Phát hiện ngôn từ công kích trên mạng xã hội tiếng Việt

Nguyễn Trường Giang, Nguyễn Thanh Tùng, Hồ Đức Dũng
University of Information Technology, Ho
Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{giangntse,tungntse}@fpt.edu.vn

Abstract

Chứng kiến sự phát triển nhanh chóng và mạnh mẽ của các phương tiện truyền thông xã hội tại Việt Nam như facebook, youtube, tiktok,. . . Điều này cũng đồng nghĩa với việc tạo ra một lượng dữ liệu khổng lồ từ người dùng. Do đó, việc xuất hiện những phát ngôn, bình luận mang thiên hướng tiêu cực, công kích và xúc phạm trở nên lớn hơn và khó kiểm soát. Trong bối cảnh này, nghiên cứu về việc phát hiện lời nói bạo lực trở thành một phần quan trọng, nhằm giúp cải thiện môi trường trực tuyến và thúc đẩy cuộc trò chuyện lành mạnh hơn. Trong nghiên cứu này, chúng tôi đã thu thập, gán nhãn và trình bày một tập ngữ liệu về việc phát hiện lời nói căm thù bao gồm 7300 mẫu văn bản. Chúng tôi cũng khám phá tính hiệu quả của bốn mô hình riêng biệt: Logistic Regres- sion, Support Vector Machine (SVM), Bidirec- tional Long Short-Term Memory (BiLSTM) và ViSoBERT, một Pre-trained Language Model (PLM) dựa trên transformer xử lý tốt trên các dữ liệu truyền thông xã hội tiếng Việt. Các thử nghiệm của chúng tôi bao gồm các giai đoạn đào tạo và đánh giá, sử dụng độ đo F1 và Ac- curacy để đánh giá tính hiệu quả của mô hình trong việc xác định lời nói căm thù. Kết quả cho thấy mức độ thành công khác nhau giữa các mô hình, trong đó ViSoBERT thể hiện hiệu suất đặc biệt hứa hẹn. Nghiên cứu này góp phần hiểu biết sâu hơn về các phương pháp phát hiện lời nói căm thù và làm nổi bật tiềm năng của các mô hình dựa trên transformer trong việc giải quyết các sắc thái ngôn ngữ trong việc phát hiện lời nói căm thù trong văn bản tiếng Việt.

Disclaimer: Bài viết này chứa đựng những bình luận thực tế của người dùng trên mạng xã hội nên có thể chứa những từ ngữ nhạy cảm, thô tục.

1 Giới thiệu

Sự phổ biến của mạng xã hội đã biến đổi cách giao tiếp, cho phép mọi người từ các nền văn hóa đa dạng có thể kết nối và chia sẻ ý kiến trên quy mô toàn cầu. Tuy nhiên, sự kết nối chưa từng có này cũng đồng nghĩa với sự gia tăng lo ngại về những câu từ tiêu cực, định kiến và ngôn ngữ xúc phạm trực tuyến. Tác động của cuộc trò chuyện tiêu cực này đối với cá nhân và cộng đồng làm nổi bật nhu cầu quan trọng về các công cụ mạnh mẽ có khả năng nhận diện và giảm thiểu những biểu hiện có hại này. Hate speech - Lời nói căm thù trên mạng xã hội đề cập đến những biểu hiện xúc phạm hoặc đe dọa các cá nhân hoặc nhóm dựa trên các đặc điểm như chủng tộc, sắc tộc, tôn giáo, giới tính hoặc các đặc điểm khác. Nó có thể biểu hiện dưới nhiều hình thức khác nhau, bao gồm chữ viết, hình ảnh hoặc video và thường gắn liền với việc kích động bạo lực, đưa cá nhân hoặc tổ chức vào một tình huống tiêu cực thậm chí nguy hiểm đến tính mạng.

Có những yếu tố nào tạo nên lời lẽ căm ghét và khi nào nó khác biệt so với ngôn ngữ xúc phạm? Hiện tại định nghĩa về cả hai ngôn từ này vẫn đang còn có những sự nhập nhằng tùy vào cách nhìn nhận của mỗi người mà có thể có những định nghĩa riêng khác nhau đối với chúng. Nhưng đa số đều đồng tình rằng đó là loại ngôn ngữ tiêu cực có thể nhắm tới các cá nhân, tổ chức và gây hại cho họ Walker (1994).

Trong nhiều quốc gia, bao gồm Vương quốc Anh, Canada và Pháp, có các luật cấm lời lẽ căm ghét, thường được định nghĩa là lời nói nhằm vào các nhóm thiểu số một cách có thể thúc đẩy bạo lực hoặc gây rối xã hội. Những người bị kết án vì sử dụng lời lẽ căm ghét thường phải đối mặt với những mức phạt lớn và thậm chí là án tù. Những luật lệ này được mở rộng đến internet và mạng xã hội, khiến nhiều trang web tạo ra các quy định riêng chống lại lời lẽ căm ghét. Cả Facebook và Twitter đã đáp ứng các chỉ trích về việc không làm đủ để ngăn chặn lời lẽ căm ghét trên các trang web của họ bằng cách thiết lập chính sách để cấm việc sử dụng nền tảng của họ để tấn công vào những người dựa trên các đặc điểm như chủng tộc, dân tộc, giới

tính và tình dục, hoặc đe dọa bạo lực đối với người khác.

