Bài toán phân loại văn bản truyền thông xã hội trên tiếng Việt

Nguyễn Thị Phương,Nguyễn Đức Duy Anh, Trang Hoàng Nhựt,Nguyễn Ngọc Quí, Lê Thị Minh Hiền

***Tóm tắt nội dung*—Phân loại văn bản là một chủ đề phổ biến của xử lý ngôn ngữ tự nhiên, và nó hiện đang là chủ đề thu hút nhiều người nghiên cứu trên thế giới. Có nhiều nghiên cứu về lĩnh vực này bằng các ngôn ngữ khác nhau nhưng chỉ giới hạn ở tiếng Việt. Do đó, nghiên cứu của nhóm nhằm mục đích phân loại các văn bản tiếng Việt trên mạng xã hội từ ba bộ dữ liệu chuẩn tiếng Việt khác nhau (UIT-VSFC, UIT-VSMEC, UIT-ViCTSD). Các mô hình mà chúng tôi sử dụng và tối ưu hóa trong nghiên cứu này bao gồm: SVM+TFIDF, Naive Bayes+TFIDF, Naive Bayes + PhoW2V, BiLSTM + PhoW2V, BiGRU + PhoW2V, PhoBERT. Trải qua quá trình tìm hiểu và thực hiện nhóm thấy được phương pháp PhoBERT phù hợp cho cả ba bộ dữ liệu. Với kết quả được thấy rõ như sau: Đối với bộ dữ liệu UIT-VSFC thì phương pháp PhoBERT không tiền xử lý cho kết quả có thông số đánh giá F1-Score trên khía cạnh cảm xúc là 92.98% cao hơn bài báo [11]. Tuy nhiên, trên khía cạnh chủ đề lại cho thông số F1-Score cao nhất đối với phương pháp BERT đã qua tiền xử lý là 89.38%. Đối với bộ dữ liệu UIT-VSMEC phương pháp PhoBERT không tiền xử lý cho kết quả có thông số đánh giá F1-Score là 62.71% một con số khá lý tưởng so với bài đánh giá đầu tiên trên bộ dữ liệu[4]. Còn phương pháp PhoBERT sau tiền xử lý trên bộ dữ liệu UIT-ViCTSD có thông số đánh giá cao nhất với F1-Score cho Tính xây dựng và Tính tiêu cực lần lượt là: 82.68% và 72.92% cao hơn khá nhiều so với cùng phương pháp PhoBERT được thực hiện trong bài báo[10] với tính Xây dựng cao hơn 4.09% và tính Tiêu cực cao hơn 13.52%.**

I. Giới thiệu

Phân loại văn bản hay còn được gọi là Text Classification là một bài toán thuộc về lĩnh vực Xử lý ngôn ngữ tự nhiên dưới dạng văn bản. Ứng dụng của công nghệ phân loại văn bản đang phát triển mạnh mẽ ở rất nhiều lĩnh vực như học thuật, kinh doanh, y tế, ... Trong bài báo cáo này, nhóm thực hiện phân loại văn bản trên một số khía cạnh được xem là đáng được quan tâm trong xã hội ngày nay.

Đầu tiên là trong lĩnh vực giáo dục. Hiện nay, chất lượng giáo dục đang ngày càng được chú ý và việc phấn đấu nâng cao chất lượng đào tạo luôn là ưu tiên hàng đầu của các cơ sở giáo dục. Để có thể đánh giá đúng thực trạng hiện tại và thấy được các khía cạnh cần được cải thiện thì mỗi cơ sở phải nắm bắt được suy nghĩ của chính những người trực tiếp tiếp xúc và hưởng lợi từ nó, đặc biệt là sinh viên. Nguồn có thể cung cấp dữ liệu lớn nhất Việt Nam cho việc đánh giá chất lượng giáo dục hiện tại đó là:[16] Phản hồi từ sinh viên. Việc thu thập những đánh giá này về các khía cạnh như Giảng viên, Giáo trình, Cơ sở vật chất và những khía cạnh khác sẽ giúp cho nhà trường có thể hiểu hơn về những suy nghĩ cũng như mong muốn của các sinh viên về mỗi khía cạnh. Ngoài ra, dựa vào những phản hồi này, chúng ta có thể thấy được thái cực cảm xúc mà nó bộc lộ (tích cực, tiêu cực hoặc trung tính), từ đó có thể dễ dàng khoanh vùng các khía cạnh cần được cải thiện.

Tiếp đến là việc phân tích cảm xúc. Ngôn ngữ không chỉ bao gồm thông tin mà con mang theo cái hồn của người viết. Muốn biết được cảm xúc của một người thì việc xem qua những nội dung mà họ viết cũng có thể tự đoán được phần nào. Có thể thấy, cảm xúc quyết định chất lượng cuộc sống của mỗi người và việc thúc đẩy những cảm xúc tích cực cũng như xua đi các cảm xúc tiêu cực trong một ngày là điều ai ai cũng muốn. Ngoài biểu hiện ở khuôn mặt, thì còn có rất nhiều nguồn khác để sử dụng nhằm phân tích cảm xúc. Trong đó, văn bản được xem là một trong những nguồn được xem là phổ biến nhất vì tiềm năng mà nó đem lại trong các lĩnh vực như: tiếp thị, bảo mật, tâm lý học, trí tuệ nhân tạo, ... [4]

Cuối cùng là trong khía cạnh mạng xã hội. Sự gia tăng của truyền thông xã hội đã dẫn đến sự gia tăng bình luận trên các diễn đàn trực tuyến, tình trạng thiếu kiểm soát khi bình luận cũng ngày một tăng. Một số bình luận có thông tin không liên quan cũng như không cung cấp kiến thức cho người đọc. Bên cạnh đó, cũng có một vài bình luận mang sắc thái khá tiêu cực và ảnh hưởng không tốt đến người dùng. Điều này ảnh hưởng đến chất lượng của các cuộc thảo luận trực tuyến. Vì vậy, việc kiểm soát chất lượng của các bình luận trên các trang mạng xã hội cũng ngày càng trở nên cấp thiết. Việc lọc ra những bình luận tích cực, mang tính đóng góp xây dựng, cũng như ngăn chặn những thông tin tiêu cực là vô cùng hữu ích để cải thiện chất lượng cuộc thảo luận trực tuyến, đồng thời mang lại kiến thức cho người dùng.

