TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

HUỲNH ANH TUẤN - 52000291

SỬ DỤNG CÁC MÔ HÌNH TRANSFORMER ĐỂ PHÁT HIỆN CÁC TIN GIẢ TIẾNG VIỆT TRÊN MẠNG XÃ HỘI

CHUYÊN ĐỀ NGHIÊN CỨU 1
NGÀNH KHOA HỌC MÁY TÍNH

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

HUỲNH ANH TUẤN - 52000291

SỬ DỤNG CÁC MÔ HÌNH TRANSFORMER ĐỂ PHÁT HIỆN CÁC TIN GIẢ TIẾNG VIỆT TRÊN MẠNG XÃ HỘI

CHUYÊN ĐỀ NGHIÊN CỨU 1
NGÀNH KHOA HỌC MÁY TÍNH

Người hướng dẫn

TS. Trần Thanh Phước

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn sâu sắc đến TS. Trần Thanh Phước – Giảng viên Khoa Công nghệ thông tin, Trường Đại học Tôn Đức Thắng. Thầy đã tận tình hướng dẫn, đưa ra những góp ý và hỗ trợ em trong suốt quá trình thực hiện nghiên cứu này.

Bên cạnh đó, em cũng xin gửi lời cảm ơn đến toàn thể quý Thầy, Cô tại Trường Đại học Tôn Đức Thắng, đặc biệt là các Thầy, Cô thuộc Khoa Công nghệ thông tin đã giảng dạy và truyền đạt kiến thức trong suốt quá trình học tập để em có nền tảng vững chắc để thực hiện nghiên cứu.

Mặc dù đã dành nhiều thời gian thực hiện nhưng sẽ không tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý quý báu từ Thầy, Cô để có thể hoàn thiện hơn nữa trong tương lai.

Một lần nữa, em xin chân thành cảm ơn quý Thầy, Cô đã luôn đồng hành và hỗ trợ em trên con đường học tập tại trường.

TP. Hồ Chí Minh, ngày tháng năm 2024

Tác giả

(Ký tên và ghi rõ họ tên)

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Trần Thanh Phước. Các nội dung nghiên cứu, kết quả trong nghiên cứu này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong luận văn còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung nghiên cứu của mình. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm 2024

Tác giả

(Ký tên và ghi rõ họ tên)

SỬ DỤNG CÁC MÔ HÌNH TRANSFORMER ĐỂ PHÁT HIỆN CÁC TIN GIẢ TIẾNG VIỆT TRÊN MẠNG XÃ HỘI

TÓM TẮT

Sự lan truyền tin tức giả trên mạng xã hội đã trở thành một vấn đề nghiêm trọng, dẫn đến thông tin sai lệch và gây hại cho xã hội. Dự án này nhằm phát triển một hệ thống để phân tích và phân loại tin tức giả bằng tiếng Việt sử dụng các mô hình Transformer, đặc biệt là các mô hình được huấn luyện và tối ưu hóa cho tiếng Việt như PhoBERT, ViBERT và ViSoBERT.

Để thực hiện nghiên cứu này, chúng tôi đã thu thập một tập dữ liệu bao gồm các bài viết bằng tiếng Việt trên nền tảng mạng xã hội Facebook và một số bài viết từ các nguồn tin tức Việt Nam, bao gồm các chủ đề như tin tức đời sống, tin tức, thể thao, thời tiết và chính trị. Tuy nhiên, vẫn còn những thách thức do sự mất cân bằng dữ liệu giữa số lượng tin thật và tin giả. Các bài viết đã được gán nhãn thủ công là thật hoặc giả, sau đó trải qua quá trình tiền xử lý dữ liệu và được huấn luyện bằng các mô hình Transformer. Chúng tôi cũng đã áp dụng kỹ thuật tiền xử lý dữ liệu TF-IDF để tối ưu hóa hiệu suất của mô hình.

Để đánh giá hiệu suất của các mô hình, chúng tôi đã sử dụng nhiều chỉ số đánh giá như Accuracy, Precision, Recall, F1 Score và AUC. Kết quả của chúng tôi cho thấy PhoBERT vượt trội hơn các mô hình Transformer khác trong việc phát hiện tin tức giả tiếng Việt, đạt được độ chính xác và độ tin cậy cao lên đến trên 94%

Báo cáo này cung cấp cái nhìn tổng quan về bối cảnh, mục tiêu, phương pháp và kết quả của dự án. Qua đó nêu rõ những đóng góp của dự án đối với lĩnh vực phát hiện tin tức giả, đồng thời đề xuất hướng phát triển nghiên cứu trong tương lai.

SỬ DỤNG CÁC MÔ HÌNH TRANSFORMER ĐỂ PHÁT HIỆN CÁC TIN GIẢ TIẾNG VIỆT TRÊN MẠNG XÃ HỘI

ABSTRACT

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC CHỮ VIẾT TẮT

AUC	Area Under the Curve
BiLSTM	Bidirectional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
CBOW	Continuous Bag of Words
CNN	Convolutional Neural Network
DistilBERT	Distilled BERT
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
MLM	Masked Language Model
NLP	Natural Language Processing
NSP	Next Sentence Prediction
RNN	Recurrent Neural Network
RoBERTa	A Robustly Optimized BERT Pretraining Approach
TF-IDF	Term Frequency - Inverse Document Frequency
vELECTRA	Vietnamese ELECTRA
ViBERT	Vietnamese BERT
ViSoBERT	Vietnamese Social BERT
Word2Vec	Word to Vector

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1 Bối cảnh và lý do chọn đề tài

Trong thời đại công nghệ thông tin đang phát triển, mạng xã hội đã trở thành công cụ đăng tải và tiếp nhận thông tin trong cuộc sống hàng ngày của hàng triệu người trên toàn thế giới. Các nền tảng như Facebook, Twitter,… mang lại nhiều giá trị tiện ích như là nơi cập nhật các thông tin hằng ngày từ các trang báo mạng; đồng thời là phương tiện để kết nối bạn bè, gia đình, đồng nghiệp;…Tuy nhiên, cùng với đó là rủi ro về việc lan truyền thông tin thiếu kiểm chứng, tin giả và tin đồn thất thiệt cũng tăng nhanh chóng, mang đến những tác động tiêu cực đến xã hội.

Trên thế giới, những tin tức giả chưa được kiểm soát không chỉ truyền tải sai sự thật, gây hiểu lầm, hoang mang trong cộng đồng mà còn có khả năng gây ra những hậu quả nghiêm trọng như mâu thuẫn và xung đột. Các nghiên cứu cho thấy rằng tin giả thường lan truyền với tốc độ nhanh hơn rất nhiều so với tin tức chính xác, bởi nó được viết theo cấu trúc thu hút sự chú ý của người xem. Những thông tin sai lệch này không chỉ ảnh hưởng đến nhận thức của người đọc mà còn gây khó khăn cho họ khi phải đưa ra quyết định trước các sự kiện xã hội quan trọng.

Tại Việt Nam, tin giả cũng đã ngày càng trở nên phổ biến, từ những thông tin sai lệch về dịch bệnh như COVID-19, cho đến các tin đồn về những vụ tai nạn giao thông nghiêm trọng hay những vụ án giật gân, lôi kéo sự chú ý của công chúng. Ngoài ra, tin giả còn xuất hiện trong các lĩnh vực nhạy cảm hơn như chính trị, với những nội dung có tính chất gây chia rẽ, làm xáo trộn lòng tin của người dân vào chính phủ và các cơ quan chức năng. Những loại tin tức này có thể tác động trực tiếp đến tâm lý người dân, gây ra sự bất ổn trong xã hội Việt Nam.

Nhận thấy vấn đề phân tích và phát hiện tin giả là một nhiệm vụ cấp bách và cần thiết để hỗ trợ duy trì sự ổn định của xã hội. Do đó, chúng tôi mong muốn kết hợp những tiến bộ trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và khai thác dữ liệu để tạo ra các giải pháp hiệu quả trong nhiệm vụ phát hiện tin giả và đóng góp cho cộng đồng.

Trong những năm gần đây, học sâu đã được công nhận là một công cụ mạnh mẽ trong lĩnh vực trí tuệ nhân tạo, đặc biệt là trong xử lý ngôn ngữ tự nhiên (NLP). Tuy nhiên, các mô hình học sâu truyền thống thường dựa vào xử lý dữ liệu tuần tự [1], điều này có thể gây hạn chế khi đối mặt với các nhiệm vụ ngôn ngữ phức tạp. Sau đó, sự ra đời của một kiến trúc mới là Transformers đã cách mạng hóa NLP bằng cách sử dụng các cơ chế chú ý [2], cho phép xử lý ngữ cảnh và các mối quan hệ trong văn bản một cách hiệu quả hơn.

Transformer là một trong những kiến trúc tiên tiến trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), các mô hình điển hình như BERT và RoBERTa đã chứng minh hiệu suất vượt trội trong việc xử lý các nhiệm vụ ngôn ngữ tự nhiên, nhờ vào khả năng phân tích và hiểu ngữ nghĩa của văn bản một cách toàn diện và hiệu quả. Khả năng này đặc biệt quan trọng trong việc phát hiện tin giả, khi mà các mô hình truyền thống thường gặp khó khăn trong việc xử lý những ngữ cảnh phức tạp và đa dạng của ngôn ngữ.

Tại Việt Nam, nhiều mô hình nghiên cứu đã được thực hiện và phát triển dành riêng cho đặc thù ngôn ngữ Tiếng Việt như ViBERT, ViSoBERT, và PhoBERT. ViBERT tinh chỉnh BERT để phù hợp với cấu trúc ngữ nghĩa phức tạp của Tiếng Việt [3]. ViSoBERT nghiên cứu các đặc điểm của Tiếng Việt trên mạng xã hội [4], trong khi PhoBERT được huấn luyện trên lượng dữ liệu lớn Tiếng Việt [5], mang lại hiệu quả cao trong các bài toán phân loại. Những mô hình này không chỉ tối ưu hóa khả năng xử lý ngữ nghĩa mà còn nắm bắt được những sắc thái và ý nghĩa ẩn sâu trong văn bản Tiếng Việt.

Chính vì vậy, chúng tôi quyết định chọn đề tài "Sử dụng các mô hình Transformer để phân tích và phát hiện tin giả bằng tiếng Việt". Thông qua nghiên cứu này, chúng tôi kỳ vọng sẽ phát triển một giải pháp hiệu quả, góp phần vào việc giảm thiểu tác động của tin giả trong xã hội. Chúng tôi hy vọng rằng nghiên cứu này sẽ có những đóng góp hữu ích trong việc phát hiện và ngăn chặn tin giả tại Việt Nam.

1.2 Mục tiêu và phạm vi nghiên cứu

Với bối cảnh tin giả đang lan rộng và ngày càng phức tạp như hiện nay, việc nghiên cứu các giải pháp công nghệ để phát hiện và ngăn chặn thông tin sai lệch là vô cùng cần thiết. Chính vì vậy, trong nghiên cứu này chúng tôi tập trung vào việc tận dụng các mô hình Transformer là BERT và các biến thể để phát hiện tin tức giả, đặc biệt các biến thể được thiết kế dành riêng cho ngôn ngữ tiếng Việt như là ViBERT, ViSoBERT và PhoBERT.

Tuy nhiên, chúng tôi đang đối mặt với những thách thức lớn do thiếu hụt các bộ dữ liệu quy mô lớn chứa cả tin thật và tin giả tiếng Việt. Để khắc phục điều này, chúng tôi quyết định thu thập một bộ dữ liệu mới về tin giả bằng tiếng Việt trong năm 2024. Bộ dữ liệu gồm các thông tin về nhiều lĩnh vực được thu thập từ nhiều nguồn khác nhau, bao gồm các nền tảng mạng xã hội, các trang tin tức và các nguồn thông tin trực tuyến đáng tin cậy. Điều này giúp đảm bảo rằng dữ liệu dùng để huấn luyện mô hình có độ phong phú và chính xác cao, hỗ trợ quá trình phân tích và phát hiện tin giả một cách hiệu quả nhất.

Cuối cùng thông qua bộ dữ liệu đã thu được, chúng tôi áp dụng các kỹ thuật NLP và huấn luyện các mô hình Transformer để thử nghiệm, sau đó đánh giá và so sánh mức độ hiểu quả của các mô hình với nhau nhằm xác định mô hình nào tối ưu nhất cho nhiệm vụ phát hiện tin giả trong bối cảnh ngôn ngữ tiếng Việt.

Như vậy mục tiêu toàn diện của của chúng tôi là khai thác sức mạnh của các mô hình Transformer để phát triển một mô hình có độ chính xác trên 90% trong việc phát hiện tin tức giả trên các nền tảng truyền thông xã hội, đặc biệt là Facebook - nền tảng mạng xã hội phổ biến nhất tại Việt Nam, đồng thời cung cấp một bộ dữ liệu mới về tin giả bằng tiếng Việt phục vụ cho các nghiên cứu và ứng dụng sau này.

Cuối cùng, nghiên cứu của chúng tôi giới hạn trong phạm vi phát hiện tin tức giả bằng ngôn ngữ tiếng Việt và không bao gồm các ngôn ngữ khác. Bên cạnh đó, chúng tôi cũng chưa áp dụng phân tích các thông tin giả qua hình ảnh hoặc video,

mà chỉ tập trung vào phân tích văn bản. Hơn nữa, nghiên cứu này không xem xét các khía cạnh như động cơ hoặc nguồn gốc của tin giả, mà chỉ tập trung vào việc phát hiện nội dung sai lệch dựa trên nội dung tin tức hiện có. Chúng tôi cũng không phân tích các thông tin về luật pháp, hiến pháp và những thông tin vĩ mô mà chỉ tập trung vào các tin tức về đời sống, kinh tế, thời tiết, thể thao, và các thông tin chính trị đơn giản nhằm đảm bảo không đưa ra các dự đoán sai lệch nghiêm trọng.

1.3 Cấu trúc báo cáo

Báo cáo được chia thành 5 chương, mỗi chương tập trung vào các khía cạnh khác nhau của quá trình nghiên cứu như sau:

- ***Chương 1: Giới thiệu***

Chương này sẽ mở đầu bằng việc trình bày tổng quan về đề tài nghiên cứu, lý do chọn đề tài và tầm quan trọng của việc phát hiện tin tức giả trong bối cảnh các nền tảng mạng xã hội ngày càng phát triển. Đặc biệt, chương này sẽ nhấn mạnh đến tác động tiêu cực của tin tức giả đến xã hội, và vai trò của các mô hình Transformer trong việc giúp phát hiện thông tin sai lệch. Từ đó, mục tiêu và phạm vi nghiên cứu sẽ được làm rõ, tạo nền tảng cho các chương tiếp theo.

- ***Chương 2: Cơ sở lý thuyết***

Chương này sẽ tập trung vào việc xem xét các công trình nghiên cứu trước đó liên quan đến các mô hình Transformer và phát hiện tin tức giả. Nội dung sẽ bao gồm tổng hợp và đánh giá các mô hình đã được áp dụng trên thế giới và trong nghiên cứu tại Việt Nam. Bằng cách so sánh các phương pháp và mô hình, chúng tôi sẽ làm rõ những điểm mạnh, điểm yếu và những đóng góp của các công trình trước đây, từ đó làm cơ sở cho phương pháp nghiên cứu trong đề tài này.

Đồng thời chương 2 cũng cung cấp các kiến thức nền tảng thiết yếu cho đề tài. Cụ thể, phần này sẽ đi sâu vào các khái niệm cơ bản về xử lý ngôn ngữ tự nhiên và các mô hình Transformer, như BERT, RoBERTa, ViBERT, ViSoBERT và PhoBERT. Việc giới thiệu các kiến thức nền tảng này nhằm đảm bảo rằng người

đọc có một nền tảng vững chắc để hiểu rõ các phương pháp và kỹ thuật được áp dụng trong nghiên cứu. Điều này không chỉ giúp làm rõ các khái niệm cốt lõi mà còn tạo điều kiện cho việc theo dõi và đánh giá các phần nghiên cứu chi tiết trong các chương tiếp theo của đồ án.

- ***Chương 3: Phương pháp phát hiện tin giả***

Chương 3 sẽ tập trung vào phân tích chi tiết về phương pháp nghiên cứu mà đề tài áp dụng, từ việc xây dựng mô hình tổng thể cho đến các bước triển khai cụ thể nhằm phát triển hệ thống hoàn chỉnh. Chúng tôi sẽ giới thiệu cách tiếp cận toàn diện, bao gồm quá trình thu thập và tiền xử lý dữ liệu, tiếp đó là các bước quan trọng như huấn luyện và đánh giá các mô hình.

