TÓM TẮT CÔNG TRÌNH

Trong bối cảnh các thế lực thù địch ngày càng gia tăng hoạt động chống phá trên không gian mạng, đặc biệt thông qua các bình luận mang tính xuyên tạc, phản động nhằm công kích chế độ và Đảng Cộng sản Việt Nam, công trình này được thực hiện với mục tiêu xây dựng một hệ thống có khả năng nhận diện hiệu quả các bình luận phản động trên mạng xã hội tiếng Việt.

Công trình gồm hai đóng góp chính: (1) xây dựng bộ dữ liệu chuyên biệt với gần 19.000 bình luận, được thu thập từ Facebook, TikTok, Reddit và Threads, gán nhãn theo ba nhóm: Phản động (PD), Không phản động (KPD), và Không liên quan (KLQ); (2) đề xuất và so sánh toàn diện các mô hình học máy và học sâu hiện đại như Logistic Regression, CNN, LSTM, các mô hình Transformer (PhoBERT, CafeBERT, XLM-R) và mô hình ngôn ngữ lớn (Vistral-7B, SeaLLM, Qwen3-32B) trong bài toán phân loại.

Quy trình xử lý dữ liệu được thiết kế chặt chẽ gồm: chuẩn hóa Unicode, loại bỏ nhiễu, chuẩn hóa từ vựng (1.247 biến thể được xử lý), tóm tắt bài viết cung cấp ngữ cảnh, gán nhãn tự động bằng Gemini API và kiểm tra thủ công bằng giao diện do nhóm phát triển. Các mô hình được huấn luyện và đánh giá qua các chỉ số F1-score, đặc biệt tập trung vào khả năng phát hiện bình luận phản động – vốn chiếm tỷ lệ nhỏ trong tập dữ liệu (~11%).

Kết quả cho thấy mô hình Vistral-7B đạt hiệu suất vượt trội (F1-macro = 0.74, F1-PD = 0.69), khẳng định vai trò của LLMs chuyên biệt khi được fine-tune. Ngược lại, các mô hình zero-shot cho hiệu quả rất thấp. Qua đó, nhóm sinh viên tìm hiểu rút ra rằng việc tinh chỉnh mô hình trên dữ liệu đặc thù và hiểu ngữ cảnh chính trị – xã hội là yếu tố quyết định thành công.

Công trình không chỉ có ý nghĩa học thuật trong việc đóng góp dữ liệu, phương pháp và đánh giá, mà còn mang lại giá trị ứng dụng thiết thực trong công tác kiểm

duyệt nội dung, góp phần bảo vệ an ninh tư tưởng, giữ gìn môi trường mạng lành mạnh.

Chương 1. ĐẶT VẤN ĐỀ

1.1. Bối cảnh đề tài

Trong khoảng thời gian vừa qua, nước ta vừa trải qua kỉ niệm 50 năm ngày Giải phóng miền Nam, thống nhất đất nước, đây là một mốc son chói lọi trong lịch sử dựng nước và giữ nước của dân tộc Việt Nam. Trong bối cảnh cả nước dâng cao niềm tự hào dân tộc và cùng nhìn lại chặng đường phát triển vẻ vang, không gian mạng đã trở thành một mặt trận tư tưởng sôi động. Tuy nhiên, đây cũng chính là thời điểm cho các thế lực thù địch, phản động lợi dụng để đẩy mạnh hoạt động chống phá. Chúng đã thông qua sự kiện này để đăng tải các thông tin có nội dung sai trái, xuyên tạc sự thật lịch sử, phủ nhận thành quả cách mạng và công kích trực diện vào vai trò lãnh đạo của Đảng Cộng sản Việt Nam [1].

Mặt khác, việc kiểm soát và ngăn chặn các bình luận chống phá đang đối mặt với những thách thức, khó khăn. Khi sự kiểm duyệt thủ công bởi con người không thể bao quát hết được khối lượng thông tin khổng lồ và luôn chậm một nhịp so với tốc độ lan truyền của các thông tin sai sự thật. Thêm vào đó, các đối tượng có tư tưởng chống phá nhà nước thường sử dụng lối diễn đạt tinh vi, biến hóa liên tục với từ lóng, teencode, ẩn dụ, mỉa mai và các ký hiệu đặc biệt nhằm tránh né các bộ lọc từ khóa truyền thống từ hệ thống kiểm duyệt [2]. Các hệ thống kiểm duyệt tự động hiện có của các nền tảng xuyên quốc gia thường không được tối ưu cho bối cảnh chính trị và đặc thù ngôn ngữ tiếng Việt, dẫn đến hiệu quả thấp. Thực trạng này đặt ra một yêu cầu cấp bách về đề xuất ra một giải pháp nhận diện hiệu quả hơn để hỗ trợ xử lý loại nội dung độc hại này, từ đó xoá bỏ đi các thông tin sai trái, thù địch.

Bên cạnh đó, phần lớn các công trình tại Việt Nam thường tập trung vào phát hiện *tin giả* (fake news) và phát hiện *ngôn từ thù ghét* (hate speech) với nhiều mô hình và bộ dữ liệu được công bố [3] [4], tuy nhiên việc ứng dụng công nghệ này vào các bài toán đặc thù về an ninh chính trị vẫn còn là một khoảng trống lớn. Hiện tại, vẫn

chưa có công trình nào đi sâu vào bài toán nhận diện chống phá Nhà nước Việt Nam – một chủ đề chính trị đặc thù và nhạy cảm, đòi hỏi các mô hình trí tuệ nhân tạo phải có sự am hiểu sâu sắc về ngữ cảnh chính trị, lịch sử và các hình thức biểu đạt tinh vi như ẩn dụ, mỉa mai hay teencode, tất cả đều là những thách thức lớn đối với các phương pháp phân tích ngôn ngữ tự động.

1.2. Lý do chọn đề tài

Xuất phát từ chính yêu cầu cấp bách đó, đề tài này được ra đời nhằm mục đích thử nghiệm và so sánh hiệu quả giữa các phương pháp học máy trong việc nhận diện bình luận có dấu hiệu "phản động". Cụ thể, công trình sẽ so sánh hiệu suất của các mô hình truyền thống với những kiến trúc hiện đại như *Transformers và Mô hình Ngôn ngữ lớn* (LLMs). Mục tiêu là tìm ra giải pháp tối ưu cho bài toán này trên ngữ cảnh mạng xã hội tiếng Việt.

Việc lựa chọn đề tài này được thúc đẩy bởi hai giá trị cốt lõi:

Giá trị khoa học: Công trình có hai đóng góp chính.

- Thứ nhất, xây dựng và công bố một bộ dữ liệu mới, chuyên biệt về nội dung chống phá nhà nước trên mạng xã hội tiếng Việt, được thu thập, gán nhãn và kiểm tra thủ công để làm tài nguyên cho cộng đồng.
- Thứ hai, đề tài cung cấp một phép đối sánh toàn diện, kiểm chứng hiệu quả của các mô hình từ truyền thống đến LLMs trên miền dữ liệu đặc thù này, qua đó lấp đầy khoảng trống nghiên cứu quan trọng trong lĩnh vực.

Giá trị thực tiễn: Đề tài mang kỳ vọng tạo ra một giải pháp công nghệ hữu ích, có khả năng hỗ trợ các cơ quan chức năng trong việc sàng lọc, phát hiện sớm các nội dung độc hại, góp phần bảo vệ nền tảng tư tưởng của Đảng, giữ vững an ninh văn hóa và xây dựng một không gian mạng quốc gia trong sạch, lành mạnh.

Chương 2. TỔNG QUAN TÀI LIỆU

2.1. Tổng quan về tuyên truyền chống phá nhà nước trên không gian mạng

Tuyên truyền chống phá Nhà nước là hành vi phát tán các thông tin, tài liệu, luận điệu xuyên tạc, phỉ báng Chính quyền nhân dân và Nhà nước Cộng hòa xã hội chủ nghĩa Việt Nam, nhằm gây hoang mang, nghi ngờ, bất mãn trong quần chúng, góp phần kích động chia rẽ nội bộ, làm suy giảm niềm tin của nhân dân vào chế độ.

Trong thực tiễn, tuyên truyền chống phá Nhà nước trên không gian mạng có thể thể hiện qua việc phổ biến những thông tin sai sự thật về chủ trương, chính sách của Đảng, Nhà nước; xuyên tạc, phủ nhận vai trò lãnh đạo của Đảng Cộng sản Việt Nam; cổ xúy các luận điệu phản động như đặt “nhân quyền cao hơn chủ quyền”; kích động biểu tình, lật đổ chế độ thông qua các chiến dịch tuyên truyền có quy mô trên không gian mạng. Hành vi này có thể thực hiện công khai hoặc bí mật, dưới nhiều hình thức khác nhau [5], [6].

Về mặt hành vi pháp lý, theo Điều 16 luật An ninh mạng Việt Nam [7], các hành vi tuyên truyền chống phá nhà nước này bao gồm:

a) Tuyên truyền xuyên tạc, phỉ báng chính quyền nhân dân.

b) Chiến tranh tâm lý, kích động chiến tranh xâm lược, chia rẽ, gây thù hận giữa các dân tộc, tôn giáo và nhân dân các nước.

c) Xúc phạm dân tộc, quốc kỳ, quốc huy, quốc ca, vĩ nhân, lãnh tụ, danh nhân, anh hùng dân tộc.

Âm mưu và thủ đoạn của các thế lực thù địch trong việc tuyên truyền chống phá trên mạng thường đi kèm với các chiến dịch "chiến tranh tâm lý", phát tán tin giả, xuyên tạc các sự kiện, gây rối loạn thông tin, và lợi dụng các tổ chức xã hội dân sự,

các diễn đàn, mạng xã hội để tập hợp lực lượng phản đối, kích động "bất tuân dân sự", nhằm mục đích làm suy yếu nền tảng tư tưởng của Đảng và chế độ [5], [8].

Có thể thấy rằng, đây là một dạng nội dung độc hại (Harmful Content) đặc biệt phức tạp, không chỉ tấn công vào các cá nhân hay tổ chức mà còn nhắm đến nền tảng tư tưởng, sự ổn định chính trị và an ninh quốc gia. Khác với các dạng *ngôn từ thù ghét* (Hate Speech) hay *tin giả* (Fake News) thông thường, nội dung chống phá thường được thể hiện một cách tinh vi, có tổ chức, sử dụng các thủ đoạn đan cài thật - giả và khai thác các chủ đề nhạy cảm để thao túng dư luận [2].

2.2. Xu hướng nghiên cứu hiện đại về phát hiện nội dung độc hại

Phát triển nội dung độc hại là một đề tài nổi bật trong lĩnh vực xử lý ngôn ngữ tự nhiên trong hơn một thập kỷ qua. Ban đầu, các phương pháp tiếp cận chủ yếu dựa vào học máy truyền thống (Classical Machine Learning), sử dụng các thuật toán như Support Vector Machines (SVM) [9], Naive Bayes [10] kết hợp với các đặc trưng được thiết kế thủ công như Bag-of-Words và TF-IDF. Các phương pháp này tỏ ra hiệu quả trong việc nhận diện các nội dung độc hại thể hiện rõ ràng qua từ khóa (explicit hate speech), song gặp phải hạn chế lớn trong việc nắm bắt ngữ nghĩa và các sắc thái tinh vi của ngôn ngữ [11].

Tiếp đến là sự ra đời của học sâu (Deep Learning), sự phát triển của các mô hình học sâu đã mang lại một hướng tiếp cận mới nhằm giải quyết những hạn chế của phương pháp dựa trên đặc trưng thủ công. Bằng cách sử dụng kỹ thuật nhúng từ (word embeddings) cùng với các mô hình như Mạng nơ-ron tích chập (CNN) [12] hay Mạng bộ nhớ dài-ngắn hạn (LSTM) [13] có khả năng tự động học các đặc trưng biểu diễn từ dữ liệu văn bản, giúp hiểu được ngữ cảnh của từ và các mẫu câu phức tạp hơn. Nhiều công trình thực nghiệm cách tiếp cận này giúp cải thiện đáng kể hiệu suất phát hiện các dạng nội dung độc hại ẩn ý (implicit hate speech) so với phương pháp truyền thống [14] [15]. Tuy nhiên, cũng có nhiều công trình cho ra kết quả ngược lại [16]

Tiếp theo nữa là sự ra đời kiến trúc Transformer và các mô hình ngôn ngữ tiền huấn luyện (PLMs) như BERT [17] và RoBERTa [18]. Với cơ chế tự chú ý (self-

attention), các mô hình này có khả năng nắm bắt các mối quan hệ ngữ nghĩa và ngữ cảnh ở tầm xa, cho phép chúng hiểu sâu sắc hơn về văn bản [19]. Việc tinh chỉnh (fine-tuning) các mô hình đã được tiền huấn luyện trên các bộ dữ liệu về nội dung độc hại đã mang lại hiệu suất vượt trội so với các mô hình truyền thống và dần nhanh chóng trở thành phương pháp tiêu chuẩn cho xử lý ngôn ngữ tự nhiên trên nhiều các tác vụ khác nhau, đặc biệt là phát hiện nội dung độc hại [20] [21].

Gần đây nhất, xu hướng nghiên cứu đang tập trung vào việc khai thác tiềm năng của các Mô hình Ngôn ngữ lớn (LLMs). Với khả năng suy luận phức tạp và học trong ngữ cảnh (in-context learning), các LLMs mở ra hướng tiếp cận mới để giải quyết các trường hợp khó nhất, chẳng hạn như phát hiện sự mỉa mai, các thông điệp ngầm, hay các dạng nội dung độc hại đòi hỏi kiến thức nền về văn hóa-xã hội [22]. Chính vì những năng lực này, nhóm sinh viên xem đây là hướng tiếp cận hứa hẹn nhất để giải quyết các dạng nội dung độc hại tinh vi mà các mô hình trước đây thường bỏ sót.