Lấy cơ sở từ những định nghĩa này, chúng tôi đặt ra định nghĩa cho lời lẽ căm ghét như ngôn ngữ được sử dụng để thể hiện sự căm ghét đối với một nhóm cụ thể hoặc có ý định mỉa mai, làm nhục, hoặc xúc phạm các thành viên của nhóm đó. Trong những trường hợp cực kỳ nghiêm trọng, đây có thể là ngôn ngữ đe dọa hoặc kích động bạo lực, nhưng nếu giới hạn định nghĩa của chúng tôi chỉ vào những trường hợp như vậy, sẽ loại trừ một phần lớn lời lẽ căm ghét. Quan trọng là, định nghĩa của chúng tôi không bao gồm tất cả các trường hợp của ngôn ngữ xúc phạm, vì người ta thường sử dụng các thuật ngữ mà theo đó có thể làm tổn thương nhóm nào đó một cách khác biệt chất lượng.

Mục tiêu của nghiên cứu của chúng tôi là phát triển một mô hình tiên tiến cho việc phân loại lời lẽ căm ghét, không chỉ dừng lại ở việc xác định lời lẽ căm ghét rõ ràng mà còn bao gồm sự tinh tế vốn có trong cảnh quan giao tiếp trực tuyến đang phát triển liên tục. Bằng cách hiểu rõ sự phức tạp của ngôn ngữ xúc phạm, mô hình đề xuất nhằm đóng góp vào việc nâng cao cơ chế quản lý nội dung, tạo điều kiện cho một môi trường trực tuyến Wang et al. (2014), an toàn và tích cực hơn. Công trình nghiên cứu trước đó về phát hiện lời lẽ căm ghét đã nhận diện vấn đề này, nhưng nhiều nghiên cứu vẫn có xu hướng lẫn lộn lời lẽ căm ghét và ngôn ngữ xúc phạm.

2 Các công trình liên quan

Phân tích ngôn ngữ chủ quan trên mạng xã hội đã được nghiên cứu sâu rộng và áp dụng trong các lĩnh vực khác nhau, từ phân tích tâm trạng
J. M. Soler and Roblizo (2012) đến phát hiện châm biếm Bouazizi and Ohtsuki (2016), hoặc phát hiện tin đồn Z. Zhao and Mei (2015) vv.

Tuy nhiên, có tương đối ít nghiên cứu (so với các chủ đề đã đề cập) đã được thực hiện trong lĩnh vực phát hiện lời nói căm ghét. Một số nghiên cứu này nhắm đến các câu trên mạng toàn cầu như công trình của hate speech on the world wide Web (2012) và (Djuric and Bhamidipati, 2015). Công trình đầu tiên đạt được độ chính xác phân loại bằng 94% với điểm F1 bằng 63,75% trong nhiệm vụ phân loại nhị phân, trong khi công trình thứ hai đạt độ chính xác bằng 80%.

Luu et al. (2021) đã giới thiệu ViHSD - một bộ dữ liệu đã được gán nhãn để phát hiện tự động lời nói bạo lực trên mạng xã hội Việt Nam. Bộ dữ liệu này chứa hơn 30,000 bình luận, mỗi bình luận được gán một trong ba nhãn: CLEAN, OFFENSIVE, hoặc HATE. Họ đã đánh giá bộ dữ liệu này bằng cách sử dụng mô hình deep learning và các pre- trained model dựa trên transfomers.

N. D. Gitari and Long (2015) trích xuất các câu từ một số ’trang web chứa nhiều lời nói căm ghét tại Hoa Kỳ. Họ chú thích mỗi câu vào một trong ba lớp: ’cực kỳ căm ghét (Strongly Hateful)’, ’nhẹ nhàng căm ghét (Weakly Hateful)’, và ’không căm ghét (Non-Hateful)’. Họ sử dụng các đặc điểm ngữ nghĩa và đặc điểm mô hình ngữ pháp, thực hiện phân loại trên một bộ kiểm tra và đạt được một F1-score bằng 65,12%.

Tuy nhiên, một số công trình khác nhắm đến việc phát hiện các câu nói căm ghét trên Twitter.C. No- bata and Chang (2016) nhằm đến việc phát hiện các tweet căm ghét chống lại người da đen. Họ sử dụng đặc điểm unigram với độ chính xác bằng 76% trong nhiệm vụ phân loại nhị phân. Rõ ràng, sự tập trung vào lời nói căm ghét đối với một giới tính, nhóm dân tộc, sắc tộc cụ thể khác nhau khiến cho các unigram được thu thập liên quan đến nhóm đó. Do đó, từ điển unigram được xây dựng không thể tái sử dụng để phát hiện lời nói căm ghét đối với các nhóm khác với hiệu suất tương tự. (Burnap and Williams, 2015)sử dụng các phụ thuộc kiểu (tức là, mối quan hệ giữa các từ) cùng với đặc điểm ’bag of words’ (BoW) để phân biệt lời nói căm ghét khỏi lời nói trong sạch.