- ***Chương 4: Thực nghiệm***

Chương này sẽ trình bày quá trình thực nghiệm và đánh giá kết quả. Cụ thể, chương sẽ mô tả tập dữ liệu được sử dụng, các tham số thử nghiệm, và phương pháp đánh giá mô hình. Đồng thời so sánh kết quả đánh giá độ chính xác của các mô hình được với nhau, và phân tích, thảo luận về kết quả một cách chi tiết để tìm ra những ưu điểm và nhược điểm.

- ***Chương 5: Kết luận và hướng phát triển***

Chương cuối cùng sẽ tóm tắt các kết quả chính đạt được từ nghiên cứu, đánh giá khả năng áp dụng của hệ thống phát hiện tin giả trong thực tế. Ngoài ra, chúng tôi cũng sẽ đề xuất các hướng nghiên cứu tiếp theo nhằm cải thiện mô hình và mở rộng phạm vi ứng dụng của hệ thống trong các lĩnh vực khác.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Các nghiên cứu liên quan

Việc phát hiện tin giả đang là một chủ đề được nghiên cứu rất nhiều do sự gia tăng của thông tin sai lệch trên toàn thế giới. Nhiều nghiên cứu đã thử nhiều cách khác nhau để giải quyết vấn đề này.

Trong những ngiên cứu phân tích tin giả, Agarwal và cộng sự (2021) đã sử dụng một lớp Bi-LSTM với chức năng chú ý để phân loại tin tức tiếng Anh dựa trên ngữ cảnh [6]. Monti và cộng sự (2019) đã nghiên cứu mạng nơ-ron đồ thị, sử dụng Graph CNN bốn lớp để dự đoán tin tức bằng cách kết hợp thông tin về hoạt động của người dùng và bài viết [7]. Trong khi đó, Qi và cộng sự (2019) nhấn mạnh tầm quan trọng của nội dung hình ảnh, và phát triển một mô hình đa miền sử dụng CNN và RNN để phân tích đặc điểm hình ảnh, giúp phân biệt giữa tin giả và tin thật [8].

Mặc dù các mô hình như Bi-LSTM, Graph CNN và CNN/RNN đã chứng minh hiệu quả trong việc phát hiện tin giả, nhưng chúng vẫn gặp một số hạn chế như thời gian tính toán lớn, phụ thuộc vào cấu trúc mạng người dùng, và gặp khó khăn đối với mối quan hệ phức tạp giữa từ ngữ trong văn bản. Trong khi đó, vào năm 2017, Vaswani và cộng sự (2017) đã giới thiệu một kiến trúc mới là Transformer, sử dụng cơ chế tự chú ý để xử lý hiệu quả dữ liệu tuần tự [9]. Kể từ đó, các mô hình Transformer đã nhận được sự quan tâm và nghiên cứu ngày càng nhiều , đặt nền tảng cho các ứng dụng xử lý ngôn ngữ tự nhiên hiện đại.

Kể từ đó, nhiều mô hình Transformer đã được phát triển ra để thực hiện các nhiệm vụ xử lý ngôn ngữ tự nhiên, điển hình là BERT (Bidirectional Encoder Representations from Transformers) được Devlin và cộng sự giới thiệu lần đầu vào năm 2018 [10]. BERT là mô hình có khả năng chú ý hai chiều, giúp mô hình hiểu rõ hơn ngữ cảnh của các từ trong câu. BERT sử dụng hai phương pháp là: Masked Language Model để dự đoán các token bị che khuất [MASK], nhằm huấn luyện mô hình hiểu ngữ nghĩa của từ; và Next Sentence Prediction để dự đoán hai câu ngẫu nhiên có phải là hai câu liên tiếp trong văn bản không.

Dựa trên nền tảng của BERT, Liu và cộng sự đã phát triển RoBERTa vào năm 2019, cải thiện hiệu quả huấn luyện và hiệu suất trên các bài kiểm tra NLP, RoBERTa có một số thay đổi như bỏ qua phương pháp NSP của BERT để đơn giản hóa quy trình huấn luyện [11]. Cũng trong năm 2019, Sanh và cộng sự đã giới thiệu DistilBERT, một phiên bản nhỏ gọn và nhanh hơn của BERT [12], phù hợp cho các ứng dụng yêu cầu phản hồi nhanh.

Năm 2020, Clark và cộng sự đã nghiên cứu và phát triển một mô hình có kỹ thuật huấn luyện khác với BERT, nhưng vẫn dựa trên kiến trúc Transformer, đó là ELECTRA [13]. Mô hình này sủ dụng phương pháp Replaced Token Detection thay vì Masked Language Modeling. Phương pháp này không thay thế token bằng [MASK] mà thay thế một số token bằng các từ hoặc token hợp lý khác do một hàm tạo mẫu thực hiện, sau đó sẽ phân loại chúng. Qua thí nghiệm, mô hình này cho hiệu quả ngay cả khi sử dụng một lượng tài nguyên tính toán nhỏ hơn so với BERT.

Tại Việt Nam, Nguyễn Quốc Đạt và Nguyễn Anh Tuấn đã phát triển PhoBERT vào năm 2020, một mô hình Transformer được huấn luyện trên một tập văn bản lớn bằng tiếng Việt [5]. Điều này đã tạo ra một bước tiến lớn cho các nhiệm vụ xử lý ngôn ngữ tự nhiên bằng tiếng Việt. Kết quả của nghiên cứu thấy PhoBERT thường xuyên cho kết quả tốt hơn so với các mô hình đa ngôn ngữ khác khi áp dụng cho Tiếng Việt. Cũng trong năm 2020, Bùi Thế Việt và cộng sự đã phát triển hai mô hình là ViBERT và vELECTRA dựa trên kiến trúc BERT và ELECTRA, đây cũng là các mô hình Transformer được huấn luyện riêng cho Tiếng Việt với các tập dữ liệu huấn luyện có kích thước rất lớn [3]. Hai mô hình này đã đạt được độ chính xác trên 95% khi thực hiện gán nhãn phân loại trên hai tập dữ liệu VLSP 2010 và VLSP 2013.

Các mô hình PhoBERT, ViBERT và vELECTRA đã nâng cao hiệu suất nhiều nhiệm vụ NLP cụ thể cho tiếng Việt như phân loại từ, phân tích phụ thuộc, nhận diện thực thể có tên, và suy luận ngữ nghĩa. Gần đây, nhiều nghiên cứu đã tập trung vào việc sử dụng các mô hình này và các kỹ thuật học sâu khác để phát hiện

tin giả bằng tiếng Việt. Một trong những nghiên cứu nổi bật là của Nguyễn Cao Minh Hiếu và cộng sự đã đề xuất trong cuộc thi ReINTEL 2020. Họ đã phát triển một mô hình kết hợp PhoBERT với các chỉ số thời gian và tương tác cộng đồng như số lượt chia sẻ, lượt thích và bình luận. Mô hình StackNet của họ đạt được điểm AUC là 0.9521, đứng đầu bảng xếp hạng của ReINTEL [14].

Năm 2021, Phạm Ngọc Đông và cộng sự đã đề xuất một phương pháp kết hợp PhoBERT với TF-IDF để tạo ra word embedding và sử dụng CNN để trích xuất đặc trưng [15]. Phương pháp này đạt được điểm AUC là 0.9538. Tuy nhiên, sự phụ thuộc vào tập dữ liệu ReINTEL có thể hạn chế sự đa dạng của kết quả. Trong năm 2022 tiếp theo, Nguyễn Thị Cẩm Vân và cộng sự đã giới thiệu v3MFND, một mô hình phát hiện tin giả đa miền đa phương tiện sâu, tích hợp văn bản, hình ảnh và video để cải thiện độ chính xác [16], nhưng sự phức tạp của mô hình có thể ảnh hưởng đến khả năng áp dụng thời gian thực.

Năm 2022, Võ Trung Hùng và cộng sự đã áp dụng các mô hình CNN và RNN để phân loại tin tức thành bốn nhóm khác nhau, đạt được tỷ lệ chính xác 85% [17]. Dù vậy, kích thước nhỏ của tập dữ liệu của họ có thể làm giảm tính tổng quát của kết quả. Trong khi đó, Khoa Đăng Phạm và cộng sự đã phát triển một mô hình mới dựa trên vELECTRA, họ sử dụng các đặc trưng tiền chế và đạt được điểm AUC là 0.9575 trên tập dữ liệu ReINTEL [18]. Tuy nhiên, sự phụ thuộc vào những đặc trưng này có thể gây khó khăn trong việc thích ứng với các tình huống khác.

Năm 2023, Nguyễn Quốc Nam và cộng sự đã phát triển một mô hình mới cho văn bản trên mạng xã hội Tiếng Việt là ViSoBERT, mô hình được huấn luyện trên tập dữ liệu gồm các văn bản trên các mạng xã hội lớn như Facebook, Tiktok và Youtube [4]. Kết quả nghiên cứu của họ chỉ ra rằng ViSoBERT vượt trội hơn các mô hình ngôn ngữ tiền huấn luyện trước đó cho Tiếng Việt như viBERT, vELECTRA, PhoBERT khi thử nghiệm trên nhiều nhiệm vụ liên quan đến mạng xã hội bao gồm nhận diện cảm xúc, phát hiện phát ngôn thù địch… khi đánh giá trên tập dữ liệu của họ.

Các nghiên cứu này cho thấy các mô hình Transformer, đặc biệt là PhoBERT, rất hiệu quả trong việc phát hiện tin giả bằng tiếng Việt. Chúng cũng nhấn mạnh rằng việc kết hợp dữ liệu văn bản với hình ảnh, video có thể cải thiện hiệu suất. Tuy nhiên, vẫn còn những thách thức lớn như kích thước tập dữ liệu, nguồn thu dữ liệu, sự đa dạng và độ phức tạp tính toán mà các nghiên cứu trong tương lai cần giải quyết.

2.2 Kiến thức nền tảng

Để triển khai hiệu quả nghiên cứu phát hiện tin giả tiếng Việt trên các nền tảng mạng xã hội sử dụng các mô hình Transformer, đặc biệt là PhoBERT và các biến thể BERT khác, việc có hiểu biết vững chắc về các kiến thức nền tảng sau đây là rất quan trọng:

2.2.1 Xử lý Ngôn ngữ Tự nhiên (NLP)

Xử lý ngôn ngữ tự nhiên (NLP) là một nhánh của học máy cho phép máy tính hiểu, phân tích và tương tác với ngôn ngữ của con người [1]. NLP đóng vai trò quan trọng trong việc giúp máy tính xử lý và phân tích văn bản bằng các kỹ thuật học máy và học sâu.

Những nhiệm vụ chính trong NLP bao gồm:

- **Phân tích cú pháp:** giúp xác định ý nghĩa của một từ, cụm từ hoặc câu bằng cách phân tích cú pháp của các từ và áp dụng các quy tắc ngữ pháp được lập trình sẵn**.** Phân tích cú pháp được thực hiện theo hai hình thức:

- Phân tích phụ thuộc tập trung phân tích ngữ pháp của câu bằng cách xác định chủ ngữ, vị ngữ, tân ngữ,… và xem xét cách chúng liên hệ với nhau để tạo nên ý nghĩa tổng thể.
- Phân tích thành phần là một cấu trúc dạng cây biểu diễn cấu trúc phân cấp của câu. Tức là biểu diễn mối quan hệ giữa các thành phần trong câu như cụm danh từ, cụm động từ,…

- **Phân tích ngữ nghĩa:** giúp xác định ý nghĩa, cách sử dụng của các từ và cụm từ dựa vào ngữ cảnh cụ thể của câu văn; từ đó hiểu sâu hơn về ngữ nghĩa cũng như nội dung mà văn bản truyền tải.
- **Nhận dạng thực thể có tên (NER):** nhằm xác định và phân loại các thực thể cụ thể trong một đoạn văn như tên riêng của người, địa điểm, thời gian, tiền tệ,…
- **Phân loại văn bản:** gán nhãn cho văn bản theo các danh mục như tích cực/tiêu cực, thư rác/không phải thư rác, hoặc tin thật/tin giả. Trong nhiệm vụ này, NLP sẽ trích xuất thông tin từ văn bản, phân tích ý nghĩa của nó và chuyển đổi văn bản thành các đặc trưng có thể đưa vào các mô hình học máy hoặc thuật toán học sâu để thực hiện việc phân loại. Ví dụ, các kỹ thuật xử lý ngôn ngữ như bag of words, TF-IDF, và nhúng từ (word embeddings) hỗ trợ trong việc chuyển đổi văn bản sang dạng số hóa. Sau đó, các mô hình học máy như Naive Bayes và SVM (Máy Vector Hỗ Trợ)... có thể được huấn luyện để phân loại văn bản vào các danh mục như tích cực hay tiêu cực, thư rác hay không phải thư rác, và tin thật hay tin giả.

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

TF-IDF là một phương pháp được sử dụng phổ biến trong Xử lý Ngôn ngữ Tự nhiên (NLP) và khai thác văn bản [19], [20]. Nó giúp đánh giá mức độ quan trọng của một từ trong một tài liệu bằng cách xem xét cả tần suất mà từ đó xuất hiện trong tài liệu và tần suất của từ đó trong toàn bộ tập hợp tài liệu. Nói cách khác, TF-IDF cho phép chúng ta xác định những từ nổi bật hơn trong một tài liệu so với các tài liệu khác trong cùng một tập hợp.

TF-IDF được tính bằng tích số của hai giá trị là Term Frequency (TF) và Inverse Document Frequency (IDF) như tong công thức (2.1).

$$\mathrm{TF-IDF}\,(t,d,D) \;=\; \mathrm{TF}\,(t,d) \;\times\; \mathrm{IDF}\,(t,D) \tag{2.1}$$

- **Term Frequency (TF):** là tần suất xuất hiện của một từ "t" trong một tài liệu "d", được tính như công thức (2.2).

$$\text{TF}\ (t, d)\ =\ \frac{\text{Số lần } t \text{ xuất hiện trong } d}{\text{Tổng số từ trong } d} \tag{2.2}$$

- **Inverse Document Frequency (IDF)**: đo lường tầm quan trọng của một từ dựa trên tần suất xuất hiện của nó trong tập hợp các tài liệu. Một cách dễ hiểu hơn IDF đo lường mức độ phổ biến của một từ để xem xét từ đó có giá trị đặc biệt hay không, được tính như công thức (2.3).

$$\text{IDF}\ (t, D)\ =\ \log\left(\frac{\text{Tổng số tài liệu } |D|\ +\ 1}{\text{Số tài liệu chứa } t\ +\ 1}\right) \tag{2.3}$$

Lưu ý: 1 được cộng vào để tránh việc chia cho 0 khi từ đó không xuất hiện trong bất kỳ tài liệu nào.

Khi IDF thấp, nghĩa là từ này xuất hiện ở rất nhiều tài liệu, có thể từ này không mang lại nhiều giá trị trong văn bản vì nó xuất hiện quá phổ biến. Ví dụ như những từ nối, giới từ, đại từ chỉ định, đây là những từ không quan trọng.

Khi IDF cao, nghĩa là từ này ít khi xuất hiện trong tập các tài liệu, có khả năng mang lại các thông tin quan trọng và đặc trưng, giúp cho việc phân loại tài liệu hiệu quả hơn.

Trong nghiên cứu này, TF-IDF sẽ chuyển đổi văn bản thành các vector đặc trưng. Sau đó, các vector này có thể được kết hợp với các mô hình Transformer nhằm nâng cao khả năng phân loại tin thật và tin giả. TF-IDF giúp mô hình tập trung vào các từ khóa quan trọng và giảm thiểu tác động của các từ phổ biến nhưng ít mang lại thông tin trong quá trình huấn luyện.

2.2.3 Word2Vec

Word2Vec là một phương pháp học biểu diễn từ (word embeddings) được nhóm nghiên cứu của Google phát triển vào năm 2013. Word2Vec thường được xây

dựng dựa trên dữ liệu là một tập văn bản có kích thước lớn và trả về kết quả là một không gian vector, có thể lên đến hàng trăm chiều. Trong không gian này, mỗi từ trong tập dữ liệu (corpus) sẽ được đại diện bởi một vector. Trong không gian vectơ này, các từ từ có ngữ cảnh, ngữ nghĩa tương tự, thường xuất hiện cùng nhau sẽ có các vectơ với vị trí gần nhau.