Tuy nhiên, một thực tế quan trọng là phần lớn các tài nguyên, bộ dữ liệu và các mô hình tiên tiến kể trên đều được phát triển và tối ưu chủ yếu cho tiếng Anh, ngôn ngữ có nguồn tài nguyên dồi dào nhất (high-resource language). Trong khi đó, việc phát hiện nội dung độc hại ở các ngôn ngữ ít tài nguyên (low-resource languages) như tiếng Việt lại đối mặt với những thách thức lớn hơn nhiều. Sự thiếu hụt các bộ dữ liệu được gán nhãn quy mô lớn, sự phức tạp về cấu trúc ngôn ngữ và các đặc thù văn hóa là những rào cản chính [23] Cụ thể, các thách thức này bao gồm:

- **Thiếu hụt tài nguyên dữ liệu:** Đây là trở ngại lớn nhất. Việc thiếu các bộ dữ liệu lớn, được gán nhãn chất lượng cao khiến việc huấn luyện và đánh giá các mô hình học sâu trở nên khó khăn và kém tin cậy[23].
- **Sự phức tạp về ngôn ngữ và văn hóa:** Các ngôn ngữ ít tài nguyên thường có các hiện tượng ngôn ngữ phức tạp như hoán ngữ (code-mixing), từ lóng, phương ngữ và các sắc thái văn hóa riêng biệt mà các mô hình được tiền huấn luyện trên dữ liệu tiếng Anh không thể nắm bắt được.

- **Sự thiên vị trong thu thập dữ liệu:** Dữ liệu được thu thập từ một nền tảng hoặc một nhóm người dùng cụ thể có thể không đại diện cho toàn bộ cộng đồng, dẫn đến việc mô hình học phải các đặc trưng sai lệch và hoạt động kém hiệu quả khi triển khai trong thực tế [23].

Thực trạng này tạo ra một sự chênh lệch đáng kể trong năng lực kiểm duyệt nội dung tự động giữa các ngôn ngữ, đòi hỏi các công trình phải tập trung vào việc xây dựng tài nguyên và phương pháp phù hợp cho từng bối cảnh ngôn ngữ cụ thể.

2.3. Tình hình nghiên cứu và ứng dụng tại Việt Nam

Trong bối cảnh nghiên cứu hiện tại, lĩnh vực phát hiện nội dung độc hại tại Việt Nam cũng đã có những bước tiến đáng kể với sự xuất hiện của các mô hình ngôn ngữ tiền huấn luyện chuyên biệt cho tiếng Việt. Các công trình gần đây tập trung chủ yếu vào việc áp dụng kiến trúc Transformer, trong đó PhoBERT-CNN đã đạt được F1-score 67.5% trên dataset ViHSD và 98.5% trên HSDVLSP [24]. Đặc biệt, ViSoBERT được tiền huấn luyện trên dữ liệu mạng xã hội Việt Nam đã thể hiện hiệu suất vượt trội so với cả PhoBERT và XLM-R chuẩn [25], trong khi ViHateT5 với kiến trúc unified text-to-text đạt Macro F1 từ 68.7% đến 86.4% trên các benchmark khác nhau [25].

Phạm vi nghiên cứu hiện tại đã mở rộng khá đáng kể từ các tác vụ truyền thống sang nhiều loại nội dung độc hại khác nhau. Bên cạnh nhận diện ngôn ngữ thù ghét với ViHSD (33,400 bình luận) [26] và VOZHSD (hơn 10 triệu comments) [25], các nhà nghiên cứu Việt Nam đã phát triển các bộ dữ liệu cho nhận diện tin giả như CTUJS Fake News (25,000 bài viết chính trị) [27], cũng như nhận diện spam với ViSpamReviews đạt 86.9%/72.2% F1-score cho binary/multi-class classification [28]. Tuy nhiên, lĩnh vực này vẫn đối mặt với những thách thức lớn bao gồm thiếu dataset chất lượng cao cho nội dung chính trị-tuyên truyền, khó khăn trong xử lý các biến thể ngôn ngữ phức tạp như từ lóng và teencode, cũng như vấn đề data imbalance đòi hỏi Inter-Annotator Agreement cao trong quá trình gán nhãn.

2.4. Vấn đề tồn tại và phương án giải quyết

Qua quá trình tìm hiểu tài liệu, việc phát hiện bình luận tuyên truyền chống phá nhà nước trên mạng xã hội tiếng Việt đang đối mặt với những thách thức đáng kể. Đây là một lĩnh vực cần được quan tâm đặc biệt nhưng lại thiếu những công trình chuyên sâu.

Điều có thể quan sát được từ các công trình hiện tại là nội dung tuyên truyền chống phá có độ phức tạp cao hơn nhiều so với ngôn ngữ thù ghét hay tin giả thông thường. Khác với các loại nội dung độc hại khác thường sử dụng từ khóa rõ ràng, nội dung tuyên truyền chống phá thường được "ngụy trang" dưới nhiều hình thức tinh vi như ẩn dụ, mỉa mai, so sánh ngầm, và đặc biệt là sử dụng từ lóng cũng như teencode. Qua khảo sát tài liệu, các phương pháp truyền thống dựa trên từ khóa hoặc học máy cơ bản không thể xử lý hiệu quả những sắc thái ngữ nghĩa phức tạp này. Một khó khăn lớn trong quá trình tìm hiểu là sự thiếu hụt nghiêm trọng các bộ dữ liệu chất lượng cao được gán nhãn cho lĩnh vực này. Mặc dù đã có những dataset quan trọng như ViHSD cho ngôn từ thù ghét [26] hay CTUJS Fake News cho tin giả [27], nhưng vẫn chưa có một bộ dữ liệu chuẩn nào chuyên biệt cho bài toán phát hiện tuyên truyền chống phá tại Việt Nam. Điều này tạo ra rào cản lớn trong việc phát triển và đánh giá các mô hình. Từ các công trình hiện có, mặc dù có nhiều công trình về phát hiện nội dung độc hại nói chung, nhưng vẫn tồn tại một khoảng cách lớn trong việc áp dụng vào bài toán cụ thể về an ninh chính trị tại Việt Nam. Các hệ thống hiện tại chưa được tối ưu hóa cho bối cảnh đặc thù của Việt Nam, và số lượng công trình trong nước tập trung vào chủ đề này vẫn còn hạn chế.

Dựa trên những vấn đề đã xác định, công trình này đề xuất một phương pháp tiếp cận có hệ thống gồm hai thành phần chính. Bước đầu tiên và quan trọng nhất là xây dựng một bộ dữ liệu mới chuyên biệt cho bài toán này. Kế hoạch bao gồm việc thu thập các bình luận từ các nền tảng mạng xã hội phổ biến tại Việt Nam như Facebook, YouTube, và các diễn đàn trực tuyến. Sau đó sẽ thực hiện quá trình gán nhãn thủ công bởi nhóm sinh viên, dựa trên một bộ quy tắc rõ ràng và nhất quán.

Để giải quyết vấn đề về độ phức tạp ngôn ngữ, công trình sẽ thực hiện một so sánh toàn diện các kiến trúc khác nhau thay vì chỉ dựa vào một mô hình duy nhất. Quá trình sẽ bắt đầu với các mô hình học máy truyền thống như Random Forest, Naive Bayes, và Logistic Regression kết hợp TF-IDF để thiết lập ngưỡng cơ sở. Tiếp theo, công trình sẽ triển khai các mô hình học sâu như CNN và LSTM để đánh giá khả năng nắm bắt ngữ cảnh tuần tự. Đối với các mô hình Transformer, công trình sẽ tập trung vào PhoBERT, CafeBERT (chuyên biệt cho tiếng Việt) và XLM-RoBERTa (đa ngôn ngữ) để khai thác khả năng hiểu ngữ cảnh sâu, dựa trên kết quả tích cực của. Cuối cùng, công trình sẽ thử nghiệm với các Large Language Models như Vistral-7B (chuyên biệt cho tiếng việt), SeaLLMs3-7B (chuyên biệt cho ngôn ngữ trong khối đông nam á) và Qwen3-32B (đa ngôn ngữ) để khám phá khả năng xử lý các trường hợp phức tạp nhất, theo hai phương thức là zero-shot và fine-tuning. Qua cách tiếp cận này, công trình hy vọng có thể đóng góp vào việc lấp đầy khoảng trống nghiên cứu hiện tại và cung cấp những hiểu biết có giá trị cho việc phát triển các hệ thống phát hiện nội dung độc hại hiệu quả trong thực tế.

Chương 3. MỤC TIÊU - PHƯƠNG PHÁP

3.1. Mục tiêu công trình

Mục tiêu của công trình là phát triển một hệ thống có khả năng nhận diện và phân loại bình luận trên mạng xã hội Việt Nam theo ba nhãn: PHAN_DONG, KHONG_PHAN_DONG và KHONG_LIEN_QUAN. Bên cạnh đó, công trình tập trung xây dựng bộ dữ liệu tiếng Việt được gán nhãn chính xác, đồng thời đánh giá và so sánh hiệu quả của các phương pháp phân loại khác nhau, bao gồm cả các mô hình học máy truyền thống và mô hình hiện đại dựa trên khả năng hiểu ngữ cảnh phức tạp của bình luận, nhằm tìm ra giải pháp tối ưu phù hợp với đặc điểm ngôn ngữ và văn hóa Việt Nam.

3.2. Phương pháp nghiên cứu

Sinh viên tìm hiểu dựa trên ba nội dung chính:

1) Xây dựng tiêu chí đánh giá bình luận có nội dung chống phá Nhà nước trong bối cảnh chính trị Việt Nam

2) Phương pháp xây dựng bộ dữ liệu bình luận chống phá nhà nước trên mạng xã hội tiếng Việt

3) Phương pháp đánh giá hiệu quả các phương pháp *học máy truyền thống*, mô hình dựa trên kiến trúc *Transformers* cũng như các *mô hình ngôn ngữ lớn* (LLMs) trong việc huấn luyện và phân loại nhằm phát hiện chính xác các bình luận thuộc các nhóm nhãn đã xác định.

3.2.1. Tiêu chí đánh giá bình luận có nội dung phản động

Dựa trên các biểu hiện chống phá nhà nước đã đề ra ở Mục 2.1, công trình lập ra 3 tiêu chí đánh giá bình luận có nội dung phản động như sau.

Phản động:

Phản động là tiêu chí đánh giá các bình luận mang nội dung có chủ đích truyền bá, ủng hộ hành vi chống đối chính quyền, xóa bỏ chế độ, lật đổ bộ máy chính quyền Nhà nước hiện nay. Các biểu hiện cơ bản có thể kể đến:

- Công kích, xuyên tạc, phản đối Đảng, Nhà nước, chính quyền.
- Phủ nhận hoặc bóp méo các sự kiện lịch sử quan trọng của đất nước (30/4, Điện Biên Phủ, Cách mạng Tháng 8...).
- Xúc phạm các lãnh đạo, biểu tượng quốc gia (Bác Hồ, quốc kỳ, quốc ca,...)
- Tuyên truyền chống phá chế độ xã hội chủ nghĩa, ủng hộ việc khôi phục chế độ cũ.
- Đồng tình, ủng hộ quan điểm phản động trong bài đăng.
- Kích động bạo lực, thù hận đối với Nhà nước.

Ví dụ một vài bình luận được đánh giá *Phản động*: “giải tán Đảng Cộng sản Việt Nam”, “chủ nghĩa cộng sản là một chủ nghĩa bịp bợm”, “cộng sản chủ trương tam vô”.

Không phản động:

Không phản động là tiêu chí đánh giá các bình luận mang nội dung trung lập, bảo vệ chủ trương, chính sách của Đảng. Chỉ trích, góp ý chính sách trong khuôn khổ pháp luật. Các biểu hiện bao gồm như sau.

- Phản biện ôn hòa, góp ý xây dựng về chính sách
- Đề cập sự kiện, cột mốc lịch sử đúng sự thật, khách quan
- Ủng hộ, bảo vệ Đảng, Nhà nước, chính quyền
- Phê phán các quan điểm, đối tượng có tư tưởng phản động
- Thảo luận trung lập về vấn đề chính trị-xã hội
- Đề xuất cải thiện trong khuôn khổ pháp luật
- Bày tỏ lòng yêu nước một cách tích cực

Ví dụ một vài bình luận được đánh giá *Không phản động*: “tẩy chay kênh phản động khốn nạn này đi mọi người ơi”, “đây là giọng điệu của bọn ba que”, “đúng là một lũ phản động xuyên tạc lịch sử của đất nước nghiệp quật nhé”

Không liên quan:

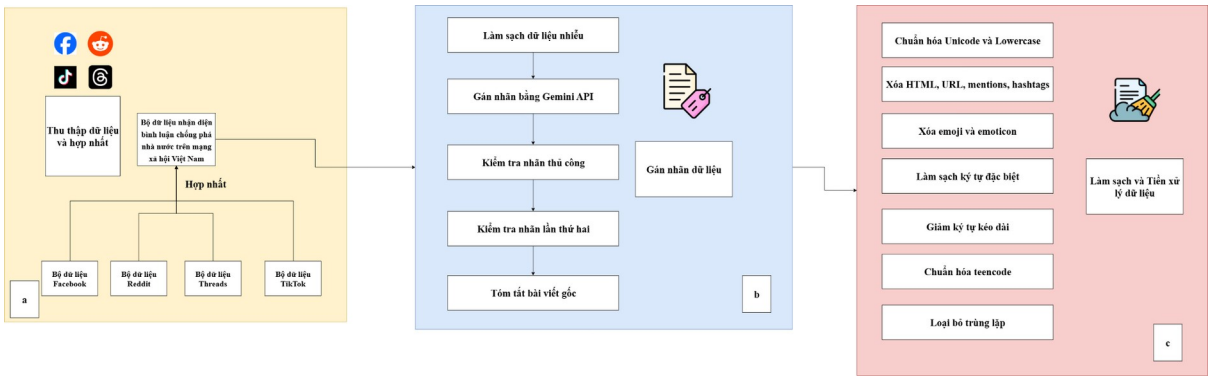
Không liên quan là tiêu chí đánh giá các bình luận mang nội dung không chứa nội dung chính trị, không đề cập đến chế độ, chính quyền hoặc các vấn đề bạo lực, chống phá Nhà nước. Các biểu hiện sẽ bao gồm:

- Không đề cập đến chính trị, chính sách, lãnh đạo
- Spam, quảng cáo, emoji đơn thuần
- Hỏi thông tin cá nhân, trò chuyện phiếm
- Bàn về thể thao, giải trí, đời sống không liên quan chính trị
- Bình luận kỹ thuật, hướng dẫn không có ý nghĩa chính trị

Ví dụ một vài bình luận được đánh giá *Không phản động*: “minh dinh tỗ cha đứa nào post hình rồi lại xóa”, “e ở đâu vậy e chị ở bến tre nè e có dịp nào về bến tre alo chị nhé”, “ngày xưa phạt 3tr còn j phạt 5tr”.

