PHÁT HIỆN PHÁT NGÔN TIÊU CỰC TRÊN MẠNG XÃ HỘI SỬ DỤNG MÔ HÌNH HỌC SÂU VÀ SỬA LỖI CHÍNH TẢ

Nguyễn Thị Thanh Thủy, Nguyễn Ngọc Điệp
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Sự phát triển của mạng xã hội hiện nay kéo theo xu hướng tự do thể hiện quan điểm cá nhân, kèm theo đó là các phát ngôn tiêu cực ngày càng gia tăng gây nhiều hậu quả xấu đối với xã hội. Việc phát triển các hệ thống nhằm phát hiện phát ngôn tiêu cực là rất cấp thiết, tuy nhiên do tính phức tạp, đa dạng, có nhiều đặc trưng liên quan tới ngôn ngữ, văn hóa của loại văn bản là các bình luận trên mạng xã hội, việc phát hiện chính xác phát ngôn tiêu cực còn gặp nhiều khó khăn, bao gồm cả tiếng Việt. Một số tiếp cận nổi trội gần đây để giải quyết bài toán này là các phương pháp tiên tiến dựa trên kỹ thuật học sâu, được sử dụng nhiều trong lĩnh vực xử lý ngôn ngữ tự nhiên. Bài báo này đề xuất một phương pháp phát hiện phát ngôn tiêu cực trên mạng xã hội sử dụng các kỹ thuật học sâu, trong đó kết hợp các kỹ thuật nhúng từ và nhúng ký tự khác nhau như charCNN, word2vec, BERT và mô hình BiLSTM. Đồng thời, chúng tôi cũng đề xuất phương pháp để tăng cường độ chính xác cho dữ liệu đầu vào là sửa lỗi chính tả tiếng Việt trong bước tiền xử lý dữ liệu. Kết quả cho thấy mô hình đề xuất có độ chính xác tốt hơn so với các mô hình cơ sở khác khi thử nghiệm trên tập dữ liệu ViHSD với các bình luận tiếng Việt trên mạng xã hội.

Từ khóa: phát ngôn tiêu cực, sửa lỗi chính tả, tiếng Việt, BiLSTM, BERT.

I. GIỚI THIỆU

"Phát ngôn tiêu cực" (hate speech) là một thuật ngữ thường được sử dụng để chỉ những phát ngôn chứa những lời lẽ xúc phạm, khêu gợi sự căm ghét, thù hận hoặc ác cảm đối với một người hoặc một nhóm người, có khả năng gây ảnh hưởng xấu đến sự yên bình của xã hội. Theo Ủy ban châu Âu, thuật ngữ này bao gồm tất cả các hình thức phát ngôn gây nên sự thù hận dựa trên sắc tộc, xuất xứ, cũng như tất cả các phát ngôn lăng mạ, kỳ thị, thiếu lòng khoan dung đối với sự khác biệt. Cùng với sự phát triển của mạng xã hội, đi kèm theo là khả năng ẩn danh, các phát ngôn tiêu cực ngày càng xuất hiện phổ biến. Đôi khi, cũng không có lý do cụ thể nào về phân biệt chủng tộc, tôn giáo, giới tính, mà đơn giản chỉ là một người, hay một nhóm người muốn chửi, muốn lăng nhục khi không đồng quan điểm. Người càng nổi tiếng càng dễ trở thành nạn nhân của những phát ngôn kiểu này. Rõ ràng, các phát ngôn tiêu cực gây hại không chỉ cho những nạn nhân mà còn cho toàn bộ cộng đồng xã hội. Do đó, cần phải có công cụ để kiểm soát các bài đăng trực tuyến để phát hiện và lọc bỏ các nội dung phát ngôn tiêu cực này. Tuy nhiên, do tính phức tạp và đa dạng của văn bản trong mạng xã hội, việc xác định các phát ngôn tiêu cực là một công việc có nhiều thách thức. Ví dụ như mạng xã hội Facebook đã phải loại bỏ 1,8 tỷ bình luận vi phạm quy chuẩn cộng đồng, đến từ hơn 100 ngôn ngữ trên toàn thế giới, trong đó có tiếng Việt [1].

Về cơ bản, việc xác định phát ngôn tiêu cực trên mạng xã hội là bài toán phân loại văn bản trong xử lý ngôn ngữ tự nhiên, tương tự như bài toán phân tích quan điểm. Một số mô hình hiệu quả như Long short-term memory (LSTM) [2], [3], mô hình dựa trên BERT [4] và tốt hơn nữa là mô hình kết hợp giữa BERT và CNN [5] đã được giới thiệu để giải quyết bài toán này. Các mô hình dựa trên BERT đã tận dụng được tri thức học được từ các văn bản có sẵn, cho phép phân loại nhanh chóng nội dung với phát ngôn tiêu cực, có thể áp dụng trên nhiều ngôn ngữ. Tuy nhiên, tính hiệu quả của chúng chưa thực sự cao do thiếu ngữ cảnh, thiếu tri thức về văn hóa bản địa. Hơn nữa, các hệ thống đã có chưa được nghiên cứu đầy đủ cho ngôn ngữ tiếng Việt sử dụng trên mạng xã hội, từ đó dẫn đến việc khó có thể sử dụng để giải quyết bài toán này. Cụ thể, khi viết văn bản tiếng Việt trong một ngữ cảnh không chính thức, không sợ bị kiểm duyệt như các tin nhắn ngắn, các đoạn bình luận ngắn trên mạng xã hội, vì một số lý do, nhiều người thường viết văn bản mà không sử dụng dấu, thậm chí viết tắt, hoặc gõ sai chính tả nhưng không sửa lại. Đó có thể là do thực hiện việc gõ văn bản như vậy sẽ tiết kiệm thời gian đáng kể hơn nhiều, nhất là trên thiết bị di động, bất kể phương pháp nhập liệu đang sử dụng, hoặc thậm chí do muốn thể hiện mình. Ví dụ: khi sử dụng phương pháp gõ Telex phổ biến, để gõ cụm từ "*Đường lên thiên đường*", người dùng cần gõ đầy đủ chuỗi Telex với 31 ký tự "*Dduowngf leen thieen dduowngf*"; tuy nhiên thay vào đó, trên mạng xã hội, người ta có thể gõ chuỗi "*Đg leen thjeen đường*" với 25 ký tự Latin, hoặc chuỗi không dấu như "*Duong len thien duong*" với 21 ký tự. Ngoài ra, một số người dùng khác, đặc biệt là người dùng lớn tuổi, không biết cách gõ văn bản tiếng Việt đúng cách do không được học và sử dụng các phần mềm gõ tiếng Việt, hoặc có thể là do không có sẵn phần mềm này. Thêm nữa, các văn bản dạng này còn có rất nhiều từ và ký tự đặc biệt do người dùng tự thêm vào như kiểu "*anh ơiiiiiii*", "*thik j*" hoặc các từ lóng, từ tiếng Anh hay các biểu tượng cảm xúc để gây ấn tượng như hình mặt cười, yêu thích, trái tim, v.v. Do vậy, với dữ liệu văn bản

Tác giả liên hệ: Nguyễn Ngọc Điệp,
Email: diepnguyenngoc@ptit.edu.vn
Đến tòa soạn: 10/2023, chỉnh sửa: 11/2023, chấp nhận đăng: 12/2023.

tiếng Việt đầu vào không chính xác, chưa được huấn luyện trước đó như vậy thì việc áp dụng các mô hình ngôn ngữ hiện đại, được huấn luyện trước sẽ không được hiệu quả lắm.