Word2Vec có hai kiến trúc chính: **Skip-Gram** và **Continuous Bag of Words (CBOW)**.

- **Skip-Gram** là phương pháp dự đoán các từ xung quanh thông qua một từ trung tâm cho trước. Khi một từ được chọn làm trung tâm, mô hình sẽ dự đoán các từ xung quanh nó trong một phạm vi ngữ cảnh được xác định trước (ví dụ: 2 từ trước và 2 từ sau).
- **CBOW** là phương pháp hoàn toàn ngược lại với Skip-Gram, thay vì dự đoán các từ ngữ cảnh từ một từ trung tâm, nó dự đoán từ trung tâm dựa trên các từ ngữ cảnh xung quanh. CBOW nhận các từ trong ngữ cảnh làm đầu vào và cố gắng dự đoán từ trung tâm. Ví dụ, trong câu "Học toán giúp phát triển tư duy", CBOW sẽ dự đoán từ trung tâm "phát triển" dựa trên các từ xung quanh nó là "Học", "toán", "tư duy".

Word2Vec có thể nâng cao hiệu quả trong bài toán phân loại tin giả nhờ khả năng chuyển từ ngữ thành các vector ngữ nghĩa, Word2Vec giúp mô hình phân loại phát hiện các điểm đặc trưng của tin giả, như ngôn ngữ giật gân hoặc phóng đại. Chúng tôi hy vọng khi kết hợp với các mô hình Transformer, Word2Vec sẽ cải thiện độ chính xác bằng cách cung cấp thông tin ngữ cảnh và ý nghĩa từ ngữ đầy đủ hơn so với các phương pháp chỉ dùng Transformer riêng biệt.

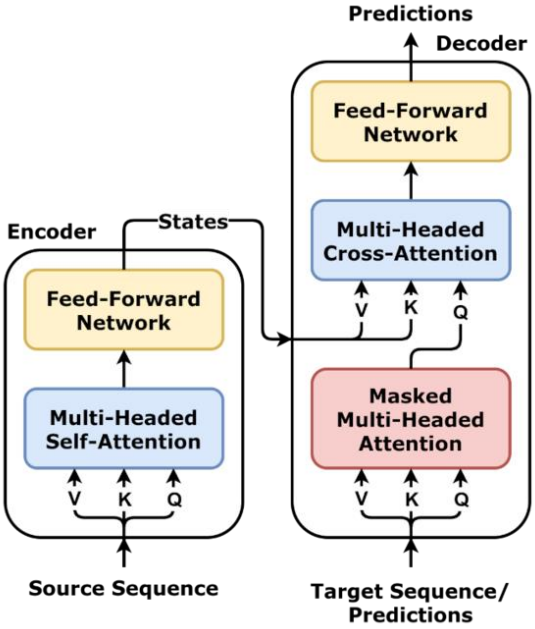
2.2.4 Mô hình Transformer

Mô hình Transformer đã thực sự tạo ra một bước đột phá trong lĩnh vực xử lý ngôn ngữ tự nhiên, được giới thiệu bởi Vaswani và cộng sự (2017) trong bài báo "Attention Is All You Need" [9]. Transformer nổi bật với kiến trúc tự chú ý (self-attention), có khả năng hiểu mối quan hệ giữa các từ trong một câu mà không cần

phải tuân theo thứ tự tuần tự như các mô hình trước đây, chẳng hạn như RNN hay LSTM.

RNN và LSTM được phát triển từ những năm đầu thế kỷ 20, nhưng chúng có nhược điểm là xử lý dữ liệu theo từng bước một, xử lý từng từ trong câu theo thứ tự. Điều này làm cho quá trình tính toán mất nhiều thời gian, đặc biệt là với những văn bản dài. Hơn nữa, các mạng RNN và LSTM cũng gặp khó khăn trong việc nhớ và giữ ngữ cảnh khi khoảng cách giữa các thông tin trong chuỗi quá xa nhau. Ví dụ, nếu một câu quá dài, các từ đầu câu có thể không còn ảnh hưởng nhiều đến từ cuối câu.

Ngược lại, mô hình Transformer có khả năng xử lý song song tất cả dữ liệu đầu vào cùng một lúc [9]. Đây là một thay đổi dựa vào kiến trúc thay vì tăng tốc bằng cách thêm GPU như nhiều phương pháp khác. Cơ chế tự chú ý (self-attention) của Transformer cũng đã khắc phục được khó khăn trong ghi nhớ ngữ cảnh bằng cách liên kết các từ và hiểu được mối quan hệ ngữ cảnh dù chúng có cách xa nhau trong văn bản.



Hình 1. Sơ đồ hoạt động của Encoder và Decoder [21]

Hình 1 mô tả sơ đồ hoạt động của mô hình Transformer, trong đó bao gồm hai thành phần chính là:

- **Encoder**: Encoder nhận vào một chuỗi các từ và chuyển đổi chúng thành các vector ngữ nghĩa. Mỗi encoder được tạo thành từ nhiều lớp xếp chồng lên nhau, với hai thành phần chính trong mỗi lớp: cơ chế tự chú ý (self-attention) và mạng nơ-ron truyền thẳng (feedforward neural network). Cơ chế tự chú ý giúp mô hình tập trung vào các từ quan trọng trong chuỗi trong khi lọc bỏ những từ ít liên quan hơn. Sau đó, mạng nơ-ron truyền thẳng sẽ xử lý các vector đã được điều chỉnh theo trọng số chú ý để tạo ra các biểu diễn ngữ nghĩa sâu hơn.
- **Decoder**: Decoder hoạt động tương tự như encoder nhưng có một vài tính năng bổ sung. Nó sử dụng cơ chế tự chú ý để tập trung vào đầu vào mục tiêu mà nó đang xử lý. Ngoài ra, nó còn sử dụng cơ chế chú ý chéo (cross-attention) để kết nối với đầu ra của encoder. Thiết lập này cho phép decoder tạo ra các biểu diễn ý nghĩa dựa trên cả chuỗi đầu vào gốc và chuỗi đầu ra mà nó đã tạo ra.

Sự kết hợp giữa Encoder và Decoder cho phép Transformer xử lý các nhiệm vụ ngôn ngữ như dịch máy, tóm tắt văn bản, tạo văn bản và phân loại văn bản một cách linh hoạt và hiệu quả.

2.2.5 BERT (Bidirectional Encoder Representations from Transformers)

BERT là một mô hình ngôn ngữ tiền huấn luyện được thiết kế để hiểu ngữ cảnh của từ theo cả hai hướng (từ trái sang phải và từ phải sang trái) trong một câu, do Google giới thiệu vào tháng 10 năm 2018 [10]. Mô hình này học cách hiểu văn bản bằng cách tự học từ dữ liệu, chuyển văn bản thành một chuỗi các vector.

BERT là một mô hình Transformer chỉ sử dụng phần "encoder", bao gồm 4 phần chính sau đây:

- Tokenizer: chuyển đổi văn bản thành chuỗi các số nguyên gọi là "tokens".

- Embedding: chuyển các token thành các vector giá trị thực, giúp biến các token từ dạng rời rạc thành không gian Euclidean với kích thước thấp hơn.
- Encoder: là một chồng các khối Transformer với cơ chế tự chú ý (self-attention) mà không có lớp che chắn nguyên nhân (causal masking), giúp hiểu ngữ cảnh của văn bản.
- Task Head: chuyển đổi các vector đại diện cuối cùng thành các token mã hóa one-hot bằng cách dự đoán phân phối xác suất trên các loại token.

BERT được đào tạo trên hai nhiệm vụ chính:

- **Mô hình Ngôn ngữ Ẩn (MLM):** Trong nhiệm vụ này, một số từ trong câu được thay thế bằng ký hiệu [MASK], và mô hình cần dự đoán các từ bị ẩn dựa trên ngữ cảnh xung quanh. Điều này giúp mô hình học cách hiểu ý nghĩa của các từ trong câu mà không có đầy đủ thông tin, từ đó cải thiện khả năng nắm bắt ý nghĩa của từ trong nhiều tình huống khác nhau.
- **Dự đoán Câu Kế Tiếp (NSP):** Nhiệm vụ này yêu cầu mô hình dự đoán liệu một câu cho trước có phải là câu tiếp theo của câu trước đó hay không, nhằm nâng cao khả năng hiểu biết của mô hình về mối liên hệ giữa các câu. Điều này rất quan trọng trong việc xử lý văn bản dài và phức tạp.

BERT đã đạt được kết quả xuất sắc trong nhiều nhiệm vụ NLP như phân loại văn bản, nhận diện thực thể và trả lời câu hỏi, mang lại sự cải thiện rõ rệt so với các mô hình trước đó và trở thành một trong những mô hình ngôn ngữ lớn điển hình.

2.2.6 RoBERTa (A Robustly Optimized BERT Pretraining Approach)

RoBERTa là một biến thể của BERT, được phát triển bởi các nhà nghiên cứu tại Facebook AI. RoBERTa cũng là một mô hình ngôn ngữ dựa trên transformer sử dụng self-attention để xử lý chuỗi đầu vào và tạo ra các biểu diễn ngữ cảnh của các từ trong một câu [11].

Tuy nhiên RoBERTa được tối ưu hóa để cải thiện hiệu suất bằng cách thay đổi một số phương pháp huấn luyện như:

- Loại bỏ nhiệm vụ Dự đoán Câu Kế Tiếp (NSP): RoBERTa chỉ tập trung vào nhiệm vụ Masked Language Modeling (MLM), và bỏ qua nhiệm vụ dự đoán câu tiếp theo (NSP) như ở BERT, điều này có thể giúp cho mô hình học ngữ cảnh tốt hơn.
- Tăng kích thước tập dữ liệu và thời gian huấn luyện: RoBERTa sử dụng một tập dữ liệu lớn hơn nhiều so với BERT để huấn luyện. Cụ thể, RoBERTa được đào tạo trên một tập dữ liệu gồm 160GB văn bản từ các nguồn như BookCorpus, Wikipedia, Common Crawl, và OpenWebText; lớn hơn 10 lần so với tập dữ liệu được sử dụng để đào tạo BERT. Cùng với đó là thời gian huấn luyện dài với nhiều vòng huấn luyện (epochs) để mô hình có thể tiếp cận tối đa khả năng của nó.

Kết quả cho thấy RoBERTa đã thể hiện hiệu suất vượt trội trong các nhiệm vụ NLP như dịch ngôn ngữ, phân loại văn bản và trả lời câu hỏi. Nó cũng đã được sử dụng làm mô hình cơ sở cho nhiều mô hình NLP thành công khác và đã trở thành lựa chọn phổ biến cho các ứng dụng nghiên cứu.

2.2.7 PhoBERT

PhoBERT là một biến thể của BERT được phát triển nhằm giải quyết những hạn chế của BERT khi xử lý tiếng Việt [5]. Mặc dù BERT rất mạnh trong việc xử lý ngôn ngữ tự nhiên, nhưng vì nó chủ yếu được huấn luyện trên dữ liệu tiếng Anh, nên hiệu quả khi áp dụng cho các ngôn ngữ khác không cao. Do được đào tạo hoàn toàn trên dữ liệu văn bản tiếng Việt, PhoBERT nắm bắt rất tốt các đặc điểm ngữ nghĩa và cú pháp cụ thể của ngôn ngữ này.

PhoBERT tích hợp các cải tiến từ RoBERTa, chẳng hạn như loại bỏ nhiệm vụ Dự đoán Câu Kế Tiếp (NSP) và chỉ sử dụng Mô hình Ngôn ngữ Ẩn (MLM) [5] – một nhiệm vụ giúp mô hình hiểu ngữ cảnh của các từ trong câu bằng cách dự đoán từ bị ẩn, đồng thời được đào tạo trên tập dữ liệu quy mô lớn.

PhoBERT được huấn luyện với một tập dữ liệu lớn có kích thước khoảng 20 GB, bao gồm khoảng 1 GB được trích xuất từ Wikipedia tiếng Việt và khoảng 19 GB từ các bài báo tiếng Việt [5]. Điều này đảm bảo rằng mô hình có thể học từ nhiều nguồn khác nhau, bao gồm cả ngôn ngữ chính quy và ngôn ngữ báo chí, giúp nó có khả năng hiểu tốt các bối cảnh ngôn ngữ khác nhau trong thực tế.

Trước khi dữ liệu được đưa vào bộ mã hóa Byte-Pair Encoding (BPE), PhoBERT sử dụng công cụ RDRSegmenter từ VnCoreNLP để tách từ. Điều này giúp mô hình xử lý tốt hơn các vấn đề về ngữ nghĩa và cú pháp trong tiếng Việt, vì tiếng Việt là ngôn ngữ dính kết (agglutinative language), nghĩa là một từ có thể chứa nhiều thành phần khác nhau, làm phức tạp việc tách từ.

Nhờ vào việc huấn luyện và tinh chỉnh trên một tập dữ liệu tiếng Việt lớn và đa dạng, PhoBERT có khả năng xử lý tốt hơn các cấu trúc phức tạp của tiếng Việt so với BERT và RoBERTa. Điều này giúp cho PhoBERT được chọn để sử dụng trong nhiều bài toán xử lý ngôn ngữ tự nhiên dành cho Tiếng Việt như: phân loại văn bản, trả lời câu hỏi, nhận diện thực thể có tên, dịch máy, tóm tắt văn bản,…

Cụ thể trong nhiệm vụ phân loại tin tức thật và giả bằng tiếng Việt, PhoBERT thể hiện hiệu suất vượt trội trong việc phân tích và hiểu ngữ nghĩa của văn bản Tiếng Việt. Do được huyến luyện từ một tập dữ liệu lớn và đa dạng, PhoBERT có thể phân biệt giữa tin tức thật và giả thông qua sự khác biệt về ngữ cảnh, ngữ pháp và phong cách ngôn ngữ. Từ đó giúp PhoBERT nâng cao độ chính xác của mình trong việc phát hiện tin giả Tiếng Việt so với các mô hình khác.

2.2.8 ViBERT (Vietnamese BERT)

ViBERT là một biến thể của mô hình BERT, do **FPT AI phát triển, mô hình này** được thiết kế và tinh chỉnh cho tiếng Việt. ViBERT được xây dựng dựa trên mô hình BERT, sử dụng kiến trúc Transformer với **encoder** và **self-attention** [3]. Nó cũng duy trì khả năng học ngữ cảnh của từ từ cả hai hướng, từ đó giúp mô hình nắm bắt ngữ nghĩa chính xác hơn so với các mô hình xử lý tuần tự.

ViBERT được huấn luyện trên một khối lượng lớn dữ liệu Tiếng Việt, bao gồm báo chí, mạng xã hội, tài liệu pháp lý, sách văn học, và các nguồn khác. Cũng giống như BERT, ViBERT được đào tạo trên hai nhiệm vụ chính là Masked Language Model (MLM) và Next Sentence Prediction (NSP) mà không loại bỏ nhiệm vụ Dự đoán Câu Kế Tiếp (NSP) như PhoBERT và RoBERTa.

Khác với PhoBERT, ViBERT sử dụng phương pháp **WordPiece Tokenization** để tác từ tiếng Việt thành các thành phần nhỏ hơn (subwords). Tuy nhiên, ViBERT không có quá trình xử lý tách từ (word segmentation) chuyên biệt cho tiếng Việt trước khi áp dụng WordPiece.

Mục tiêu chính của ViBERT là cải thiện khả năng xử lý ngôn ngữ tự nhiên (NLP) đối với tiếng Việt, bao gồm các tác vụ như phân loại văn bản, nhận diện thực thể, dịch tự động, và các ứng dụng khác.

2.2.9 ViSoBERT (Vietnamese Social BERT)

ViSoBERT (Vietnamese Social BERT) là một mô hình ngôn ngữ tiếng Việt được phát triển nhằm cải thiện khả năng xử lý ngữ nghĩa và hiểu ngữ cảnh trong các bài viết trên mạng xã hội. ViSoBERT cũng được phát triển bởi FPT AI [4].

ViSoBERT cũng dựa trên kiến trúc BERT và ViBERT, nhưng được tinh chỉnh với dữ liệu mạng xã hội tiếng Việt, nguồn dữ liệu huấn luyện ViSoBERT là từ các nền tảng mạng xã hội tiếng Việt, bao gồm Facebook, Twitter, và các diễn đàn trực tuyến khác. Dữ liệu này giúp mô hình hiểu và phân tích các đặc điểm ngôn ngữ đặc thù của mạng xã hội như ngữ điệu, từ viết tắt, và các cách diễn đạt không chính thức.