3.2.2. Phương pháp xây dựng bộ dữ liệu

Quy trình xây dựng bộ dữ liệu bao gồm các giai đoạn như sau:



Hình 3.1: Các giai đoạn xây dựng bộ dữ liệu

3.2.2.1. Phương pháp thu thập dữ liệu

Sinh viên sẽ tiến hành thu thập nội dung bài viết và bình luận từ các nền tảng mạng xã hội phổ biến như Facebook, TikTok, Threads và Reddit. Các bài viết được lựa chọn liên quan đến các chủ đề *chính trị, lịch sử, kỉ niệm ăn mừng các cột mốc của đất nước*, đặc biệt là các cuộc thảo luận xoay quanh về kỉ niệm 50 năm ngày 30/4, với mức độ tương tác cao (trung bình trên 100 bình luận mỗi bài viết) và không giới hạn nội dung bình luận từ người dùng. Sau khi hoàn tất quá trình thu thập, các bình luận được loại bỏ tên người dùng khỏi dữ liệu nhằm bảo vệ danh tính và đảm bảo tính ẩn danh của người dùng.

Sau khi hoàn tất quá trình thu thập dữ liệu từ các bài viết được lựa chọn, sinh viên tiến hành hợp nhất các file dữ liệu riêng lẻ từ mỗi nền tảng mạng xã hội thành một file tổng hợp. Quá trình này được thực hiện đồng nhất cho cả bốn trang mạng xã hội là *Facebook, TikTok, Threads và Reddit*. Kết quả cuối cùng là một file csv tổng hợp chứa toàn bộ dữ liệu thu thập bao gồm các cột: post_raw, comment_raw, created_date, platform.

3.2.2.2. Phương pháp gán nhãn dữ liệu

Bộ dữ liệu thô thu thập trước đó sẽ thông qua các bước sau để đảm bảo chất lượng cho nhãn được gán.

- **Làm sạch dữ liệu nhiễu:**

Để chuẩn bị cho quá trình gán nhãn, bộ dữ liệu thô được làm sạch thông qua nhiều bước được tự động hóa. Cụ thể, các bình luận quá dài (>300 từ) hoặc quá ngắn (≤3 từ) sẽ bị loại bỏ, trừ trường hợp các bình luận ngắn chứa những từ khóa chính trị đặc thù (cs, vnch, phản động, v.v.) được định nghĩa trước. Bên cạnh đó, sinh viên cũng thực hiện các bước tiền xử lý làm sạch dữ liệu nhiễu như *chuẩn hóa Unicode về NFC, chuyển về lowercase, loại bỏ emoji, liên kết URL, thẻ HTML, mentions, hashtags, xóa dấu câu và khoảng trắng thừa* để chuẩn bị cho quá trình gán nhãn bằng Gemini. Cuối cùng, số lượng bình luận cho mỗi bài viết được cân bằng lại (mặc định tối đa 750 bình luận/bài, có thể điều chỉnh số lượng) để giảm sự sai lệch do một vài

bài viết có quá nhiều tương tác, trong đó các bình luận chứa từ khóa nhạy cảm được ưu tiên giữ lại.

- **Tóm tắt bài viết gốc:**

Để cung cấp ngữ cảnh rõ ràng cho việc gán nhãn bình luận và tối ưu hóa tài nguyên xử lý, nội dung của mỗi bài viết gốc được tóm tắt bằng cách gọi Gemini API. Một câu prompt (hướng dẫn) chi tiết được thiết kế để hướng dẫn mô hình A.I tạo ra một bản tóm tắt có cấu trúc gồm 3 phần chính:

1) Nội dung sơ lược của bài viết.

2) Vấn đề chính được đề cập.

3) Đánh giá bài viết có chứa nội dung phản động hoặc tin giả hay không.

- **Gán nhãn bằng Gemini API:**

Quá trình gán nhãn tự động được thực hiện bằng cách gọi Gemini API, với mô hình được lựa chọn là *gemini-2.5-flash*. Các bình luận được gửi đến API theo từng batch, kèm theo bản tóm tắt của bài viết gốc để cung cấp ngữ cảnh.

Câu prompt (hướng dẫn) để gán nhãn thuộc kiểu *zero-shot prompting*, trong đó mô hình được cung cấp hướng dẫn rõ ràng về nhiệm vụ phân loại 3 nhãn (PHAN_DONG, KHONG_PHAN_DONG, KHONG_LIEN_QUAN) mà không cần các ví dụ cụ thể.

```
"""You are an expert in Vietnamese social-media content moderation and political sentiment analysis.

Your task is to classify Vietnamese comments into three categories based on their political stance:

PHAN_DONG: Comments expressing anti-government, anti-communist, or pro-reactionary views
- Criticizes current Vietnamese government/party/socialist system
- Uses derogatory terms for the government: "việt cộng", "đảng cướp", "cs thối", "độc tài"
- Supports opposing political forces: "VNCH", "dmcs", reactionary groups
- Spreads false information about Vietnamese history or achievements

KHONG_PHAN_DONG: Comments expressing pro-government or anti-reactionary views
- Supports Vietnamese government/party/socialist system
- Defends Vietnamese revolutionary history and achievements
- Criticizes reactionary/anti-government forces
- Uses terms like "phản động", "cali", "vệ quốc", "ba que" to criticize opposition

KHONG_LIEN_QUAN: Comments unrelated to politics
- Sports, entertainment, daily life discussions
- Spam, advertisements, personal disputes
- No political content whatsoever

IMPORTANT RULES:
- Consider the article context when analyzing comments
- When a comment contains both pro and anti-government elements, prioritize the stronger sentiment
- Respond ONLY with a JSON object containing comment classifications
- Use Vietnamese political context and terminology for accurate classification"""
```

Hình 3.2: Prompt hướng dẫn gán nhãn cho Gemini

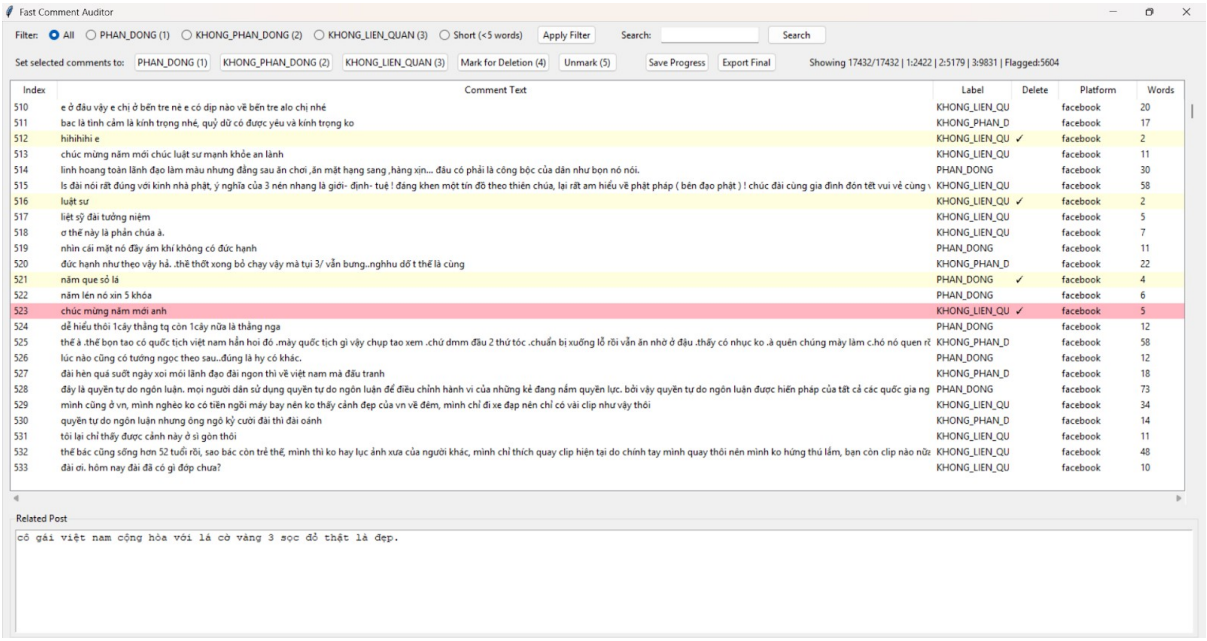
Sau khi nhận kết quả từ API, một bước hậu xử lý sử dụng biểu thức chính quy (regex) được áp dụng để ghi đè các nhãn dựa trên sự hiện diện của các từ khóa chính trị đặc biệt: nếu phát hiện từ khóa chống chính quyền thì gán nhãn PHAN_DONG, nếu phát hiện từ khóa chống phản động (và không có từ khóa chống chính quyền) thì gán nhãn KHONG_PHAN_DONG. Điều này giúp tăng cường độ chính xác cho các trường hợp đặc thù của tiếng Việt mà AI có thể bỏ sót.

```
# Fixed ANTI_GOVT_PATTERNS
ANTI_GOVT_PATTERNS = [
    # Original patterns (longer first)
    r'\b(việt\s*cộng|đảng\s*cướp|độc\s*tài|csvn|xứ\s*vẹm|cộng\s*phỉ|đbrr)\b',
    # Phục quốc
    r'\b(phục\s*quốc)\b',
    # phụt quốc
    r'\b(phụt\s*quốc)\b',
    # Enhanced patterns from JSON map (longer first)
    r'\b(vịt\s*cộng|vịt\s*cọng|bò\s*dát\s*vàng|red\s*bull|cộng\s*sản\s*thổ\s*phỉ)\b',
    r'\b(cộng\s*sả|cọng\s*sả|cạn\s*sổng|cơm\s*sườn|cộng\s*nô|súc\s*nô)\b',
    r'\b(béc\s*hù|hochominh|csthophi|đacosa)\b',

    # Escaped special characters
    r'\b(v\+|việt\+|viet\+|vịt\s*\+)\b',
    r'\b(bò\s*đỏ|bo\s*do|redbull)\b',

    # Short patterns last (more specific context)
    r'\b(cs|béc|đẻng|đẽng)\b'
]

# Fixed ANTI_REACTIONARY_PATTERNS
ANTI_REACTIONARY_PATTERNS = [
    # Original patterns (longer first)
    r'\b(ba\s*que|3\s*que|phản\s*động)\b',
    r'\b(cali)\b.*\b(phản\s*quốc|bán\s*nước)\b',

    # Enhanced patterns (longer first)
    r'\b(3\s*\/\/\/|phổng\s*đạn|bắc\s*kầy|băc\s*kì|bac\s*kì|bac\s*ki|parkầy)\b',
    r'\b(ba\s*kẻ|3\s*gạch|3\s*xẹt|parque|parwe|bac\s*ky)\b',
    r'\b(backy|parky|bakye|bakey|parkey|parke)\b',
    r'\b(barqe|bakue|3soc|becgie|bẹc\s*giê)\b',
    r'\b(ka\s*li|calo|calu|fandong)\b',

    # Short patterns last
    r'\b(3q|3que|\/\/\/|bake|parq|baq|kali|cal|ali)\b'
]
```

Hình 3.3: Các biểu thức chính quy để phát hiện từ khóa chính trị cho Gemini

- **Kiểm tra nhãn thủ công:**

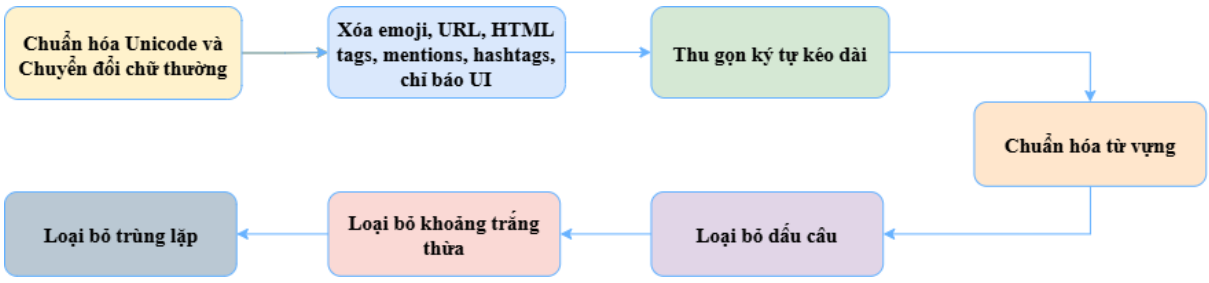
Sau khi quá trình gán nhãn tự động hoàn tất, sinh viên thực hiện kiểm tra bằng công cụ có giao diện đồ họa (GUI) được phát triển bằng thư viện Tkinter và chỉnh sửa nhãn thủ công. Quá trình này đảm bảo chất lượng và độ tin cậy của bộ dữ liệu trước khi đưa vào các bước phân tích sâu hơn.



Hình 3.4: Kiểm tra nhãn thủ công

3.2.2.3. Phương pháp tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một giai đoạn nền tảng trong quy trình xử lý ngôn ngữ tự nhiên (NLP), đóng vai trò cốt lõi trong việc chuyển đổi dữ liệu thô vốn chứa nhiều nhiễu và sự thiếu nhất quán, thành một định dạng sạch và có cấu trúc. Mục đích của quá trình này là loại bỏ các thành phần phi ngữ nghĩa, chuẩn hóa các biến thể từ vựng, và đồng nhất hóa cấu trúc văn bản, qua đó tối ưu hóa dữ liệu đầu vào để các mô hình học máy có thể học và suy luận một cách hiệu quả nhất. Trong công trình này, quy trình tiền xử lý sẽ được thực hiện qua các giai đoạn tuần tự được minh họa chi tiết trong *Hình 3.5*.



Hình 3.5: Các bước tiền xử lý dữ liệu

1. Chuẩn hóa Unicode và Chuyển đổi chữ thường

Tất cả văn bản bình luận được chuyển đổi thành chữ thường để đảm bảo tính nhất quán (ví dụ: "Phản Động" và "phản động" được xem là một). Đồng thời, sinh viên áp dụng chuẩn hóa Unicode NFC (Normalization Form C) để hợp nhất các ký tự có cùng biểu diễn nhưng khác mã Unicode, một vấn đề thường gặp trong tiếng Việt.

