**Chương 1. Giới thiệu**

Viết tầm 8 đến 10 trang giấy, size 13, gồm các nội dung:

**1.1 Hiện trạng (hoặc Bối cảnh, hoặc động cơ thực hiện)**

giải thích tại sao mình làm đề tài này

**1.2. Mục tiêu của đồ án**

Trình bày mục tiêu của đồ án

Những kết quả dự kiến đạt được trong đồ án

GIới hạn phạm vi của đồ án

**1.3. Cấu trúc đồ án**

Trình bày cấu trúc đồ án, chương 1 làm gì, ……

**Chương 2. Kiến thức nền tảng**

Phần này trình bày các kiến thức nền tảng phục vụ cho đề tài

**2.1. Công trình liên quan**

Liệt kê lại các công trình liên quan đến đề tài này (ưu tiên những công trình tầm 5 năm trở lại đây)

Viết tầm 10 đến 20 công trình

Mỗi công trình review tầm 3 đến 5 dòng

Làm rõ được họ đã làm gì và mình sẽ làm gì ko trùng lắp với họ thì càng tốt

**2.2  Kiến thức nền tảng**

Phần này có thể viết chung với phần 2 Công trình liên quan

Viết về các kiến thức nền tảng của bài báo

Ví dụ viết bài báo về Trả lời tự động sử dụng BERT thì cần các kiến thức nền tảng như: QA, Transfer learning, BERT, các biến thể mô hình BERT, …

**Chương 3. Phương pháp (hoặc mô hình bài toán, ….)**

Trình bày phương pháp thực hiện bài toán, thường gồm các bước

Mô hình tổng quát

Giải thích từng bước

Vi dụ minh hoạ

**Chương 4. Thực nghiệm**

Bao gồm các nội dung như:

Dữ liệu thực nghiệm: trình bày về dữ liệu cho thực nghiệm (bao nhiêu cặp câu, chia ra Train bao nhiêu, dev bao nhiêu, test bao nhiêu, ..). Nếu bài báo đóng góp phần khởi tạo dữ liệu thì phải tách riêng mục này thành nội dung lớn như một chương chẳng hạn

Công cụ đánh giá: Trình bày các độ đo mà mình sử dụng để đánh giá kết quả thực nghiệm

Kết quả thực nghiệm: trình bày kết quả thực nghiệm, thường là bảng, hình

Thảo luận: Giải thích tại sao kết quả thực nghiệm lại như vậy, thường sử dụng các ví dụ trong bộ test để minh chứng cho kết quả thực nghiệm

**Chương 5. Ứng dụng minh hoạ**

Chương này có thể sáp nhập vào chương 4 nếu nó quá ngắn

Trình bày ứng dụng minh hoạ (có thể là web, windows, …) để thể hiện ứng dụng demo nhờ dựa vào các phương pháp trên.

**Chương 6. Kết luận và hướng phát triển**

Trình bày từ 2 đến 5 trang giấy, tóm tắt kết quả đạt được và dự định nghiên cứu tiếp theo mình sẽ làm gì

6.1. Kết luận

6.2. Hướng phát triển

**Tài liệu tham khảo**

Liệt kê các tài liệu tham khảo ở đây

**CHƯƠNG 1: GIỚI THIỆU**

* 1. **Bối cảnh**

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, 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.

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, RoBERTa, và PhoBERT 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.

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. **Mục tiêu**

Với bối cảnh tin giả đang lan rộng và ngày càng phức tạp như trên, 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 là sử dụng PhoBERT - một biến thể được thiết kế dành riêng cho ngôn ngữ tiếng Việt.

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 sẽ 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 xử lý ngôn ngữ tự nhiên (NLP) và các chọn mô hình Transformer phù hợp 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. Mỗi mô hình đều có những đặc điểm riêng, và việc phân tích kỹ lưỡng hiệu suất của chúng sẽ giúp lựa chọn mô hình phù hợp nhất cho bài toán.

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 cũng như đóng góp thêm về dữ liệu để phát triển các nghiên cứu sau này.

* 1. **Cấu trúc**

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, 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 nghiên cứu***

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 các nội dung liên quan đến 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**

* 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. 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. 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.

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. 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.

Nhiều mô hình Transformer đã được phát triển ra để thực hiện các nhiệm vụ xử lý ngôn ngữ tự nhiên như BERT (Bidirectional Encoder Representations from Transformers) được Devlin và cộng sự giới thiệu lần đầu vào năm 2018, đây là mô hình có khả năng chú ý hai chiều, giúp mô hình hiểu rõ hơn ngữ cảnh của các từ trong câu. 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. Cũng trong năm 2019, Sanh và cộng sự đã giới thiệu DistilBERT, một phiên bản nhỏ gọn và nhanh hơn của BERT, phù hợp cho các ứng dụng yêu cầu phản hồi nhanh.

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. 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. PhoBERT đã giúp cải thiện hiệu suất trên 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 PhoBERT 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.

Năm 2021, Phạm Ngọc Đông và cộng sự đã đề xuất một phương pháp kết hợp PhoBERT với TF-IDF (Tần suất Thuật ngữ - Đảo ngược Tần suất Tài liệu) để tạo ra word embedding và sử dụng CNN để trích xuất đặc trưng. 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, nhưng sự phức tạp của mô hình có thể ảnh hưởng đến khả năng áp dụng thời gian thực. Khoa Đăng Phạm và cộng sự (2023) đã phát triển một mô hình mới dựa trên mô hình vELECTRA trước đó do Bùi Thế Việt và cộng sự đề xuất năm 2020, họ sử dụng các đặc trưng tiền chế và đạt được điểm AUC là 0.9575 trên tập dữ liệu ReINTEL. Tuy nhiên, sự phụ thuộc vào những đặc trưng này có thể gây khó khăn trong việc thích ứng với các tình huống khác. Trong khi đó, Võ Trung Hùng và cộng sự (2022) 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%. Dù vậy, kích thước nhỏ của tập dữ liệu của họ có thể làm giảm tính tổng quát của kết quả.

