**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. Transformers, một kiến trúc mới, 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. Điều này khiến Transformers trở nên đặc biệt có giá trị trong các nhiệm vụ như phát hiện tin tức giả, nơi việc hiểu rõ ngôn ngữ và ngữ cảnh với nhiều sắc thái là rất quan trọng.

Trong nghiên cứu này, chúng tôi tập trung vào việc tận dụng các mô hình Transformer để phát hiện tin tức giả. Cụ thể, chúng tôi sử dụng PhoBERT, một biến thể nâng cao của mô hình BERT 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 công nghệ học máy cho phép máy tính hiểu, tương tác và giải thích ngôn ngữ con người. NLP bao gồm nhiều nhiệm vụ như phân tích cú pháp, phân tích ngữ nghĩa, nhận diện thực thể và phân loại văn bản.

Trong bối cảnh phân loại văn bản, NLP trích xuất thông tin từ văn bản, xử lý ngữ nghĩa và biểu diễn văn bản dưới dạng các đặc trưng phù hợp để đưa vào các mô hình học máy hoặc học sâu. Các kỹ thuật như Bag of Words, TF-IDF (Tần suất từ - Tần suất tài liệu ngược) và nhúng từ giúp chuyển đổi văn bản thành dạng số. Sau đó, các mô hình học máy như Naive Bayes, SVM (Máy vector hỗ trợ) và các mô hình khác có thể được đào tạo để phân loại văn bản thành các danh mục như cảm xúc tích cực hoặc tiêu cực, thư rác hoặc không phải thư rác, và tin thật hoặc tin giả.

**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 tồn tại một số hạn chế, bao gồm việc dữ liệu còn hạn chế và mất mát thông tin do cấu trúc ngôn ngữ tiếng Việt, một số bài viết sử dụng từ viết tắt, ngữ pháp đa dạng, uyển ngữ hoặc một số bài viết gồm một nửa thông tin thật và giả. Điều này có thể dẫn đến dự đoán sai từ mô hình. Bên cạnh đó chúng tôi chỉ đang nghiên cứu và phân loại tin tức dựa trên nội dung của bài đăng mà chưa sử dụng đến các dữ liệu khác như số lượng tương tác và các bình luận, đây là một nguồn dữ liệu khá lớn và cũng rất quan trọng. Do đó, trong tương lai, chúng tôi sẽ tiếp tục thu thập dữ liệu và kết hợp thêm việc phân tích bình luận trên cả các bài viết thật và giả để hiểu rõ hơn về cảm xúc và thái độ của người dùng đối với hai loại thông tin, nhằm đóng góp cho kết quả dự đoán chính xác hơn.

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): Tỷ lệ giữa số dự đoán đúng trên tổng số dự đoán, phản ánh hiệu suất tổng thể của mô hình, mặc dù có thể không phản ánh chính xác khi dữ liệu không cân bằng.

Độ chính xác (Precision): Tỷ lệ giữa số dự đoán dương tính đúng trên tổng số dự đoán dương tính, 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, đo lường 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.

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.