Nhận dạng thực thể được đặt tên trong văn bản Tiếng Việt sử dụng học máy và ứng dụng trong đảm bảo an ninh mạng

Nguyễn Ngọc Toàn, Lê Xuân Tuấn, Lương Thế Dũng, Trần Nghi Phú

***Tóm tắt*— Nhận dạng thực thể được đặt tên (hoặc "được gán nhãn": Named Entity Recognization - NER) trong văn bản Tiếng Việt hiện đang là một bài toán có nhiều thách thức do không có nhiều tập dữ liệu chuẩn hoặc có nhưng chưa đủ lớn và các mô hình nhận dạng thường được xây dựng chủ yếu dựa trên phương pháp học sâu. Trong bài báo này, nhóm tác giả trình bày một cách tiếp cận có hệ thống trong việc xây dựng mô hình nhận dạng thực thể của các văn bản Tiếng Việt, từ bước thu thập, xây dựng tập dữ liệu cho đến áp dụng, tinh chỉnh các mô hình học máy. Đặc biệt, nhóm tác giả cũng đề xuất các kịch bản ứng dụng và chứng minh khả năng áp dụng mô hình nhận dạng thực thể được đặt tên hiệu quả trong giải quyết các bài toán đảm bảo an ninh, an toàn thông tin mạng. Cụ thể, nhóm tác giả xây dựng tập dữ liệu gồm hơn 5000 văn bản thu thập từ các mạng xã hội sử dụng Tiếng Việt, đặt tên và gán 1 trong 4 nhãn định trước cho các thực thể; sau đó, áp dụng mô hình huấn luyện trước XLM-RoBERTa với các tham số ban đầu được tinh chỉnh phù hợp để nhận dạng các thực thể này. Kết quả thử nghiệm cho thấy hệ thống là hiệu quả với khả năng nhận dạng thực thể của mô hình và đạt độ đo F1- measure lên đến 95.6%, trội hơn so với một số hệ thống NER cho văn bản Tiếng Việt trên cùng tập dữ liệu mà nhóm tác giả đã xây dựng. Mô hình đề xuất cũng đã được ứng dụng trong xây dựng các hệ thống hỗ trợ đảm bảo an ninh mạng hiện nay.**

***Abstract*— Named Entity Recognition (NER) in Vietnamese documents is currently a challenging task because of the lacking of standard datasets, or these datasets might be not large enough. Moveover, recognition models are often built mainly based on deep learning methods. In this paper, we present a systematic approach in building entity recognition models of Vietnamese documents, beginning with collecting and building data sets then applying and refining machine learning models. In addition to that, we also propose some scenarios of application which proof the capability of our model in dealing with information security problems. Specifically, we built a dataset of more than 5000 documents collected from social networks using Vietnamese, naming and assigning 1 of 4 predefined labels to the entities in the documents and then apply the pre-training model XLM-RoBERTa with the appropriate fine-tuned initial parameters to recognize these entities. Preliminary results show that the proposed system is effective with the ability to recognize the entity of the model and achieve the F1- measure up to 95.6%, which is better than some NER systems curently available for Vietnamese documents on the same dataset which we have built. The proposed model has been used in building support systems for cybersecurity protection currently.**

Từ khóa— nhận dạng thực thể; hệ thống NER; học máy; văn bản Tiếng Việt; tiêu cực; phản động.

Keywords— named entity recognition; NER system; machine learning; Vietnamese text; negative; reactionary.

I. GIỚI THIỆU

Cùng với sự phát triển của cuộc cách mạng công nghiệp lần thứ 4, số lượng người sử dụng Internet ngày càng tăng cùng với sự phát triển nhanh chóng của các mạng xã hội đã kéo theo việc rất nhiều văn bản và hình ảnh được đăng tải trên các trang mạng này. Các thông tin được lưu trữ trên các trang web, các nền tảng mạng xã hội theo cách thức phi cấu trúc, vì vậy việc tìm kiếm thông tin liên quan từ dữ liệu này sẽ mất nhiều

Bài báo được nhận ngày 04/6/2022. Bài báo được nhận xét bởi phản biện thứ nhất ngày 16/6/2022 và được chấp nhận đăng ngày 20/6/2022. Bài báo được nhận xét bởi phản biện thứ hai ngày 27/7/2022 và được chấp nhận đăng ngày 28/7/2022.

thời gian. Tự động xác định và phân loại các thực thể đã được được đặt tên là một nhiệm vụ quan trọng đối với các bài toán xử lý ngôn ngữ tự nhiên, bao gồm khai thác thông tin, truy xuất thông tin và dịch máy. Việc ứng dụng các bài toán khai thác thông tin nhằm chuyển văn bản không có cấu trúc thành một dạng có cấu trúc là yêu cầu cần thiết cho máy tính xử lý và đóng vai trò quan trọng trong việc truy xuất thông tin, trả lời câu hỏi, tóm tắt văn bản,… [5]. Nhận dạng thực thể là một bài toán con của khai thác thông tin và cần được nghiên cứu để giải quyết.

Nhận dạng thực thể được đặt tên (hoặc "được gán nhãn": Named Entity Recognization - NER) là một nhiệm vụ xử lý ngôn ngữ tự nhiên cơ bản để trích xuất các thực thể từ dữ liệu phi cấu trúc. Đây là bước tiền đề cho các bài toán phức tạp trong trích xuất thông tin như trích xuất quan hệ, trích xuất sự kiện,… Báo cáo của D.Wu và các cộng sự [6] đã chỉ ra rằng 60% tổng số truy vấn trong công cụ tìm kiếm là các thực thể có tên. Vì vậy, việc xác định các thực thể được đặt tên từ văn bản phi cấu trúc được nhiều nhà nghiên cứu quan tâm khi xử lý truy vấn và sử dụng trong hệ thống trả lời tự động.

Nhận dạng thực thể được đặt tên trong phân tích dữ liệu mạng xã hội là một công việc trong hoạt động trích xuất, phân tích thông tin mạng xã hội. Trong đó, nhiệm vụ tập trung vào việc tìm kiếm và phân loại các thành phần trong văn bản để đưa chúng vào những loại xác định trước như là tên người, tổ chức, địa điểm, thời gian, số lượng, giá trị tiền tệ và nhiều loại giá trị khác.