Trong nghiên cứu này, chúng tôi đề xuất sử dụng mô hình học sâu kết hợp bao gồm word2vec (sử dụng Fasttext [6] , BERT [7], and BiLSTM [3] để giải quyết bài toán phát hiện phát ngôn tiêu cực trong các bình luận tiếng Việt trên mạng xã hội. Mô hình kết hợp này hiệu quả trong các lĩnh vực ngôn ngữ nghèo tài nguyên, nhiều ký hiệu, từ viết tắt không có trong từ điển, với việc kết hợp khả năng tích hợp các đặc trưng cho từ, trích xuất từ mô hình ngôn ngữ có ngữ cảnh như BERT, và mô hình ngôn ngữ phi ngữ cảnh nhưng lại đặc biệt hiệu quả cho các từ mới và phức tạp là Fasttext cùng các đặc trưng từ mức ký tự. Ngoài ra, chúng tôi cũng đề xuất việc tiền xử lý tiếng Việt hiệu quả cho loại văn bản là các bình luận ngắn trên mạng xã hội, đó là sử dụng công cụ khôi phục chính tả bên cạnh các bước tiền xử lý ngôn ngữ thông thường (ví dụ như loại bỏ các từ dừng và các ký tự đặc biệt). Kết quả thực nghiệm trong phần sau cho thấy, mô hình hoạt động có sự cải thiện độ chính xác đáng ghi nhận trên tập dữ liệu văn bản phát ngôn tiêu cực trên mạng xã hội.

Phần còn lại của bài báo được tổ chức như sau. Phần II mô tả các nghiên cứu liên quan. Phần III trình bày đề xuất phương pháp phát hiện phát ngôn tiêu cực trên mạng xã hội. Kết quả và những phân tích thực nghiệm được trình bày trong phần Phần IV. Cuối cùng, Phần V là kết luận bài báo và định hướng nghiên cứu.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong những năm gần đây đã có nhiều nghiên cứu về phát hiện ngôn ngữ tiêu cực trong mạng xã hội. Hướng tiếp cận ban đầu dựa trên các mô hình học máy cơ bản như SVM, Random Forest như trong nghiên cứu của Davidson và cộng sự [8] hoặc của Martin và cộng sự [9]. Gần đây các phương pháp hiệu quả hơn cho bài toán này dựa trên các mô hình học sâu được đề xuất, ví dụ như áp dụng các mô hình huấn luyện trước cho nhiều ngôn ngữ như BERT, RoBERTa [4]. Trong nghiên cứu [5], các tác giả đã kết hợp BERT (Bidirectional Long Short-Term Memory) và CNN để tạo ra một mô hình mạnh mẽ hơn trong việc phát hiện phát ngôn tiêu cực. Tương tự như vậy, nghiên cứu [10] đề xuất kết hợp PhoBERT và CNN để tối ưu cho ngôn ngữ tiếng Việt. Một số nghiên cứu khác đề xuất một mô hình sử dụng mạng nơ-ron BiLSTM [3], hoặc kết hợp các mạng nơ-ron CNN, BiLSTM [2]. Để xử lý các bình luận không chuẩn với các từ không dấu tiếng Việt, nghiên cứu [11] đề xuất việc khôi phục dấu để tăng cường độ chính xác cho dữ liệu đầu vào của mô hình. Việc này giúp cải thiện độ chính xác của mô hình phát hiện phát ngôn tiêu cực một cách rõ rệt.

Nghiên cứu này cũng sử dụng mô hình học sâu kết hợp ưu điểm của BiLSTM tương tự như nghiên cứu [3]. Tuy nhiên chúng tôi còn kết hợp thêm ưu điểm của các các phương pháp biểu diễn từ và ký tự khác nhau để mô hình hiệu quả hơn, bao gồm BERT với khả năng biểu diễn ngôn ngữ có ngữ cảnh, kết hợp với khả năng biểu diễn từ hiệu quả cho các từ mới, ít xuất hiện như Fasttext và charCNN. Các kết hợp này tạo ra mô hình hiệu quả cho việc hiểu văn bản và phát hiện hiệu quả phát ngôn tiêu cực tiếng Việt trên mạng xã hội. Tương tự nghiên cứu [11], chúng tôi cũng thêm một bước trong tiền xử lý để cải thiện độ chính xác của dữ liệu, bao gồm cả khôi phục dấu của các từ tiếng Việt không dấu và khôi phục lại các lỗi chính tả, các từ cố tình viết sai, các từ theo kiểu ngôn ngữ teen.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Ý tưởng chính cho việc đề xuất kiến trúc mô hình phát hiện phát ngôn tiêu cực trong các bình luận trên mạng xã hội là kết hợp nhiều phương pháp biểu diễn từ hiệu quả vào kiến trúc mạng nơ-ron học sâu, bao gồm đặc trưng ngữ cảnh BERT, đặc trưng biểu diễn từ theo kiểu *n*-gram với Fasttext, đặc trưng biểu diễn từ ở mức ký tự. Các đặc trưng biểu diễn từ phi ngữ cảnh biểu diễn mỗi từ bằng một véc-tơ hữu ích trong miền dữ liệu ngôn ngữ hay được sử dụng trên mạng xã hội do các từ ngữ sử dụng thường xuyên, có thể coi như là từ mới trong từ điển của giới trẻ (ngôn ngữ tuổi teen / teen code), ít phụ thuộc vào ngữ cảnh hay các mối liên hệ trong văn bản. Tuy nhiên, các văn bản này không thể tránh khỏi sự mơ hồ, đa nghĩa, phụ thuộc ngữ cảnh, và khi đó đặc trưng ngữ cảnh BERT sẽ rất hữu ích khi có thể biểu diễn chính xác ngữ nghĩa của từ trong câu. Ngoài ra, các đặc trưng dựa trên Fasttext kết hợp với đặc trưng của từ ở mức ký tự rất hiệu quả trong biểu diễn các từ mới. Thêm nữa, đối với các từ viết không quy chuẩn như viết không dấu, viết sai chính tả, sự kết hợp với mô đun sửa lỗi chính tả sẽ giúp ích rất nhiều cho đầu vào của mô hình biểu diễn ngôn ngữ được chính xác, giúp mô hình hoạt động hiệu quả.