Một PLM chuyên biệt dành cho dữ liệu ngôn ngữ truyền thông xã hội tiếng Việt dựa trên kiến trúc XLM-R, ViSoBERT Nguyen et al. (2023) được huấn luyện trên một tập dữ liệu lớn chất lượng và đa dạng của các văn bản mạng xã hội Việt Nam, thực hiện tốt trên nhiều tác vụ quan trọng trên các văn bản mạng xã hội tiếng việt: Nhận diện cảm xúc, phát hiện ngôn ngữ thù địch, phân tích tình cảm, phát hiện đánh giá tin rác và xác định đoạn văn chứa nội dung bạo lực. Đây cũng chính là mô hình chính đại diện cho phương pháp huấn luyện dựa trên các PLM – một trong 3 phương pháp huấn luyện chính mà chúng tôi hướng tới.

Bộ dữ liệu ViCTSD Nguyen et al. (2021) được xây dựng để xác định tính độc hại trong nhận xét của người dùng. Xuất phát từ các trang web tin tức trực tuyến, nơi người dùng bình luận và thường thể hiện bản thân theo phong cách trịnh trọng (formal), mức độ phản cảm có thể không quá lộ liễu như trong các bộ dữ liệu khác. Do đó, việc tiến hành phát hiện tính độc hại trên tập dữ liệu này đặt

một thách thức đối với các mô hình ngôn ngữ.

Một bộ ngữ liệu đại diện cho kho văn bản có chú thích đầu tiên của con người để xác định các span thù hận và xúc phạm trong văn bản tiếng Việt - ViHOS Hoang et al. (2023), cung cấp tác vụ syllable-level cho HSD tiếng Việt. Với hơn 11 nghìn bình luận và khoảng 26 nghìn span chú thích. Các phương pháp tinh chỉnh dựa trên BERT hiện tại thường áp dụng gắn thẻ trình tự IOB để xử lý trước dữ liệu, coi nhiệm vụ này là nhiệm vụ phân loại.

Ngoài ra, còn có các phương pháp phân tích dựa trên những ngôn ngữ khác nhau, đặt biệt là Zampieri ((2020) phân tích trên 5 bộ ngôn ngữ khác nhau: Tiếng Anh, Tiếng Ả Rập, Tiếng Đan Mạch, Tiếng Hy Lạp và Tiếng Thổ Nhĩ Kỳ.

3 Dữ liệu

Chúng tôi đã khảo sát và thu thập bình luận của người dùng từ nhiều fanpage và bài viết trên nền tảng mạng xã hội Facebook Việt Nam liên quan đến những chủ đề khác nhau về các vấn đề như chính trị, game giải trí, bóng đá, người nổi tiếng,. Chúng
tôi chọn lọc những bài viết có lượng tương tác tiêu cực cao. Sử dụng thư viện Selenium để tự động thu thập, sau đó tổng hợp và khảo sát mô hình, xây dựng guildline gán nhãn.

3.1 Xây dựng guildline

Chúng tôi hướng tới xây dựng cấu trúc của bộ ngữ liệu và guildline dựa trên nghiên cứu "A Large- scale Dataset for Hate Speech Detection on Viet- namese Social Media Texts" Luu et al. (2021). Bộ ngữ liệu chứa 3 nhãn: CLEAN, OFFENSIVE và HATE, trong đó có 2 nhãn biểu thị các bình luận mang tính chất tiêu cực và 1 nhãn mô tả cho những bình luận bình thường. Các bình luận thường được viết dưới dạng informal. Guildline và ý nghĩa chi tiết của 3 tập nhãn được biểu diễn trong bảng 2. Mặc dù số lượng người gán nhãn của bộ ngữ liệu chỉ là 1 nhưng việc xây dựng guildline phục vụ cho việc gán nhãn vẫn rất cần thiết để có thể định rõ các tiêu chí và quy định cho việc gán nhãn, từ đó giảm thiểu sự không nhất quán trong mọi thời điểm khác nhau và đảm bảo rằng quá trình gán nhãn được thực hiện một cách có hệ thống và đáng tin cậy.

