Quát	11
Hình 2. Kiến trúc Transformer	24
Hình 3. Kiến trúc BERT và quy trình Pre – training/Fine – tuning	26
Hình 4. Kiến trúc RoBERTa	28
Hình 5. Phương pháp tiếp cận được đề xuất trong Phân loại tin tức thật giả.....	31
Hình 6. Biểu đồ phân bố số lượng mẫu tin thật/giả của các tập dữ liệu.....	32
Hình 7. Biểu đồ Top 15 chủ đề của tập ISOT có tỷ lệ tin giả cao nhất phân tích từ BERTopic	38
Hình 8. Biểu đồ Top 15 chủ đề của tập dữ liệu Fake News có tỷ lệ tin giả cao nhất phân tích từ BERTopic.....	42
Hình 9. Biểu đồ Top 15 chủ đề của tập dữ liệu Fake or Real News có tỷ lệ tin giả cao nhất phân tích từ BERTopic	44
Hình 10. Biểu đồ Top 15 chủ đề của tập dữ liệu Fake News Detection có tỷ lệ tin giả cao nhất phân tích từ BERTopic	46
Hình 11. Quy trình xử lý và phân loại của mô hình đề xuất	48
Hình 12. Đồ thị Training và Validation Accuracy/Loss của mô hình RoBERTa_Adamax_10 trên tập ISOT	54
Hình 13. Ma trận nhầm lẫn/chuẩn hóa của mô hình RoBERTa_Adamax_10 trên tập dữ liệu ISOT	55
Hình 14. Đồ thị Training và Validation Accuracy/Loss của mô hình RoBERTa_Adam_5 trên tập Fake News	57
Hình 15. Ma trận nhầm lẫn/chuẩn hóa mô hình RoBERTa_Adam_5 trên tập dữ liệu Fake News	58
Hình 16. Đồ thị Training và Validation Accuracy/Loss của mô hình RoBERTa_Adamax_10 trên tập Fake or Real News.....	60
Hình 17. Ma trận nhầm lẫn/chuẩn hóa mô hình RoBERTa_Adamax_10 trên tập dữ liệu Fake or Real	61
Hình 18. Đồ thị Training và Validation Loss của mô hình RoBERTa_AdamW_10 trên tập Fake News Detection	63
Hình 19. Ma trận nhầm lẫn chuẩn hóa mô hình RoBERTa_AdamW_10 trên tập dữ liệu Fake News Detection	64
Hình 20. Giao diện Đăng ký tài khoản mới cho người dùng	73
Hình 21. Giao diện đăng nhập vào hệ thống	74
Hình 22. Giao diện trang Hồ sơ người dùng	74
Hình 23. Giao diện Trang chủ dành cho người dùng	75
Hình 24. Quy trình và kết quả của chức năng phân loại tin tức	75
Hình 25. Giao diện Lịch sử phân loại và các thao tác liên quan	77
Hình 26. Giao diện Trang chủ (Dashboard) dành riêng cho Người quản trị	79
Hình 27. Giao diện Quản lý người dùng của Người quản trị	79
Hình 28. Giao diện Quản lý tin tức của Người quản trị	81
Hình 29. Giao diện trang “Thông tin về chúng tôi” (Giới thiệu)	84
Hình 30. Giao diện trang “Chính sách bảo mật”	85
Hình 31. Giao diện trang “Điều khoản dịch vụ”	85

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Điễn giải	Nghĩa tiếng việt
Acc	Accuracy	Độ chính xác tổng thể
AI	Artificial Intelligence	Trí tuệ nhân tạo
API	Application Programming Interface	Giao diện lập trình ứng dụng
BERT	Bidirectional Encoder Representations from Transformers	Biểu diễn mã hóa hai chiều từ Transformer
BiLSTM	Bidirectional Long Short – Term Memory	Bộ nhớ ngắn dài hạn hai chiều
BPE	Byte-Pair Encoding	Mã hóa Cặp Byte
CNN	Convolutional Neural Network	Mạng nơ – ron tích chập
CRUD	Create, Read, Update, Delete	Tạo, Đọc, Cập nhật, Xóa
CSDL		Cơ sở dữ liệu
CSV	Comma – Separated Values	
F1	F1 – score	Trung bình điều hòa của Precision (độ chính xác) và Recall (độ nhạy)
FNN	Feed – forward Neural Network	
FR	Functional Requirement	Yêu cầu chức năng
GPU	Graphics Processing Unit	Bộ xử lý đồ họa
JWT	JSON Web Token	
LSTM	Long Short – Term Memory	Bộ nhớ ngắn dài hạn
ML	Machine Learning	Học máy
MLM	Masked Language Model	Mô hình Ngôn ngữ Che (Mask)

Từ viết tắt	Điễn giải	Nghĩa tiếng việt
NFR	Non – Functional Requirement	Yêu cầu phi chức năng
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
NSP	Next Sentence Prediction	Dự đoán Câu Tiếp theo
P	Precision	Độ chính xác dương tính
PLM	Pre-trained Language Model	Mô hình ngôn ngữ được tiền huấn luyện
R	Recall	Độ nhạy
RNN	Recurrent Neural Network	Mạng nơ – ron hồi quy
RoBERTa	Robustly Optimized BERT Approach	Phương pháp BERT được tối ưu hóa một cách mạnh mẽ
SHAP	SHapley Additive exPlanations	
SVM	Support Vector Machine	Máy vector hỗ trợ
TF – IDF	Term Frequency – Inverse Document Frequency	Tần suất xuất hiện của từ – nghịch đảo tần suất tài liệu
UC	UseCase	
UI	User Interface	Giao diện người dùng
UniLSTM	Unidirectional Long Short – Term Memory	Bộ nhớ ngắn dài hạn một chiều
UX	User Experience	Trải nghiệm người dùng
VADER	Valence Aware Dictionary and sEntiment Reasoner	Bộ phân tích cảm xúc nhận biết ngữ cảnh
XAI	Explainable Artificial Intelligence	Trí tuệ nhân tạo Giải thích được

TÓM LƯỢC

Tin giả đang là một thách thức lớn trong thời đại kỷ nguyên số, nó gây ảnh hưởng trực tiếp đến nhận thức xã hội, làm hoang mang, mất niềm tin của mọi người. Vì vậy, việc phát triển một mô hình tự động có khả năng phân loại tin giả một cách chính xác và hiệu quả là rất cần thiết nhằm giảm thiểu tác động tiêu cực từ các thông tin sai lệch này. Nghiên cứu này tập trung vào việc ứng dụng và đánh giá hiệu quả của các mô hình Transformer tiên tiến, cụ thể là BERT và RoBERTa, cho bài toán phân loại tin giả. Các mô hình này đã được tinh chỉnh siêu tham số một cách có hệ thống và đánh giá trên bốn bộ dữ liệu phổ biến: ISOT Fake News Dataset, Fake News Dataset, Fake or Real News Dataset và Fake News Detection Dataset. Đồng thời, khả năng tổng quát hóa của các mô hình cũng được khảo sát thông qua các thử nghiệm kiểm tra chéo bộ dữ liệu. Kết quả thực nghiệm cho thấy các mô hình Transformer, đặc biệt là RoBERTa, đạt được hiệu suất phân loại rất cao khi được huấn luyện và kiểm thử trên cùng một bộ dữ liệu, với các chỉ số F1 – score và Accuracy vượt trội đạt đến 99.98% trên tập ISOT. Tuy nhiên, kết quả kiểm thử chéo bộ dữ liệu lại bộc lộ một thách thức đáng kể về khả năng tổng quát hóa: hiệu suất của mô hình giảm sút mạnh khi áp dụng trên các tập dữ liệu khác biệt so với tập huấn luyện, cho thấy sự quan trọng của mô hình đối với đặc trưng riêng của từng nguồn dữ liệu. Dựa trên sự cân bằng giữa hiệu suất và khả năng tổng quát hóa trung bình tốt hơn trong kiểm thử chéo, mô hình RoBERTa được huấn luyện trên tập Fake News Detection được đề xuất lựa chọn để phát triển ứng dụng thực tế. Nghiên cứu này không chỉ khẳng định tiềm năng to lớn của kiến trúc Transformer trong việc phát hiện tin giả dựa trên nội dung văn bản mà còn chỉ ra những hạn chế về tính tổng quát hóa cần được quan tâm trong các ứng dụng thực tế. Kết quả đạt được cung cấp cơ sở khoa học và thực nghiệm cho việc phát triển các công cụ kiểm chứng thông tin, hỗ trợ hoạt động báo chí số, góp phần xây dựng một không gian thông tin lành mạnh hơn.

Từ khóa: Tin giả, Phân loại tin giả, Mô hình Transformer, BERT, RoBERTa, Khả năng tổng quát hóa, Xử lý ngôn ngữ tự nhiên.

ABSTRACT

Fake news is a big challenge in the digital age, it directly affects social awareness, causing confusion and loss of trust among people. Therefore, developing an automatic model that can accurately and effectively classify fake news is essential to minimize the negative impact of this misinformation. This study focuses on the application and evaluation of the effectiveness of advanced Transformer models, specifically BERT and RoBERTa, for the problem of fake news classification. These models have been systematically tuned hyperparameters and evaluated on four popular datasets: ISOT Fake News Dataset, Fake News Dataset, Fake or Real News Dataset and Fake News Detection Dataset. At the same time, the generalization ability of the models is also examined through cross-dataset validation experiments. Experimental results show that Transformer models, especially RoBERTa, achieve very high classification performance when trained and tested on the same dataset, with outstanding F1-score and Accuracy indexes reaching 99.98% on the ISOT set. However, the cross-dataset testing results reveal a significant challenge in generalization ability: the model performance drops sharply when applied on datasets different from the training set, showing the importance of the model to the specific characteristics of each data source. Based on the balance between performance and better average generalization ability in cross-testing, the RoBERTa model trained on the Fake News Detection set is proposed to be selected for practical application development. This study not only confirms the great potential of the Transformer architecture in detecting fake news based on text content but also points out the limitations in generalization that need to be considered in practical applications. The results provide a scientific and empirical basis for developing information verification tools, supporting digital journalism activities, and contributing to building a healthier information space.

Keywords: *Fake news, Fake news classification, Transformer model, BERT, RoBERTa, Generalization ability, Natural Language Processing*

PHẦN I: GIỚI THIỆU

1. Đặt vấn đề

Công nghệ thông tin đang phát triển mạnh mẽ và được áp dụng rộng rãi trong nhiều lĩnh vực, từ sản xuất công nghiệp, y tế đến giáo dục và đời sống hàng ngày [1]. Đặc biệt, các công nghệ như trí tuệ nhân tạo (AI) và học máy (Machine Learning) được ứng dụng để hỗ trợ việc xử lý và phân tích dữ liệu lớn. Trong đó, các ứng dụng phân tích ngôn ngữ tự nhiên (NLP), học sâu (Deep Learning) và Transformer đang ngày càng được ứng dụng phổ biến nhằm mục đích trích xuất thông tin, phân loại văn bản, dịch ngôn ngữ, tóm tắt nội dung và hỗ trợ chatbot tự động. Những công nghệ này không chỉ giúp tối ưu hóa quá trình xử lý ngôn ngữ mà còn đóng vai trò quan trọng trong việc đánh giá và xác minh tính chính xác của thông tin.

Tuy nhiên, bất chấp những tiến bộ này, tin giả đã trở thành một vấn đề nghiêm trọng trên toàn cầu, đặc biệt với sự phát triển mạnh mẽ của mạng xã hội và các nền tảng trực tuyến [2]. Các thông tin sai lệch này có thể gây ảnh hưởng tiêu cực đến nhận thức xã hội, chính trị, kinh tế và thậm chí là sức khỏe, tinh thần của cộng đồng [3]. Việc phát hiện và phân loại tin tức thật – giả là một thách thức lớn, vì các tin giả này ngày càng tinh vi, chúng được tạo ra bằng cách chỉnh sửa nội dung hoặc sử dụng AI để tạo ra tin tức giả một cách chuyên nghiệp. Hiện nay, việc kiểm chứng tin tức chủ yếu dựa vào các tổ chức kiểm chứng hoặc các chuyên gia, tuy nhiên phương pháp này tốn nhiều thời gian và không thể xử lý khối lượng thông tin khổng lồ trên Internet.

Do đó, nhu cầu sử dụng công nghệ AI để tự động phát hiện và phân loại tin tức thật – giả ngày càng trở nên cấp thiết. Các mô hình AI có khả năng phân tích nội dung, nguồn tin, cách sử dụng ngôn từ và các yếu tố hình ảnh để xác định mức độ đáng tin cậy của một bài báo điện tử hay bản tin. Việc ứng dụng AI trong phân loại tin tức không chỉ giúp người dùng tiếp cận thông tin chính xác, mà còn góp phần ngăn chặn sự lan truyền của tin giả, nâng cao nhận thức cộng đồng và bảo vệ sự minh bạch trong môi trường truyền thông số.

2. Những liên cứu liên quan

Trong bài báo [4], Võ Trung Hưng và cộng sự đã trình bày việc xây dựng một công cụ phát hiện tin giả cho tiếng Việt dựa trên các kỹ thuật học sâu. Cụ thể, nghiên cứu đề xuất và thử nghiệm hai mô hình mạng nơ – ron là CNN (Mạng nơ – ron tích chập) và RNN (Mạng nơ – ron hồi quy) cho bài toán phân loại văn bản tin tức. Các mô hình này được huấn luyện và đánh giá trên tập dữ liệu tiếng Việt gồm 2400 bài báo về chủ đề Chính trị và Covid – 19 do nhóm tự xây dựng (bao gồm 2000 bài cho huấn luyện và 400 bài cho kiểm thử), với các nhãn thật/giả được

thu thập từ các trang web chính phủ và các trang đối lập. Kết quả thực nghiệm được báo cáo cho thấy công cụ có khả năng phân loại tin tức thật/giả với độ chính xác khoảng 85% trên tập dữ liệu kiểm thử. Nghiên cứu này khẳng định tiềm năng của CNN và RNN trong việc giải quyết vấn đề tin giả tiếng Việt và đề xuất hướng cải thiện trong tương lai bằng cách mở rộng dữ liệu và tinh chỉnh tham số mô hình.

Trong nghiên cứu [5], Võ Trung Hùng và cộng sự đã khảo sát và thực nghiệm các phương pháp phát hiện tin giả tự động cho tiếng Việt dựa trên phân tích nội dung. Nghiên cứu này đã so sánh hiệu quả của mô hình học máy truyền thống SVM (Support Vector Machine) với các mô hình học sâu gồm CNN (Mạng nơ – ron tích chập) và RNN (Mạng nơ – ron hồi quy). Các mô hình được huấn luyện và đánh giá trên bộ dữ liệu tiếng Việt tự xây dựng gồm khoảng 12.700 bài báo huấn luyện và 17.000 bài báo kiểm thử thuộc hai chủ đề chính là Chính trị và Covid – 19, được thu thập từ các nguồn báo chính thống và blog/Facebook. Kết quả thực nghiệm cho thấy mô hình RNN đạt hiệu quả cao nhất trong các mô hình được thử nghiệm, với điểm F1 – score là 0.77 cho chủ đề Chính trị và 0.73 cho chủ đề Covid – 19. Nghiên cứu kết luận rằng các phương pháp dựa trên nội dung, đặc biệt là RNN, có khả năng ứng dụng tốt cho việc phát hiện tin giả tiếng Việt, đồng thời cũng thảo luận các thách thức về việc thu thập, cập nhật dữ liệu và sự cần thiết phải tích hợp nhiều yếu tố phân tích khác trong tương lai để nâng cao độ chính xác.

Trong bài báo [6], tác giả Võ Đức Vinh và Đỗ Phúc đã đề xuất các mô hình học sâu để phát hiện tin giả tự động cho tiếng Việt. Nghiên cứu này đã khảo sát và đánh giá hiệu quả của ba mô hình học sâu gồm LSTM, BiLSTM, và CNN – BiLSTM (Mạng nơ–ron tích chập kết hợp LSTM hai chiều). Các mô hình được huấn luyện và đánh giá trên bộ dữ liệu tiếng Việt hỗn hợp (gồm khoảng 28.162 mẫu huấn luyện, 7.040 mẫu kiểm định và 8.800 mẫu kiểm thử), được xây dựng bằng cách thu thập từ các nguồn báo chính thống và nguồn tin giả trong nước, kết hợp với dữ liệu dịch từ tiếng Anh. Kết quả thực nghiệm cho thấy mô hình CNN–BiLSTM đạt hiệu quả cao nhất trong các mô hình được thử nghiệm, với điểm AUC (Area Under The Curve) là 0.966 và F1-score là 0.97. Nghiên cứu kết luận rằng các phương pháp học sâu này có khả năng ứng dụng tốt cho việc phát hiện tin giả tiếng Việt, đồng thời cũng thảo luận các thách thức về việc xây dựng dữ liệu, đặc thù ngôn ngữ, xác minh nguồn tin và đề xuất tích hợp các kỹ thuật như suy luận ngôn ngữ tự nhiên hay đồ thị tri thức trong tương lai để nâng cao độ chính xác.

Trong bài nghiên cứu [7], Patwa và cộng sự đã đề xuất mô hình SVM trên tập dữ liệu COVID – 19 Fake News Dataset, bao gồm 10,700 bài đăng và bài viết từ các nền tảng mạng xã hội như Facebook, Twitter, Instagram. Trong đó, 5,600

tin thật được thu thập từ các tài khoản chính thức như Tổ chức Y tế Thế giới (WHO) và các thông báo từ các tổ chức y tế như Trung tâm Kiểm soát và Phòng ngừa Dịch bệnh Hoa Kỳ (CDC). Còn lại, 5,100 tin giả là các bài viết liên quan đến COVID – 19 được lan truyền trên mạng xã hội và bị xác định là giả bởi các trang kiểm chứng thông tin uy tín như PolitiFact, Snopes và Boomlive. Mô hình trên đạt độ chính xác 93.32%, cho thấy hiệu quả cao trong việc phát hiện tin giả về COVID – 19. Đây là một trong những nghiên cứu quan trọng, góp phần nâng cao nhận thức về vấn đề tin giả trong đại dịch, đồng thời hỗ trợ phát triển các hệ thống kiểm chứng tin tức tự động dựa trên máy học. Trong nghiên cứu [8], tác giả đã đề xuất ba mô hình học sâu để phân loại tin tức giả, bao gồm CNN, LSTM, và BERT. Các mô hình này được huấn luyện trên tập dữ liệu COVID – 19 Fake News Dataset, sử dụng phương pháp tiền huấn luyện mô hình ngôn ngữ và biểu diễn từ phân tán để cải thiện hiệu suất. Kết quả thử nghiệm cho thấy BERT đạt độ chính xác cao nhất, lên đến 98.41%, vượt trội so với các mô hình CNN và LSTM. Điều này nhấn mạnh tính hiệu quả của các mô hình Transformer trong việc phát hiện tin tức giả, đặc biệt khi kết hợp với các kỹ thuật học không giám sát từ dữ liệu chưa gán nhãn.

Trong nghiên cứu [9], tác giả đã đề xuất hai mô hình học máy để phân loại tin tức giả, bao gồm Linear SVM và Decision Tree. Để biểu diễn đặc trưng của văn bản, tác giả sử dụng phương pháp tần suất xuất hiện của từ – nghịch đảo tần suất tài liệu (TF – IDF), giúp đánh giá mức độ quan trọng của từng từ trong một văn bản so với toàn bộ tập dữ liệu. Nghiên cứu được thực hiện trên tập dữ liệu ISOT Fake News Dataset gồm 44,919 tin tức, 23,502 tin tức giả và 21,417 tin tức thật, một tập dữ liệu phổ biến trong nghiên cứu về phát hiện tin tức giả. Kết quả tốt nhất đạt độ chính xác 92%, chứng tỏ tính hiệu quả của phương pháp được đề xuất trong việc phân loại tin tức thật và giả. Trong bài báo [10], nhóm tác giả đã thử nghiệm 23 mô hình phân loại khác nhau, bao gồm các mô hình như ZeroR, CV Parameter Selection (CVPS), Weighted Instances Handler Wrapper (WIHW), Decision Tree (DT), và một số phương pháp khác. Mục tiêu của họ là xác định phương pháp đạt được hiệu suất tốt nhất trong việc phân loại tin tức giả. Sau khi thực hiện các thử nghiệm, họ báo cáo rằng kết quả tốt nhất đạt được vượt trội hơn so với kết quả trong nghiên cứu [9], đạt được độ chính xác 96.8%, precision 96.3%, recall 97.3% và F1 – score 96.8%. Những kết quả này chứng tỏ hiệu quả cao của phương pháp được đề xuất, đồng thời cũng cho thấy sự cải thiện rõ rệt so với các nghiên cứu trước đó trong việc phát hiện tin tức giả.

Trong bài báo [11], Ahmad và cộng sự đã so sánh các phương pháp: Logistic Regression (LR), LSVM, Multilayer Perceptron (MLP), and KNN trên nhiều tập dữ liệu khác nhau, bao gồm ISOT Fake News Dataset[9], Fake News

Dataset[12], Fake News Detection Dataset[13], và một tập dữ liệu được tạo từ sự kết hợp của chúng. Khi thử nghiệm trên tập dữ liệu ISOT, thuật toán Random Forest đạt kết quả vượt trội so với nghiên cứu trước đó [9], [10], với độ chính xác 99%, độ chuẩn xác 99%, độ nhạy 100% và điểm F1 đạt 99%. Ngoài ra, thuật toán Random Forest cũng thể hiện hiệu suất tốt trên tập dữ liệu thứ ba và thứ tư, lần lượt đạt độ chính xác 95%, precision 98%, recall 93%, F1 – score 95% trên tập dữ liệu thứ ba và độ chính xác 91%, precision 92%, recall 91%, F1 – score 91% trên tập dữ liệu thứ tư. Bên cạnh đó, nghiên cứu cũng chỉ ra rằng khi sử dụng thuật toán Bagging Classifier kết hợp với Decision Tree trên tập dữ liệu Fake News Dataset, mô hình đạt độ chính xác 94%, precision 94%, recall 95% và F1 – score 94%.

Mimura và cộng sự [14] đã đề xuất mô hình Transformers (BERT) để thực hiện phân loại tin tức thật – giả trên hai tập dữ liệu phổ biến: ISOT Fake News Dataset và COVID – 19 Fake News Dataset. Mô hình được huấn luyện với mục tiêu khai thác ngữ cảnh từ văn bản nhằm nâng cao khả năng phân loại tin tức chính xác hơn so với các phương pháp truyền thống. Kết quả thực nghiệm cho thấy BERT đạt độ chính xác lên đến 99.84% trên tập dữ liệu ISOT và 90.24% trên tập dữ liệu COVID – 19. Sự chênh lệch này cho thấy mô hình hoạt động hiệu quả hơn trên các tập dữ liệu ổn định, trong khi dữ liệu về COVID – 19 có thể chứa nhiều yếu tố nhiễu hoặc tin tức có nội dung phức tạp hơn. Nhìn chung, nghiên cứu này chứng minh rằng BERT có khả năng xử lý các tập dữ liệu lớn, nắm bắt tốt ngữ cảnh văn bản, và là một phương pháp tiềm năng trong lĩnh vực phát hiện tin giả.

Trong nghiên cứu [15], các tác giả đã đề xuất mô hình FNDNet, một mạng nơ – ron tích chập sâu (CNN) để phát hiện tin giả. Mô hình này được huấn luyện và đánh giá trên tập dữ liệu Benchmarked, đạt độ chính xác 98.36% trên tập kiểm tra. Kết quả nghiên cứu cho thấy mô hình CNN có tiềm năng mạnh mẽ trong việc phân loại tin giả trên mạng xã hội. Trong bài báo [16], Bahad và cộng sự đã đề xuất một số kiến trúc học sâu như CNN, RNN, UniLSTM, BiLSTM để thực hiện phân loại tin giả. Kết quả thực nghiệm cho thấy mô hình BiLSTM vượt trội khi đạt độ chính xác 98.75% trên tập dữ liệu Fake News Detection[13] gồm 4,005 tin tức, trong đó 2,135 tin tức giả và 1,870 tin tức thật, nghiên cứu cũng đạt độ chính xác 91.48% với mô hình UniLSTM trên tập dữ liệu Fake or Real News[17]. Điều này cho thấy việc sử dụng các phiên bản LSTM có thể cải thiện đáng kể độ chính xác trong nhiệm vụ phân loại tin giả.

Trong bài báo [18], Deepak và Chitturi đã đề xuất mô hình LSTM kết hợp với khai thác dữ liệu bổ sung để phát hiện tin giả trên tập dữ liệu Fake or Real News[17] bao gồm 3154 tin tức giả và 3161 tin tức thật. Các đặc trưng bổ sung bao gồm tên miền tin tức, tác giả bài viết và tiêu đề nhằm cải thiện hiệu suất phân

loại. Họ đã thử nghiệm nhiều phương pháp nhúng từ như Bag of Words (BoW), Word2Vec và GloVe, kết hợp với Feed – forward Neural Network (FNN) và LSTM. Kết quả thực nghiệm độ chính xác của FNN tăng từ 83.3% lên 84.3%, trong khi LSTM có mức cải thiện đáng kể từ 83.7% lên 91.3%. Kết quả này khẳng định rằng việc kết hợp khai thác dữ liệu bổ sung giúp tăng cường hiệu suất mô hình so với các phương pháp thuần NLP.

3. Mục tiêu đề tài

Mục tiêu chính của nghiên cứu là xây dựng một mô hình trí tuệ nhân tạo dựa trên các mô hình biến đổi – Transformer tiên tiến, cụ thể là BERT và RoBERTa để phân loại tin thật hay giả. Để đạt được điều này, nghiên cứu tập trung vào việc ứng dụng và tinh chỉnh (fine – tuning) các kỹ thuật Transformer nhằm đạt được độ chính xác cao và khả năng khái quát hóa tốt cho mô hình. Quá trình này bao gồm việc thực hiện tinh chỉnh các siêu tham số quan trọng và đánh giá hiệu quả của phương pháp đề xuất. Cuối cùng, hiệu suất của các mô hình được đánh giá một cách định lượng trên bộ dữ liệu tổng hợp do Sastrawan và cộng sự công bố [19], nhằm chứng minh tính ứng dụng và hiệu quả thực tiễn của phương pháp.

Bên cạnh đó, một mục tiêu phụ là xây dựng ứng dụng web “Fake News Detector” sử dụng mô hình đã được huấn luyện hiệu quả nhất, nhằm minh họa khả năng ứng dụng thực tế và cung cấp công cụ cho người dùng thử nghiệm.

4. Đối tượng và phạm vi nghiên cứu

4.1 Đối tượng nghiên cứu

Đối tượng nghiên cứu là tập dữ liệu về tin tức thật và giả do Sastrawan và cộng sự tổng hợp và công bố [19]. Tập dữ liệu này bao gồm các bài viết và bài đăng trên mạng xã hội liên quan đến nhiều chủ đề khác nhau, được thu thập từ các nguồn đa dạng.

4.2 Phạm vi nghiên cứu

Phạm vi nghiên cứu tập trung vào việc tinh chỉnh siêu tham số và thực hiện các thực nghiệm với các mô hình Transformer cụ thể là BERT và RoBERTa. Hiệu suất phân loại của các mô hình này sẽ được đánh giá chi tiết thông qua các chỉ số đo lường tiêu chuẩn (accuracy, precision, recall, F1 – score) để xác định tính hiệu quả và khả thi của phương pháp đề xuất trên các tập dữ liệu đã xác định – bộ dữ liệu Sastrawan [19].

5. Phương pháp nghiên cứu

Nghiên cứu này áp dụng kết hợp các phương pháp sau:

Nghiên cứu lý thuyết:

Tổng quan các tài liệu, bài báo khoa học từ các nguồn uy tín liên quan đến lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), các mô hình phân

loại dựa trên kiến trúc Transformer (như BERT, RoBERTa), và các công trình nghiên cứu trước đây về phân loại tin tức thật/giả.

Tìm hiểu và phân tích đặc điểm của các bộ dữ liệu sẽ được sử dụng trong thực.

Nghiên cứu thực nghiệm:

Áp dụng các mô hình Transformer tiên tiến (BERT, RoBERTa) để xây dựng hệ thống phân loại tin tức.

Thực hiện huấn luyện, tinh chỉnh siêu tham số cho các mô hình trên các tập dữ liệu đã chọn.

Đánh giá hiệu quả của các mô hình thông qua các chỉ số đo lường định lượng phổ biến như Accuracy, Precision, Recall và F1 – score.

Phân tích và so sánh kết quả để đưa ra kết luận về hiệu suất và tính khả thi của phương pháp.

6. Nội dung nghiên cứu

Để đạt được các mục tiêu đã đề ra trong luận văn, các nội dung nghiên cứu chính đã được thực hiện bao gồm:

Nghiên cứu tổng quan:

Tìm hiểu về thực trạng tin tức giả và tầm quan trọng của việc phát hiện chúng.

Nghiên cứu các phương pháp phát hiện tin tức giả đã có, đặc biệt là các phương pháp ứng dụng học sâu và xử lý ngôn ngữ tự nhiên.

Đi sâu vào tìm hiểu cơ sở lý thuyết về kiến trúc Transformer, cơ chế Attention, và các mô hình ngôn ngữ dựa trên Transformer như BERT và RoBERTa, cũng như kỹ thuật fine – tuning.

Thu thập và xử lý dữ liệu:

Lựa chọn hoặc thu thập tập dữ liệu phù hợp cho bài toán phân loại tin tức thật/giả (tiếng Anh).

Xây dựng và huấn luyện mô hình:

Đề xuất và lựa chọn mô hình nền tảng (pre – trained model) dựa trên RoBERTa và BERT phù hợp với ngôn ngữ và đặc điểm của dữ liệu.

Thiết kế kiến trúc mô hình cho bài toán phân loại nhị phân (tin thật/tin giả) bằng cách thêm lớp phân loại (Softmax) vào mô hình nền tảng.

Thực hiện quá trình tinh chỉnh (fine – tuning) mô hình trên tập dữ liệu đã chuẩn bị.

Thử nghiệm và lựa chọn các siêu tham số tối ưu (optimizer, số epochs) thông qua quá trình huấn luyện và đánh giá.

Đánh giá và so sánh kết quả:

Xây dựng quy trình đánh giá hiệu năng mô hình dựa trên các độ đo phổ biến (Accuracy, Precision, Recall, F1 – score).

Tiến hành đánh giá mô hình đã huấn luyện trên tập dữ liệu kiểm thử (test set).

Phân tích, so sánh kết quả và thảo luận về hiệu quả của phương pháp đề xuất.

Phát triển ứng dụng minh họa:

Xây dựng một ứng dụng web đơn giản để minh họa khả năng ứng dụng thực tế của mô hình đã huấn luyện tốt nhất, cho phép người dùng nhập văn bản và nhận kết quả phân loại.

7. Bố cục của quyển luận văn

- ❖ PHẦN I: GIỚI THIỆU
- ❖ PHẦN II: NỘI DUNG
- ❖ PHẦN III: KẾT LUẬN

PHẦN II: NỘI DUNG

CHƯƠNG 1: ĐẶC TẢ YÊU CẦU

1.1 Giới thiệu

1.1.1 Mục đích

Tài liệu này nhằm mục đích đặc tả chi tiết các yêu cầu chức năng và phi chức năng cho hệ thống ứng dụng web “Fake News Detector”. Tài liệu này đóng vai trò là cơ sở tham chiếu cho quá trình thiết kế, phát triển, kiểm thử và nghiệm thu sản phẩm cuối cùng. Đồng thời, nó giúp các bên liên quan, bao gồm nhóm phát triển, người hướng dẫn và hội đồng đánh giá, hiểu rõ về phạm vi, mục tiêu và các khả năng của hệ thống được xây dựng trong khuôn khổ luận văn này.

1.1.2 Phạm vi sản phẩm

Hệ thống “Fake News Detector” là một ứng dụng web được phát triển với mục tiêu chính là cung cấp công cụ hỗ trợ người dùng phân loại tin tức (kết quả phân loại là tiếng Anh) là thật hay giả dựa trên việc ứng dụng các mô hình học sâu Transformer tiên tiến (cụ thể là RoBERTa và BERT).

Chức năng bao gồm (In – scope):

Tiếp nhận đầu vào là tiêu đề và nội dung văn bản tin tức từ người dùng.

Gọi API để xử lý văn bản (phát hiện ngôn ngữ, dịch nếu cần, tiền xử lý) và thực hiện dự đoán bằng mô hình đã huấn luyện.

Trả về kết quả phân loại (nhận Thật/Giả), tỷ lệ phần trăm xác suất tương ứng.

Cung cấp thông tin bổ sung như điểm phân tích cảm xúc (sentiment analysis).

Cung cấp giải thích cho dự đoán (qua các từ khóa ảnh hưởng từ SHAP – nếu người dùng yêu cầu).

Các chức năng quản lý tài khoản cơ bản cho người dùng: Đăng ký, Đăng nhập, Đăng xuất, Quản lý hồ sơ cá nhân (cập nhật thông tin, ảnh đại diện, đổi mật khẩu).

Lưu trữ và cho phép người dùng xem lại lịch sử các tin tức đã phân loại.

Cho phép người dùng báo cáo kết quả phân loại bị sai sót.

Cung cấp giao diện quản trị (Admin) để: Quản lý danh sách người dùng (xem, thêm, sửa, xóa) và Quản lý danh sách tin tức đã phân loại cùng các báo cáo sai sót liên quan (xem, xóa tin, duyệt báo cáo).

Xây dựng ứng dụng web minh họa cho các chức năng trên.

