**1. Introduction**

Trong bối cảnh hiện đại hóa toàn cầu, các nền tảng mạng xã hội đang ngày càng trở nên phổ biến, đi kèm với đó là những tác động tích cực và tiêu cực đáng kể. Trong đó, sự lan truyền nhanh chóng của tin tức giả trên mạng xã hội đã nổi lên như một vấn đề xã hội nghiêm trọng, khi những thông tin sai lệch được phát tán gây ra không ít hiểu lầm, thậm chí có thể dẫn đến những xung đột trên toàn cầu.

Tại Việt Nam, trong những năm gần đây, nhiều tin tức giả đã gây xôn xao dư luận, điển hình như các tin liên quan đến dịch bệnh, tai nạn giao thông, những thông tin sai lệch trong đời sống, hay thậm chí cả những thông tin mang tính chất phản động về chính trị. Những loại tin tức giả này thường lan truyền nhanh chóng trong cộng đồng, gây hoang mang dư luận và ảnh hưởng đến đời sống của người dân. Vì vậy, việc nghiên cứu và phát hiện tin tức giả là một nhiệm vụ cần thiết nhằm hỗ trợ và duy trì sự ổn định của xã hội. Chính vì lẽ đó, chúng tôi đã chọn chủ đề này làm đối tượng nghiên cứu của mình.

Trong những năm gần đây, deep learning 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 deep learning truyền thống thường dựa vào xử lý dữ liệu tuần tự, 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, cho phép xử lý ngữ cảnh và các mối quan hệ trong văn bản một cách hiệu quả hơn. Những lợi thế này giúp Transformers hoạt động tốt hơn trong việc hiểu rõ ngôn ngữ và ngữ cảnh của văn bản, từ đó giải quyết tốt các nhiệm vụ phân loại văn bản ví dụ như phát hiện tin tức giả.

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 là sử dụng PhoBERT - một biến thể được thiết kế dành riêng cho ngôn ngữ tiếng Việt. Mục tiêu của chúng tôi là phát triển một hệ thống hiệu quả nhằm xác định tin tức giả mạo trên các nền tảng truyền thông xã hội, đặc biệt là Facebook - nền tảng mạng xã hội phổ biến nhất tại Việt Nam.

Bằng cách khai thác sức mạnh của PhoBERT, chúng tôi hy vọng sẽ cải thiện độ chính xác và hiệu quả trong việc phát hiện tin tức giả tiếng Việt. 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 đã tiến hành thu thập dữ liệu từ các trang Facebook chính thống của các cơ quan báo chí tại Việt Nam và các nguồn dữ liệu giả từ các trang mạo danh, các nguồn chống đối và các trang tin sai lệch, bao gồm nhiều lĩnh vực từ đời sống xã hội đến chính trị. Để thực hiện công việc này, chúng tôi đã sử dụng các công cụ khác nhau, bao gồm Selenium để thu thập dữ liệu, sau đó tiến hành xử lý dữ liệu thông qua các bước làm sạch và mã hóa.

Cấu trúc của phần còn lại của bài báo này được sắp xếp như sau: Phần 2 sẽ xem xét các công trình liên quan đến các mô hình Transformer và phát hiện tin tức giả, tập trung vào các phương pháp và mô hình áp dụng cho nghiên cứu tại Việt Nam. Tiếp theo, phần 3 sẽ trình bày chi tiết về phương pháp đề xuất, bao gồm mô hình tổng thể và các bước cụ thể để phát triển hệ thống. Phần 4 sẽ tập trung vào các thiết lập thử nghiệm, mô tả tập dữ liệu, kết quả và thảo luận. Cuối cùng, phần 5 sẽ tóm tắt những phát hiện của chúng tôi và đề xuất các hướng nghiên cứu trong tương lai.

1. **Background**

Để triển khai hiệu quả dự án 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. 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):** Phát hiện các tên riêng trong văn bản như tên người, địa điểm, hoặc tổ chức.

- 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 phân loại văn bản, NLP được sử dụng để 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.

NLP được sử dụng trong nhiều sản phẩm và dịch vụ hàng ngày. Một số cách phổ biến nhất mà NLP được sử dụng là thông qua trợ lý kỹ thuật số được kích hoạt bằng giọng nói trên điện thoại thông minh, các chương trình quét email được sử dụng để xác định thư rác và các ứng dụng dịch thuật giải mã các ngôn ngữ nước ngoài.