3.2 Tổng quan về bộ ngữ liệu

Bộ ngữ liệu chứa 7,380 bình luận. Mỗi bình luận đã được gán thành 3 nhãn khác nhau: CLEAN (0), OFFENSIVE (1), HATE (2). Cấu trúc của bộ ngữ liệu bao gồm 2 thuộc tính: "Comment" và "Labels" tương ứng với bình luận và nhãn của chúng. Bảng 3 hiển thị một số ví dụ về bình luận và cấu trúc của bộ ngữ liệu. Sau cùng, chúng tôi tiến hành chia bộ ngữ liệu thành 3 tập riêng biệt gồm: tập huấn luyện (train), tập thẩm định (val), tập kiểm thử (test) tương ứng với tỷ lệ 7:1:2. Hình 1 cung cấp một cái nhìn tổng quan về phân phối của các nhãn trên các tập dữ liệu. Tuy có một số dấu hiệu về mất cân bằng dữ liệu, nhưng không đến mức quá nghiêm trọng, điều này có thể ảnh hưởng đến hiệu suất của mô hình nhưng không gây ra vấn đề lớn.

Thời điểm hiện tại, có khá nhiều bộ ngữ liệu phục vụ chủ đề phát hiện và phân loại ngôn từ xúc phạm, độc hại cho tiếng việt như các bộ ngữ liệu thuộc mục 2. Bảng 1 thể hiện so sánh cấu trúc của bộ ngữ liệu của chúng tôi với những bộ dữ liệu khác liên quan đến chủ đề ngôn từ tiêu cực trên mạng xã hội tiếng Việt. Bộ ngữ liệu của chúng tôi xây dựng có những đặc điểm đặc điểm nổi bật, giúp cho việc xử lý, nhận diện câu từ độc hại hiệu quả hơn như:

- **Mức độ đa dạng bình luận cao:** bộ ngữ liệu này được chúng tôi thu thập từ các *drama*, *scandal* mới nổi lên đầu năm 2024 này với nhiều chủ đề khác nhau do đó những bình luận mới mẻ hơn, có mức độ đa dạng cao.

- **Tránh thiên vị nhãn khi huấn luyện và dự đoán:** hình 1 biểu diễn số lượng của các nhãn trong từng tập dữ liệu, ta có thể nhận thấy tỉ lệ giữa nhãn có số lượng thấp nhất và nhãn có số lượng cao nhất là không quá 15% điều này dẫn đến dữ liệu sẽ không bị mất cân bằng và các mô hình sẽ nhận diện với hiệu suất đồng đều trên cả 3 nhãn.

- **Guildline được thiết kế với nhiều trường hợp:** giúp cho việc gán nhãn trở nên tốt hơn và các bình luận sẽ giảm tối đa vấn đề nhập nhằng trong dữ liệu.

4 Thí nghiệm

Trong nghiên cứu này, chúng tôi muốn so sánh và đánh giá hiệu suất của các mô hình dự đoán trên bộ ngữ liệu về ngôn ngữ thù địch tiếng Việt. Mục tiêu của chúng tôi là có cái nhìn tổng quan hơn về các ưu điểm và hạn chế của mỗi mô hình khi đối mặt với các đặc trưng và ngữ cảnh đặc biệt của ngôn ngữ công kích. Bốn mô hình khác nhau đã được sử dụng, Hai trong số đó là mô hình dựa trên phương pháp máy học thông thường, bao gồm các

	OUR corpus	ViHSD	ViHOS	UIT-ViCTSD
Số lượng nhãn	3	3	-	2
Số lượng bình luận	7,380	33,400	11,056	10,000
Lượng từ vựng	10,937	29,183	14,182	18,561
Nguồn thu thập	Facebook	Facebook & Youtube	ViHSD	VnExpress
Độ đồng thuận	-	0.52 (Cohen Kappa)	0.72 (Cohen Kappa)	0.58 (Fleiss' Kappa)

Table 1: So sánh các bộ ngữ liệu phổ biến về chủ đề ngôn từ tiêu cực tiếng Việt

Nhãn	Mô tả	Ví dụ
CLEAN	Những bình luận không mang tính chất công kích, lăng mạ. Có thể chứa các từ như "mày", "tao", "bố mày".	Bình luận 1: "Em bán xôi kiếm từng đồng lẻ mà nó cũng cướp kg tha" Bình luận 2: "Bố lạy mày"
OFFENSIVE	Những bình luận chửi tục, chửi đổng không nhắm vào một cá nhân hay tổ chức. Những bình luận có nhắm vào các tổ chức, cá nhân nhưng theo thiên hướng nói đểu, nói xoáy. Những từ ngữ mang tính xúc phạm như "nguyền rủa", "quả báo".	Bình luận 1: "Ý thức như đb" Bình luận 2: "Quả báo sẽ ko chừa 1 ai...đợi đi sẽ đến sớm với mày thôi"
HATE	Những bình luận mang tính công kích cá nhân, tổ chức. Bình luận kỳ thị, công kích bằng tiếng lóng, nghĩa bóng. Phân biệt chủng tộc, phân biệt vùng miền.	Bình luận 1: "Bố tk óc chó ngáo đá tưởng vậy là ngầu" Bình luận 2: "Cái thằng khỉ đầu chó này nữa" Bình luận 3: "Đ*t m* mày thằng mọi đen không biết làm trọng tài bóng đá thì đi bốc cứt mà ăn đừng làm ô uế bóng đá"