Mô hình này nhằm mục đích phân tích và hiểu rõ hơn về ngữ cảnh mạng xã hội, nơi mà ngôn ngữ được sử dụng linh hoạt và không theo các quy tắc chính thức. ViSoBERT có thể được sử dụng cho các ứng dụng như phát hiện tin giả, phân tích cảm xúc, và nhiều tác vụ NLP khác có liên quan đến dữ liệu mạng xã hội tiếng Việt.

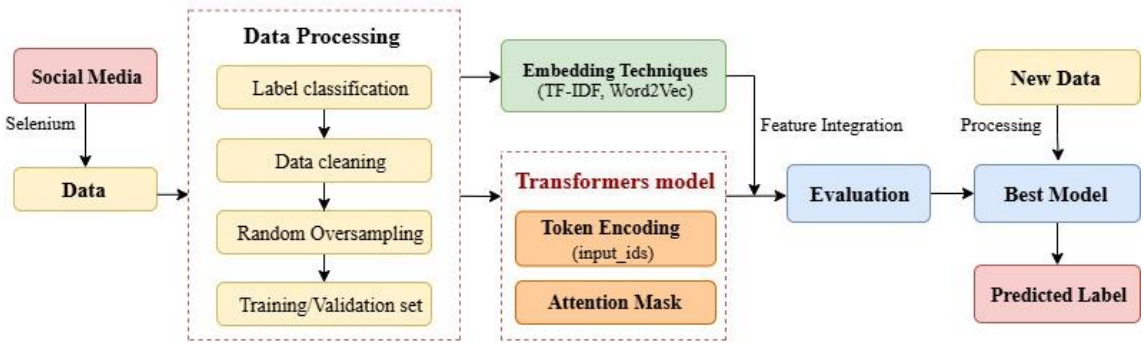
CHƯƠNG 3: PHƯƠNG PHÁP PHÁT HIỆN TIN GIẢ

3.1 Mô hình tổng quát

Hệ thống của chúng tôi được chia thành bốn giai đoạn chính, được thể hiện tổng quát trong **Hình 2**: (1) Thu thập dữ liệu, (2) Xử lý dữ liệu, (3) Huấn luyện mô hình, và (4) Đánh giá mô hình.

- **Thu thập dữ liệu:** Ở giai đoạn đầu tiên, chúng tôi đã thu thập dữ liệu từ các bài đăng trên Facebook, bao gồm cả các trang tin tức chính thống và các trang thường xuyên đăng thông tin sai lệch về các chủ đề như tin tức thời sự, đời sống, và chính trị. Chúng tôi thu thập các chi tiết như tác giả, nội dung bài đăng, liên kết bài đăng, và cả các bình luận. Giai đoạn này rất quan trọng vì tập dữ liệu thu thập được sẽ ảnh hưởng lớn đến kết quả của nghiên cứu.
- **Xử lý dữ liệu:** Ở giai đoạn này, dữ liệu thu thập được sẽ trải qua một loạt các bước tiền xử lý bao gồm làm sạch, chuẩn hóa văn bản và bước quan trọng nhất là gán nhãn các bài viết với nhãn thật hoặc giả. Sau khi tiền xử lý, dữ liệu sẽ được chia thành tập huấn luyện và tập kiểm tra và sẵn sàng cho bước huấn luyện mô hình.
- **Huấn luyện mô hình:** Ở giai đoạn tiếp theo, chúng tôi tiến hành huấn luyện các mô hình Transformer như: BERT, RoBERTa, PhoBERT, ViBERT và ViSoBERT. Chúng tôi áp dụng các kỹ thuật huấn luyện khác nhau cho từng mô hình để tối ưu hóa hiệu suất, bao gồm tinh chỉnh siêu tham số và sử dụng các kỹ thuật như kiểm tra chéo. Sau khi huấn luyện, chúng tôi dùng các mô hình để dự đoán nhãn cho các văn bản trong tập thử nghiệm và tính toán các giá trị như độ chính xác và so sánh kết quả trong bước tiếp theo.
- **Đánh giá mô hình:** Giai đoạn cuối cùng chúng tôi đánh giá hiệu suất của các mô hình đã huấn luyện. Chúng tôi đánh giá dựa trên độ chính xác, độ tinh cậy, độ nhạy, F1-Score và AUC khi so sánh nhãn dự đoán và nhãn thực tế

của tập huấn luyện. Dựa trên kết quả đánh giá, chúng tôi có thể tinh chỉnh thêm mô hình hoặc điều chỉnh các kỹ thuật tiền xử lý để cải thiện hiệu suất.



Hình 2. Mô hình tổng quát của hệ thống

3.2 Thu thập dữ liệu

Do các tập dữ liệu về tin tức và các bài đăng trên mạng xã hội tiếng Việt còn hạn chế hoặc các tập dữ liệu hiện có có thể không còn phù hợp với bối cảnh hiện tại, chúng tôi đã quyết định thu thập dữ liệu một bộ dữ liệu mới để nghiên cứu và hy vọng có thể đóng góp vào nguồn tài nguyên dữ liệu để hỗ trợ các nghiên cứu mới trong tương lai. Quá trình thu thập dữ liệu của chúng tôi gồm các bước dưới đây:

Đầu tiên, chúng tôi đã chọn lọc các bài đăng một cách thủ công. Đối với tin tức thật, chúng tôi xác định các trang tin tức chính thức của Việt Nam trên Facebook, bao gồm các kênh truyền thông lớn, trang thông báo của chính phủ và các nguồn đáng tin cậy khác. Đây là những nguồn đáng tin cậy để thu thập tin tức chính xác. Để thu thập tin tức giả, chúng tôi tìm các trang báo lá cải và các nhóm Facebook thường xuyên đăng tải tin tức giật gân và lan truyền thông tin sai lệch về chính trị và xã hội.

Sau khi chọn lọc các nguồn dữ liệu cần thiết, chúng tôi đã sử dụng Selenium để tự động thu thập dữ liệu bằng cách mô phỏng các hành động của người dùng trên trình duyệt web và thông qua đó trích xuất dữ liệu. Cuối cùng, chúng tôi thu thập được hai tập dữ liệu: một cho tin tức thật và một cho tin tức giả, mỗi tập dữ liệu có cấu trúc như trong **Hình 3**.

date	author_id	content	label	link	comment_list
29/07/2024 13:45	https://www.facebook	Vụ xe bán tải cố vượt rào chắn, bị tàu hỏa tông ở E	0	https://www.fật" }, {	"comment_id": "c36", "author": "Trần Phúc Hậu",
30/07/2024 23:58	https://www.facebook	TPHCM: Hơn 4.600 ca mắc sốt xuất huyết, nhiều 'đ	0	https://www.fuyễn",	"content": "Ánh Tây coi chừng bối nhen" }, { "comr
30/07/2024 22:59	https://www.facebook	Nóng: Ngộ độc hàng loạt tại trụ sở công ty mẹ TikT	0	https://www.for": "Lê Nguyễn Bảo Thư",	"content": "Nhii Mai Kim Ngân Hồng
31/07/2024 14:50	https://www.facebook	Ngày mai: Giá xăng trong nước có thể giảm lần thú	0	https://www.făng nào" }, {	"comment_id": "c6", "author": "Lắm Chuyện
31/07/2024 12:30	https://www.facebook	Pin dự phòng của hành khách bốc cháy tại nhà ga :	0	https://www.fềm quá" }, {	"comment_id": "c20", "author": "Thang Vo",
31/07/2024 10:50	https://www.facebook	Thương tâm quá: Trong lúc chở nhau trên xe máy '	0	https://www.fl": "c14",	"author": "Vũ Hà", "content": "Nam mô a di đà phậ
30/07/2024 22:50	https://www.facebook	THƯƠNG TÂM HÀ GIANG: ĐẤT ĐÁ LĂN TỪ TALUY D	0	https://www.fiuy",	"content": "A di đà Phật" }, { "comment_id": "c9",
30/07/2024 22:30	https://www.facebook	NỮ TÀI XẾ Ô TÔ ĐẠP NHẦM CHÂN GA GÂY TNGT LI	0	https://www.fạy mẹ trẻ ..may k chết ng ..."	}, { "comment_id": "c12", "au

Hình 3. Cấu trúc và nội dung của hai tập dữ liệu

Tuy nhiên, quá trình thu thập dữ liệu đã gặp phải một số thách thức, bao gồm:

- **Thời gian thu thập hạn chế:** do thời gian hạn chế, và việc mô phỏng duyệt web để thu thập dữ liệu tốn khá nhiều thời gian, điều này ảnh hưởng đến khối lượng dữ liệu chúng tôi có thể thu thập.
- **Khó khăn trong việc tìm kiếm nguồn tin giả:** Một số bài viết đã bị gỡ bỏ sau khi bị báo cáo, điều này làm giảm số lượng tin tức giả mà chúng tôi thu thập được.

Do đó, tập dữ liệu mà chúng tôi thu thập được có sự chênh lệch về số lượng tin thật và giả. Chúng tôi vẫn đang tiếp tục thu thập các tin tức mới và bài viết mới trong thời gian có thể và phù hợp với thời hạn báo cáo.

3.3 Xử lý dữ liệu

Ở giai đoạn này, chúng tôi đã thực hiện xử lý dữ liệu qua các bước sau:

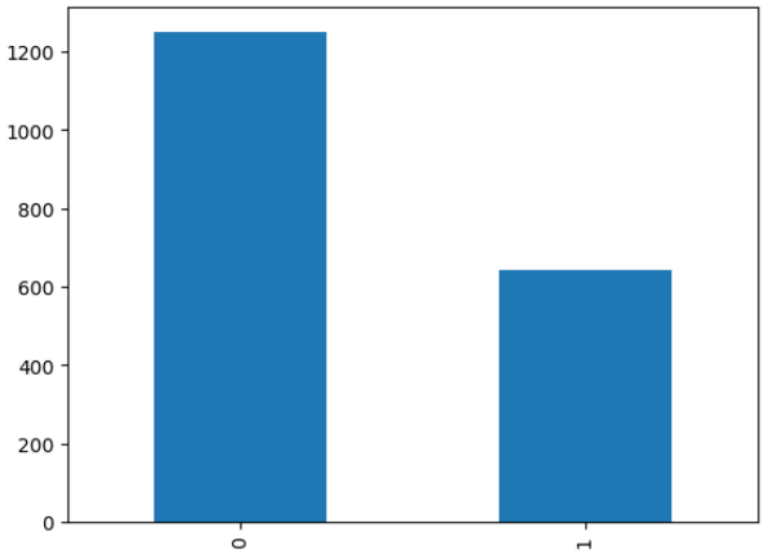
- **Loại bỏ các trường hợp không hợp lệ**: các trường hợp dữ liệu không hợp lệ gồm các dòng bị trống do lỗi khi thu thập, định dạng dữ liệu không hợp lệ hoặc bị trùng lặp.
- **Chuyển đổi văn bản**: tất cả các văn bản được chuyển đổi về dạng chữ thường, các ký tự đặc biệt, liên kết URL cũng được loại bỏ. Việc này làm giảm sự nhiễu loạn thông tin, phân biệt chữ hoa chữ thường và giúp tập trung vào nội dung chính của các bài đăng.
- **Chọn lọc các trường dữ liệu**: dữ liệu thu thập được gồm nhiều thông tin như nội dung bài viết, người đăng, liên kết, danh sách bình luận,… tuy nhiên hiện tại chúng tôi chỉ chọn các trường dữ liệu quan trọng để sử dụng cho phân tích

đó là nội dung bài viết và nhãn phân loại. Dữ liệu còn lại được lưu trữ để mở rộng nghiên cứu trong tương lai.

	content	label
0	vụ xe bán tải cố vượt rào chắn bị tàu hỏa tông...	0
1	tphcm hơn 4600 ca mắc sốt xuất huyết nhiều điể...	0
2	nóng ngộ độc hàng loạt tại trụ sở công ty mẹ t...	0
3	ngày mai giá xăng trong nước có thể giảm lần t...	0
4	pin dự phòng của hành khách bốc cháy tại nhà g...	0

Hình 4. Dữ liệu sau khi được xử lý

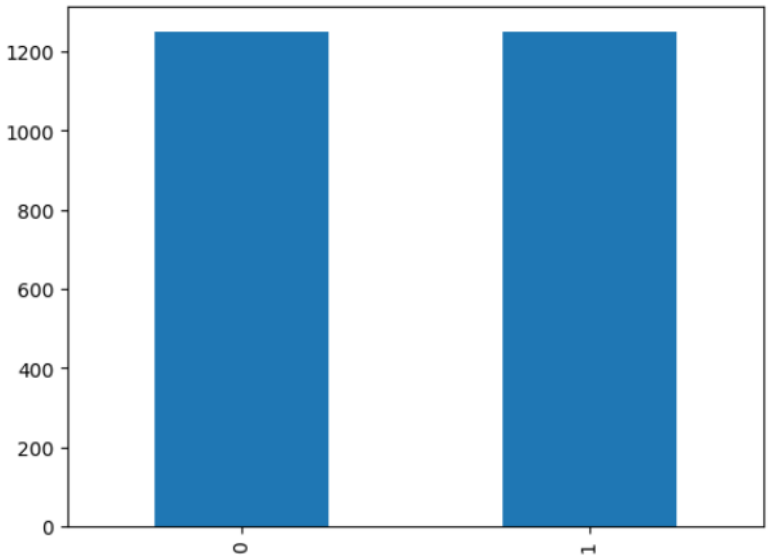
Sau khi hoàn thành quá trình xử lý dữ liệu qua các bước trên, chúng tôi đã thu thập được hai tập dữ liệu chính: tin thật và tin giả. Mỗi tập dữ liệu chứa hai trường thông tin đã chọn là nội dung tin tức và nhãn phân loại tương ứng, như minh họa ở **Hình 4**.



Hình 5. Kích thước của hai tập dữ liệu sau khi thu thập

Tuy nhiên số lượng mẫu tin giả có số lượng ít hơn đáng kể so với số lượng tin thật (như được thể hiện trong biểu đồ ở **Hình 5** với nhãn 0 là tin thật và nhãn 1 là tin giả). Điều này có thể dẫn đến sự thiên lệch trong quá trình huấn luyện mô hình và kết quả không chính xác. Để giải quyết vấn đề này, chúng tôi đã thực hiện hai giải pháp:

- **Tìm thêm nguồn dữ liệu bổ sung**: Chúng tôi đã trích thêm các tin tức giả từ tập dữ liệu *VFND - Vietnamese Fake News Datasets* [22], được thu thập trong luận án của tác giả Ho Quang Thanh. Tuy nhiên, vì tập dữ liệu này được thu thập từ khoảng thời gian năm 2019 đến 2020, nên để không ảnh hưởng đến tập dữ liệu mà chúng tôi đã thu thập, chúng tôi chỉ chọn từ VFND những tin tức không thay đổi theo thời gian, chẳng hạn như kiến thức khoa học đã được chứng minh là sai hoặc tin tức về mê tín dị đoan, lối sống lệch lạc. Dữ liệu bổ sung này chiếm không quá 20% tổng số tin giả mà chúng tôi đã thu thập trong bộ dữ liệu của mình.
- **Sử Dụng Kỹ Thuật Tăng Cường Ngẫu Nhiên (Random Oversampling)**: Chúng tôi đã sử dụng kỹ thuật Random Oversampling từ thư viện "imbalanced-learn" để cân bằng dữ liệu. Kỹ thuật này giúp tăng số lượng mẫu của các nhãn ít hơn bằng cách sao chép ngẫu nhiên các mẫu hiện có cho đến khi số lượng mẫu của các nhãn được cân bằng [23]. Điều này làm giảm sự thiên lệch và giúp cải thiện độ chính xác của mô hình.



Hình 6. Kích thước của 2 tập dữ liệu sau khi xử lý

Biểu đồ **Hình 6** thể hiện tỷ lệ của hai nhãn dữ liệu sau khi thực hiện hai giải pháp trên để cân bằng (nhãn 0 là tin thật và nhãn 1 là tin giả). Việc cân bằng các nhãn giúp đảm bảo rằng mô hình không bị thiên lệch về lớp số lượng nhiều hơn,

giúp nâng cao độ chính xác trong việc phân loại cả hai loại tin tức. Điều này giúp mô hình đưa ra kết quả phân loại đáng tin cậy và chính xác hơn so với khi dữ liệu bị mất cân bằng.