Hiện nay, các nghiên cứu về nhận dạng thực thể được đặt tên đang dựa vào các tập dữ liệu cho các ngôn ngữ phổ biến như tiếng Anh, tiếng Pháp, tiếng Đức mà chưa có nhiều nghiên cứu chuyên sâu cho Tiếng Việt. Một số nghiên cứu như trong các công bố [1, 2] đã giải quyết các bài toán nhận dạng thực thể sử dụng các phương pháp học sâu trong miền có giám sát. Bên cạnh đó, phương pháp học sâu cũng được sử dụng trong bài toán khác như phân lớp văn bản [3],nhận dạng hình ảnh [4],…

Các bài toán gán nhãn câu, chuỗi văn bản hiện đang chỉ dừng lại ở các phương pháp dựa trên học máy truyền thống hoặc học sâu. Trong nghiên cứu này, nhóm tác giả áp dụng mô hình luyện trước RoBERTa XLM cho việc nhận dạng thực thể bằng cách lựa chọn các tham số ban đầu tối ưu. Thực nghiệm hệ thống đề xuất được thực hiện với tập dữ liệu Tiếng Việt gồm hơn 5000 văn bản thu thập trực tiếp từ mạng xã hội đã cho kết quả độ đo F1-measure lên đến 95.6%, trội hơn các mô hình đã công bố trước đây.

Về tổng thể, bài báo này có những đóng góp chính như sau:

(1) Thu thập, xây dựng tập dữ liệu hơn 5000 văn bản Tiếng Việt từ các mạng xã hội hiện nay.

(2) Đề xuất hệ thống nhận dạng thực thể được đặt tên hiệu quả áp dụng mô hình luyện trước RoBERTa XLM với việc lựa chọn các tham số ban đầu tối ưu.

(3) Thử nghiệm và đánh giá khả năng nhận dạng của mô hình trên tập dữ liệu đã thu thập được.

Bài báo được cấu trúc gồm 6 phần: Phần II sẽ đánh giá tổng quan một số nghiên cứu liên quan. Trong phần III nhóm tác giả trình bày cách thu thập, xây dựng tập dữ liệu (xác định thực thể, đặt tên/gán nhãn). Phần IV mô tả cụ thể phương pháp nhận dạng thực thể được đặt tên sử dụng trong hệ thống đề xuất. Phần V của bài báo trình bày về các thử nghiệm và đánh giá kết quả thu được. Phần cuối cùng kết thúc bài báo với một số định hướng nghiên cứu trong tương lai.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Nhận dạng thực thể được đặt tên (NER) là một trong các lĩnh vực nghiên cứu phổ biến về xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP). Đây là việc gắn nhãn các chuỗi văn bản, trong đó mỗi mã thông báo trong một chuỗi được gán nhãn thích hợp với các lớp ngữ nghĩa tương ứng của nó. Với một văn bản được mã hoá, mục tiêu là xác định các lớp khác nhau trong văn bản như vị trí, cơ quan, sự kiện,… Có nhiều giải pháp khác nhau đã được đề xuất để trích xuất và phân loại các thực thể được đặt tên trong một văn bản. Mỗi phương pháp đều có những ưu điểm và hạn chế nhất định. Hiện nay, ngày càng có nhiều nhà nghiên cứu quan tâm đến bài toán phân loại văn bản thu thập từ mạng xã hội (bao gồm tin tức và các bài đăng trên mạng xã hội) theo cách phân tích sắc thái để xác định xem một tin tức hoặc bài

báo là tích cực, bình thường, tiêu cực hay thậm chí là phản động mà người đọc quan tâm. Bài toán này có ý nghĩa quan trọng, vì các thông tin, tin tức có tác động to lớn đến tâm lý, cảm xúc và có khả năng chi phối hành động của con người. Một số loại thông tin xấu, độc hại như tin thất nghiệp, thông tin phá sản có thể ảnh hưởng đến thị trường chứng khoán và nền kinh tế của một quốc gia, khu vực. Ngoài ra, với sự phát triển của mạng xã hội và các dịch vụ mạng, các thế lực thù địch đã triệt để lợi dụng không gian mạng thực hiện các hành vi phát tán thông tin giả, thông tin xấu độc, hại trên hàng nghìn website, blog, hàng trăm tờ báo, nhà xuất bản, kênh truyền hình gây ảnh hưởng trực tiếp đến an ninh mạng của một quốc gia.

Về cơ bản, có thể coi trích rút thực thể là việc tìm kiếm và phân lớp các từ (cụm từ) trong văn bản vào các nhóm thực thể như tên người (person), tên địa điểm (location), tên tổ chức (organization), ngày tháng (date), thời gian (time), tỷ lệ (percentage), tiền tệ (monetary),… Gần đây, trích rút thực thể được mở rộng sang nhiều lớp khác như tên protein, chủ đề bài báo, tên tạp chí,... Từ năm 1995, hội thảo quốc tế với chuyên đề "Hiểu thông điệp" (Message Understanding Conference - MUC) lần thứ 6 đã bắt đầu tổ chức nhằm đánh giá các hệ thống NER cho tiếng Anh. Tại hội thảo CoNLL năm 2002 và 2003, các nhà nghiên cứu đã xây dựng và đánh giá các hệ thống NER cho tiếng Hà Lan, Tây Ban Nha, Đức và tiếng Anh. Trong nhiệm vụ đánh giá cho các hệ thống này, các nhà nghiên cứu xét 4 loại thực thể được đặt tên gồm tên người, tên tổ chức, tên địa danh và các tên khác. Sau đó, các cuộc thi về NER vẫn được thường xuyên tổ chức như GermEval 2014 Named Entity Recognition [16] cho tiếng Đức.