Table 2: Xây dựng guildline

kỹ thuật như Logistic Regression và Support Vector Machines, được lựa chọn vì tính hiệu quả và tính linh hoạt trong việc xử lý các vấn đề phân loại văn bản. Kế tiếp là một mô hình đến từ phương pháp học sâu (Deep learning) cụ thể là mạng nơ-ron hồi quy nhằm khai thác sâu hơn các đặc trưng ẩn của dữ liệu. Và cuối cùng, một mô hình đến từ một PLM dựa trên kiến trúc transformers. Tận dụng khả năng hiểu ngôn ngữ, ngữ cảnh mạnh mẽ của các mô hình đã được huấn luyện trên lượng lớn dữ liệu từ trước, đây là phương pháp tiếp cận có tiềm năng thể hiện hiệu suất tốt nhất trong tất cả các tiếp cận của chúng tôi đối với bộ ngữ liệu.

Chúng tôi thực hiện việc huấn luyện và đánh giá hiệu suất mô hình thông qua hai hướng tiếp cận. Thứ nhất là đào tào với dữ liệu đã được tiền xử lý từ trước và thứ hai là dữ liệu chưa được trải qua quá trình làm sạch. Điều này giúp đánh giá tổng qaun được việc tiền xử lý ảnh hưởng thế nào đến các dữ liệu đến từ mạng xã hội trực tuyến.

4.1 Tiền xử lý dữ liệu

Sau quá trình khảo sát bộ dữ liệu, chúng tôi nhận thấy còn nhiều vấn đề có thể xử lý trước khi đưa vào mô hình huấn luyện:

- **Đưa văn bản về dạng viết thường:** Trong quá trình thu thập dữ liệu, chúng tôi đã thực hiện việc loại bỏ các thực thể tên riêng nên để mô hình phân biệt chữ in hoa và chữ thường là không cần thiết. Và nhằm giảm độ phức tạp của dữ liệu cũng như giúp các bước tiền xử lý sau được thực hiện tốt hơn, chúng tôi quyết định đưa các bình luận về dạng viết thường.

- **Xóa các đường dẫn có chứa trong bình luận:** thông thường đối với các bình luận,

Tên cột	Kiểu dữ liệu	Mô tả	Giá trị
Comment	String	Bình luận của người dùng.	bỏ con mà éo liên quan là sao?
Labels	Float	Nhãn tương ứng.	1

Table 3: Ví dụ về một mẫu dữ liệu

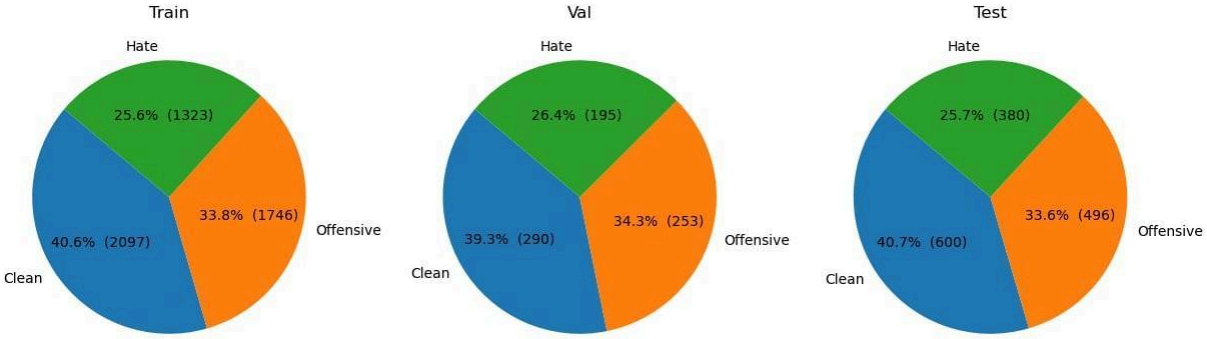

Figure 1: Phân bố nhãn của mỗi tập dữ liệu

bài viết do người dùng tạo ra đôi lúc sẽ chứa một số đường dẫn nhằm nhiều mục đích khác nhau như quảng cáo, đính kèm thông tin,. . . . Do những đường dẫn này thường không mang ý nghĩa trong câu, nên chúng tôi quyết định xóa bỏ chúng để giảm nhiễu dữ liệu trước khi đưa vào mô hình huấn luyện.

- **Xóa các kí tự đặc biệt cũng như những biểu tượng cảm xúc:** Trong bài toán phát hiện lời nói căm thù, các kí tự đặc biệt và biểu tượng cảm xúc thường không mang ý nghĩa trong câu nhiều như các bài toán phân tích tình cảm (sentiment analysis). Nên chúng tôi cân nhắc đến việc xóa chúng mà không lo ảnh hưởng xấu đến hiệu suất của mô hình.