3.4 Huấn luyện các mô hình Tranformer

Trong giai đoạn này, chúng tôi xây dựng và huấn luyện các mô hình Transformer sau khi đã có được các tập dữ liệu. Các mô hình chính mà chúng tôi lựa chọn bao gồm BERT, RoBERTa, và đặc biệt là ViBERT, ViSoBERT, PhoBERT – các mô hình được thiết kế dành riêng cho Tiếng Việt. Các mô hình này nổi bật nhờ khả năng xử lý ngôn ngữ tự nhiên với số lượng lớn dữ liệu và đã đạt được các kết quả tốt trong các nghiên cứu trước đó.

Đối với mỗi mô hình, chúng tôi sử dụng các tokenizer tương ứng với từng mô hình để chuyển đổi các văn bản thành các chuỗi số liệu đầu vào mà mô hình có thể xử lý được. Các tokenizer này không chỉ đơn thuần chia từ mà còn thực hiện việc mã hóa các token dưới dạng số, giúp mô hình nhận diện các từ, cụm từ, và cấu trúc câu.

Chúng tôi điều chỉnh các thông số phù hợp với từng mô hình thông qua tinh chỉnh siêu tham số (hyperparameter tuning). Các tham số như *learning rate*, *batch size*, và *number of epochs* được điều chỉnh để đảm bảo mô hình học tốt từ dữ liệu huấn luyện và tránh hiện tượng quá khớp (overfitting) có thể khiến hiệu suất mô hình bị giảm. Để ngăn ngừa tình trạng này, chúng tôi đã sử dụng kỹ thuật early stopping, cho phép chúng tôi dừng quá trình huấn luyện khi mô hình không còn cải thiện, tránh việc mô hình trở nên phức tạp một cách không cần thiết.

Ngoài việc sử dụng các mô hình Transformer độc lập, chúng tôi cũng triển khai phương pháp kết hợp các mô hình Transformer với các kỹ thuật nhúng từ (word embedding techniques) trong NLP. Hai phương pháp chúng tôi lựa chọn bao gồm các đặc trưng tần suất từ TF-IDF (Term Frequency-Inverse Document Frequency) và các mối quan hệ ngữ nghĩa từ Word2Vec (Word to Vector). Mục tiêu của việc này là tận dụng tối đa khả năng hiểu ngữ nghĩa mạnh mẽ của mô hình

Transformer, đồng thời bổ sung thêm thông tin về tần suất từ và các mối quan hệ ngữ nghĩa, giúp mô hình nắm bắt được các yếu tố quan trọng từ nhiều góc độ. Chúng tôi cũng hy vọng rằng sự kết hợp này sẽ nâng cao đáng kể khả năng phân loại tin giả so với việc chỉ sử dụng Transformer.

Cuối cùng, chúng tôi dùng các mô hình đã huấn luyện để dự đoán nhãn cho tập dữ liệu kiểm tra. Các nhãn do mô hình phân loại sẽ được so sánh với nhãn thực tế, sau đó đánh giá hiệu quả của từng mô hình cũng như so sánh chúng với nhau thông qua các chỉ số như A*ccuracy, Precision, Recall,* và *F1-score.*

3.5 Cách kết hợp Transformers và TF-IDF

Trong nghiên cứu của chúng tôi, để kết hợp TF-IDF với các mô hình Transformer, chúng tôi đã chọn cách tích hợp TF-IDF như một tầng bổ sung sau khi đã tạo ra các đặc trưng từ Transformer. Chúng tôi chọn cách làm này thay vì TF-IDF được đưa vào ngay từ đầu như một phần đầu vào của mô hình. Cách làm này sẽ giữ cho quá trình học ngữ cảnh và quan hệ giữa các từ trong câu được xử lý hoàn toàn bởi Transformer, trong khi các đặc trưng dựa trên tần suất xuất hiện từ (TF-IDF) sẽ được bổ sung vào giai đoạn cuối của quá trình huấn luyện.

Đầu tiên, chúng tôi đã chuyển đổi các văn bản thành các đặc trưng (input_ids và attention_mask) thông qua tokenizer của các mô hình Transformer. Đồng thời chúng tôi cũng dùng TF-IDF để tạo ra một tập các đặc trưng riêng biệt từ các văn bản, tập đặc trưng riêng biệt này lưu trữ tần suất xuất hiện của các từ trong ngữ cảnh cụ thể.

Để kết hợp cả hai loại đặc trưng này, chúng tôi chọn các đặc trưng từ token [CLS] của Transformer, token này đại diện cho toàn bộ câu và chứa thông tin ngữ cảnh tổng hợp từ tất cả các từ trong câu. Các đặc trưng này là kết quả từ hidden state của layer cuối cùng. Sau đó, chúng tôi lấy các đặc trưng này kết hợp với các vector TF-IDF,tạo thành một tập hợp đặc trưng kết hợp cả thông tin ngữ cảnh và tần suất từ. Cuối cùng tập hợp các đặc trưng này được đưa vào một tầng fully connected để tạo ra kết quả dự đoán của mô hình.

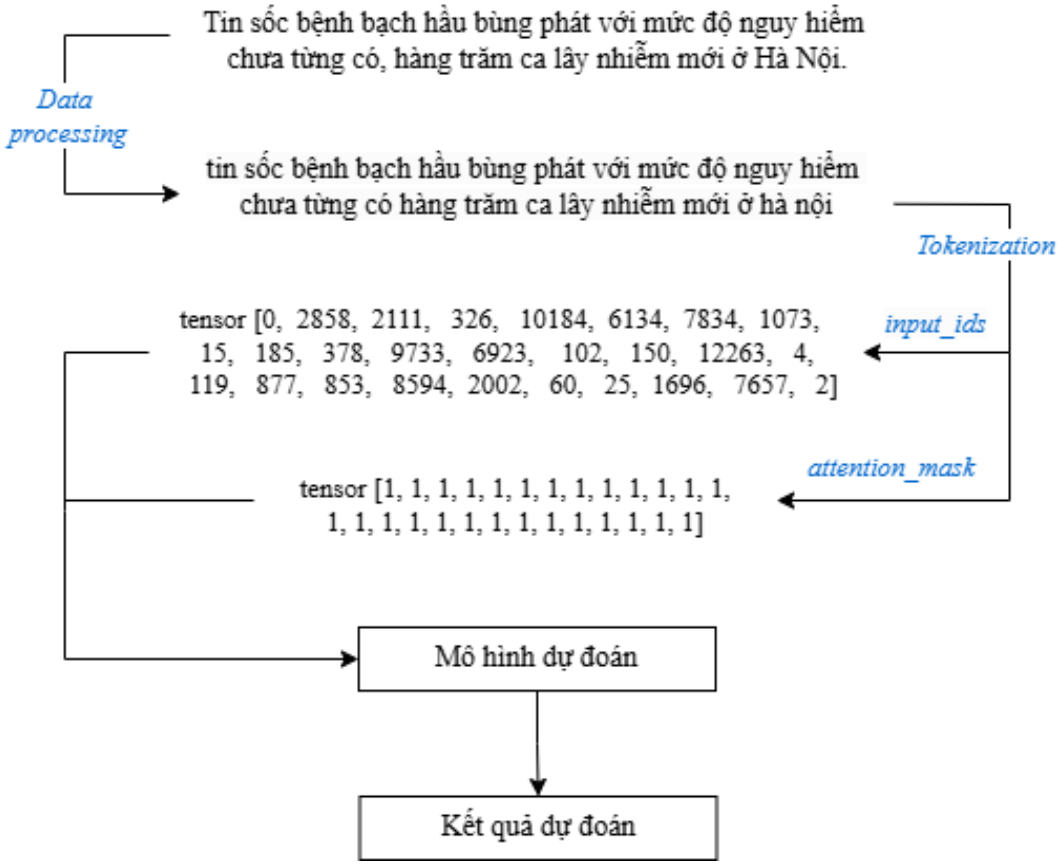
Chúng tôi đã cho các mô hình Transformer hoạt động độc lập trong việc cung cấp các đặc trưng ngữ cảnh cho văn bản, trong khi TF-IDF là một bộ đặc trưng được tạo ra riêng biệt lưu trữ thông tin về tần suất từ và không liên quan đến ngữ cảnh. TF-IDF được chúng tôi tích hợp ở giai đoạn cuối trước khi đưa ra kết quả dự đoán, có nghĩa là TF-IDF không ảnh hưởng trực tiếp đến quá trình xử lý của các mô hình Transformer, mà nó chỉ là một phần bổ sung thêm vào quá trình kết hợp thông tin để cải thiện khả năng dự đoán của mô hình.

Hình 7. Tần suất các từ trong tập dữ liệu tin thật

Hình 8. Tần suất các từ trong tập dữ liệu tin giả

Hình 7 và **Hình 8** thể hiện tần suất của các từ xuất hiện trong các tin tin tức thật và giả trong tập dữ liệu mà chúng tôi đã thu thập được.

3.6 Ví dụ minh họa



Hình 9. Ví dụ về quy trình mô hình Transformer dự đoán tin giả

Các mô hình được huấn luyện với dữ liệu trong tập huấn luyện, giúp chúng nắm bắt được các đặc điểm thường thấy trong tin thật và tin giả. Khi gặp một tin mới, mô hình sẽ so sánh với những cấu trúc, ngữ cảnh đã học để đưa ra dự đoán tin mới này là thật hay giả.

Ví dụ như trong dữ liệu huấn luyện có các tin tức về bệnh bạch hầu như: "Hà Nội triển khai các biện pháp đề phòng bệnh bạch hầu xâm nhập" và "Trong 6 tháng đầu năm 2024, Việt Nam ghi nhận 5 trường hợp mắc bệnh bạch hầu" là những thông tin chính xác. Trong khi đó "Sốc, hiện tượng chưa từng thấy. Hàng nghìn người xếp hàng chờ xét nghiệm bệnh bạch hầu" là thông tin giả. Tin này đã dựa vào một dịch bệnh đang xảy ra trong thực tế sau đó phóng đại sự thật với các cụm từ

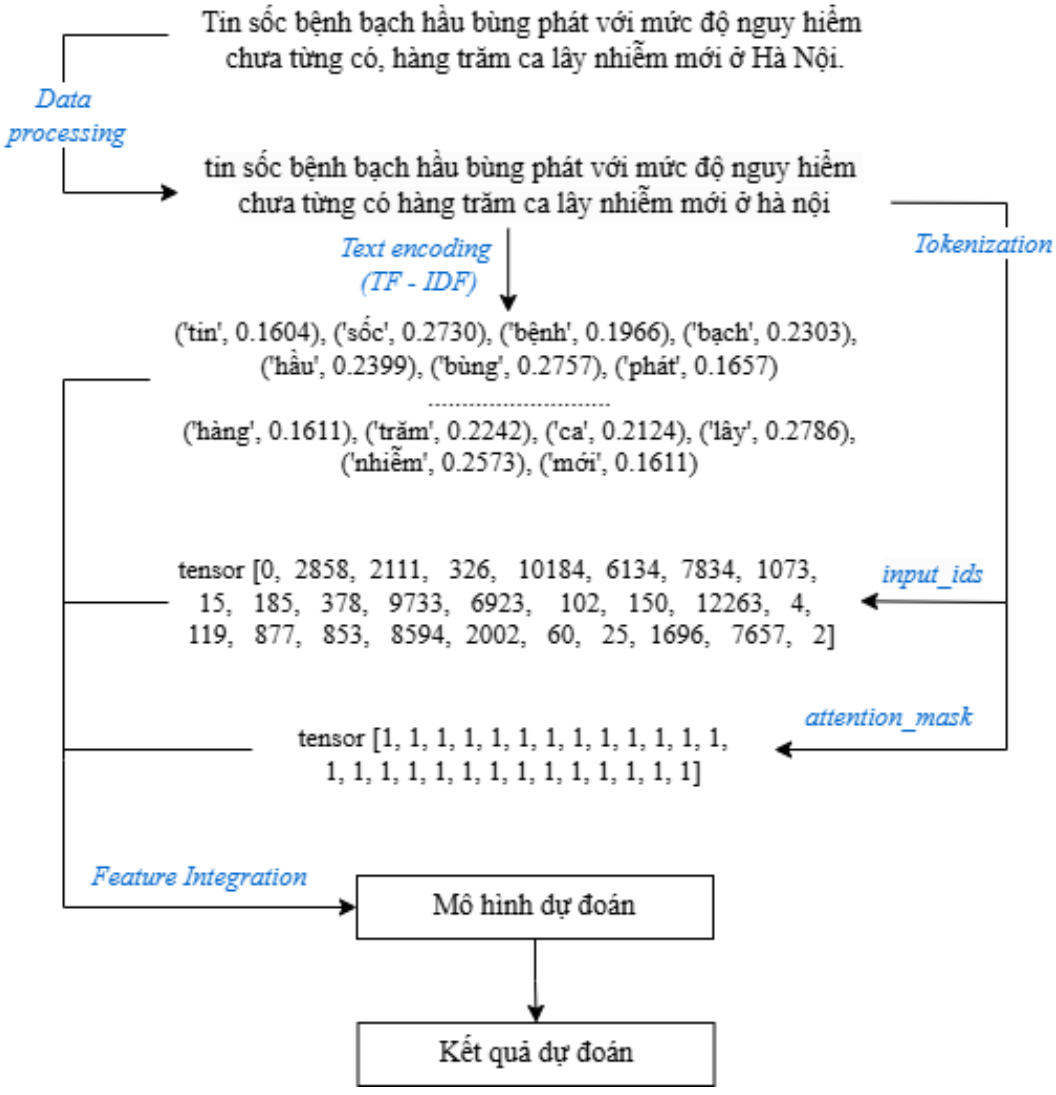
như “sốc”, “hàng nghìn người”, “hiện tượng chưa từng thấy”,… Đây là những ngôn từ thường xuyên xuất hiện trong tin giả, nhằm tạo ra các thông tin có ngữ nghĩa giật gân, gây sốc, đánh vào tâm lý của người đọc để được lan truyền nhanh chóng và hiếm khi xuất hiện trong các tin thật, chính thống.

Qua quá trình huấn luyện, mô hình học cách phân biệt giữa cách sử dụng ngôn ngữ và cấu trúc câu điển hình của tin thật và tin giả. Khi đó mô hình nhận ra rằng tin giả thường có ngữ nghĩa nhấn mạnh tác động tiêu cực, cường điệu hóa vấn đề nhằm tạo sự hoang mang, bàn tán của dư luận trên mạng xã hội.

Hình 9 mô tả quy trình mà các mô hình Transformer dự đoán một tin tức là thật hay giả. Khi mô hình gặp một tin mới, ví dụ như “Tin sốc bệnh bạch hầu bùng phát với mức độ nguy hiểm chưa từng có, hàng trăm ca lây nhiễm mới ở Hà Nội”. Đầu tiên qua bước tiền xử lý, mô hình sẽ loại bỏ các kí tự đặc biệt và các từ không hợp lệ (nếu có), sau đó chuyển đổi toàn bộ văn bản thành chữ thường. Tiếp theo các mô hình Transformer sẽ mã hóa các văn bản thành các đặc trưng dưới dạng input_ids và attention_mask để chuẩn bị cho quá trình phân tích ngữ nghĩa và cấu trúc ngữ pháp. Mô hình dựa vào các cụm từ như "tin sốc", "hàng trăm", “nguy hiểm chưa từng có” là những dấu hiệu chỉ ra rằng câu này có ngữ nghĩa cường điệu, ngữ cảnh của câu cũng cho thấy đây là tin thu hút sự chú ý, giống như các tin giả đã được huấn luyện trước đó. Dựa vào sự giống nhau về ngữ nghĩa và cấu trúc với các tin giả khác, mô hình có xu hướng dự đoán rằng tin mới này là "Giả".

Đối với các mô hình Transformer khi được kết hợp với TF-IDF như ở **Hình 10**, các bước tiền xử lý và mã hóa văn bản thành các đặc trưng dưới dạng input_ids và attention_mask vẫn được giữ nguyên như khi chỉ sử dụng các mô hình Transformer. Tuy nhiên TF-IDF sẽ bổ sung thêm một lớp đặc trưng dựa trên tần suất từ, giúp mô hình nhận diện rõ hơn những cụm từ thường xuyên xuất hiện trong tin giả. Điều này cho phép mô hình phân loại tin tức chính xác hơn, đặc biệt trong các trường hợp có nhiều yếu tố ngữ nghĩa giống nhau giữa tin thật và tin giả, nhưng sự khác biệt nằm ở tần suất xuất hiện của các từ quan trọng. Những từ trong các

cụm từ như "tin sốc", "hàng trăm ca lây nhiễm", “nguy hiểm chưa từng có” thường xuất hiện trong tin giả và sẽ có tần suất cao hơn trong các tài liệu thuộc lớp tin giả.