Nhóm thực hiện phân tích trên ba bộ dữ liệu tiếng Việt, gồm: UIT-VSFC, UIT-VSMEC và UIT-ViCTSD. Nhóm áp dụng 5 phương pháp: SVM + TF-IDF, Naive Bayes + TF-IDF/PhoW2V, BiLSTM + PhoW2V, BiGRU + PhoW2V và PhoBERT vào 3 bộ dữ liệu để xem xét đâu là phương pháp phù hợp nhất đối với mỗi bộ dữ liệu thông qua các thông số đánh giá như Precision, Recall, F1-Score,...Đồng thời xây dựng một ứng dụng phân loại văn bản trên tiếng Việt

II. Công trình liên quan

A. Công trình liên quan đến bộ dữ liệu UIT-VSFC

Năm 2018, Nguyen Van Kiet và các đồng tác giả đã cho ra đời bộ dữ liệu Vietnamese Students' Feedback Corpus for Sentiment Analysis (UIT-VSFC) [16]. Bộ dữ liệu có hơn 16,000 phản hồi từ sinh viên được thu thập trong 8 học kỳ. Đây là bộ dữ liệu miễn phí với 2 nhiệm vụ là phân tích cảm xúc của phản hồi và phân tích chủ đề được đề cập đến. Bộ dữ liệu đã được đánh giá trên 2 phương pháp Naive Bayes và Maximum Entropy classifier. Kết quả thu được từ MaxEn được đánh giá cao hơn với chỉ số F1-score đối với khía cạnh cảm xúc là 88% và đối với khía cạnh chủ đề là 84%.

Cũng trong thời gian đó, Vu Duc Nguyen cùng các đồng tác giả cho ra mắt một công trình nghiên cứu mang tên Variants of Long Short-Term Memory for Sentiment Analysis on Vietnamese Students' Feedback Corpus [12]. Nghiên cứu này đã kết hợp hai mô hình Long Short-Term Memory và Deep Tree-Long Short-Term Memory vào SVM cho mục đích phân tích cảm xúc đối với bộ dữ liệu UIT-VSFC[16][12]. Kết quả đạt được là chỉ số F1-Score đạt được lên đến 90,2% và accuracy là 90,7%

Ngoài ra, trong năm 2018, Phu X. V. Nguyen cùng các đồng tác giả cũng cho ra mắt một công trình lên quan tới bộ dữ liệu trên Deep Learning versus Traditional Classifiers on Vietnamese Student' Feedback Corpus [11]. Trong đó, bộ dữ liệu được đánh giá trên 4 phương pháp Naive bayes, Maximum Entropy, Long Short-Term Memory và Bi-Directional Long Short-Term Memory. Từ đó thấy được Bi-LSTM là phương pháp tốt nhất với F1-Score cho 2 nhiệm vụ phân tích cảm xúc và chủ đề lần lượt là 92% và 89.6%. Đồng thời, nhóm cũng phát triển một ứng dụng để phân tích phản hồi từ sinh viên và cung cấp các báo cáo tổng quan cho quản trị viên [11].

B. Công trình liên quan đến bộ dữ liệu UIT-VSMEC

Năm 2020, Vong Anh Ho cùng các đồng tác giả đã công bố bộ dữ liệu Vietnamese Social Media Emotion Corpus (UIT-VSMEC) với 6.927 câu bình luận được gắn nhãn cảm xúc. Và dựa trên bộ dữ liệu đó, họ đánh giá và đo lường trên bốn thuật toán. Sau cùng với mô hình CNN (deep learning) đã cho ra hiệu suất cao nhất với điểm F1- score có trọng số là 59,74% [4].

Cùng năm 2020, sau công bố bộ dữ liệu Vietnamese Social Media Emotion Corpus (UIT-VSMEC) từ tháng 1 thì đến tháng 10 Khang Phuoc-Quy Nguyen và Kiet Van Nguyen đã công bố thêm một bài báo [8] nhằm sử dụng bộ dữ liệu UIT-VSMEC để thực hiện nghiên cứu dùng các kỹ thuật xử lý dữ liệu phù hợp, Multinomial Logistic Regression (MLR) [4] đã đạt F1 - score tốt nhất là 64,40%, cải thiện đáng kể 4,66% so với mô hình CNN vào tháng 1 khi áp dụng vào bộ dữ liệu UIT-VSMEC là 59,74% điểm F1 - score [4].

Cũng trong năm, vào tháng 12 Huy Duc Huynh và các đồng tác giả đã dùng bộ dữ liệu UIT-VSMEC để thực hiện nghiên cứu với Các mô hình deep learning nâng cao bao gồm CNN [8], LSTM và các biến thể của chúng [5]. Bên cạnh đó các tác giả còn triển khai BERT một mô hình chưa bao giờ được áp dụng cho tập dữ liệu UIT-VSMEC. Các thử nghiệm muốn tìm ra một mô hình phù hợp cho từng tập dữ liệu cụ thể. Có khi sẽ sử dụng một mô hình đơn lẻ hoặc có khi sử dụng kết hợp các mô hình lại với nhau. Và cuối cùng kết quả cho thấy mô hình tổng hợp đã đạt được hiệu suất tốt nhất trên tập dữ liệu UIT-VSMEC với 65,79% điểm F1- score [5].