Hệ thống xử lý ngôn ngữ tự nhiên có thể hoạt động nhanh chóng và hiệu quả: sau khi các mô hình NLP được đào tạo đúng cách, nó có thể đảm nhận các nhiệm vụ hành chính, giải phóng nhân viên để làm việc hiệu quả hơn. Các lợi ích có thể bao gồm:

Khám phá thông tin chi tiết nhanh hơn: Các tổ chức có thể tìm thấy các mô hình, xu hướng và mối quan hệ ẩn giữa các phần nội dung khác nhau. Việc truy xuất dữ liệu văn bản hỗ trợ thông tin chi tiết và phân tích sâu hơn, cho phép đưa ra quyết định sáng suốt hơn và đưa ra các ý tưởng kinh doanh mới.

Tiết kiệm ngân sách nhiều hơn: Với khối lượng lớn dữ liệu văn bản phi cấu trúc có sẵn, NLP có thể được sử dụng để tự động hóa việc thu thập, xử lý và tổ chức thông tin với ít nỗ lực thủ công hơn.

Truy cập nhanh vào dữ liệu của công ty: Một doanh nghiệp có thể xây dựng cơ sở kiến thức về thông tin tổ chức để có thể truy cập hiệu quả bằng tìm kiếm AI. Đối với đại diện bán hàng, NLP có thể giúp nhanh chóng trả về thông tin có liên quan, cải thiện dịch vụ khách hàng và giúp chốt doanh số.

NLP bao gồm nhiều kỹ thuật phân tích ngôn ngữ của con người. Một số kỹ thuật phổ biến nhất mà bạn có thể gặp trong lĩnh vực này bao gồm:

Phân tích tình cảm: Một kỹ thuật NLP phân tích văn bản để xác định tình cảm của nó, chẳng hạn như "tích cực", "tiêu cực" hoặc "trung lập". Phân tích tình cảm thường được các doanh nghiệp sử dụng để hiểu rõ hơn phản hồi của khách hàng.

Tóm tắt: Một kỹ thuật NLP tóm tắt một văn bản dài hơn, để giúp người đọc có thời gian dễ hiểu hơn. Một số văn bản phổ biến được tóm tắt bao gồm báo cáo và bài viết.

Trích xuất từ khóa: Một kỹ thuật NLP phân tích văn bản để xác định các từ khóa hoặc cụm từ quan trọng nhất. Trích xuất từ ​​khóa thường được sử dụng cho mục đích tối ưu hóa công cụ tìm kiếm (SEO), giám sát phương tiện truyền thông xã hội và thông tin kinh doanh.

Phân tích mã thông báo: Quá trình chia nhỏ các ký tự, từ hoặc từ phụ thành "mã thông báo" có thể được phân tích bởi một chương trình. Phân tích mã thông báo hỗ trợ các tác vụ NLP phổ biến như mô hình hóa từ, xây dựng vốn từ vựng và tần suất xuất hiện của từ.

NLP kết hợp sức mạnh của ngôn ngữ học tính toán với các thuật toán học máy và học sâu. Ngôn ngữ học tính toán là một ngành ngôn ngữ học sử dụng khoa học dữ liệu để phân tích ngôn ngữ và lời nói. Nó bao gồm hai loại phân tích chính: phân tích cú pháp và phân tích ngữ nghĩa.

Phân tích ngữ nghĩa sử dụng đầu ra cú pháp để rút ra ý nghĩa từ các từ và diễn giải ý nghĩa của chúng trong cấu trúc câu.

Học tự giám sát (SSL) nói riêng rất hữu ích để hỗ trợ NLP vì NLP yêu cầu một lượng lớn dữ liệu được gắn nhãn để đào tạo các mô hình trí tuệ nhân tạo (AI) tiên tiến. Vì các tập dữ liệu được gắn nhãn này yêu cầu chú thích tốn thời gian—một quy trình liên quan đến việc gắn nhãn thủ công của con người—nên việc thu thập đủ dữ liệu có thể cực kỳ khó khăn. Các phương pháp tiếp cận tự giám sát có thể tiết kiệm thời gian và chi phí hơn vì chúng thay thế một số hoặc toàn bộ dữ liệu đào tạo được gắn nhãn thủ công.

Ba phương pháp tiếp cận khác nhau đối với NLP bao gồm:

NLP dựa trên quy tắc: Các ứng dụng NLP đầu tiên là các cây quyết định nếu-thì đơn giản, yêu cầu các quy tắc được lập trình sẵn. Chúng chỉ có thể cung cấp câu trả lời để phản hồi các lời nhắc cụ thể, chẳng hạn như phiên bản gốc của Moviefone. Vì không có khả năng học máy hoặc AI trong NLP dựa trên quy tắc, nên chức năng này bị hạn chế rất nhiều và không thể mở rộng quy mô.