2. Loại bỏ các thành phần nhiễu

Nhiều thành phần không mang giá trị ngữ nghĩa trong các bình luận trên mạng xã hội đã được loại bỏ, bao gồm:

+ **Biểu tượng cảm xúc** (Emoji & Emoticons): Xóa bỏ cả emoji tiêu chuẩn và các biểu tượng cảm xúc dạng văn bản (ví dụ: :)), :<).
+ **Đường dẫn, Thẻ HTML, và các tương tác mạng xã hội:** Loại bỏ các URL, thẻ HTML, các lượt đề cập (@username), và hashtags (#hashtag).
+ **Các chỉ báo giao diện** (UI Indicators): Xóa các văn bản do nền tảng tự động thêm vào như "xem thêm", "đã chỉnh sửa", "see more", v.v.

3. Chuẩn hóa các ký tự lặp lại (Reduce Elongated Characters)

Để xử lý hiện tượng người dùng cố ý kéo dài một từ để nhấn mạnh (ví dụ: "ghê quáaaaaa"), chúng tôi chuẩn hóa các chuỗi ký tự bị lặp lại. Bất kỳ ký tự nào lặp lại từ 3 lần trở lên đều được rút gọn về 2 lần (ví dụ: quáaaaaa → quáa).

4. Chuẩn hóa từ vựng (Lexical Normalization)

Sinh viên tự xây dựng từ điển chứa các từ viết tắt, tiếng lóng, và các biến thể teencode để chuẩn hóa văn bản. Quá trình này thay thế các từ không chuẩn bằng dạng chuẩn tương ứng của chúng (ví dụ: ko → không, vncs→ việt nam cộng sản).

5. Loại bỏ dấu câu (Remove Punctuation)

Tất cả các dấu câu và ký tự đặc biệt không cần thiết đều bị loại bỏ. Quy tắc được áp dụng là chỉ giữ lại các ký tự thuộc bảng chữ cái tiếng Việt, tiếng Anh, chữ số và khoảng trắng.

6. Loại bỏ khoảng trắng thừa (Whitespace Stripping)

Mọi khoảng trắng thừa (nhiều hơn một khoảng trắng liên tiếp) được thay thế bằng một khoảng trắng duy nhất. Các khoảng trắng ở đầu và cuối văn bản cũng được cắt bỏ để đảm bảo dữ liệu sạch sẽ.

7. Loại bỏ các bình luận trùng lặp (Deduplication)

Sau khi văn bản đã được làm sạch và chuẩn hóa hoàn toàn cho các bình luận, chúng tôi tiến hành loại bỏ các bình luận có nội dung giống hệt nhau. Bước này giúp giảm thiểu sự thừa thãi dữ liệu và đảm bảo mỗi mẫu trong tập dữ liệu là duy nhất.

Kết quả kì vọng: Sau quá trình tiền xử lý dữ liệu, từ các bình luận gốc, sinh viên thu được bộ dữ liệu cuối cùng được chuẩn hóa đồng nhất và loại bỏ các yếu tố gây nhiễu có thể ảnh hưởng đến độ chính xác của mô hình học máy. Chất lượng của dữ liệu ở bước này là yếu tố tiên quyết, đảm bảo tính khách quan và độ tin cậy của các kết quả đánh giá mô hình về sau.

3.2.3. Phương pháp thực nghiệm mô hình

3.2.3.1. Các mô hình xử lý ngôn ngữ tự nhiên truyền thống

(a) Random Forest Classifier

Random Forest [29] là một thuật toán ensemble learning dựa trên việc kết hợp nhiều cây quyết định (decision trees) để tạo ra một mô hình dự đoán mạnh mẽ và ổn định hơn. Thuật toán này hoạt động bằng cách xây dựng nhiều cây quyết định trên các tập con khác nhau của dữ liệu huấn luyện và sử dụng voting mechanism để đưa ra quyết định cuối cùng. Trong bối cảnh phân loại văn bản, Random Forest có khả năng xử lý hiệu quả các đặc trưng có chiều cao như TF-IDF vectors, đồng thời giảm thiểu overfitting thông qua tính ngẫu nhiên trong quá trình xây dựng cây. Ưu điểm nổi bật của Random Forest bao gồm khả năng xử lý dữ liệu nhiễu, không yêu cầu chuẩn hóa đặc trưng, và cung cấp điểm quan trọng của đặc trưng (feature importance scores) để hiểu được tầm quan trọng của các từ khóa trong quá trình phân loại.

(b) Multinomial Naive Bayes Multinomial

Naive Bayes [10] là một biến thể của thuật toán Naive Bayes được thiết kế đặc biệt cho dữ liệu có đặc trưng rời rạc như word counts hoặc TF-IDF. Thuật toán này dựa trên giả định "ngây thơ" rằng các đặc trưng (từ) độc lập với nhau khi cho trước nhãn lớp, mặc dù giả định này thường không đúng trong thực tế nhưng mô hình vẫn hoạt động hiệu quả trong nhiều tác vụ phân loại văn bản. Multinomial Naive Bayes tính toán xác suất thuộc về mỗi lớp dựa trên tần suất xuất hiện của các từ trong tài liệu, sử dụng định lý Bayes để đưa ra dự đoán cuối cùng. Mô hình này đặc biệt phù hợp cho bài toán phân loại văn bản do tính đơn giản, tốc độ huấn luyện nhanh, và khả năng hoạt động tốt ngay cả với dữ liệu huấn luyện ít.

(c) Logistic Regression

Logistic Regression [30] là một thuật toán học máy tuyến tính sử dụng hàm sigmoid để mô hình hóa xác suất thuộc về các lớp khác nhau. Trong bối cảnh phân loại đa lớp, mô hình sử dụng phương pháp một-so-với-còn-lại hoặc hồi quy logistic đa thức để xử lý nhiều lớp đồng thời. Thuật toán này hoạt động bằng cách học một tập các trọng số (weights) cho mỗi đặc trưng đầu vào, sau đó kết hợp tuyến tính các đặc trưng này để tính toán log-odds và cuối cùng là xác suất thuộc về mỗi lớp. Logistic Regression có ưu điểm là khả năng diễn giải cao, có thể cung cấp xác suất dự đoán cho mỗi lớp, và hoạt động hiệu quả với các đặc trưng tuyến tính như TF-IDF features.

(d) Convolutional Neural Network (CNN)

CNN [12] là một kiến trúc deep learning ban đầu được thiết kế cho xử lý ảnh nhưng đã được chứng minh hiệu quả trong nhiều tác vụ xử lý ngôn ngữ tự nhiên [16]. Trong phân loại văn bản, CNN sử dụng các bộ lọc tích chập với kích thước khác nhau để trích xuất các n-gram features từ chuỗi từ đầu vào. Quá trình này bao gồm việc áp dụng phép tích chập trên từ nhúng, theo sau là gộp cực đại để lựa chọn các đặc trưng quan trọng nhất. CNN có khả năng nắm bắt các mẫu cục bộ trong văn bản như cụm từ hoặc n-gram, đồng thời tính bất biến dịch chuyển giúp mô hình nhận diện các mẫu bất

kể vị trí xuất hiện trong văn bản. Kiến trúc này đặc biệt hiệu quả trong việc phát hiện các mẫu từ khóa và đặc trưng cấp cụm từ quan trọng cho phân loại.

(e) Long Short-Term Memory (LSTM)

LSTM [13] là một kiến trúc mạng nơ-ron hồi quy (RNN) được thiết kế để giải quyết vấn đề vanishing gradient trong RNN truyền thống. LSTM sử dụng cơ chế gates (forget gate, input gate, output gate) để kiểm soát luồng thông tin, cho phép mô hình học và ghi nhớ các dependencies dài hạn trong chuỗi dữ liệu. Trong bối cảnh phân loại văn bản, LSTM xử lý tuần tự từng từ trong câu và duy trì một hidden state chứa thông tin về ngữ cảnh đã thấy trước đó. Điều này cho phép mô hình hiểu được cấu trúc ngữ pháp và semantic dependencies trong văn bản. Bidirectional LSTM, một biến thể của LSTM, xử lý chuỗi theo cả hai hướng (từ trái sang phải và từ phải sang trái) để capture đầy đủ hơn ngữ cảnh xung quanh mỗi từ.

3.2.3.2. Cấu hình thực nghiệm

Các mô hình được cấu hình như sau:

- **Đối với các mô hình học máy truyền thống:** dữ liệu được tiền xử lý sử dụng TfidfVectorizer để chuyển đổi text thành numerical features. Dataset được chia theo tỷ lệ train/test = 90/10 với stratified sampling để đảm bảo phân bố nhãn cân bằng. Random Forest Classifier được cấu hình với 600 estimators và random_state=2025 để đảm bảo tính reproducible. Multinomial Naive Bayes sử dụng cấu hình mặc định, trong khi Logistic Regression được thiết lập với random_state=0 cho consistency.
- **Đối với các mô hình deep learning:** dữ liệu được chia thành train/validation/test với tỷ lệ 80/10/10. Vocabulary được xây dựng với kích thước tối đa 13,000 từ, sequence length được cố định ở 256 tokens. CNN model sử dụng embedding dimension 256, filter sizes [3,4,5] với 100 filters cho mỗi size, và Global Max Pooling để trích xuất features. LSTM model được triển khai với Bidirectional LSTM có 96 units, spatial dropout 0.2, và dropout rates 0.3 cho cả dropout và recurrent dropout. Cả hai mô hình đều sử dụng Adam optimizer với learning rate 1e-4, batch size 32, và early stopping với

patience=2 để tránh overfitting. Để xử lý vấn đề class imbalance, SMOTE (Synthetic Minority Oversampling Technique) được áp dụng với random_state=2025 nhằm cân bằng phân bố các lớp trong tập huấn luyện.

3.2.3.3. Các mô hình pre-trained

(a) PhoBERT (Vietnamese BERT)

PhoBERT [31] là mô hình BERT được pre-train chuyên biệt cho tiếng Việt, được phát triển bởi VinAI Research. Mô hình này sử dụng kiến trúc Transformer encoder với 12 layers, 768 hidden dimensions, và 12 attention heads cho phiên bản base, hoặc 24 layers, 1024 hidden dimensions, và 16 attention heads cho phiên bản large. PhoBERT được pre-train trên một corpus lớn gồm 20GB dữ liệu tiếng Việt từ Wikipedia và các nguồn web khác, sử dụng Masked Language Modeling (MLM) task. Điểm đặc biệt của PhoBERT là việc sử dụng RDRSegmenter để tách từ tiếng Việt trước khi tokenization, giúp mô hình hiểu tốt hơn cấu trúc ngôn ngữ Việt Nam. Mô hình sử dụng SentencePiece tokenizer với vocabulary size 64,000 tokens, được tối ưu hóa đặc biệt cho các đặc thù của tiếng Việt như từ ghép và các biến thể chính tả.

(b) CafeBERT

CafeBERT [32] là một mô hình BERT được fine-tune chuyên biệt cho dữ liệu mạng xã hội tiếng Việt, được phát triển bởi UIT-NLP Lab. Khác với PhoBERT được pre-train trên dữ liệu formal, CafeBERT được huấn luyện trên dữ liệu từ các nền tảng mạng xã hội như Facebook, bao gồm cả các biến thể ngôn ngữ informal như teencode, từ lóng, và các cách viết tắt phổ biến. Mô hình này sử dụng kiến trúc tương tự PhoBERT nhưng với vocabulary được mở rộng để bao gồm các từ và cụm từ đặc trưng của ngôn ngữ mạng xã hội. CafeBERT đặc biệt hiệu quả trong việc xử lý các biểu thức cảm xúc, emoticons, và các pattern ngôn ngữ không chuẩn thường xuất hiện trong bình luận mạng xã hội. Điều này làm cho CafeBERT trở thành lựa chọn phù hợp cho bài toán phát hiện nội dung tuyên truyền chống phá, nơi mà ngôn ngữ thường được mã hóa và sử dụng các biểu thức ngầm.

(c) XLM-RoBERTa

XLM-RoBERTa [33] (Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach) là một mô hình đa ngôn ngữ được phát triển bởi Facebook AI Research. Mô hình này được pre-train trên dữ liệu từ 100 ngôn ngữ khác nhau, bao gồm cả tiếng Việt, với tổng dung lượng 2.5TB text data. XLM-RoBERTa sử dụng kiến trúc RoBERTa (một phiên bản cải tiến của BERT) với các tối ưu hóa như loại bỏ Next Sentence Prediction task, sử dụng dynamic masking, và training với batch size lớn hơn. Mô hình large có 24 layers, 1024 hidden dimensions, và 16 attention heads với vocabulary size 250,000 tokens sử dụng SentencePiece tokenizer. Ưu điểm của XLM-RoBERTa là khả năng transfer learning across languages và hiểu biết về ngữ cảnh đa ngôn ngữ, điều này có thể hữu ích trong việc phát hiện nội dung tuyên truyền chống phá khi có sự pha trộn ngôn ngữ hoặc sử dụng từ ngoại lai.

Cấu hình thực nghiệm

Cả ba mô hình được cấu hình với các tham số huấn luyện tương đồng để đảm bảo tính công bằng trong so sánh. Dataset được chia theo tỷ lệ train/validation/test = 80/10/10 với stratified sampling để duy trì phân bố nhãn cân bằng. Dữ liệu đầu vào được chuẩn bị bằng cách kết hợp summary và comment_clean với các special tokens tương ứng của từng mô hình (CLS và SEP tokens).

PhoBERT được cấu hình với MAX_LENGTH=256, learning_rate=1e-5, batch_size=32 cho training và 64 cho evaluation, num_epochs=5 với early stopping patience=2. Mô hình sử dụng FP16 precision để tối ưu hóa memory usage và gradient_accumulation_steps=1.

CafeBERT sử dụng cấu hình tương tự nhưng với MAX_LENGTH=384 để accommodate các sequence dài hơn thường gặp trong dữ liệu mạng xã hội, batch_size=12 cho training và 24 cho evaluation với gradient_accumulation_steps=3 để maintain effective batch size. Mô hình sử dụng BF16 precision và num_epochs=6.

XLM-RoBERTa được cấu hình với MAX_LENGTH=384, learning_rate=1e-5, batch_size=32 cho training và 64 cho evaluation, num_epochs=6 với FP16 precision. Gradient_accumulation_steps=1 do model capacity lớn.