C. Công trình liên quan đến bộ dữ liệu UIT-ViCTSD

Năm 2021, Nguyễn Thành Luân và các đồng tác giả đã tạo bộ dữ liệu Vietnamese Constructive and Toxic Speech Detection (UIT-ViCTSD)[9] đã được xét duyệt của IEA/AIE 2021 và sẽ công bố trong thời gian sắp tới. Bộ dữ liệu gồm 10,000 bình luận, chia làm 10 chủ đề khác nhau thu thập trên VnExpress.net. Kết quả mà bộ dữ liệu này đạt được với PhoBERT của chính nhóm nghiên cứu thực hiện là 78,59% F1-score cho việc xác định các bình luận là Constructiveness (Có tính xây dựng) và 59,40% F1-score cho việc xác định Toxicity (Tiêu cực).

III. Bộ dữ liệu

A. UIT-VSFC

Bộ dữ liệu đầu tiên được sử dụng là bộ UIT-VSFC[1]: Vietnamese Students' Feedback Corpus for Sentiment Analysis. Bộ dữ liệu bao gồm 16,175 câu phản hồi từ sinh viên được thu thập ở cuối các học kỳ từ năm 2013 đến năm 2016[16]. Bộ dữ liệu đề cập đến 2 nhiệm vụ trong một phản hồi, bao gồm: cảm xúc và chủ đề của phản hồi. Bộ dữ liệu này được chia thành 3 phần là Train, Dev và Test với tỉ lệ phần trăm tương ứng là 70%, 10% và 20%. Bảng I sẽ cho chúng ta thấy sự phân bổ các câu trong 3 tập Train, Dev, Test với khía cạnh cảm xúc và bảng II là sự thống kê với khía cạnh chủ đề

Bảng I: Phân bổ các câu đối với cảm xúc

	Train	Dev	Test	Tổng
Positive	5,643	805	1,590	8,038
Neutral	5,325	705	167	6,197
Negative	458	73	1,409	1,940
Tổng	11,426	1,583	3,166	16,175

Bảng II: Phân bổ các câu đối với chủ đề

	Train	Dev	Test	Tổng
Lecturer	8,166	1,151	2,290	11,607
Curriculum	2,201	267	572	3,040
Facility	497	95	145	737
Others	562	70	159	791
Tổng	11,426	1,583	3,166	16,175

B. UIT-VSMEC

Vietnamese Social Media Emotion Corpus (UIT-VSMEC) - một bộ dữ liệu đầu tiên ghi nhận các nhãn cảm xúc (enjoyment, sadness, anger, surprise, fear, disgust và other) trên 6.927 câu bình luận của phương tiện truyền thông Facebook [4]. Bộ dữ liệu UIT-VSMEC còn được chia thành tỷ lệ 80:10:10, trong đó 80% kho ngữ liệu là bộ Train, 10% là bộ Dev và phần còn lại là bộ Test. Do đó, UIT-VSMEC là một kho ngữ liệu nhãn không cân bằng, để đảm bảo rằng các câu trong nhãn có khối lượng thấp được phân phối đầy đủ trong mỗi bộ, thì các nhà nghiên cứu đã sử dụng phương pháp lấy mẫu phân tầng [15] bằng cách sử dụng hàm train_test_split() được hỗ trợ bằng thư viện learn scikit để phân phối chúng thành các bộ Train, Dev và Test. Kết quả được trình bày trong Bảng III

Bảng III: Phân bổ các câu trong 3 tập

	Train	Dev	Test	Tổng
Enjoyment	1,573	205	193	1,965
Disgust	1,064	141	132	1,338
Sadness	938	92	116	1,149
Anger	395	38	40	480
Fear	317	38	46	395
Surprise	242	36	37	309
Other	1,019	132	129	1,291
Tổng	5,548	686	693	6,927

C. UIT-ViCTSD

Bô thứ ba được sử dụng là bộ UIT-ViCTSD[10]: Vietnamese Constructive and Toxic Speech Detection, gồm 10.000 bình luận và 10 chủ đề khác nhau bao gồm: "Entertainment" (Giải trí), "Education" (Giáo dục), "Science" (Khoa học),

"Business"(Kinh doanh), "Cars"(Xe cộ), "Law" (Pháp luật), "Health"(Sức khỏe), "World"(Thế giới), "Sports"(Thể thao), and "News"(Tin tức)[14]. Chia ra làm ba tập train, valid, test theo tỷ lệ 7: 2: 1. Bộ dữ liệu liên quan đến hai khía cạnh của một bình luận: tính xây dựng và tính tiêu cực của bình luận. Định nghĩa các nhãn dán trong bộ dữ liệu, bao gồm tính xây dựng: Constructive (1), Non-Constructive (0). Sự phân bổ các câu trong 3 tập trong khía cạnh Tính xây dựng và Tính tiêu cực được thể hiện ở Bảng IV và Bảng V

Bảng IV: Phân bổ các câu ở khía cạnh Tính xây dựng

	Train	Valid	Test	Tổng
Constructive	2,503	729	364	3,596
Non-constructive	4,497	1,271	636	6,404
Tổng	7,000	2,000	1,000	10,000

Bảng V: Phân bổ các câu ở khía cạnh Tính tiêu cực

	Train	Valid	Test	Tổng
Toxic	759	232	110	1,101
Non-toxic	6,241	1,768	890	8,899
Tổng	7,000	2,000	1,000	10,000

D. Tiền xử lý

Đối với mỗi bộ dữ liệu, nhóm thực hiện các tiền xử lý, bao gồm: thực hiện chuyển đổi bảng mã thành unicode, tiến hành chuẩn hóa từ, chuẩn hóa câu, xử lý các từ viết tắt và sử dụng VnCoreNLP để xử lý tokenization.