Chức năng không bao gồm (Out – of – scope):

Tự động thu thập tin tức từ các trang web hoặc mạng xã hội.

Phân tích các yếu tố khác ngoài văn bản (hình ảnh, video, đường link nguồn...).

Kiểm chứng thông tin dựa trên cơ sở dữ liệu tri thức thực tế (fact – checking database) theo thời gian thực.

Hỗ trợ phân loại cho các ngôn ngữ khác ngoài tiếng Anh một cách chuyên sâu (mặc dù có bước dịch tự động sơ bộ).

Các tính năng mạng xã hội phức tạp (bình luận, chia sẻ tin tức trong ứng dụng...).

1.1.3 Đối tượng sử dụng

Người dùng cuối (End User): Bất kỳ ai có nhu cầu kiểm tra nhanh tính xác thực của một đoạn văn bản tin tức. Đối tượng này có thể là sinh viên, nhà nghiên cứu, nhà báo, hoặc người dùng Internet thông thường muốn đánh giá thông tin trước khi tin tưởng hoặc chia sẻ. Yêu cầu có kiến thức cơ bản về sử dụng web.

Quản trị viên (Administrator): Người chịu trách nhiệm duy trì hệ thống ở mức độ cơ bản, quản lý người dùng và xem xét các phản hồi, báo cáo từ người dùng để góp phần cải thiện hệ thống (thu thập các trường hợp mô hình dự đoán sai).

1.2 Mô tả tổng quan

1.2.1 Bối cảnh và Mục tiêu ứng dụng

Trước thách thức ngày càng tăng của tin tức giả mạo (fake news) lan truyền trên các nền tảng trực tuyến, nhu cầu về các công cụ hỗ trợ người dùng tự động xác minh thông tin trở nên cấp thiết. Để đáp ứng nhu cầu này, ứng dụng web “Fake News Detector” được phát triển.

Mục tiêu chính của ứng dụng là cung cấp một nền tảng trực tuyến, dễ sử dụng, nơi người dùng có thể nhập văn bản tin tức và nhận được kết quả (tiếng Anh) phân loại tự động (Thật/Giả). Hệ thống ứng dụng các mô hình học sâu Transformer (cụ thể là RoBERTa/BERT đã được fine – tuning) để phân tích nội dung văn bản. Qua đó, ứng dụng giúp người dùng đánh giá sơ bộ tính xác thực của tin tức thông qua việc cung cấp dự đoán, tỷ lệ xác suất, điểm cảm xúc và các thông tin giải thích liên quan.

1.2.2 Mô tả chung về Hệ thống

Hệ thống được thiết kế theo kiến trúc Client – Server gồm các thành phần chính:

Frontend (Client): Xây dựng bằng React.js, sử dụng React Bootstrap để tạo giao diện người dùng đáp ứng, thân thiện. Axios được dùng để giao tiếp với

Backend qua API. React Router DOM quản lý điều hướng, Context API quản lý trạng thái.

Backend (Server): Xây dựng bằng Node.js với framework Express.js, cung cấp các RESTful API. Sử dụng Mongoose để tương tác với CSDL MongoDB. JWT và Bcryptjs đảm bảo xác thực và bảo mật mật khẩu. Multer xử lý tải file ảnh đại diện.

API Phân loại (Classification API): Xây dựng bằng Python với framework Flask. API này nhận văn bản từ Backend Node.js, thực hiện tiền xử lý, gọi mô hình RoBERTa/BERT đã huấn luyện để dự đoán, tính toán cảm xúc (VADER) và giải thích (SHAP), sau đó trả kết quả về cho Backend Node.js.

Cơ sở dữ liệu (Database): Sử dụng MongoDB (NoSQL) để lưu trữ thông tin người dùng, lịch sử phân loại, các báo cáo sai sót.

1.2.3 Các chức năng chính

Hệ thống cung cấp các nhóm chức năng cốt lõi sau:

Quản lý tài khoản: Đăng ký, đăng nhập, đăng xuất, xem/cập nhật hồ sơ, đổi mật khẩu.

Phân loại tin tức: Nhập liệu văn bản, nhận kết quả dự đoán thật/giả, xác suất, cảm xúc, giải thích SHAP.

Quản lý lịch sử: Xem lại các tin đã phân loại, xem chi tiết, xóa khỏi lịch sử.

Báo cáo sai sót: Gửi phản hồi về kết quả phân loại chưa đúng.

Quản trị hệ thống: Quản lý người dùng (Thêm, sửa, xóa), quản lý tin tức (Xem, xóa, duyệt báo cáo).

Truy cập thông tin: Xem các trang Giới thiệu, Chính sách Bảo mật, Điều khoản Dịch vụ.

1.2.4 Đặc điểm người dùng

Người dùng cuối được giả định có kỹ năng sử dụng máy tính và trình duyệt web ở mức cơ bản. Họ cần hiểu tiếng Anh để nhập liệu và diễn giải kết quả phân loại tin tức. Họ mong muốn có một công cụ nhanh chóng, dễ sử dụng để kiểm tra sơ bộ thông tin. Quản trị viên cần có hiểu biết cơ bản về quản lý người dùng và nội dung trên nền tảng web.

1.2.5 Các ràng buộc

Công nghệ: Việc phát triển bị ràng buộc bởi các công nghệ đã lựa chọn (React, Node.js, Python, MongoDB, RoBERTa/BERT...).

Dữ liệu: Chất lượng và đặc điểm của mô hình phân loại phụ thuộc lớn vào bộ dữ liệu Sastrawan [19] được sử dụng để huấn luyện và đánh giá.

Ngôn ngữ: Mô hình cốt lõi tập trung xử lý và phân loại văn bản tiếng Anh.

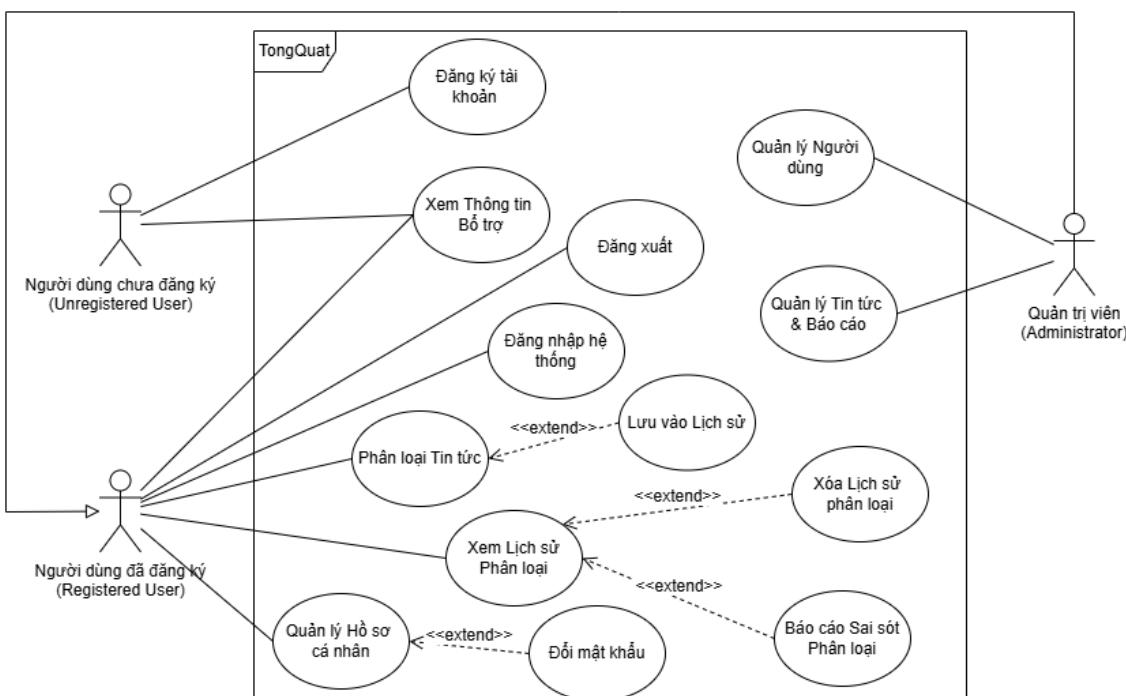
Phạm vi phân tích: Chỉ phân tích nội dung text, không xử lý đa phương tiện (hình ảnh, video).

Kiểm chứng: Hệ thống không thực hiện kiểm chứng thông tin đối chiếu với nguồn tin thực tế bên ngoài. Kết quả chỉ dựa trên phân tích của mô hình học sâu.

Tài nguyên: Hiệu năng API phân loại (đặc biệt là bước tính SHAP) có thể bị ảnh hưởng bởi cấu hình máy chủ triển khai.

1.2.6 Sơ đồ UseCase Tổng quát

Sơ đồ Use Case tổng quát của hệ thống “Fake News Detector”, thể hiện các tác nhân chính và các nhóm chức năng (UseCase) chính mà họ tương tác, được trình bày trong Hình 1.



Hình 1. UseCase Tổng Quát

1.3 Yêu cầu cụ thể

1.3.1 Các Tác nhân (Actors)

Người dùng chưa đăng ký (Unregistered User): Người dùng truy cập hệ thống nhưng chưa có tài khoản.

Người dùng đã đăng ký (Registered User): Người dùng đã có tài khoản và đăng nhập vào hệ thống.

Quản trị viên (Administrator): Người dùng có quyền quản lý hệ thống, người dùng và nội dung. (Ké thừa quyền của Người dùng đã đăng ký).

1.3.2 Chi tiết Use Cases

UC01: Đăng ký tài khoản

Tác nhân: Người dùng chưa đăng ký

Mục tiêu: Cho phép người dùng tạo một tài khoản mới để sử dụng các chức năng của hệ thống.

Luồng sự kiện chính:

Người dùng chọn chức năng/truy cập trang Đăng ký.

Hệ thống hiển thị form đăng ký yêu cầu nhập Tên đăng nhập, Email, Mật khẩu, Xác nhận mật khẩu.

Người dùng nhập đầy đủ thông tin bắt buộc, tick vào ô đồng ý điều khoản và nhấn nút “Đăng ký”.

Hệ thống kiểm tra tính hợp lệ của dữ liệu (định dạng email, mật khẩu khớp...).

Hệ thống kiểm tra Tên đăng nhập và Email có bị trùng trong CSDL không. Nếu trùng, báo lỗi và dừng luồng.

Nếu thông tin hợp lệ và không trùng, hệ thống mã hóa mật khẩu (sử dụng Bcryptjs).

Hệ thống lưu thông tin tài khoản mới (username, email, mật khẩu đã mã hóa, vai trò mặc định là ‘user’) vào CSDL MongoDB.

Hệ thống hiển thị thông báo đăng ký thành công cho người dùng.

Hệ thống chuyển hướng người dùng đến trang Đăng nhập.

Ngoại lệ:

Nếu kiểm tra ở bước 4 hoặc 5 thất bại, hệ thống hiển thị thông báo lỗi cụ thể cho người dùng và giữ nguyên trang đăng ký.

Nếu có lỗi CSDL ở bước 7, hệ thống hiển thị thông báo lỗi chung.

UC02: Đăng nhập hệ thống

Tác nhân: Người dùng chưa đăng ký/đã đăng ký, Quản trị viên

Mục tiêu: Cho phép người dùng đã có tài khoản truy cập vào các chức năng dành riêng cho họ hoặc vai trò của họ.

Luồng sự kiện chính:

Người dùng truy cập trang Đăng nhập.

Hệ thống hiển thị form đăng nhập yêu cầu nhập Username hoặc Email, và Mật khẩu.

Người dùng nhập thông tin và nhấn nút “Đăng nhập”.

Hệ thống tìm kiếm tài khoản dựa trên username hoặc email.

Nếu không tìm thấy, chuyển đến Luồng thay thế 1.

Hệ thống so sánh mật khẩu người dùng nhập với mật khẩu đã mã hóa lưu trong CSDL. Nếu không khớp, chuyển đến Luồng thay thế 2.

Nếu mật khẩu khớp, hệ thống tạo một JSON Web Token (JWT) chứa thông tin định danh người dùng (user ID, username, role).

Hệ thống trả JWT về cho client để lưu trữ.

Hệ thống xác định vai trò người dùng từ thông tin tài khoản.

Hệ thống chuyển hướng người dùng đến trang chủ tương ứng với vai trò (Trang chủ User hoặc Trang Admin).

Luồng thay thế 1 (Không tìm thấy tài khoản):

Hệ thống hiển thị thông báo lỗi “Tài khoản không tồn tại”.

Use case kết thúc.

Luồng thay thế 2 (Sai mật khẩu):

Hệ thống hiển thị thông báo lỗi “Mật khẩu không chính xác”.

Use case kết thúc.

Ngoại lệ:

Hệ thống xử lý và thông báo lỗi nếu có vấn đề khi tạo JWT hoặc lỗi CSDL.

UC03: Quản lý Hồ sơ cá nhân

Tác nhân: Người dùng đã đăng ký

Mục tiêu: Cho phép người dùng xem và cập nhật thông tin cá nhân, ảnh đại diện và đổi mật khẩu.

Điều kiện tiên quyết: Người dùng đã đăng nhập.

Luồng sự kiện chính (Xem/Cập nhật thông tin):

Người dùng chọn chức năng/truy cập trang Hồ sơ cá nhân.

Hệ thống truy vấn và hiển thị thông tin hiện tại của người dùng (Họ tên, Số điện thoại, Giới tính, Email, Username, Ảnh đại diện).

Người dùng chỉnh sửa các thông tin cho phép (Họ tên, Số điện thoại, Giới tính).

Người dùng nhấn nút “Lưu thay đổi”.

Hệ thống kiểm tra tính hợp lệ của dữ liệu nhập mới (nếu có ràng buộc).

Hệ thống cập nhật thông tin mới vào CSDL.

Hệ thống hiển thị thông báo cập nhật thành công và tải lại thông tin hồ sơ.

Luồng thay thế (Thay đổi ảnh đại diện):

Người dùng chọn chức năng thay đổi ảnh đại diện.

Hệ thống hiển thị cửa sổ cho phép chọn file ảnh từ máy tính.

Người dùng chọn file ảnh và tải lên.

Hệ thống (có thể) hiển thị công cụ cho phép người dùng cắt (crop) ảnh theo tỷ lệ mong muốn.

Người dùng xác nhận ảnh đã cắt.

Hệ thống lưu ảnh đại diện mới và cập nhật đường dẫn ảnh trong CSDL.

Hệ thống hiển thị ảnh đại diện mới trên trang hồ sơ.

UC04: Phân loại Tin tức

Tác nhân: Người dùng đã đăng ký

Mục tiêu: Nhận được dự đoán về tính thật/giả, điểm cảm xúc và các từ khóa ảnh hưởng (nếu có) cho một đoạn văn bản tin tức do người dùng cung cấp.

Điều kiện tiên quyết: Người dùng đã đăng nhập.

Luồng sự kiện chính:

Người dùng truy cập trang Phân loại tin tức.

Hệ thống hiển thị các ô nhập liệu cho Tiêu đề và Nội dung.

Người dùng nhập/dán Tiêu đề và Nội dung, sau đó nhấn nút “Phân loại”.

Frontend gửi yêu cầu chứa Tiêu đề và Nội dung đến Backend Node.js.

Backend Node.js gọi API Python/Flask, truyền văn bản cần xử lý và cờ yêu cầu giải thích (explain flag).

API Python thực hiện:

Phát hiện ngôn ngữ.

Dịch sang tiếng Anh (nếu cần).

Tiền xử lý văn bản tiếng Anh.

Đưa văn bản đã xử lý vào mô hình RoBERTa/BERT để dự đoán nhãn và xác suất.

Tính điểm cảm xúc (VADER).

Nếu cờ ‘explain’ là true, tính toán SHAP values và lấy top N từ ảnh hưởng đến dự đoán ‘Fake’.

Trả kết quả (xác suất thật/giả, điểm cảm xúc, top words SHAP) về cho Backend Node.js.

Backend Node.js nhận kết quả từ API Python và trả về cho Frontend.

Frontend hiển thị kết quả cho người dùng: Nhãn dự đoán (Real/Fake), Xác suất (%), Điểm cảm xúc, Top words SHAP (nếu có).

<<include>> UC04 – INC1: Lưu vào Lịch sử (Hệ thống tự động lưu thông tin phân loại này vào lịch sử của người dùng).

Ngoại lệ:

Nếu văn bản nhập vào không hợp lệ (quá ngắn, rỗng), hệ thống báo lỗi.

Nếu API Python gặp lỗi xử lý/dự đoán, hệ thống báo lỗi cho người dùng.

UC05: Xem Lịch sử Phân loại

Tác nhân: Người dùng đã đăng ký

Mục tiêu: Cho phép người dùng xem lại danh sách các tin tức đã phân loại trước đó.

Điều kiện tiên quyết: Người dùng đã đăng nhập.

Luồng sự kiện chính:

Người dùng truy cập trang Lịch sử phân loại.

Hệ thống truy vấn CSDL để lấy danh sách các mục tin tức đã được phân loại bởi người dùng này, sắp xếp theo thời gian (mới nhất trước).

Hệ thống hiển thị danh sách các mục, mỗi mục hiển thị thông tin tóm tắt (Tiêu đề, Ngày phân loại, Nhãn dự đoán).

Người dùng có thể chọn xem chi tiết một mục.

Hệ thống hiển thị đầy đủ nội dung và kết quả phân loại của mục đã chọn.

UC06: Báo cáo Sai sót Phân loại

Tác nhân: Người dùng đã đăng ký

Mục tiêu: Cho phép người dùng phản hồi về một kết quả phân loại mà họ cho là không chính xác.

Điều kiện tiên quyết: Người dùng đang xem chi tiết một mục trong Lịch sử phân loại (UC05).

<<extend>> UC05: Xem Lịch sử Phân loại

Luồng sự kiện chính:

Tại màn hình chi tiết lịch sử, người dùng nhấn nút “Báo cáo sai sót”.

Hệ thống hiển thị form báo cáo yêu cầu nhập: Lý do báo cáo, Nguồn kiểm chứng (link), Bình luận thêm (tùy chọn).

Người dùng điền thông tin và nhấn nút “Gửi báo cáo”.

Hệ thống lưu thông tin báo cáo (nội dung báo cáo, ID tin tức liên quan, ID người dùng báo cáo, thời gian, trạng thái mặc định ‘Pending’) vào CSDL.

Hệ thống hiển thị thông báo gửi báo cáo thành công cho người dùng.

UC07: Đăng xuất

Tác nhân: Người dùng đã đăng ký

Mục tiêu: Kết thúc phiên làm việc hiện tại của người dùng.

Điều kiện tiên quyết: Người dùng đã đăng nhập.

Luồng sự kiện chính:

Người dùng nhấn vào nút/link “Đăng xuất”.

Hệ thống (Client – side/Frontend) xóa bỏ JWT đã lưu trữ.

Hệ thống chuyển hướng người dùng về trang Đăng nhập hoặc trang chủ công khai.

UC08: Xem Thông tin Bổ trợ

Tác nhân: Người dùng chưa đăng ký, Người dùng đã đăng ký

Mục tiêu: Cung cấp thông tin về ứng dụng, chính sách bảo mật và điều khoản dịch vụ.

Luồng sự kiện chính:

Người dùng nhấn vào link tương ứng (Giới thiệu, Chính sách bảo mật, Điều khoản dịch vụ) thường ở footer hoặc menu.

Hệ thống hiển thị nội dung của trang thông tin tương ứng.

UC09: Quản lý Người dùng

Tác nhân: Quản trị viên

Mục tiêu: Cho phép Admin xem, tìm kiếm, thêm, sửa, xóa tài khoản người dùng trong hệ thống.

Điều kiện tiên quyết: Người dùng đăng nhập với vai trò Quản trị viên.

Luồng sự kiện chính:

Admin truy cập trang Quản lý Người dùng.

Hệ thống hiển thị danh sách người dùng (phân trang nếu cần), kèm các chức năng tìm kiếm/ lọc.

Admin thực hiện một trong các hành động sau:

Tìm kiếm: Nhập tiêu chí (username/email/tên) và nhấn
Tìm kiếm → Hệ thống lọc và hiển thị lại danh sách.

Xem chi tiết: Nhấn vào một người dùng → Hệ thống
hiển thị thông tin chi tiết.

Thêm mới: Nhấn nút “Thêm người dùng mới” → Hệ
thống hiển thị form thêm người dùng → Admin nhập thông
tin (username, email, pass, ...) → Nhấn Lưu → Hệ thống kiểm
tra, mã hóa pass, lưu vào CSDL, thông báo thành công.

Sửa: Nhấn nút “Sửa” của một người dùng → Hệ thống
hiển thị form với thông tin hiện tại → Admin chỉnh sửa thông
tin (có thể cả vai trò) → Nhấn Lưu → Hệ thống kiểm tra, cập
nhật CSDL, thông báo thành công.

Xóa: Nhấn nút “Xóa” của một người dùng → Hệ thống hiển
thị hộp thoại xác nhận → Admin xác nhận → Hệ thống xóa người
dùng khỏi CSDL, thông báo thành công.

UC10: Quản lý Tin tức

Tác nhân: Quản trị viên

Mục tiêu: Cho phép Admin xem lại các tin tức đã phân loại trong hệ
thống và xem xét, xử lý các báo cáo sai sót do người dùng gửi.

Điều kiện tiên quyết: Người dùng đăng nhập với vai trò Quản trị viên.

Luồng sự kiện chính:

Admin truy cập trang Quản lý Tin tức.

Hệ thống hiển thị danh sách các tin tức đã được phân loại
(phân trang, tìm kiếm theo tiêu đề).

Admin thực hiện một trong các hành động sau:

Xem chi tiết tin tức: Nhấn vào một tin tức → Hệ thống hiển
thị nội dung và kết quả phân loại gốc.

Xóa tin tức: Nhấn vào một tin tức → Nhấn nút “Xóa” → Xác
nhận → Hệ thống xóa tin tức.

Xem báo cáo của tin tức: Nhấn vào một tin tức → Nhấn nút
“Báo cáo” → Hệ thống hiển thị danh sách các báo cáo sai sót liên
quan đến tin tức đó.

Duyệt báo cáo: Admin chọn một báo cáo → Xem chi tiết nội
dung báo cáo (lý do, nguồn...) → Admin cập nhật trạng thái báo cáo
(Pending → Approved/Rejected).

Tìm kiếm tin tức: Nhập tiêu đề → Hệ thống lọc danh sách.

1.3.3 Yêu cầu Phi chức năng (Non – Functional Requirements – NFR)

1.3.3.1 Hiệu năng (Performance – PERF)

NFR – PERF – 01: Thời gian phản hồi trung bình của API phân loại tin tức (không bao gồm bước tính toán giải thích SHAP) phải dưới 5 giây đối với các yêu cầu thông thường trong điều kiện tải bình thường.

NFR – PERF – 02: Thời gian phản hồi khi yêu cầu giải thích SHAP có thể lâu hơn đáng kể (do tính toán phức tạp), hệ thống nên cân nhắc hiển thị chỉ báo đang xử lý cho người dùng nếu thời gian chờ dự kiến kéo dài.

NFR – PERF – 03: Giao diện ứng dụng web (frontend) phải tải nhanh và phản hồi các tương tác người dùng cơ bản (như điều hướng, mở form) một cách mượt mà trên các trình duyệt phổ biến.

NFR – PERF – 04: Hệ thống backend và CSDL cần đảm bảo khả năng xử lý được lượng truy cập đồng thời ở mức cơ bản mà không bị suy giảm hiệu năng nghiêm trọng.

1.3.3.2 Bảo mật (Security – SEC)

NFR – SEC – 01: Mật khẩu người dùng phải được băm (hash) bằng thuật toán an toàn (như Bcryptjs) trước khi lưu vào cơ sở dữ liệu. Mật khẩu gốc không bao giờ được lưu trữ.

NFR – SEC – 02: Hệ thống phải sử dụng JSON Web Tokens (JWT) để quản lý và xác thực phiên làm việc của người dùng sau khi đăng nhập. JWT phải có cơ chế hết hạn (expiration time).

NFR – SEC – 03: Phải triển khai cơ chế phân quyền dựa trên vai trò (role – based access control) để đảm bảo chỉ Quản trị viên mới truy cập được các chức năng quản trị. Middleware phía backend phải kiểm tra quyền truy cập cho các API nhạy cảm.

NFR – SEC – 04: Backend phải thực hiện xác thực và làm sạch (validate and sanitize) dữ liệu đầu vào từ người dùng để giảm thiểu rủi ro từ các tấn công phổ biến như Cross – Site Scripting (XSS) hoặc NoSQL Injection.

NFR – SEC – 05: Trong môi trường triển khai chính thức (production), toàn bộ giao tiếp giữa client và server phải được mã hóa bằng HTTPS.

1.3.3.3 Tính khả dụng (Usability – USA)

NFR – USA – 01: Giao diện người dùng phải được thiết kế trực quan, đơn giản, dễ hiểu và dễ thao tác cho đối tượng người dùng mục tiêu.

NFR – USA – 02: Bố cục, màu sắc, font chữ và các thành phần giao diện phải nhất quán trên toàn bộ ứng dụng.

NFR – USA – 03: Ứng dụng phải có thiết kế đáp ứng (responsive), tự động điều chỉnh giao diện để hiển thị tốt trên các thiết bị có kích thước màn hình khác nhau (desktop, tablet, mobile).

NFR – USA – 04: Hệ thống phải cung cấp các thông báo trạng thái và lỗi rõ ràng, dễ hiểu cho người dùng sau các hành động quan trọng (ví dụ: thông báo thành công, cảnh báo lỗi nhập liệu, thông báo lỗi hệ thống...).

1.3.3.4 Độ tin cậy (Reliability – REL)

NFR – REL – 01: Hệ thống phải hoạt động ổn định, các chức năng cốt lõi phải thực hiện đúng như đặc tả trong điều kiện hoạt động bình thường.

NFR – REL – 02: Hệ thống cần có khả năng xử lý các lỗi dự kiến (ví dụ: mất kết nối tới API phân loại, lỗi CSDL) một cách hợp lý, tránh làm sập toàn bộ ứng dụng và thông báo lỗi phù hợp cho người dùng.

NFR – REL – 03: Kết quả phân loại của mô hình phải nhất quán đối với cùng một đầu vào (trong điều kiện mô hình không thay đổi).

1.3.3.5 Khả năng bảo trì (Maintainability – MNT)

NFR – MNT – 01: Mã nguồn của các thành phần (Frontend React, Backend Node, API Python) cần được tổ chức theo cấu trúc module/component rõ ràng, logic để dễ dàng đọc hiểu, sửa đổi và mở rộng.

NFR – MNT – 02: Sử dụng các quy ước đặt tên (naming conventions) nhất quán cho biến, hàm, class, file.

NFR – MNT – 03: Các đoạn code phức tạp hoặc logic quan trọng cần có comment giải thích phù hợp.

1.3.4 Yêu cầu Dữ liệu (Data Requirements – DAT)

NFR – DAT – 01: Dữ liệu dùng để huấn luyện và đánh giá mô hình phân loại chủ yếu là bộ dữ liệu tổng hợp Sastrawan [19], tập trung vào ngôn ngữ tiếng Anh.

NFR – DAT – 02: Hệ thống sử dụng cơ sở dữ liệu MongoDB để lưu trữ các tập dữ liệu (collections) chính sau:

users: Thông tin tài khoản người dùng (username, email, password hash, họ tên, sđt, giới tính, avatar path, role, timestamps...).

classification_history: Lịch sử các lần phân loại của người dùng (userId, originalText (hoặc textId), predictedLabel, realProbability, fakeProbability, sentimentScore, shapWords (nếu có), timestamp...).

error_reports: Các báo cáo sai sót từ người dùng (userId, historyId (hoặc textId), reason, sourceLink, comment, status ('Pending', 'Approved', 'Rejected'), timestamp...).

NFR – DAT – 03: Cần đảm bảo các ràng buộc về tính duy nhất cho các trường quan trọng như username và email trong collection users.

NFR – DAT – 04: Dữ liệu nhạy cảm (mật khẩu) phải được mã hóa khi lưu trữ. Dữ liệu cá nhân khác cần được xử lý tuân thủ các nguyên tắc bảo mật cơ bản.

1.3.5 Yêu cầu Giao diện (Interface Requirements)

1.3.5.1 Giao diện Người dùng (User Interface – UI)

NFR – UI – 01: Giao diện được phát triển bằng React.js, sử dụng các component từ thư viện React Bootstrap để đảm bảo tính thẩm mỹ, nhất quán và khả năng đáp ứng.

NFR – UI – 02: Giao diện phải tương thích tốt với các trình duyệt web hiện đại phổ biến (Chrome, Firefox, Edge phiên bản mới nhất).

NFR – UI – 03: Cung cấp đầy đủ các màn hình/trang tương ứng với các Use Case đã mô tả (Đăng ký, Đăng nhập, Phân loại, Lịch sử, Hồ sơ, Admin Users, Admin News/Reports, Giới thiệu, Chính sách, Điều khoản...).

NFR – UI – 04: Sử dụng các yếu tố điều khiển (controls) chuẩn và quen thuộc (buttons, input fields, dropdowns, tables, modals...) để người dùng dễ dàng tương tác.

NFR – UI – 05: Thiết kế đáp ứng (responsive) đảm bảo trải nghiệm tốt trên nhiều loại thiết bị.

1.3.5.2 Giao diện Lập trình Ứng dụng (Application Programming Interface – API)

NFR – API – 01: Backend Node.js phải cung cấp tập hợp các RESTful API endpoints cho Frontend sử dụng để thực hiện các chức năng (/api/users/register, /api/users/login, /api/users/profile, /api/news/classify, /api/news/history, /api/report, /api/admin/list,...).

NFR – API – 02: API Phân loại Python/Flask phải cung cấp một endpoint (/classify) nhận dữ liệu văn bản (và cờ ‘explain’) qua phương thức POST và trả về kết quả dự đoán dưới định dạng JSON.

NFR – API – 03: Các API yêu cầu quyền truy cập (xem hồ sơ, phân loại, quản trị) phải được bảo vệ bằng cơ chế xác thực JWT (kiểm tra token trong header Authorization). Middleware phía backend phải thực hiện việc này.

NFR – API – 04: Định dạng trao đổi dữ liệu chính giữa Frontend, Backend Node.js và API Python là JSON.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Tin giả (Fake News)

Sự phát triển nhanh chóng của Internet và tốc độ lan truyền của thông tin trên các nền tảng số tuy mang lại nhiều lợi ích nhưng đồng thời cũng tạo ra một thách thức nghiêm trọng đó là vấn nạn tin giả. Việc nhận biết và đối phó với tin giả trở nên phức tạp hơn bởi chính sự đa dạng trong cách định nghĩa và thể hiện của nó [3]. Về bản chất, tin giả khác với thông tin sai lệch đơn thuần bởi yếu tố chủ đích. Nó là những nội dung được tạo ra nhằm có ý gây hiểu lầm, thao túng hoặc lừa dối người đọc [20]. Thông tin này không chỉ sai sự thật mà còn thường ngụy trang dưới hình thức của một tin tức thật để tăng độ tin cậy. Sự đa dạng của nó còn thể hiện qua nhiều hình thức, từ nội dung châm biếm dễ gây hiểu nhầm đến các thông tin bịa đặt hoàn toàn.

Trong bối cảnh đó, việc phát hiện tin giả tự động trở thành một nhu cầu cấp thiết. Các phương pháp tiếp cận dựa trên Xử lý Ngôn ngữ Tự nhiên (NLP) đóng vai trò trung tâm, bởi chúng có khả năng phân tích các đặc trưng bên trong của văn bản – những dấu hiệu mà con người đôi khi bỏ qua nhưng lại có thể được các mô hình máy học nhận diện. Tin giả ngày càng tinh vi, các nghiên cứu đã chỉ ra một số đặc điểm ngôn ngữ và văn phong thường liên quan [9], [14]: đó có thể là việc sử dụng ngôn ngữ quá giật gân, giàu cảm xúc; lối hành văn thiên vị, thiếu khách quan; sự thiếu váng hoặc mơ hồ trong việc trích dẫn nguồn; hay các tiêu đề câu khách không phản ánh đúng nội dung.

Những đặc trưng về văn phong, cấu trúc và ngữ nghĩa này, dù tinh tế đến đâu, cũng tạo thành các mẫu mà các mô hình học sâu hiện đại, đặc biệt là kiến trúc Transformer với cơ chế chú ý (attention mechanism), có thể học được từ dữ liệu lớn. Khả năng phân tích ngữ cảnh sâu và mối quan hệ giữa các từ giúp các mô hình này không chỉ dựa vào tần suất từ đơn lẻ mà còn nhận biết được các tín hiệu phức tạp hơn liên quan đến tính xác thực của thông tin, mà không cần đến các bộ quy tắc hay danh sách từ khóa định sẵn.

Do đó, để phục vụ cho việc xây dựng và đánh giá mô hình trong khuôn khổ luận văn này, tin giả (Fake News) được định nghĩa hoạt động là:

Những văn bản được trình bày dưới dạng tin tức thời sự nhưng chưa đựng thông tin cốt lõi sai lệch so với sự thật khách quan hoặc được tạo ra/lan truyền với chủ đích rõ ràng nhằm đánh lừa hoặc gây hiểu lầm cho người đọc.

Tuy nhiên, cần nhận thức rõ các thách thức trong việc phát hiện tự động: sự tinh vi của tin giả, sự phụ thuộc vào ngữ cảnh và kiến thức nền, khó khăn trong việc xác định chủ đích, và yêu cầu về dữ liệu huấn luyện lớn, chất lượng cao. Hiểu

rõ những giới hạn này là cần thiết để đánh giá đúng mức hiệu quả của các mô hình được phát triển.