Cả ba mô hình đều sử dụng warmup_ratio=0.15, weight_decay=0.02, và Adam optimizer. Để xử lý class imbalance, oversampling technique đã được thử áp dụng cho các mô hình (tuy nhiên khiến độ chính xác suy giảm nên đã bị loại bỏ). Evaluation metrics bao gồm accuracy, balanced_accuracy, macro F1-score, weighted F1-score, và per-class precision/recall, với macro F1-score được sử dụng làm metric chính cho model selection thông qua early stopping mechanism.

3.2.3.4. Các mô hình ngôn ngữ lớn (Large Language Models)

(a) Vistral-7B

Vistral-7B [34] là một mô hình ngôn ngữ lớn được phát triển đặc biệt cho tiếng Việt, dựa trên kiến trúc Mistral-7B. Mô hình này có 7 tỷ tham số và được fine-tune chuyên biệt cho các tác vụ trò truyện trong tiếng Việt. Điểm nổi bật của Vistral là việc được huấn luyện trên một lượng lớn dữ liệu tiếng Việt chất lượng cao, bao gồm cả dữ liệu formal và informal, giúp mô hình hiểu tốt các biểu thức địa phương và ngữ cảnh văn hóa Việt Nam. Mô hình hỗ trợ context window lên đến 32K tokens và được tối ưu hóa cho các tác vụ instruction-following, làm cho nó phù hợp cho việc phân loại nội dung dựa trên hướng dẫn chi tiết.

(b) SeaLLMs-v3-7B

SeaLLMs-v3-7B [35] là phiên bản thứ ba của dòng mô hình SeaLLMs, được thiết kế đặc biệt cho các ngôn ngữ Đông Nam Á, bao gồm tiếng Việt. Mô hình này có 7 tỷ tham số và được phát triển dựa trên kiến trúc Llama-2 với những cải tiến đáng kể cho multilingual performance. SeaLLMs-v3 được pre-train trên một dataset đa ngôn ngữ lớn với sự tập trung đặc biệt vào các ngôn ngữ Đông Nam Á, sau đó được fine-tune với instruction data chất lượng cao. Điểm mạnh của mô hình này là khả năng cross-lingual understanding, cho phép nó hiểu được các sắc thái văn hóa và ngôn ngữ đặc trưng của khu vực. Mô hình sử dụng advanced tokenization strategies để xử lý hiệu quả các đặc điểm hình thái phức tạp của tiếng Việt và các ngôn ngữ tương tự trong khu vực.

(c) Qwen3-32B

Qwen3-32B [36] là một mô hình ngôn ngữ lớn với 32 tỷ tham số, thuộc dòng Qwen3 được phát triển bởi Alibaba Cloud. Mô hình này sử dụng kiến trúc Transformer tiên tiến với các cải tiến về cơ chế chú ý (attention mechanism) và kỹ thuật chuẩn hóa (normalization techniques). Qwen3-32B được pre-train trên một corpus khổng lồ bao gồm dữ liệu từ nhiều ngôn ngữ, trong đó có tiếng Việt, với tổng dung lượng hàng nghìn tỷ tokens. Mô hình này nổi bật với khả năng suy luận phức tạp và hiểu biết sâu về ngữ cảnh, nhờ vào kiến trúc quy mô lớn và kĩ thuật huấn luyện tiên tiến như curriculum learning và mixture of experts. Qwen2332B hỗ trợ context window lên đến 128K tokens và được tối ưu hóa cho đa dạng tác vụ đích thông qua extensive instruction tuning. Mô hình này đặc biệt hiệu quả trong việc xử lý các tác vụ yêu cầu suy luận logic phức tạp và hiểu biết ngữ cảnh sâu sắc.

Cấu hình thực nghiệm

Cả ba mô hình được triển khai với phương pháp fine-tuning sử dụng QLoRA (Quantized Low-Rank Adaptation) [37] do hạn chế phần cứng. Dataset được chia theo tỷ lệ train/validation/test = 80/10/10 với oversampling technique áp dụng cho minority class "PHAN_DONG" để cân bằng phân bố dữ liệu.

Vistral-7B-Chat được cấu hình với 4-bit quantization sử dụng BitsAndBytesConfig, MAX_SEQ_LENGTH_SFT=1024 tokens dựa trên phân tích 98th percentile của sequence lengths. LoRA configuration bao gồm rank r=32, alpha=64, dropout=0.02, targeting các linear layers chính (q_proj, k_proj, v_proj, o_proj, up_proj, down_proj, gate_proj). Training arguments: learning_rate=3e-5, batch_size=32 cho training và 64 cho evaluation, num_epochs=2 với early stopping patience=2. Mô hình sử dụng paged_adamw_8bit optimizer và gradient checkpointing để tối ưu memory.

SeaLLMs-v3-7B sử dụng cấu hình tương tự với 4-bit quantization và LoRA parameters giống hệt Vistral. Training setup: learning_rate=6e-5, batch_size=24 cho training và 48 cho evaluation với gradient_accumulation_steps=2, num_epochs=2. Mô hình được train với BF16 precision và sử dụng Unsloth framework để tăng tốc quá trình training và inference.

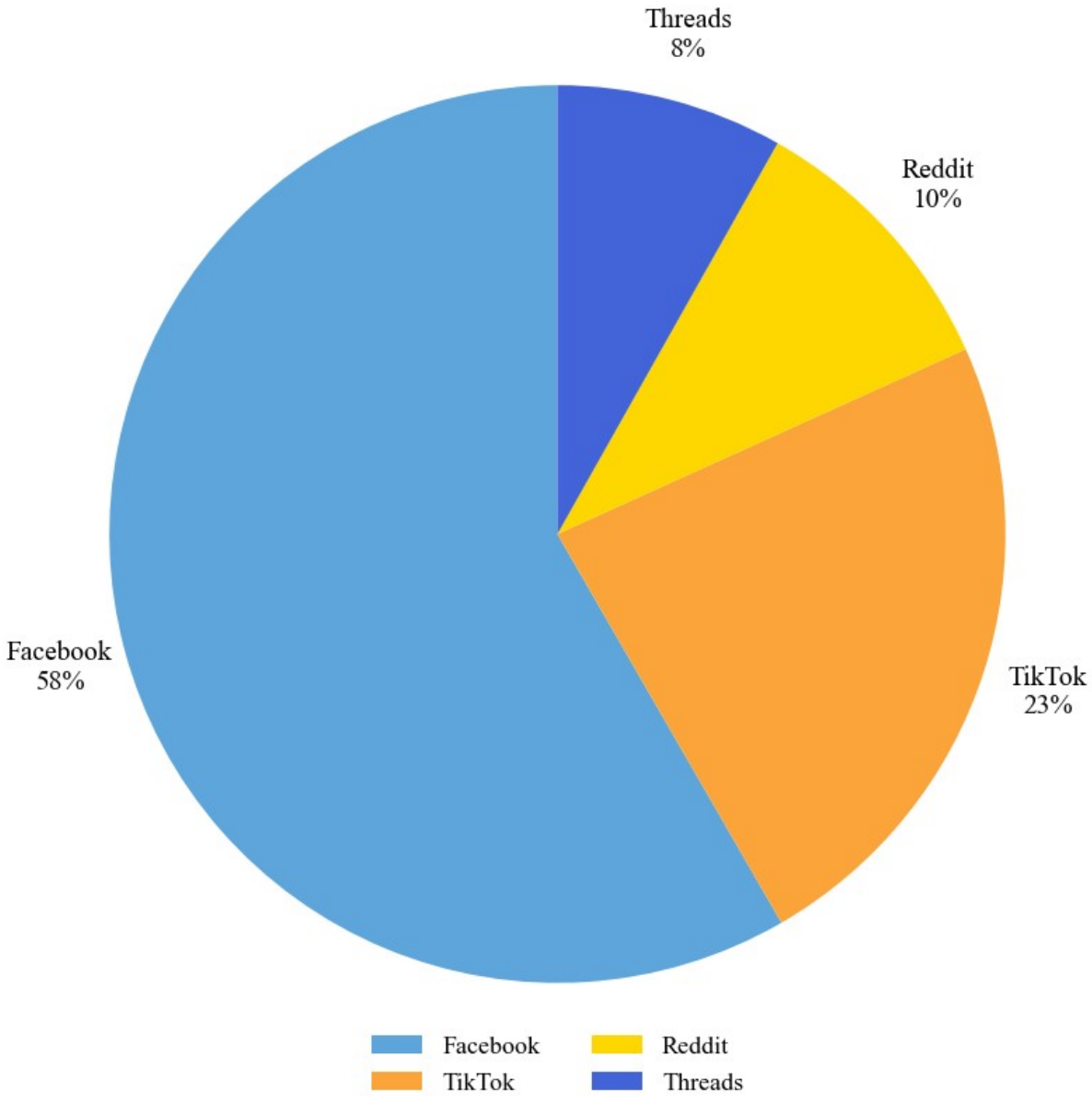
Qwen3-32B do kích thước lớn hơn được cấu hình với memory-efficient settings: batch_size=16 cho training và 32 cho evaluation với gradient_accumulation_steps=2,learning_rate=6e-5, MAX_SEQ_LENGTH_SFT=650 tokens để phù hợp với memory constraints. Mô hình sử dụng BF16 precision và extensive gradient checkpointing. Định dạng gợi nhắc Alpaca (Alpaca prompt format) được sử dụng thay vì chat template để tối ưu hóa hiệu suất.

Đối với quá trình suy luận (inference) và đánh giá (evaluation), cả ba mô hình đều được test với zero-shot learning trước khi fine-tuning và few-shot learning sau fine-tuning. Generation parameters bao gồm: max_new_tokens=20 cho classification task, temperature=0.1, top_p=0.95, top_k=40, repetition_penalty=1.05. Việc trích xuất nhãn được thực hiện thông qua phương pháp khớp mẫu (pattern matching) với các biến thể của POSSIBLE_LABELS để đảm bảo khả năng phân tích dự đoán một cách ổn định. Các chỉ số đánh giá bao gồm accuracy, precision, recall, F1-score cho từng lớp và hiệu suất tổng thể, với detailed classification reports để phân tích hiệu suất trên từng nhãn cụ thể.

Chương 4. KẾT QUẢ - THẢO LUẬN

4.1. Kết quả bộ dữ liệu thu thập

Sau khi hoàn tất quá trình thu thập, sinh viên xây dựng được một bộ dữ liệu thô về bình luận có nội dung chống phá Nhà nước trên mạng xã hội tiếng Việt với 19,280 bình luận, với tỉ lệ bình luận trên các trang mạng xã hội như *Hình 4.1* minh hoạ.



Hình 4.1: Minh họa tỉ lệ bình luận thuộc các trang mạng xã hội trong tập dữ liệu

Sau quá trình tiền xử lý dữ liệu, từ 19,280 bình luận gốc, sinh viên thu được bộ dữ liệu cuối cùng gồm 18,912 mẫu duy nhất. Phân tích cho thấy sự phân bổ nhãn mất

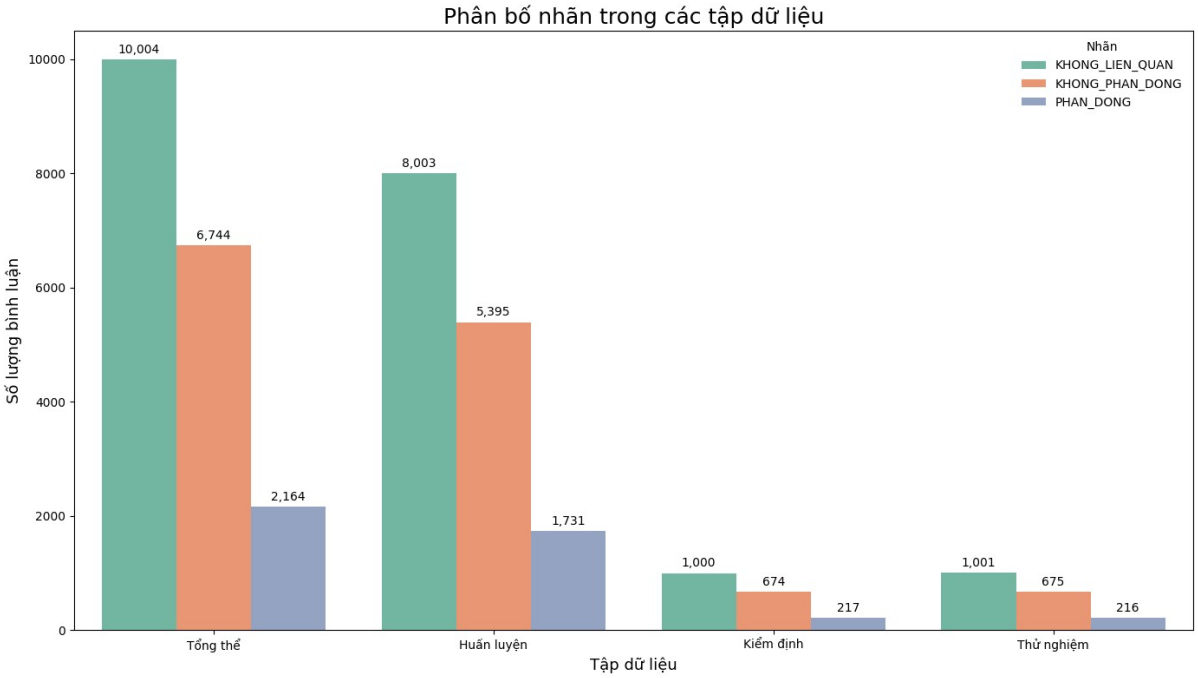
cân bằng đáng kể, với lớp KHONG_LIEN_QUAN chiếm 52.9%, KHONG_PHAN_DONG chiếm 35.7%, và lớp thiểu số PHAN_DONG chỉ chiếm 11.4%.

Nhận thấy sự mất cân bằng nghiêm trọng giữa các nhãn, đặc biệt là với lớp thiểu số PHAN_DONG. Để giải quyết vấn đề này, nhóm sinh viên đã thử nghiệm các kỹ thuật cân bằng dữ liệu khác nhau, phù hợp với từng loại mô hình.