NLP thống kê: Được phát triển sau này, NLP thống kê tự động trích xuất, phân loại và gắn nhãn các thành phần của dữ liệu văn bản và giọng nói, sau đó chỉ định một khả năng thống kê cho từng ý nghĩa có thể có của các thành phần đó. Điều này dựa vào máy học, cho phép phân tích ngôn ngữ học tinh vi như gắn thẻ từ loại.

NLP thống kê đã giới thiệu kỹ thuật thiết yếu là ánh xạ các thành phần ngôn ngữ—chẳng hạn như từ và quy tắc ngữ pháp—thành biểu diễn vectơ để ngôn ngữ có thể được mô hình hóa bằng cách sử dụng các phương pháp toán học (thống kê), bao gồm mô hình hồi quy hoặc mô hình Markov. Điều này đã thông báo cho các phát triển NLP ban đầu như trình kiểm tra chính tả và nhắn tin T9 (Văn bản trên 9 phím, được sử dụng trên điện thoại Touch-Tone).

NLP học sâu: Gần đây, các mô hình học sâu đã trở thành chế độ NLP thống trị, bằng cách sử dụng khối lượng lớn dữ liệu thô, không có cấu trúc—cả văn bản và giọng nói—để trở nên chính xác hơn bao giờ hết. Học sâu có thể được coi là sự phát triển hơn nữa của NLP thống kê, với sự khác biệt là nó sử dụng các mô hình mạng nơ-ron. Có một số tiểu loại mô hình:

Mô hình Sequence-to-Sequence (seq2seq): Dựa trên mạng nơ-ron hồi quy (RNN), chúng chủ yếu được sử dụng để dịch máy bằng cách chuyển đổi cụm từ từ một miền (như tiếng Đức) thành cụm từ của miền khác (như tiếng Anh).

Mô hình biến đổi: Chúng sử dụng mã hóa ngôn ngữ (vị trí của từng mã thông báo—từ hoặc từ phụ) và tự chú ý (ghi lại các mối quan hệ và phụ thuộc) để tính toán mối quan hệ giữa các phần ngôn ngữ khác nhau với nhau. Mô hình biến đổi có thể được đào tạo hiệu quả bằng cách sử dụng học tự giám sát trên cơ sở dữ liệu văn bản lớn. Một cột mốc trong mô hình biến đổi là biểu diễn bộ mã hóa hai chiều của Google từ bộ biến đổi (BERT), đã trở thành và vẫn là cơ sở cho cách thức hoạt động của công cụ tìm kiếm của Google.

Mô hình tự hồi quy: Loại mô hình biến đổi này được đào tạo cụ thể để dự đoán từ tiếp theo trong một chuỗi, thể hiện bước tiến lớn về khả năng tạo văn bản. Ví dụ về LLM tự hồi quy bao gồm GPT, Llama, Claude và Mistral nguồn mở.

Các mô hình nền tảng: Các mô hình nền tảng được xây dựng sẵn và tuyển chọn có thể đẩy nhanh quá trình triển khai nỗ lực NLP và tăng cường sự tin tưởng vào hoạt động của nó. Ví dụ, các mô hình nền tảng IBM Granite™ có thể áp dụng rộng rãi trong nhiều ngành. Chúng hỗ trợ các tác vụ NLP bao gồm tạo nội dung và trích xuất thông tin chi tiết. Ngoài ra, chúng tạo điều kiện cho việc tạo ra được tăng cường truy xuất, một khuôn khổ để cải thiện chất lượng phản hồi bằng cách liên kết mô hình với các nguồn kiến thức bên ngoài. Các mô hình cũng thực hiện nhận dạng thực thể được đặt tên bao gồm xác định và trích xuất thông tin chính trong văn bản.

Một số tác vụ NLP thường giúp xử lý dữ liệu giọng nói và văn bản của con người theo cách giúp máy tính hiểu được những gì nó đang tiếp nhận. Một số tác vụ này bao gồm:

Các tác vụ ngôn ngữ

Giải quyết đồng tham chiếu là nhiệm vụ xác định xem hai từ có đề cập đến cùng một thực thể hay không và khi nào. Ví dụ phổ biến nhất là xác định người hoặc vật mà một đại từ nhất định đề cập đến (chẳng hạn như "she" = "Mary"). Nhưng nó cũng có thể xác định một phép ẩn dụ hoặc thành ngữ trong văn bản (chẳng hạn như trường hợp "bear" không phải là một loài động vật, mà là một người to lớn và nhiều lông).