2.2 Xử lý ngôn ngữ tự nhiên

Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing – NLP) là một lĩnh vực liên ngành, nằm ở giao điểm của Khoa học máy tính, Trí tuệ nhân tạo (AI) và Ngôn ngữ học. Mục tiêu cốt lõi của NLP là phát triển các phương pháp và công nghệ cho phép máy tính có khả năng “hiểu”, diễn giải, xử lý, và thậm chí tạo ra ngôn ngữ mà con người sử dụng hàng ngày (dưới cả dạng văn bản và giọng nói) một cách có ý nghĩa và hữu ích.

Tầm quan trọng của NLP ngày càng trở nên rõ rệt trong kỷ nguyên thông tin số. Các ứng dụng của nó hiện diện trong vô số công cụ và dịch vụ thiết yếu, từ máy tìm kiếm, dịch tự động, trợ lý ảo, chatbot, đến các hệ thống phân tích dữ liệu văn bản phức tạp hơn. Đặc biệt, trong bối cảnh đối phó với vấn nạn tin tức giả (fake news) đang lan tràn trên không gian mạng, NLP đóng vai trò trung tâm. Nhiều nghiên cứu trước đây (như sẽ được tổng hợp và phân tích kỹ hơn trong phần các nghiên cứu liên quan) đã tập trung khai thác các kỹ thuật NLP để xây dựng những hệ thống có khả năng tự động nhận diện và phân loại tin tức giả dựa trên các đặc trưng về nội dung, văn phong, hoặc cấu trúc lan truyền.

Tuy nhiên, việc xây dựng các hệ thống NLP hiệu quả luôn đối mặt với những thách thức không nhỏ, xuất phát từ chính bản chất phức tạp, đa nghĩa và luôn biến đổi của ngôn ngữ tự nhiên. Các hiện tượng như từ đồng âm khác nghĩa, cấu trúc câu đa dạng, sự phụ thuộc vào ngữ cảnh, ý nghĩa ẩn dụ, mỉa mai, hay sự xuất hiện liên tục của từ mới, tiếng lóng đòi hỏi các mô hình phải có khả năng nắm bắt ngữ nghĩa sâu sắc và linh hoạt.

Để vượt qua những thách thức này, các phương pháp tiếp cận trong NLP đã không ngừng phát triển. Từ những hệ thống dựa trên luật ngữ pháp và từ điển ban đầu, lĩnh vực này đã chuyển sang các mô hình thống kê và học máy truyền thống. Gần đây nhất, sự trỗi dậy của học sâu (Deep Learning) [21], đặc biệt là các mạng nơ – ron và kiến trúc Transformer [22] với cơ chế tập trung (attention mechanism), đã tạo ra một bước đột phá ngoạn mục. Kiến trúc này là nền tảng cho sự ra đời của các mô hình ngôn ngữ lớn được huấn luyện trước (Pre – trained Language Models – PLMs) như BERT [23] và RoBERTa [24]. Các công trình nghiên cứu tiên tiến trong lĩnh vực phát hiện tin tức giả những năm gần đây, bao gồm cả những nghiên cứu làm nền tảng cho luận văn này, đã chứng minh hiệu quả vượt trội của việc ứng dụng và tinh chỉnh (fine – tuning) các PLMs này. Khả năng nắm bắt ngữ cảnh hai chiều và biểu diễn ngữ nghĩa sâu sắc của chúng đặc biệt phù hợp với việc phân tích và phân loại các văn bản tin tức phức tạp.

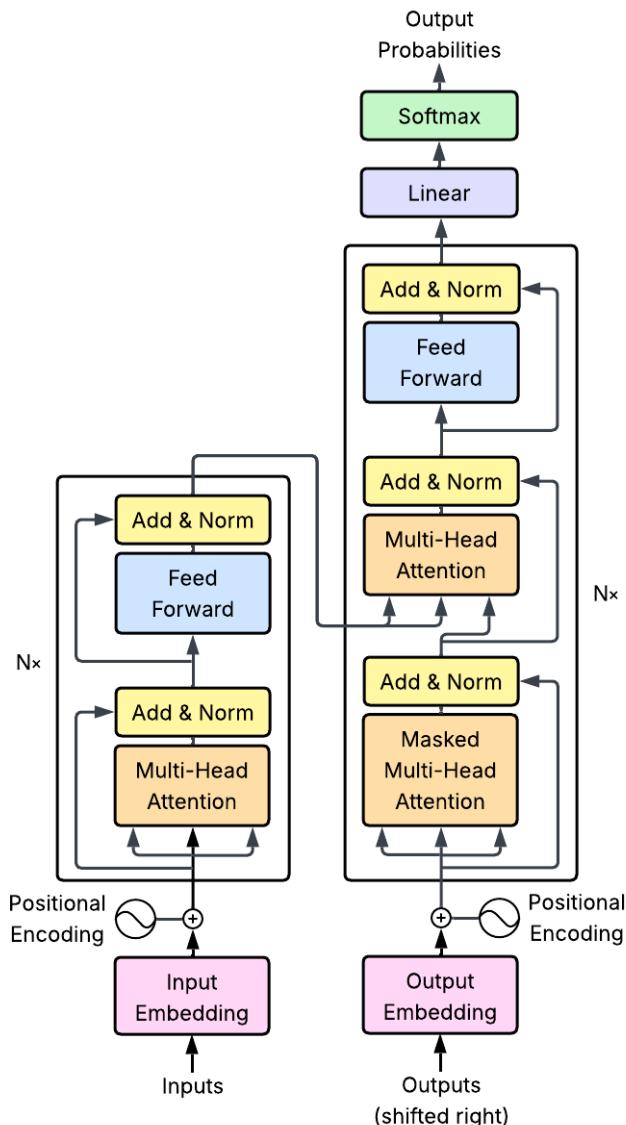
Tóm lại, NLP không chỉ là một lĩnh vực nghiên cứu hấp dẫn mà còn là một công cụ mạnh mẽ để giải quyết các vấn đề thực tiễn quan trọng. Việc hiểu và áp dụng các kỹ thuật NLP hiện đại, đặc biệt là các mô hình dựa trên Transformer, là chìa khóa để xây dựng các giải pháp hiệu quả cho bài toán phát hiện tin tức giả trong luận văn này.

2.3 Transformer

Trong những năm gần đây, Transformers đã trở thành kiến trúc cốt lõi trong các mô hình xử lý ngôn ngữ tự nhiên (NLP) hiện đại, mang lại những cải tiến đáng kể trong nhiều tác vụ như phân loại văn bản, dịch máy, tóm tắt văn bản, và đặc biệt là phát hiện tin tức giả. Mô hình Transformer, được giới thiệu lần đầu bởi Vaswani và cộng sự [22] đã thay thế phần lớn các kiến trúc tuần tự truyền thống như Mạng nơ – ron hồi quy (RNN) hay LSTM trong nhiều tác vụ NLP phức tạp, nhờ vào việc loại bỏ cơ chế hồi quy và thay thế hoàn toàn bằng cơ chế chú ý (attention), đặc biệt là Self – Attention (Tự chú ý). Kiến trúc Transformer được trình bày trong Hình 2.

Thành phần cốt lõi tạo nên sức mạnh của Transformer là cơ chế Multi – head Self – Attention (Tự chú ý Đa đầu). Cơ chế này cho phép mô hình cân nhắc tầm quan trọng của tất cả các từ khác trong chuỗi đầu vào khi tính toán biểu diễn cho một từ cụ thể, bất kể khoảng cách giữa chúng. Điều này khắc phục hạn chế của RNN/LSTM trong việc xử lý các phụ thuộc xa.

Cụ thể, cơ chế self – attention hoạt động dựa trên ba vector được học từ mỗi vector biểu diễn đầu vào (input embedding) của token: Query (Q), Key (K), và Value (V). Ý tưởng là dùng vector Query của một token để “truy vấn” mức độ phù hợp (điểm attention) với các vector Key của tất cả các token khác (và chính nó) trong chuỗi. Điểm attention này sau đó được sử dụng để tính trọng số cho các vector Value tương ứng. Biểu diễn đầu ra cho token đó là tổng có trọng số của các vector Value.



Hình 2. Kiến trúc Transformer

Phiên bản attention được sử dụng phổ biến nhất trong Transformer là Scaled Dot – Product Attention, được tính theo công thức (1):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

$$\text{Softmax}_i = \frac{\exp(z_i)}{\sum(\exp(z_j))} \quad (2)$$

Trong đó:

QK^T là phép nhân ma trận giữa Query và Key chuyển vị, tạo ra ma trận điểm attention thô.

$\sqrt{d_k}$ là căn bậc hai của số chiều của vector Key (và Query), dùng để chuẩn hóa (scale) điểm attention, giúp quá trình huấn luyện ổn định hơn.

softmax là hàm được áp dụng lên ma trận điểm attention đã chuẩn hóa. Hàm softmax (2) biến đổi các điểm số thành một phân phối xác suất, đảm bảo các trọng số attention cộng lại bằng 1. Trọng số này thể hiện mức độ “chú ý” mà mô hình nêu dành cho mỗi token khác khi biểu diễn token hiện tại.

Kết quả cuối cùng được nhân với ma trận Value (V) để tạo ra vector biểu diễn đầu ra có trọng số.

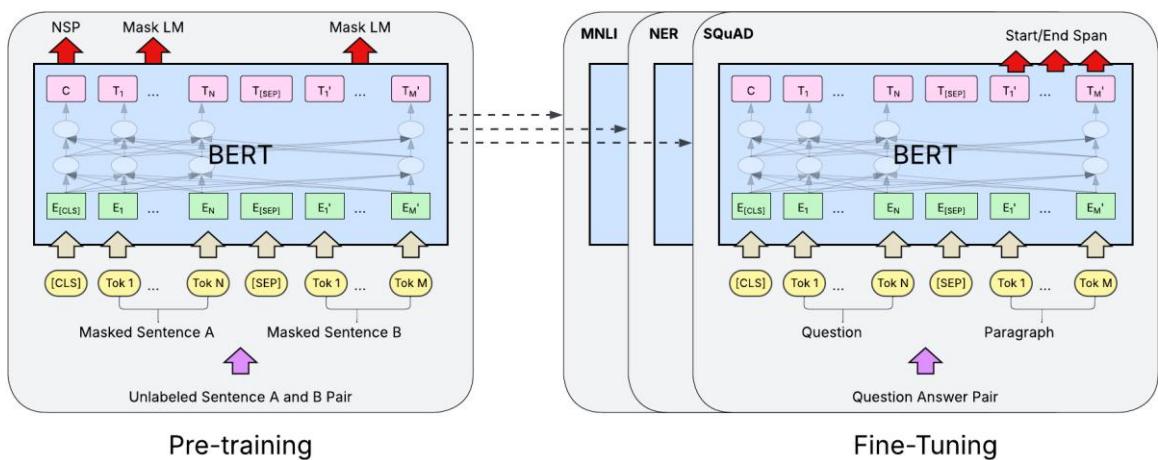
Để mô hình có thể học được các loại quan hệ khác nhau giữa các từ (ví dụ: quan hệ ngữ pháp, quan hệ ngữ nghĩa...), Transformer sử dụng cơ chế Multi – head Attention. Thay vì chỉ tính attention một lần, nó thực hiện phép chiếu tuyến tính (linear projection) để tạo ra nhiều bộ (h heads) Q, K, V khác nhau từ đầu vào ban đầu. Scaled Dot – Product Attention được tính toán song song cho từng “head”. Kết quả đầu ra của các head này sau đó được ghép lại (concatenate) và chiếu tuyến tính một lần nữa để ra kết quả cuối cùng của lớp Multi – head Attention. Việc này cho phép mô hình cùng lúc tập trung vào các không gian biểu diễn con khác nhau (different representation subspaces) tại các vị trí khác nhau.

Ngoài Multi – head Self – Attention, mỗi lớp trong bộ mã hóa (Encoder) và giải mã (Decoder) của Transformer còn chứa một Mạng nơ – ron truyền thẳng theo vị trí (Position – wise Feed – Forward Network). Mạng này bao gồm hai lớp tuyến tính với hàm kích hoạt ReLU ở giữa, được áp dụng độc lập cho từng vị trí trong chuỗi. Bên cạnh đó, kiến trúc Transformer còn sử dụng Kết nối dư (Residual Connections) xung quanh mỗi thành phần (Attention và Feed – Forward) và Chuẩn hóa lớp (Layer Normalization) sau mỗi kết nối dư để hỗ trợ việc huấn luyện các mô hình sâu và ổn định hơn. Do Transformer không có tính tuần tự, nó cần thông tin về vị trí của token, thông tin này thường được cung cấp thông qua Mã hóa vị trí (Positional Encoding) cộng trực tiếp vào embedding đầu vào.

2.4 BERT

BERT (Bidirectional Encoder Representations from Transformers) được Devlin và cộng sự [23] giới thiệu, đánh dấu một bước đột phá trong NLP. Điểm cốt lõi làm nên thành công của BERT là khả năng học biểu diễn ngữ cảnh hai chiều (bidirectional context learning) một cách sâu sắc. Không giống các mô hình trước đó thường chỉ xem xét ngữ cảnh bên trái hoặc bên phải, BERT sử dụng cơ chế Masked Language Model (MLM) trong quá trình tiền huấn luyện (pre – training) để cho phép mô hình hiểu ý nghĩa của một từ dựa trên đồng thời cả ngữ cảnh đứng trước và đứng sau nó trong câu.

Về mặt kiến trúc, BERT được xây dựng chỉ dựa trên phần bộ mã hóa (Encoder) của kiến trúc Transformer gốc. Mô hình BERT bao gồm nhiều lớp mã hóa Transformer giống nhau được xếp chồng lên nhau (BERT – base có 12 lớp, BERT – large có 24 lớp). Mỗi lớp này chứa hai thành phần chính là Multi – head Self – Attention và Position – wise Feed – Forward Network, cùng với các kết nối dư và chuẩn hóa lớp. Cơ chế Multi – head Self – Attention cho phép BERT cân nhắc mối quan hệ giữa tất cả các cặp token trong chuỗi đầu vào. Kiến trúc và quy trình Pre – training/Fine – tuning của BERT được minh họa chi tiết trong Hình 3.



Hình 3. Kiến trúc BERT và quy trình Pre – training/Fine – tuning

Đầu vào của BERT được xử lý đặc biệt để tích hợp nhiều loại thông tin. Nó là tổng hợp của ba loại embedding:

Token Embeddings: Biểu diễn các từ hoặc các đơn vị nhỏ hơn (sub – token) theo kỹ thuật WordPiece.

Segment Embeddings: Dùng để phân biệt giữa hai câu khi đầu vào là một cặp câu, gán một embedding cho câu A và một embedding khác cho câu B.

Positional Embeddings: Cung cấp thông tin về vị trí của mỗi token trong chuỗi, do kiến trúc Transformer vốn không có tính tuần tự.

BERT được tiền huấn luyện trên một kho dữ liệu văn bản cực lớn với hai nhiệm vụ không giám sát đồng thời:

Masked Language Model (MLM): Che (mask) ngẫu nhiên khoảng 15% số token trong chuỗi đầu vào và yêu cầu mô hình dự đoán các token gốc đã bị che dựa vào ngữ cảnh xung quanh (cả trái và phải). Đây là chìa khóa cho việc học biểu diễn hai chiều.

Next Sentence Prediction (NSP): Nhận đầu vào là một cặp câu (A, B) và yêu cầu mô hình dự đoán xem câu B có phải là câu kế tiếp thực sự của câu A trong văn bản gốc hay không. Nhiệm vụ này giúp mô hình hiểu mối quan hệ giữa các câu.

Sau khi tiền huấn luyện, BERT có thể được tinh chỉnh cho các tác vụ NLP cụ thể như phân loại văn bản, nhận dạng thực thể, trả lời câu hỏi, bao gồm cả phát hiện tin tức giả, bằng cách thêm một lớp đầu ra đơn giản và huấn luyện tiếp trên dữ liệu có nhãn của tác vụ đó.

2.5 RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) được Liu và cộng sự [24] giới thiệu, không phải là một kiến trúc mới hoàn toàn mà là một phiên bản tối ưu hóa mạnh mẽ phương pháp tiền huấn luyện của BERT, nhằm cải thiện hiệu suất trên các tác vụ NLP. RoBERTa giữ nguyên kiến trúc cơ bản dựa trên bộ mã hóa Transformer như BERT nhưng thực hiện những thay đổi quan trọng trong chiến lược huấn luyện và chuẩn bị dữ liệu.

Những cải tiến chính của RoBERTa so với BERT bao gồm:

Huấn luyện trên dữ liệu lớn hơn nhiều: RoBERTa được huấn luyện trên một bộ dữ liệu lớn hơn đáng kể so với BERT.

Thời gian huấn luyện dài hơn và batch size lớn hơn: Tăng cường quá trình huấn luyện để mô hình hội tụ tốt hơn.

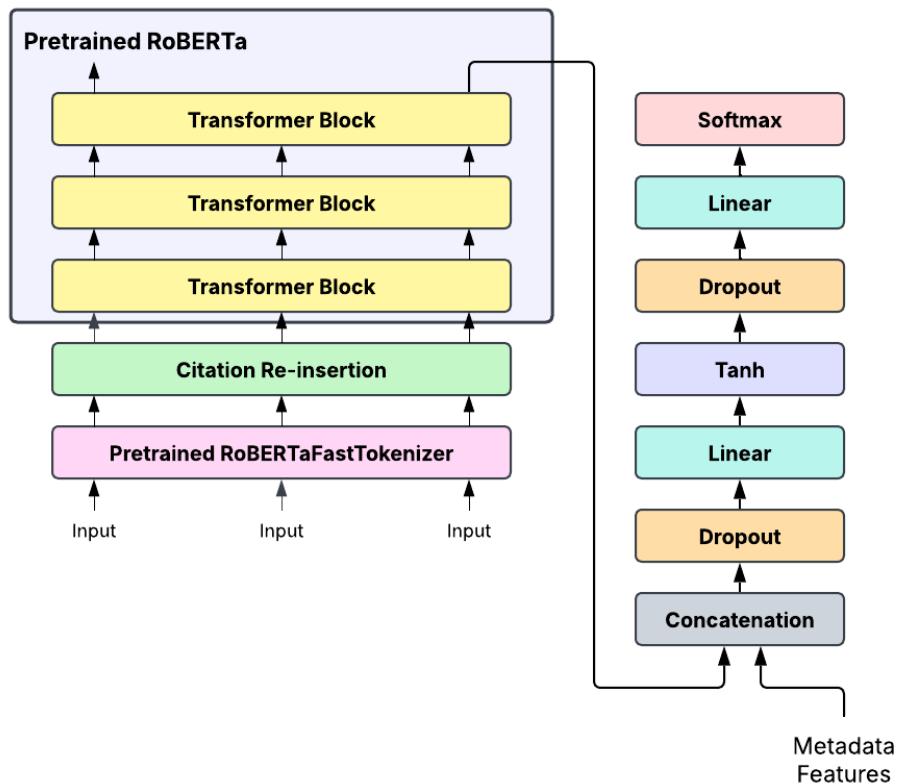
Loại bỏ nhiệm vụ Next Sentence Prediction (NSP): Các tác giả RoBERTa phát hiện ra rằng việc loại bỏ nhiệm vụ NSP và chỉ huấn luyện với các chuỗi dài đầy đủ từ một hoặc nhiều tài liệu có thể cải thiện hiệu suất trên các tác vụ downstream.

Sử dụng Dynamic Masking: Thay vì thực hiện che token một lần duy nhất trong quá trình tiền xử lý dữ liệu như BERT, RoBERTa sử dụng kỹ thuật che động (dynamic masking), nghĩa là mẫu token bị che được tạo ra khác nhau trong mỗi epoch huấn luyện khi dữ liệu được đưa vào mô hình.

Sử dụng Byte – Pair Encoding (BPE) cấp độ byte: Thay vì WordPiece, RoBERTa sử dụng BPE với bộ từ vựng lớn hơn và dựa trên byte, giúp xử lý các từ chưa biết hiệu quả hơn và không cần các ký tự đặc biệt cho từ không xác định.

Do loại bỏ NSP, đầu vào của RoBERTa trong quá trình fine – tuning thường chỉ cần mã định danh của các token (input IDs) và attention mask (để chỉ định token nào cần chú ý, token nào là padding), không còn cần segment embeddings. Nhờ những tối ưu hóa này, RoBERTa thường xuyên đạt được kết quả vượt trội

hơn BERT trên nhiều bộ dữ liệu và tác vụ đánh giá NLP chuẩn. Kiến trúc của RoBERTa áp dụng trong nghiên cứu này được trình bày trong Hình 4.



Hình 4. Kiến trúc RoBERTa

2.6 SHAP (SHapley Additive exPlanations)

Trong bối cảnh các mô hình học sâu, đặc biệt là các mô hình Transformer phức tạp như BERT và RoBERTa, thường được xem như các “hộp đen” (black boxes), việc hiểu và diễn giải tại sao mô hình đưa ra một dự đoán cụ thể trở nên rất quan trọng, đặc biệt trong các lĩnh vực nhạy cảm như phát hiện tin giả. SHAP (SHapley Additive exPlanations) là một phương pháp giải thích mạnh mẽ và thống nhất, dựa trên lý thuyết trò chơi (game theory) và giá trị Shapley, nhằm giải thích đầu ra của bất kỳ mô hình học máy nào.[25]

Ý tưởng cốt lõi của SHAP là xem xét đầu ra dự đoán của mô hình cho một mẫu dữ liệu cụ thể (ví dụ: một bài báo) như là kết quả của một “trò chơi” mà các “người chơi” chính là các đặc trưng đầu vào (input features). Trong bài toán xử lý ngôn ngữ tự nhiên, các “đặc trưng” này thường tương ứng với các token (từ hoặc mảng từ) trong văn bản đầu vào. SHAP tính toán giá trị đóng góp (contribution) riêng lẻ của từng token vào dự đoán cuối cùng so với một dự đoán cơ sở (base prediction - thường là dự đoán trung bình trên tập dữ liệu). Giá trị đóng góp này được gọi là SHAP value.

Một SHAP value dương cho một token đối với một lớp dự đoán (lớp “Fake”) có nghĩa là token đó đẩy dự đoán về phía lớp “Fake”. Ngược lại, SHAP value âm có nghĩa là token đó đẩy dự đoán ra xa lớp “Fake” (về phía lớp “Real”). Độ lớn của SHAP value thể hiện mức độ ảnh hưởng của token đó.

Ưu điểm chính của SHAP là nó cung cấp các đảm bảo về mặt lý thuyết như tính nhất quán (consistency) và tính chính xác cục bộ (local accuracy), đảm bảo rằng tổng các SHAP value của tất cả các token cộng với giá trị dự đoán cơ sở sẽ bằng đúng dự đoán thực tế của mô hình cho mẫu dữ liệu đó. Bằng cách xác định các token có SHAP value dương cao nhất cho lớp “Fake”, chúng ta có thể hiểu được những từ ngữ nào trong văn bản đang khiến mô hình nghiêng về dự đoán là tin giả, từ đó tăng tính minh bạch và tin cậy cho mô hình. Trong nghiên cứu này, SHAP được sử dụng để cung cấp thông tin giải thích này cho người dùng.

2.7 Fine – tuning và Vai trò của Lớp Softmax trong Phân loại

Các mô hình ngôn ngữ lớn như BERT và RoBERTa thể hiện sức mạnh vượt trội nhờ quá trình tiền huấn luyện trên lượng dữ liệu khổng lồ, giúp chúng học được các biểu diễn ngôn ngữ sâu sắc và giàu ngữ nghĩa. Tuy nhiên, để áp dụng chúng vào một tác vụ cụ thể như phân loại tin giả, cần phải thực hiện bước tinh chỉnh (fine – tuning).

Để thực hiện phân loại, người ta thường thêm một hoặc một vài lớp đơn giản vào phía trên cấu trúc Transformer của mô hình pre – trained. Một kiến trúc phổ biến cho bài toán phân loại văn bản là sử dụng vector biểu diễn đầu ra của token đặc biệt [CLS] (thường đứng đầu mỗi chuỗi đầu vào trong BERT/RoBERTa) làm đại diện cho toàn bộ chuỗi. Vector này sau đó được đưa qua một lớp tuyến tính (Linear layer, hay còn gọi là Fully Connected layer) để chiếu nó vào không gian có số chiều bằng số lớp cần phân loại (trong trường hợp này là 2 lớp: tin giả và tin thật).

Lớp cuối cùng và đóng vai trò quyết định trong việc đưa ra dự đoán xác suất cho mỗi lớp chính là lớp Softmax. Hàm softmax nhận đầu vào là các điểm số thô (logits) từ lớp tuyến tính và chuyển đổi chúng thành một phân phối xác suất hợp lệ, sao cho tổng xác suất của tất cả các lớp bằng 1. Công thức của hàm softmax cho lớp i trong tổng số C lớp (3):

$$P(\mathbf{y} = \mathbf{i} | \mathbf{x}) = \text{Softmax}_i = \frac{\exp(z_i)}{\sum_{j=1}^C (\exp(z_j))} \quad (3)$$

Trong đó:

z_i là logit (điểm số thô) cho lớp i .

$\exp(z_i)$ là hàm mũ của logit lớp i .

Luận văn: Transformer để phân loại tin giả

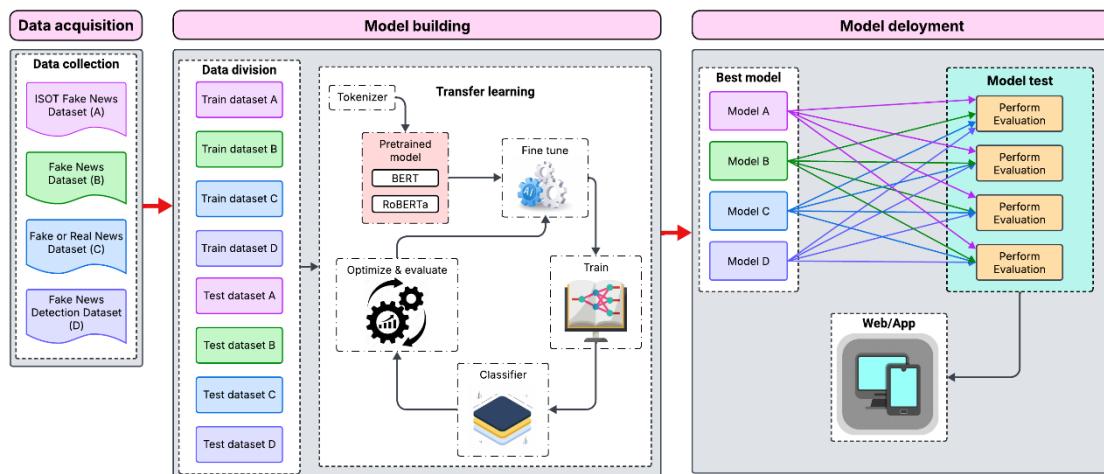
Mẫu số là tổng hàm mũ của tất cả các logit từ lớp 1 đến lớp C.

$P(y=i|x)$ là xác suất dự đoán rằng đầu vào x thuộc về lớp i.

Sau khi có được các xác suất này, mô hình thường sẽ dự đoán lớp có xác suất cao nhất là nhãn cho đầu vào. Ví dụ, nếu $P(\text{Giả}) > P(\text{Thật})$, tin tức sẽ được phân loại là giả. Lớp softmax đóng vai trò then chốt trong việc diễn giải đầu ra của mạng nơ – ron thành các dự đoán xác suất rõ ràng, phù hợp cho bài toán phân loại.

CHƯƠNG 3: PHƯƠNG PHÁP THỰC NGHIỆM

3.1 Tổng quan



Hình 5. Phương pháp tiếp cận được đề xuất trong Phân loại tin tức thật giả

Quy trình phát triển hệ thống Phân loại tin tức thật giả bao gồm ba giai đoạn chính: tìm kiếm dữ liệu, xây dựng và huấn luyện mô hình, triển khai mô hình.

Đầu tiên, dữ liệu được thu thập từ bài báo “Detection of fake news using deep learning CNN–RNN based methods”[21]. Gồm bốn tập dữ liệu ISOT Fake News Dataset, Fake News Dataset, Fake or Real News Dataset và Fake News Detection Dataset. Các tập dữ liệu này cung cấp dữ liệu phong phú và đa dạng về tin tức thật và giả, được tổng hợp từ nhiều bài nghiên cứu trước đó, giúp đánh giá khả năng của mô hình trong việc xử lý dữ liệu lớn và nhỏ.

Dữ liệu đã được các tác giả của bài báo tổng hợp và tiền xử lý kỹ lưỡng, bao gồm làm sạch dữ liệu (loại bỏ dữ liệu trống, trùng lặp), tăng cường dữ liệu (sinh thêm mẫu tin tức), và tiền xử lý (loại bỏ ký tự đặc biệt, từ dừng, tách từ ghép), và cân bằng dữ liệu để giảm thiểu sự chênh lệch giữa số lượng tin tức thật và giả. Ngoài ra, dữ liệu còn được chuẩn hóa và kiểm tra tính nhất quán nhằm đảm bảo chất lượng đầu vào tốt nhất cho quá trình huấn luyện mô hình. Quá trình tổng hợp và xử lý dữ liệu này đóng vai trò quan trọng trong việc đảm bảo rằng mô hình được huấn luyện trên các tập dữ liệu đáng tin cậy, từ đó nâng cao độ chính xác và khả năng tổng quát hóa của mô hình trong quá trình triển khai thực tế.

Tiếp theo là quá trình xây dựng và huấn luyện mô hình. Mô hình được phát triển dựa trên các kiến trúc transformers như BERT và RoBERTa, tận dụng khả năng hiểu ngữ cảnh và biểu diễn ngôn ngữ để phân loại tin tức thật – giả. Hiệu suất của mô hình được đánh giá dựa trên các chỉ số tương tự như khi huấn luyện để đảm bảo rằng mô hình không chỉ học tốt trên tập dữ liệu huấn luyện mà còn hoạt động hiệu quả trên tập dữ liệu kiểm tra. Mỗi tập dữ liệu sẽ được sử dụng để huấn

luyện riêng biệt các mô hình (BERT, RoBERTa). Mô hình có hiệu suất cao nhất trên mỗi tập dữ liệu sẽ được chọn.

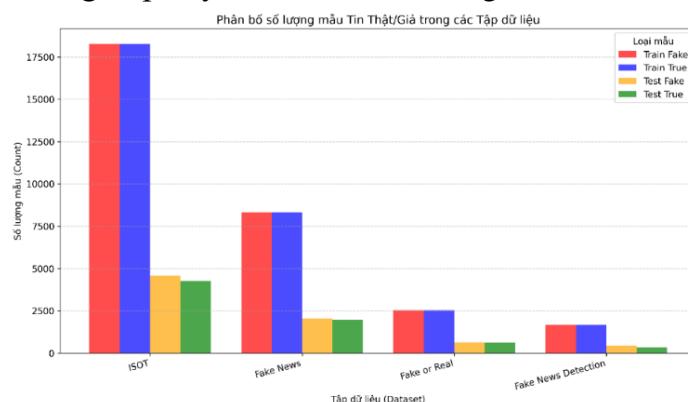
Sau khi lựa chọn bốn mô hình tốt nhất từ bốn tập dữ liệu, nghiên cứu tiến hành kiểm tra chéo, trong đó mỗi mô hình được đánh giá trên tập kiểm tra của ba tập dữ liệu còn lại. Mục tiêu của bước này là đánh giá khả năng tổng quát hóa của từng mô hình khi áp dụng vào dữ liệu chưa từng thấy. Điều này giúp xác định mô hình nào hoạt động ổn định nhất trên nhiều nguồn dữ liệu khác nhau và chọn làm mô hình đề xuất. Mô hình đề xuất này sau đó có thể được tích hợp vào các ứng dụng thực tế như website hoặc ứng dụng di động để hỗ trợ người dùng phân loại tin tức.

Quy trình phát triển này không chỉ đảm bảo tính chính xác và hiệu quả của hệ thống phân loại tin tức thật giả, mà còn tạo ra một mô hình linh hoạt và mạnh mẽ, có khả năng áp dụng vào các bài toán phân loại khác nhau trong tương lai. Nghiên cứu tương lai sẽ tập trung vào việc tinh chỉnh kiến trúc và khám phá các kỹ thuật mới nhằm nâng cao hiệu quả nhận diện, đồng thời giảm thiểu chi phí tính toán và số lượng tham số, cùng khả năng mở rộng dễ dàng cho các tác vụ phân loại khác nhau.

3.2 Phân loại tin tức thật giả

3.2.1 Tập dữ liệu

Trong đề tài nghiên cứu này, mô hình được huấn luyện và đánh giá chủ yếu dựa trên bộ dữ liệu tổng hợp do Sastrawan và cộng sự công bố [19]. Bộ dữ liệu này được xây dựng bằng cách thu thập, làm sạch, tiền xử lý và có thể bao gồm cả tăng cường dữ liệu từ bốn tập dữ liệu gốc phổ biến trong lĩnh vực phát hiện tin giả là: ISOT Fake News Dataset[9], Fake News Dataset[12], Fake or Real News Dataset[17] và Fake News Detection Dataset[13]. Việc tổng hợp và xử lý này nhằm tạo ra một nguồn dữ liệu đa dạng, phù hợp cho việc đánh giá các mô hình trên nhiều loại tin tức khác nhau. Biểu đồ Phân bố số lượng mẫu tin thật/giả trong bộ dữ liệu tổng hợp này được thể hiện trong Hình 6.