Phần dưới đây sẽ trình bày lý thuyết về một số mô hình học sâu có liên quan, sau đó là mô tả bài toán và đề xuất phương pháp phát hiện phát ngôn tiêu cực trong các bình luận trên mạng xã hội dựa trên kết hợp nhiều đặc trưng biểu diễn từ và kiến trúc mạng nơ-ron BiLSTM. Mô hình đề xuất gồm 2 phần chính: (1) xây dựng véc-tơ từ được biểu diễn theo các cách khác nhau và (2) kiến trúc mạng nơ-ron học sâu để đưa ra các dự đoán từ các đặc trưng từ kết hợp.

Về biểu diễn từ, chúng tôi sử dụng các phương pháp trích xuất khác nhau: biểu diễn từ mức ký tự dựa trên mạng CNN, Fasttext để biểu diễn từ theo *n*-gram và mô hình BERT để biểu diễn từ theo ngữ cảnh. Sau đó, kết hợp các đặc trưng có được của các phương pháp biểu diễn từ thành một véc-tơ tổng trước khi cung cấp cho kiến trúc mạng học sâu BiLSTM. Mạng có các lớp BiLSTM để biểu diễn câu và suy luận nhãn tương ứng.

Để hiểu chi tiết hơn về kiến trúc của mô hình đề xuất, trước hết chúng tôi giới thiệu sơ bộ về các mô hình học sâu có liên quan như phần A dưới đây, sau đó mô tả về bài toán và mô hình đề xuất.

A. Một số mô hình học sâu

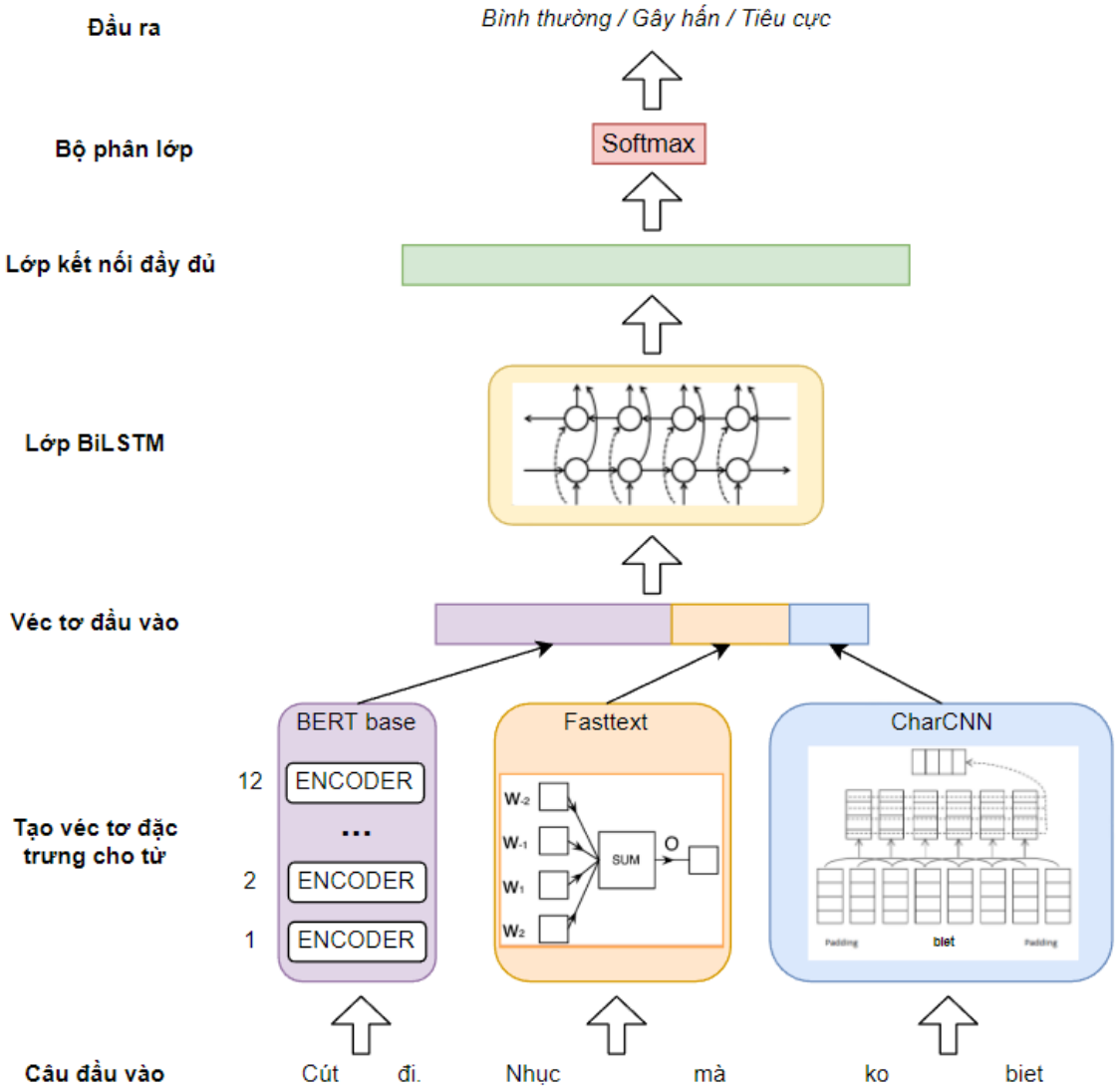
1) Mạng nơ-ron tích chập (CNN)

CNN [12] là mạng rất nổi tiếng do có hiệu năng cao và ít sử dụng các tham số học. Mạng này bao gồm ba loại tầng

là tầng tích chập, tầng gộp và tầng kết nối đầy đủ. Trong tầng tích chập của mạng CNN, phép toán tích chập được thực hiện bằng cách sử dụng một số bộ lọc trượt qua đầu vào và học đặc trưng từ dữ liệu đầu vào. Tầng gộp được sử dụng để kết hợp thông tin qua các vùng không gian kề nhau bằng cách giảm kích thước của tầng trước đó. Có các loại tầng gộp khác nhau bao gồm gộp cực tiểu, gộp cực đại và gộp trung bình. Mạng được kết nối với một tầng kết nối

Mấu chốt của mạng LSTM là tế bào trạng thái, chạy xuyên suốt tất cả các nút mạng, giúp thông tin có thể dễ dàng di chuyển và không bị thay đổi. Việc thêm hoặc bớt thông tin cần thiết cho tế bào trạng thái được thực hiện (sàng lọc) bởi các cổng.