Hình 10. Ví dụ về quy trình Transformer dự đoán tin giả kết hợp cùng TF – IDF

Những đặc trưng của văn bản sẽ được kết hợp cùng với đặc trưng tần suất từ giúp cho mô hình không chỉ có khả năng hiểu ngữ nghĩa của câu và mà còn có khả năng phân tích tần suất xuất hiện của các từ quan trọng, và phân tích kết quả dự đoán cuối cùng.

CHƯƠNG 4: THỰC NGHIỆM

4.1 Dữ liệu thực nghiệm

Sau khi hoàn thành xử lý dữ liệu thông qua làm sạch và cân bằng dữ liệu như đã trình bày ở phần 3.2 và 3.3, chúng tôi đã thu được tập dữ liệu bao gồm các bài đăng trên mạng xã hội và tin tức từ các nguồn tiếng Việt. Tập dữ liệu chứa hơn 5.000 mẫu, bao gồm cả tin thật và tin giả từ nhiều chủ đề khác nhau. Sau đó, chúng tôi chia ngẫu nhiên tập dữ liệu thành tập huấn luyện và tập kiểm tra với tỷ lệ 80/20, kết quả thu được gồm tập dữ liệu dùng để huấn luyện có 4.000 mẫu và tập dùng để kiểm tra có 1.000 mẫu. Cách tiếp cận này cho phép mô hình học các mẫu một cách hiệu quả, tăng khả năng hoạt động tốt hơn trên dữ liệu mới, chưa được huấn luyện trước đó.

Tập dữ liệu được đạt tên là “Vietnamese News Dataset”và có thể được truy cập tại: https://github.com/huynhtuan0106/Vietnamese-News-Dataset-Version2

4.2 Công cụ đánh giá

Trước khi tìm hiểu về các thước đo và cách tính toán chúng, chúng ta cần hiểu rõ các khái niệm sau:

- **True Positive (TP):** là số lượng trường hợp trong tập kiểm tra mà được mô hình dự đoán thuộc lớp “Positive” và nhãn thực tế cũng thuộc lớp “Positive”.
- **True Negative (TN):** là số lượng trường hợp trong tập kiểm tra mà được mô hình dự đoán thuộc lớp “Negative” và nhãn thực tế cũng thuộc lớp “Negative”.
- **False Positive (FP):** là số lượng trường hợp trong tập kiểm tra mà được mô hình dự đoán thuộc lớp “Positive” nhưng nhãn thực tế lại thuộc lớp “Negative”.

- **False Negative (FN):** là số lượng trường hợp trong tập kiểm tra mà được mô hình dự đoán thuộc lớp "Negative" nhưng nhãn thực tế lại thuộc lớp "Positive".

Trong nghiên cứu này, với tin thật được gán nhãn 0 và tin giả được gán nhãn 1, cách xác định positive và negative sẽ như sau:

- **Positive class:** Tin giả (nhãn 1). Đây là lớp mà mô hình sẽ cố gắng phát hiện (dự đoán là dương tính).
- **Negative class:** Tin thật (nhãn 0). Đây là lớp mà mô hình dự đoán là không phải tin giả (dự đoán là âm tính).

Kết quả phân loại sẽ được đánh giá bằng các thước đo sau:

- **Accuracy (Độ chính xác):**

Accuracy đo lường tỷ lệ dự đoán đúng trong tổng số dự đoán, dễ hiểu hơn là tỷ lệ giữa số mẫu mô hình dự đoán đúng và tổng số mẫu trong tập dữ liệu kiểm tra, được tính như công thức (4.1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

Accuracy phản ánh hiệu suất tổng thể của mô hình, mặc dù có thể không phản ánh chính xác khi dữ liệu không cân bằng. Giả sử khi một tập dữ liệu có số positive lớn hơn rất nhiều so với negative, thì khi mô hình phân loại chỉ trả về tất cả dự đoán là positive cũng có thể đạt accuracy là cao.

- **Precision:**

Precision đo lường tỷ lệ số mẫu mô hình dự đoán chính xác là positive trong số những mẫu được mô hình dự đoán là positive (bao gồm cả đúng và sai), chỉ ra mức độ chính xác của các dự đoán positive, được tính như công thức (4.2).

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

Đây là chỉ số quan trọng khi cần giảm thiểu số lượng FP (False Postive), trong các trường hợp dự đoán nhầm là Postive sẽ gây tác động lớn. Ví dụ như dự đoán một thư là spam trong khi thư đó là không spam.

- **Recall (Khả năng truy hồi):**

Recall đo lường khả năng của mô hình trong việc tìm ra tất cả các giá trị positive thực sự. Nghĩa là tỷ lệ giữa số dự đoán positive đúng trên tổng số positve thực tế. Recall cao đồng nghĩa với việc bỏ sót các mẫu thực sự positive là thấp.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

Công thức (4.3) là công thức tính giá trị của Recall. Đây là chỉ số quan trọng khi cần giảm thiểu số lượng FN (False Negatives), trong các trường hợp dự đoán nhầm là Negative sẽ gây tác động lớn. Ví dụ trong nghiên cứu này, FN là trường hợp dự đoán một mẫu là tin thật trong khi đó là tin giả.

- **F1-Score:**

F1-Score là trung bình điều hòa giữa Precision và Recall, được tính như công thức (4.4). Chỉ số này giúp cân bằng hai thước đo này, đặc biệt hữu ích trong trường hợp dữ liệu không cân bằng. F1-Score càng cao tương ứng precision và recall càng cao, mô hình phân loại càng tốt.

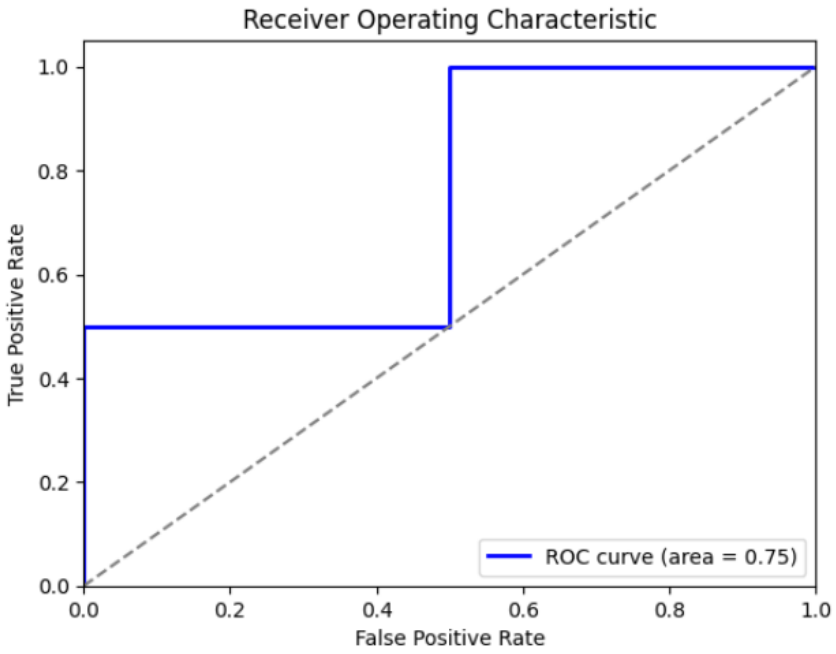
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.4}$$

- **AUC (Area Under the Curve):**

AUC đo lường khả năng phân biệt giữa các lớp của mô hình. Đại diện cho diện tích dưới đường cong ROC [24], một đồ thị thể hiện mối quan hệ giữa **True Positive Rate** (tức là Recall theo công thức 4.5) và **False Positive Rate** (tỷ lệ các trường hợp âm tính được dự đoán nhầm là dương tính theo công thức 4.6) trên các ngưỡng phân loại khác nhau.

$$True\ Positive\ Rate\ = \frac{TP}{TP\ +\ FN} \tag{4.5}$$

$$False\ Positive\ Rate\ = \frac{FP}{FP\ +\ TN} \tag{4.6}$$



Hình 11. Minh họa đường Receiver Operating Characteristic

Hình 11 minh họa đường Receiver Operating Characteristic (ROC) cùng với diện tích dưới. Trong đồ thị này, trục hoành đại diện cho False Positive Rate (FPR), còn trục tung thể hiện True Positive Rate (TPR). AUC cao cho thấy mô hình có khả năng phân biệt tốt giữa các lớp dương tính và âm tính.

4.3 Kết quả thực nghiệm

Bảng 1 và Bảng 2 trình bày kết quả đánh giá bằng các độ đo trên của các mô hình khi được huấn luyện trên cùng một tập dữ liệu huấn luyện và kiểm tra. Trong đó Bảng 1 thể hiện kết quả của các mô hình Transformer riêng biệt và Bảng 2 thể hiện kết quả của các mô hình Transformer khi kết hợp cùng với các kỹ thuật nhúng từ (word embedding techniques) như TF-IDF hoặc Word2Vec.

Bảng 1. Kết quả đánh giá các mô hình Transformer

Model	Accuracy	Precision	Recall	F1-Score	AUC
BERT	0.898204	0.901235	0.890244	0.895706	0.940220
RoBERTa	0.842315	0.8199230	0.869919	0.844181	0.897880
ViBERT	0.916168	0.914634	0.918699	0.916662	0.965901
PhoBERT	**0.944112**	0.936000	0.951220	**0.943548**	0.980376
ViSoBERT	0.930140	0.930612	0.926829	0.928717	0.980982

Bảng 2. Kết quả đánh giá các mô hình Transformer kết hợp TF-IDF/Word2Vec

Model	Accuracy	Precision	Recall	F1-Score	AUC
PhoBERT + TF-IDF	0.942116	**0.949020**	0.937984	0.943470	0.975387
ViSoBERT + TF-IDF	0.916168	0.904762	0.926829	0.915663	0.972358
PhoBERT + Word2Vec	**0. 944112**	0.932540	**0.955285**	**0.943775**	**0.982704**
ViSoBERT + Word2Vec	0.934132	0.917647	0.951220	0.934132	0.971720

- **RoBERTa**: Đạt hiệu suất phân loại khá kém

Tuy RoBERTa là một mô hình khá phổ biến trong xử lý ngôn ngữ tự nhiên, tuy nhiên trong trường hợp phân loại tin giả tiếng Việt đã không đạt được kết quả như mong đợi. Với **độ chính xác (Accuracy) đạt 84.23% và AUC là 0.897**, thấp hơn tất cả các mô hình còn lại.

Các chỉ số còn lại của RoBERTa lần lượt là **Precision: 81.99%, Recall: 86.99%, F1-Score: 84.42%**, tất cả đều dưới 90%. Điều này cho thấy RoBERTa

không phù hợp để xử lý phân loại tin tức giả Tiếng Việt, đặc biệt là trong nghiên cứu này, khi các tin tức khá phức tạp. Nguyên nhân có thể là do RoBERTa được huấn luyện trên tập dữ liệu Tiếng Anh, và không thể áp dụng cho một ngôn ngữ có đặc điểm riêng, ngữ nghĩa phức tạp như là Tiếng Việt.

- **BERT**: Đạt hiệu suất tốt hơn so với RoBERTa

BERT có **độ chính xác (Accuracy) 89.82%** và **AUC là 0.940**, cho thấy mô hình này có khả năng phân biệt giữa các loại tin tức tốt hơn RoBERTa. **Precision của BERT đạt 90.12% và Recall đạt 89,02%**, cho thấy BERT có khả năng phát hiện nhiều hơn các trường hợp tin giả so với RoBERTa. Tuy nhiên đây vẫn là các giá trị thấp, mô hình không phát hiện hiệu quả các tin giả và tin thật, có thể gây ra nhiều sai sót nghiêm trọng.

Tổng quan cho thấy BERT mang lại hiệu suất ổn trong việc phân loại tin giả Tiếng Việt nhưng vẫn chưa phải là mô hình tốt trong các mô hình đã thử nghiệm và không phải là lựa chọn tốt cho đề tài này. Do BERT không được huấn luyện và nghiên cứu trên một tập dữ liệu Tiếng Việt.

- **ViBERT**: Đạt hiệu suất tốt hơn BERT và RoBERTa

ViBERT cho kết quả khá tốt với **Accuracy đạt 91.62%** và **Precision đạt 91.46%.** Hai giá trị này cho thấy mô hình có khả năng dự đoán chính xác số lượng lớn các tin giả so với BERT và RoBERTa, tuy nhiên nó không vượt qua được ViSoBERT hay PhoBERT.

Recall của ViBERT đạt 91.87%, mặc dù khá tốt nhưng vẫn thấp hơn một số mô hình khác, cho thấy ViBERT vẫn còn phân loại nhầm một số trường hợp tin giả là tin thật. Điều này khiến mô hình này kém hiệu quả hơn trong việc phát hiện toàn diện các trường hợp tin thật và giả.

Tuy nhiên với **F1-Score đạt 91.67% và AUC đạt 0.966**, đây vẫn có thể được xem là một mô hình mạnh mẽ, đạt độ chính xác và mang tính ổn định.

- **PhoBERT**: Đạt độ chính xác cao nhất trong các mô hình

PhoBERT đã đạt hiệu suất phân loại cao với **độ chính xác (Accuracy) là 94.41%**. Điều này cho thấy PhoBERT có khả năng phân loại tin thật và tin giả trong ngữ cảnh tiếng Việt một cách hiệu quả.

Với **Recall đạt 95,12%** (cao thứ 2 trong tất cả mô hình) và **Precision đạt 93,60%**. Chứng minh PhoBERT không chỉ có khả năng tốt trong việc nhận dạng chính xác các tin tức giả mà còn xác định đúng khá nhiều các tin tức là thật.

F1-Score đạt 94.35% và **AUC của mô hình đạt 0.980**. PhoBERT đạt được sự cân bằng xuất sắc giữa việc phân biệt tin giả và tin thật và có khả năng phân biệt giữa mạnh mẽ giữa hai loại thông tin.

PhoBERT đã cho thấy khả năng phát hiện tin giả tốt hơn khi vượt qua các mô hình được thiết kế riêng cho Tiếng Việt còn lại như ViBERT hay ViSoBERT.

- **PhoBERT + TF-IDF**: Đạt hiệu suất thấp hơn so với PhoBERT

Việc kết hợp PhoBERT với phương pháp TF-IDF đã không mang lại hiệu suất cao hơn khi chỉ sử dụng PhoBERT.

Với ***Accuracy* đạt 94.21%, F1-Score đạt 94.34%, Recall đạt 93.80%**, và chỉ số **AUC đạt 0.975** cho thấy đây là mô hình có độ chính xác, sự cân bằng cũng như khả năng phân biệt giữa hai loại tin tức khá tương đồng với khi chỉ dùng PhoBERT, tuy nhiên có phần kém hơn một chút.

Tuy nhiên, mô hình này đạt được kết quả **Precision là 94.90%** - là giá trị cao nhất trong tất cả mô hình. Kết quả này cho thấy PhoBERT + TF-IDF ít gán nhầm tin thật thành tin giả, nhưng lại có nhiều trường hợp gán nhầm tin giả thành tin thật hơn khi chỉ dùng PhoBERT, điều này mang lại rủi ro nhiều hơn vì khi một tin giả được nhận diện nhầm thành tin thật sẽ có thể mang lại nhiều tác động tiêu cực hơn so với trường hợp ngược lại.

Vì vậy đây là mô hình phù hợp cho các hệ thống cần tối ưu Precision, ưu tiên việc giảm thiểu tối đa các sai sót về dự đoán tin thật là tin giả, nhưng không tối ưu cho trường hợp ngược lại.

- **PhoBERT + Word2Vec**: Đạt hiệu suất cao nhất cùng với PhoBERT

PhoBERT đã đạt hiệu suất phân loại cao nhất tương tự như PhoBERT với **Accuracy là 94.41%**. Điều này cho thấy khi kết hợp cùng Word2Vec đã không làm giảm có khả năng phân loại tin thật và tin giả trong ngữ cảnh tiếng Việt của PhoBERT.

Với **Recall đạt 95,52%, F1-Score đạt 94.38%** và **AUC của mô hình đạt 0.983** (đạt 3 giá trị cao nhất trong tất cả mô hình) và **Precision đạt 93,25%**. Việc dẫn đầu ở các chỉ số này và vượt qua cả PhoBERT đã chứng minh việc kết hợp PhoBERT và Word2Vec không chỉ tăng khả năng phân biệt mà còn duy trì sự cân bằng giữa tin tức thật và giả.