- **Chuẩn hóa các teencode, từ viết tắt:** Teen- code là một cách biến đổi chữ viết thông thường thành một phiên bản khác trông trẻ trung, năng động hơn. Kiểu chữ này thường đi kèm với việc sử dụng từ viết tắt, ký hiệu thường được sử dụng trong các bình luận trên mạng xã hội. Có rất nhiều teencode mang cùng một ý nghĩa, ví dụ: “thg”, “thk”, “thag” (từ gốc: “thằng”) đều có mục đích chung để chỉ một ai đó. Để có thể chuyển đổi các teen- code thành từ gốc, Chúng tôi sử dụng một pre-trained sequence to sequence model là BARTpho Tran et al. (2021) để fine-tuning trên ViLexNorm Nguyen et al. (2024), đây là một bộ ngữ liệu dành riêng cho việc chuẩn hóa từ vựng trên dữ liệu truyền thông xã hội tiếng việt. Đánh giá trên tập kiểm thử đạt hiệu suất khả quan với Error Reduction Rate (ERR) Van Der Goot (2019) và Accuracy lần lượt là 0.805 và 0.94, chúng tôi áp dụng trên bộ dữ liệu của chúng tôi để sinh ra một bộ dữ liệu mới đã được chuẩn hóa từ vựng. Tiếp theo, các từ mang tính chất *Hate* và *Offensive* như *dm*, *cc*,... của bộ ngữ liệu sẽ được chuẩn hóa thông qua một từ điển mà chúng tôi định nghĩa trong quá trình gán nhãn để chuẩn hóa hoàn toàn teencode và từ viết tắt trong bộ ngữ liệu. Điều này cũng giúp cho việc tách từ ở giai đoạn sau được trở nên tốt hơn khi mà các từ teencode đã trở vê dạng văn bản chuẩn.

- **Tách từ trong tiếng việt:** Các từ đơn, từ phức, cụm từ làm tiếng việt trở nên đa dạng hơn hết. Việc tách từ có thể giúp làm giảm độ phức tạp trong việc xử lý các văn bản, từ đó mô hình có thể tập trung vào các đơn vị từ ngữ cụ thể thay vì phải xử lý các đoạn văn bản liền mạch dài. Chúng tôi sử dụng bộ tách từ tiếng việt từ thư viện pyvi, giúp việc hiểu ngữ cảnh của các mô hình được trở nên hiệu quả hơn.

4.2 Tiếp cận mô hình

4.2.1 Phương pháp học máy truyền thống

Trong nghiên cứu này, chúng tôi sử dụng Support Vector Machine (SVM) và Logistic Regression đại
diện cho các mô hình học máy cơ bản.

SVM nổi bật với khả năng xử lý tốt các không gian đặc trưng có số chiều lớn, đặc biệt hữu ích trong phân loại văn bản nơi mỗi từ có thể được xem

là một chiều. SVM có khả năng tìm ra siêu phẳng phân tách tối ưu, giúp tối đa hóa khoảng cách giữa các lớp khác nhau, từ đó tăng cường khả năng phân loại chính xác, đặc biệt trong các trường hợp dữ liệu không tuyến tính nhờ sử dụng các kernel trick. Đối với Logistic Regression, mặc dù đơn giản hơn, lại rất mạnh mẽ và hiệu quả trong nhiều tác vụ phân loại văn bản. Logistic Regression dễ triển khai, huấn luyện nhanh và có khả năng mở rộng tốt với các tập dữ liệu lớn. Nó cung cấp đầu ra xác suất, cho phép mô hình không chỉ dự đoán lớp mà còn đánh giá độ chắc chắn của dự đoán, hữu ích cho các ứng dụng cần xác suất dự đoán. Thêm vào đó, Logistic Regression dễ dàng được hiểu và giải thích, giúp dễ dàng phân tích và diễn giải kết quả của mô hình.

4.2.2 Phương pháp học sâu

Mô hình BiLSTM (Bidirectional Long Short-Term Memory) có nhiều điểm mạnh đáng chú nên chúng tôi đã thử nghiệm và xây dựng một mô hình cuối cùng gồm 2 lớp BiLSTM và 3 lớp tuyến tính với activation là 'relu' để phục vụ cho việc phân loại văn bản.

BiLSTM có khả năng học và nắm bắt thông tin từ cả hai hướng của chuỗi dữ liệu, tức là từ trước ra sau và từ sau ra trước. Điều này đặc biệt quan trọng trong xử lý ngôn ngữ tự nhiên, nơi ngữ cảnh từ các từ xung quanh có thể ảnh hưởng lớn đến ý nghĩa của một từ cụ thể. Bằng cách sử dụng cả thông tin trước và sau từ trong câu, BiLSTM có thể nắm bắt được các phụ thuộc dài hạn và các mối quan hệ phức tạp giữa các từ, giúp cải thiện độ chính xác của mô hình.

Hơn nữa, BiLSTM đặc biệt hiệu quả trong việc xử lý các chuỗi dữ liệu có thứ tự và ngữ cảnh phức tạp, như văn bản và ngôn ngữ tự nhiên, do đó thường vượt trội trong các nhiệm vụ phân loại văn bản so với các mô hình không có khả năng xử lý ngữ cảnh hai chiều.