Một LSTM có 3 cổng để duy trì và điều khiển trạng thái của tế bào. Mỗi cổng được kết hợp bởi một tầng mạng sigmoid và một phép nhân. Đầu ra của tầng sigmoid là một



Hình 1. Kiến trúc mô hình học sâu để phát hiện phát ngôn tiêu cực trên mạng xã hội

đầy đủ ở phía cuối để các đặc trưng có thể được ánh xạ phân loại.

2) Mạng bộ nhớ dài-ngắn (LSTM)

Mạng LSTM [6] là một dạng đặc biệt của mạng nơ-ron hồi quy (RNN), được đưa ra để giải quyết vấn đề triệt tiêu gradient trong RNN. LSTM có khả năng học được các phụ thuộc xa, có thể ghi nhớ có chọn lọc các mẫu trong một thời gian dài mà không cần phải huấn luyện (trong khi RNN chỉ có thể xử lý dữ liệu ngắn hạn). LSTM có kiến trúc là dạng chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron, trong đó mỗi mô-đun có 4 tầng tương tác với nhau (khác với RNN chuẩn chỉ có 1 tầng mạng nơ-ron).

giá trị trong khoảng [0, 1], mô tả lượng thông tin cho phép qua. Nếu đầu ra là 1 thì cho tất cả các thông tin đi qua, nếu đầu ra là 0 thì không cho thông tin nào qua cả.

3) Các dạng kiến trúc hai chiều

Để hiểu ngữ cảnh tốt hơn và giải quyết những điểm mơ hồ trong văn bản, các cấu trúc hồi quy hai chiều (bidirectional) được sử dụng để học thông tin trong quá khứ và cả tương lai. Mỗi cấu trúc này có hai loại kết nối, trong đó một loại đi về phía trước theo thời gian và loại còn lại đi lùi lại theo thời gian. Các kết nối nhằm trợ giúp trong việc học các biểu diễn trong quá khứ và tương lai.

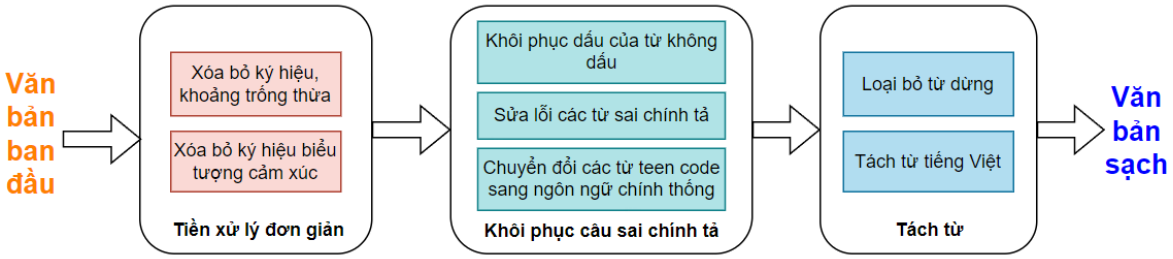
Một số dạng mô-đun có thể có cấu trúc hồi quy hai chiều là RNN, LSTM hoặc GRU.

4) BERT

BERT [7] là viết tắt của cụm từ Bidirectional Encoder Representation from Transformer, có nghĩa là mô hình biểu diễn từ theo hai chiều, ứng dụng kỹ thuật Transformer. BERT được thiết kế để huấn luyện trước các biểu diễn từ. Điểm đặc biệt ở BERT đó là nó có thể điều hòa cân bằng ngữ cảnh theo cả 2 chiều trái và phải.

các hậu tố và tiền tố. Sử dụng Fasttext cũng cho phép biểu diễn ý nghĩa cho các từ không phổ biến, các từ ngữ theo ngôn ngữ tuổi teen trong cách viết không chính thức như trong các bình luận trên mạng xã hội.

Đặc trưng mức từ có được từ Fasttext là đặc trưng phi ngữ cảnh, do đó không mã hóa được các từ đa nghĩa, phụ thuộc ngữ cảnh trong câu. Để nắm bắt được mối tương quan giữa các từ trong một câu, cần sử dụng khả năng biểu diễn từ theo ngữ cảnh từ mô hình BERT [7]. Đây là kỹ thuật học máy dựa trên các Transformer được dùng cho



Hình 2. Các bước tiền xử lý cho dữ liệu bình luận trên mạng xã hội

Cơ chế tập trung (attention) của Transformer sẽ truyền toàn bộ các từ trong câu văn bản đồng thời vào mô hình một lúc mà không cần quan tâm đến chiều của câu. Do đó Transformer được xem như là huấn luyện hai chiều (bidirectional). Đặc điểm này cho phép mô hình học được bối cảnh của từ dựa trên toàn bộ các từ xung quanh nó bao gồm cả từ bên trái và từ bên phải.

Mô hình BERT của Google được huấn luyện trên một kho dữ liệu lớn của văn bản không gán nhãn, bao gồm toàn bộ Wikipedia (lên tới 2500 triệu từ) và Book Corpus (lên tới 800 triệu từ). Khi huấn luyện trên kho dữ liệu lớn như vậy, mô hình học và có được sự hiểu biết thực sự sâu sắc về cách thức hoạt động của ngôn ngữ.

B. Mô tả bài toán

Giả sử cho một văn bản T gồm các câu bình luận trong mạng xã hội. Mỗi câu S đầu vào trong văn bản T được biểu diễn thành một chuỗi các từ (token) $S=w_1w_2...w_n$, với n là số các từ có trong câu. Với mỗi câu đầu vào S, đầu ra của mô hình là một nhãn E là một loại phát ngôn (bình thường/gây hấn/tiêu cực) tương ứng.

C. Mô hình đề xuất

Mô hình đề xuất cho việc phát hiện phát ngôn tiêu cực trong các bình luận trên mạng xã hội là mô hình dựa trên kiến trúc BiLSTM, khai thác sự kết hợp của các đặc trưng ngữ cảnh và phi ngữ cảnh, đặc trưng biểu diễn từ theo mức ký tự. Hình 1 trình bày kiến trúc của mô hình này.