Mô hình này phù hợp cho các hệ thống yêu cầu độ chính xác cao và khả năng phát hiện các tin giả chính xác. Với Recall cao nhất, mô hình này đảm bảo rằng rất ít tin giả bị bỏ sót, giảm thiểu rủi ro cho xã hội.

- **ViSoBERT:** Mô hình đạt hiệu suất tốt

ViSoBERT cho khả năng phân loại tin giả ở mức tốt với **độ chính xác (Accuracy) đạt 93.01%.**

Các chỉ số **Precision: 93.06%**, **Recall: 92.68%**, và **F1-Score: 92.87%** đều ở mức ổn định và tương đối gần bằng nhau, tuy nhiên vẫn thấp hơn PhoBERT và PhoBERT + TF-IDF/Word2Vec. Điều này có nghĩa là trong một số trường hợp, ViSoBERT có thể nhầm lẫn giữa tin tức thật và giả, chưa tối ưu hóa tốt như PhoBERT.

Vì vậy trong các trường hợp tin tức phức tạp, yêu cầu cao về độ chính xác và khả năng phát hiện, PhoBERT vẫn là lựa chọn tốt hơn.

- **ViSoBERT + TF-IDF:** Đạt hiệu suất chưa tốt bằng ViSoBERT

ViSoBERT khi kết hợp cùng TF-IDF mang lại kết quả không tốt bằng khi chỉ sử dụng ViSoBERT thông thường.

Các chỉ số **Accuracy: 91.62%, Precision: 90.48%**, **Recall: 92.68%**, và **F1-Score: 91.57%** nếu xét riêng đây vẫn là một mô hình mang lại kết quả ổn định, nhưng so với tổng thể ViSoBERT + TF-IDF thấp hơn các mô hình thiết kế dành cho Tiếng Việt còn lại như ViBERT, PhoBERT và ViSoBERT thông thường.

- **ViSoBERT + Word2Vec:** Đạt hiệu suất tốt hơn ViSoBERT

Ngược lại với ViSoBERT + TF-IDF. Khi kết hợp cùng Word2Vec, mô hình này đã mang lại kết quả tốt hơn khi chỉ sử dụng ViSoBERT thông thường, tuy nhiên vẫn không vượt qua được PhoBERT.

Chỉ số **Accuracy đạt 93.41%** cao hơn so với ViSoBERT, tuy nhiên **Precision giảm còn 91.76%** cho thấy dù tăng độ chính xác tuy nhiên đã có nhiều trường hợp tin thật được phân loại là giả hơn.

Ngược lại mô hình này đạt **Recall: 95.12%**, và **F1-Score: 93.41%** cao hơn so với cả ViSoBERT và ViSoBERT + TF-ID. Nếu xét riêng đây vẫn là một mô hình mang lại kết quả ổn định, và đặc biệt Recall có giá trị cao (bằng với PhoBERT) cho thấy mô hình phát hiện được nhiều tin giả hơn và có ít khả năng gán nhầm một tin giả là tin thật.

4.4 Nhận xét

Từ kết quả đánh giá trên, có thể thấy rằng mỗi mô hình có những điểm mạnh và yếu riêng. Trong khi RoBERTa và BERT thể hiện khả năng phân loại tin giả ở mức độ tương đối; ViBERT và ViSoBERT đạt được hiệu suất phân loại cao hơn do được thiết kế dựa trên những đặc thù của ngôn ngữ Tiếng Việt; và PhoBERT đã dẫn đầu về hiệu suất phân loại, giúp phát hiện tin giả một cách chính xác hơn trong ngữ cảnh tiếng Việt.

Tuy nhiên có một điều cần lưu ý là khi tích hợp **TF-IDF** vào PhoBERT và ViSoBERT, kết quả **Accuracy** và **Recall** có xu hướng giảm so với khi sử dụng riêng các mô hình này. Nguyên nhân có thể đến từ một số lý do sau:

- **Thông tin ngữ cảnh bị mất:**

PhoBERT và **ViSoBERT** là các mô hình **Transformer** dựa trên kiến trúc **BERT**, chúng được huấn luyện để hiểu ngữ cảnh của từ ngữ trong câu, tức là chúng có khả năng nắm bắt được mối quan hệ ngữ nghĩa giữa các từ nhờ vào cơ chế **self-attention**. Điều này giúp chúng hiểu rõ hơn về nội dung và ý nghĩa của các câu trong văn bản.

Khi kết hợp với **TF-IDF**, một phương pháp thống kê truyền thống chỉ dựa trên tần suất xuất hiện của các từ mà không xem xét ngữ cảnh của chúng trong câu, mô hình có thể bị mất đi một phần khả năng hiểu ngữ cảnh sâu hơn. Điều này làm suy giảm hiệu suất của mô hình vì TF-IDF chỉ chú trọng đến việc tìm ra các từ có tần suất xuất hiện quan trọng, nhưng lại không phản ánh được mối quan hệ ngữ nghĩa giữa chúng.

- **Không phù hợp trong nghiên cứu này:**

TF-IDF hoạt động tốt khi các từ khóa riêng lẻ có thể mang nhiều ý nghĩa và đóng vai trò quan trọng trong phân loại, ví dụ trong các văn bản đơn giản, tin tức ngắn hoặc các bài toán phân loại truyền thống. Trong những trường hợp như vậy, TF-IDF có thể hữu ích vì nó giúp tách biệt các từ khóa quan trọng khỏi các từ phổ biến không quan trọng, do đó có thể hỗ trợ mô hình Transformer trong việc phân loại.

Tuy nhiên, trong nghiên cứu của chúng tôi, các tin tức đa phần là các văn bản phức tạp và dài, ngữ cảnh và cách các từ liên kết với nhau lại đóng vai trò quan trọng hơn so với tần suất từ xuất hiện. Lúc này việc dựa vào tần suất từ khóa (TF-IDF) thường không đủ để nắm bắt ngữ nghĩa thực sự. PhoBERT và ViSoBERT đã

được thiết kế để hiểu ngữ cảnh, vì vậy khi thêm TF-IDF, nếu không được xử lý cẩn thận, có thể làm giảm khả năng của mô hình trong việc tận dụng ngữ nghĩa toàn cục.

Khi kết hợp TF-IDF, mô hình có thể bị dựa vào các đặc trưng về tần suất từ, khiến khả năng phân biệt dựa trên ngữ cảnh bị giảm sút. Điều này dẫn đến mô hình khó khăn hơn trong việc phát hiện đúng tin giả, nhất là những tin có nội dung dài và phức tạp.

Ngược lại khi được kết hợp cùng **Word2Vec**, các mô hình Transformer đã không bị giảm đi độ chính xác của mình. Cụ thể là PhoBERT + Word2Vec vẫn duy trì độ chính xác tương đương PhoBERT, trong khi đó ViSoBERT + Word2Vec đã tăng độ chính xác hơn một chút so với ViSoBERT.

Các chỉ số Recall và F1-Score cũng tăng thêm, tuy nhiên Precision của cả hai mô hình đã giảm đi so với PhoBERT và ViSoBERT riêng biệt, điều này cho thấy rằng mô hình đã cẩn thận hơn và phân loại được nhiều trường hợp là tin giả hơn, nhưng cũng đã làm tăng số lượng các tin thật bị dự đoán là giả.

Word2Vec đã góp phần làm tăng độ chính xác của các mô hình thông qua việc biểu diễn từ ngữ dựa trên ngữ nghĩa và ngữ cảnh gần nhất của chúng, qua đó hỗ trợ Transformer trong việc nắm bắt ngữ nghĩa sâu hơn mà không mất đi thông tin ngữ cảnh quan trọng như khi sử dụng TF-IDF.

4.5 Thảo luận

Kết quả thử nghiệm cho thấy PhoBERT, PhoBERT + TF-IDF và ViSoBERT là các mô hình mang lại hiệu quả cao trong việc phân tích và phân loại tin giả. Trong đó PhoBERT đã đạt được hiệu suất tốt nhất và cách xa các mô hình còn lại với điểm số Accuracy, Recall và F1-Score cao nhất, thể hiện khả năng phân loại xuất sắc của nó. Trong khi đó PhoBERT + TF-IDF đạt điểm Precision cao nhất, và độ chính xác cũng gần như bằng PhoBERT. ViSoBERT cũng thể hiện rất tốt với điểm số Accuracy, Precision, Recall và F1-Score cao.

Bảng 3. So sánh kết quả dự đoán của các mô hình trên tập dữ liệu kiểm tra

	Tin tức/ Bài viết	Nhãn thực tế	Nhãn dự đoán của mô hình								
			RoBERTa	BERT	ViBERT	ViSoBERT	PhoBERT	PhoBERT + TF-IDF	ViSoBERT + TF-IDF	PhoBERT + Word2Vec	ViSoBERT + Word2Vec
1	Hà Nội gặp khó khăn di dời người dân ra khỏi vùng lũ.	**Real**	Real	Real	Real	Real	Real	Real	Real	Real	Real
2	Hiện trường kinh hoàng xe tải cố vượt đường ray khiến tàu hỏa trật bánh, ít nhất 100 người thương vong, hành khách hoảng loạn.	**Fake**	Fake	Fake	Fake	Fake	Fake	Fake	Fake	Fake	Fake
3	Ngày 29/7, Ban tổ chức Thế vận hội Olympic 2024 đã hủy buổi tập bơi 3 môn phối hợp do chất lượng nước sông Seine không đạt yêu cầu, sau lần hủy tương tự vào ngày 28/7.	**Real**	Fake	Fake	Real	Real	Real	Real	Real	Real	Real
4	Sáng nay ngày 07/10/2024, Hà Nội là thành phố ô nhiễm không khí nhất thế giới.	**Real**	Fake	Fake	Real	Real	Real	Real	Real	Real	Real

5	Tai nạn sập hầm lò đặc biệt nghiêm trọng ở Quảng Ninh khiến 5 công nhân tử vong,... đảng bộ và công đoàn bù nhìn chưa bao giờ lo cho điều kiện lao động của người dân.	**Fake**	Real	Real	Fake	Fake	Fake	Fake	Fake	Fake	Fake
6	Chiều 31/7, trong lúc làm việc, nhiều nhân viên một ngân hàng ở TP Biên Hòa có dấu hiệu mệt mỏi, khó thở, tim đập nhanh... đã được đưa đến Bệnh viện. Nguyên nhân ban đầu có thể do ngộ độc khí CO.	**Real**	Fake	Fake	Fake	Real	Real	Real	Fake	Real	Real
7	Nữ nhân viên làm việc tại Samsung lây nhiễm HIV cho 16 người.	**Fake**	Real	Real	Fake	Real	Fake	Real	Real	Fake	Real
8	Tin mới: Máy bay Vietnam Airlines phải hạ cánh khẩn cấp vì hành khách đánh nhau, đập vỡ kính cửa sổ.	**Fake**	Real	Real	Real	Real	Fake	Real	Real	Fake	Fake

9	Tin nóng tại Bình Thuận: Tai nạn xe khách nghiêm trọng làm tài xế tử vong, 11 người nhập viện cấp cứu.	**Real**	Real	Real	Real	Real	Real	Fake	Fake	Real	Real
10	Uống nước ngọt trước cổng trường, hàng loạt học sinh tại các trường Hà Nội nhập viện.	**Fake**	Real	Real	Real	Real	Real	Real	Real	Real	Real

Trong hầu hết các trường hợp tin thật và giả đơn giản, tất cả các mô hình đều cho kết quả dự đoán chính xác như nhau. Ví dụ như trường hợp thứ nhất: ***"Hà Nội gặp khó khăn khi di dời người dân ra khỏi vùng lũ"***, đây là tin thật và đều được tất cả mô hình cho kết quả dự đoán đúng. Tương tự với những tin giả có thể dễ dàng nhận thấy như trường hợp 2: ***"Hiện trường kinh hoàng xe tải cố vượt đường ray khiến tàu hỏa trật bánh, ít nhất 100 người thương vong, hành khách hoảng loạn"***, tin này dựa vào một tai nạn có thật, nhưng đã đưa ra thông tin sai về hậu quả, số người thương vong, làm phóng đại sự thật, tin này cũng được tất cả các mô hình dự đoán chính xác là giả.

Tuy nhiên, đối với một số tin tức phức tạp hơn như trường hợp 3: ***"Ngày 29/7, Ban tổ chức Thế vận hội Olympic 2024 đã hủy buổi tập bơi 3 môn phối hợp do chất lượng nước sông Seine không đạt yêu cầu, sau lần hủy tương tự vào ngày 28/7"*** và trường hợp 4: ***"Sáng nay ngày 07/10/2024, Hà Nội là thành phố ô nhiễm không khí nhất thế giới"*** đều là hai tin thật. Nhưng chỉ có các mô hình được phát triển dành riêng cho Tiếng Việt như ViBERT, ViSoBERT, PhoBERT và các phiên bản kết hợp cùng TF-IDF, Word2Vec cho ra dự đoán đúng, trong khi đó BERT và RoBERTa đã cho ra kết quả ngược lại.

Đặc biệt hơn, đối với một số thông tin giả được được viết theo kiểu nửa thật nửa giả, hoặc điều hướng dư luận để gây hiểu lầm, các mô hình được nghiên cứu dành riêng cho Tiếng Việt đã thể hiện khả năng hiểu ngữ nghĩa của câu một cách sâu sắc hơn và cho ra kết quả chính xác. Ví dụ, tin tức số 5: ***"Tai nạn sập hầm lò đặc biệt nghiêm trọng ở Quảng Ninh khiến 5 công nhân tử vong,... đảng bộ và công đoàn bù nhìn chưa bao giờ lo cho điều kiện lao động của người dân"*** chứa phần tin thật là "Tai nạn sập hầm lò ở Quảng Ninh khiến 5 công nhân tử vong," nhưng phần bổ sung "đảng bộ và công đoàn bù nhìn chưa bao giờ lo cho điều kiện lao động của người dân" là không chính xác và chưa được kiểm chứng, đây là một thông tin được thêm vào nhằm làm rối loạn xã hội. Trong tình huống này, ViBERT, ViSoBERT, PhoBERT đã phân loại chính xác đây là tin giả, trong khi BERT và RoBERTa lại bị đánh lừa bởi phần tin thật trong bài viết. Điều này cho thấy BERT và RoBERTa không được tối ưu hóa cho ngôn ngữ và ngữ cảnh tiếng Việt, dẫn đến khả năng phân loại sai lầm trong những thông tin như thế này.

Bên cạnh đó các mô hình được thiết kế dành riêng cho Tiếng Việt cũng có những trường hợp phân loại sai. Ví dụ với tin tức số 6: ***"Chiều 31/7, trong lúc làm việc, nhiều nhân viên một ngân hàng ở TP Biên Hòa có dấu hiệu mệt mỏi, khó thở, tim đập nhanh... đã được đưa đến Bệnh viện. Nguyên nhân ban đầu có thể do ngộ độc khí CO"***, đây là thông tin thật liên quan đến một sự cố sức khỏe nghiêm trọng tại nơi làm việc. Và kết quả là ngoài hai mô hình là BERT, RoBERTa dự đoán sai thì ViBERT và ViSoBERT + TF-IDF cũng dự đoán sai đây là một thông tin giả. Nguyên nhân có thể là vì việc đây là tình huống hiếm gặp và thiếu sự rõ ràng.

Mặc khác cũng có những trường hợp có rất ít mô hình cho kết quả dự đoán đúng và đặc biệt là các mô hình có kết hợp với Word2Vec đã dự đoán chính xác hơn các mô hình còn lại. Ở tin tức số 7: ***"Nữ nhân viên làm việc tại Samsung lây nhiễm HIV cho 16 người"***, đây là một thông tin giả được lan truyền rộng rãi và đã được đính chính lại bằng nhiều thông tin chính thống khác. Tuy nhiên chỉ có ViBERT, PhoBERT và PhoBERT + Word2Vec nhận diện đúng đây là tin giả, các

mô hình khác như BERT, RoBERTa và kể cả ViSoBERT, ViSoBERT + TF-IDF/Word2Vec, PhoBERT + TF-IDF đã nhận diện sai rằng đây là một tin thật. Một trường hợp khác nằm ở tin tức số 8: ***"Tin mới: Máy bay Vietnam Airlines phải hạ cánh khẩn cấp vì hành khách đánh nhau, đập vỡ kính cửa sổ"***, đây cũng là một thông tin sai được lan truyền trên mạng xã hội Việt Nam, sự thật là việc này diễn ra tại một quốc gia khác và của một hãng hàng không khác. Tuy nhiên chỉ có PhoBERT, PhoBERT + Word2Vec và ViSoBERT + Word2Vec nhận diện đúng đây là tin giả, trong khi các mô hình còn lại cho rằng đây là thông tin thật.