IV. Phương pháp thử nghiệm

A. Support Vector Machine (SVM)

SVM là một thuật toán thuộc nhóm Supervised Learning (Học có giám sát) được sử dụng khá phổ biến trong các bài toán phân loại [6]. Mục tiêu của SVM là tìm ra một siêu phẳng (là một hàm tương tự như phương trình đường thẳng) trong không gian có N chiều. Siêu phẳng này sẽ chia không gian thành các miền khác nhau và mỗi miền sẽ chưa một loại dữ liệu. Tuy nhiên, trong một không gian sẽ tồn tại rất nhiều siêu phẳng khác nhau. Vì vậy, chúng ta cần phải tìm ra được siêu phẳng tối ưu nhất cho bài toán của mình. SVM cố gắng tối ưu thuật toán bằng cách tìm cách maximize giá trị của khoảng cách giữa miền siêu phẳng đến điểm k_x gần nhất tương ứng với mỗi miền dữ liệu. SVM sử dụng thuật ngữ Margin để biểu diễn giá trị này.

Margin = $\frac{2}{||W||}$ [6]

Trong đó: ||W|| là độ dài của vector W: ||W|| = $\sqrt{w_1^2 + w_2^2 + ... + w_n^2}$

B. Term Frequency – Inverse Document Frequency TF-IDF

TF-IDF là 1 kỹ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu.
Term frequency thường được chia cho độ dài văn bản (tổng số từ trong một văn bản).

$$ft(t,d) = \frac{f(t,d)}{\max\{f(w,d) : w \epsilon d\}}[17]$$

Trong đó:

- tf(t, d): tần suất xuất hiện của từ t trong văn bản d
- f(t, d): Số lần xuất hiện của từ t trong văn bản d
- $\max[f(w,d) : w \epsilon d]$: Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d

Nó giúp đánh giá tầm quan trọng của một từ. Như thế chúng ta cần giảm độ quan trọng của những từ.

$$idf(t,D) = \log \frac{|D|}{|\{d \epsilon D : t \epsilon d\}|}[17]$$

Trong đó:

- idf(t, D): giá trị idf của từ t trong tập văn bản
- |D|: Tổng số văn bản trong tập D
- $|\{d \epsilon D : t \epsilon d\}|$: Thể hiện số văn bản trong tập D có chứa từ t.

Cụ thể, chúng ta có công thức tính TF-IDF

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)[17]$$

C. Naive Bayes Classification

Naive Bayes Classification là một thuật toán thuộc nhóm Supervised Learning được sử dụng trong các bài toán phân loại. Nó hoạt động dựa trên việc tính toán xác suất áp dụng định lý Bayes [2]. Xét bài toán phân loại với N classes 1, 2, ..., N. Giả sử có một điểm dữ liệu x $\epsilon \mathbb{R}^d$. Hãy tính xác suất để điểm dữ liệu này rơi vào class n.

$$p(y = c|x)[2]$$

Từ đó có thể giúp xác định class của điểm dữ liệu đó bằng cách chọn ra class có xác suất cao nhất:

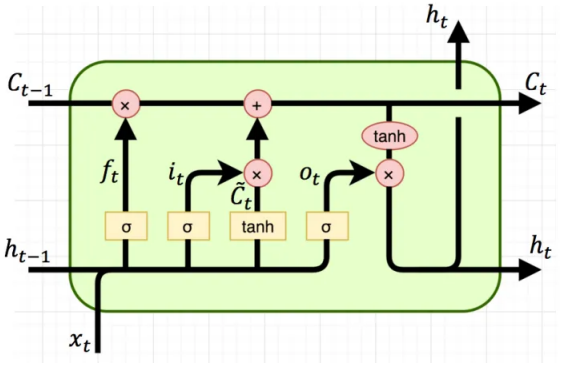
$$n = arg \max_{n \in \epsilon 1,2,..,n)} p(n|x)[2]$$

Tương đương: n = arg $\max_c$p(x|c)p(c)

D. Bidirectional Long Short-Term Memory (BiLSTM)

- Long Short-Term Memory (LSTM)

Là một kiến trúc đặc biệt của RNN (Recurrent Neural Network) đã được sử dụng khá rộng rãi trong bài toán phân loại [1]. Vì LSTM có thể học được những thông tin nào cần được lưu trữ và những thông tin nào không cần lưu trữ vì vậy nó có khả năng học được sự phụ thuộc trong dài hạn (long-term dependencies).