1) Trích xuất đặc trưng

Đặc trưng mức từ có được từ phương pháp nhúng từ với kỹ thuật Fasttext được Facebook giới thiệu [6], có hiệu suất tốt hơn mô hình Word2vec [13] trong nhiều ứng dụng. Nguyên nhân là Fasttext biểu thị mỗi từ dưới dạng *n*-gram ký tự thay vì học trực tiếp véc-tơ cho các từ, từ đó giúp nắm bắt ý nghĩa của các từ ngắn hơn và cho phép biểu diễn

việc học chuyển giao trong xử lý ngôn ngữ tự nhiên do các nhà nghiên cứu tại Google đề xuất. Mô hình này là một mô hình học trước (pre-trained), cung cấp các véc-tơ đại diện theo ngữ cảnh 2 chiều của từ trong câu. Trong mô hình đề xuất, chúng tôi sử dụng một biến thể của BERT là mô hình RoBERTa [14] cỡ nhỏ huấn luyện cho đa ngôn ngữ, gồm 12 lớp là 12 bộ mã hóa (encoder) của mô hình Transformer, mỗi lớp tạo ra một véc-tơ 768 chiều để mã hóa một từ. Vì mỗi lớp trong RoBERTa nắm bắt các cấp độ ngữ cảnh khác nhau, nên sẽ hợp lý hơn khi sử dụng nhúng từ nhiều lớp hơn là chỉ sử dụng lớp cuối cùng. Do đó, chúng tôi nối 3 lớp cuối cùng để tạo thành biểu diễn 2034 (768*3) cho một từ.

Các đặc trưng mức từ rất phổ biến và đạt được nhiều thành công trong các ứng dụng xử lý ngôn ngữ tự nhiên. Tuy nhiên, các đặc trưng này cũng tồn tại một số điểm yếu trong xử lý văn bản là các bình luận trên mạng xã hội vì những văn bản này chứa rất nhiều các ký tự, ký hiệu và các từ không có trong từ điển như ngôn ngữ tuổi teen. Những cụm từ này hầu như không có ý nghĩa trong ngôn ngữ tự nhiên nhưng mang nhiều thông tin. Do đó, chúng tôi sử dụng mô hình CharCNN để trích xuất các véc-tơ nhúng cấp độ ký tự. Mô hình bao gồm các lớp tích chập 1D, maxpooling và lớp kết nối đầy đủ. Mô hình nhận các chuỗi ký tự đầu vào ở dạng one-hot rồi chuyển qua một lớp embedding để ánh xạ các véc-tơ vào một không gian 30 chiều mới. Lớp tích chập 1D tiếp theo bao gồm 30 kernel với kích cỡ là 3 để duyệt trên các véc-tơ này. Sau đó véc-tơ được làm phẳng và chuyển qua một lớp kết nối đầy đủ 128 unit.

Kết hợp các véc-tơ đặc trưng của ba phương pháp biểu diễn từ ở trên bằng cách nối lại với nhau tạo thành một véc-tơ nhiều chiều biểu diễn cho mỗi từ. Véc-tơ này là đầu vào cho mạng nơ-ron sâu BiLSTM được mô tả dưới đây.

2) Kiến trúc mạng nơ-ron BiLSTM

Nhiệm vụ phát hiện phát ngôn tiêu cực trong văn bản là các bình luận trong mạng xã hội được xây dựng dưới dạng bài toán phân loại nhiều đầu ra. Một mạng gồm 2 lớp BiLSTM được sử dụng để chuyển các véc-tơ biểu diễn từ (token) thành véc-tơ biểu diễn câu.

Về cơ bản, LSTM truyền thông tin theo một hướng chỉ có thông tin quá khứ trong lớp không cho phép biết thông tin từ hướng các lớp mạng LSTM hai chiều (BiLSTM) học từ cả hai hướng, cho phép tạo ra các đặc trưng véc-tơ phong phú so với các mô hình LSTM một chiều [6]. Việc áp dụng mô hình này cho phép nắm bắt được nhiều ngữ cảnh nhất có thể, đồng thời còn có thể ngăn ngừa mất mát thông tin. Như thể hiện trong kiến trúc ở Hình 1, các véc-tơ đầu vào kết hợp từ ba phương pháp biểu diễn từ được đưa vào theo cả hai hướng của LSTM. Các đầu ra của BiLSTM lại được sử dụng trong lớp mạng kết nối đầy đủ trước khi vào lớp Softmax nhằm suy luận nhãn cho các từ ban đầu.

3) Tiền xử lý dữ liệu

Chúng tôi sử dụng bộ dữ liệu ViHSD [15] với tổng cộng 33,400 bình luận. Đây là bộ dữ liệu được thu thập từ các trang mạng xã hội, nên chúng chứa các bình luận đa dạng và phức tạp. Đặc biệt, nhiều bình luận trong cả hai bộ dữ liệu chứa các ký tự Unicode không chuẩn, ngôn ngữ tuổi teen, ký hiệu cảm xúc, từ viết tắt và từ chứa ký tự lặp lại. Ngoài ra, nhiều bình luận được viết không dấu. Do đó, chúng tôi tiến hành xây dựng một quy trình tiền xử lý dữ liệu để cải thiện chất lượng của bộ dữ liệu trước khi sử dụng chúng để huấn luyện các mô hình phân loại. Đồng thời, chúng tôi cũng kết hợp sử dụng công cụ khôi phục lỗi chính tả của câu dựa trên công cụ ChatGPT [16] để khôi phục lại các câu. Công cụ có khả năng xử lý các lỗi chính tả, các từ viết tắt đơn giản, và có khả năng khôi phục dấu câu hiệu quả. Ví dụ (1):

Câu đầu vào ChatGPT:

"Vậy đủ chậm chuaaa. Chac chan Coronaviruswuhan qua di thi The Gioi se khong de yen Trung Cong."

Câu đầu ra:

"Vậy đủ chậm chưa? Chắc chắn rồi, khi Coronavirus Wuhan qua đi, thế giới sẽ không để yên Trung Quốc."

Quá trình tiền xử lý bắt đầu với bước tiền xử lý dữ liệu văn bản đơn giản, bao gồm việc xóa bỏ các ký hiệu thừa không cần thiết, các URL, các ký hiệu thể hiện biểu tượng cảm xúc (emoticons). Tiếp tục là bước khôi phục câu sai chính tả, gồm khôi phục dấu đối với các từ không có dấu, sửa lỗi các từ bị viết sai dấu, sai chính tả và chuyển đổi các từ kiểu ngôn ngữ tuổi teen sang ngôn ngữ chính thống. Bước cuối cùng là thực hiện loại bỏ từ dừng và tách từ tiếng Việt. Mục tiêu là đạt được đầu ra là văn bản sạch, theo chuẩn tiếng Việt, phù hợp với các mô hình ngôn ngữ hiện đại, được huấn luyện trước. Hình 2 mô tả tổng quan về quy trình tiền xử lý dữ liệu này.