Đối với Tiếng Việt, Cuộc thi VLSP (Association for Vietnamese Language and Speech Processing) 2016 [7] đã được đưa ra lần đầu tiên để đánh giá chất lượng các công cụ NER. Đến năm 2021, VLSP 2021 [8] được tổ chức tại Trường Đại học Công nghệ thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh là cuộc thi thứ hai nhằm đưa ra được một đánh giá khách quan về chất lượng các công cụ NER, khuyến khích phát triển các hệ thống trích rút thực thể được đặt tên đạt độ chính xác cao. So với cuộc thi thứ nhất tại VLSP 2016, tập dữ liệu lần này đa dạng, phong phú hơn và được tập hợp theo một số lĩnh vực nhằm có các đánh giá chi tiết hơn về các hệ thống NER.

Nhận dạng thực thể được đặt tên là một bước có ý nghĩa quan trọng tiền đề và cơ sở khi giải quyết các bài toán phức tạp hơn trong xử lý ngôn ngữ tự nhiên. Thực tế đã chứng minh rằng trước khi nhận dạng được quan hệ giữa các thực thể chúng ta phải xác định được đâu là các thực thể tham gia vào mối quan hệ đó. Ban đầu, NER được xem là một thao tác đơn giản trong các bài toán khai thác thông tin. Tuy nhiên, hiện nay NER có một vai trò quan trọng quyết định đến các bài toán phức tạp khác như truy vấn thông tin (Information Retrieval - IR) hoặc các hệ hỏi đáp (Question Answering Systems - QA).

Nhìn chung, các phương pháp được đề xuất để giải quyết cho bài toán trích rút thực thể được chia thành hai nhóm chính, bao gồm nhóm các phương pháp dựa trên tri thức và nhóm các phương pháp dựa trên kỹ thuật học máy. Các hệ thống dựa trên tri thức chủ yếu dựa trên tập luật được xây dựng một cách thủ công, sử dụng các ngôn ngữ đặc thù như văn phạm JAPE. Ý tưởng của kỹ thuật học máy là học các đặc trưng (sử dụng để mô tả thuộc tính của từ) của mẫu dương (lớp tương ứng với các nhãn quan tâm chẳng hạn như tên người, tên địa điểm) và mẫu âm (lớp không tương ứng với nhãn quan tâm, quy ước là nhãn O) từ tài liệu đã gán nhãn. Vì vậy, việc sử dụng kỹ thuật học máy trong xử lý ngôn ngữ Tiếng Việt sẽ đem lại hiệu quả tốt hơn dựa trên tri thức.

Đối với hướng tiếp cận học máy, nhiều kỹ thuật học máy đã được áp dụng cho bài toán trích rút thông tin như mô hình trường ngẫu nhiên có điều kiện (Conditional Random Fields - CRE), máy vectơ hỗ trợ (Support Vector Machine - SVM) [11, 12], mô hình Markov ẩn (Hidden Markov Model - HMM) [9], mô hình Markov Entropy cực đại (Maximum Entropy Markov Model - MEMM) [10],… Với bản chất của quá trình trích rút thực thể là gán nhãn các từ, cụm từ trong văn bản với loại thực thể tương ứng nên chúng ta có thể coi bài toán trích rút thực thể là bài toán phân lớp dữ liệu.

Trong nghiên cứu [13] đã sử dụng trường ngẫu nhiên có điều kiện dùng để nhận dạng Tiếng Việt có tên các thực thể gồm các loại sau: Người, vị trí, tổ chức, thời gian, tài chính và có F1-measure là 81%. Pham và các cộng sự [14] đã mô tả hệ thống sử dụng phương pháp học máy SVM cho nhận dạng thực thể tiếng Việt trong lĩnh vực y tế với F1-measure là 83%. Các nghiên cứu về nhận dạng thực thể được đặt tên Tiếng Việt còn chưa nhiều, khả năng nhận dạng chưa cao do sự đa dạng về ngữ nghĩa. Vì vậy, bài báo sẽ đánh giá, đề xuất hệ thống, xây dựng tập dữ liệu và tiến hành thực nghiệm để xác định các mô hình huấn luyện hiệu quả trong trích rút thực thể phục vụ nhận dạng thực thể được đặt tên trong văn bản Tiếng Việt.

III. XÂY DỰNG TẬP DỮ LIỆU

Tập dữ liệu được bài báo xây dựng bao gồm 5000 mẩu tin tức, bài đăng được thu thập từ các kênh truyền thông xã hội. Tập dữ liệu được thu thập trực tiếp từ các tài khoản trang cá nhân, trang tin, nhóm trên mạng xã hội, nhờ hệ thống thu thập dữ liệu thuộc đề tài nghiên cứu mà các tác giả đang tham gia thực hiện (như trong phần Lời cảm ơn). Tập dữ liệu gồm các nhãn thực thể như Bảng 1 dưới đây:

BẢNG 1. THỰC THỂ TRONG VĂN BẢN TIẾNG VIỆT THU THẬP TỪ MẠNG XÃ HỘI

Tên thực thể	Mô tả thực thể	Các loại thực thể
Tên người (PERSON)	Các loại tên người	▪ Tên, tên đệm và họ của một người. ▪ Tên hiệu (biệt hiệu), bí danh (mật danh), biệt danh. ▪ Tên các nhân vật hư cấu.
Địa chỉ (LOCATION)	Các thực thể có tọa độ địa lý nhất định, ghi lại được trên bản đồ (trừ các địa danh tưởng tượng)	▪ Tên gọi các hành tinh: Mặt Trăng, Mặt Trời, Trái Đất,... ▪ Tên gọi các thực thể mang yếu tố địa lý tự nhiên và địa lý lịch sử (quốc gia, vùng lãnh thổ, châu lục), các vùng quần cư (làng, thị trấn, thành phố, tỉnh, giáo khu, giáo xứ). ▪ Tên gọi các thực thể tự nhiên: Đèo, núi, dãy núi, rừng, sông, suối, hồ, biển, vịnh, vũng, eo biển, đại dương, thung lũng, cao nguyên, đồng bằng, bãi biển, khu bảo tồn thiên nhiên, khu sinh thái,... ▪ Tên gọi các thực thể là công trình xây dựng, công trình kiến trúc công cộng: Cầu, đường, lâu đài, quảng trường, bảo tàng, trường học, nhà trẻ, thư viện, bệnh viện, nhà hát, nhà máy,... ▪ Tên gọi địa điểm, địa chỉ thương mại: Nhà hàng, khách sạn, hiệu thuốc, quán bar,... ▪ Một số địa danh tưởng tượng khác: Vườn Địa Đàng, Sông Ngân, Cầu Ô Thước,…
Thời gian (DATETIME)	Các thực thể có liên quan đến mốc thời gian	▪ Mốc thời gian trên đồng hồ: Giờ, phút, giây,… ▪ Mốc thời gian theo lịch: Ngày, tuần, tháng, năm,... ▪ Mốc thời gian quy ước năm: Thập kỷ, thế kỷ, thiên niên kỷ,…
Từ tiêu cực, phản động (TC_PD)	Các hashtag, cụm từ chứa nội dung tiêu cực hoặc chống phá Đảng Cộng sản Việt Nam và Nhà nước Cộng hòa xã hội chủ nghĩa Việt Nam	▪ Những hành động mang tính chất tiêu cực: bắt giữ, phóng hỏa, khủng bố, đe dọa,... ▪ Những cụm từ mang ý nghĩa chống phá Đảng và Nhà nước: Đảng bán nước, Hội Anh Em Dân chủ, sự cai trị độc tài, tổ chức phản động,…

IV. PHƯƠNG PHÁP ĐỀ XUẤT

A. Đặt tên và biểu diễn thực thể

Tên người (PERSON) được xem là tên riêng. Trong Tiếng Việt, tên người được viết hoa nhằm mục đích chỉ ra rằng người đó chỉ có một, không giống với các tên khác. Trong các văn bản Tiếng Việt, tên người thường đi trước là các danh từ chung như ông, bà, anh, chị, chú, bác, thằng, chủ tịch, giám đốc, trưởng phòng,… Các danh từ này được dùng để chỉ hoặc gọi một người theo mối quan hệ và không mang tính cố định. Vì vậy, các danh từ này không nằm trong cấu tạo tên người. Tên người thường gồm các dạng tên người đầy đủ, tên dạng rút ngọn, tên danh nhân, nhân vật lịch sử, tên hiệu, tên tự, bí danh,…

Tên người dạng đầy đủ: Là tên riêng chỉ ra từng cá nhân và ở dạng đầy đủ, bao gồm 3 phần là họ, tên đệm và tên. Thông thường, chữ cái đầu tiên của các âm tiết được viết hoa do không phân biệt họ, tên đệm và tên do việc riêng hoá. Ví dụ minh hoạ trong Bảng 2 và Hình 1.

BẢNG 2. TÊN NGƯỜI ĐẦY ĐỦ

Họ	Chữ đệm	Tên
Nguyễn	Xuân	Phúc
Hồ	Chí	Minh
Võ	Nguyên	Giáp
Nguyễn	Phú	Trọng

▷ "*thủ tướng Nguyễn Xuân Phúc*"

thủ tướng	O	O
Nguyễn	B-PER	O
Xuân	I-PER	O
Phúc	I-PER	O

Hình 1. Biểu diễn tên người dạng đầy đủ

Tên người dạng rút gọn 2 thành phần: Là tên gồm 2 thành phần là họ và tên hoặc tên và họ hoặc tên đệm và tên. Ví dụ như Hình 2 dưới đây:

▷ "*nhạc sĩ Đỗ Nhuận*"

nhạc sĩ	O	O
Đỗ	B-PER	O
Nhuận	I-PER	O

▷ "*footballer Alan Shearer*"

footballer	O	O
Alan	B-PER	O
Shearer	I-PER	O

Hình 2. Biểu diễn tên người dạng rút gọn

Tên người dạng rút gọn 1 thành phần: Là tên gọi chỉ gồm 1 thành phần là tên. Trường hợp có các danh từ chung là từ xưng hô đứng trước bộ phận tên (hoặc họ với các ngôn ngữ Ấn-Âu) thì các danh từ này không được coi là thuộc tên người. Trường hợp danh từ chung chỉ chức vụ, công việc,... được dùng để gọi thay cho tên người đảm nhiệm chức vụ, công việc đó trong một không gian cụ thể (ngữ cảnh của câu chuyện) thì cũng được coi là tên người (có thể viết hoa theo phong cách) như trong Bảng 3.

BẢNG 3. TÊN NGƯỜI DẠNG RÚT GỌN

Từ xưng hô	Tên
anh	Thành
chị	Thoa
bé	Manh
ông	Hải
bạn	Hồng
cô	Tấm
bà	Lan
chú	Hiếu

▷ "*anh Nam là giáo viên*"

anh	O	O
Nam	B-PER	O
là	O	O
giáo viên	O	O

▷ "*chính phủ Obama*"

chính phủ	O	O
Obama	B-PER	O

Hình 3. Biểu diễn tên gọi

Tên danh nhân, nhân vật lịch sử: Là tên được cấu tạo bằng cách kết hợp giữa bộ phận là danh từ chung chỉ chức vụ, công việc hoặc từ tôn sùng (Y) với bộ phận là tên (CapWord - từ đỉnh) theo dạng:

$$X = Y + CapWord \qquad (1)$$

Ví dụ như Đề Thám, Đội Cấn, Lý Cường, Ông Đùng, Bà Đà, Thánh Gióng, Đức Phật Như Lai,…

▷ *"con cháu Bà Trưng, Bà Triệu"*

con cháu	O	O
Bà	B-PER	O
Trưng	I-PER	O
,	O	O
Bà	B-PER	O
Triệu	I-PER	O

Hình 4. Biểu diễn tên danh nhân, nhân vật lịch sử

Tên người dưới dạng tên hiệu, tên tự, bí danh. Ví dụ như Hình 5:

▷ *"Ức Trai là tên hiệu của Nguyễn Trãi"*

Ức	B-PER	O
Trai	I-PER	O
là	O	O
tên hiệu	O	O
của	O	O
Nguyễn	B-PER	O
Trãi	I-PER	O

Hình 5. Biểu diễn tên người dạng tên hiệu

Tên địa lý (Location) được định nghĩa bao gồm tên quốc gia, tên địa phương phân chia theo khu vực địa lý, tên gọi chỉ thực thể địa lý tự nhiên, tên địa lý được cấu tạo giữa một hoặc hai danh từ chỉ phương hướng, tên được cấu tạo giữa một danh từ chung hoặc một danh từ chỉ hướng với một danh từ chung hoặc một danh từ chỉ hướng, tên các công trình xây dựng, kiến trúc. Tên địa lý được biểu diễn như Hình 6, 7, 8, 9, 10 và Hình 11.