Nhận dạng thực thể có tên (NER) xác định các từ hoặc cụm từ là các thực thể hữu ích. NER xác định "London" là một địa điểm hoặc "Maria" là tên của một người.

Gắn thẻ từ loại, còn được gọi là gắn thẻ ngữ pháp, là quá trình xác định từ hoặc đoạn văn bản thuộc loại từ loại nào, dựa trên cách sử dụng và ngữ cảnh của từ hoặc đoạn văn đó. Ví dụ, từ loại xác định "make" là động từ trong "I can make a paper plane" và là danh từ trong "What make of car do you own?"

Giải nghĩa từ là lựa chọn nghĩa của một từ cho một từ có nhiều nghĩa có thể. Điều này sử dụng một quy trình phân tích ngữ nghĩa để xem xét từ trong ngữ cảnh. Ví dụ, giải nghĩa từ giúp phân biệt nghĩa của động từ "make" trong "make the grade" (to attain) so với "make a bet" (to place). Việc sắp xếp "I will be merry when I married Mary" đòi hỏi một hệ thống NLP tinh vi.

Nhiệm vụ hỗ trợ người dùng

Nhận dạng giọng nói, còn được gọi là chuyển giọng nói thành văn bản, là nhiệm vụ chuyển đổi dữ liệu giọng nói thành dữ liệu văn bản một cách đáng tin cậy. Nhận dạng giọng nói là một phần của bất kỳ ứng dụng nào tuân theo lệnh thoại hoặc trả lời các câu hỏi được nói. Điều khiến nhận dạng giọng nói trở nên đặc biệt khó khăn là cách mọi người nói - nhanh, chạy các từ với nhau, với sự nhấn mạnh và ngữ điệu khác nhau.

Tạo ngôn ngữ tự nhiên (NLG) có thể được mô tả là ngược lại với nhận dạng giọng nói hoặc chuyển giọng nói thành văn bản: NLG là nhiệm vụ đưa thông tin có cấu trúc vào ngôn ngữ giao tiếp của con người. Nếu không có NLG, máy tính sẽ khó có thể vượt qua bài kiểm tra Turing, trong đó máy tính cố gắng bắt chước cuộc trò chuyện của con người. Các tác nhân đàm thoại như Alexa của Amazon và Siri của Apple đã thực hiện tốt điều này và hỗ trợ khách hàng theo thời gian thực.

Hiểu ngôn ngữ tự nhiên (NLU) là một tập hợp con của NLP tập trung vào việc phân tích ý nghĩa đằng sau các câu. NLU cho phép phần mềm tìm ra ý nghĩa tương tự trong các câu khác nhau hoặc xử lý các từ có ý nghĩa khác nhau.

Phân tích tình cảm cố gắng trích xuất các phẩm chất chủ quan—thái độ, cảm xúc, mỉa mai, bối rối hoặc nghi ngờ—từ văn bản. Điều này thường được sử dụng để định tuyến thông tin liên lạc đến hệ thống hoặc người có nhiều khả năng đưa ra phản hồi tiếp theo nhất.

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

TF-IDF (Term Frequency-Inverse Document Frequency) là một kỹ thuật thường được sử dụng trong xử lý ngôn ngữ tự nhiên (NLP) và khai thác dữ liệu văn bản. Nó là một thước đo thống kê nhằm đánh giá tầm quan trọng của một từ đối với một tài liệu trong một tập hợp các tài liệu hoặc một bộ sưu tập văn bản (corpus).

**Trong nghiên cứu này, chúng tôi đã sử dụng TF-IDF như một bước tiền xử lý để biến đổi văn bản thành các vector đặc trưng. Những vector này sau đó có thể được kết hợp với các mô hình Transformer để cải thiện 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 ảnh hưởng của các từ phổ biến nhưng không mang nhiều thông tin trong quá trình huấn luyện mô hình.**

**2.2.2 Mô hình Transformer**

Mô hình Transformer đại diện cho một bước đột phá trong NLP, được giới thiệu bởi Vaswani và các cộng sự (2017) trong bài báo "Attention Is All You Need". Điểm nổi bật của Transformer nằm ở kiến trúc tự chú ý (self-attention), cho phép mô hình học các mối quan hệ giữa các từ trong câu mà không phụ thuộc vào thứ tự tuần tự như các mô hình trước đây như RNN hoặc LSTM.