Hình 6. Biểu đồ phân bố số lượng mẫu tin thật/giả của các tập dữ liệu

Số lượng chi tiết cho các tập huấn luyện (train) và kiểm thử (test) của từng bộ dữ liệu (sau khi được xử lý bởi Sastrawan và cộng sự) được trình bày trong Bảng 1.

Bảng 1. Số lượng chi tiết cho các tập huấn luyện (train) và kiểm thử (test) được sử dụng để nghiên cứu

Tập dữ liệu	Số lượng tập huấn luyện			Số lượng tập kiểm thử			Tổng số lượng	Ghi chú xử lý
	Fake	True	Tổng	Fake	True	Tổng		
ISOT Fake News Dataset [19]	18,266	18,266	36,538	4,585	4,269	8,856	45,394	Đã được làm sạch, tăng cường, tiền xử lý
Fake News Dataset [19]	8,323	8,323	16,660	2,056	1,979	4,044	20,704	Đã được làm sạch, tăng cường, tiền xử lý
Fake or Real News Dataset [19]	2,536	2,536	5,076	635	625	1,262	6,338	Đã được làm sạch, tăng cường, tiền xử lý
Fake News Detection Dataset [19]	1,670	1,670	3,340	450	348	798	4,138	Đã được làm sạch, tăng

								cường, tiền xử lý
--	--	--	--	--	--	--	--	-------------------

3.2.2 Phân tích dữ liệu

Để hiểu rõ hơn về đặc điểm, cấu trúc chủ đề và các đặc trưng ngôn ngữ tiềm ẩn trong các tập dữ liệu được sử dụng, nghiên cứu đã tiến hành các bước phân tích khám phá dữ liệu (EDA) sử dụng kỹ thuật mô hình hóa chủ đề BERTopic và phân tích tần suất từ bằng CountVectorizer.

3.2.2.1 Phân tích Chủ đề bằng BERTopic

Kỹ thuật mô hình hóa chủ đề BERTopic được áp dụng riêng biệt trên từng tập dữ liệu con (ISOT Fake News Dataset, Fake News Dataset, Fake Or Real News Dataset, Fake News Detection Dataset) trong bộ dữ liệu tổng hợp [19]. Mục tiêu là tự động khám phá các chủ đề tiềm ẩn, xác định các cụm nội dung chính và đánh giá sự phân bố của tin tức thật/giả trong các chủ đề đó. Kết quả từ mô hình BERTopic bao gồm các Topic ID và danh sách từ khóa tiêu biểu cho mỗi chủ đề. Để thuận tiện cho việc phân tích và diễn giải, nghiên cứu đã tiến hành đặt tên gợi nhớ cho từng chủ đề (như trình bày trong cột ‘Tên Chủ đề (Diễn giải)’ ở các Bảng 2 – 5). Quá trình đặt tên này chủ yếu dựa trên việc phân tích ý nghĩa và mối liên hệ ngữ nghĩa của các từ khóa tiêu biểu do BERTopic cung cấp. Đồng thời, nghiên cứu này cũng đã kiểm tra nội dung của một số văn bản đại diện thuộc mỗi chủ đề để đảm bảo tính chính xác và phù hợp của tên gọi được diễn giải, kết hợp với kiến thức nền về các sự kiện, nhân vật được đề cập. Kết quả phân loại chủ đề chi tiết cho từng tập dữ liệu được trình bày trong các Bảng 2, Bảng 3, Bảng 4, Bảng 5.

Bảng 2. Kết quả phân loại chủ đề trên tập dữ liệu ISOT Fake News

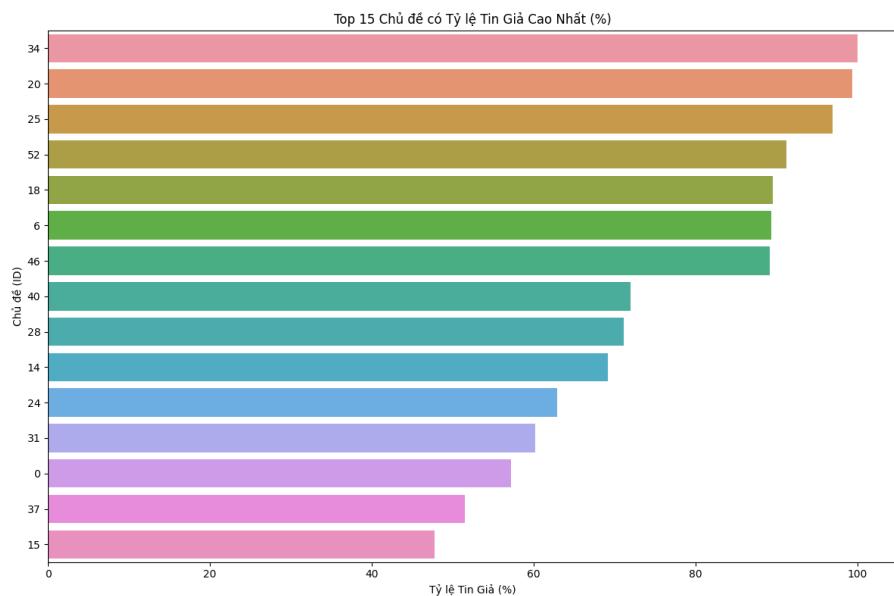
Topic (ID)	ISOT Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Diễn giải)	Tỷ lệ gi
-1	-1_trump_say_clinton_president	Ngoại lai (BERTopic không phân loại được)	
0	0_trump_say_republican_president	Chính trị Hoa Kỳ (Trump & Đảng Cộng hòa)	57.19%

Topic (ID)	ISOT Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Điễn giải)	Tỷ lệ giả
1	1_korea_north_korean_nuclear	Vũ khí hạt nhân Triều Tiên	9.67%
2	2_merkel_germany_german_party	Chính trị Đức & Châu Âu (Merkel)	21.07%
3	3_kurdish_iraq_iraqi_turkey	Xung đột Iraq/Syria/Thổ Nhĩ Kỳ (Người Kurd)	4.33%
4	4_pakistan_mugabe_zimbabwe_mnangagwa	Chính trị Pakistan & Zimbabwe (Có vẻ gồm 2 khu vực)	2.38%
5	5_eu_brexit_britain_may	Brexit (Anh & EU)	0.43%
6	6_gun_officer_shoot_police	Bạo lực súng đạn & Cảnh sát (Hoa Kỳ)	89.35%
7	7_israel_jerusalem_palestinian_israeli	Xung đột Israel – Palestine	15.05%
8	8_saudi_lebanon_arabia_hariri	Chính trị Trung Đông (Saudi, Lebanon, Qatar)	8.60%
9	9_china_taiwan_xi_chinese	Quan hệ Trung Quốc – Đài Loan	8.26%
10	10_london_britain_british_uk	Tin tức Nội địa Anh (London, Manchester)	27.33%
11	11_myanmar_rohingya_bangladesh_rakhine	Khủng hoảng Rohingya (Myanmar)	0%
12	12_catalan_spain_catalonia_independence	Độc lập Catalonia (Tây Ban Nha)	0%
13	13_venezuela_maduro_venezuelan_colombia	Khủng hoảng Venezuela	12.50%
14	14_abortion_parenthood_woman_plan	Tranh cãi Phá thai (Hoa Kỳ)	69.23%
15	15_syria_syrian_assad_chemical	Nội chiến Syria (Assad & Vũ khí hóa học)	47.81%
16	16_ban_order_court_travel	Sắc lệnh Cấm Nhập cảnh (Hoa Kỳ)	25.51%

Topic (ID)	ISOT Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Điễn giải)	Tỷ lệ giả
17	17_cuba_cuban_havana_castro	Quan hệ Hoa Kỳ – Cuba	33.00%
18	18_nfl_anthem_player_kaepernick	Biểu tình Chào cờ NFL (Kaepernick)	89.55%
19	19_france_macron_french_paris	Chính trị Pháp (Macron)	25.38%
20	20_boiler_acr_room_pm	Show ‘Boiler Room’ (Radio ACR)	99.42%
21	21_duterte_philippine_drug_manila	Philippines (Duterte & Cuộc chiến ma túy)	2.76%
22	22_nato_defense_european_alliance	NATO & An ninh Châu Âu	12.06%
23	23_yemen_houthis_saudi_coalition	Chiến tranh Yemen	5.80%
24	24_moore_alabama_roy_jones	Bầu cử Thượng viện Alabama (Roy Moore)	62.88%
25	25_bundy_oregon_finicum_federal	Vụ Chiếm đóng Oregon (Bundy)	96.95%
26	26_irland_irish_northern_border	Biên giới Ireland & Brexit	2.34%
27	27_afghanistan_afghan_taliban_kabul	Tình hình Afghanistan (Taliban)	6.78%
28	28_flint_water_snyder_michigan	Khủng hoảng Nước sạch Flint (Michigan)	71.19%
29	29_illinois_rauner_budget_governor	Chính trị & Ngân sách Illinois	7.63%
30	30_australia_turnbull_australian_sydney	Chính trị Úc (Turnbull)	6.03%
31	31_blasio_mayor_city_de	Chính trị New York City (De Blasio)	60.18%
32	32_odinga_kenya_kenyatta_election	Bầu cử Kenya (Odinga & Kenyatta)	0.93%

Topic (ID)	ISOT Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Điễn giải)	Tỷ lệ giả
33	33_puerto_rico_debt_creditor	Khủng hoảng Nợ Puerto Rico	3.81%
34	34_wire_patrick_henningsen_episode	Nội dung từ 21st Century Wire (Henningsen)	100%
35	35_temer_brazil_brazilian_lula	Chính trị Brazil (Temer, Lula, Tham nhũng)	4.85%
36	36_irma_hurricane_florida_storm	Bão Irma (Florida)	10.10%
37	37_pope_francis_vatican_catholic	Giáo hoàng Francis & Vatican	51.52%
38	38_indonesia_indonesian_jakarta_bali	Tin tức Indonesia	1.01%
39	39_egypt_egyptian_cairo_sisi	Tình hình Ai Cập (Sisi)	6.38%
40	40_iran_iranian_hostage_prisoner	Ván đè Tù nhân/Con tin Iran	72.04%
41	41_trudeau_canada_canadian_morneau	Chính trị Canada (Trudeau)	35.96%
42	42_california_brown_state_voter	Chính trị California	40.96%
43	43_nafta_trade_mexico_canada	Đàm phán NAFTA/USMCA	2.47%
44	44_vietnam_china_vietnamese_sea	Biển Đông (Việt Nam – Trung Quốc)	15.58%
45	45_zuma_anc_ramaphosa_africa	Chính trị Nam Phi (ANC)	1.32%
46	46_pelosi_nancy_fence_democrat	Chính trị Hoa Kỳ (Nancy Pelosi & Dân chủ)	89.19%
47	47_puerto_rico_maría_hurricane	Bão Maria (Puerto Rico)	38.03%
48	48_thailand_thai_bangkok_yingluck	Chính trị Thái Lan	2.90%

Topic (ID)	ISOT Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Điễn giải)	Tỷ lệ giả
49	49_italy_berlusconi_pd_star	Chính trị Ý	1.49%
50	50_zika_virus_microcephaly_cdc	Virus Zika & Y tế công cộng	17.74%
51	51_cambodia_hun_sen_kem	Chính trị Campuchia (Hun Sen)	3.23%
52	52_baltimore_gray_police_freddie	Vụ Freddie Gray (Baltimore)	91.23%
53	53_nigeria_boko_haram_buhari	Boko Haram (Nigeria)	0%
54	54_newzealand_labour_new_peter	Chính trị New Zealand	0%
55	55_christie_jersey_governor_chris	Chính trị New Jersey (Chris Christie)	44.44%
56	56_migrant_libya_italy_libyan	Khủng hoảng Di cư (Địa Trung Hải)	1.92%



Hình 7. Biểu đồ Top 15 chủ đề của tập ISOT có tỷ lệ tin giả cao nhất phân tích từ BERTopic

Phân tích kết quả BERTopic trên tập ISOT (Bảng 2, Hình 7): Tập dữ liệu ISOT bao phủ rộng rãi các vấn đề, đặc biệt là các vấn đề chính trị quốc tế và Hoa Kỳ. Phân tích tỷ lệ tin giả cho thấy các chủ đề liên quan đến các nguồn tin

thay thế hoặc các sự kiện gây tranh cãi như Topic 34 (100% – Nội dung từ 21st Century Wire), Topic 20 (99.42% – Show ‘Boiler Room’), Topic 25 (96.95% – Vụ Chiếm đóng Oregon) gần như hoàn toàn là tin giả. Tỷ lệ giả cũng rất cao ở các chủ đề xã hội và chính trị Hoa Kỳ như Topic 18 (89.55% – Biểu tình Chào cờ NFL), Topic 46 (89.19% – Chính trị Hoa Kỳ – Pelosi & Dân chủ), Topic 6 (89.35% – Bạo lực súng đạn & Cảnh sát), Topic 28 (71.19% – Khủng hoảng Nước sạch Flint), Topic 14 (69.23% – Tranh cãi Phá thai), Topic 24 (62.88% – Bầu cử Alabama) và Topic 31 (60.18% – Chính trị NYC). Chủ đề về vấn đề Tù nhân/Con tin Iran (Topic 40) cũng có tỷ lệ giả cao (72.04%). Ngược lại, các chủ đề tin tức quốc tế như Khủng hoảng Rohingya (Topic 11), Độc lập Catalonia (Topic 12) hay Chính trị New Zealand (Topic 54) không chứa tin giả nào trong tập dữ liệu này. Nhóm ngoại lai (Topic – 1) với từ khóa trump, say, clinton, president phản ánh các nội dung chính trị chung, mô hình BERTopic khó phân loại.

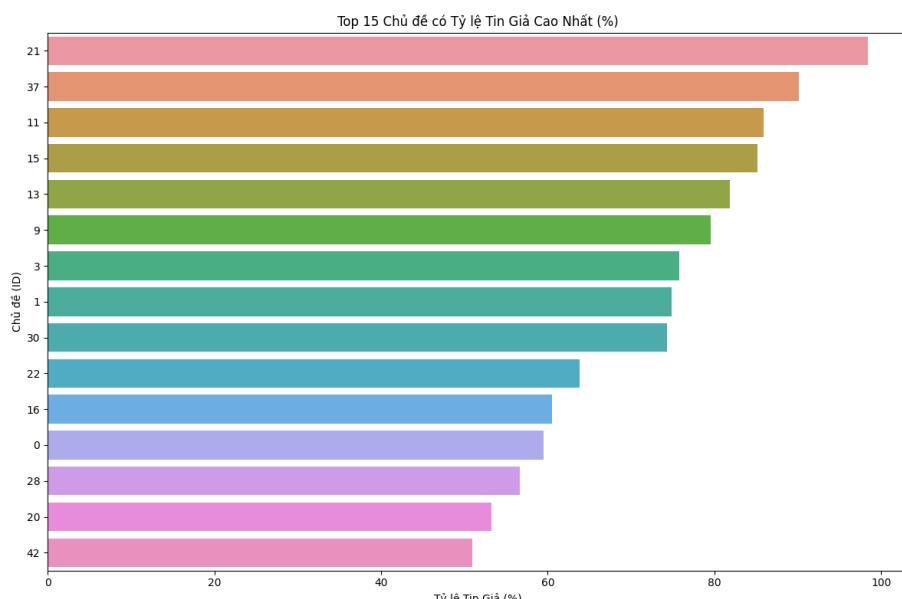
Bảng 3. Kết quả phân loại chủ đề trên tập dữ liệu Fake News

Topic (ID)	Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Diễn giải)	Tỷ lệ giả
- 1	_1_trump_say_clinton _president	Ngoại lai (BERTopic không phân loại được)	
0	0_clinton_trump_hillary_email	Bầu cử Hoa Kỳ 2016 (Clinton & Trump, Emails)	59.53%
1	1_russia_russian_syria_syrian	Nga & Cuộc chiến Syria	74.87%
2	2_film_mr_show_like	Phim ảnh / Chương trình TV	17.57%
3	3_cancer_food_drug_study	Nghiên cứu Ung thư & Sức khỏe (Thực phẩm, Thuốc)	75.81%
4	4_muslim_refugee_say_islam ic	Người tị nạn Hồi giáo & Hồi giáo	36.25%
5	5_border_immigration_mexico _mexican	Biên giới Hoa Kỳ – Mexico & Nhập cư	14.82%
6	6_police_officer_say_mr	Cảnh sát & Thực thi pháp luật	27.02%
7	7_health_obamacare_care_ins urance	Y tế Hoa Kỳ (Obamacare)	27.33%

Topic (ID)	Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Diễn giải)	Tỷ lệ giả
8	8_game_player_team_season	Thể thao (Đội tuyển, Mùa giải)	13.42%
9	9_gold_market_rate_bank	Tài chính & Thị trường (Vàng, Ngân hàng)	79.58%
10	10_israel_palestinian_israeli_jeerusalem	Xung đột Israel – Palestine	43.7%
11	11_pipeline_dakota_stand_rock	Biểu tình Đường ống Dakota Access (Standing Rock)	85.86%
12	12_britain_european_brexit_union	Brexit (Anh & EU)	48.98%
13	13_earth_space_planet_nasa	Vũ trụ & Thiên văn học (NASA)	81.87%
14	14_court_judge_gorsuch_justice	Tòa án Tối cao Hoa Kỳ (Neil Gorsuch)	17.58%
15	15_life_mind_energy_love	Triết lý sống / Tâm linh / Năng lượng (Trùu tượng)	85.19%
16	16_mosul_iraqi_iraq_isi	Chiến sự Mosul (Iraq & ISIS)	60.49%
17	17_china_chinese_beijing_xi	Chính trị Trung Quốc (Xi Jinping)	32.3%
18	18_france_french_le_paris	Tin tức / Chính trị Pháp (Có thể liên quan Le Pen)	27.15%
19	19_olympic_rio_athlete_olympics	Thể vận hội Olympic (Rio 2016)	5.63%
20	20_afghan_taliban_india_afghanistan	Tình hình Afghanistan (Taliban & Ấn Độ)	53.24%
21	21_halloween_ghost_leftright_swipe	Halloween / Chủ đề ma quái	98.35%
22	22_saudi_arabia_yemen_yemeni	Saudi Arabia & Chiến tranh Yemen	63.81%

Topic (ID)	Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Diễn giải)	Tỷ lệ giả
23	23_restaurant_cook_recipe_food	Âm thực & Nấu ăn	23.71%
24	24_turkey_erdogan_turkish_coup	Chính trị Thổ Nhĩ Kỳ (Erdogan & Đảo chính 2016)	37.5%
25	25_korea_north_korean_south	Quan hệ Liên Triều	20.88%
26	26_car_uber_tesla_vehicle	Công nghiệp Ô tô (Uber, Tesla)	16.67%
27	27_child_school_say_parent	Trẻ em, Trường học & Phụ huynh	15.73%
28	28_company_security_information_use	An ninh thông tin & Doanh nghiệp	56.63%
29	29_climate_pruitt_paris_change	Biến đổi khí hậu & Thỏa thuận Paris	29.87%
30	30_veteran_bonus_soldier_araldo	Cựu chiến binh / Quân đội (Có thể liên quan nhân vật Arnaldo)	74.32%
31	31_iran_iranian_pasdaran_iran	Tin tức Iran (IRGC/Pasdaran)	41.1%
32	32_job_worker_work_wage	Việc làm & Lao động	21.13%
33	33_gun_firearm_rifle_amendment	Quyền sở hữu súng (Hoa Kỳ)	40%
34	34_fox_kelly_ailes_news	Nội bộ Fox News (Kelly & Ailes)	21.74%
35	35_germany_german_merkel_migrant	Khủng hoảng Di cư ở Đức (Merkel)	31.34%
36	36_abortion_parenthood_plan_woman	Tranh cãi Phá thai (Hoa Kỳ)	20.97%
37	37_soros_george_election_protest	George Soros & Hoạt động chính trị/biểu tình	90.16%
38	38_california_state_los_angels	Tin tức California (Los Angeles)	15.25%

Topic (ID)	Fake News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Diễn giải)	Tỷ lệ giả
39	39_sweden_swedish_migrant_stockholm	Người di cư ở Thụy Điển	19.3%
40	40_venezuela_maduro_venezuelan_colombia	Khủng hoảng Venezuela	32.73%
41	41_storm_flood_hurricane_water	Thiên tai (Bão, Lũ lụt)	21.82%
42	42_duterte_philippine_manila_china	Philippines (Duterte & Quan hệ với Trung Quốc)	50.98%
43	43_school_education_student_teacher	Giáo dục & Trường học	24%



Hình 8. Biểu đồ Top 15 chủ đề của tập dữ liệu Fake News có tỷ lệ tin giả cao nhất phân tích từ BERTopic

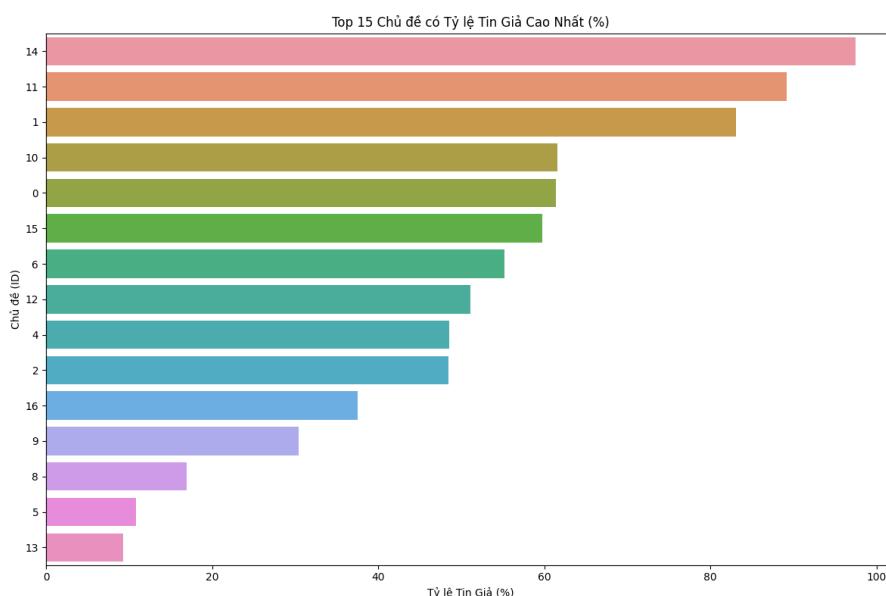
Phân tích kết quả BERTopic trên tập Fake News (Bảng 3, Hình 8): Tập dữ liệu Fake News thể hiện sự đa dạng chủ đề, bao gồm chính trị, xã hội, khoa học, sức khỏe và đời sống. Phân tích tỷ lệ tin giả cho thấy một số kết quả rất đáng chú ý. Chủ đề về Halloween/Ma quái (Topic 21) có tỷ lệ giả gần như tuyệt đối (98.35%), có thể do chứa các câu chuyện giải trí bị đặt hoặc nội dung đặc thù. Các chủ đề liên quan đến George Soros (Topic 37, 90.16%), Tranh cãi Phá thai (Topic 36, 90.16%), Biểu tình Đường ống Dakota (Topic 11, 85.86%), Tòa

án Tối cao Hoa Kỳ (Topic 14, 85.19%), và Xung đột Israel – Palestine (Topic 10, 85.86%) đều có tỷ lệ tin giả rất cao, phản ánh tính chất nhạy cảm và dễ bị xuyên tạc của các vấn đề này. Đáng ngạc nhiên, các chủ đề tưởng chừng ít mang màu sắc chính trị trực tiếp như Vũ trụ/NASA (Topic 13, 81.87%), Thể thao (Topic 8, 79.58%), và Phim ảnh/TV (Topic 2, 75.81%) cũng có tỷ lệ giả rất đáng kể trong tập dữ liệu này. Các chủ đề khác có tỷ lệ giả trên 70% bao gồm Nga & Syria (Topic 1, 74.87%), Biến đổi khí hậu (Topic 29, 74.32%), và Cựu chiến binh (Topic 30, 74.32%). Chủ đề trung tâm về Bầu cử Hoa Kỳ 2016 (Topic 0) có tỷ lệ giả ở mức khá cao (59.53%). Nhiều chủ đề chính trị hoặc sự kiện quốc tế khác có tỷ lệ giả thấp hơn đáng kể trong tập dữ liệu này. Nhóm ngoại lai (Topic – 1) gồm các văn bản không được phân loại vào chủ đề cụ thể, với các từ khóa chung chung liên quan đến chính trị Hoa Kỳ như trump, say, clinton, president.

Bảng 4. Kết quả phân loại chủ đề trên tập dữ liệu Fake or Real News

Topic (ID)	Fake or Real News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Điễn giải)	Tỷ lệ giả
-1	– 1_trump_say_would_clinton	Ngoại lai (BERTopic không phân loại được)	
0	0_russia_syria_russian_say	Nga & Cuộc chiến Syria	61.36%
1	1_clinton_email_fbi_hillary	Vụ emails của Hillary Clinton (FBI)	83.1%
2	2_trump_donald_say _campaign	Chiến dịch tranh cử của Donald Trump (2016)	48.46%
3	3_cruz_trump_say_republican	Chính trị Đảng Cộng hòa (Cruz & Trump)	1.61%
4	4_police_say_gun_officer	Cảnh sát & Vụ việc liên quan súng	48.56%
5	5_court_say_house_marriage	Pháp lý Hôn nhân (Tòa án)	10.82%
6	6_gold_bank_rate_market	Tài chính & Thị trường (Vàng, Ngân hàng)	55.24%
7	7_sander_clinton_democratic _campaign	Bầu cử sơ bộ Đảng Dân chủ 2016 (Sanders & Clinton)	5.97%
8	8_obama_president_say _house	Tổng thống Obama & Chính sách	16.94%

Topic (ID)	Fake or Real News Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Diễn giải)	Tỷ lệ giả
9	9_clinton_hillary_campaign _say	Chiến dịch tranh cử của Hillary Clinton (2016)	30.43%
10	10_vote_election_poll_voter	Bầu cử, Bỏ phiếu & Thăm dò ý kiến	61.61%
11	11_zika_cancer_health_drug	Tin tức Y tế (Zika, Ung thư, Thuốc)	89.19%
12	12_climate_coal_change _emission	Biến đổi khí hậu & Than đá	51.11%
13	13_iran_deal_nuclear_iranian	Thỏa thuận Hạt nhân Iran (JCPOA)	9.3%
14	14_world_people_one_human	Các vấn đề Toàn cầu, Nhân loại	97.44%
15	15_israel_netanyahu _palestinian_israeli	Xung đột Israel – Palestine (Netanyahu)	59.72%
16	16_health_insurance_care _obamacare	Y tế Hoa Kỳ	37.5%



Hình 9. Biểu đồ Top 15 chủ đề của tập dữ liệu Fake or Real News có tỷ lệ tin giả cao nhất phân tích từ BERTopic

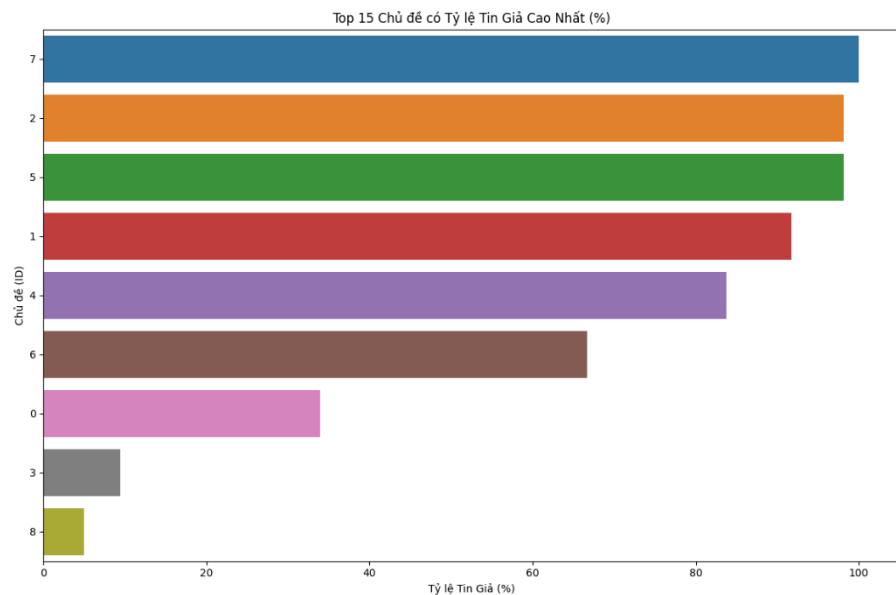
Phân tích kết quả BERTopic trên tập Fake or Real News (Bảng 4, Hình 9): Tập dữ liệu này tập trung rõ rệt vào các sự kiện chính trị Hoa Kỳ giai đoạn

2016. Phân tích tỷ lệ tin giả cho thấy Topic 14 (với các từ khóa rất chung như world, people, one, human, được diễn giải là ‘Các vấn đề Toàn cầu, Nhân loại’) có tỷ lệ giả cao bất thường (97.44%). Điều này gợi ý rằng các văn bản được gom vào chủ đề rất tổng quát này trong bộ dữ liệu Fake or Real News phần lớn là tin giả, có thể chúng thuộc loại nội dung spam, các bài viết triết lý/chủ quan không phải tin tức, hoặc một loại hình thông tin sai lệch đặc thù nào đó mà mô hình gom lại. Các chủ đề khác có tỷ lệ giả cao rõ rệt bao gồm Tin tức Y tế (chủ yếu về Zika, Ung thư – Topic 11, 89.19%) và Vụ emails của Hillary Clinton (Topic 1, 83.1%). Các vấn đề quốc tế như Nga & Cuộc chiến Syria (Topic 0, 61.36%), Xung đột Israel – Palestine (Topic 15, 59.72%) cùng các chủ đề về Bầu cử/Thăm dò ý kiến (Topic 10, 61.61%), Tài chính (Topic 6, 55.24%) và Biến đổi khí hậu (Topic 12, 51.11%) cũng có tỷ lệ tin giả chiếm đa số. Đáng chú ý, các chủ đề nói trực tiếp về chiến dịch tranh cử của Trump (Topic 2, 48.46%) hay Clinton (Topic 9, 30.43%) lại có tỷ lệ giả dưới 50%, thấp hơn đáng kể so với chủ đề về vụ emails (Topic 1). Các chủ đề liên quan đến nội bộ Đảng Cộng hòa (Topic 3, 1.61%) và bầu cử sơ bộ Đảng Dân chủ (Topic 7, 5.97%) có tỷ lệ giả rất thấp trong tập dữ liệu này. Topic – 1 (ngoại lai) vẫn chứa các phát biểu chính trị chung.

Bảng 5. Kết quả phân loại chủ đề trên tập dữ liệu Fake News Detection

Topic(ID)	Fake News Detection Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Diễn giải)	Tỷ lệ giả
- 1	- 1_blackhawks_goal _rohingya_myanmar	Ngoại lai (BERTopic không phân loại được)	
0	0_say_state_new_trump	Tin tức / Phát biểu chung (Trump)	33.92%
1	1_week_game_yard_play	Thể thao (Kết quả/Trận đấu hàng tuần)	91.75%
2	2_red_warn_list_happen	Cảnh báo / Danh sách nguy hiểm / Dự đoán	98.18%
3	3_cup_world_photo_sport	Thể thao (World Cup & Hình ảnh)	9.44%
4	4_season_game_go_hit	Thể thao (Mùa giải, Trận đấu)	83.72%
5	5_light_room_ledpowered _diy	Dự án Tự làm (DIY) – Đèn LED	98.18%

Topic(ID)	Fake News Detection Dataset		
	Tên chủ đề (BERTopic)	Tên Chủ đề (Diễn giải)	Tỷ lệ giả
6	6_nfl_anthem_player_protest	Biểu tình Chào cờ NFL	66.67%
7	7_campaign_prolegalization_fentanyl_painkiller	Vận động hợp pháp hóa thuốc / Opioids (Fentanyl)	100%
8	8_catalan_independence_catalonia_puigdemont	Độc lập Catalonia (Tây Ban Nha)	5%



Hình 10. Biểu đồ Top 15 chủ đề của tập dữ liệu Fake News Detection có tỷ lệ tin giả cao nhất phân tích từ BERTopic

Phân tích kết quả BERTopic trên tập Fake News Detection (Bảng 5, Hình 10): Dựa trên kết quả tập dữ liệu này cho thấy một sự phân bố tin giả khá đặc biệt. Chủ đề về Vận động hợp pháp hóa thuốc / Opioids (Topic 7) có tỷ lệ giả tuyệt đối (100%). Các chủ đề về Cảnh báo / Danh sách nguy hiểm / Dự đoán (Topic 2) và Dự án Tự làm – Đèn LED (Topic 5) cũng có tỷ lệ giả cực kỳ cao (đều 98.18%). Các chủ đề liên quan đến Thể thao (Topic 1 – Hàng tuần, 91.75%; Topic 4 – Mùa giải, 83.72%) và phát biểu chung về Trump (Topic 0, 91.75%) cũng nằm trong nhóm có tỷ lệ giả rất cao. Chủ đề về Biểu tình Chào cờ NFL (Topic 6) vẫn có tỷ lệ giả đáng kể (66.67%), mặc dù không còn là 100% như phân tích trước đó. Ngược lại, chủ đề về Thể thao World Cup (Topic 3) lại có tỷ lệ giả rất thấp (9.44%), cho thấy sự khác biệt lớn giữa các nội dung thể thao

khác nhau trong tập dữ liệu này. Chủ đề Độc lập Catalonia (Topic 8) chỉ có 5% tin giả. Nhóm ngoại lai (Topic – 1) gồm các văn bản không được phân loại, với từ khóa pha trộn giữa thể thao (blackhawks, goal), chính trị (myanmar) và khủng hoảng nhân đạo (rohingya).