▷ *"nhà nước Việt Nam"*

nhà nước	O	O
Việt Nam	B-LOC	O

▷ *"công dân Bosnia và Herzegovina"* ("Bosnia and Herzegovina" là tên quốc gia)

công dân	O	O
Bosnia	B-LOC	O
và	I-LOC	O
Herzegovina	I-LOC	O

Hình 6. Biểu diễn tên quốc gia

▷ *"chủ tịch Thành phố Hà Nội"*

chủ tịch	O	O
Thành phố	B-LOC	O
Hà Nội	I-LOC	O

▷ *"Thị xã Sông Công"* ("Sông Công" đã trở thành một địa danh hành chính)

Thị xã	B-LOC	O
Sông Công	I-LOC	O

Hình 7. Tên địa phương phân chia theo khu vực địa lý

Hồ	B-LOC	O
Gươm	I-LOC	O
Hồ	B-LOC	O
Hoàn Kiếm	I-LOC	O
Quần đảo	B-LOC	O
Hoàng Sa	I-LOC	O

Hình 8. Tên gọi chỉ thực thể địa lý tự nhiên

▷ *"khu vực Đông Nam Á"*

khu vực	O	O
Đông	B-LOC	O
Nam	I-LOC	O
Á	I-LOC	O

Hình 9. Tên địa lý được cấu tạo giữa một hay hai danh từ chỉ phương hướng

Miền	B-LOC	O
Nam	I-LOC	O
Nam	B-LOC	O
Bộ	I-LOC	O

Hình 10. Tên địa lý được cấu tạo giữa một danh từ chung hoặc một (hai) danh từ chỉ hướng với một danh từ chung hoặc một danh từ chỉ hướng

▷ *"tại Khu Đô thị Time City"*

tại	O	O
Khu	B-LOC	O
Đô thị	I-LOC	O
Time	I-LOC	O
City	I-LOC	O

Hình 11. Tên các công trình xây dựng, kiến trúc được cấu tạo giữa một danh từ chung chỉ loại của công trình

B. Xây dựng mô hình huấn luyện

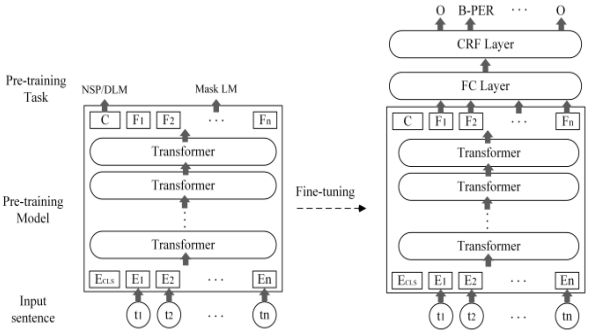
Trong hệ thống nhận dạng thực thể Tiếng Việt, bài báo đề xuất sử dụng mô hình XLM-RoBERTa XML. RoBERTa XML là một phiên bản phát triển cho đa ngôn ngữ của RoBERTa với 100 ngôn ngữ khác nhau. Trong đó, RoBERTa là một mô hình transformer được tiền huấn luyện trên một bộ dữ liệu lớn theo phương pháp tự giám sát. Mô hình này chỉ được huấn luyện với các văn bản thô, không được gán nhãn với cơ chế tự động tạo đầu vào và nhãn từ các

văn bản ban đầu. Mô hình sẽ được tiền huấn luyện với mô hình Masked Language Model (MLM). Trong mô hình MLM, 15% các từ của câu đầu vào được thay thế bởi [MASK] trước khi truyền vào mô hình đại diện cho những từ bị ẩn (masked). Mô hình sẽ dựa trên các từ không ẩn (non-masked) xung quanh [MASK] và đồng thời là ngữ cảnh của [MASK] để dự báo giá trị gốc của từ bị ẩn đi. Số lượng từ được ẩn đi được lựa chọn là một số bé (15%) để mô hình vẫn hiểu được tỷ lệ bối cảnh từ những từ không bị ẩn (chiếm 85%). Bản chất của kiến trúc BERT vẫn là một mô hình seq2seq gồm bộ mã hóa 2 pha giúp nhúng các từ đầu vào và bộ giải mã giúp tìm ra phân phối xác suất của các từ ở đầu ra. Kiến trúc Transfomer-Encoder được giữ lại trong tác vụ Masked ML. Sau khi thực hiện tự giám sát và chuyển tiếp, sẽ thu được các véc-tơ nhúng ở đầu ra là O_1, O_2,…, O_n.

Bên cạnh đó, để tính toán phân phối xác suất cho từ đầu ra, nhóm tác giả thêm một lớp kết nối đầy đủ (Fully Connected) ngay sau lớp Transformer-Encoder dựa trên hàm Softmax. Do số lượng units của lớp Fully Connected phải bằng với số biến mục tiêu nên trong mô hình nhóm tác giả sử dụng là 8 nhằm tương ứng với số chủ đề. Cuối cùng, nhóm tác giả thu được véc tơ nhúng (embedding) của mỗi một từ tại vị trí [MASK], tương ứng với véc-tơ giảm chiều của véc-tơ O_i sau khi đi qua lớp Fully Connected.