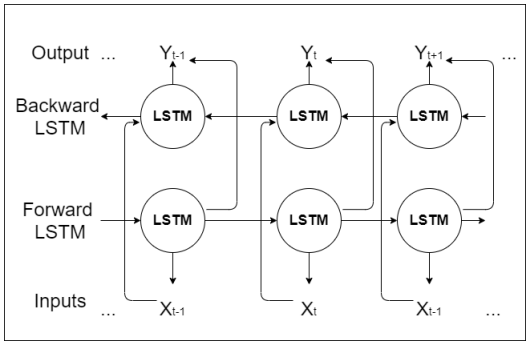


Hình 1: Cấu trúc của một cell state BiLSTM

Hình 1 trên mô tả cấu trúc của một cell state, trong đó:
+ Input: C_{t-1}, h_{t-1}, x_t: x_t là input state thứ t của model, C_{t-1} và h_{t-1} là output của layer trước đó [1]
+ Output: Ct và ht: c gọi là cell state, h gọi là hidden state
+ Forget state: $f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f$ [1]
+ Input gate: $i_t = \sigma(U_i * x_t + W_t * h_{t-1} + b_i)$ [1]
+ New memory cell:C_{tanh} = tanh$(W_C.[h_{t-1}, x_t] + b_C$ [1]
+ Cập nhật cell state mới:$C_t = f_t * C_{t-1} + i_t * C_{tanh}$ [1]
+ Cập nhật hidden state mới:$h_t = o_t * tanh(C_t)$ [1]

- Bidirectional Long Short-Term Memory (BiLSTM)

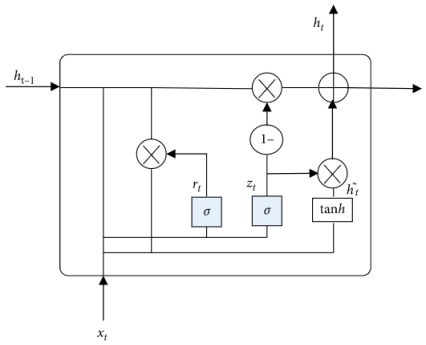
Việc phân loại văn bản phụ thuộc không chỉ vào các thông tin phía trước của từ đang xét mà còn cả các thông tin phía sau. Tuy nhiên, một kiến trúc LSTM truyền thống với một lớp duy nhất chỉ có thể dự đoán nhãn của từ hiện tại dựa trên thông tin có được từ các từ nằm trước đó. BiLSTM đã được tạo ra để khắc phục điểm yếu trên. Một kiến trúc BiLSTM thường chứa 2 mạng LSTM đơn được sử dụng đồng thời và độc lập để mô hình hoá chuỗi đầu vào theo 2 hướng: từ trái sang phải (forward LSTM) và từ phải sang trái (backward LSTM) [18]



Hình 2: Sơ đồ cấu trúc BiLSTM

E. Bidirectional Gated Recurrent Unit (BiGRU)

Gated Recurrent Unit (GRU) [3] được coi là một biến thể của LSTM vì cả hai được thiết kế tương tự nhau. Trong một số trường hợp, kết quả có thể tốt tương tự nhau. Để giải quyết vấn đề mất mát gradient của mạng RNN truyền thống, GRU có hai cổng là update gate và reset gate đó chính là hai vector quyết định thông tin nào sẽ được truyền cho đầu ra. Điều đặc biệt là nó có thể được đào tạo để giữ thông tin từ lâu trước đó, không hề xóa thông tin không liên quan đến dự đoán đầu ra.



Hình 3: Sơ đồ cấu trúc cells state BiGRU

Công thức:

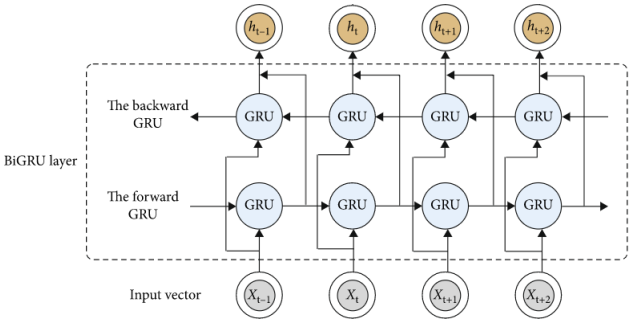
$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{h-1})[13]$$

$$r_t = \sigma(W^{(z)}x_t + U^{(z)}h_{h-1})[13]$$

$$h'_t = tanh(Wx_t + r_t * Uh_{t-1}[13]$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h'_t[13]$$

BiGRU là mô hình kết hợp nhiều GRU với nhau theo hai chiều, có khả năng tìm hiểu mối liên quan giữa các yếu tố ảnh hưởng trong quá khứ, tương lai và cả hiện tại. Điều này có lợi hơn cho việc trích xuất các đặc điểm của dữ liệu được trích xuất. Cấu trúc của BiGRU được thể hiện trong Hình 5.



Hình 4: Sơ đồ cấu trúc BiGRU

F. PhoBERT

PhoBERT là một pre-trained được huấn luyện monolingual language (tức là chỉ huấn luyện dành riêng cho tiếng Việt). Việc huấn luyện dựa trên kiến trúc và cách tiếp cận giống RoBERTa [7].

Tương tự như BERT, PhoBERT cũng có 2 phiên bản là $PhoBERT_base$ với 12 transformers block và $PhoBERT_large$ với 24 transformers block.

PhoBERT được train trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipedia corpus và 19GB còn lại lấy từ Vietnamese news corpus. Đây là một lượng dữ liệu khả ổn để train một mô hình như BERT.

PhoBERT sử dụng RDRSegmenter của VnCoreNLP để tách từ cho dữ liệu đầu vào trước khi qua BPE encoder.