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 và metadata có thể cải thiện hiệu suất. Tuy nhiên, vẫn còn những thách thức lớn như kích thước tập dữ liệu, sự đa dạng và độ phức tạp tính toán mà các nghiên cứu trong tương lai cần giải quyết.

ViBERT và VisoBERT

* 1. **Kiến thức nền tảng**

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):** 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 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.

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

TF-IDF là một phương pháp được sử dụng rộng rãi trong Xử lý Ngôn ngữ Tự nhiên (NLP) và khai thác văn bản [17], [18]. Nó giúp đánh giá mức độ quan trọng của một từ trong một tài liệu, dựa trên cả tần suất của từ đó trong tài liệu và tần suất của từ 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: Term Frequency (TF) và Inverse Document Frequency (IDF).

Term Frequency (TF): là tần suất xuất hiện của một từ “t” trong một tài liệu “d”

**Công thức:**

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.

**Công thức:**

**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, chúng tôi sử dụng TF-IDF như một bước tiền xử lý để chuyển đổi văn bản thành các vector đặc trưng. Sau đó, các vector này 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 tác động của các từ phổ biến nhưng ít mang lại thông tin trong quá trình huấn luyện mô hình.

***2.2.2 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". 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. 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.

Mô hình Transformer bao gồm hai thành phần chính:

- **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.

***2.2.3 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. 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 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. 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 mô hình ngôn ngữ lớn điển hình.

***2.2.4 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.

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.

***2.2.5 PhoBERT***

PhoBERT là một biến thể của BERT được phát triển nhằm giải quyết những hạn chế của BERT khi xử lý tiếng Việt. 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) – 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 trên một tập dữ liệu đa dạng 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. 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. Từ đó PhoBERT có thể được 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 đóng vai trò quan trọng trong việc phân tích và hiểu ngữ nghĩa của văn bản Tiếng Việt. Bằng cách học 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 nâng cao độ chính xác của PhoBERT trong việc phát hiện tin giả Tiếng Việt so với các mô hình khác.

**CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU**

**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 1: (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 sử dụng dữ liệu đã xử lý để huấn luyện các mô hình Transformer: BERT, RoBERTa, và PhoBERT. 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.

**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 thu thập dữ liệu tự động, mô phỏng các hành động của người dùng như duyệt trang web và trích xuất dữ liệu. Cuối cùng, chúng tôi đã thu thập được hai tập dữ liệu: một cho tin tức thật và một cho tin tức giả, như được minh họa trong Hình 2.

****

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.



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 3. 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 4 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*, 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. Điều này làm giảm sự thiên lệch và giúp cải thiện độ chính xác của mô hình.

|  |  |
| --- | --- |
| *Figure 4. The size of 2 data sets after collection.* | *Figure 5.The size of 2 data sets after processing.* |

Biểu đồ Hình 5 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. Ba mô hình chính mà chúng tôi lựa chọn bao gồm BERT, RoBERTa, và PhoBERT. 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 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 PhoBERT với các đặc trưng TF-IDF (Term Frequency-Inverse Document Frequency). 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 hy vọng rằng mô hình này sẽ không chỉ hiểu được các khái niệm ngữ nghĩa mà còn xác định được những từ hoặc cụm từ nào trong văn bản có ảnh hưởng nhiều đến việc phân loại tin thật và tin giả. 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 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 accuracy, precision, recall, và F1-score.

**3.5 Ví dụ minh họa**

Các mô hình được huấn luyện với dữ liệu trong tập huấn luyện, giúp chung 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 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 xã. Trong khi đó “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. 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 các cụm từ và ngữ nghĩa 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 hiếm khi xuất hiện trong các tin thật, chính thống.

Khi mô hình gặp một tin mới, chẳng hạn như là “Sốc, hiện tượng chưa từng thấy. Hàng nghìn người xếp hàng chờ xét nghiệm bệnh bạch hầu”, mô hình sẽ dựa trên những gì đã học để phân tích. Các cụm từ như "sốc", "hàng nghìn người", “hiện tượng chưa từng thấy” là những dấu hiệu thường xuất hiện trong các tin giả. Vì vậy, mô hình có xu hướng dự đoán rằng tin mới này là "Giả", dựa trên sự tương đồng về cấu trúc và ngữ nghĩa với các mấu tin giả đã được huấn luyện trước đó.

**CHƯƠNG 4: THỰC NGHIỆM**

**4.1 Dữ liệu thực nghiệm**

Sau khi hoàn thành các bước xử lý dữ liệu, bao gồm 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 1.400 mẫu, bao gồm cả tin thật và tin giả từ 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, kết quả thu được tập dữ liệu huấn luyện có 1.124 mẫu và tập kiểm tra với 282 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 nhìn thấy.

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

**4.2 Công cụ đánh giá**

**4.3 Kết quả thực nghiệm**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC** |
| **PhoBERT**  **+ TF-IDF** | **0.888112** | **0.863014** | **0.913043** | **0.887324** | 0.922689 |
| **PhoBERT** | 0.872340 | 0.850649 | 0.909722 | 0.879195 | **0.947665** |
| **BERT** | 0.787234 | 0.850000 | 0.708333 | 0.772727 | 0.858343 |
| **RoBERTa** | 0.741135 | 0.844660 | 0.604167 | 0.704453 | 0.834541 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**4.4 Thảo luận**

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.

**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.

**Evaluation tool**

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.

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.