Hàm mất mát (Loss Function) của RoBERTa sẽ bỏ qua mất mát từ những từ không bị ẩn đi và chỉ tính mất mát của những từ bị che dấu. Do đó, mô hình sẽ hội tụ lâu hơn nhưng đây là đặc tính bù trừ để gia tăng tri thức về ngữ cảnh văn bản. Việc lựa chọn ngẫu nhiên 15% số lượng các từ bị che dấu cũng tạo ra vô số các kịch bản đầu vào cho mô hình huấn luyện và cho phép mô hình học cách biểu diễn câu theo cả hai chiều.

Sau quá trình tiền huấn luyện, mô hình thu được đã có được tri thức ngữ nghĩa phong phú từ kho ngữ liệu huấn luyện trước không được gắn nhãn thông qua phương pháp học máy không giám sát. Sau đó, chúng tôi sử dụng cách tiếp cận tinh chỉnh (fine tuning) để áp dụng mô hình tiền huấn luyện trong các nhiệm vụ tiếp theo. Lớp Fully Connection (FC) và lớp Conditional Random Field (CRF) được thêm vào sau đầu ra của mô hình huấn luyện trước đó (pretrained model). Các vectơ đầu ra pretrained model có thể được coi là biểu diễn của các câu đầu vào. Do đó, nhóm tác giả sử dụng một lớp FC để có được các biểu diễn cấp cao hơn và trừu tượng hơn. Các từ của chuỗi đầu ra phụ thuộc và liên kết mạnh mẽ với nhau. Ví dụ: "PERSON" phải xuất hiện sau "TC_PD". Cùng với đó, lớp CRF cũng được thêm vào để đảm bảo thứ tự đầu ra của các từ.



Hình 12. Tinh chỉnh mô hình

V. THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Dữ liệu thử nghiệm bao gồm 5000 văn bản Tiếng Việt được thu thập và biểu diễn như mô tả trong mục III và được chia theo tỉ lệ train/val/test tương ứng là 0.7/0.15/0.15.

Nhóm tác giả sử dụng độ đo F1 (F1-measure) để đánh giá kết quả hệ thống nhận dạng thực thể có tên theo công thức dưới đây:

$$F_1 = \frac{2*P*R}{P+R} \quad (2)$$

Trong đó: P là Độ chính xác (Precision) được tính theo công thức:

$$P = \frac{NE_{true}}{NE_{sys}} \quad (3)$$

R là Độ bao phủ (Recall) được tính theo công thức:

$$R = \frac{NE_{true}}{NE_{ref}} \quad (4)$$

Với NE-ref là số thực thể trong dữ liệu gốc; NE-sys là số thực thể được đưa ra bởi hệ thống; NE-true là số thực thể được hệ thống gán nhãn đúng.

Kết quả thử nghiệm hệ thống với tập dữ liệu thu thập được như Bảng 4.

BẢNG 4. KẾT QUẢ THỬ NGHIỆM HỆ THỐNG NHẬN DẠNG THỰC THỂ TIẾNG VIỆT

Tag	F1 Measure
LOCATION	0.935
DATETIME	0.912
PERSON	0.956
TC_PD	0.7

Từ Bảng 4 chúng ta có thể thấy rằng mô hình nhận dạng rất tốt đối với thực thể LOCATION, DATETIME, PERSON khi chỉ số F1 của cả 3 nhãn này đều trên 90%, trong đó PERSON có độ đo F1 lên đến 95.6%. Với nhãn TC_PD (tiêu cực, phản động) do tính đa dạng, không thống nhất và khó nắm bắt nên chỉ số F1 của nhãn này chỉ đạt mức 70%, tuy nhiên đây vẫn là một mức cao cho một bài toán khó trong nhận dạng thực thể tiếng Việt đối với dữ liệu mạng xã hội.

Để đánh giá hiệu quả của mô hình nhận dạng đã đề xuất, nhóm tác giả đã thử nghiệm mô hình của Dat Ba Nguyen và cộng sự [13] cùng với tập dữ liệu chúng tôi đã thu thập kết quả nhận dạng đối với 2 thực thể là "tên người" và "địa chỉ" của 2 mô hình như Bảng 5. Mô hình nhận dạng của nhóm tác giả đã cho kết quả tốt hơn với độ đo F1 vượt trội đối với 2 thực thể tiếng Việt nêu trên.

BẢNG 5. SO SÁNH KẾT QUẢ THỬ NGHIỆM HỆ THỐNG NHẬN DẠNG THỰC THỂ TIẾNG VIỆT

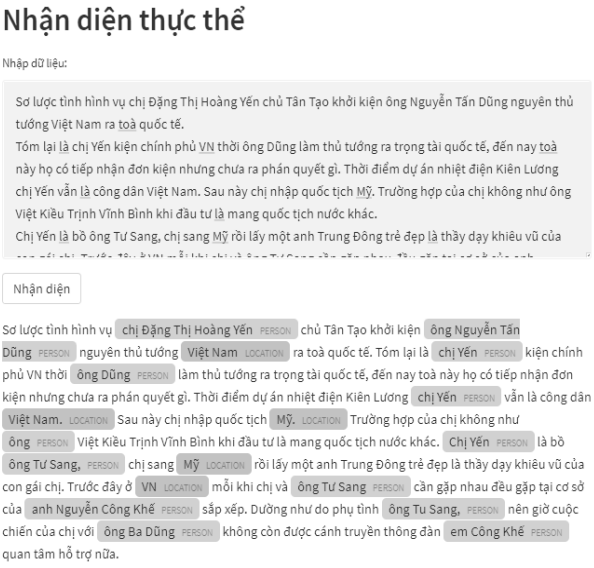
Tag	F1 Measure	
	Mô hình Nguyen và cộng sự	Mô hình chúng tôi đề xuất
PERSON	0.845	0.956
LOCATION	0.886	0.935

VI. KỊCH BẢN ỨNG DỤNG TRONG ĐẢM BẢO AN NINH, AN TOÀN THÔNG TIN

Kịch bản 1: Phát hiện các bài viết trên mạng xã hội sử dụng tiếng Việt đề cập đến các yếu nhân, lãnh đạo các cơ quan, tổ chức.