V. Thí nghiệm

Nhóm đã thực hiện huấn luyện các mô hình với các thông số: Số từ trong từ điển, số từ trong một câu, số lớp, embedding dim, batch size, epoch và learning rate. Cụ thể:

A. Đối với bộ dữ liệu UIT-VSFC

Bảng VI: Các thông số thí nghiệm trên khía cạnh cảm xúc trong UIT-VSFC

Thông số	SVM	Naive Bayes	BiLSTM	BiGRU	PhoBERT
Số lượng từ trong câu	100	100	128	128	256
Số lớp	-	-	3	3	-
Embedding dim	-	-	300	300	-
Batch size	-	-	50	50	32
Epoch	-	-	10	10	3
learning rate	-	-	0.001	0.001	0.00005
Thiết bị sử dụng	CPU	CPU	GPU	GPU	TPU

Bảng VI liệt kê các thông số mà nhóm thực hiện đối với bộ dữ liệu UIT-VSFC trên khía cạnh cảm xúc. Còn trên khía cạnh chủ đề, nhóm đã thay đổi một vài thông số gồm: số lớp thành 4, Epoch đối với BiLSTM từ 30 chuyển thành 20 và learning rate chỉ còn 0.00002

B. Đối với bộ dữ liệu UIT-VSMEC

Bảng VII liệt kê các thông số mà nhóm thực hiện đối với bộ dữ liệu UIT-VSMEC

Bảng VII: Các thông số thí nghiệm với UIT-VSMEC

Thông số	SVM	Naive Bayes	BiLSTM	BiGRU	PhoBERT
Số lượng từ trong câu	100	100	128	128	
Số lớp	-	-	7	7	-
Embedding dim	-	-	300	300	-
Batch size	-	-	20	64	
Epoch	-	-	10	10	3
learning rate	-	-	0.001	0.001	0.00005
Thiết bị sử dụng	CPU	CPU	GPU	GPU	TPU

C. UIT-ViCTSD

Bảng VIII: Các thông số thí nghiệm trên tính xây dựng trong UIT-ViCTSD

Thông số	SVM	Naive Bayes	BiLSTM	BiGRU	PhoBERT
Số lượng từ trong câu	100	100	128	128	
Số lớp	-	-	2	2	-
Embedding dim	-	-	300	300	-
Batch size	-	-	50	50	
Epoch	-	-	10	10	3
learning rate	-	-	0.001	0.001	0.00005
Thiết bị sử dụng	CPU	CPU	GPU	GPU	TPU

Bảng VIII thể hiện các thống số thí nghiệm đối với bộ UIT-ViCTSD trên tính xây dựng. Đối với tính tiêu cực thì nhóm chỉ thay đổi thông số learning rate trong BiGRU thành 0.0001.

VI. Kết quả

Sau khi chạy thực nghiệm 5 phương pháp với ba bộ dữ liệu thì thu được kết quả với các thông số đánh như Accuracy, F1-Score, Precision, Recall.

A. Bộ dữ liệu UIT-VSFC

- Trên khía cạnh cảm xúc

Bảng IX: Kết quả trên khía cạnh cảm xúc không bao gồm tiền xử lý

Phương pháp	Nhãn	P	R	F1
SVM+TFIDF	Negative	80.52%	96.81%	87.91%
	Neutral	42.55%	11.98%	18.69%
	Positive	94.32%	84.53%	89.15%
	Trung Bình	86.16%	86.16%	86.16%
Naive Bayes+TFIDF	Negative	81.00%	95.88%	87.81%
	Neutral	66.67%	1.20%	2.35%
	Positive	92.24%	86.73%	89.40%
	Trung Bình	86.29%	86.29%	86.29%
Naive Bayes + PhoW2V	Negative	64.52%	46.20%	53.85%
	Neutral	9.17%	61.08%	15.95%
	Positive	62.87%	41.32%	49.87%
	Trung Bình	44.53%	44.53%	44.53%
BiLSTM + PhoW2V	Negative	90.44%	93.33%	91.86%
	Neutral	48.23%	40.72%	44.16%
	Positive	92.68%	91.57%	92.12%
	Trung Bình	90.33%	90.33%	90.33%
BiGRU + PhoW2V	Negative	88.32%	96.59%	92.27%
	Neutral	73.58%	23.35%	35.45%
	Positive	93.38%	92.33%	92.85%
	Trung Bình	90.14%	90.14%	90.14%
PhoBERT	Negative	93.72%	96.31%	94.99%
	Neutral	64.71%	39.52%	49.07%
	Positive	94.12%	95.66%	94.88%
	Trung Bình	92.98%	92.98%	**92.98%**

Bảng X: Kết quả trên khía cạnh cảm xúc bao gồm tiền xử lý

Phương pháp	Nhãn	P	R	F1
SVM+TFIDF	Negative	80.61%	96.81%	87.97%
	Neutral	43.75%	12.57%	19.53%
	Positive	94.32%	84.59%	89.19%
	Trung Bình	86.22%	86.22%	86.22%
Naive Bayes+TFIDF	Negative	81.00%	95.88%	87.81%
	Neutral	66.67%	1.20%	2.35%
	Positive	92.24%	86.73%	89.40%
	Trung Bình	86.29%	86.29%	86.29%
Naive Bayes + PhoW2V	Negative	64.38%	46.06%	53.7%
	Neutral	9.16%	61.08%	15.93%
	Positive	62.84%	41.26%	49.81%
	Trung Bình	44.4%	44.4%	44.4%
BiLSTM + PhoW2V	Negative	87.02%	97.09%	91.78%
	Neutral	65.75%	28.74%	40.00%
	Positive	94.67%	90.57%	92.57%
	Trung Bình	90.30 %	90.30%	90.30%
BiGRU + PhoW2V	Negative	90.85%	93.75%	92.28%
	Neutral	57.41%	37.13%	45.09%
	Positive	92.27%	93.08%	92.67%
	Trung Bình	90.65%	90.65%	90.65%
PhoBERT	Negative	93.56%	95.81%	94.67%
	Neutral	64.81%	41.92%	50.91%
	Positive	93.75%	95.22%	94.48%
	Trung Bình	92.67%	92.67%	**92.67%**