4.2.3 Phương pháp tiếp cận từ PLM

Để tiếp cận bộ ngữ liệu phục vụ cho bài toán phát hiện câu từ công kích trên mạng xã hội tiếng việt, chúng tôi quyết định sử dụng ViSoBERT – một mô hình dựa trên BERT được tinh chỉnh trên dữ liệu mạng xã hội tiếng Việt nên sẽ có nhiều điểm mạnh đặc biệt trong bài toán này.

Đầu tiên, việc tinh chỉnh trên dữ liệu mạng xã hội tiếng Việt giúp ViSoBERT hiểu sâu sắc hơn về ngữ cảnh và cách sử dụng ngôn ngữ trong các cuộc trò chuyện thực tế của người dùng. Điều này đặc biệt quan trọng trong phát hiện ngôn từ công kích, vì các biểu hiện thù địch thường mang tính ngữ cảnh cao và có thể chứa các từ lóng, từ viết tắt hoặc các biểu hiện không chính thức (informal) khác.

Thứ hai, nhờ vào cấu trúc của BERT với cơ chế tự chú ý (self-attention), ViSoBERT có khả năng nắm bắt các mối quan hệ phức tạp và sự phụ thuộc dài hạn giữa các từ trong câu giúp mô hình nhận diện được các sắc thái tinh tế của ngôn từ thù địch mà các mô hình đơn giản hơn có thể bỏ sót. Bên cạnh đó, ViSoBERT có thể xử lý tốt các câu dài và phức tạp, thường gặp trong các bình luận và bài viết trên mạng xã hội, nơi ngôn từ thù địch có thể xuất hiện dưới nhiều hình thức khác nhau.

Cuối cùng, việc sử dụng mô hình dựa trên BERT, vốn đã được huấn luyện trước trên một lượng lớn dữ liệu văn bản, mang lại cho ViSoBERT một nền tảng hiểu biết vững chắc về ngôn ngữ, từ đó cải thiện hiệu suất tổng thể trong việc phát hiện ngôn từ công kích. Kết hợp tất cả những điểm mạnh này, Visobert trở thành một công cụ mạnh mẽ và hiệu quả trong việc phát hiện và xử lý ngôn từ công kích trên mạng xã hội tiếng Việt.

5 Phân tích kết quả

Chúng tôi đã thực hiện huấn luyện và tinh chỉnh nhiều lần trên bộ dữ liệu với 4 mô hình kể trên và cuối cùng đặt được hiệu suất khá tốt trong bài toán phát hiện câu từ công kích.

Bảng 5 so sánh hiệu suất của các mô hình trên bộ ngữ liệu của chúng tôi gồm 2 độ đo đánh giá là Accuracy và F1-Score, 2 phương pháp tiếp cận là huấn luyện và đánh giá dữ liệu không qua tiền xử lý và đã qua tiền xử lý. Riêng đối với PLM ViSoBERT, nhằm tận dụng tối đa sức mạnh của mô hình trong việc xử lý các dữ liệu trên mạng xã hội tiếng việt, chúng tôi quyết định không sử dụng các phương pháp tiền xử lý nào mà chúng tôi đã định nghĩa tại mục 4.1 bởi mô hình đã có những phương pháp để xử lý dữ liệu cho riêng mình.

Kết quả không nằm ngoài dự đoán của chúng tôi, khi mà mô hình có hiệu suất tốt nhất là pre-trained model ViSoBERT với điểm accuracy và F1-score lần lượt là 0.798, 0.793. Từ đó có thể nhận thấy rằng các mô hình đã được tiền huấn luyện thường mang lại hiệu quả cao nhờ vào khả năng xử lý ngôn ngữ tự nhiên tiên tiến và sự hiểu biết sâu rộng về ngữ cảnh ngôn ngữ. Theo sau đó là mô hình BiLSM và cuối cùng là các mô hình học máy truyền thống. Các mô hình đều đạt hiệu suất khá tốt khi F1-score

Bình luận	Nhãn thực	Nhãn dự đoán
mn nghĩ ntn vs cái tên loạn thế **âm binh**?	0	1
con chó nó cảm nhận được **bọn xấu** mà không nói được	0	2
sự thật là bỏ con cắn thầy phản bạn vong ơn bội nghĩa =))) được quả phó bang là ngọc bình thông minh dễ sợ	1	0
Quân ác nhơn sẽ có một ngày nào **sẽ tàn tật**	2	0

Table 4: Bảng mô tả các bình luận và dự đoán

luôn đạt trên mức 0.7 chứng tỏ được khả năng ứng dụng trong các bài toán phát hiện câu từ công kích tốt.