Công tác phát hiện chủ đề bài viết hoặc tự động thu thập, thống kê các bài viết có nội dung đề cập đến tên người (thường là các yếu nhân, lãnh đạo Đảng và Nhà nước) là một yêu cầu quan trọng trong đảm bảo an ninh, an toàn thông tin. Việc nhận dạng chính xác được tên người trong các bài viết tiếng Việt trên mạng xã hội sẽ góp phần xây dựng các hệ thống phát hiện các bài viết phục vụ mục đích bảo vệ an ninh mạng.

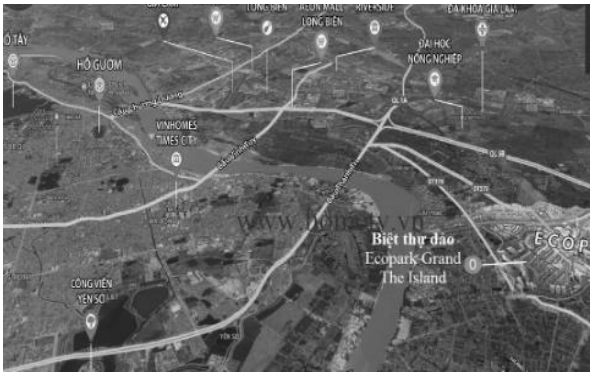
Mô hình nhận dạng thực thể "tên người" được ứng dụng trong xây dựng hệ thống phát hiện tên người, thống kê các bài viết có nội dung đề cập đến tên người như Hình 13.



Hình 13. Hệ thống nhận diện tên người trong bài viết trên mạng xã hội sử dụng mô hình đề xuất

Kịch bản 2: Xây dựng bản đồ di chuyển của đối tượng thông qua nội dung bài viết đối tượng đăng tải trên mạng xã hội.

Việc mô phỏng lại quá trình di chuyển của các đối tượng cần theo dõi, giám sát là công việc cần thiết là quan trọng của cơ quan chức năng trong điều tra tội phạm. Với việc tự động thu thập, nhận dạng các địa điểm đối tượng đã đi qua hoặc thông tin lại từ các bài đăng trên mạng xã hội, cơ quan chức năng nó tiến hành phân tích để khoanh vùng vị trí đối tượng có thể xuất hiện, thực hiện các biện pháp chuyên biệt khác trong đấu tranh với các loại tội phạm hiện nay. Vì vậy, thông qua nội dung các bài viết, mô hình nhận dạng thực thể "địa chỉ" có thể ứng dụng tốt trong bài toán xây dựng bản đồ di chuyển của đối tượng. Ví dụ về việc xây dựng hệ thống mô phỏng bản đồ di chuyển của đối tượng được thể hiện qua Hình 14.



Hình 14. Ví dụ bản đồ di chuyển xây dựng từ các địa điểm phân tích được

Kịch bản 3: Đánh giá mức độ tiêu cực, phản động của một bài viết trên mạng xã hội.

Thông qua mô hình nhận dạng thực thể "từ tiêu cực, phản động", chúng ta có thể xây dựng hệ thống phân loại mức độ tiêu cực, phản động của một bài viết tiếng Việt trên mạng xã hội. Từ đó ứng dụng trong công tác phòng chống tội phạm xâm phạm an ninh mạng.

Mô hình nhận dạng thực thể "từ tiêu cực, phản động" được ứng dụng trong hệ thống phân tích bài viết trên mạng xã hội như Hình 15.

Hình 15.Nhận dạng thực thể "từ tiêu cực, phản động" trong bài viết trên trang Facebook cá nhân

Ngoài 3 kịch bản trên, việc ứng dụng mô hình nhận dạng thực thể được đặt tên trong văn bản Tiếng Việt trong các bài toán đảm bảo an ninh, an toàn thông tin mạng là rất quan trọng và có nhiều ý nghĩa thực tiễn.

VII. KẾT LUẬN

Với sự phát trển của mạng xã hội, nhận dạng các từ khoá, nội dung có tính chất tiêu cực, phản động ngày càng trở nên quan trọng để hỗ trợ các cơ quan chức năng xử lý các vấn đề liên quan. Bài báo đã xây dựng một tập dữ liệu 5000 bài viết trên mạng xã hội sử dụng Tiếng Việt và đưa ra một cách tiếp cận có hệ thống trong việc xây dựng mô hình nhận dạng thực thể được đặt tên đối với các văn bản Tiếng Việt. Kết quả thử nghiệm cho thấy hệ thống hoạt động tốt và có hiệu quả các hơn so với các hệ thống tương tự đã công bố trước đây. Tuy nhiên, với sự đa dạng của các nhãn "tiêu cực, phản động" trong các bài viết, hashtag trên mạng xã hội sử dụng Tiếng Việt nên hệ thống cần có những nghiên cứu sâu hơn về đối tượng này.

Trong tương lai, nhóm tác giả tiếp tục thu thập, tinh chỉnh tập dữ liệu bài viết tiếng Việt trên mạng xã hội để xây dựng các tập dữ liệu lớn hơn, đầy đủ hơn. Bên cạnh đó là tìm hiểu, cải tiến hệ thống, thay đổi các mô hình cơ sở để tận dụng tri thức cho các bài toán mới tốt hơn.

LỜI CẢM ƠN

Nghiên cứu được hỗ trợ từ Đề tài khoa học công nghệ cấp quốc gia mã số ĐTĐL.CN.46/20-C, Bộ Khoa học và Công nghệ Việt Nam. Nhóm tác giả xin cảm ơn Ban Chủ nhiệm đề tài đã hỗ trợ cả về học thuật và tài chính. Đặc biệt, tác giả Nguyễn Ngọc Toàn được tài trợ bởi Tập đoàn Vingroup – Công ty CP và hỗ trợ bởi Chương trình học bổng Thạc sĩ, Tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), Viện Nghiên cứu Dữ liệu lớn, mã số VINIF.2021.TS.128.