Dựa vào số liệu từ Bảng IX và Bảng X với micro F1-score ta thấy PhoBERT không tiến hành tiền xử lý có thông số đánh giá cao nhất với F1-Score trên khía cạnh cảm xúc là: 92.98%. So với phương pháp BiLSTM được thực hiện trong bài báoDeep Learning versus Traditional Classifiers on Vietnamese Students' Feedback Corpus [11] (công trình có kết quả cao nhất) thì cao hơn 0.98%

- Trên khía cạnh chủ đề

Bảng XI: Kết quả trên khía cạnh chủ đề không bao gồm tiền xử lý

Phương pháp	Nhãn	P	R	F1
SVM+TFIDF	Lec	90.50%	90.70%	90.60%
	Curr	61.68%	78.50%	69.08%
	Fac	95.51%	58.62%	72.65%
	Others	61.11%	20.75%	30.99%
	TB	83.51%	83.51%	83.51%
Naive Bayes+TFIDF	Lec	85.29%	95.98%	90.32%
	Curr	66.93%	58.39%	62.37%
	Fac	97.67%	57.93%	72.73%
	Others	100.00%	2.52%	4.91%
	TB	82.75%	82.75%	82.75%
Naive Bayes + PhoW2V	Lec	82.25%	33.58%	47.69%
	Curr	27.35%	41.78%	33.06%
	Fac	100.00%	17.93%	30.41%
	Others	8.49%	71.07%	15.17%
	TB	36.22%	36.22%	36.22%
BiLSTM + PhoW2V	Lec	93.27%	92.66%	92.97%
	Curr	71.00%	79.20%	74.88%
	Fac	90.54%	92.41%	91.47%
	Others	61.90%	40.88%	49.24%
	TB	87.71%	87.71%	87.71%
BiGRU + PhoW2V	Lec	92.54%	93.23%	92.89%
	Curr	71.81%	75.70%	73.70%
	Fac	96.27%	88.97%	92.47%
	Others	57.38%	44.03%	49.82%
	TB	87.49%	87.49%	87.49%
PhoBERT	Lec	93.82%	94.80%	94.31%
	Curr	76.48%	79.02%	77.73%
	Fac	92.36%	91.72%	92.04%
	Others	59.83%	44.03%	50.72%
	TB	89.13%	89.13%	**89.13%**

Bảng XII: Kết quả trên khía cạnh chủ đề bao gồm tiền xử lý

Phương pháp	Nhãn	P	R	F1
SVM+TFIDF	Lec	90.58%	90.66%	90.62%
	Curr	61.73%	78.67%	69.18%
	Fac	95.51%	58.62%	72.65%
	Others	60.71%	21.38%	31.63%
	TB	83.54%	83.54%	83.54%
Naive Bayes+TFIDF	Lec	85.29%	95.98%	90.32%
	Curr	66.93%	58.39%	62.37%
	Fac	97.67%	57.93%	72.73%
	Others	100.00%	2.52%	4.91%
	TB	82.75%	82.75%	82.75%
Naive Bayes + PhoW2V	Lec	82.53%	33.62%	47.78%
	Curr	27.44%	41.78%	33.13%
	Fac	100.00%	17.93%	30.41%
	Others	8.53%	71.70%	15.25%
	TB	36.29%	36.29%	36.29%
BiLSTM + PhoW2V	Lec	92.43%	93.32%	92.87%
	Curr	75.27%	73.43%	74.34%
	Fac	97.62%	84.83%	90.77%
	Others	47.06%	50.31%	48.63%
	TB	87.46%	87.46%	87.46%
BiGRU + PhoW2V	Lec	93.00%	93.00%	93.00%
	Curr	75.00%	71.00%	73.00%
	Fac	83.00%	94.00%	88.00%
	Others	47.00%	46.00%	46.00%
	TB	86.79%	86.79%	86.79%
PhoBERT	Lec	93.92%	94.50%	94.21%
	Curr	75.66%	80.42%	77.97%
	Fac	89.33%	92.41%	90.85%
	Others	69.23%	45.28%	54.75%
	TB	89.38%	89.38%	**89.38%**

Dựa vào số liệu từ Bảng XI và Bảng XII với micro F1-score ta thấy PhoBERT đã tiến hành tiền xử lý có thông số đánh giá cao nhất với F1-Score trên khía cạnh chủ đề là: 89.38%.
So với phương pháp BiLSTM được thực hiện trong bài báoDeep Learning versus Traditional Classifiers on Vietnamese Students' Feedback Corpus [11] (công trình có kết quả cao nhất) thì thấp hơn 0.22%

B. Bộ dữ liệu UIT-VSMEC

- Khi không áp dụng tiền xử lý

Bảng XIII: Kết quả đối với UIT-VSMEC khi không tiến hành tiền xử lý

Phương pháp	Accuracy	F1-score
Naive Bayes+PhoW2V	12.70%	8.98%
Naive Bayes+TF-IDF	45.60%	40.26%
SVM+TF-IDF	45.02%	43.74%
BiGRU + PhoW2V	51.23%	51.45%
BiLSTM + PhoW2V	45.17%	46.10%
PhoBERT	**63.06%**	**62.71%**

- Thực hiện tiền xử lý

Bảng XIV: Kết quả đối với UIT-VSMEC tiến hành tiền xử lý

Phương pháp	Accuracy	F1-score
Naive Bayes+PhoW2V	12.70%	8.98%
Naive Bayes+TF-IDF	46.46%	41.07%
SVM+TF-IDF	45.02%	43.87%
BiGRU + PhoW2V	53.39%	52.63%
BiLSTM + PhoW2V	48.48%	46.92%
PhoBERT	**60.32%**	**60.31%**