3.2.2.2 Phân tích Từ vựng Đặc trưng bằng CountVectorizer

Để có cái nhìn sơ bộ về sự khác biệt từ vựng giữa hai nhãn, nghiên cứu đã sử dụng CountVectorizer để đếm tần suất xuất hiện của các từ trong các văn bản thuộc nhãn “Giả” (Fake) và “Thật” (True) trên từng bộ dữ liệu con. Bằng cách so sánh tần suất, nghiên cứu xác định được những từ xuất hiện nhiều trong các tin giả nhưng lại ít xuất hiện trong tin thật. Dưới đây là danh sách một số từ ngữ đặc trưng tiêu biểu được tìm thấy cho nhãn “Giả” trên mỗi tập dữ liệu:

ISOT Fake News Dataset: america, ask, believe, black, candidate, case, child, claim, continue, fact, family, fbi, good, hillary, hillary clinton, image, know, leave, like, live, look, man, medium, muslim, pay, point, post, public, question, really, run, story, thing, think, try, tweet, video, watch, way, woman, world.

Fake News Dataset: america, article, believe, big, black, continue, control, email, fact, fbi, great, happen, hillary, hillary clinton, mean, military, money, nation, october, order, place, post, power, russia, russian, source, syria, war.

Fake or Real News Dataset: ask, big, bush, congress, conservative, court, cruz, deal, debate, democrat, democratic, gop, iran, leader, month, nt, percent, plan, poll, primary, question, race, rubio, run, sander, security, senate, talk, washington, week.

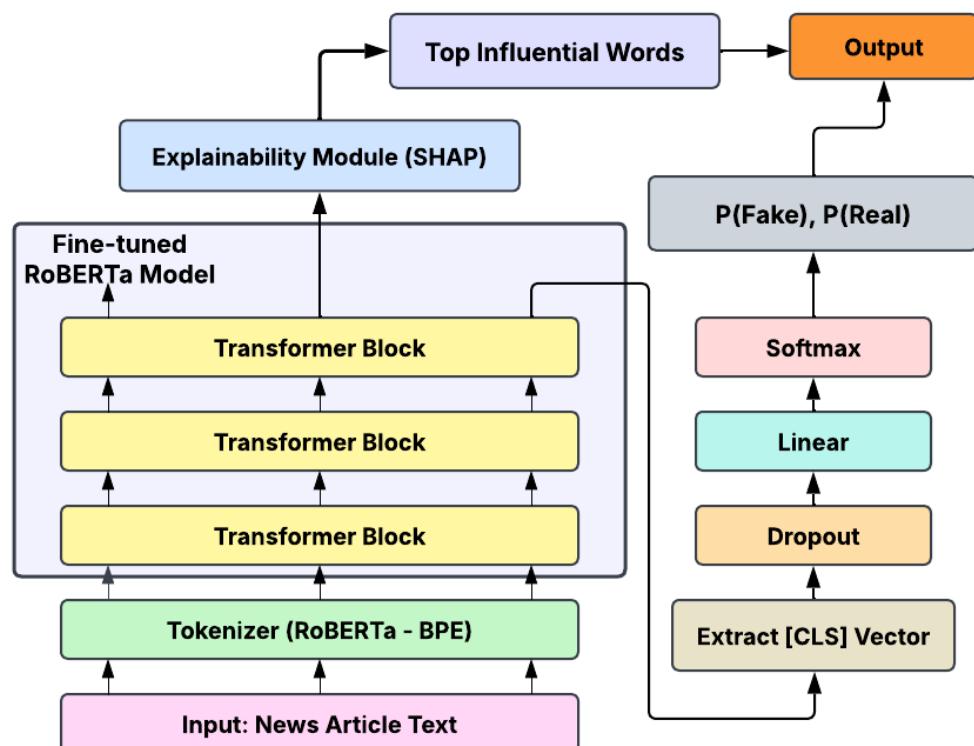
Fake News Detection Dataset: allow, america, appear, article, attack, bad, bear, best, content, contentad, data, defense, event, fact, fan, field, follow, free, great, happen, health, hit, http, jet, law, line, link, long, lose, nfl, night, option, pass, point, post, provide, raven, reader, room, season, second, smart, source, source http, steelers, video, vs, war, win, yard.

Phân tích các danh sách này cho thấy một số xu hướng từ vựng thường liên kết với nhãn “Giả”. Ví dụ, các tên riêng chính trị gia như ‘hillary’, ‘clinton’, ‘trump’, ‘bush’, ‘cruz’, ‘rubio’, ‘sander’ xuất hiện nhiều trong các bộ dữ liệu tập trung vào chính trị Hoa Kỳ (ISOT, Fake News, Fake or Real). Các thuật ngữ chính trị như ‘congress’, ‘conservative’, ‘democrat’, ‘democratic’, ‘gop’, ‘senate’ đặc trưng cho tin giả trong Fake or Real. Một số động từ hoặc danh từ mang tính khẳng định, bình luận hoặc liên quan đến thông tin như ‘believe’,

‘fact’, ‘claim’, ‘article’, ‘post’, ‘source’, ‘question’ cũng khá phổ biến trong nhãn giả ở nhiều bộ dữ liệu. Ngoài ra, các từ liên quan đến các chủ đề gây tranh cãi như ‘muslim’ (ISOT), ‘fbi’ (ISOT, Fake News), ‘russia’/‘russian’ (Fake News), ‘iran’ (Fake or Real), ‘nfl’ (Fake News Detection) cũng nằm trong top từ phân biệt. Tuy nhiên, cần lưu ý rằng phương pháp dựa trên tần suất từ đơn giản này chỉ cung cấp góc nhìn bề mặt và có thể bị ảnh hưởng bởi chủ đề chính của tin tức hơn là bản chất thật/giả. Nó chưa thể hiện được ngữ cảnh hay sắc thái mà các mô hình học sâu như Transformer sẽ khai thác để phân loại chính xác hơn.

3.2.3 Xây dựng phương pháp

Để giải quyết bài toán phân loại tin giả, nghiên cứu này đề xuất và áp dụng một phương pháp dựa trên kiến trúc Transformer tiên tiến, cụ thể là tinh chỉnh (fine – tuning) mô hình RoBERTa đã được tiền huấn luyện (pre – trained). RoBERTa được lựa chọn làm mô hình nền tảng do khả năng nắm bắt ngữ cảnh hai chiều sâu sắc và hiệu quả vượt trội đã được chứng minh trên nhiều tác vụ Xử lý Ngôn ngữ Tự nhiên (NLP). Quy trình tổng thể để xử lý và phân loại một văn bản tin tức đầu vào bằng mô hình đề xuất được minh họa trong Hình 11.



Hình 11. Quy trình xử lý và phân loại của mô hình đề xuất
Luồng xử lý này bao gồm các bước chính như sau:

Tiền xử lý và Tokenization: Đầu tiên, văn bản tin tức gốc (Input Text) được đưa vào bộ Tokenizer tương ứng với mô hình RoBERTa (sử dụng kỹ thuật mã hóa Byte – Pair Encoding – BPE). Quá trình này chuẩn hóa và tách văn bản thành các đơn vị nhỏ hơn gọi là token (có thể là từ hoặc các mảnh từ – subword). Sau đó, các token này được chuyển đổi thành dạng số mà mô hình có thể hiểu được, bao gồm:

Input IDs: Dãy các mã định danh số duy nhất cho mỗi token.

Attention Mask: Một dãy nhị phân (0 hoặc 1) cho mô hình biết cần chú ý vào những token nào (token thật) và bỏ qua những token nào (token đệm – padding). Để đảm bảo tất cả đều vào có cùng kích thước, các chuỗi token sẽ được cắt bớt (truncation) hoặc thêm token đệm (padding) cho đến khi đạt một độ dài tối đa (256 tokens) xác định.

Trích xuất Đặc trưng Ngữ cảnh bởi RoBERTa: Input IDs và Attention Mask sau đó được đưa làm đầu vào cho Mô hình RoBERTa đã được Fine – tune. Dữ liệu sẽ đi qua các lớp Transformer Encoder xếp chồng của RoBERTa. Tại đây, nhờ cơ chế Multi – head Self – Attention, mô hình sẽ tính toán và tạo ra các vector biểu diễn ngữ cảnh (contextualized embeddings) cho mỗi token, nắm bắt mối quan hệ và ý nghĩa của từ trong ngữ cảnh của toàn bộ chuỗi.

Lấy Vector Đại diện cho Văn bản: Đối với các tác vụ phân loại văn bản, một phương pháp phổ biến là sử dụng vector biểu diễn của một token đặc biệt làm đại diện cho toàn bộ chuỗi. Trong nghiên cứu này, vector ẩn (hidden state) ở lớp cuối cùng của RoBERTa tương ứng với token [CLS] (thường được tự động thêm vào đầu chuỗi bởi tokenizer) được trích xuất để làm vector đặc trưng tổng hợp cho văn bản đầu vào.

Phân loại và Giải thích (nếu được yêu cầu): Vector đặc trưng [CLS] này sau đó được đưa qua phần đầu phân loại (classification head). Đồng thời, nếu người dùng yêu cầu giải thích (explain_flag là true), mô hình và dữ liệu đầu vào cũng được sử dụng để tính toán độ quan trọng của từng token bằng SHAP.

Lớp Dropout: Một lớp Dropout có thể được áp dụng lên vector [CLS] để giảm overfitting.

Lớp Linear: Vector được đưa qua một lớp tuyến tính duy nhất để ánh xạ sang không gian 2 chiều (logits) cho hai lớp “Tin Giả” và “Tin Thật”.

Tính toán SHAP (khi explain_flag = true):

Một hàm dự đoán (`shap_predict_fn`) được định nghĩa để nhận đầu vào là các đoạn văn bản, thực hiện tokenization (với độ dài tối đa phù hợp cho SHAP, ở nghiên cứu này là 256 tokens), và trả về logits từ mô hình RoBERTa.

Thư viện shap được sử dụng để tạo một Explainer dựa trên hàm dự đoán và tokenizer.

Explainer được áp dụng lên văn bản đã tiền xử lý để tính toán SHAP values cho mỗi token.

Các SHAP value dương (đóng góp vào dự đoán “Fake”) được xác định và sắp xếp.

Các token tương ứng với top N SHAP value dương cao nhất (sau khi loại bỏ các token đặc biệt và tiền tố `G` của RoBERTa tokenizer) được trích xuất làm “Top words SHAP”.

Lớp Softmax: Hàm Softmax được áp dụng lên vector logits (từ bước 2) để chuyển đổi thành phân phối xác suất $[P(\text{Giả}), P(\text{Thật})]$.

Dự đoán Nhãn: Nhãn cuối cùng được dự đoán là lớp có xác suất cao hơn.

Kết quả trả về: Hệ thống trả về nhãn dự đoán, xác suất, điểm cảm xúc và danh sách “Top words SHAP” (nếu được yêu cầu và tính toán thành công).

Quá trình fine – tuning chính là bước huấn luyện để điều chỉnh các trọng số trong toàn bộ mô hình RoBERTa và lớp phân loại mới thêm vào, giúp mô hình học cách phân biệt giữa tin thật và tin giả dựa trên bộ dữ liệu huấn luyện cụ thể. Kiến trúc mô hình tổng thể này được kỳ vọng sẽ tận dụng được sức mạnh biểu diễn ngôn ngữ của RoBERTa để đạt hiệu quả phân loại cao.

CHƯƠNG 4: KẾT QUẢ THỰC NGHIỆM

4.1 Thiết lập môi trường

Nghiên cứu đã triển khai mô hình nghiên cứu trong môi trường Kaggle, một nền tảng phổ biến để phát triển và kiểm thử các mô hình học máy. Cấu hình phần cứng của môi trường thực nghiệm bao gồm CPU (Central Processing Unit) Intel Xeon với 2 lõi, GPU NVIDIA Tesla P100 với 16GB bộ nhớ HBM2 (High Bandwidth Memory 2) và 29GB RAM (Random Access Memory). Nghiên cứu sử dụng Python phiên bản 3.10.12 và dung lượng ổ đĩa là 57GB. Với cấu hình này, nghiên cứu đảm bảo khả năng xử lý hiệu quả các tác vụ tính toán trong suốt quá trình huấn luyện mô hình.

4.2 Cơ sở đánh giá

Nghiên cứu này thực hiện phân loại tin tức thật giả với mục tiêu xác định một mẫu tin tức là “Thật” (True) hay “Giả” (Fake). Để thực hiện đánh giá hiệu quả của các mô hình, nghiên cứu đã lựa chọn các độ đo phổ biến trong bài toán phân loại nhị phân bao gồm: Accuracy, Precision, Recall, và F1 – score. Trong đó, nghiên cứu quy ước lớp “Tin Giả” (Fake News) là lớp dương tính (positive) và lớp “Tin Thật” (True News) là lớp âm tính (negative). Chi tiết về các độ đo này sẽ trình bày bên dưới.

Quá trình phân loại tin tức có thể cho ra các kết quả dựa trên ma trận nhầm lẫn (confusion matrix) như sau:

True Positives (TP): Các mẫu tin tức Giả được dự đoán chính xác là Giả.

True Negatives (TN): Các mẫu tin tức Thật được dự đoán chính xác là Thật.

False Positives (FP): Các mẫu tin tức Thật nhưng bị dự đoán sai là Giả.

False Negatives (FN): Các mẫu tin tức Giả nhưng bị dự đoán sai là Thật.

Các chỉ số đánh giá hiệu suất mô hình được tính toán dựa trên các giá trị trên, bao gồm:

Accuracy (Độ chính xác tổng thể): Đo lường tổng thể độ chính xác của mô hình bằng cách tính tỷ lệ tổng số dự đoán đúng (cả Thật và Giả) trên toàn bộ tập dữ liệu. Độ chính xác được tính bằng công thức (4):

$$\text{Acc} = \frac{TN + TP}{TN + FN + TP + FP} \quad (4)$$

Precision (Độ chính xác dương tính): Đo lường tỷ lệ các mẫu được dự đoán là Giả mà thực sự là Giả. Nó cho biết mức độ đáng tin cậy khi mô hình dự đoán một tin là Giả. Công thức tính Precision là công thức (5):

$$\mathbf{P} = \frac{TP}{TP + FP} \quad (5)$$

Recall (Độ nhạy): Đo lường tỷ lệ các mẫu thực sự là Giả mà mô hình dự đoán đúng là Giả. Nó cho biết khả năng mô hình “phát hiện” được tin Giả trong tập dữ liệu. Công thức tính Recall là công thức (6):

$$\mathbf{R} = \frac{TP}{TP + FN} \quad (6)$$

F1 – score: Là trung bình điều hòa của Precision và Recall. Đây là độ đo hữu ích khi cần cân bằng giữa Precision và Recall, đặc biệt trong trường hợp dữ liệu mất cân bằng hoặc khi cả FP và FN đều quan trọng. Công thức tính F1 – score là công thức (7):

$$\mathbf{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

4.3 Điều chỉnh siêu tham số và đào tạo mô hình

Để tối ưu hóa hiệu suất cho nhiệm vụ phân loại tin tức thật giả, nghiên cứu thực hiện quy trình điều chỉnh siêu tham số trên các tập dữ liệu đã được chia theo tỷ lệ 8:2 cho huấn luyện (training) và kiểm thử (testing). Quá trình này bao gồm việc thử nghiệm các kết hợp khác nhau của Optimizer (Adam, Adamax, AdamW) và số lượng epoch huấn luyện (3, 5, 10) cho cả hai mô hình nền tảng là BERT và RoBERTa. Hiệu suất của mỗi cấu hình được đánh giá trên tập kiểm thử dựa trên các chỉ số F1 – score (F1), Recall (R), Precision (P) và Accuracy (Acc).

Kết quả chi tiết trên tập dữ liệu ISOT được trình bày đầy đủ trong Bảng 6. Một điểm đáng chú ý là cả hai mô hình BERT và RoBERTa đều đạt hiệu suất cực kỳ cao trên tập dữ liệu này, với hầu hết các cấu hình thử nghiệm đều cho Accuracy và F1 – score vượt ngưỡng 0.99. Điều này cho thấy đặc điểm dữ liệu ISOT có thể tương đối rõ ràng, giúp các mô hình Transformer mạnh mẽ dễ dàng phân tách tin thật và giả, tuy nhiên cũng cần lưu ý khi xem xét khả năng khai thác dữ liệu.

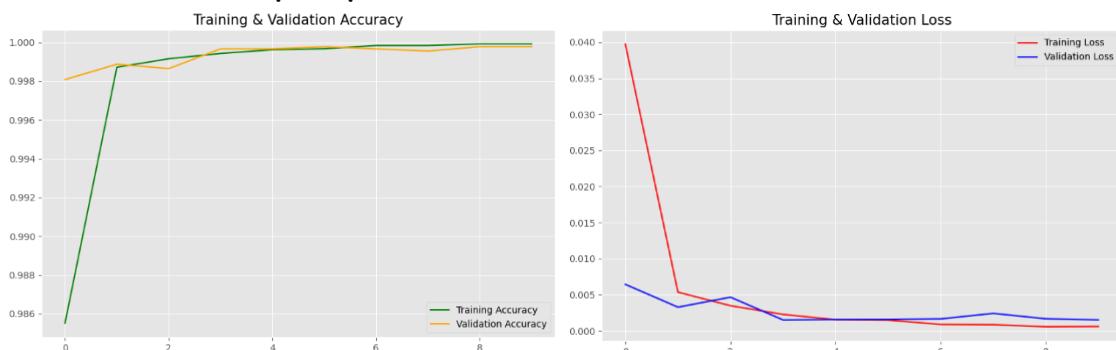
Bảng 6. Kết quả đánh giá hiệu suất trên tập dữ liệu ISOT Fake News Dataset

Mô hình	Optimizer	Epoch	F1	R	P	Acc
BERT	Adam	3	0.9992	0.9986	0.9998	0.9992
		5	0.9994	0.9988	1	0.9994
		10	0.9997	0.9993	1	0.9997
	Adamax	3	0.9992	0.9988	0.9995	0.9992
		5	0.9994	0.9991	0.9998	0.9994
		10	0.9997	0.9998	0.9995	0.9997
	AdamW	3	0.9991	0.9984	0.9998	0.9991
		5	0.9988	0.9998	0.9979	0.9989
		10	0.9995	0.9991	1	0.9996
	Adam	3	0.9994	1	0.9988	0.9994
		5	0.9997	0.9998	0.9995	0.9997
		10	0.9991	0.9986	0.9995	0.9991
	Adamax	3	0.9993	0.9995	0.9991	0.9993
		5	0.9987	0.9998	0.9977	0.9988
		10	0.9998	0.9995	1	0.9998
	AdamW	3	0.9991	0.9986	0.9995	0.9991
		5	0.9992	0.9986	0.9998	0.9992
		10	0.9995	0.9993	0.9998	0.9996

Phân tích sâu hơn vào Bảng 6, cấu hình RoBERTa kết hợp với Optimizer Adamax và huấn luyện trong 10 epochs cho kết quả vượt trội nhất, đạt F1 – score là 0.9998 và Accuracy là 0.9998. So sánh giữa các optimizer, Adamax và Adam

nhìn chung cho kết quả rất tốt và ổn định, trong khi AdamW trong một số trường hợp cụ thể (như BERT + AdamW tại epoch=5) cho F1 – score (0.9988) thấp hơn đôi chút so với các optimizer còn lại ở cùng mốc epoch. Về so sánh giữa hai mô hình, cả BERT và RoBERTa đều chứng tỏ hiệu quả vượt trội, với kết quả tốt nhất của BERT (F1=0.9997) chỉ thấp hơn không đáng kể so với RoBERTa (F1=0.9998). Dựa trên hiệu suất cao nhất đạt được, cấu hình RoBERTa + Adamax (epoch=10) được lựa chọn là mô hình tối ưu cho tập dữ liệu ISOT trong nghiên cứu này.

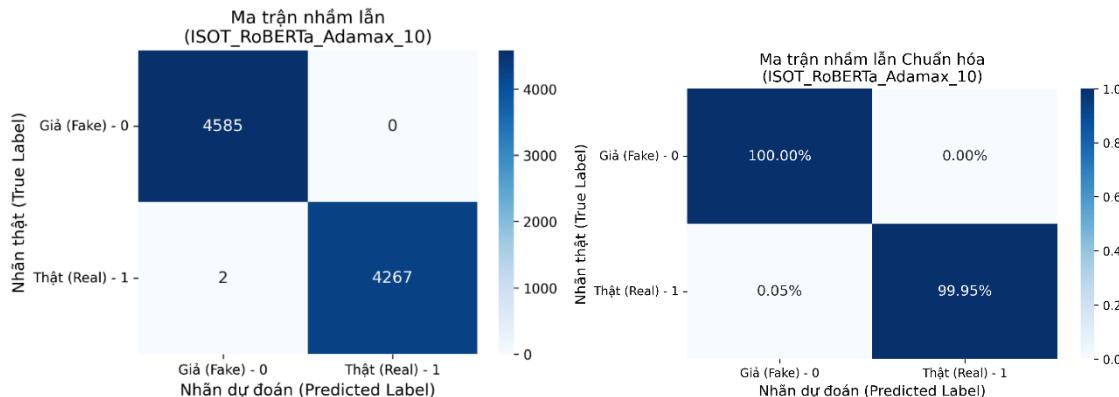
Hình 12 minh họa quá trình huấn luyện của mô hình RoBERTa + Adamax (epoch=10) trên tập dữ liệu ISOT thông qua đồ thị accuracy và loss qua các epoch. Nhìn chung, quá trình huấn luyện diễn ra rất hiệu quả và ổn định. Các đường loss (cả training màu đỏ và validation màu xanh dương) đều giảm mạnh trong những epoch đầu tiên và nhanh chóng hội tụ về mức rất thấp, duy trì ổn định ở các epoch sau. Tương ứng, các đường accuracy (training màu xanh lá và validation màu cam) cũng tăng vọt lên giá trị cực kỳ cao (xấp xỉ 1) và đạt trạng thái ổn định từ khoảng epoch 4 – 5. Quan trọng hơn, khoảng cách giữa đường training và validation trên cả hai đồ thị loss và accuracy đều rất nhỏ, cho thấy mô hình học tốt các đặc trưng của dữ liệu và không có dấu hiệu rõ ràng của hiện tượng overfitting trong 10 epoch huấn luyện. Kết quả này cũng cố việc lựa chọn 10 epochs là phù hợp, mặc dù hiệu suất cao có thể đã đạt được sớm hơn.



Hình 12. Đồ thị Training và Validation Accuracy/Loss của mô hình RoBERTa_Adamax_10 trên tập ISOT

Hình 13 trình bày ma trận nhầm lẫn và chuẩn hóa cho mô hình RoBERTa + Adamax (epoch = 10) trên tập dữ liệu ISOT. Kết quả cho thấy hiệu suất gần như hoàn hảo, phản ánh các chỉ số F1 và Accuracy cực kỳ cao đã được báo cáo trước đó. Quan sát đường chéo chính, số lượng dự đoán đúng cho cả lớp Tin Giả (TP = 4585) và Tin Thật (TN = 4267) chiếm tuyệt đối đa số các trường hợp. Đáng chú ý nhất, mô hình không bỏ sót bất kỳ tin giả nào (FN = 0), tương ứng với tỷ lệ Recall 100% cho lớp tin giả (như thấy rõ trên ma trận chuẩn hóa). Đồng thời, chỉ có 2 trường hợp tin thật bị phân loại nhầm thành tin giả (FP = 2). Điều này cho thấy mô

hình không chỉ cực kỳ hiệu quả trong việc phát hiện tin giả mà còn gần như không đưa ra cảnh báo sai cho các tin tức thật trên bộ dữ liệu ISOT này.



Hình 13. Ma trận nhầm lẩn/chuẩn hóa của mô hình RoBERTa_Adamax_10 trên tập dữ liệu ISOT

Tiếp tục quá trình tinh chỉnh siêu tham số, nghiên cứu đã áp dụng quy trình tương tự cho tập dữ liệu Fake News Dataset. Kết quả chi tiết cho các cấu hình thử nghiệm được trình bày trong Bảng 7. Nhìn chung, hiệu suất trên tập dữ liệu này vẫn ở mức cao (đa số các chỉ số chính đều trên 0.97 – 0.98), tuy nhiên không đạt đến mức gần như tuyệt đối như trên tập ISOT. Điều này gợi ý rằng Fake News Dataset có thể chứa nhiều mẫu dữ liệu phức tạp hơn hoặc có sự chồng lấn nhất định về đặc điểm ngôn ngữ giữa tin thật và tin giả, gây khó khăn hơn đôi chút cho các mô hình.

Khi phân tích chi tiết Bảng 7 để xác định cấu hình tối ưu, một điểm cần lưu ý là sự khác biệt nhỏ giữa chỉ số F1 – score và Accuracy cao nhất. Cụ thể, cấu hình RoBERTa kết hợp Optimizer Adam và huấn luyện trong 5 epochs đạt được F1 – score cao nhất một cách rõ rệt (0.9986). Đây là một kết quả rất ấn tượng, cho thấy khả năng cân bằng tốt giữa Precision (0.9860) và Recall (0.9913) của mô hình này. Mặt khác, Accuracy cao nhất lại thuộc về cấu hình BERT + Adam (epoch=3) với giá trị 0.9888, chỉ nhỉnh hơn không đáng kể so với mức Accuracy 0.9884 của RoBERTa + Adam (epoch=5).

Xét về ảnh hưởng của số epoch, cấu hình tốt nhất RoBERTa + Adam lại đạt đỉnh ở 5 epochs và việc tăng lên 10 epochs làm giảm đáng kể hiệu suất (F1 giảm xuống 0.9846). Xu hướng này không nhất quán trên tất cả các cấu hình, với optimizer Adamax, cả BERT và RoBERTa đều cho thấy hiệu suất tăng dần khi tăng số epoch từ 3 lên 10. Điều này cho thấy sự tương tác phức tạp giữa kiến trúc mô hình, optimizer và thời gian huấn luyện trên bộ dữ liệu này.

Về phía các optimizer, Adam tỏ ra hiệu quả khi kết hợp với RoBERTa để đạt F1 cao nhất và kết hợp với BERT để đạt Accuracy cao nhất. AdamW cũng cho

kết quả cạnh tranh ở một số cấu hình (ví dụ RoBERTa + AdamW epoch=10 đạt F1=0.9883). Adamax nhìn chung cho kết quả hơi thấp hơn một chút trên tập dữ liệu này so với Adam và AdamW ở các mốc epoch thấp và trung bình, nhưng cũng có cải thiện ở 10 epochs.

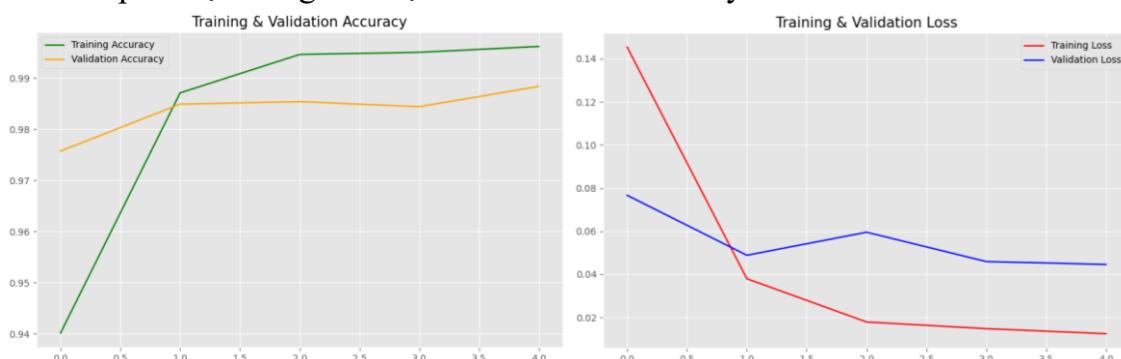
So sánh giữa hai mô hình, RoBERTa thể hiện ưu thế rõ ràng về chỉ số F1 – score (0.9986 so với mức cao nhất của BERT là 0.9890), cho thấy khả năng tổng quát tốt hơn trong việc cân bằng Precision/Recall trên tập dữ liệu này. Mặc dù BERT đạt Accuracy nhỉnh hơn một chút ở một cấu hình khác, sự vượt trội về F1 của RoBERTa là đáng kể hơn. Do đó, dựa trên việc ưu tiên chỉ số F1 – score – một thước đo quan trọng cho hiệu suất tổng thể trong các bài toán phân loại – cấu hình RoBERTa + Adam (epoch=5) được xác định là mô hình tối ưu cho tập dữ liệu Fake News Dataset.

Bảng 7. Kết quả đánh giá hiệu suất trên tập dữ liệu Fake News Dataset

Mô hình	Optimizer	Epoch	F1	R	P	Acc
BERT	Adam	3	0.9890	0.9825	0.9956	0.9888
		5	0.9856	0.9801	0.9915	0.9854
		10	0.9879	0.9883	0.9874	0.9876
	Adamax	3	0.9803	0.9805	0.9801	0.9799
		5	0.9813	0.9849	0.9778	0.9809
		10	0.9873	0.9840	0.9907	0.9871
	AdamW	3	0.9848	0.9776	0.9921	0.9846
		5	0.9881	0.9878	0.9883	0.9879
		10	0.9870	0.9791	0.9951	0.9869
RoBERTa	Adam	3	0.9811	0.9844	0.9778	0.9807
		5	0.9986	0.9913	0.9860	0.9884
		10	0.9846	0.9806	0.9887	0.9844
	Adamax	3	0.9763	0.9728	0.9799	0.9760

Mô hình	Optimizer	Epoch	F1	R	P	Acc
AdamW		5	0.9793	0.9878	0.9708	0.9787
		10	0.9849	0.9864	0.9835	0.9846
		3	0.9837	0.9825	0.9849	0.9834
		5	0.9818	0.9708	0.9930	0.9817
		10	0.9883	0.9869	0.9898	0.9881

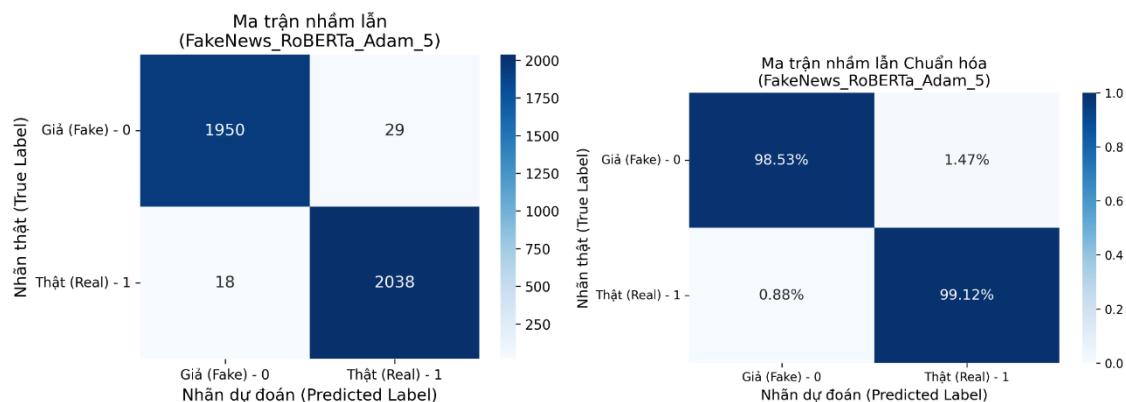
Quan sát đồ thị huấn luyện Hình 14, có thể thấy mô hình học rất nhanh trên tập dữ liệu Fake News. Đường training loss (đỏ) giảm cực kỳ nhanh chóng trong epoch đầu tiên và tiếp tục giảm đều ở các epoch sau, trong khi training accuracy (xanh lá) cũng tăng vọt và đạt mức trên 0.99 từ epoch thứ hai. Tuy nhiên, đường validation loss (xanh dương) sau khi giảm ở epoch đầu lại có xu hướng dao động và không tiếp tục giảm sâu thêm, thậm chí còn tăng nhẹ ở epoch 2. Tương tự, validation accuracy (cam) tăng mạnh ban đầu lên khoảng 0.985 nhưng sau đó chỉ dao động nhẹ quanh mức này. Sự khác biệt ngày càng tăng giữa đường training và validation trên cả hai đồ thị, đặc biệt là việc validation loss không tiếp tục giảm sâu, cho thấy mô hình có thể đã bắt đầu có dấu hiệu overfitting nhẹ sau epoch đầu tiên hoặc thứ hai. Mặc dù vậy, mô hình vẫn đạt được hiệu suất tổng thể rất cao trên tập kiểm định trong 5 epoch này, phù hợp với kết quả F1 – score và Accuracy cao đã được ghi nhận trong Bảng 7. Điều này cũng lý giải tại sao việc huấn luyện thêm đến 10 epoch lại làm giảm hiệu suất của cấu hình này.