TÀI LIỆU THAM KHẢO

[1]. G.I.Parisi, J.Tani, C.Weber and S.Wermter, "Lifelong learning of human actions with deep neural network self-organization", Neural Networks 96, pp.137-149, 2017. *https://doi.org/10.1016/j.neunet.2017.09.001.*

[2]. T.H. Pham and P. Le-Hong, "End-to-end Recurrent Neural Network Models for Vietnamese Named Entity Recognition: Word-level vs. Characterlevel", 2017, *arXiv preprint arXiv:1705.04044.*

[3]. L.Shu, H.Xu and B.Liu, "Doc: Deep open classification of text documents", 2017, *arXiv preprint arXiv:1709.08716.*

[4]. A.A.Rusu, N.C. Rabinowitz, G.Desjardins, H.Soyer, J.Kirkpatrick, K. Kavukcuoglu, and R.Hadsell, "Progressive neural networks", 2016, *arXiv preprint arXiv:1606.0467.*

[5]. N.Patil, A.S.Patil and B.Pawar, "Survey of named entity recognition systems with respect to Indian and foreign languages". Int. J. Comput. Appl. 134, pp.21–26, 2016.*doi=10.1.1.736.1297*

[6]. D. Wu, Y.Zhang, S.Zhao, T.Liu, "Identification of web query intent based on query text and web knowledge", In Proceedings of the 2010 First International Conference on Pervasive Computing, Signal Processing and Applications, Harbin, China, 17–19; pp. 128–131, 2010. *doi: 10.1109/PCSPA.2010.40.*

[7]. VLSP 2016, [Online] https://vlsp.org.vn/vlsp2016.

[8]. VLSP 2021, [Online] https://vlsp.org.vn/vlsp2021.

[9]. D. Bikel, S. Miller, R. Schwartz, R. Weischedel, "A High- Performance Learning Name-finder", Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 194–201, 1998. *arXiv preprint cmp-lg/9803003.*

[10]. A. Borthwick, J. Sterling, E. Agichtein, R,. Grishman, "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition", Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, Canada , 1998. *https://aclanthology.org/W98-1118.pdf*

[11]. Y. Wu, T. Fan, Y. Lee, S. Yen, "Extracting Named Entities Using Support Vector Machines", Bremer, E.G., Hakenberg, J., Han, E.-H(S.), Berrar, D., Dubitzky, W. (eds.) KDLL 2006. LNCS (LNBI), vol. 3886, pp. 91–103, 2006. *https://doi.org/10.1007/11683568_8*

[12]. A. Mansouri, L. Affendey, A. Mamat, "Named Entity Recognition Using a New Fuzzy Support Vector Machine", Proceedings of the International Journal of Computer Science and Network Security, IJCSNS 8(2), pp.320–325, 2008. *https://www.researchgate.net/profile/Lilly-Affendey/publication/251928363_Named_Entity_Recognition_Using_a_New_Fuzzy_Support_Vector_Machine/links/544854050cf22b3c14e30cc5/Named-Entity-Recognition-Using-a-New-Fuzzy-Support-Vector-Machine.pdf*

[13]. T.C. Nguyen, O.T. Tran, H.X. Phan, T.Q. Ha, "Named Entity Recognition in Vietnamese Free-Text and Web Documents Using Conditional Random Fields", Proceedings of the Eighth Conference on Some Selection Problems of Information Technology and Telecommunication, Hai Phong, Viet Nam, 2005.*doi=10.1.1.300.3597*

[14]. Pham, T., Kawazoe, A., Dinh, D., Collier, N.: Construction of Vietnamese Corpora for Named Entity Recognition. In: Conference RIAO 2007, Pittsburgh PA, U.S.A, May 30-June 1, 2007. *doi=10.1.1.106.7855*

[15]. Q.Tri Tran, et al. "Named entity recognition in Vietnamese documents." Progress in Informatics Journal 5, pp. 14-17, 2007.

[16]. GermEval 2014 NER: [Online] https://sites.google.com/site/germeval2014ner/.

SƠ LƯỢC VỀ TÁC GIẢ

Nguyễn Ngọc Toàn

Đơn vị công tác: Học viện An ninh nhân dân.

Email: ngoctoan.hvan@gmail.com

Quá trình đào tạo: Nhận bằng Cử nhân tại Học viện An ninh nhân dân vào năm 2015; Thạc sĩ An toàn thông tin năm 2019 và đang học nghiên cứu sinh chuyên ngành An toàn thông tin tại Học viện Kỹ thuật mật mã.

Hướng nghiên cứu hiện nay: Phát hiện mã độc IoT; học máy trong đảm bảo an toàn thông tin; hệ thống giám sát mạng xã hội, giám sát an ninh mạng.

Lê Xuân Tuấn

Đơn vị công tác: Học viện An ninh nhân dân.

Email: tuanlx.psa@gmail.com

Quá trình đào tạo: Nhận bằng Cử nhân Tin học tại Học viện An ninh nhân dân vào năm 2001; Thạc sĩ Công nghệ thông tin tại Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội vào năm 2006; Tiến sĩ Công nghệ thông tin tại PSL Research University - Cộng hòa Pháp năm 2017.

Hướng nghiên cứu hiện nay: Hệ thống phức hợp; học máy trong đảm bảo an toàn thông tin; hệ thống giám sát mạng xã hội.

Lương Thế Dũng

Đơn vị công tác: Học viện Kỹ thuật mật mã.

Email: theduongluong1@gmail.com

Quá trình đào tạo: Nhận bằng Cử nhân Công nghệ thông tin tại Học viện Kỹ thuật Quân sự vào năm 2001; Tiến sĩ chuyên ngành Bảo đảm toán học cho máy tính và hệ thống tính toán tại Viện Khoa học và Công nghệ Quân sự vào năm 2011.

Hướng nghiên cứu hiện nay: Quyền riêng tư dữ liệu; mật mã học; khai phá dữ liệu; học máy trong đảm bảo an toàn thông tin.

Trần Nghi Phú

Đơn vị công tác: Học viện An ninh nhân dân.

Email: tnphvan@gmail.com

Quá trình đào tạo: Nhận bằng Tiến sĩ tại Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội vào năm 2021.

Hướng nghiên cứu hiện nay: Phát hiện mã độc IoT; phát hiện bất thường; học máy trong đảm bảo an toàn thông tin