- **Đối với các mô hình học máy truyền thống (như Logistic Regression, Random Forest)**: Do các mô hình này hoạt động trên không gian đặc trưng số hóa (ví dụ: TF-IDF), kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) đã được áp dụng. Phương pháp này tạo ra các mẫu dữ liệu tổng hợp mới cho lớp thiểu số.
- **Đối với các mô hình học sâu và Transformer:** Do các mô hình này làm việc trực tiếp với dữ liệu văn bản, một phương pháp oversampling đơn giản hơn (sao chép các mẫu thuộc lớp thiểu số) đã được sử dụng.

Tuy nhiên, kết quả từ các đợt huấn luyện sơ bộ cho thấy cả hai phương pháp đều không mang lại sự cải thiện hiệu suất rõ rệt trên tập kiểm định (validation set). Do đó, với mục tiêu xây dựng một mô hình có khả năng hoạt động hiệu quả trên dữ liệu thực tế, nhóm sinh viên đã quyết định giữ nguyên sự phân bổ nhãn gốc của dữ liệu và tập trung vào việc lựa chọn kiến trúc mô hình có khả năng xử lý tốt sự mất cân bằng này. Bộ dữ liệu cuối cùng gồm 18,912 mẫu duy nhất được phân chia theo tỉ lệ 8:1:1 lần lượt cho các tập huấn luyện, kiểm định và thử nghiệm.

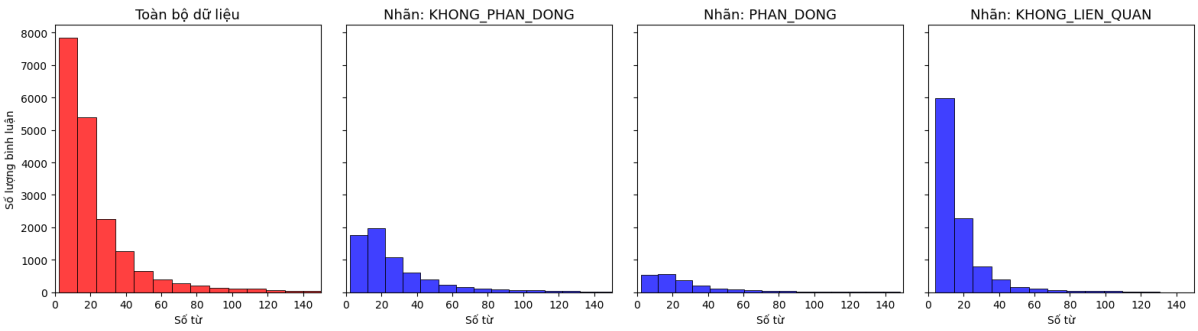
- **Tập huấn luyện (train):** 15,130 mẫu
- **Tập kiểm định (validation):** 1,891 mẫu
- **Tập thử nghiệm (test):** 1,891 mẫu



Hình 4.2: Phân bố nhãn trong các tập dữ liệu sau tiền xử lý dữ liệu

Sơ lược thông tin cơ bản về bộ dữ liệu:

Trước đó trong quá trình kiểm tra thủ công nhãn, sinh viên nhận thấy các bình luận thuộc nhãn KHONG_LIEN_QUAN hầu hết đều có độ dài bình luận rất ngắn (30 từ trở xuống) và chiếm phần lớn bộ dữ liệu. Vì thế, sinh viên tiến hành kiểm tra độ dài bình luận ở từng nhãn.



Hình 4.3: Độ dài bình luận ở từng nhãn

Một thách thức lớn trong bài toán này là việc người dùng trên mạng xã hội thường xuyên sử dụng ngôn ngữ phi chuẩn (viết tắt, teencode) để tránh các bộ lọc tự

động. Phân tích quá trình chuẩn hóa từ vựng cho thấy hiện tượng này rất phổ biến với các từ khóa chính trị (Xem **Bảng 4.1**).

Bảng 4.1: Danh sách các biến thể phổ biến trong bộ dữ liệu

Từ khóa gốc	**Các biến thể được tìm thấy trong dữ liệu**
cộng sản	cs, c.s, +s, +sản, +san, congsan, cộng sả, cọng sả, cạn sổng, cơm sườn, csản, csãn, csa, cộng sản, cộng sảng, cờ sờ
bắc kỳ	paky, parkầy, bắc kầy, bake, 3ke, 3kue, bakky, backy, backycho, parky, bakye, bakey, parkey, parke, bac ky, bac kì, băc kì
ba que	3///, 3que, 3 què, 3 quèe, 3q, 3queh, 3 quê, bá qué, bá que, ba qué, ba quế, ba quễ, baq, ba què, parque
việt nam cộng hòa	vnch, vịt ngan cọng hành
việt cộng	vc, v+, việt+, viet+, vịt cộng, vịt +, vit cong, vịt cọng, zic kọng
bò đỏ	podo, bo đo, bò đo, bò dỏ, bò dỏa, bò đỏa, pê hường, pê hường, bodo, bo do, rebull, red bull, redbull
phản động	fandong, phổng đạn

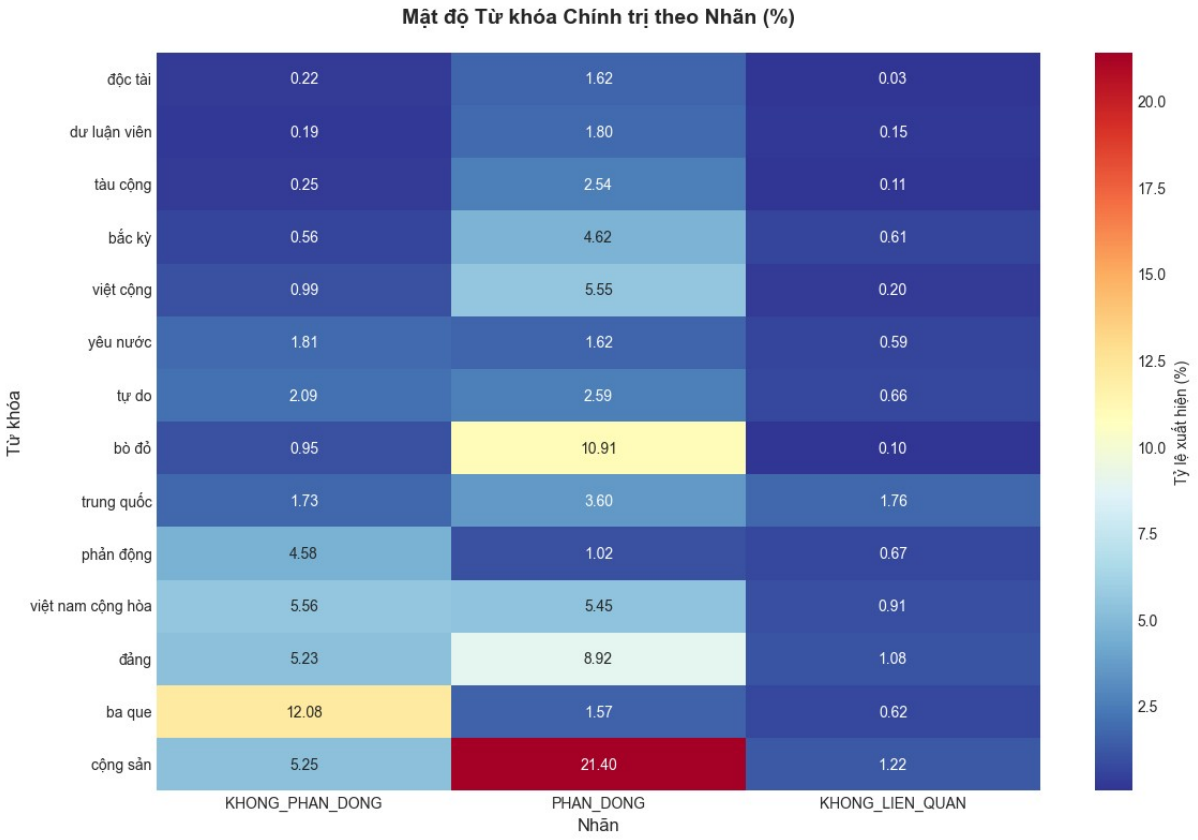
Trong quá trình kiểm tra nhãn thủ công, sinh viên thực hiện ghi chú lại bộ từ điển chuẩn hoá bao gồm 1,247 mục, tổng hợp đa dạng các biến thể từ vựng liên quan đến các từ khóa chính trị thường xuất hiện trong bộ dữ liệu. Quá trình chuẩn hóa từ vựng (lexical normalization) đã được áp dụng thành công cho 53,53% tổng số bình luận (10.321 trên tổng số 19.280 bình luận gốc), cho thấy mức độ ảnh hưởng đáng kể của hiện tượng ngôn ngữ phi chuẩn của các bình luận thường thấy trên mạng xã hội tiếng Việt.

Một số ví dụ điển hình:

- "cs" (489 trường hợp) được chuyển thành "cộng sản" (tăng từ 486 lên 988).
- "vnch" (549 trường hợp) được chuẩn hóa thành "việt nam cộng hòa" (tăng từ 43 lên 591).
- "vc" (131 trường hợp) chuyển thành "việt cộng" (tăng từ 110 lên 210).

Đáng chú ý, từ khóa "ba que" sau chuẩn hóa đã tăng từ 206 lên 950 lượt xuất hiện (+744), cho thấy hệ thống đã nhận diện và hợp nhất hiệu quả các biến thể như “3///”, “3que”, “parque”, v.v. Kết quả này cho thấy bước chuẩn hóa từ vựng đóng vai trò quan trọng trong việc nâng cao độ chính xác của mô hình phân loại, giúp tập trung và phân bổ đúng các từ khóa theo nhãn mục tiêu.

Sau quá trình tiền xử lý, chuẩn hoá từ vựng, các từ vựng chính trị nhạy cảm được chuẩn hoá đã xuất hiện rõ rệt hơn. Để làm rõ hơn về tần suất xuất hiện các từ khoá này, sinh viên thực hiện phân tích tần suất xuất hiện của các từ này như *Hình 4.4.*



Hình 4.4: Mật độ sử dụng các từ khoá chính trị theo nhãn

Có thể thấy rõ sự khác biệt ở từng nhãn, với các từ khoá chính trị đặc trưng ở từng nhãn như sau:

- Từ khóa đặc trưng ở nhãn PHAN_DONG: 'bò đỏ', 'việt cộng', 'bắc kỳ'
- Từ khóa đặc trưng ở nhãn KHONG_PHAN_DONG: 'ba que', 'phản động'
- Từ khóa đặc trưng ở nhãn KHONG_LIEN_QUAN: 'trung quốc', 'tự do', 'yêu nước'

4.2. Kết quả thực nghiệm

4.2.1. Các chỉ số đánh giá

Để đánh giá và so sánh hiệu năng của các mô hình một cách khách quan, công trình sử dụng các chỉ số đo lường tiêu chuẩn trong bài toán phân loại văn bản. Do tính chất mất cân bằng của dữ liệu gốc, việc đánh giá không chỉ dựa vào độ chính xác tổng thể (Accuracy) mà tập trung chủ yếu vào các chỉ số Precision, Recall và F1-Score trên

từng lớp. Trong đó, F1-Score của lớp thiểu số PHAN_DONG được xem là thước đo quan trọng nhất, phản ánh năng lực thực sự của mô hình trong việc phát hiện các nội dung cần quan tâm.

- **True Positive (TP):** Số mẫu được mô hình dự đoán đúng.
- **False Positive (FP):** Số mẫu bị mô hình dự đoán nhầm vào lớp này.
- **False Negative (FN):** Số mẫu của lớp này bị mô hình bỏ sót.
- **True Negative (TN):** Số mẫu không thuộc lớp này và được mô hình dự đoán đúng.

Dựa trên các giá trị này, các chỉ số đánh giá chính được định nghĩa như sau:

Precision (Độ chính xác): Đo lường mức độ đáng tin cậy của các dự đoán.

$$Precision = \frac{TP}{TP + FP}$$

Recall (Độ phủ): Đo lường khả năng của mô hình trong việc phát hiện tất cả các mẫu của một lớp.

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

F1-score: Trung bình điều hòa của Precision và Recall, cân bằng cả hai yếu tố.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Do sự mất cân bằng dữ liệu, đặc biệt là với lớp PHAN_DONG (chiếm ~11.4%), các chỉ số F1-macro và F1-Score cho từng lớp sẽ được sử dụng làm thước đo chính để đánh giá hiệu quả, bên cạnh độ chính xác (Accuracy) tổng thể.

4.2.2. Hiệu năng thực nghiệm của các mô hình

4.2.2.1. Các mô hình học máy truyền thống và học sâu

Nhóm này bao gồm các mô hình học máy cổ điển và các kiến trúc học sâu tuần tự ban đầu. Chúng được sử dụng để thiết lập một ngưỡng hiệu suất cơ bản (baseline), làm cơ sở để đánh giá mức độ cải thiện của các phương pháp phức tạp hơn.

Bảng 4.2: Kết quả của các mô hình học máy truyền thống và học sâu

Models	F1_PD	F1_KPD	F1_KLQ	Accuracy	F1-macro
Random Forest	0.33	0.62	0.79	0.70	0.58
Multinomial Naive Bayes	0.07	0.56	0.77	0.66	0.47
Logistic Regression	**0.45**	0.65	**0.80**	**0.72**	**0.63**
CNN	0.42	**0.67**	0.79	0.71	0.62
LSTM	0.41	0.63	0.78	0.69	0.61

Kết quả từ Bảng 4.1 cho thấy các mô hình học máy truyền thống và học sâu gặp nhiều khó khăn trong việc nhận diện lớp PHAN_DONG. Đặc biệt là mô hình Multinomial Naive Bayes với F1-PD chỉ 0.07, gần như không có khả năng nhận diện. Các mô hình học sâu tuần tự như CNN (F1-PD=0.42) và LSTM (F1-PD=0.41) cho thấy sự cải thiện nhưng không đáng kể. Logistic Regression là mô hình hoạt động tốt nhất trong nhóm này, đạt F1-PD là 0.45. Nhìn chung, các mô hình trong nhóm này đều chưa đủ khả năng nắm bắt các sắc thái ngữ nghĩa phức tạp để nhận diện hiệu quả lớp thiểu số PHAN_DONG.

4.2.2.2. Các mô hình Transformer và Ngôn ngữ lớn

Nhóm này bao gồm các mô hình dựa trên kiến trúc Transformer và các Mô hình Ngôn ngữ Lớn đã được tinh chỉnh (fine-tuned) trên bộ dữ liệu của công trình. Đây là nhóm đại diện cho các phương pháp tiên tiến nhất hiện nay.