Hình 14. Đồ thị Training và Validation Accuracy/Loss của mô hình RoBERTa_Adam_5 trên tập Fake News

Ma trận nhầm lẫn cho mô hình RoBERTa + Adam (epoch = 5) trên tập dữ liệu Fake News (Hình 15) tiếp tục cho thấy hiệu suất tổng thể rất tốt, với số lượng lớn các dự đoán đúng (TP = 1950, TN = 2038). Tuy nhiên, so với tập ISOT, mô

hình đã bắt đầu mắc phải nhiều lỗi hơn. Cụ thể, có 29 trường hợp tin giả bị bỏ sót và phân loại nhầm thành tin thật ($FN = 29$), và 18 trường hợp tin thật bị phân loại nhầm thành tin giả ($FP = 18$). Nhìn vào ma trận chuẩn hóa, tỷ lệ bỏ sót tin giả (FN rate) là khoảng 1.47% ($29 / (1950 + 29)$), tương ứng với Recall khoảng 98.53% cho lớp tin giả. Tỷ lệ tin thật bị phân loại nhầm (FP rate) thấp hơn một chút, khoảng 0.87% ($18 / (18 + 2038)$). Điều này cho thấy, trên tập dữ liệu này, mô hình vẫn hoạt động rất hiệu quả nhưng có xu hướng bỏ sót tin giả nhiều hơn một chút so với việc báo động nhầm tin thật.



Hình 15. Ma trận nhầm lẫn/chuẩn hóa mô hình RoBERTa_Adam_5 trên tập dữ liệu Fake News

Ké tiếp, nghiên cứu tiến hành phân tích hiệu suất trên tập dữ liệu Fake or Real News, với kết quả được trình bày trong Bảng 8. Các chỉ số F1 – score và Accuracy tốt nhất chỉ dao động quanh mức 0.95. Điều này khẳng định mạnh mẽ rằng Fake or Real News là một tập dữ liệu thử thách hơn đáng kể, có thể do chất lượng dữ liệu, sự tinh vi trong cách viết tin giả, hoặc sự tương đồng lớn hơn giữa các mẫu tin thật và giả.

Mặc dù đối mặt với dữ liệu khó hơn, RoBERTa tiếp tục thể hiện ưu thế so với BERT. Cấu hình mang lại hiệu suất cao nhất trên tập dữ liệu này là RoBERTa kết hợp với Optimizer Adamax và huấn luyện trong 10 epochs, đạt F1 – score là 0.9515 và Accuracy là 0.9508. Đây là mức hiệu suất tốt nhất ghi nhận được trong tất cả các thử nghiệm trên tập Fake or Real News.

Về tương tác giữa optimizer và số epoch trên tập dữ liệu này, với Adamax hiệu suất của RoBERTa có xu hướng tăng lên khi tăng thời gian huấn luyện từ 5 lên 10 epochs (đạt đỉnh ở 10), trong khi đó, với Adam, hiệu suất của RoBERTa lại có xu hướng giảm nhẹ khi tăng từ 5 lên 10 epochs (sau khi đạt đỉnh ở epoch 3 và 5). Điều này có thể gợi ý rằng Adamax ổn định hơn hoặc phù hợp hơn cho việc huấn luyện dài hơn trên dữ liệu có thể nhiều hoặc phức tạp hơn như tập này, trong

khi Adam có thể hội tụ nhanh hơn nhưng lại dễ bị overfitting khi huấn luyện quá lâu với RoBERTa trên tập này.

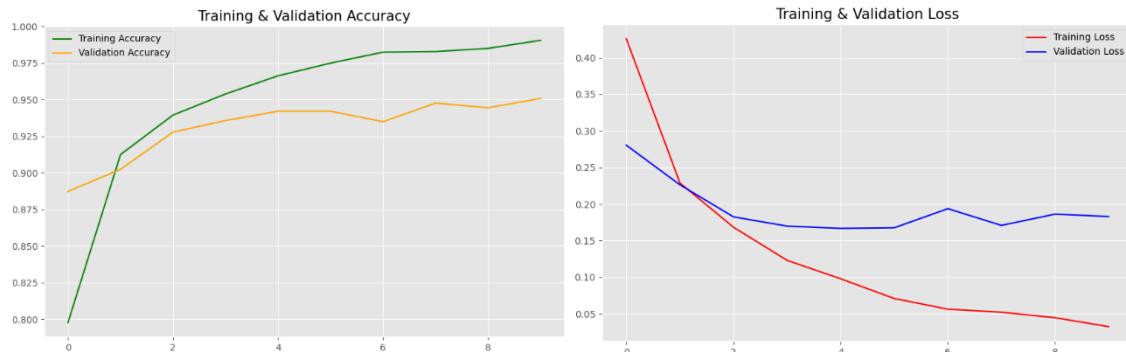
So sánh với mô hình BERT, cấu hình tốt nhất của BERT trên tập này được xác định là BERT + Adam (epoch=5), đạt F1=0.9415 và Acc=0.9413. Mức này thấp hơn đáng kể so với kết quả tốt nhất của RoBERTa. Như vậy, dựa trên việc đạt được cả F1 – score và Accuracy cao nhất, cấu hình RoBERTa + Adamax (epoch=10) được xác định là lựa chọn tối ưu cho tập dữ liệu Fake or Real News, cho thấy sự kết hợp cụ thể giữa mô hình, optimizer và thời gian huấn luyện là rất quan trọng để đạt hiệu quả tốt nhất trên các tập dữ liệu có độ khó khác nhau.

Bảng 8. Kết quả đánh giá hiệu suất trên tập dữ liệu Fake or Real News Dataset

Mô hình	Optimizer	Epoch	F1	R	P	Acc
BERT	Adam	3	0.9259	0.8850	0.9706	0.9286
		5	0.9415	0.9370	0.9460	0.9413
		10	0.9398	0.9465	0.9332	0.9389
	Adamax	3	0.9163	0.9480	0.8866	0.9127
		5	0.9204	0.8835	0.9606	0.9230
		10	0.9270	0.8898	0.9675	0.9294
	AdamW	3	0.9317	0.9134	0.9508	0.9324
		5	0.9247	0.8898	0.9625	0.9270
		10	0.9196	0.8740	0.9703	0.9230
RoBERTa	Adam	3	0.9504	0.9496	0.9511	0.9500
		5	0.9507	0.9559	0.9355	0.9500
		10	0.9398	0.9213	0.9590	0.9405
	Adamax	3	0.9222	0.9339	0.9109	0.9206
		5	0.9426	0.9575	0.9282	0.9413
		10	0.9515	0.9575	0.9456	0.9508

Mô hình	Optimizer	Epoch	F1	R	P	Acc
AdamW		3	0.9427	0.9197	0.9669	0.9437
		5	0.9474	0.9213	0.9750	0.9484
		10	0.9494	0.9307	0.9689	0.9500

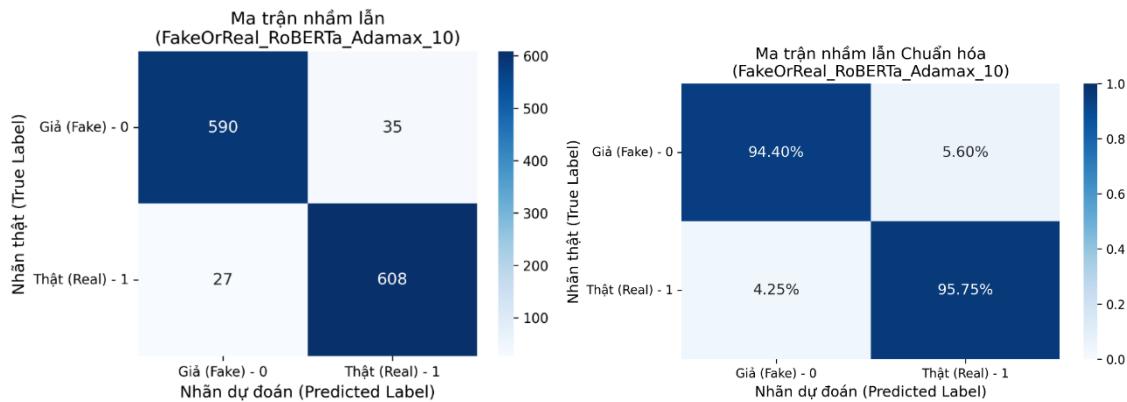
Quan sát đồ thị huấn luyện Hình 16, có thể thấy quá trình học trên tập dữ liệu Fake or Real News phức tạp hơn so với hai tập trước. Mặc dù training accuracy (xanh lá) vẫn tăng liên tục và đạt mức cao (gần 0.99), validation accuracy (cam) lại tăng chậm hơn đáng kể, dao động nhiều hơn và chỉ đạt đỉnh khoảng 0.95 ở các epoch cuối. Khoảng cách giữa đường training và validation accuracy trở nên rất rõ rệt sau khoảng 2 epoch đầu tiên. Tương tự, đồ thị Loss cho thấy trong khi training loss (đỏ) giảm đều đặn xuống mức thấp, thì validation loss (xanh dương) chỉ giảm đến khoảng epoch thứ 3 rồi sau đó lại có xu hướng dao động và không giảm thêm, thậm chí có lúc tăng nhẹ. Sự phân kỳ rõ ràng giữa đường training và validation trên cả hai đồ thị (accuracy và loss) là dấu hiệu mạnh mẽ cho thấy hiện tượng overfitting đã xảy ra tương đối sớm trong quá trình huấn luyện trên tập dữ liệu này. Mặc dù vậy, cấu hình này vẫn đạt được kết quả tốt nhất trên tập kiểm định ở 10 epoch so với các cấu hình khác đã thử nghiệm (như trong Bảng 9), điều này có thể cho thấy sự cần thiết phải huấn luyện đủ dài để mô hình học được đặc trưng phức tạp của bộ dữ liệu này, ngay cả khi phải chấp nhận một mức độ overfitting nhất định.



Hình 16. Đồ thị Training và Validation Accuracy/Loss của mô hình RoBERTa_Adamax_10 trên tập Fake or Real News

Đối với tập Fake or Real News (Hình 17), ma trận nhầm lẫn của mô hình RoBERTa_Adamax_10 phản ánh rõ ràng hiệu suất tổng thể thấp hơn so với hai tập dữ liệu trước, cũng cố nhận định về độ phức tạp cao hơn của dữ liệu này. Mặc dù số dự đoán đúng (TP = 590, TN = 608) vẫn chiếm đa số, số lượng lỗi đã tăng

lên đáng kể. Cụ thể, số tin giả bị bỏ sót ($FN = 35$) cao hơn số tin thật bị phân loại nhầm ($FP = 27$). Phân tích ma trận chuẩn hóa cho thấy tỷ lệ bỏ sót tin giả (Recall lớp Fake) giảm xuống còn khoảng 94.40% ($590 / (590 + 35)$). Trong khi đó, tỷ lệ tin thật bị phân loại nhầm (FP rate) tăng lên khoảng 4.25% ($27 / (27 + 608)$). Kết quả này chỉ ra rằng mô hình gặp nhiều khó khăn hơn đáng kể trên tập dữ liệu này, đặc biệt là trong việc nhận diện đầy đủ các tin tức giả, đồng thời cũng mắc nhiều lỗi hơn trong việc phân loại tin thật.



Hình 17. Ma trận nhầm lẫn/chuẩn hóa mô hình RoBERTa_Adamax_10 trên tập dữ liệu Fake or Real

Cuối cùng, nghiên cứu đánh giá hiệu suất trên tập dữ liệu Fake News Detection Dataset, với kết quả chi tiết được trình bày trong Bảng 9. Hiệu suất tổng thể trên tập dữ liệu này nhìn chung khá cao, với các chỉ số F1 và Accuracy tốt nhất đạt khoảng 0.98, cao hơn so với tập Fake or Real News nhưng chưa đạt đến mức gần như tuyệt đối của tập ISOT hay mức cao nhất của tập Fake News. Điều này cho thấy độ khó của tập dữ liệu này nằm ở mức trung bình – cao trong số các tập được khảo sát.

Phân tích chi tiết Bảng 9 cho thấy cấu hình RoBERTa kết hợp với Optimizer AdamW và huấn luyện trong 10 epochs mang lại hiệu suất toàn diện tốt nhất, đạt F1 – score cao nhất là 0.9800 và đồng thời cũng có Accuracy cao nhất là 0.9825. Cấu hình này cũng cho thấy sự cân bằng tốt giữa Precision (0.9744) và Recall (0.9856).

Về ảnh hưởng của số epoch đối với RoBERTa trên tập dữ liệu này, có sự khác biệt giữa các optimizer. Tương tự như trên tập Fake or Real News, Adamax và AdamW đều cho thấy xu hướng cải thiện hiệu suất khi tăng số epoch từ 3 lên 10, với kết quả tốt nhất đạt được ở 10 epochs. Ngược lại, RoBERTa + Adam lại đạt hiệu suất F1 cao nhất ở 5 epochs (0.9768) và có xu hướng giảm nhẹ khi tăng

lên 10 epochs. Điều này tiếp tục củng cố nhận định rằng sự lựa chọn optimizer có ảnh hưởng đến thời gian huấn luyện tối ưu.

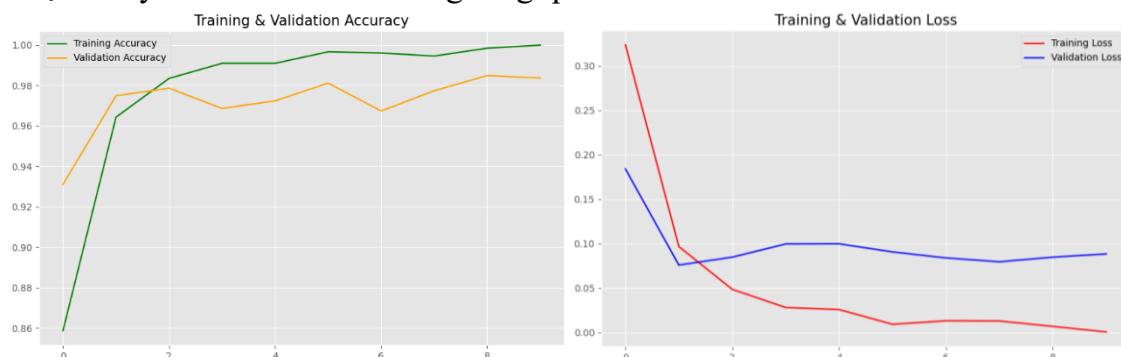
So sánh các optimizer cho RoBERTa, AdamW nổi lên là lựa chọn hàng đầu trên tập dữ liệu này khi đạt được cả F1 và Accuracy cao nhất ở 10 epochs. Adamax cũng cho kết quả rất cạnh tranh ở 10 epochs ($F1=0.9771$), trong khi Adam đạt đỉnh sớm hơn ở 5 epochs. Khi so sánh giữa hai kiến trúc mô hình, RoBERTa tiếp tục duy trì lợi thế so với BERT. Cấu hình tốt nhất của RoBERTa ($F1=0.9800$, $Acc=0.9825$) vượt trội hơn cấu hình tốt nhất của BERT là BERT + Adam (epoch=3) ($F1=0.9714$, $Acc=0.9749$). Tóm lại, dựa trên kết quả F1 – score và Accuracy cao nhất, cấu hình RoBERTa + AdamW (epoch=10) được xác định là mô hình tối ưu cho tập dữ liệu Fake News Detection Dataset.

Bảng 9. Kết quả đánh giá hiệu suất trên tập dữ liệu Fake News Detection Dataset

Mô hình	Optimizer	Epoch	F1	R	P	Acc
BERT	Adam	3	0.9714	0.9770	0.9659	0.9749
		5	0.9517	0.9914	0.9151	0.9561
		10	0.9608	0.9856	0.9342	0.9649
	Adamax	3	0.9118	0.9799	0.8525	0.9173
		5	0.9576	0.9741	0.9417	0.9624
		10	0.9644	0.9741	0.9549	0.9687
	AdamW	3	0.9582	0.9885	0.9297	0.9624
		5	0.9556	0.9885	0.9247	0.9599
		10	0.9542	0.9885	0.9223	0.9586
RoBERTa	Adam	3	0.9692	0.9943	0.9454	0.9724
		5	0.9768	0.9684	0.9854	0.9800
		10	0.9701	0.9799	0.9606	0.9737
	Adamax	3	0.9616	0.9713	0.9521	0.9662
		5	0.9685	0.9713	0.9657	0.9724

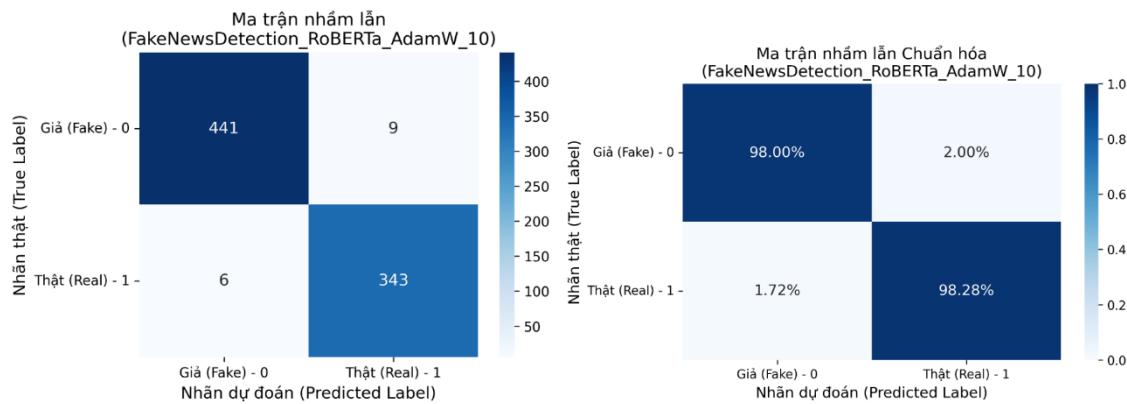
Mô hình	Optimizer	Epoch	F1	R	P	Acc
AdamW	AdamW	10	0.9771	0.9828	0.9716	0.9800
		3	0.9714	0.9741	0.9686	0.9749
		5	0.9787	0.9914	0.9664	0.9812
		10	0.9815	0.9885	0.9745	0.9837

Quá trình huấn luyện mô hình RoBERTa + AdamW (10 epochs) trên tập Fake News Detection được thể hiện qua . Đồ thị cho thấy mô hình học rất nhanh trên tập huấn luyện, với training loss (đỏ) giảm mạnh về gần 0 và training accuracy (xanh lá) tăng lên đến 1.0 (100%) sau 9 epoch. Tuy nhiên, các đường validation lại cho thấy một diễn biến khác biệt rõ rệt. Validation loss (xanh dương) đạt giá trị cực tiểu rất sớm (ngay tại epoch đầu tiên), sau đó tăng lên đáng kể ở các epoch 2-3 và dao động ở mức cao hơn (quanh 0.08 - 0.10) trong suốt phần còn lại của quá trình huấn luyện mà không có dấu hiệu cải thiện thêm. Tương tự, validation accuracy (cam) cũng đạt gần mức định (khoảng 0.98) ngay tại epoch thứ 2 và sau đó dao động mà không cải thiện đáng kể, trong khi training accuracy tiếp tục tăng mạnh. Sự phân kỳ rất rõ nét giữa đường training và validation trên cả hai đồ thị, đặc biệt là việc validation loss tăng lên rõ rệt sau các epoch đầu, là bằng chứng mạnh mẽ cho thấy hiện tượng overfitting xảy ra khá sớm và rõ rệt khi huấn luyện mô hình này trên tập dữ liệu Fake News Detection. Mặc dù cấu hình 10 epoch này đạt điểm số F1 và Accuracy cao nhất trên tập kiểm thử trong bảng kết quả (Bảng 9), đồ thị huấn luyện cho thấy tiềm năng hiệu suất tốt nhất trên dữ liệu chưa biết có thể đã đạt được ở các epoch rất sớm (epoch 1 hoặc 3). Việc huấn luyện kéo dài đến 10 epoch đã khiến mô hình học quá khớp với dữ liệu training, điều này cần được lưu ý khi xem xét khả năng tổng quát hóa của nó.



Hình 18. Đồ thị Training và Validation Loss của mô hình RoBERTa_AdamW_10 trên tập Fake News Detection

Trên tập Fake News Detection (Hình 19), mô hình RoBERTa_AdamW_10 cho thấy hiệu suất tốt trở lại, vượt trội so với tập Fake or Real News. Ma trận nhầm lẫn ghi nhận số lỗi tương đối thấp, với chỉ 9 tin giả bị bỏ sót ($FN = 9$) và 6 tin thật bị phân loại nhầm ($FP = 6$). Số lượng dự đoán đúng $TP = 441$ (Tin Giả) và $TN = 343$ (Tin Thật) chiếm ưu thế rõ rệt. Từ ma trận chuẩn hóa, có thể thấy Recall cho lớp tin giả đạt mức cao khoảng 98.00% ($441 / (441 + 9)$), cho thấy khả năng phát hiện tin giả tốt. Tỷ lệ báo động giả (FP rate) cũng được giữ ở mức thấp, khoảng 1.72% ($6 / (6 + 343)$).



Hình 19. Ma trận nhầm lẫn chuẩn hóa mô hình RoBERTa_AdamW_10 trên tập dữ liệu Fake News Detection

Nhìn chung, qua quá trình thực nghiệm và điều chỉnh siêu tham số trên bốn tập dữ liệu khác nhau, có thể thấy rằng không có một cấu hình duy nhất nào là tối ưu cho tất cả. Mô hình RoBERTa thường xuyên cho thấy ưu thế về hiệu suất so với BERT, tuy nhiên, optimizer và số epoch huấn luyện tốt nhất lại phụ thuộc đáng kể vào đặc tính của từng tập dữ liệu cụ thể. Điều này nhấn mạnh tầm quan trọng của việc thực hiện tinh chỉnh siêu tham số một cách cẩn thận cho từng bài toán và bộ dữ liệu riêng biệt để đạt được kết quả tốt nhất.

4.4 So sánh và Thảo luận kết quả

Sau khi xác định được các cấu hình mô hình BERT và RoBERTa tối ưu cho từng bộ dữ liệu thông qua quá trình tinh chỉnh siêu tham số, nghiên cứu sẽ tiến hành so sánh hiệu suất đạt được với các kết quả từ những nghiên cứu trước đó đã công bố trên cùng các bộ dữ liệu. Việc so sánh này nhằm mục đích đánh giá hiệu quả của các mô hình Transformer được đề xuất trong bối cảnh các phương pháp đã có. Kết quả so sánh chi tiết được trình bày trong các Bảng: Bảng 10, Bảng 11, Bảng 12, Bảng 13.

Trên tập dữ liệu ISOT (Bảng 10): Các mô hình đề xuất của nghiên cứu, cả BERT ($F1=0.9997$, $Acc=0.9997$) và RoBERTa ($F1=0.9998$, $Acc=0.9998$), đều cho thấy hiệu suất vượt trội. Kết quả này không chỉ cao hơn đáng kể so với các

phương pháp dựa trên máy học truyền thống như Decision Tree + TF – IDF [10] hay Linear SVM + TF – IDF [11], mà còn nhỉnh hơn các phương pháp học sâu trước đó như Random Forest + LIWC [9], CNN, ResNet và BiLSTM sử dụng word embedding như fastText hoặc GloVe [21]. Đáng chú ý, kết quả của nghiên cứu này cũng cao hơn so với một nghiên cứu khác sử dụng BERT [14] ($F1/Acc=0.9984$). Điều này khẳng định sức mạnh của kiến trúc Transformer và hiệu quả của quá trình tinh chỉnh siêu tham số trên bộ dữ liệu ISOT, với mô hình RoBERTa đạt được kết quả gần như hoàn hảo.

Bảng 10. So sánh mô hình đề xuất với các nghiên cứu trước đó trên tập dữ liệu ISOT fake news

Nghiên cứu	Mô hình phân loại	Word embedding	F1	R	P	Acc
Ozbay & Alatas [10]	Decision tree	TF – IDF	0.9680	0.9730	0.9630	0.9680
Ahmad et al.[11]	Linear SVM	TF – IDF	–	–	–	0.9200
Ahmed et al.[9]	Random forest	LIWC	0.9900	1	0.9900	0.9900
Sastrawan et al.[21]	CNN	fastText	0.9989	0.9988	0.9989	0.9988
	ResNet	GloVe	0.9990	0.9990	0.9991	0.9990
	BiLSTM	GloVe	0.9995	0.9995	0.9995	0.9995
Mimura & Ishimaru [14]	BERT		0.9984	–	–	0.9984
Ours	BERT		0.9997	0.9998	0.9995	0.9997
Ours	RoBERTa		0.9998	0.9995	1	0.9998

Trên tập dữ liệu Fake News (Bảng 11): Mô hình RoBERTa của nghiên cứu này ($F1=0.9986$) tiếp tục thể hiện sự vượt trội rõ rệt, đạt hiệu suất cao hơn đáng kể so với tất cả các nghiên cứu được liệt kê, bao gồm cả các mô hình học sâu mạnh như CNN [15], [21], ResNet và BiLSTM [21] (với F1 cao nhất trước đó là 0.9865).

Mô hình BERT của nghiên cứu này ($F1=0.9890$) cũng cho kết quả rất cạnh tranh và cao hơn phần lớn các nghiên cứu trước đó. Kết quả này cho thấy các mô hình Transformer, đặc biệt là RoBERTa với cấu hình tối ưu, mang lại cải thiện đáng kể về hiệu suất trên tập dữ liệu này.

Bảng 11. So sánh mô hình đề xuất với các nghiên cứu trước đó trên tập dữ liệu Fake news

Nghiên cứu	Mô hình phân loại	Word embedding	F1	R	P	Acc
Kaliyar et al.[15]	CNN	GloVe	0.9812	0.9688	0.9940	0.9836
Ahmad et al.[11]	Bagging + Decision tree	LIWC	0.9400	0.9500	0.9400	0.9400
Sastrawan et al.[21]	CNN	fastText	0.9691	0.9698	0.9686	0.9692
	ResNet	fastText	0.9810	0.9809	0.9811	0.9811
	BiLSTM	fastText	0.9865	0.9866	0.9864	0.9865
Ours	BERT		0.9890	0.9825	0.9956	0.9888
Ours	RoBERTa		0.9986	0.9913	0.9860	0.9884

Trên tập dữ liệu Fake or Real News (Bảng 12): Đây là tập dữ liệu mà các mô hình nói chung đạt hiệu suất thấp hơn. Tuy nhiên, mô hình RoBERTa của nghiên cứu này ($F1=0.9515$) vẫn đạt được kết quả tốt nhất khi so sánh với các nghiên cứu trước đó, vượt qua các mô hình LSTM/UniLSTM [18], [16] và các kiến trúc CNN/ResNet/BiLSTM [21]. Đáng chú ý, RoBERTa của nghiên cứu này nhỉnh hơn so với mô hình BiLSTM mạnh nhất trước đó ($F1=0.9459$) [21]. Mô hình BERT của nghiên cứu này ($F1=0.9415$) cũng cho kết quả cạnh tranh, gần tương đương với BiLSTM [21] nhưng thấp hơn RoBERTa. Điều này một lần nữa khẳng định ưu thế của RoBERTa trên tập dữ liệu phức tạp hơn này.

Bảng 12. So sánh mô hình đè xuất với các nghiên cứu trước đó trên tập dữ liệu Fake or Real news

Nghiên cứu	Mô hình phân loại	Word embedding	F1	R	P	Acc
Deepak & Chitturi [18]	LSTM	Word2Vec	–	–	–	0.9130
Bahad et al.[16]	UniLSTM	GloVe	–	–	–	0.9148
Sastrawan et al.[21]	CNN	fastText	0.9189	0.9193	0.9188	0.9190
	ResNet	fastText	0.8888	0.8911	0.8936	0.8888
	BiLSTM	GloVe	0.9459	0.9464	0.9188	0.9460
Ours	BERT		0.9415	0.9370	0.9460	0.9413
Ours	RoBERTa		0.9515	0.9575	0.9456	0.9508

Trên tập dữ liệu Fake News Detection (Bảng 13): Mô hình RoBERTa ($F1=0.9815$) và BERT ($F1=0.9714$) của nghiên cứu vẫn đạt hiệu suất cao, vượt trội so với Random Forest + LIWC [11] ($F1=0.9500$). Tuy nhiên, khi so sánh với các mô hình học sâu khác trong nghiên cứu của Sastrawan et al. [16], kết quả của nghiên cứu này lại thấp hơn so với mô hình ResNet ($F1=0.9899$) và đặc biệt là BiLSTM ($F1=0.9924$). Ngoài ra, mô hình BiLSTM của Bahad et al. [16] cũng được báo cáo đạt Accuracy 0.9875, cao hơn mức Accuracy của RoBERTa nghiên cứu này (0.9837). Sự khác biệt này có thể xuất phát từ nhiều yếu tố như sự khác biệt nhỏ trong quá trình tiền xử lý dữ liệu, cách chia tập train/test, hoặc các chi tiết triển khai mô hình ResNet/BiLSTM cụ thể trong nghiên cứu [16] và [21] mà nghiên cứu chưa tái tạo hoàn toàn. Mặc dù mô hình Transformer của nghiên cứu không đạt kết quả cao nhất trên tập dữ liệu này so với một số báo cáo trước, hiệu suất đạt được vẫn ở mức rất cao và cạnh tranh.

Bảng 13. So sánh mô hình đề xuất với các nghiên cứu trước đó trên tập dữ liệu Fake news detection

Nghiên cứu	Mô hình phân loại	Word embedding	F1	R	P	Acc
Bahad et al.[16]	BiLSTM	GloVe	—	—	—	0.9875
Ahmad et al.[11]	Random forest	LIWC	0.9500	0.9300	0.9800	0.9500
Sastrawan et al.[21]	CNN	GloVe	0.9820	0.9809	0.9832	0.9824
	ResNet	GloVe	0.9899	0.9905	0.9889	0.9897
	BiLSTM	fastText	0.9924	0.9919	0.9926	0.9923
Ours	BERT		0.9714	0.9770	0.9659	0.9749
Ours	RoBERTa		0.9815	0.9885	0.9745	0.9837

Tổng kết so sánh: Nhìn chung, các mô hình dựa trên kiến trúc Transformer như BERT và đặc biệt là RoBERTa được đề xuất trong luận văn này đã chứng tỏ hiệu quả mạnh mẽ trên phần lớn các tập dữ liệu thử nghiệm, thường xuyên đạt được kết quả vượt trội với các nghiên cứu trước đó, bao gồm cả các phương pháp máy học truyền thống và các kiến trúc học sâu như CNN, LSTM, BiLSTM. Mô hình RoBERTa thường cho thấy ưu thế hơn so với BERT. Tuy nhiên, kết quả trên tập dữ liệu Fake News Detection cho thấy hiệu suất của mô hình phụ thuộc vào sự tương tác phức tạp giữa kiến trúc, dữ liệu và các yếu tố thực nghiệm khác, và không phải lúc nào các mô hình mới nhất cũng đảm bảo kết quả tốt nhất trên mọi bộ dữ liệu so với các báo cáo đã có.

Sau khi các mô hình tốt nhất cho từng tập dữ liệu được chọn nghiên cứu này thực hiện kiểm thử chéo bộ dữ liệu. Đây là một bước quan trọng để đánh giá khả năng tổng quát hóa của các mô hình khi chúng đối mặt với dữ liệu khác biệt so với dữ liệu đã được huấn luyện. Cụ thể, mô hình tốt nhất được huấn luyện trên một tập dữ liệu sẽ được đánh giá hiệu suất trên các tập dữ liệu còn lại. Kết quả chi tiết được thể hiện trong Bảng 14.

Bảng 14. Kiểm thử chéo với các tập dữ liệu

Mô hình huấn luyện	Dữ liệu huấn luyện	Dữ liệu kiểm thử	F1	R	P	Acc	Avg F1
RoBERTa _Adamax _10	ISOT fake news	Fake news	0.3208	0.1999	0.8123	0.5688	0.4204
		Fake or Real news	0.3047	0.1874	0.8151	0.5691	
		Fake news detection	0.6358	0.4741	0.9649	0.7632	
RoBERTa _Adam_5	Fake news	ISOT fake news	0.6719	0.5423	0.8829	0.7446	0.5209
		Fake or Real news	0.4769	0.3165	0.9664	0.6500	
		Fake news detection	0.4138	0.2759	0.8276	0.6592	
RoBERTa _Adamax _10	Fake or Real news	ISOT fake news	0.5348	0.5172	0.5537	0.5662	0.6079
		Fake news	0.7751	0.6698	0.9198	0.8020	
		Fake news detection	0.5139	0.4253	0.6491	0.6491	
RoBERTa _AdamW _10	Fake news detection	ISOT fake news	0.8036	0.9998	0.6718	0.7644	0.7558
		Fake news	0.7521	0.8174	0.6964	0.7301	
		Fake or Real news	0.7117	0.7378	0.6874	0.6954	

Quan sát Bảng 14 nổi bật nhất là sự sụt giảm hiệu suất rất đáng kể của tất cả các mô hình khi được kiểm thử trên những tập dữ liệu khác biệt so với tập huấn luyện gốc. Các chỉ số F1 – score và Accuracy, vốn đạt mức rất cao trong thử nghiệm trên cùng bộ dữ liệu, nay đã giảm xuống đáng kể, phần lớn chỉ còn dao động trong khoảng 0.3 đến dưới 0.8. Ví dụ điển hình là mô hình

RoBERTa_Adamax_10 huấn luyện trên ISOT (vốn đạt F1 gần 1.0 trên tập test ISOT) chỉ đạt F1 – score lần lượt là 0.3208 và 0.3047 khi kiểm thử trên Fake news và Fake or Real news. Ngược lại, mô hình RoBERTa_AdamW_10 huấn luyện trên Fake News Detection cho thấy khả năng khá hơn khi kiểm thử trên ISOT ($F1=0.8036$) nhưng vẫn thấp hơn nhiều so với hiệu suất gốc của nó trên chính tập Fake News Detection.