Mô hình Transformer bao gồm hai thành phần chính: Encoder và Decoder:

**Encoder:** Encoder nhận một chuỗi từ đầu vào và biểu diễn chúng dưới dạng các vector ngữ nghĩa. Mỗi Encoder bao gồm nhiều lớp tuần tự, với hai thành phần chính trong mỗi lớp: cơ chế tự chú ý và mạng nơ-ron hồi tiếp. Cơ chế tự chú ý cho phép mô hình học các mối quan hệ ngữ nghĩa giữa các từ liên quan trong một chuỗi trong khi bỏ qua các từ không liên quan. Mạng nơ-ron hồi tiếp xử lý các vector đã chú ý này để tạo ra các biểu diễn ngữ nghĩa sâu hơn.

**Decoder:** Decoder có cấu trúc tương tự như Encoder, sử dụng tự chú ý cho đầu vào mục tiêu. Ngoài ra, nó còn sử dụng chú ý chéo (cross-attention) để kết nối với đầu ra của Encoder. Điều này cho phép Decoder tạo ra các biểu diễn ngữ nghĩa dựa trên cả chuỗi đầu vào ban đầu và chuỗi đầu ra đã được tạo ra trước đó.

Sự phối hợp giữa Encoder và Decoder cho phép Transformer xử lý các nhiệm vụ ngôn ngữ như dịch máy, tóm tắt văn bản, sinh văn bản và phân loại văn bản với sự linh hoạt và hiệu quả.

**2.2.3 BERT (Bidirectional Encoder Representations from Transformers)**

BERT là một mô hình ngôn ngữ tiền huấn luyện được thiết kế để hiểu ngữ cảnh từ cả hai hướng (từ trái sang phải và từ phải sang trái) trong một câu. BERT được đào tạo trên hai nhiệm vụ chính: Mô hình Ngôn ngữ Ẩn (Masked Language Modeling - MLM) và Dự đoán Câu Kế Tiếp (Next Sentence Prediction - NSP).

**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

**Dự đoán Câu Kế Tiếp (NSP):** Nhiệm vụ này yêu cầu mô hình dự đoán xem một câu cho trước có phải là câu kế tiếp của câu trước đó không, nhằm cải thiện khả năng của mô hình trong việc hiểu các mối quan hệ giữa các câu.

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.

**2.2.4 RoBERTa (A Robustly Optimized BERT Pretraining Approach)**

RoBERTa là một biến thể của BERT được tối ưu hóa để cải thiện hiệu suất bằng cách loại bỏ nhiệm vụ Dự đoán Câu Kế Tiếp (NSP) và sử dụng tập dữ liệu huấn luyện lớn hơn. RoBERTa áp dụng phương pháp mô hình ngôn ngữ ẩn (MLM) với các cải tiến trong huấn luyện và dữ liệu. RoBERTa đã thể hiện hiệu suất vượt trội trong các nhiệm vụ NLP như phân loại văn bản và nhận diện thực thể, nhờ vào các siêu tham số và dữ liệu huấn luyện được tối ưu hóa.

**2.2.5 PhoBERT**

PhoBERT là một biến thể của BERT được đào tạo hoàn toàn trên dữ liệu văn bản tiếng Việt, cho phép mô hình nắm bắt tốt hơn 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), đồng thời được đào tạo trên tập dữ liệu quy mô lớn.

Cách tiếp cận này giúp PhoBERT hoạt động hiệu quả hơn so với các mô hình BERT hoặc RoBERTa được đào tạo trên các ngôn ngữ khác.

Kết quả thử nghiệm cho thấy PhoBERT và PhoBERT + TF-IDF là hai mô hình rất hiệu quả trong việc phân tích và phân loại tin giả. PhoBERT + TF-IDF đạt được hiệu suất tốt nhất với điểm số Accuracy, Precision và AUC cao nhất, thể hiện khả năng phân loại xuất sắc của nó. PhoBERT cũng thể hiện rất tốt với điểm số Accuracy, Precision, Recall và F1 cao.

Tuy nhiên, cần lưu ý rằng mặc dù PhoBERT + TF-IDF có Recall hơi thấp hơn so với PhoBERT, nhưng nó vẫn duy trì điểm F1 cao, phản ánh sự cân bằng tốt giữa Precision và Recall. Sự cân bằng này cho thấy PhoBERT + TF-IDF thận trọng hơn, có thể bỏ sót một số tin tức chính thống nhưng lại cung cấp dự đoán chính xác hơn tổng thể.