Đối với bước *Khôi phục câu sai chính tả* trong quá trình tiền xử lý, chúng tôi sử dụng công cụ ChatGPT [16]. Công cụ hoạt động rất hiệu quả trong việc sửa lỗi chính tả của câu, dù vẫn còn tồn tại một số lỗi như xóa bỏ từ (xem ví dụ (1)). Đối với việc tách từ tiếng Việt, chúng tôi sử dụng thư viện PyVi [17], là thư viện sử dụng phổ biến trong các nghiên cứu về xử lý ngôn ngữ tự nhiên cho tiếng Việt.

IV. THỰC NGHIỆM VÀ KẾT QUẢ

A. Tập dữ liệu

Nghiên cứu này sử dụng bộ dữ liệu ViHSD [15] với tổng cộng 33,400 bình luận để thử nghiệm mô hình đề xuất trong nhiệm vụ phát hiện phát ngôn tiêu cực trên mạng xã hội. Các bình luận nằm trong 3 tệp csv, bao gồm tệp dữ liệu huấn luyện, tệp dữ liệu đánh giá và tệp dữ liệu kiểm tra. Số lượng bình luận được chia cho các tập dữ liệu này được phân phối theo tỉ lệ 7-2-1. Mỗi dòng dữ liệu của các file được gán một trong 3 nhãn: "CLEAN (0)", "OFFENSIVE (1)" hoặc "HATE (2)". Phân bố các nhãn trong mỗi tập dữ liệu là giống nhau. Số lượng nhãn "CLEAN" là nhiều nhất, chiếm tỷ lệ 82,7%. Sau đó là nhãn "HATE" với 10.53% và cuối cùng là nhãn "OFFENSIVE" với 6,77%.

Bảng I. Thống kê dữ liệu bình luận trong bộ dữ liệu ViHSD

	CLEAN	**OFFENSIVE**	**HATE**	**TOTAL**
TRAIN	19,886	1,606	2,556	2,2784
DEV	2,190	212	270	2,672
TEST	5,548	444	688	6.680

- Phát ngôn nhãn HATE chứa từ ngữ lăng mạ, thường mang mục đích sỉ nhục cá nhân hoặc nhóm, và có thể có ngôn ngữ tiêu cực, mỉa mai và xúc phạm. Một phát ngôn được gán nhãn HATE nếu nó (1) nhắm vào cá nhân hoặc nhóm dựa trên đặc điểm của họ; (2) thể hiện một ý định gây hại rõ ràng hoặc khêu gợi sự căm ghét; (3) có thể sử dụng hoặc không sử dụng các từ ngữ xúc phạm hoặc lăng mạ.
- Phát ngôn nhãn OFFENSIVE là phát ngôn gây hấn, xúc phạm nhưng không phải phát ngôn tiêu cực (phát ngôn tiêu cực là một bài viết/bình luận có thể chứa các từ ngữ xúc phạm nhưng không nhắm vào cá nhân hoặc nhóm dựa trên đặc điểm của họ).
- Phát ngôn nhãn CLEAN là một phát ngôn bình thường, không phải là bài bình luận gây xúc phạm và cũng không tiêu cực. Đó là cuộc trò chuyện, thể hiện cảm xúc một cách bình thường, không chứa từ ngữ xúc phạm hoặc giọng điệu tiêu cực.

B. Thiết lập thực nghiệm

Hiệu năng của mô hình trích xuất được đo bằng độ đo F_1, được tính từ độ chính xác (precision), độ bao phủ (recall) theo các công thức như sau:

$$Precision = \frac{|A \cap B|}{|A|}$$

$$Recall = \frac{|A \cap B|}{|B|}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Các tham số A và B ở công thức trên tương ứng là tập các nhãn được phát hiện và tập các nhãn đúng (được gán nhãn bởi người gán nhãn). Thử nghiệm được thực hiện với việc huấn luyện mô hình dựa trên tập dữ liệu huấn luyện (train), tối ưu mô hình dựa trên tập dữ liệu đánh giá (valid) và đánh giá mô hình dựa trên tập dữ liệu kiểm tra (test).

Chúng tôi áp dụng cơ chế *mini-batch* để huấn luyện mô hình đề xuất, trong đó: *batch size* là 128; bộ tối ưu Adam optimizer được sử dụng với *learning rate* là $1e^{-5}$, độ dài tối đa của câu đầu vào là 80. Chúng tôi cũng áp dụng cơ chế dừng sớm để ngăn mô hình bị tình trạng quá khớp. Cụ thể, quá trình huấn luyện sẽ dừng khi hiệu suất trên tập dữ liệu kiểm chứng không được cải thiện nào trong ít nhất 5 *epoch* liên tiếp. Số Transformer block là 12, với kích thước của véc-tơ trạng thái ẩn là 768. Mô hình BERT đã đào tạo trước sử dụng cho đa ngôn ngữ *multilingual-bert-base*.

C. Kết quả thực nghiệm

Phần dưới đây sẽ mô tả các thực nghiệm để đánh giá các đặc trưng quan trọng cũng như hiệu năng của mô hình phát hiện bình luận tiêu cực trong mạng xã hội đã đề xuất khi so sánh với các mô hình cơ sở khác.

1) Đánh giá hiệu năng của các kết hợp đặc trưng khác nhau

Chúng tôi sử dụng kết hợp một số đặc trưng đã đề xuất cho mô hình để hiểu rõ hơn về đóng góp của từng đặc trưng đối với độ chính xác của mô hình phát hiện phát ngôn tiêu cực. Nhiều cách kết hợp đặc trưng biểu diễn từ khác nhau đã trình bày ở trên được thực hiện, hoặc bỏ lần lượt từng đặc trưng biểu diễn từ trong tổ hợp các đặc trưng, sau đó kết hợp với mạng nơ-ron với các lớp BiLSTM để trích xuất thực thể. Để thuận tiện cho việc so sánh kết quả với các mô hình trong các nghiên cứu khác ở phần sau, trong phần này chúng tôi chỉ thực hiện việc tiền xử lý dữ liệu đơn giản, chưa có khôi phục lỗi câu sai chính tả.