Sự sụt giảm mạnh mẽ này cho thấy các mô hình có xu hướng học các đặc điểm rất đặc trưng của tập dữ liệu huấn luyện, và những đặc điểm này không dễ dàng tổng quát hóa tốt sang các bộ dữ liệu khác nhau về nguồn tin, chủ đề, hay phong cách viết. Để đánh giá khả năng tổng quát hóa trung bình, Bảng 14 cũng trình bày chỉ số Average F1 – score (Avg F1), được tính bằng trung bình cộng F1 – score của một mô hình huấn luyện trên ba tập dữ liệu kiểm thử còn lại.

Phân tích cột Avg F1 cho thấy:

Mô hình huấn luyện trên ISOT có khả năng tổng quát hóa trung bình kém nhất với $\text{Avg F1} \approx 0.4204$.

Mô hình huấn luyện trên Fake News và Fake or Real News cho kết quả khá hơn một chút $\text{Avg F1} \approx 0.5209$ và ≈ 0.6079 .

Mô hình được huấn luyện trên Fake News Detection Dataset RoBERTa_AdamW_10 thể hiện khả năng tổng quát hóa trung bình tốt nhất khi đạt $\text{Avg F1} \approx 0.7558$.

Một điểm đáng chú ý là mô hình (RoBERTa_AdamW_10 huấn luyện trên Fake News Detection), mặc dù thể hiện dấu hiệu overfitting khá rõ trong quá trình huấn luyện trên tập dữ liệu gốc của nó (như đã phân tích ở mục 4.3), lại cho thấy khả năng tổng quát hóa trung bình (Avg F1 cao nhất) tốt hơn các mô hình khác (mô hình huấn luyện trên ISOT vốn có kết quả gần như hoàn hảo trên tập test gốc nhưng Avg F1 lại thấp nhất). Điều này có thể được lý giải bởi sự khác biệt về đặc tính của các tập dữ liệu. Tập ISOT có thể chứa các đặc điểm hoặc nguồn tin rất riêng biệt, khiến mô hình học được các quy tắc quá chuyên biệt và hoạt động kém khi gặp dữ liệu khác. Ngược lại, tập Fake News Detection, dù có thể “nhiều” hơn và khiến mô hình overfit với các chi tiết của nó, lại có thể chứa các đặc trưng ngôn ngữ hoặc các dạng tin giả phổ biến hơn, giúp mô hình có khả năng nhận diện tốt hơn trên dữ liệu đa dạng hơn từ các nguồn khác.

Kết quả kiểm thử chéo này cung cấp thông tin then chốt cho việc lựa chọn mô hình triển khai thực tế. Mặc dù một số mô hình (huấn luyện trên ISOT, Fake News) đạt hiệu suất rất cao trên tập dữ liệu nguồn, khả năng tổng quát hóa kém khi đối mặt với dữ liệu khác biệt cho thấy chúng không phải là lựa chọn tối ưu cho ứng dụng cần xử lý tin tức đa dạng. Do đó, dựa trên khả năng tổng quát hóa trung

bình tốt nhất ($\text{Avg F1} \approx 0.7558$) thể hiện qua kiểm thử chéo, nghiên cứu này đề xuất lựa chọn mô hình RoBERTa_AdamW_10 (huấn luyện trên Fake News Detection Dataset) làm nền tảng cho ứng dụng “Fake News Detector”. Lựa chọn này ưu tiên sự ổn định và hiệu quả trên nhiều loại dữ liệu không biết trước, dù hiệu suất có thể không phải là cao nhất tuyệt đối trên một tập dữ liệu đơn lẻ. Tuy nhiên, sự sụt giảm hiệu suất chung khi kiểm thử chéo vẫn nhán mạnh sự cần thiết của các hướng nghiên cứu tiếp theo để nâng cao hơn nữa tính tổng quát và độ tin cậy của hệ thống phát hiện tin giả.

4.5 Phát triển ứng dụng

Mục tiêu của đề tài là xây dựng một ứng dụng web hoàn chỉnh, có khả năng hỗ trợ người dùng phân loại tin tức thật/giả một cách tự động. Để hiện thực hóa mục tiêu này, ứng dụng “Fake News Detector” đã được phát triển dựa trên kiến trúc client – server hiện đại, tích hợp các công nghệ tiên tiến:

Frontend: Giao diện người dùng (UI) được xây dựng bằng thư viện React.js, kết hợp với React Bootstrap để đảm bảo tính đáp ứng (responsive), trực quan và thân thiện trên nhiều thiết bị. Thư viện Axios đảm nhiệm việc thực hiện các yêu cầu HTTP đến backend một cách hiệu quả. React Router DOM được sử dụng để quản lý việc điều hướng mượt mà giữa các trang trong ứng dụng. Context API của React đóng vai trò quan trọng trong việc quản lý trạng thái chung, như thông tin người dùng đăng nhập và số lượt sử dụng chức năng phân loại còn lại.

Backend: Nền tảng Node.js cùng với framework Express.js được chọn để xây dựng các API endpoint theo chuẩn RESTful, tạo sự linh hoạt và dễ dàng tích hợp. Mongoose được sử dụng làm ODM (Object Data Modeling), giúp đơn giản hóa việc tương tác với cơ sở dữ liệu MongoDB. Quá trình xác thực người dùng và phân quyền được xử lý an toàn bằng JSON Web Tokens (JWT) kết hợp với middleware tùy chỉnh. Thư viện Bcryptjs đảm bảo mật khẩu người dùng được mã hóa một cách mạnh mẽ trước khi lưu trữ. Multer được tích hợp để xử lý việc tải lên tập tin, cụ thể là ảnh đại diện của người dùng.

Cơ sở dữ liệu: Hệ quản trị cơ sở dữ liệu NoSQL MongoDB được lựa chọn để lưu trữ linh hoạt thông tin người dùng, lịch sử các lần phân loại tin tức, và các báo cáo về kết quả phân loại chưa chính xác do người dùng gửi lên.

API Phân loại: Một thành phần quan trọng là API riêng biệt, được xây dựng bằng Python với framework Flask, chuyên trách thực thi mô hình học máy. API này tiếp nhận văn bản (tiêu đề và nội dung tin tức) từ backend

Node.js, thực hiện các bước cần thiết như dịch thuật (nếu cần), tiền xử lý dữ liệu, sau đó đưa vào mô hình và trả về kết quả phân loại chi tiết (bao gồm xác suất tin thật/giả, phân tích cảm xúc, và các từ khóa có ảnh hưởng đến dự đoán).

Ứng dụng được thiết kế với các chức năng cốt lõi, bao gồm:

Quản lý tài khoản: Cho phép người dùng đăng ký, đăng nhập, cập nhật thông tin hồ sơ cá nhân, và thay đổi mật khẩu.

Phân loại tin tức: Người dùng có thể nhập liệu (tiêu đề, nội dung), nhận kết quả phân loại (thật/giả), xem xác suất dự đoán, phân tích cảm xúc và các từ khóa ảnh hưởng.

Lịch sử phân loại: Xem lại các tin tức đã được phân loại trước đó.

Báo cáo sai sót: Gửi phản hồi nếu người dùng cho rằng kết quả phân loại chưa chính xác.

Khu vực quản trị (Admin): Giao diện riêng dành cho quản trị viên để quản lý người dùng, xem xét các tin tức đã phân loại và duyệt các báo cáo sai sót.

Quá trình phát triển ứng dụng bao gồm các giai đoạn chính: thiết kế giao diện người dùng (UI/UX), xây dựng các component phía frontend, phát triển các API phía backend, thiết kế lược đồ cơ sở dữ liệu và thực hiện kiểm thử toàn diện các chức năng.

4.5.1 Đối tượng sử dụng

Ứng dụng hướng đến những người dùng có nhu cầu kiểm tra và phân loại tin tức thật hay giả một cách tự động và nhanh chóng.

4.5.2 Chức năng của ứng dụng

- Đăng ký tài khoản người dùng mới.
- Đăng nhập vào hệ thống bằng tài khoản đã đăng ký.
- Phân loại tin tức:
 - o Thực hiện phân loại để xác định tin tức là thật hay giả.
 - o Xem lại lịch sử các tin tức đã được phân loại.
 - o Báo cáo sai sót nếu kết quả phân loại chưa chính xác.
- Quản lý tài khoản cá nhân (cập nhật thông tin, đổi mật khẩu).
- (Dành cho Admin) Quản lý người dùng và quản lý tin tức/báo cáo.

4.5.3 Kết quả

Sau quá trình phân tích, thiết kế và triển khai, ứng dụng “Fake News Detector” đã được hoàn thiện, đáp ứng các mục tiêu và chức năng đã đề ra. Các chức năng chính cho cả người dùng thông thường và người quản trị (Admin) đều

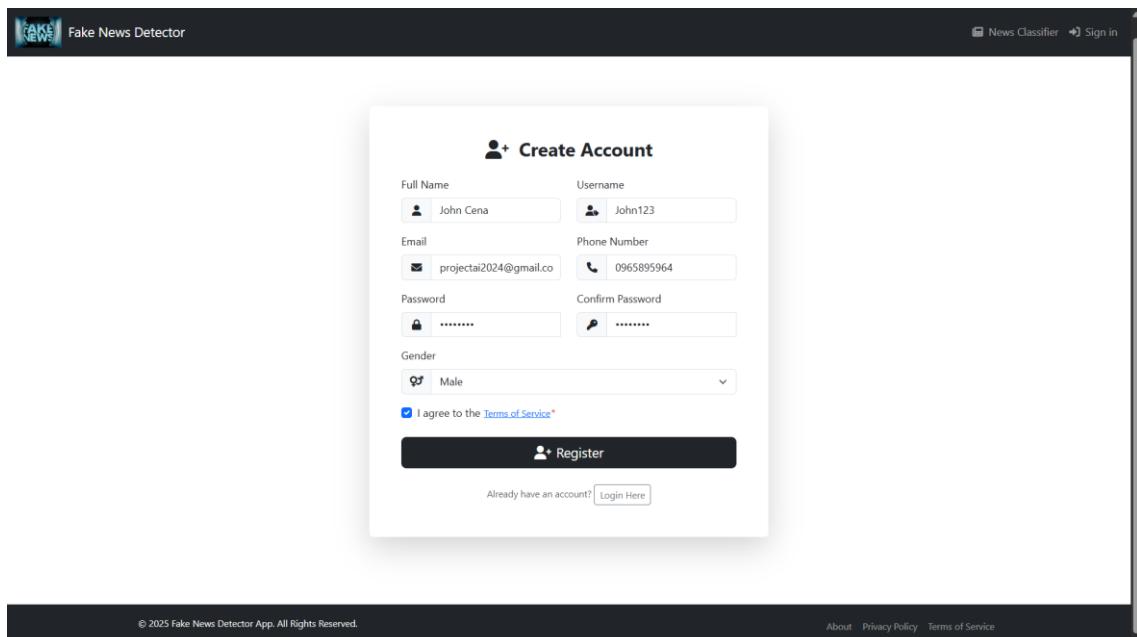
Luận văn: Transformer để phân loại tin giả

đã được triển khai thành công với giao diện thân thiện, trực quan và đáp ứng tốt. Kết quả cụ thể của từng chức năng được minh họa chi tiết qua các hình ảnh và mô tả đi kèm dưới đây.

Chức năng dành cho Người dùng:

Đăng ký và Đăng nhập: Người dùng có thể dễ dàng tạo tài khoản mới và đăng nhập an toàn vào hệ thống bằng username/email và mật khẩu đã được mã hóa.

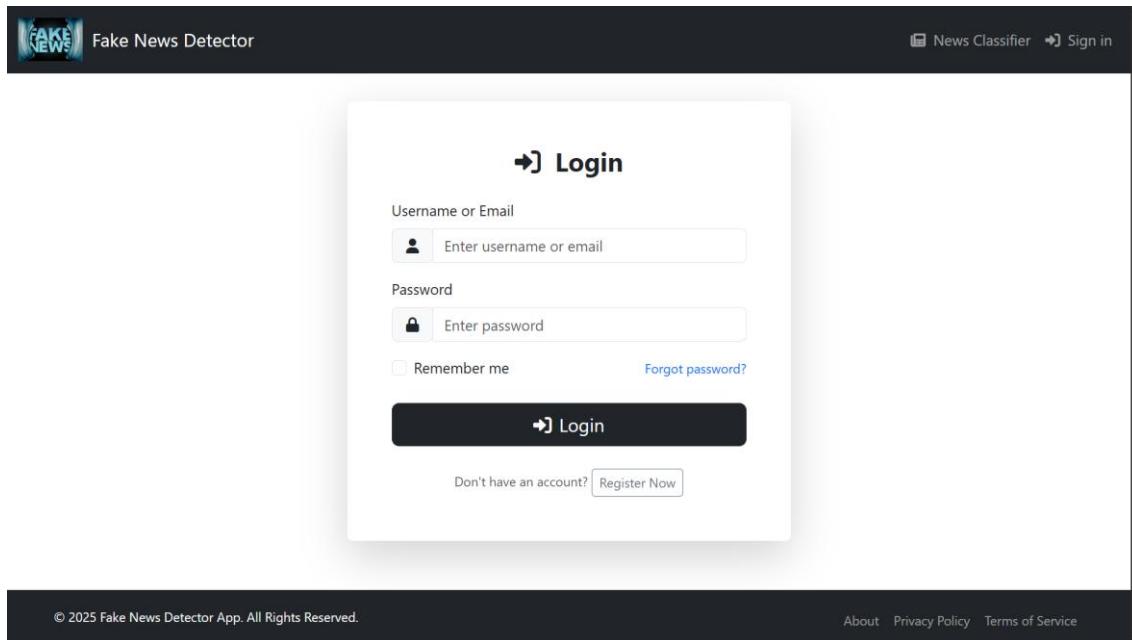
Giao diện cho phép người dùng nhập thông tin cần thiết để tạo một tài khoản mới.(Hình 20)



Hình 20. Giao diện Đăng ký tài khoản mới cho người dùng

Giao diện cho phép người dùng đã có tài khoản đăng nhập vào ứng dụng.(Hình 21)

Luận văn: Transformer để phân loại tin giả

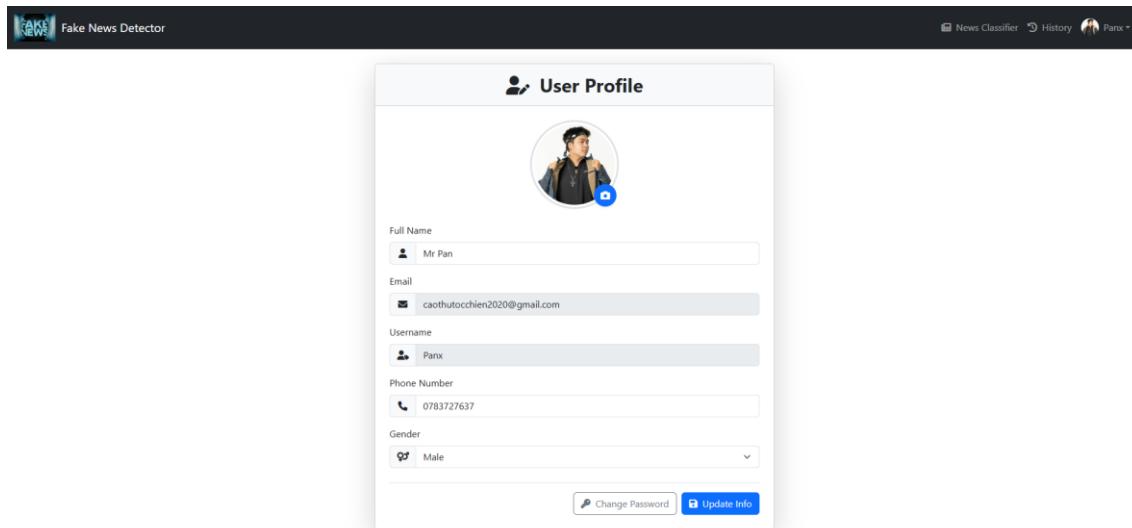


© 2025 Fake News Detector App. All Rights Reserved. About Privacy Policy Terms of Service

Hình 21. Giao diện đăng nhập vào hệ thống

Quản lý Hồ sơ: Người dùng có thể quản lý thông tin cá nhân cơ bản, cập nhật ảnh đại diện và thay đổi mật khẩu.

Trang quản lý thông tin cá nhân, cho phép cập nhật họ tên, số điện thoại, giới tính và thay đổi ảnh đại diện (với chức năng cắt ảnh). Chức năng đổi mật khẩu cũng được tích hợp tại đây.(Hình 22)



© 2025 Fake News Detector App. All Rights Reserved.

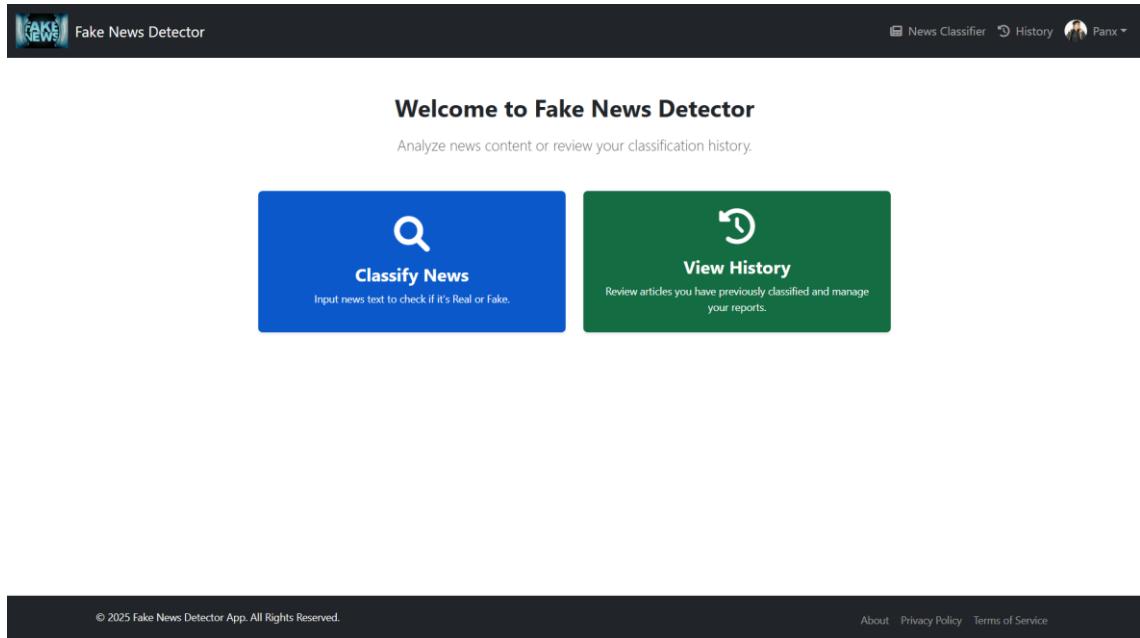
About Privacy Policy Terms of Service

Hình 22. Giao diện trang Hồ sơ người dùng

Trang chủ: Giao diện chính sau khi đăng nhập, cung cấp lối truy cập nhanh đến các chức năng.

Luận văn: Transformer để phân loại tin giả

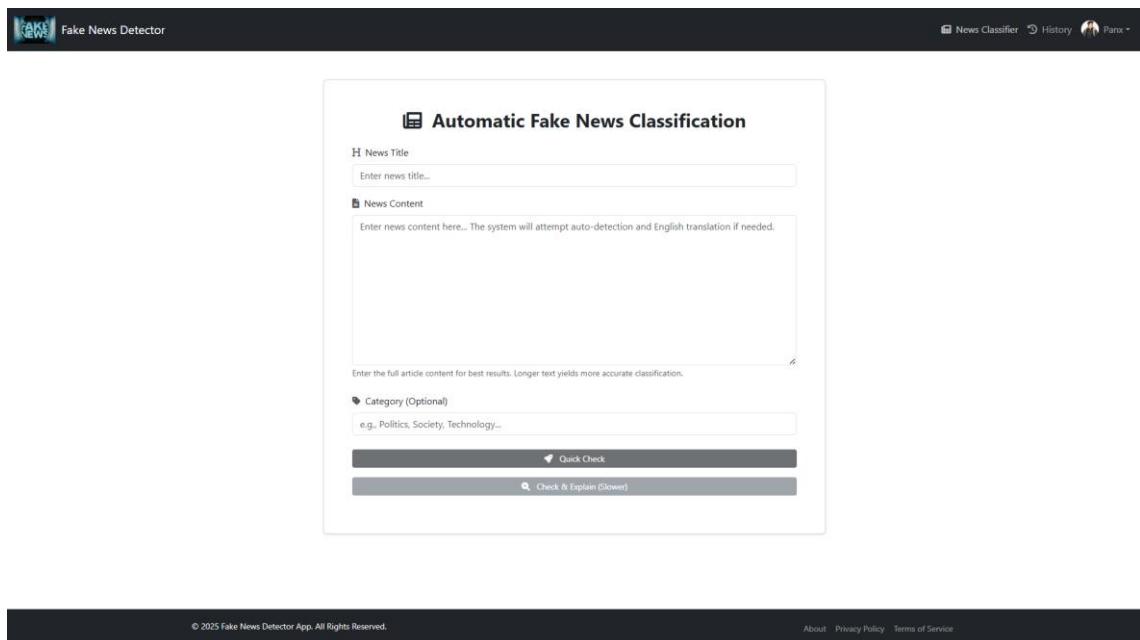
Giao diện trang chủ thân thiện, hiển thị các thông tin tổng quan và điều hướng đến các chức năng chính.(Hình 23)



Hình 23. Giao diện Trang chủ dành cho người dùng

Phân loại Tin tức: Chức năng cốt lõi cho phép người dùng nhập tin tức và nhận kết quả phân loại chi tiết.

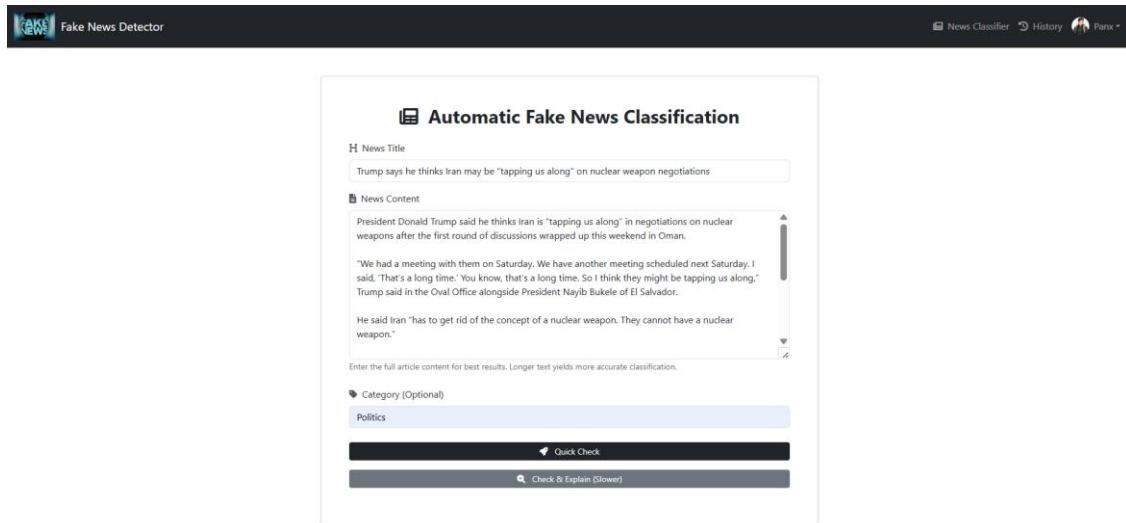
Quy trình và kết quả của chức năng phân loại tin tức.(Hình 24)



Hình 24. Quy trình và kết quả của chức năng phân loại tin tức

Người dùng nhập tiêu đề và nội dung của tin tức cần kiểm tra vào các trường tương ứng. (Hình 24.a)

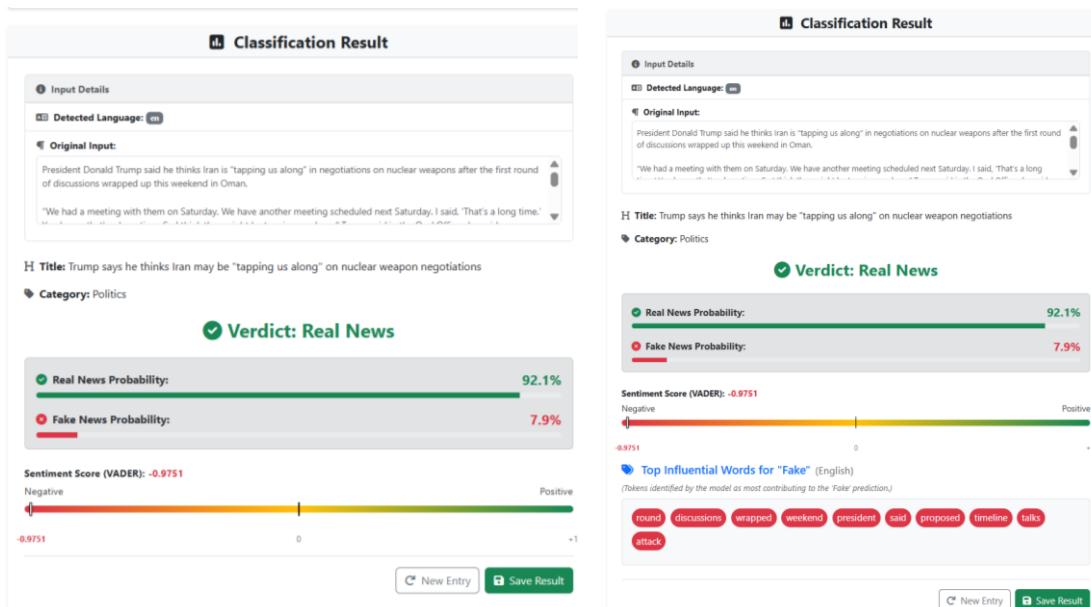
Luận văn: Transformer để phân loại tin giả



Hình 24.a. Giao diện nhập liệu

Hiển thị kết quả phân loại nhanh chóng, bao gồm nhãn dự đoán (Real/Fake) và tỷ lệ phần trăm xác suất. (Hình 24.b)

Cung cấp kết quả phân tích sâu hơn, bao gồm cả phân tích cảm xúc và các yếu tố ảnh hưởng đến dự đoán. (Hình 24.c)



Hình 24.b. Kết quả kiểm tra nhanh

Lịch sử Phân loại: Người dùng có thể xem lại, báo cáo sai sót hoặc xóa các tin tức đã phân loại.

Trang hiển thị danh sách các tin tức người dùng đã phân loại trước đó và các thao tác liên quan.(Hình 25)

Hình 24.c. Kết quả kiểm tra và phân tích chi tiết

Luận văn: Transformer để phân loại tin giả

The screenshot shows the 'Classification History' section of the Fake News Detector app. It lists three news articles:

- Trump says he thinks Iran may be "tapping us along" on nuclear weapon negotiations** (Category: Politics, Real, 14/04/2025 23:52)
Influential Words (for "Fake"): round, discussions, wrapped, weekend, president, said, proposed
Full Content:

President Donald Trump said he thinks Iran is "tapping us along" in negotiations on nuclear weapons after the first round of discussions wrapped up this weekend in Oman. "We had a meeting with them on Saturday. We have another meeting scheduled next Saturday. I said, 'That's a long time.' You know, that's a long time. So I think they might be tapping us along." Trump said in the Oval Office alongside President Nayib Bukele of El Salvador.

He said Iran "has got rid of the concept of a nuclear weapon. They cannot have a nuclear weapon."

After the meeting, Iranian Foreign Minister Abbas Araghchi told state media that the two sides "got very close" to reaching a framework for negotiations. Trump's has threatened military strikes as a consequence of failure and Tehran has warned any attack on it would drag the US into a broader Middle Eastern conflict.

"I think they're tapping us along because they were so used to dealing with stupid people in this country." Trump said of Iran's proposed timeline for talks.
- Ex-January 6 prosecutors urge attorney disciplinary board to investigate Trump's controversial pick to be DC's top prosecutor** (Category: Politics, Real, 14/04/2025 19:11)
Influential Words (for "Fake"):
confrontation, will, demand, all, expected, always, however
Full Content:

Ex-January 6 prosecutors urge attorney disciplinary board to investigate Trump's controversial pick to be DC's top prosecutor
- Trump's showdown with China deepens, with huge stakes for the economy** (Category: Politics, Fake, 14/04/2025 18:51)
Influential Words (for "Fake"):
damage, us, economy, direct, confrontation, presidents, conflict
Full Content:

Trump's showdown with China deepens, with huge stakes for the economy

At the bottom, there is a footer with links: © 2025 Fake News Detector App. All Rights Reserved., About, Privacy Policy, Terms of Service, News Classifier, History, Panx.

Hình 25. Giao diện Lịch sử phân loại và các thao tác liên quan
Chức năng cho phép xem lại nội dung đầy đủ của một mục
tin tức đã lưu trong lịch sử.(Hình 25.a)

The screenshot shows the 'Classification History' section of the Fake News Detector app. It displays two news articles, with the first one expanded to show its full content:

- Trump says he thinks Iran may be "tapping us along" on nuclear weapon negotiations** (Category: Politics, Real, 14/04/2025 23:52)
Influential Words (for "Fake"): round, discussions, wrapped, weekend, president, said, proposed
Full Content:

President Donald Trump said he thinks Iran is "tapping us along" in negotiations on nuclear weapons after the first round of discussions wrapped up this weekend in Oman. "We had a meeting with them on Saturday. We have another meeting scheduled next Saturday. I said, 'That's a long time.' You know, that's a long time. So I think they might be tapping us along." Trump said in the Oval Office alongside President Nayib Bukele of El Salvador.

He said Iran "has got rid of the concept of a nuclear weapon. They cannot have a nuclear weapon."

After the meeting, Iranian Foreign Minister Abbas Araghchi told state media that the two sides "got very close" to reaching a framework for negotiations. Trump's has threatened military strikes as a consequence of failure and Tehran has warned any attack on it would drag the US into a broader Middle Eastern conflict.

"I think they're tapping us along because they were so used to dealing with stupid people in this country." Trump said of Iran's proposed timeline for talks.
- Ex-January 6 prosecutors urge attorney disciplinary board to investigate Trump's controversial pick to be DC's top prosecutor** (Category: Politics, Real, 14/04/2025 19:11)
Influential Words (for "Fake"):
confrontation, will, demand, all, expected, always, however
Full Content:

Ex-January 6 prosecutors urge attorney disciplinary board to investigate Trump's controversial pick to be DC's top prosecutor

Hình 25.a. Chức năng xem lại nội dung chi tiết của một tin tức đã được phân loại
trong lịch sử

Chức năng cho phép người dùng gửi báo cáo nếu cho rằng
kết quả phân loại không chính xác, kèm theo nguồn kiểm chứng và
bình luận.(Hình 25.b)

Luận văn: Transformer để phân loại tin giả

The screenshot shows the 'Classification History' section of the Fake News Detector. It lists three news entries:

- Trump says he thinks Iran may be "tapping us along"**: Category: Politics. Real (Fake: 7.9% | Real: 92.1%). Influential Words (for "Fake"): round, discussions, wrapped, weekend, president, said. Article: Trump says he thinks Iran may be "tapping us along" on nuclear weapon negotiations.
- Ex-January 6 prosecutors urge attorney disciplinary prosecutor**: Category: Politics. Real (Fake: 0.1% | Real: 99.9%). Influential Words (for "Fake"): confirmation, voter, democrat, ate, pledged, delay, former. Article: Ex-January 6 prosecutors urge attorney disciplinary prosecutor.
- Trump's showdown with China deepens, with huge stakes for the economy**: Category: Politic. Fake (Fake: 99.8% | Real: 0.2%). Influential Words (for "Fake"): confirmation, vote, democrat, ate, pledged, delay, former. Article: Trump's showdown with China deepens, with huge stakes for the economy.

A modal window titled 'Report Classification' is open for the third entry. It contains the following fields:

- Article: Trump says he thinks Iran may be "tapping us along" on nuclear weapon negotiations
- Prediction Label: Real
- What is the correct classification? Real Fake
- Source URL (for verification): <https://edition.cnn.com/politics/live-news/trump-presidency-tz>
- Additional Comments (Optional): Good job

Buttons at the bottom of the modal include 'Cancel' and 'Submit Report'.

Hình 25.b. Chức năng báo cáo sai sót về kết quả phân loại, cho phép người dùng cung cấp nguồn gốc tin tức và bình luận
Chức năng cho phép xóa một mục tin tức khỏi lịch sử cá nhân.(Hình 25.c)

The screenshot shows the 'Classification History' section again. A modal window titled 'Confirm Deletion' is open for the third news entry. It contains the following text:

- Are you sure you want to delete this news entry?
- "Trump's showdown with China deepens, with huge stakes for the economy"
- This action cannot be undone.

Buttons at the bottom of the modal include 'Cancel' and 'Delete'.

Hình 25.c. Chức năng xóa một tin tức khỏi lịch sử phân loại cá nhân
Chức năng dành cho Người quản trị (Admin):

Trang tổng quan (Dashboard): Điểm truy cập trung tâm cho các chức năng quản trị.