Trong hầu hết các trường hợp tin thật và giả đơn giản, không mơ hồ khó hiểu ví dụ như “Hà Nội gặp khó khăn khi di dời người dân ra khỏi vùng lũ” là tin thật thì cả PhoBERT và PhoBERT + TFIDF kều cho ra kết quả chính xác như nhau. Tuy nhiên trong một số trường hợp, tin tức được viết lách theo kiểu nửa thật nửa giả, hoặc thổi phồng sự thật để gây hiểu lầm, PhoBERT + TF-IDF đã chứng tỏ khả năng dự đoán chính xác hơn. Ví dụ, bài viết "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 đặc biệt nghiêm trọng 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. Trong tình huống này, PhoBERT + TF-IDF đã phân loại chính xác đây là tin giả, trong khi PhoBERT lại bị đánh lừa bởi phần tin thật trong bài viết. Sở dĩ PhoBERT + TF-IDF nhận diện đúng được trường hợp này là nhờ vào khả năng TF-IDF trong việc nhấn mạnh các từ khóa quan trọng và giảm thiểu ảnh hưởng của những từ phổ biến nhưng thiếu thông tin. TF-IDF giúp mô hình nhận ra rằng phần bổ sung không có giá trị thông tin xác thực và không nên được coi là thật, từ đó giúp phân loại tin chính xác hơn.

Ngược lại, BERT không đạt hiệu suất tốt như PhoBERT và PhoBERT TF-IDF. Mặc dù BERT cho thấy hiệu suất khá ổn định với điểm số Accuracy, Precision và F1 trung bình, nhưng điểm Recall và AUC thấp hơn chỉ ra rằng nó vẫn có tỷ lệ lỗi phân loại cao. Cuối cùng Roberta là mô hình có hiệu suất kém nhất trong số các mô hình. Mặc dù nó có Recall hoàn hảo, nghĩa là phát hiện tất cả các tin tức chính thống, nhưng điểm Precision, Accuracy và AUC rất thấp cho thấy số lượng lớn các kết quả dương tính giả, làm cho dự đoán của Roberta về tin tức chính thống không đáng tin cậy.

**CORPUS**

Sau khi hoàn thành các bước xử lý dữ liệu, bao gồm làm sạch và cân bằng dữ liệu như đã thảo luận ở phần 3.2 và 3.3, chúng tôi đã thu được bộ dữ liệu bao gồm các bài đăng trên mạng xã hội và tin tức từ các nguồn tiếng Việt. Bộ dữ liệu chứa hơn 1.400 mẫu, bao gồm cả tin thật và tin giả trên nhiều lĩnh vực 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, tạo ra 1.124 mẫu để huấn luyện và 282 mẫu để kiểm tra. 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 nhìn thấy.

**KẾT LUẬN**

Trong nghiên cứu này, chúng tôi tập trung vào việc sử dụng các mô hình Transformer như BERT, RoBERTa và PhoBERT để phân loại tin giả tại Việt Nam. Chúng tôi đã thu thập một tập dữ liệu bao gồm các bài đăng trên Facebook từ tháng 6 đến tháng 7 năm 2024, bao gồm các chủ đề về đời sống, xã hội và chính trị. Do số lượng tin giả hạn chế, chúng tôi đã bổ sung thêm các ví dụ về tin giả từ tập dữ liệu VFND, như đã mô tả trong luận văn của Ho Quang Thanh, “VNFD Vietnamese Fake News Datasets: A Collection of Vietnamese News Articles and Facebook Posts Classified into Two Labels: Real & Fake.” Sau đó, chúng tôi áp dụng các mô hình Transformer để phân loại, và kết quả đánh giá cho thấy hai mô hình PhoBERT và PhoBERT kết hợp với TF-IDF đạt hiệu suất dự đoán cao nhất đối với tiếng Việt.

Tuy nhiên, mô hình này vẫn có một số điểm hạn chế. Một trong những vấn đề chính là dữ liệu còn chưa đủ và đôi khi thông tin bị mất do cách cấu trúc ngôn ngữ tiếng Việt, như việc sử dụng từ viết tắt, ngữ pháp khác nhau, hoặc những bài viết có phần thông tin thật và giả lẫn lộn. Điều này có thể dẫn đến dự đoán sai từ mô hình.

Ngoài ra, hiện tại chúng tôi chỉ mới nghiên cứu và phân loại tin tức dựa trên nội dung của bài đăng, mà chưa tận dụng các dữ liệu bổ sung như số lượng tương tác và bình luận. Đây là những nguồn thông tin quan trọng và khá lớn. Vì vậy, trong tương lai, chúng tôi sẽ tiếp tục thu thập dữ liệu và kết hợp việc phân tích bình luận từ cả các bài viết thật và giả. Điều này giúp chúng tôi hiểu rõ hơn về cảm xúc và thái độ của người dùng đối với hai loại thông tin, từ đó cải thiện độ chính xác của dự đoán.