Not using preprocessed method:

Model	Accuracy	F1-Score
LR	0.723	0.713
SVM	0.732	0.725
BiLSTM	0.737	0.735
ViSoBERT	**0.798**	**0.793**

Preprocessed:

Model	Accuracy	F1-Score
LR	0.744	0.735
SVM	0.762	0.753
BiLSTM	**0.763**	**0.758**

Table 5: Kết quả của các mô hình

5.1 Phân tích lỗi

Sau khi trải qua việc đánh giá mô hình, chúng tôi nhận thấy còn nhiều khó khăn mà mô hình gặp phải khi thực hiện việc dự đoán. Bảng 4 thể hiện một số ví dụ về những bình luận mà mô hình phân loại không chính xác.

Các mô hình có xu hướng nhận diện sai các bình luận bình thường mang nhãn CLEAN nhưng có chứa một số từ khóa mang thiên hướng lăng mạ, công kích thành nhãn OFFENSIVE hay HATE.

Chẳng hạn như ở bình luận “mn nghĩ ntn vs cái tên loạn thế âm bình”, từ âm bình chứa trong câu thường mang nghĩa tiêu cực, xúc phạm, chỉ những cá nhân hay tập thể là thứ mang lại vận đen, điều xui xẻo, khó khăn hay sự tức giận cho người khác, nhưng ở ngữ cảnh của bình luận, từ “âm binh” chỉ thuộc một phần trong một cái tên mà người viết muốn hỏi ý kiến, cảm nghĩ của những người khác về cái tên này và hiển nhiên không mang bất kì ý nghĩa tiêu cực nào trong câu.

Ngoài ra, các mô hình cũng gặp nhiều khó khăn trong việc phân loại một số bình luận sử dụng tiếng lóng, tiếng dân tộc, hay tiếng địa phương để công kích, xúc phạm người khác. Ví dụ, từ địa phương như "Quân ác nhơn" thay vì "Quân ác nhân" cũng gây khó khăn cho mô hình trong việc nhận diện đúng ý nghĩa và ngữ cảnh của bình luận.

6 Kết luận

Tổng kết lại, chúng tôi đã giới thiệu một bộ ngữ liệu gồm 7.300 bình luận được thu thập từ trang mạng xã hội và gán chúng thành 3 nhãn. Huấn luyện và đánh giá nhiều mô hình với nhiều kiến trúc khác nhau trên bộ ngữ liệu. Đưa ra các phương pháp tiền xử lý và chúng cho thấy được độ hiệu quả khi hiệu suất của các mô hình được cải thiện rõ rệt. Chúng tôi hy vọng nghiên cứu của chúng tôi cũng như bộ ngữ liệu này có thể giúp ích được cho những vấn đề nan giải trong quá trình phát hiện các câu từ lăng mạ, công kích. Giúp cho không gian môi trường trên mạng xã hội trở nên trong sạch và hợp thuần phong mĩ tục hơn.

References

M. Bouazizi and T. O. Ohtsuki. 2016. A pattern-based approach for sarcasm detection on twitter. In *IEEE Access*, volume 4, pages 5477–5488.

P. Burnap and M. L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. In *Policy Internet 7(2)*, pages 223—-242.

A. Thomas Y. Mehdad C. Nobata, J. Tetreault and Y. Chang. 2016. Abusive language detection in on- line user content. In *Proc. AAAI*, pages 1621—-1622.

J.; Morris R.; Grbovic M.; Radosavljevic V.; Djuric, N.; Zhou and N. Bhamidipati. 2015. Hate speech detection with comment embeddings. In *WWW*, pages 29–30.

Detecting hate speech on the world wide Web. 2012. W. warner and j. hirschberg. In *Proc. 2nd Workshop Lang. Social Media*, pages 19–26.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023.

Vihos: Hate speech spans detection for vietnamese. *arXiv preprint arXiv:2301.10186*.

F. Cuartero J. M. Soler and M. Roblizo. 2012. Twitter as a tool for predicting elections results. In *Proc. IEEE/ACM ASONAM*, pages 1194–1200.

Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Ad- vances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 415–426. Springer.

H. Damien N. D. Gitari, Z. Zuping and J. Long. 2015. A lexicon-based approach for hate speech detection. In *Int. J. Multimedia Ubiquitous Eng.*, volume 10, pages 215—-230.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu- Thuy Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Sys- tems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 572–583. Springer.

Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, and Kiet Van Nguyen. 2023. Vi- sobert: A pre-trained language model for viet- namese social media text processing. *arXiv preprint arXiv:2310.11166*.

Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Van Nguyen. 2024. Vilexnorm: A lexical normaliza- tion corpus for vietnamese social media text. *arXiv preprint arXiv:2401.16403*.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2021. Bartpho: pre-trained sequence-to- sequence models for vietnamese. *arXiv preprint arXiv:2109.09701*.

Rob Van Der Goot. 2019. Monoise: A multi-lingual and easy-to-use lexical normalization tool. In *Proceed- ings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 201–206.

Samuel Walker. 1994. *Hate speech: The history of an American controversy*. U of Nebraska Press.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twit- ter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social com- puting*, pages 415–425.

P. Resnick Z. Zhao and Q. Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proc. Int. Conf. World Wide Web*, pages 1395–1405.

Marcos Zampieri. (2020). Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *arXiv preprint arXiv:2006.07235*.