Những trường hợp trên cho thấy, mặc dù các mô hình ngôn ngữ dành riêng cho tiếng Việt như ViBERT, PhoBERT, và ViSoBERT có nhiều ưu điểm, nhưng chúng vẫn tồn tại những điểm yếu cần khắc phục, đặc biệt khi đối mặt với các tin tức có tính chất thật giả lẫn lộn.

Một số trường hợp khác, các mô hình riêng lẽ nhận diện đúng nhưng khi kết hợp cùng TF-IDF lại dẫn đến kết quả dự đoán sai. Ví dụ đối với tin tức số 9: ***"Tin nóng tại Bình Thuận: Tai nạn xe khách nghiêm trọng làm tài xế tử vong, 11 người nhập viện cấp cứu"***, đây là một tin thật. Tất cả các mô hình đã phân loại đúng, ngoài các mô hình PhoBERT và ViSoBERT kết hợp cùng với TF-IDF, cả hai mô hình này lại cho kết quả phân loại đây là tin giả. Nguyên nhân có thể do TF-IDF đã nhận diện một số cụm từ khóa thường xuất hiện trong tin giả từ các tin tức trong tập huấn luyện như những cụm từ ***"tin nóng"*** và ***"nghiêm trọng"*** nên đã dẫn đến kết quả phân loại sai. Những từ này dù có mặt trong nhiều tin tức giả nhưng trong các trường hợp khẩn cấp hoặc thật sự nghiêm trọng cũng xuất hiện trong các tin thật. Việc dựa trên tần suất xuất hiện của các từ thông qua TF-IDF đã ảnh hưởng đến khả năng hiểu ngữ cảnh của các mô hình và dẫn đến kết quả dự đoán đây là tin giả, dù đây là một tin tức thật.

Cuối cùng, cũng có những trường hợp các mô hình đều đưa ra kết quả dự đoán sai. Cụ thể là thông tin số 10: ***"Uống nước ngọt trước cổng trường, hàng loạt học sinh tại các trường Hà Nội nhập viện"***. Đây là một thông tin sai, trên thực tế

chỉ có 13 học sinh tại một trường tại Hà Nội gặp phải trường hợp trên, nhưng người đăng do chưa kiểm chứng cẩn thận đã đưa sai một số thông tin. Tất cả các mô hình đều dự đoán đây là tin thật. Nguyên nhân có thể là do tin tức này được trình bày với cấu trúc giống như các tin tức thật đã được mô hình học trước đó. Điều này cho thấy các mô hình Transformer vẫn có thể bị đánh lừa bởi những tin tức có ngữ nghĩa hợp lý nhưng lại chứa thông tin không chính xác.

Từ kết quả đánh giá và một số ví dụ trong tập dữ liệu kiểm tra trên, có thể thấy rằng các mô hình được phát triển dành riêng cho tiếng Việt như ViBERT, ViSoBERT, PhoBERT và các các mô hình được kết hợp cùng TF-IDF và Word2Vec có hiệu suất tốt hơn trong việc phát hiện tin giả so với BERT và RoBERTa. Tuy nhiên, vẫn còn các trường hợp các tin tức được đăng tải nhằm mục đích cố tình gây sự hiểu lầm, thu hút sự chú ý, các tin tức này thường được viết theo cấu trúc nửa thật nửa giả để làm nhiễu loạn thông tin, lợi dụng sự lo lắng và thói quen chia sẻ làm tin tức lan truyền nhanh chóng của người Việt Nam. Các mô hình vẫn còn gặp khó khăn đối với các tin tức này và vẫn có trường hợp tất cả đều dự đoán sai. Điều này nhấn mạnh tầm quan trọng của việc liên tục cải tiến các mô hình để tăng độ chính xác và khả năng nhận diện tin tức chính xác hơn trong tương lai.

Nhìn chung, mặc dù các nghiên cứu đã đạt được những thành công nhất định trong việc phân loại tin tức thật và giả, tuy nhiên vẫn cần phải nỗ lực không ngừng để nâng cao độ chính xác và khả năng hiểu ngữ nghĩa của các mô hình đối với ngôn ngữ con người, đặc biệt là đối với Tiếng Việt. Từ đó giúp cải thiện khả năng nhận diện tin tức giả, đặc biệt là trong bối cảnh các thông tin, tin tức được lan truyền trên mạng xã hội rất dễ dàng như hiện nay.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết quả đạt được

Trong nghiên cứu này, chúng tôi tập trung vào việc sử dụng các mô hình Transformer tiên tiến như BERT và RoBERTa, cùng với các phiên bản được huấn luyện riêng cho tiếng Việt như ViBERT, ViSoBERT, và PhoBERT, nhằm giải quyết bài toán phân loại tin giả tại Việt Nam.

Bên cạnh đó, chúng tôi cũng kết hợp cùng với phương pháp tiền xử lý dữ liệu như TF-IDF và Word2Vec. TF-IDF là một kỹ thuật lọc từ khóa quan trọng, giúp mô hình tập trung vào các từ ngữ có giá trị thông tin cao, đồng thời loại bỏ các từ phổ biến không có ý nghĩa trong phân loại. Trong khi đó, Word2Vec tạo ra các vector ngữ nghĩa cho từ dựa trên ngữ cảnh của chúng trong văn bản, giúp mô hình hiểu rõ hơn về mối liên hệ ngữ nghĩa giữa các từ. Sự kết hợp này được kỳ vọng sẽ tăng cường hiệu quả xử lý, tối ưu hóa quá trình phân loại và cải thiện độ chính xác của mô hình.

Tuy nhiên, qua các thử nghiệm thực tế kết quả cho thấy rằng việc kết hợp TF-IDF với các mô hình Transformer như PhoBERT và ViSoBERT không mang lại sự cải thiện về độ chính xác so với khi chỉ sử dụng riêng lẻ các mô hình này. Điều này cho thấy rằng các mô hình PhoBERT và ViSoBERT đã có khả năng xử lý và phân tích ngữ nghĩa mạnh mẽ và việc bổ sung TF-IDF không mang lại hiệu quả.

Ngược lại, Word2Vec đã giúp các mô hình tăng độ chính xác và khả năng nhận diện tin giả nhờ việc bổ sung thêm thông tin ngữ cảnh cho các mô hình. Nghiên cứu của chúng tôi đã mang lại những kết quả rõ ràng trong việc kết hợp các phương pháp này cùng mô hình Transformer đối với bài toán phát hiện tin giả bằng tiếng Việt.

Để có dữ liệu cho quá trình nghiên cứu, chúng tôi đã thu thập một tập dữ liệu bao gồm các bài đăng trên Facebook từ tháng 6 đến tháng 10 năm 2024, bao gồm

các chủ đề về đời sống, xã hội và chính trị. Sau đó chúng tôi tiến hành xử lý dữ liệu và áp dụng các mô hình Transformer để phân loại.

Kết quả chúng tôi thu được cho thấy mô hình **PhoBERT** và **PhoBERT + Word2Vec** đã đạt hiệu suất phân loại dẫn đầu với độ chính xác là 94.41%. Tiếp theo sau đó là **PhoBERT + TF-IDF** và **ViSoBERT + Word2Vec** với độ chính xác lần lượt là 94.21% và 93.41%. **ViSoBERT** cũng đạt được độ chính xác cao là 93.01% nhưng khi kết hợp cùng với TF-IDF đã giảm độ chính xác chỉ còn 91.62%. Trong khi đó ViBERT có độ chính xác ở mức khá. Cuối cùng, BERT và RoBERTa không thể đạt hiệu suất cao như các mô hình được tối ưu hóa riêng cho tiếng Việt cho thấy sự hạn chế khi áp dụng trên ngôn ngữ này.

5.2 Những điểm hạn chế

Tuy nhiên, nghiên cứu của chúng tôi vẫn có một số điểm hạn chế. Một trong những thách thức chính mà chúng tôi gặp phải là vấn đề về dữ liệu. Mặc dù đã bổ sung dữ liệu từ tập VFND [22], lượng tin giả vẫn còn hạn chế và không đủ phong phú để mô hình có thể học tốt mọi trường hợp. Điều này là rất quan trọng trong việc huấn luyện các mô hình bởi vì tập dữ liệu càng đa dạng và phong phú thì khả năng nhận diện và phân loại thông tin càng chính xác.

Ngoài ra, do cấu trúc phức tạp của tiếng Việt, nhiều thông tin có thể bị mất mát trong quá trình xử lý dữ liệu. Ví dụ, các yếu tố như từ viết tắt, ngữ pháp được diễn giải theo văn nói hằng ngày, hay các bài viết có nội dung thật và giả lẫn lộn có thể gây khó khăn cho mô hình. Những yếu tố này có thể tác động đến khả năng học của mô hình và đưa ra những dự đoán sai, ảnh hưởng đến hiệu suất tổng thể.

Hơn nữa, việc xác định đầu vào và phân loại các tin tức, bài viết là thật và giả cần phải có nhiều kiến thức và sự hiểu biết về tình hình và bối cảnh thực tế tại Việt Nam. Một số tin tức hoặc thông tin có thể thay đổi tính chất từ thật sang giả hoặc ngược lại sau một khoảng thời gian, sau khi được điều tra và cập nhật, hoặc kiến thức đã được thay đổi,… Điều này cho thấy rằng cần phải liên tục cập nhật về

các sự kiện diễn ra trong xã hội để không ảnh hưởng đến khả năng phân loại của các mô hình.

5.3 Hướng phát triển

Hiện tại nghiên cứu của chúng tôi chỉ tập vào việc phân loại tin tức dựa trên nội dung của bài đăng mà chưa khai thác hết các trường dữ liệu khác, chẳng hạn như số lượng tương tác, phản hồi và bình luận từ người dùng trên các nền tảng mạng xã hội. Đây cũng là một trong những dữ liệu quan trọng có thể mang lại những dấu hiệu quan trọng để nhận diện tin tức giả của mô hình bởi vì số lượng lượt thích, lượt chia sẻ, hay nội dung bình luận thường phản ánh phần nào mức độ phổ biến hoặc mức độ gây tranh cãi của tin tức.

Cụ thể hơn, các bài viết chứa tin tức giả thường có xu hướng tác động mạnh đến cảm xúc của người đọc, dẫn đến số lượng bình luận hoặc chia sẻ tăng cao bất thường. Những bình luận này có thể chứa nhiều thông tin về cảm xúc và thái độ của người dùng, điều này rất hữu ích để phân loại và phân tích. Một bài viết tin giả có thể thu hút nhiều bình luận tiêu cực, hoặc gây ra sự tranh luận gay gắt giữa những người tin và không tin. Ngược lại, các bài viết chứa tin thật thường có xu hướng ít gây ra những phản ứng cực đoan hơn và có thể đi kèm với bình luận mang tính chất thảo luận, đồng tình hoặc bổ sung thêm thông tin.

Dữ liệu về tương tác và bình luận có thể cung cấp các góc nhìn khác về cách người dùng tiếp nhận và phản hồi thông tin, từ đó giúp tăng cường khả năng phát hiện tin tức giả mạo. Các bình luận thường chứa đựng những tín hiệu cảm xúc rõ ràng, chẳng hạn như sự phẫn nộ, hoang mang, hoặc nghi ngờ, vốn là những dấu hiệu thường gặp trong các bài viết chứa tin giả. Những dữ liệu này nếu được phân tích kỹ lưỡng có thể giúp hệ thống hiểu rõ hơn về cách mà người dùng tương tác với hai loại thông tin thật và giả, từ đó làm tăng độ chính xác của dự đoán.

Chính vì vậy trong tương lai chúng tôi sẽ tiếp tục mở rộng nghiên cứu bằng cách không chỉ tập trung vào nội dung bài viết mà còn xem xét kỹ lưỡng hơn dữ

liệu tương tác và bình luận. Chúng tôi dự định phát triển các kỹ thuật NLP để phân tích về cảm xúc và thái độ của người dùng thông qua các bình luận. Điều này sẽ giúp cải thiện đáng kể độ chính xác trong việc phân loại tin tức, bởi hệ thống không chỉ dựa vào nội dung văn bản gốc, mà còn khai thác được thêm các phản ứng của cộng đồng đối với thông tin đó. Việc kết hợp này sẽ tạo ra một cách tiếp cận đa chiều, mạnh mẽ hơn trong việc đối phó với vấn đề tin tức giả..

TÀI LIỆU THAM KHẢO

[1] K. Chowdhary and K. R. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[2] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Adv Neural Inf Process Syst*, vol. 34, pp. 15908–15919, 2021.

[3] T. O. Tran and P. Le Hong, "Improving sequence tagging for Vietnamese text using transformer-based neural models," in *Proceedings of the 34th Pacific Asia conference on language, information and computation*, 2020, pp. 13–20.

[4] Q.-N. Nguyen, T. C. Phan, D.-V. Nguyen, and K. Van Nguyen, "ViSoBERT: A pre-trained language model for Vietnamese social media text processing," *arXiv preprint arXiv:2310.11166*, 2023.

[5] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," *arXiv preprint arXiv:2003.00744*, 2020.

[6] A. Agarwal and P. Meel, "Stacked Bi-LSTM with attention and contextual BERT embeddings for fake news analysis," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 2021, pp. 233–237.

[7] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, 2019.

[8] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE international conference on data mining (ICDM)*, IEEE, 2019, pp. 518–527.

[9] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[10] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.

[13] K. Clark, “Electra: Pre-training text encoders as discriminators rather than generators,” *arXiv preprint arXiv:2003.10555*, 2020.

[14] T. N. Hieu, H. C. N. Minh, H. T. Van, and B. V. Quoc, “ReINTEL Challenge 2020: Vietnamese Fake News Detection usingEnsemble Model with PhoBERT embeddings,” in *Proceedings of the 7th international workshop on Vietnamese language and speech processing*, 2020, pp. 1–5.

[15] N.-D. Pham, T.-H. Le, T.-D. Do, T.-T. Vuong, T.-H. Vuong, and Q.-T. Ha, “Vietnamese fake news detection based on hybrid transfer learning model and TF-IDF,” in *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, 2021, pp. 1–6.

[16] C.-V. Nguyen Thi, T.-T. Vuong, D.-T. Le, and Q.-T. Ha, “v3mfnd: A deep multi-domain multimodal fake news detection model for Vietnamese,” in *Asian Conference on Intelligent Information and Database Systems*, Springer, 2022, pp. 608–620.

[17] T. H. Vo, T. L. T. Phan, and K. C. Ninh, “DEVELOPMENT OF A FAKE NEWS DETECTION TOOL FOR VIETNAMESE BASED ON DEEP LEARNING TECHNIQUES.,” *Eastern-European Journal of Enterprise Technologies*, vol. 119, no. 2, 2022.

[18] K. D. Pham, D. Van Thin, and N. L.-T. Nguyen, “Improving Vietnamese Fake News Detection based on Contextual Language Model and Handcrafted Features,” *Science and Technology Development Journal*, vol. 26, no. 2, pp. 2705–2712, 2023.

[19] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, “The effect of the TF-IDF algorithm in times series in forecasting word on social media,” *Indones. J. Electr. Eng. Comput. Sci*, vol. 22, no. 2, p. 976, 2021.

[20] B. Trstenjak, S. Mikac, and D. Donko, “KNN with TF-IDF based framework for text categorization,” *Procedia Eng*, vol. 69, pp. 1356–1364, 2014.

[21] Wikipedia, “Transformer (deep learning architecture),”. [Online]. Available: https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture). [Truy cập: 22 Tháng 9, 2024].

[22] Ho Quang Thanh and ninh-pm-se, “thanhhocse96/vfnd-vietnamese-fake-news-datasets: Tập hợp các bài báo tiếng Việt và các bài post Facebook phân loại 2 nhãn Thật & Giả (228 bài),” Feb. 2019. [Online]. Available: https://github.com/thanhhocse96/vfnd-vietnamese-fake-news-datasets. [Truy cập: 8 Tháng 8, 2024].

[23] A. Moreo, A. Esuli, and F. Sebastiani, “Distributional random oversampling for imbalanced text classification,” in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 805–808.

[24] J. Myerson, L. Green, and M. Warusawitharana, “Area under the curve as a measure of discounting,” *J Exp Anal Behav*, vol. 76, no. 2, pp. 235–243, 2001.