Kết quả phân loại sẽ được đánh giá bằng các thước đo: Độ chính xác (Accuracy), Độ chính xác (Precision), Khả năng truy hồi (Recall), Điểm F1 (F1 Score), và AUC.

Độ chính xác (Accuracy): tỉ lệ giữa số mẫu dự đoán đúng và tổng số mẫu trong tập dữ liệu kiểm tra, 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 nhiềuu 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.

Độ chính xác (Precision): tỉ lệ số mẫu true positive trong số những mẫu được mô hình phân loại là positive, chỉ ra mức độ chính xác của các dự đoán dương tính.

Khả năng truy hồi (Recall): Tỷ lệ giữa số dự đoán dương tính đúng trên tổng số dương tính thực tế. Recall cao đồng nghĩa với việc bỏ sót các mẫu thực sự positive là thấp, thể hiện khả năng của mô hình trong việc nhận diện tất cả các trường hợp dương tính.

Điểm F1 (F1 Score): Trung bình điều hòa giữa Độ chính xác và Khả năng truy hồi, giúp cân bằng hai thước đo này, đặc biệt hữu ích trong trường hợp dữ liệu không cân bằng. F-1 score càng cao tương ứng precision và recall càng cao, mô hình phân loại càng tốt.

AUC (Diện tích dưới đường cong): Đại diện cho diện tích dưới đường cong ROC (Đường đặc trưng hoạt động của bộ thu), một đồ thị thể hiện mối quan hệ giữa Tỷ lệ Dương tính Thực (True Positive Rate) và Tỷ lệ Dương tính Giả (False Positive Rate) trên các ngưỡng phân loại khác nhau. AUC đo lường khả năng của mô hình trong việc phân biệt giữa các lớp; AUC càng cao thì mô hình càng có khả năng phân biệt tốt giữa các lớp dương tính và âm tính.

● **Roberta**: Hiệu suất phân loại của Roberta khá kém, với Độ chính xác (Accuracy) chỉ đạt 0.741 và AUC là 0.835. Mặc dù Precision đạt mức ổn là 0.845, nhưng Recall chỉ ở mức 0.604, cho thấy mô hình bỏ lỡ nhiều trường hợp tin giả. F1 Score là 0.704 cho thấy mặc dù mô hình hoạt động ở mức hợp lý, nhưng không hiệu quả bằng các mô hình khác.

● **BERT**: Hoạt động tốt hơn Roberta trong phân loại, với Độ chính xác đạt 0.787 và AUC là 0.858. Precision là 0.850 và Recall là 0.708, cho thấy hiệu suất cân bằng giữa việc phát hiện tin giả và tin thật. F1 Score là 0.773 cho thấy BERT là mô hình mạnh nhưng vẫn chưa phải là tốt nhất trong các mô hình đã thử nghiệm.

● **PhoBERT**: Đạt hiệu suất cao nhất với Độ chính xác 0.872 và AUC là 0.948, cho thấy khả năng phân biệt tin thật và tin giả rất xuất sắc. F1 Score là 0.879 cho thấy mô hình này cân bằng tốt giữa Precision và Recall, mặc dù hơi thấp hơn so với PhoBERT + TF-IDF.

● **PhoBERT + TF-IDF**: Đạt Độ chính xác cao nhất là 0.898 trong số các mô hình, với Precision (0.863), Recall (0.913), và AUC (0.923) đều rất tốt. Mô hình này cân bằng tốt giữa độ chính xác và khả năng phát hiện, khiến nó trở thành mô hình hiệu quả nhất cho nhiệm vụ này.

" (A particularly severe mining accident in Quảng Ninh killing 5 workers), but the additional part "…đảng bộ và công đoàn bù nhìn chưa bao giờ lo cho điều kiện lao động của người dân" (the puppet party and trade unions have never cared about workers' conditions) is inaccurate and unverified. In this situation, PhoBERT + TF-IDF correctly classified

Bảng 2 cho thấy một số trường hợp tiêu biểu trích từ tập huấn luyện, hầu hết các trường hợp đơn giản của tin thật và tin giả, chẳng hạn như câu 1 mang tính chất thông báo và câu 4 là thông tin bịa đặt chứa nhiều từ ngữ thu hút sự chú ý cả 4 mô hình đều cho kết quả chính xác. Đối với các trường hợp tin tức chứa nhiều thông tin hơn, các mô hình như BERT và RoBERTa đã có nhiều phân loại sai, khiến cho hai mô hình này có hiệu suất thấp và trở nên không đáng tin cậy.