Giao diện chính dành riêng cho Admin, cung cấp cái nhìn tổng quan và truy cập các mục quản lý.(Hình 26)

Luận văn: Transformer để phân loại tin giả

The screenshot shows the Admin Dashboard of the Fake News Detector application. At the top, there's a header with the logo 'FAKE NEWS' and the text 'Fake News Detector'. On the right side of the header, there are links for 'Dashboard', 'User Management', 'News Management', a user icon 'B2111894', and 'Admin'. Below the header, the title 'Admin Dashboard' is displayed with a small gear icon. A sub-instruction 'Welcome to the admin area. Choose a section to manage:' is present. Two main buttons are shown: 'User Management' (grey background, icon of three people) and 'News Management' (blue background, icon of a document with a magnifying glass). The 'User Management' button has the sub-instruction 'View, add, edit, and manage user accounts and roles.' and the 'News Management' button has the sub-instruction 'Review reported news articles, view details, and manage content status.'

Hình 26. Giao diện Trang chủ (Dashboard) dành riêng cho Người quản trị
Quản lý Người dùng: Cho phép Admin quản lý toàn bộ tài khoản người dùng trong hệ thống.

Khu vực quản lý người dùng, bao gồm xem danh sách, tìm kiếm, thêm, sửa, xóa người dùng.(Hình 27)

The screenshot shows the 'User Management' page. At the top, there's a search bar with placeholder text 'Search by Username, Email, or Full Name...' and a 'Search' button. To the right of the search bar is a green button labeled 'Add New User' with a person icon. Below the search bar is a table with columns: 'Avatar', 'Username', 'Email', 'Full Name', 'Role', and 'Actions'. The table contains four rows of data:

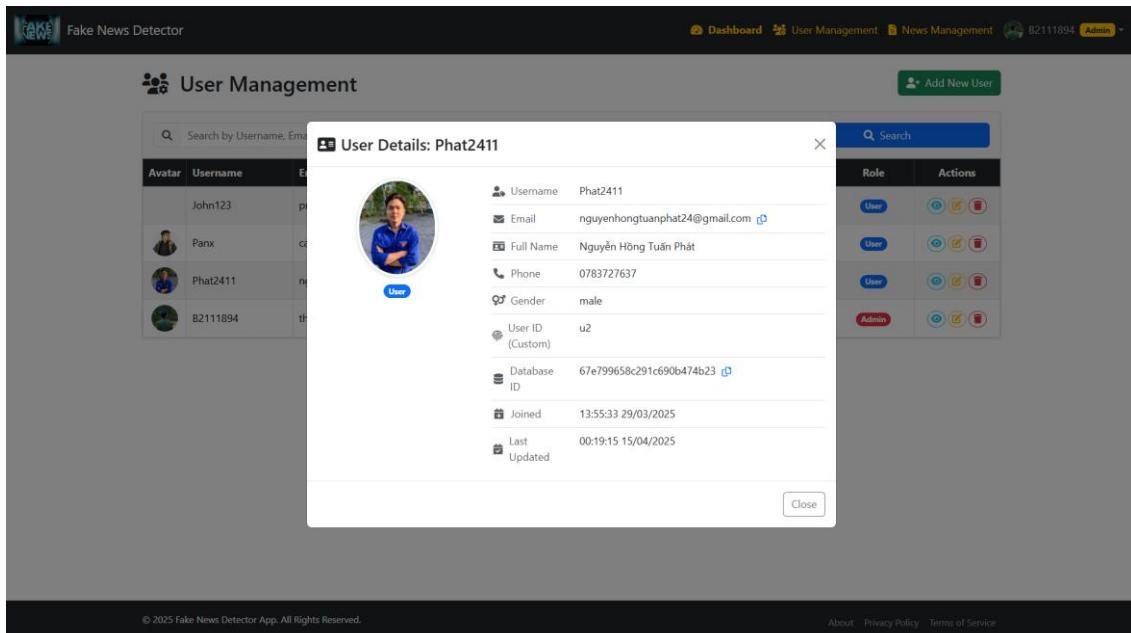
Avatar	Username	Email	Full Name	Role	Actions
[Avatar]	John123	projectai2024@gmail.com	John Cena	User	[Edit] [Delete]
[Avatar]	Panx	caothutocchien2020@gmail.com	Mr Pan	User	[Edit] [Delete]
[Avatar]	Phat2411	nguyenhongtuanphat24@gmail.com	Nguyễn Hồng Tuấn Phát	User	[Edit] [Delete]
[Avatar]	B2111894	thanhnenngunguoij747@gmail.com	Antony	Admin	[Edit] [Delete]

Hình 27. Giao diện Quản lý người dùng của Người quản trị
Chức năng cho phép Admin tạo một tài khoản người dùng mới.(Hình 27.a)

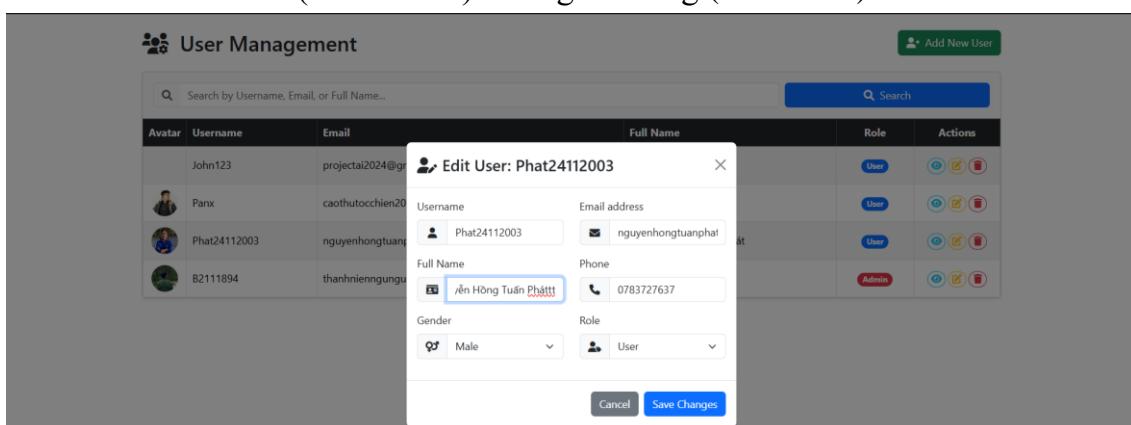
The screenshot shows a 'User Management' page with a modal dialog titled '+ Add New User'. The dialog contains fields for 'Username' (Troy77), 'Email address' (phatB2111894@student.hcmst.edu.vn), 'Password' (hidden), 'Confirm Password' (hidden), 'Full Name' (Troy), 'Phone' (0873287327), 'Gender' (Male), and 'Role' (User). There are 'Cancel' and 'Add User' buttons at the bottom of the dialog. In the background, the main 'User Management' table is visible with one row highlighted for 'Panx'.

Hình 27.a. Chức năng thêm một người dùng mới vào hệ thống
Chức năng cho phép Admin xem thông tin chi tiết của một người dùng cụ thể.(Hình 27.b)

Luận văn: Transformer để phân loại tin giả

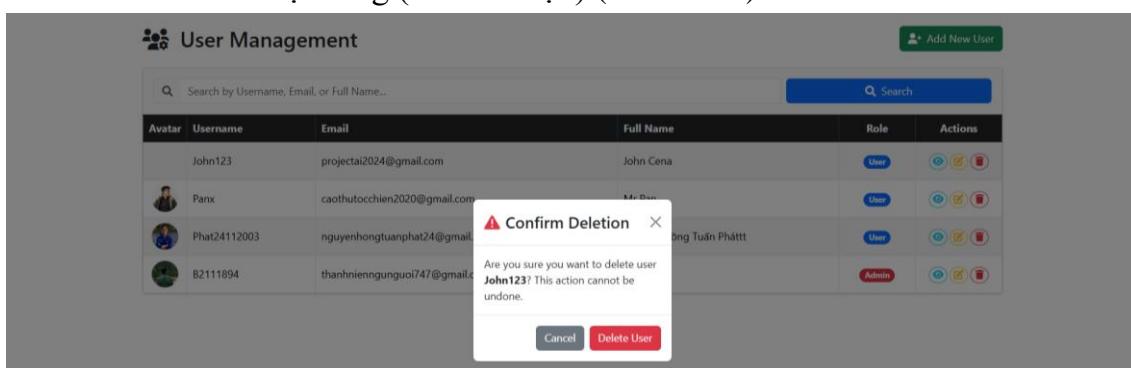


Hình 27.b. Chức năng xem thông tin chi tiết của một người dùng
Chức năng cho phép Admin chỉnh sửa thông tin và thay đổi
vai trò (user/admin) của người dùng.(Hình 27.c)



Hình 27.c. Chức năng chỉnh sửa thông tin và vai trò (user/admin) của người
dùng

Chức năng cho phép Admin xóa một tài khoản người dùng
khỏi hệ thống (có xác nhận).(Hình 27.d)



Luận văn: Transformer để phân loại tin giả

Hình 27.d. Chức năng xóa người dùng khỏi hệ thống

Chức năng tìm kiếm nhanh người dùng trong danh sách dựa trên tên, email hoặc username.(Hình 27.e)

Fake News Detector

User Management

Pan

	Role	Actions
Mr Pan	User	[Edit, View, Delete]
Panx	User	[Edit, View, Delete]
Phat24112003	User	[Edit, View, Delete]
B2111894	Admin	[Edit, View, Delete]

Fake News Detector

User Management

Panx

Avatar	Username	Email	Full Name	Role	Actions
Panx	caothutocchien2020@gmail.com		Mr Pan	User	[Edit, View, Delete]

Hình 27.e. Chức năng tìm kiếm người dùng trong danh sách theo tên, email hoặc username

Quản lý Tin tức: Cho phép Admin xem, xóa các tin tức đã phân loại và duyệt báo cáo sai sót.

Khu vực quản lý các tin tức đã được phân loại trong hệ thống và các báo cáo liên quan.(Hình 28)

Fake News Detector

News Management

Search by News Title...

Title	Category	Prediction	Created At	Report Status	Actions
Trump says he thinks Iran may be "tapping us al...	Politics	Real (7.9% F 92.1% R)	14/04/25 23:52	Processed	[Edit, View, Delete]
Ex-January 6 prosecutors urge attorney disciplin...	Politics	Real (0.1% F 99.9% R)	14/04/25 19:11	Pending	[Edit, View, Delete]

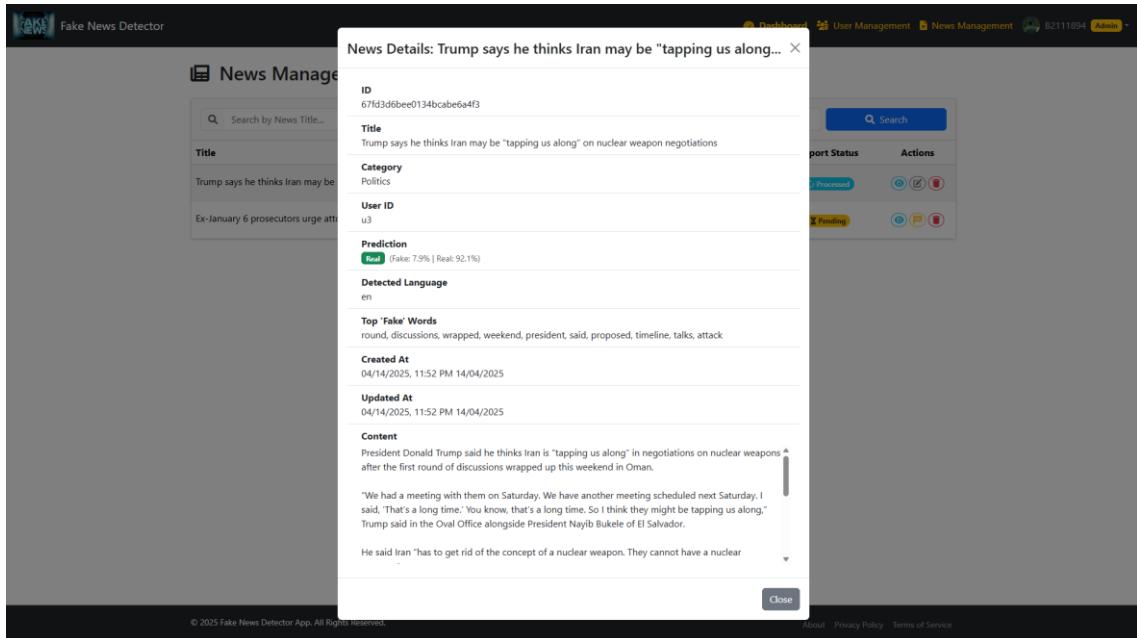
© 2025 Fake News Detector App. All Rights Reserved.

About Privacy Policy Terms of Service

Hình 28. Giao diện Quản lý tin tức của Người quản trị

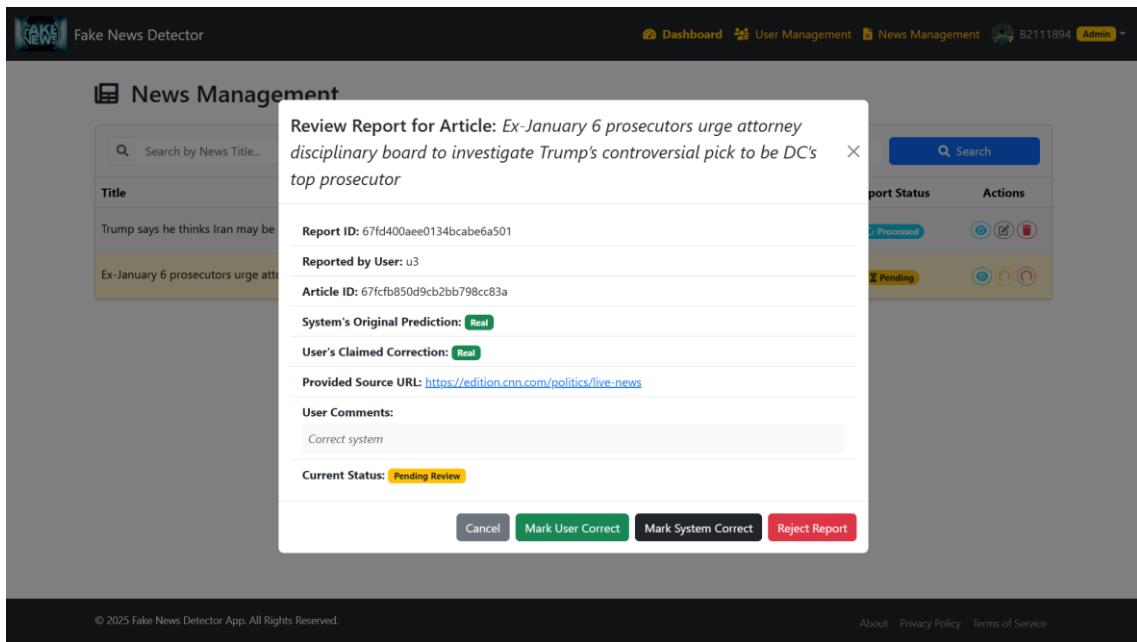
Chức năng cho phép Admin xem nội dung chi tiết của một tin tức đã được phân loại.(Hình 28.a)

Luận văn: Transformer để phân loại tin giả



Hình 28.a. Chức năng xem nội dung chi tiết của một tin tức đã được phân loại trong hệ thống

Chức năng quan trọng cho phép Admin xem xét và duyệt (chấp thuận/tù chối) các báo cáo sai sót về kết quả phân loại do người dùng gửi.(Hình 28.b)



Luận văn: Transformer để phân loại tin giả

© 2025 Fake News Detector App. All Rights Reserved. About Privacy Policy Terms of Service

Hình 28.b. Chức năng xem xét và duyệt (approve/reject) các báo cáo sai sót về tin tức do người dùng gửi lên
Chức năng tìm kiếm tin tức trong danh sách quản lý dựa theo tiêu đề.(Hình 28.c)

© 2025 Fake News Detector App. All Rights Reserved. About Privacy Policy Terms of Service

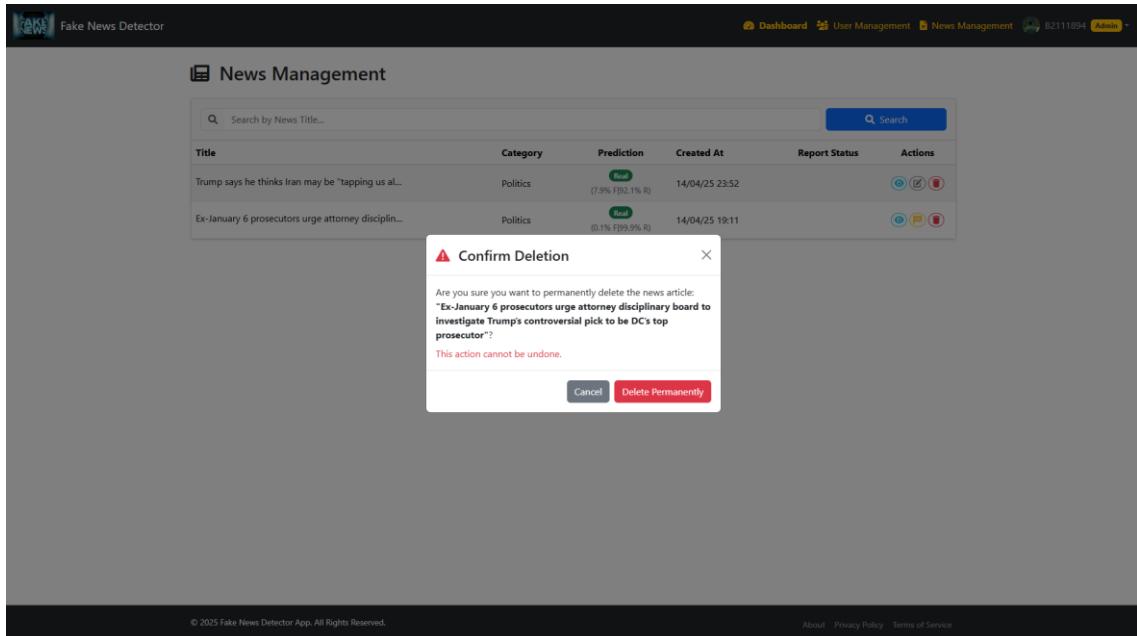
News Management

Title	Category	Prediction	Created At	Report Status	Actions
Trump says he thinks Iran may be "tapping us al...	Politics	Real (7.9% F 92.1% R)	14/04/25 23:52	Processed	

Hình 28.c. Chức năng tìm kiếm tin tức trong danh sách theo tiêu đề

Luận văn: Transformer để phân loại tin giả

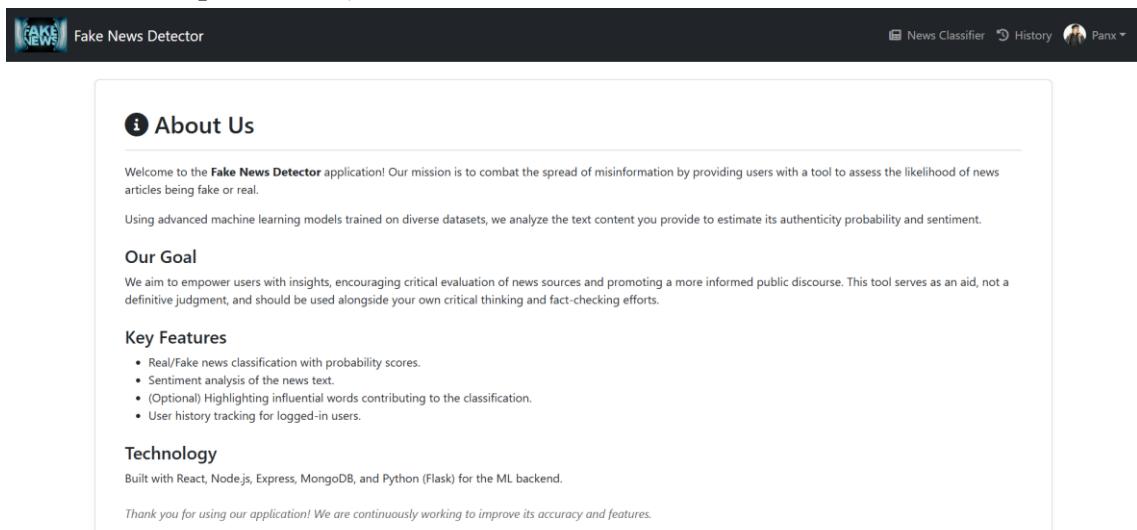
Chức năng cho phép Admin xóa một tin tức khỏi hệ thống.(Hình 28.d)



Hình 28.d. Chức năng xóa một tin tức khỏi hệ thống

Giao diện và Trải nghiệm & Trang thông tin hỗ trợ: Toàn bộ ứng dụng có giao diện được thiết kế nhất quán, hiện đại, sử dụng React Bootstrap, đảm bảo tính đáp ứng trên các kích thước màn hình khác nhau. Các trang thông tin hỗ trợ cần thiết cũng được tích hợp đầy đủ.

Trang cung cấp thông tin giới thiệu về ứng dụng và đội ngũ phát triển.(Hình 29)



Hình 29. Giao diện trang “Thông tin về chúng tôi” (Giới thiệu)

Luận văn: Transformer để phân loại tin giả

Trang trình bày các quy định về việc thu thập, sử dụng và bảo vệ thông tin người dùng.(Hình 30)

The screenshot shows the 'Privacy Policy' section of the Fake News Detector application. At the top, there is a navigation bar with the logo 'FAKE NEWS DETECTOR', 'News Classifier', 'History', and 'Panx'. Below the navigation bar, a button says 'Để thoát khỏi chế độ toàn màn hình, hãy nhấn và giữ [Thoát]'. The main content area has a title 'Privacy Policy' with a lock icon. It was last updated on April 14, 2025. A note states: 'Your privacy is important to us. This Privacy Policy explains how Fake News Detector ("we," "us," or "our") collects, uses, discloses, and safeguards your information when you use our Fake News Detector application (the "Service"). Please read this privacy policy carefully. If you do not agree with the terms of this privacy policy, please do not access the Service.' Below this, a section titled 'Information We Collect' describes what data is collected, including personal data, usage data, and news content data. Another section, 'Use of Your Information', details how information is used for account management, classification service, improving accuracy, and monitoring user behavior.

Hình 30. Giao diện trang “Chính sách bảo mật”

Trang nêu rõ các điều khoản mà người dùng cần đồng ý khi sử dụng dịch vụ.(Hình 31)

The screenshot shows the 'Terms of Service' section of the Fake News Detector application. At the top, there is a navigation bar with the logo 'FAKE NEWS DETECTOR', 'News Classifier', 'History', and 'Panx'. Below the navigation bar, a button says 'Để thoát khỏi chế độ toàn màn hình, hãy nhấn và giữ [Thoát]'. The main content area has a title 'Terms of Service' with a document icon. It was last updated on April 14, 2025. A note states: 'Please read these Terms of Service ("Terms", "Terms of Service") carefully before using the Fake News Detector application (the "Service") operated by Fake News Detector ("us", "we", or "our").' It explains that access to the service is conditioned upon acceptance of these terms. By accessing or using the service, users agree to be bound by these terms. A 'Use License' section grants permission to download materials for personal, non-commercial transitory viewing. A list of allowed actions includes modifying or copying materials, using them for commercial purposes, decompling or reverse engineering software, removing copyright notices, and transferring materials to others. A note states that the license terminates if any restrictions are violated. A 'Disclaimer' section states that materials are provided 'as is' without warranties. A note at the bottom states: 'Further, we do not warrant or make any representations concerning the accuracy, likely results, or reliability of the use of the classification materials on its website or'.

Hình 31. Giao diện trang “Điều khoản dịch vụ”

Nhìn chung, ứng dụng “Fake News Detector” đã được phát triển thành công, cung cấp một công cụ hữu ích cho người dùng trong việc đánh giá sơ bộ tính xác thực của tin tức, góp phần nâng cao nhận thức và khả năng phân biệt thông tin trong môi trường số hiện nay.

PHẦN III: KẾT LUẬN

1. Kết quả đạt được

Luận văn này đã trình bày và đánh giá các phương pháp tiếp cận dựa trên mô hình Transformer để giải quyết bài toán phân loại tin giả, hướng tới mục tiêu cung cấp công cụ hỗ trợ hiệu quả cho các nhà nghiên cứu và chuyên gia truyền thông trong việc xác định độ tin cậy của tin tức. Bằng việc sử dụng các bộ dữ liệu đa dạng về chủ đề và nguồn gốc, kết hợp với kỹ thuật Transfer Learning để tối ưu hóa các mô hình RoBERTa và BERT, nghiên cứu đã đạt được những kết quả nổi bật. Hiệu suất phân loại rất cao đã được ghi nhận, đặc biệt mô hình RoBERTa cho thấy sự vượt trội với F1 – score đạt tới 0.9998 trên ISOT Fake News Dataset và 0.9986 trên Fake News Dataset, khẳng định mạnh mẽ tiềm năng của các mô hình này.

Đóng góp chính của nghiên cứu không chỉ dừng lại ở việc đạt được độ chính xác cao mà còn bao gồm việc xác định các cấu hình siêu tham số tối ưu (optimizer, số epoch) cho từng cặp mô hình – dữ liệu cụ thể và so sánh hiệu quả với các công trình trước đó. Quan trọng hơn, nghiên cứu nhấn mạnh tính khả thi triển khai các mô hình Transformer mạnh mẽ này vào thực tế. Việc xây dựng thành công ứng dụng web “Fake News Detector” là minh chứng cho khả năng ứng dụng này, cung cấp một giao diện thân thiện để người dùng cuối có thể dễ dàng kiểm tra độ tin cậy của tin tức, qua đó góp phần hạn chế sự lan truyền thông tin sai lệch trong cộng đồng.

2. Hướng phát triển

Mặc dù đã đạt được những kết quả tích cực, hệ thống phân loại tin giả này vẫn còn nhiều tiềm năng để mở rộng và cải tiến trong tương lai.

Nâng cao hiệu suất và tính tổng quát: Để cải thiện kết quả trên các tập dữ liệu phức tạp như Fake or Real News và thu hẹp khoảng cách với một số nghiên cứu trên tập Fake News Detection, cần tiếp tục tối ưu mô hình. Có thể thử nghiệm các kỹ thuật fine – tuning nâng cao, huấn luyện trên bộ dữ liệu lớn hơn và cập nhật thường xuyên hơn, hoặc áp dụng few – shot learning để mô hình nhanh chóng thích ứng với các dạng tin giả mới. Việc tích hợp dữ liệu đa ngôn ngữ cũng là một hướng đi quan trọng để tăng phạm vi ứng dụng.

Mở rộng tính năng và ứng dụng: Phát triển API cho phép kiểm tra tin tức theo thời gian thực sẽ tăng cường tính hữu dụng của hệ thống. Xa hơn nữa là tích hợp mô hình vào các nền tảng mạng xã hội (dưới dạng plugin trình duyệt hoặc bot) để cảnh báo người dùng trực tiếp. Việc bổ sung tính năng cho phép người dùng phản hồi, báo cáo tin giả có thể được kết hợp với các kỹ thuật như học chủ động (active learning) để liên tục cải thiện mô hình.

Khám phá các hướng tiếp cận mới: Nghiên cứu có thể mở rộng sang xử lý đa phương thức (multimodal), kết hợp phân tích văn bản với hình ảnh/video đi kèm. Áp dụng các kỹ thuật Explainable AI (XAI) để diễn giải quyết định của mô hình, giúp tăng tính minh bạch và độ tin cậy. Đồng thời, việc khám phá các kiến trúc Transformer mới hơn hoặc các phương pháp lai ghép cũng là hướng đi tiềm năng.

Xây dựng tài nguyên: Việc xây dựng một cơ sở dữ liệu tin tức đã được kiểm chứng, có cấu trúc tốt sẽ là tài nguyên quý giá cho cộng đồng nghiên cứu và phát triển các công cụ kiểm duyệt hiệu quả hơn trong tương lai.

Những hướng phát triển này không chỉ giúp nâng cao chất lượng kỹ thuật của hệ thống mà còn tăng cường khả năng ứng dụng thực tiễn, góp phần tích cực vào cuộc chiến chống lại vấn nạn tin giả.

TÀI LIỆU THAM KHẢO

- [1] “Factor Graph Model Based User Profile Matching Across Social Networks | IEEE Journals & Magazine | IEEE Xplore”. Truy cập: 1 Tháng Chạp 2024. [Online]. Available at: <https://ieeexplore.ieee.org/document/8873650>
- [2] H. Allcott và M. Gentzkow, “Social Media and Fake News in the 2016 Election”, *Journal of Economic Perspectives*, vol 31, số p.h 2, tr 211–236, tháng 5 2017, doi: 10.1257/jep.31.2.211.
- [3] D. M. J. Lazer và c.s., “The science of fake news”, *Science*, vol 359, số p.h 6380, tr 1094–1096, tháng 3 2018, doi: 10.1126/science.aoa2998.
- [4] T. H. Vo, T. L. T. Phan, và K. C. Ninh, “Development of a fake news detection tool for Vietnamese based on deep learning techniques”, *Eastern-European Journal of Enterprise Technologies*, vol 5, số p.h 2(119), Art. số p.h 2(119), tháng 10 2022, doi: 10.15587/1729-4061.2022.265317.
- [5] Hùng V. T., Chi N. K., và Kiệt T. A., “Phát hiện tự động tin giả: Thành tựu và thách thức”, *UD-JST*, tr 71–78, tháng 3 2022.
- [6] D. V. Vo và P. Do, “Detecting Vietnamese fake news”, *CTU J. of Inn. & Sus. Dev.*, vol 15, số p.h Special issue: ISDS, Art. số p.h Special issue: ISDS, tháng 10 2023, doi: 10.22144/ctujoisd.2023.033.
- [7] P. Patwa và c.s., “Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts”, trong *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, và M. S. Akhtar, B.t.v, Cham: Springer International Publishing, 2021, tr 42–53. doi: 10.1007/978-3-030-73696-5_5.
- [8] A. Wani, I. Joshi, S. Khandve, V. Wagh, và R. Joshi, “Evaluating Deep Learning Approaches for Covid19 Fake News Detection”, trong *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, và M. S. Akhtar, B.t.v, Cham: Springer International Publishing, 2021, tr 153–163. doi: 10.1007/978-3-030-73696-5_15.
- [9] H. Ahmed, I. Traore, và S. Saad, “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques”, trong *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, I. Traore, I. Woongang, và A. Awad, B.t.v, Cham: Springer International Publishing, 2017, tr 127–138. doi: 10.1007/978-3-319-69155-8_9.
- [10] F. A. Ozbay và B. Alatas, “Fake news detection within online social media using supervised artificial intelligence algorithms”, *Physica A: Statistical Mechanics and its Applications*, vol 540, tr 123174, tháng 2 2020, doi: 10.1016/j.physa.2019.123174.
- [11] I. Ahmad, M. Yousaf, S. Yousaf, và M. O. Ahmad, “Fake News Detection Using Machine Learning Ensemble Methods”, *Complexity*, vol 2020, số p.h 1, tr 8885861, 2020, doi: 10.1155/2020/8885861.
- [12] “Fake News”. Truy cập: 6 Tháng Ba 2025. [Online]. Available at: <https://kaggle.com/fake-news>

- [13] “Fake News detection”. Truy cập: 6 Tháng Ba 2025. [Online]. Available at: <https://www.kaggle.com/datasets/jruvika/fake-news-detection>
- [14] M. Mimura và T. Ishimaru, “Analyzing common lexical features of fake news using multi-head attention weights”, *Internet of Things*, vol 28, tr 101409, tháng 12 2024, doi: 10.1016/j.iot.2024.101409.
- [15] R. K. Kaliyar, A. Goswami, P. Narang, và S. Sinha, “FNDNet – A deep convolutional neural network for fake news detection”, *Cognitive Systems Research*, vol 61, tr 32–44, tháng 6 2020, doi: 10.1016/j.cogsys.2019.12.005.
- [16] P. Bahad, P. Saxena, và R. Kamal, “Fake News Detection using Bi-directional LSTM-Recurrent Neural Network”, *Procedia Computer Science*, vol 165, tr 74–82, tháng 1 2019, doi: 10.1016/j.procs.2020.01.072.
- [17] J. Salminen, *joolsa/fake_real_news_dataset*. (17 Tháng Giêng 2024). Truy cập: 6 Tháng Ba 2025. [Online]. Available at: https://github.com/joolsa/fake_real_news_dataset
- [18] D. S và B. Chitturi, “Deep neural approach to Fake-News identification”, *Procedia Computer Science*, vol 167, tr 2236–2243, tháng 1 2020, doi: 10.1016/j.procs.2020.03.276.
- [19] I. K. Sastrawan, I. P. A. Bayupati, và D. M. S. Arsa, “Fake News Dataset”, vol 1, tháng 9 2021, doi: 10.17632/945z9xkc8d.1.
- [20] “Information disorder: Toward an interdisciplinary framework for research and policy making”, Council of Europe Publishing. Truy cập: 16 Tháng Tư 2025. [Online]. Available at: <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html>
- [21] I. K. Sastrawan, I. P. A. Bayupati, và D. M. S. Arsa, “Detection of fake news using deep learning CNN–RNN based methods”, *ICT Express*, vol 8, số p.h 3, tr 396–408, tháng 9 2022, doi: 10.1016/j.icte.2021.10.003.
- [22] A. Vaswani và c.s., “Attention Is All You Need”, 12 Tháng Sáu 2017, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.
- [23] J. Devlin, M.-W. Chang, K. Lee, và K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 24 Tháng Năm 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.
- [24] Y. Liu và c.s., “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, 26 Tháng Bảy 2019, *arXiv*: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.
- [25] S. Lundberg và S.-I. Lee, “A Unified Approach to Interpreting Model Predictions”, 25 Tháng Mười-Một 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.