Kết quả thử nghiệm trong Bảng II cho thấy, đặc trưng mức ký tự đóng vai trò quan trọng trong việc phân loại chính xác các phát ngôn. Hiệu suất của mô hình BiLSTM dựa trên CharCNN và CharRNN (đặc trưng số 1 và 2) tốt hơn mô hình không sử dụng đặc trưng mức ký tự này (đặc trưng số 5 và 6). So sánh giữa hai phương pháp biểu diễn từ ở mức ký tự là CharCNN và CharRNN, mô hình dựa trên CharCNN (đặc trưng số 1 và số 3) đạt được độ chính xác cao hơn từ 0,17% đến 0,84% so với các mô hình dựa trên CharRNN (đặc trưng số 2 và số 4 trong Bảng I).

Bảng II. Hiệu năng của các kết hợp đặc trưng khác nhau trong mô hình đề xuất

STT	Đặc trưng	F1-macro (%)
1	BERT-CharCNN-Fasttext	**63,48**
2	BERT-CharRNN- Fasttext	63,31
3	CharCNN- Fasttext	62,16
4	CharRNN- Fasttext	61,32
5	BERT- Fasttext	63,24
6	BERT-Glove	62,93
7	BERT-CharCNN-Glove	63,15

Đối với đặc trưng BERT, có thể thấy rằng việc thêm biểu diễn từ sử dụng BERT đã tăng hiệu suất tổng thể đáng kể, từ 62,16% (không có BERT) lên tới 63,48% (có BERT) (cặp đặc trưng số 1 và số 3). Mức tăng lên tới gần 2% khi xem xét cặp đặc trưng số 2 và số 4. Mức tăng này hơn hẳn các mức tăng còn lại của các cách kết hợp đặc trưng khác, cho thấy tầm quan trọng của BERT với khả năng biểu diễn từ theo ngữ cảnh thật sự hiệu quả.

Để đánh giá sự phù hợp của phương pháp biểu diễn dựa trên nhúng từ đối với khả năng phát hiện phát ngôn tiêu cực trong bình luận trên mạng xã hội, chúng tôi thay thế Fasttext bằng Glove, một phương pháp nhúng từ bằng véc-tơ toàn cục [13] (các đặc trưng số 1, 5 so với các đặc trưng số 6, 7 trong Bảng I). Mặc dù chênh lệch không lớn nhưng Fasttext vẫn thể hiện hiệu năng tốt hơn so với Glove. Điều này chứng tỏ biểu diễn từ kiểu *n*-gram của Fasttext phù hợp hơn với dạng văn bản bình luận trên mạng xã hội hơn như đã phân tích trước đó.

2) Đánh giá hiệu năng của mô đun Khôi phục câu sai chính tả trong Tiền xử lý dữ liệu

Đa phần các nghiên cứu về xử lý ngôn ngữ tiếng Việt đều thực hiện việc tiền xử lý dữ liệu với các bước thuộc 2 mô đun Tiền xử lý dữ liệu đơn giản và Tách từ, mà không có thực hiện khôi phục câu sai chính tả. Để đánh giá hiệu năng của quy trình Tiền xử lý dữ liệu đề xuất, với sự có mặt của mô đun *Khôi phục câu sai chính tả*, chúng tôi sẽ sử dụng mô hình học sâu đề xuất với đặc trưng có hiệu năng tốt nhất là BERT-CharCNN-Fasttext, có sử dụng Khôi phục câu sai chính tả và so sánh kết quả với chính mô hình đó nhưng không sử dụng khôi phục câu sai chính tả. Mô hình sử dụng BiLSTM và bộ phân lớp softmax với để phân lớp. Kết quả thử nghiệm trong Bảng III cho thấy, độ chính xác được thể hiện thật sự đáng kể với mức tăng lên tới hơn 2%.

Bảng III. Hiệu năng của các mô hình với đặc trưng kết hợp

STT	Tiền xử lý	F1-macro (%)
1	CÓ mô đun Khôi phục câu sai chính tả	**65,89**

2	KHÔNG CÓ mô đun Khôi phục câu sai chính tả	63,48

Bảng IV. Một số ví dụ về các bình luận bị gán sai nhãn, sau khi có Tiền xử lý khôi phục lỗi chính tả đã gán nhãn đúng

STT	Tiền xử lý đơn giản	Tiền xử lý có khôi phục lỗi chính tả
1	Lũlên mạng lừa đão	Lũ lên mạng lừa đảo
2	Con di do noi doc giong 3 vang ma	Con đĩ đó nói dóc giống 3 vàng mã
3	Dem me minh ra lam tro cuoi noi tuc tieu khg hay dau	Đem mẹ mình làm trò cười nói tục tĩu không hay đâu
4	Du me	Đù mẹ
5	Con may bi khung	Con mây bị khùng
6	Con nay tamthan nangkhung batron	Con tâm thần nặng khùng ba trợn
7	Bat dau tien trinh ban nuoc cua bon cs	Bắt đầu tiến trình bán nước của bọn cs
8	Manh thi song yeu thi chet	Mạnh thì sống yếu thì chết

Việc phân tích lỗi cho thấy rằng, nhiều lỗi dự đoán sai nhãn đến từ các câu không có dấu hoặc viết theo ngôn ngữ tuổi teen. Sau khi khắc phục các lỗi chính tả và khôi phục dấu trong câu nhờ công cụ ChatGPT, mô hình đã dự đoán chính xác nhãn. Bảng IV liệt kê một số ví dụ các bình luận bị dự đoán sai nhãn và sau khi khôi phục lỗi chính tả, mô hình đã dự đoán đúng.

2) So sánh hiệu năng của mô hình đề xuất với các mô hình khác

Để đánh giá hiệu năng của kiến trúc mạng nơ-ron đề xuất kết hợp với quá trình Tiền xử lý có khôi phục câu sai lỗi chính tả, chúng tôi sẽ so sánh kết quả với các mô hình đề xuất trong các nghiên cứu khác về phát hiện bình luận, phát ngôn tiêu cực trên mạng xã hội. Các mô hình được so sánh bao gồm TextCNN [15] , BiLSTM [3], PhoBERT-CNN [10]. Tất cả các mô hình này đều được đánh giá trên tập dữ liệu ViHSD [15].