Bảng 4.3: Kết quả của các mô hình Transformer và LLMs (Fine-tuned)

Models	F1_PD	F1_KPD	F1_KLQ	Accuracy	F1-macro
PhoBERT	0.54	0.69	0.82	0.75	0.68
CafeBERT	0.58	0.70	0.82	0.75	0.70
XLM-RoBERTa	0.55	**0.72**	**0.82**	0.76	0.70
Vistral-7B	**0.69**	0.71	0.82	**0.76**	**0.74**
SeaLLMv3-7B	0.61	0.64	0.81	0.73	0.68
Qwen3-32B	0.54	0.59	0.78	0.68	0.64

Phân tích bảng kết quả cho thấy Vistral-7B là mô hình có hiệu năng tổng thể vượt trội nhất, khi dẫn đầu ở các chỉ số quan trọng nhất bao gồm F1-macro (0.74),

Accuracy (0.76), và đặc biệt là F1-score cho lớp PD (0.69). Mặc dù vậy, mô hình XLM-RoBERTa lại chứng tỏ sự hiệu quả vượt trội trong việc nhận diện hai lớp cụ thể là KPD và KLQ, với điểm F1-score cao nhất lần lượt là 0.72 và 0.82. Các mô hình khác như CafeBERT cũng cho thấy hiệu suất cạnh tranh với điểm F1-macro là 0.70, ngang bằng với XLM-RoBERTa. Một điểm đáng chú ý là Qwen3-32B, dù có thể là mô hình lớn nhất, lại cho kết quả tổng thể thấp nhất trong nhóm (F1-macro=0.64). Nhìn chung, các kết quả này khẳng định Vistral-7B là lựa chọn tối ưu nhất cho bài toán toàn diện, đồng thời nhấn mạnh rằng việc tối ưu hóa mô hình cho ngôn ngữ và tác vụ cụ thể quan trọng hơn là chỉ dựa vào quy mô của mô hình.

4.2.2.3. Zero-shot prompting trên các mô hình ngôn ngữ lớn

Bảng 4.4: Kết quả của các mô hình LLMs Zero-shot

Method	Model	F1_PD	F1_KPD	F1_KLQ	Accuracy	F1-macro
Zero-shot	Vistral-7B	0.30	0.33	0.55	0.41	0.39
Fine-tuned	Vistral-7B	**0.69**	**0.71**	**0.82**	**0.76**	**0.74**
Zero-shot	Qwen-32B	0.33	0.07	0.72	0.48	0.37
Fine-tuned	Qwen-32B	**0.54**	**0.59**	**0.78**	**0.68**	**0.64**
Zero-shot	SeaLLM-7B	*Can't follow the instruction*				
Fine-tuned	SeaLLM-7B	**0.61**	**0.64**	**0.81**	**0.73**	**0.68**

Việc áp dụng các Mô hình Ngôn ngữ Lớn (LLMs) theo phương pháp zero-shot cho các bài toán phân loại chuyên biệt vốn tiềm ẩn nhiều thách thức. Mặc dù sở hữu kho kiến thức khổng lồ, khả năng suy luận của các mô hình trong một lĩnh vực hẹp mà không qua huấn luyện đặc thù vẫn là một dấu hỏi lớn về tính hiệu quả và độ tin cậy.

Kết quả thực nghiệm của công trình đã xác thực mạnh mẽ cho nhận định này. Dữ liệu cho thấy một sự chênh lệch hiệu suất rất lớn giữa hai phương pháp. Cụ thể, Vistral-7B và Qwen-32B đều cho hiệu suất rất thấp ở kịch bản zero-shot (F1-macro lần lượt là 0.39 và 0.37), nhưng đã có sự cải thiện vượt bậc sau khi được tinh chỉnh (F1-macro tăng lên 0.74 và 0.64). Minh chứng rõ ràng nhất là trường hợp của SeaLLM-7B, mô hình đã hoàn toàn không thể tuân theo chỉ dẫn ở chế độ zero-shot nhưng lại hoạt động hiệu quả sau khi được tinh chỉnh, đạt F1-macro 0.68.

Từ những bằng chứng thực nghiệm trên, có thể kết luận rằng phương pháp zero-shot không phải là một hướng tiếp cận khả thi cho bài toán này. Để các Mô hình Ngôn ngữ Lớn có thể giải quyết hiệu quả các tác vụ đòi hỏi sự am hiểu sâu sắc về ngữ cảnh và sắc thái, quá trình tinh chỉnh trên dữ liệu đặc thù không chỉ là một sự tối ưu hóa, mà là một yêu cầu bắt buộc để đảm bảo hiệu năng và độ tin cậy của mô hình.

4.3. Thảo luận

Kết quả thực nghiệm của công trình không chỉ đo lường hiệu năng của các mô hình mà còn cung cấp những góc nhìn sâu sắc về các yếu tố then chốt quyết định thành công trong một bài toán phức tạp như nhận diện nội dung chống phá Nhà nước. Phần thảo luận này sẽ đi vào phân tích sự phân cấp hiệu năng giữa các kiến trúc mô hình, qua đó làm rõ vai trò của việc tinh chỉnh chuyên biệt và các thách thức từ dữ liệu, đồng thời khám phá tiềm năng của các mô hình ngôn ngữ lớn (LLMs) trong việc xây dựng hệ thống AI có thể giải thích được.

4.3.1. Phân tích hiệu năng và các yếu tố ảnh hưởng

Kết quả thực nghiệm đã cho thấy một sự phân cấp hiệu năng rõ rệt, bắt nguồn từ chính những thách thức của bộ dữ liệu. Với đặc tính mất cân bằng nghiêm trọng (lớp `PHAN_DONG` chỉ chiếm 11.4%) và sự phức tạp trong ngôn ngữ (hơn 53% bình luận cần chuẩn hóa từ vựng, các mô hình học máy truyền thống đã bộc lộ hạn chế cố hữu. Việc chúng chỉ dựa vào các đặc trưng bề mặt như tần suất từ khóa đã dẫn đến thất bại, đặc biệt là với mô hình Multinomial Naive Bayes (F1-score cho lớp `PHAN_DONG` chỉ 0.07), vì không thể nắm bắt được các nội dung tinh vi được "mã hóa" bằng từ lóng, ẩn dụ và teencode. Ngay cả CNN và LSTM, dù có cải thiện, vẫn chưa đủ sức mạnh để xử lý các mối quan hệ ngữ nghĩa phức tạp này.

Chính trong bối cảnh đó, sự trỗi dậy của kiến trúc Transformer và các Mô hình Ngôn ngữ Lớn (LLMs) đã mang lại một bước đột phá. Tuy nhiên, để khai thác triệt để tiềm năng này, công trình chỉ ra rằng sự chuyên biệt hóa quan trọng hơn quy mô Điều này được thể hiện rõ qua việc Vistral-7B, một mô hình được tinh chỉnh sâu cho tiếng Việt, đã đạt hiệu suất vượt trội (F1-score cho lớp `PHAN_DONG` là 0.69), trong khi

Qwen-32B, dù lớn hơn nhiều, lại cho kết quả khiêm tốn hơn (F1-macro 0.64). Thành công của Vistral-7B củng cố giả thuyết rằng việc am hiểu sâu sắc ngữ cảnh văn hóa-chính trị và các biến thể ngôn ngữ địa phương là yếu tố quyết định.

Hơn nữa, công trình khẳng định rằng quá trình tinh chỉnh (fine-tuning) là một yêu cầu bắt buộc. Sự chênh lệch hiệu suất khổng lồ giữa phương pháp zero-shot (F1-macro của Vistral-7B là 0.39) và fine-tuned (F1-macro 0.74) đã chứng minh rằng kiến thức nền của LLMs là không đủ cho một tác vụ hẹp và nhạy cảm như vậy. Trường hợp của SeaLLM-7B, từ chỗ hoàn toàn không thể tuân theo chỉ dẫn ở chế độ zero-shot đến việc hoạt động hiệu quả sau khi tinh chỉnh, là minh chứng rõ ràng nhất. Quá trình tinh chỉnh chính là bước trang bị cho mô hình kiến thức chuyên sâu cần thiết để giải mã các sắc thái ngôn ngữ đặc thù của bài toán.

4.3.2. Tiềm năng của LLMs cho Hệ thống Phát hiện có thể Giải thích được (Explainable AI - XAI)

Vượt xa hơn khả năng phân loại chính xác, việc ứng dụng LLMs còn mở ra một hướng đi quan trọng cho các hệ thống kiểm duyệt trong thực tiễn: khả năng giải thích quyết định. Khác với các mô hình truyền thống vốn là những "hộp đen" chỉ đưa ra dự đoán, LLMs có thể được thiết kế để cung cấp một đoạn văn lý giải logic đằng sau lựa chọn đó. Tiềm năng này là nền tảng để xây dựng các hệ thống AI có thể giải thích được (Explainable AI - XAI).

Post	Comment	True Label	Predict	LLMs Explain
Ba cựu cán bộ bị khai trừ khỏi Đảng Cộng sản Việt Nam vì suy thoái tư tưởng, đạo đức, vi phạm pháp luật và tham nhũng, gây ảnh hưởng xấu đến uy tín của Đảng và chính quyền.	nhân dân đóng thuế nuôi lũ kền kền nếu bắt thì không cán bộ nào thoát vì đã là quan thì thằng nào cũng tham	PHAN DONG	PHAN DONG	Bình luận sử dụng ngôn ngữ mỉa mai ("kền kền") và khẳng định tất cả cán bộ đều tham nhũng, xuyên tạc chế độ và thể chế chính trị của Việt Nam, thể hiện thái độ chống phá nhà nước. Mặc dù có yếu tố chỉ trích tham nhũng, nhưng cách diễn đạt mang tính chất kích động, bôi nhọ hệ thống chính trị.
Bài viết ca ngợi vẻ đẹp của áo dài và cờ vàng ba sọc đỏ của Việt Nam Cộng Hòa, đồng thời lên án chế độ cộng sản Việt Nam, kêu gọi lật đổ và khẳng định sự phổ biến của cờ vàng ở nước ngoài.	úi bóng ma ngởm 50 năm rồi mà chưa siêu thoát tầm bậy còn thiếu hồn ma diệm nhu thiệu và đám đệ tử	KHONG PHAN DONG	KHONG PHAN DONG	Bình luận sử dụng ngôn ngữ miệt thị ("bóng ma ngởm", "tầm bậy") để chê bai những người ủng hộ chế độ Việt Nam Cộng Hòa. Mặc dù có tính chất công kích, nhưng nó không trực tiếp thể hiện quan điểm chống phá chính quyền hiện tại hay kêu gọi lật đổ. Thay vào đó, bình luận tập trung vào việc bôi nhọ đối phương. Do đó, nó không mang tính chất phản động đối với nhà nước Việt Nam.
Câu hỏi về việc Việt Nam có nên khoan hồng đón nhận những người từng là người Việt Nam.	ai bị treo tiền nhắn tin em chỉ chỗ làm lấy tiền lại nha em cũng mới nhân tiền về hôm qua	KHONG LIEN QUAN	KHONG LIEN QUAN	Bình luận này nói về việc treo tiền và yêu cầu liên hệ để lấy lại tiền, không liên quan đến nội dung chính của bài đăng về sự khoan dung đối với người Việt ở nước ngoài hoặc các vấn đề chính trị. Nó hoàn toàn tách biệt và không mang tính chất phản động.

Hình 4.5: Một số một số phản hồi được trích xuất từ mô hình Qwen3-32B

Những giải thích này có thể đóng vai trò như một "trợ lý AI" đắc lực, giúp người kiểm duyệt nhanh chóng nắm bắt lý do tại sao một bình luận bị gắn cờ, từ đó tăng tốc độ và độ chính xác của quyết định cuối cùng. Điều này hứa hẹn sẽ chuyển đổi các hệ thống kiểm duyệt từ công cụ thụ động thành một đối tác tương tác, minh bạch và hiệu quả hơn trong việc bảo vệ không gian mạng.

Chương 5. KẾT LUẬN - ĐỀ NGHỊ

5.1. Kết luận

Công trình này đã thực hiện một đánh giá thực nghiệm toàn diện và có hệ thống về việc áp dụng các mô hình xử lý ngôn ngữ tự nhiên hiện đại vào bài toán phát hiện nội dung tuyên truyền chống phá nhà nước trên mạng xã hội Việt Nam. Thông qua việc so sánh hiệu quả của các họ mô hình khác nhau trên một bộ dữ liệu chuyên biệt tự xây dựng, công trìnhtrình đã rút ra được những kết luận khoa học quan trọng.

Thứ nhất, kết quả thực nghiệm đã khẳng định một cách rõ ràng sự vượt trội của các kiến trúc dựa trên Transformer so với các phương pháp học máy truyền thống và học sâu tuần tự. Sự chênh lệch đáng kể về hiệu suất, đặc biệt trên lớp nội dung PHAN_DONG, cho thấy khả năng hiểu ngữ cảnh sâu của cơ chế tự chú ý là yếu tố then chốt để giải quyết bài toán phức tạp này.

Thứ hai, phát hiện cốt lõi và quan trọng nhất của công trình là giá trị của sự chuyên biệt hóa ngôn ngữ và dữ liệu. Mô hình Vistral-7B, một Mô hình Ngôn ngữ lớn được tối ưu hóa cho tiếng Việt, đã đạt hiệu suất cao nhất (F1-score là 0.69 trên lớp PHAN_DONG), vượt qua cả các mô hình Transformer chuyên biệt có hiệu suất cao khác như CafeBERT (0.58) và các LLMs đa ngôn ngữ có quy mô lớn hơn như Qwen3-32B (0.54). Điều này cung cấp bằng chứng thực nghiệm mạnh mẽ rằng, đối với một tác vụ có độ nhạy cảm cao về văn hóa và chính trị, sự am hiểu sâu sắc về ngôn ngữ và bối cảnh (đạt được thông qua tiền huấn luyện chuyên biệt) quan trọng hơn quy mô tuyệt đối của mô hình.