Mặc dù PhoBERT và PhoBERT + TF-IDF đã chứng minh khả năng dự đoán chính xác cao nhưng vẫn có một số trường hợp ngoại lệ vẫn có dự đoán sai với dạng tin có 1 nửa thông tin thật và 1 nửa là giả, ví dụ đối với câu số 5 là "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 thông tin thật là "Tai nạn sập hầm lò đặc biệt nghiêm trọng 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" lại không chính xác và chưa được kiểm chứng. Trong tình huống này, PhoBERT + TF-IDF đã phân loại đúng là tin giả, trong khi PhoBERT bị lừa bởi phần thông tin thật trong bài viết. Khả năng của PhoBERT + TF-IDF trong việc nhận diện chính xác những trường hợp như vậy là nhờ vào việc TF-IDF làm nổi bật các từ khóa quan trọng và giảm thiểu ảnh hưởng của những từ phổ biến nhưng ít thông tin. TF-IDF giúp mô hình nhận ra rằng thông tin bổ sung thiếu giá trị xác thực và không nên được coi là thật, từ đó nâng cao độ chính xác phân loại.

Tuy nhiên, điều này cũng đã khiến PhoBERT + TF IDF đã thận trọng quá mức dẫn đến đánh giá sai một số tin thật như câu số 6.

Trong giai đoạn này, chúng tôi tiến hành huấn luyện các mô hình Transformer sau khi đã có được các tập dữ liệu. Ba mô hình chính được sử dụng bao gồm BERT, RoBERTa, và PhoBERT. Đối với mỗi mô hình, chúng tôi sử dụng tokenizer tương ứng để chuyển đổi các văn bản thành các chuỗi số liệu đầu vào mà mô hình có thể xử lý được.

Chúng tôi điều chỉnh các kỹ thuật huấn luyện phù hợp với từng mô hình, bao gồm tinh chỉnh siêu tham số (hyperparameter tuning) và early stopping để tối ưu hóa hiệu suất. Trong quá trình huấn luyện, các mô hình được theo dõi và đánh giá thường xuyên để đảm bảo hiệu suất ổn định và tránh overfitting. Chúng tôi cũng thử nghiệm với các cấu hình khác nhau để tìm ra cấu hình tối ưu cho từng mô hình và mang lại kết quả chính xác nhất.

Ngoài ra, chúng tôi cũng triển khai phương pháp kết hợp PhoBERT với các đặc trưng TF-IDF. Phương pháp này nhằm khai thác tối đa khả năng hiểu ngữ nghĩa của PhoBERT, đồng thời kết hợp với thông tin về tần suất và mức độ quan trọng của từ thông qua TF-IDF. Chúng tôi tin rằng sự kết hợp này sẽ nâng cao đáng kể khả năng phân loại tin giả so với việc chỉ sử dụng PhoBERT.

Cuối cùng, chúng tôi dùng các mô hình đã huấn luyện để dự đoán nhãn cho tập testing và so sánh với nhãn thực tế để đánh giá hiệu quả của từng mô hình cũng như so sánh chúng với nhau.

#### Ví dụ

Các mô hình được huấn luyện với dữ liệu trong tập training, từ đó sẽ học được các đặc điểm thường thấy trong tin thật và tin giả. Khi gặp một tin mới, mô hình sẽ so sánh với những gì đã học để đưa ra dự đoán chính xác về việc tin đó là thật hay giả.

Ví dụ 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 các tin tức thật và “Tin sốc bệnh bạch hầu bùng phát với mức độ nguy hiểm chưa từng có, hàng trăm ca lây nhiễm mới ở Hà Nội.” là thông tin giả. Mô hình học cách phân biệt giữa các mẫu 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 các cụm từ nhấn mạnh tác động tiêu cực ví dụ như “tin sốc”, “sốc”, “hàng trăm người”, “mức độ nguy hiểm chưa từng có”,… là các từ thường xuyên xuất hiện trong tin giả và không phổ biến trong các bài viết tin thật.

Khi mô hình gặp một tin mới, ví dụ như là “Sốc, hiện tượng chưa từng thấy. Hàng nghìn người xếp hàng chờ xét nghiệm bệnh bạch hầu”. Dựa trên thông tin đã học từ các tin trong tập training, mô hình có thể nhận ra rằng tin mới này có cấu trúc và nội dung tương tự như các tin giả đã được huấn luyện trước đó. Mô hình chú ý đến các cụm từ như "sốc", "hàng nghìn người", “hiện tượng chưa từng thấy”- những từ ngữ thường xuất hiện trong các tin giả. Do đó, mô hình có khả năng cao sẽ phân loại tin mới này là "Giả".