Bảng V. Hiệu năng của các mô hình với đặc trưng kết hợp

STT	Mô hình	F1-macro (%)
1	TextCNN [15]	60,68
2	BiLSTM [3]	61,56
3	PhoBERT-CNN [10]	62,66
4	BERT-CharCNN-Fasttext	63,46
5	Mô hình đề xuất	65,89

Như thể hiện trong Bảng V, mô hình đề xuất vượt trội hơn so với các phương pháp khác trên tập dữ liệu ban đầu với giá trị F_1 là 65,89 %, tốt hơn hơn 3,23% so với mô hình PhoBERT-CNN [10] và 4,43% so với mô hình BiLSTM [3]. Mô hình BERT-CharCNN-Fasttext không có mô đun Sửa lỗi chính tả cũng có giá trị F_1 cao hơn so với các mô hình khác. Kết quả này có được do sự kết hợp nhiều phương pháp trích xuất các véc-tơ nhúng từ cả ở mức độ ký tự (CharCNN) và mức độ từ khác nhau, kể cả có ngữ cảnh (BERT) và không có ngữ cảnh (Fasttext). Ngoài ra, mạng nơ-ron dựa trên BiLSTM có thể nắm bắt các đặc điểm của cả hai hướng bao gồm phần văn bản trước phần văn bản sau trong câu, giúp tăng độ chính xác của mô hình. Một yếu tố quan trọng nữa chính là mô đun Sửa lỗi chính tả, giúp tăng giá trị F_1 lên 2,43% so với kiến trúc khi không sử dụng sửa lỗi chính tả. Việc sửa lỗi chính tả làm tăng độ chính xác đầu vào dữ liệu của mô hình BERT, giúp mô hình BERT xử lý hiệu quả hơn. Điều này cũng cho thấy tầm quan trọng của BERT với khả năng biểu diễn từ theo ngữ cảnh thật sự hiệu quả.

V. KẾT LUẬN

Nghiên cứu này đã đề xuất một mô hình học sâu để phát hiện chính xác phát ngôn tiêu cực trong các bình luận trên mạng xã hội. Mô hình đề xuất có sự kết hợp của ba phương pháp biểu diễn từ gồm BERT, Fasttext và biểu diễn từ theo mức ký tự dựa trên CNN, cùng với kiến trúc mạng BiLSTM. Kết quả của nghiên cứu cho thấy, kết hợp ưu điểm của các phương pháp biểu diễn từ khác nhau gồm: BERT – biểu diễn từ mang thông tin ngữ cảnh trong câu; Fasttext – đặc trưng phi ngữ cảnh mang thông tin ngữ nghĩa của từ, hỗ trợ tốt các từ mới trong văn bản; và đặc trưng CharCNN – ký tự mang thông tin hình thái, tiền tố và hậu tố của từ, cùng với mạng học sâu BiLSTM, tốt hơn so với các mô hình học sâu khác trong bài toán phát hiện bình luận tiêu cực trên mạng xã hội. Ngoài ra, phương pháp tiền xử lý dữ liệu Sửa lỗi chính tả được đề xuất dựa trên ChatGPT có hiệu quả đáng kể, với mức tăng gần 2,5% so với thử nghiệm trên cùng tập dữ liệu.

Trong những nghiên cứu tới, chúng tôi sẽ xem xét mô hình với ứng dụng của mô hình PhoBERT cho ngôn ngữ tiếng Việt và một số phương pháp tăng cường dữ liệu do dữ liệu bình luận tiêu cực không cân bằng.

TÀI LIỆU THAM KHẢO

[1] Statista, “Statista: Global number of hate speech-containing content removed by Facebook from 4th quarter 2017 to 2nd quarter 2021,” *Statista*. 2018. [Online]. Available: https://www.statista.com/statistics/1013804/facebook-hate-speech-content-deletion-quarter

[2] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, “Hate speech detection on vietnamese social media text using the bi-gru-lstm-cnn model,” *in Proceedings of VLSP 2019*, 2019.

[3] H. T.-T. Do, H. D. Huynh, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Hate speech detection on vietnamese social media text using the bidirectional-lstm model," *arXiv preprint arXiv:1911.03648*, 2019.
[4] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, 2020, pp. 928–940.
[5] A. Safaya, M. Abdullatif, and D. Yuret, "Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media," *arXiv preprint arXiv:2007.13184*, 2020.
[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans Assoc Comput Linguist*, vol. 5, pp. 135–146, 2017.
[7] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, p. 2.
[8] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, 2017, pp. 512–515.
[9] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, "Hate speech classification in social media using emotional analysis," in *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, 2018, pp. 61–66.
[10] K. Quoc Tran, A. Trong Nguyen, P. G. Hoang, C. D. Luu, T.-H. Do, and K. Van Nguyen, "Vietnamese hate and offensive detection using PhoBERT-CNN and social media streaming data," *Neural Comput Appl*, vol. 35, no. 1, pp. 573–594, 2023.
[11] P. Le-Hong, "Diacritics generation and application in hate speech detection on Vietnamese social networks," *Knowl Based Syst*, vol. 233, p. 107504, 2021.
[12] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit*, vol. 77, pp. 354–377, 2018.
[13] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
[14] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[15] S. T. Luu, K. Van Nguyen, and N. L.-T. Nguyen, "A large-scale dataset for hate speech detection on vietnamese social media texts," in *Advances 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, Pro*, 2021, pp. 415–426.
[16] T. Wu *et al.*, "A brief overview of ChatGPT: The history, status quo and potential future development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023.
[17] "Python Vietnamese Core NLP Toolkit." [Online]. Available: https://github.com/trungtv/pyvi

HATE SPEECH DETECTION ON SOCIAL NETWORKS USING DEEP LEARNING MODEL AND SPELLING CORRECTION

Abstract: The current development of social media is accompanied by a trend of free expression of personal opinions by netizens. However, this also leads to an increasing prevalence of hate speech, which have detrimental consequences for society. Developing systems for detecting hate speech is crucial, but due to the complexity and diversity of linguistic and cultural features in social media comments, accurately identifying hate speech remains challenging. Recently, there have been various approaches to address this issue, with deep learning methods standing out as advanced techniques commonly used in natural language processing. In this paper, we propose a method for detecting hate speech on social media using deep learning techniques, which combines various embedding techniques, including charCNN, word2vec, BERT, and BiLSTM models. Additionally, we propose a method to enhance input data accuracy by performing spelling correction during data preprocessing step. The results indicate that the proposed model achieves higher accuracy compared to other baseline models when tested on the ViHSD dataset containing hate speech from social media.

Keywords: hate speech, spelling correction, Vietnamese, BiLSTM, BERT.

Nguyễn Thị Thanh Thủy. Nhận học vị Thạc sĩ năm 2009. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, xử lý ngôn ngữ tự nhiên.

Nguyễn Ngọc Điệp. Nhận học vị Tiến sĩ năm 2017. Hiện đang công tác tại Khoa Công nghệ Thông tin 1 và Lab Học máy và ứng dụng, Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: học máy, an toàn thông tin, xử lý ngôn ngữ tự nhiên.