Cuối cùng, công trình cũng cho thấy phương pháp học không cần giám sát (zero-shot learning) chưa đủ tin cậy để triển khai trong các ứng dụng thực tế cho bài toán này. Hiệu suất thấp và tỷ lệ cảnh báo sai cao đã khẳng định rằng, mặc dù có khả năng suy luận đáng kinh ngạc, các LLMs vẫn cần được tinh chỉnh (fine-tuning) trên một bộ dữ liệu chất lượng cao để có thể hoạt động một cách chính xác và đáng tin cậy.

5.2. Đề nghị và hướng phát triển trong tương lai

Dựa trên các kết quả và hạn chế của công trình, nhóm sinh viên đề xuất một số hướng phát triển trong tương lai:

- Nâng cao chất lượng nhãn: Do bộ dữ liệu hiện tại được gán nhãn dựa trên quy trình có sự tham gia của một người, tính chủ quan trong việc gán nhãn vẫn có thể tồn tại. Hướng phát triển tiếp theo cần triển khai quy trình gán nhãn với nhiều người (multi-annotator) và tính toán các chỉ số thống nhất giữa những người gán nhãn (Inter-Annotator Agreement - IAA) như Fleiss' Kappa để đảm bảo độ tin cậy và khách quan của bộ dữ liệu.
- Mở rộng và làm phong phú bộ dữ liệu: Tiếp tục mở rộng quy mô, đặc biệt là cho lớp PHAN_DONG. Các kỹ thuật như tạo dữ liệu tổng hợp (synthetic data generation), tăng cường dữ liệu (data augmentation), và học chủ động (active learning) cần được khám phá để tăng hiệu quả thu thập.
- Khám phá các kỹ thuật tinh chỉnh (Fine-tuning) khác: công trình hiện tại sử dụng QLoRA 4-bit, một phương pháp hiệu quả về mặt tài nguyên nhưng có thể làm giảm độ chính xác. Các công trình trong tương lai cần so sánh hiệu quả giữa các phương pháp tinh chỉnh khác nhau, bao gồm tinh chỉnh toàn bộ tham số (full fine-tuning) và các kỹ thuật Parameter-Efficient Fine-Tuning (PEFT) khác để tìm ra sự cân bằng tối ưu giữa hiệu suất và chi phí tính toán.
- Áp dụng các kỹ thuật suy luận (Reasoning): Hướng dẫn LLM cách "suy nghĩ" từng bước thông qua các kỹ thuật như Chuỗi suy luận (Chain-of-Thought - CoT) prompting. Bằng cách yêu cầu mô hình giải thích logic trước khi đưa ra kết luận, phương pháp này có thể cải thiện khả năng xử lý các trường hợp phức tạp, mỉa mai và ẩn ý.
- Tận dụng các giải thích từ LLM: Sử dụng các "lý giải" (rationales) do LLMs tạo ra trong một vòng lặp cải tiến. Các giải thích này có thể được dùng để nhanh chóng xác định các mẫu dữ liệu bị gán nhãn sai, từ đó cải thiện chất lượng bộ dữ liệu. Đồng thời, việc phân tích các lý giải cũng giúp tinh chỉnh lại câu lệnh (prompt) và các chỉ dẫn hệ thống để mô hình hiểu rõ hơn về ranh giới giữa các lớp.

- Phát hiện Đa phương thức (Multimodal Detection): Mở rộng nghiên cứu để xử lý các nội dung đa phương thức (văn bản, hình ảnh, video), vì trên thực tế, tuyên truyền chống phá ngày càng được thực hiện thông qua các phương tiện như meme, hình ảnh chế, và video ngắn.
- Xây dựng cơ chế học liên tục (Continuous Learning): Phát triển các mô hình có khả năng tự cập nhật và thích ứng với sự tiến hóa không ngừng của ngôn ngữ lóng, teencode và các chiến thuật tuyên truyền mới trên không gian mạng.

TÀI LIỆU THAM KHẢO

[1] “Cảnh giác trước các thủ đoạn phá hoại, gây rối nhân dịp kỷ niệm 50 năm giải phóng miền Nam, thống nhất đất nước”, Trang tin Điện tử Đảng bộ thành phố Hồ Chí Minh. Truy cập: 2 Tháng Tám 2025. [Online]. Available at: http://hcmcpv.org.vn/tin-tuc/canh-giac-truoc-cac-thu-doan-pha-hoai-gay-roi-nhan-dip-ky-niem-50-nam-giai-phong-mien-nam-thong-nh-1491936230

[2] Ts Hoàng Quốc Cảnh - Nguyễn Hương Hạnh - Đỗ Thị Mỹ Dung, “Vạch trần phương thức, thủ đoạn lợi dụng mạng xã hội Facebook để xuyên tạc, chống phá Đảng và Nhà nước Việt Nam”. Truy cập: 31 Tháng Năm 2025. [Online]. Available at: https://www.tapchicongsan.org.vn/web/guest/nghien-cu/-/2018/972002/vach-tran-phuong-thuc%2C-thu-doan-loi-dung-mang-xa-hoi-facebook-de-xuyen-tac%2C-chong-pha-dang-va-nha-nuoc-viet-nam.aspx#

[3] P. G. Hoang, C. D. Luu, K. Q. Tran, K. V. Nguyen, và N. L.-T. Nguyen, “ViHOS: Hate Speech Spans Detection for Vietnamese”, trong *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos và I. Augenstein, B.t.v, Dubrovnik, Croatia: Association for Computational Linguistics, tháng 5 2023, tr 652–669. doi: 10.18653/v1/2023.eacl-main.47.

[4] K. D. Pham, D. V. Thin, và N. L.-T. Nguyen, “Improving Vietnamese Fake News Detection based on Contextual Language Model and Handcrafted Features”, *VNUHCM J. Sci. Technol. Dev.*, vol 26, số p.h 2, Art. số p.h 2, tháng 6 2023, doi: 10.32508/stdj.v26i1.3927.

[5] “Nhận diện âm mưu, phương thức, thủ đoạn chống phá Đảng, Nhà nước và chế độ ta của các thế lực thù địch trong và ngoài nước trên không gian mạng”, Cổng Thông tin điện tử tỉnh Tuyên Quang. Truy cập: 3 Tháng Tám 2025. [Online]. Available at: http://www.tuyenquang.gov.vn/vi/post/12899?id=12899&type=TinTuc

[6] “Tuyên truyền chống Nhà nước sẽ chịu mức án như thế nào? Và các bản án điển hình”, THƯ VIỆN PHÁP LUẬT. Truy cập: 3 Tháng Tám 2025. [Online].

Available at: https://thuvienphapluat.vn//banan/tin-tuc/tuyen-truyen-chong-nha-nuoc-se-chiu-muc-an-nhu-the-nao-va-cac-ban-an-dien-hinh-7213

[7] thuvienphapluat.vn, “Luật an ninh mạng 2018 số 24/2018/QH14 áp dụng 2025 mới nhất”, THƯ VIỆN PHÁP LUẬT. Truy cập: 2 Tháng Tám 2025. [Online]. Available at: https://thuvienphapluat.vn/van-ban/Cong-nghe-thong-tin/Luat-an-ninh-mang-2018-351416.aspx

[8] “Nhận diện và kiên quyết đấu tranh, phản bác các thế lực thù địch, phản động tuyên truyền, chống phá tấn công nền tảng tư tưởng của Đảng Cộng sản Việt Nam”. Truy cập: 3 Tháng Tám 2025. [Online]. Available at: https://tapchilichsudang.vn/nhan-dien-va-kien-quyet-dau-tranh-phan-bac-cac-the-luc-thu-dich-phan-dong-tuyen-truyen-chong-pha-tan-cong-nen-tang-tu-tuong-cua-dang-cong-san-viet-nam.html

[9] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, và B. Scholkopf, “Support vector machines”, *IEEE Intell. Syst. Their Appl.*, vol 13, số p.h 4, tr 18–28, 1998, doi: 10.1109/5254.708428.

[10] Vikramkumar, V. B, và Trilochan, “Bayes and Naive Bayes Classifier”, 3 Tháng Tư 2014, *arXiv*: arXiv:1404.0933. doi: 10.48550/arXiv.1404.0933.

[11] D. C. Asogwa, C. I. Chukwuneke, C. C. Ngene, và G. N. Anigbogu, “Hate Speech Classification Using SVM and Naive BAYES”, 21 Tháng Ba 2022. doi: 10.9790/0050-09012734.

[12] K. O’Shea và R. Nash, “An Introduction to Convolutional Neural Networks”, 2 Tháng Chạp 2015, *arXiv*: arXiv:1511.08458. doi: 10.48550/arXiv.1511.08458.

[13] S. Hochreiter và J. Schmidhuber, “Long Short-Term Memory”, *Neural Comput.*, vol 9, số p.h 8, tr 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[14] A. Bihari *và c.s.*, “Identification of Hate Speech on Social Media using LSTM”, vol 17, tr 468–474, tháng 12 2023.

[15] A. Toktarova *và c.s.*, “Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods”, *Int. J. Adv. Comput. Sci. Appl.*, vol 14, số p.h 5, 2023, doi: 10.14569/IJACSA.2023.0140542.

[16] J. S. Malik, H. Qiao, G. Pang, và A. van den Hengel, “Deep Learning for Hate Speech Detection: A Comparative Study”, 7 Tháng Chạp 2023, *arXiv*: arXiv:2202.09517. doi: 10.48550/arXiv.2202.09517.

[17] J. Devlin, M.-W. Chang, K. Lee, và K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, 24 Tháng Năm 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[18] Y. Liu *và c.s.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, 26 Tháng Bảy 2019, *arXiv*: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.

[19] A. Vaswani *và c.s.*, “Attention Is All You Need”, 2 Tháng Tám 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[20] D. Chakravorty, A. Das, và D. Saha, “Multilingual Hate Speech Detection Using Transformer-Based Deep Learning Approaches”, trong *2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2024, tr 126–131. doi: 10.1109/SNAMS64316.2024.10883805.

[21] G. Ramos *và c.s.*, “A comprehensive review on automatic hate speech detection in the age of the transformer”, *Soc. Netw. Anal. Min.*, vol 14, số p.h 1, tr 204, tháng 10 2024, doi: 10.1007/s13278-024-01361-3.

[22] A. Matarazzo và R. Torlone, “A Survey on Large Language Models with some Insights on their Capabilities and Limitations”, 9 Tháng Hai 2025, *arXiv*: arXiv:2501.04040. doi: 10.48550/arXiv.2501.04040.

[23] S. Das, A. Dutta, K. Roy, A. Mondal, và A. Mukhopadhyay, “A Survey on Automatic Online Hate Speech Detection in Low-Resource Languages”, 28 Tháng Mười-Một 2024, *arXiv*: arXiv:2411.19017. doi: 10.48550/arXiv.2411.19017.

[24] K. Q. Tran, A. T. Nguyen, P. G. Hoang, C. D. Luu, T.-H. Do, và K. V. Nguyen, “Vietnamese Hate and Offensive Detection using PhoBERT-CNN and Social Media Streaming Data”, 1 Tháng Sáu 2022, *arXiv*: arXiv:2206.00524. doi: 10.48550/arXiv.2206.00524.

[25] L. T. Nguyen, “ViHateT5: Enhancing Hate Speech Detection in Vietnamese With A Unified Text-to-Text Transformer Model”, 4 Tháng Sáu 2024, *arXiv*: arXiv:2405.14141. doi: 10.48550/arXiv.2405.14141.

[26] S. T. Luu, K. V. Nguyen, và N. L.-T. Nguyen, “A Large-Scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts”, trong *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, H. Fujita, A. Selamat, J. C.-W. Lin, và M. Ali, B.t.v, Cham: Springer International Publishing, 2021, tr 415–426.

[27] D. V. Vo và P. Do, “Detecting Vietnamese fake news”, *CTU J. Innov. Sustain. Dev.*, vol 15, số p.h Special issue: ISDS, tr 39–46, tháng 10 2023, doi: 10.22144/ctujoisd.2023.033.

[28] C. V. Dinh, S. T. Luu, và A. G.-T. Nguyen, “Detecting Spam Reviews on Vietnamese E-commerce Websites”, vol 13757, 2022, tr 595–607. doi: 10.1007/978-3-031-21743-2_48.

[29] L. Breiman, “Random Forests”, *Mach. Learn.*, vol 45, số p.h 1, tr 5–32, tháng 10 2001, doi: 10.1023/A:1010933404324.

[30] M. K. Chung, “Introduction to logistic regression”, 28 Tháng Mười 2020, *arXiv*: arXiv:2008.13567. doi: 10.48550/arXiv.2008.13567.

[31] D. Q. Nguyen và A. Tuan Nguyen, “PhoBERT: Pre-trained language models for Vietnamese”, trong *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, và Y. Liu, B.t.v, Online: Association for Computational Linguistics, tháng 11 2020, tr 1037–1042. doi: 10.18653/v1/2020.findings-emnlp.92.

[32] P. N.-T. Do, S. Q. Tran, P. G. Hoang, K. V. Nguyen, và N. L.-T. Nguyen, “VLUE: A New Benchmark and Multi-task Knowledge Transfer Learning for Vietnamese Natural Language Understanding”, 23 Tháng Ba 2024, *arXiv*: arXiv:2403.15882. doi: 10.48550/arXiv.2403.15882.

[33] A. Conneau *và c.s.*, “Unsupervised Cross-lingual Representation Learning at Scale”, 8 Tháng Tư 2020, *arXiv*: arXiv:1911.02116. doi: 10.48550/arXiv.1911.02116.

[34] T. N. Chien Van Nguyen Thuat Nguyen, Quan Nguyen, Huy Nguyen, Björn Plüster, Nam Pham, Huu Nguyen, Patrick Schramowski, “Vistral-7B-Chat - Towards a State-of-the-Art Large Language Model for Vietnamese”, 2023.

[35] W. Zhang *và c.s.*, “SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages”, 29 Tháng Bảy 2024, *arXiv*: arXiv:2407.19672. doi: 10.48550/arXiv.2407.19672.

[36] A. Yang *và c.s.*, “Qwen3 Technical Report”, 14 Tháng Năm 2025, *arXiv*: arXiv:2505.09388. doi: 10.48550/arXiv.2505.09388.

[37] T. Dettmers, A. Pagnoni, A. Holtzman, và L. Zettlemoyer, “QLoRA: Efficient Finetuning of Quantized LLMs”, 23 Tháng Năm 2023, *arXiv*: arXiv:2305.14314. doi: 10.48550/arXiv.2305.14314.