Dựa vào số liệu từ Bảng XIII và Bảng XIV với weighted F1-score ta thấy được PhoBERT không tiền xử lý cho kết quả có thông số đánh giá cao nhất với F1-Score là 62.71%
So với phương pháp tổng hợp của GRU + CNN + BiLSTM + LSTM được thực hiện trong bài báo A Simple and Efficient Ensemble Classifier Combining Multiple Neural Network Models on Social Media Datasets in Vietnamese [5] (công trình có kết quả cao nhất) thì F1-Score với 7 nhãn (có nhãn Other) thì thấp hơn 3.08%

C. Bộ dữ liệu UIT-ViCTSD

- Khi không áp dụng tiền xử lý

Bảng XV: Kết quả trên bộ dữ liệu UIT-ViCTSD không tiền xử lý

Phương pháp	**Tính xây dựng**		**Tính tiêu cực**	
	Accuracy	**F1-score**	**Accuracy**	**F1-score**
Naive Bayess+PhoW2V	78.40%	76.36%	64.20%	46.79%
Naive Bayess+TF-IDF	73.80%	67.92%	89.10%	48.02%
SVM+TF-IDF	76.00%	71.15%	89.20%	50.58%
BiGRU+ PhoW2V	78.30%	76.91%	82.70%	60.59%
BiLSTM+ PhoW2V	78.40%	76.14%	89.30%	63.86%
PhoBERT	**83.50%**	**82.19%**	**90.70%**	**72.83%**

- Thực hiện tiền xử lý

Bảng XVI: Kết quả trên bộ dữ liệu UIT-ViCTSD có tiền xử lý

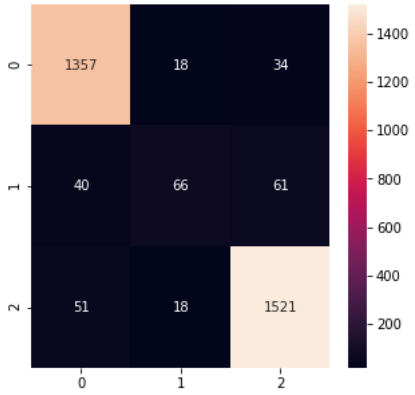
Phương pháp	Tính xây dựng		Tính tiêu cực	
	Accuracy	F1-score	Accuracy	F1-score
Naive Bayess+PhoW2V	78.10%	75.92%	64.80%	46.96%
Naive Bayess+TF-IDF	74.10%	68.58%	89.10%	48.02%
SVM+TF-IDF	76.60%	71.98%	89.20%	51.37%
BiGRU+ PhoW2V	78.70%	77.17%	87.20%	66.23%
BiLSTM+ PhoW2V	76.20%	73.05%	88.80%	65.11%
PhoBERT	**84.00%**	**82.68%**	**90.60%**	**72.92%**

Dựa vào số liệu từ Bảng XV và Bảng XVI với macro F1-score ta thấy PhoBERT sau tiền xử lý có thông số đánh giá cao nhất với F1-Score cho Tính xây dựng và Tính tiêu cực lần lượt là: 82.68% và 72.92%
So với cùng phương pháp PhoBERT được thực hiện trong bài báo Constructive and Toxic Speech Detection for Open-domain Social Media Comments in Vietnamese [10] thì F1-Score cho Tính xây dựng cao hơn 4.09% và tính Tiêu cực cao hơn 13.52%

VII. Phân tích lỗi sai

A. Đối với UIT-VSFC

- Trên khía cạnh cảm xúc

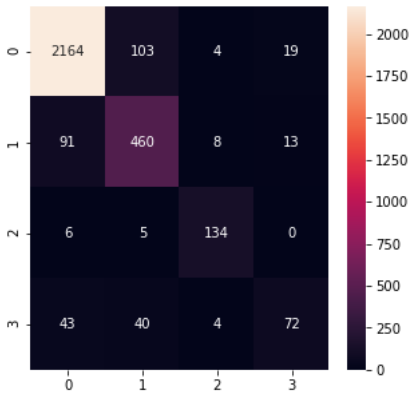


Hình 5: Confusion matrix của PhoBERT không bao gồm tiền xử lí trên khía cạnh cảm xúc

Với Negative tương ứng với 0, Neutral tương ứng với 1 và Positive tương ứng với 2.

Dựa vào Hình 5 ta thấy được, dự đoán sai cao nhất đối với Neutral với 101 nhãn dự đoán sai trên tổng 167 nhãn trong tập test. Dự đoán sai thấp nhất đối với Negative với 52 dự đoán sai trên tổng số 1409 nhãn trong tập test.

- Trên khía cạnh chủ đề

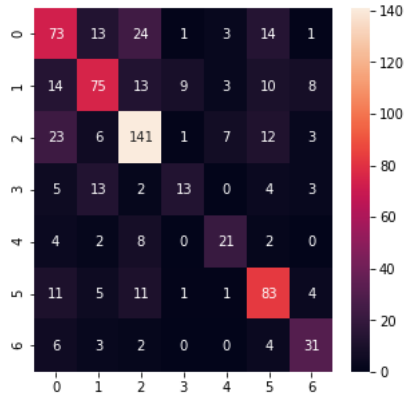


Hình 6: Confusion matrix của PhoBERT bao gồm tiền xử lí trên khía cạnh chủ đề

Với Lecturer tương ứng với 0, Curriculum tương ứng với 1, Facility tương ứng với 2 và Others tương ứng với 3

Dựa vào Hình 6 ta thấy được, dự đoán sai cao nhất đối với Others với 87 nhãn dự đoán sai trên tổng 159 nhãn trong tập test. Dự đoán sai thấp nhất đối với Lecturer với 126 dự đoán sai trên tổng số 2290 nhãn trong tập test.

B. Đối với UIT-VSMEC

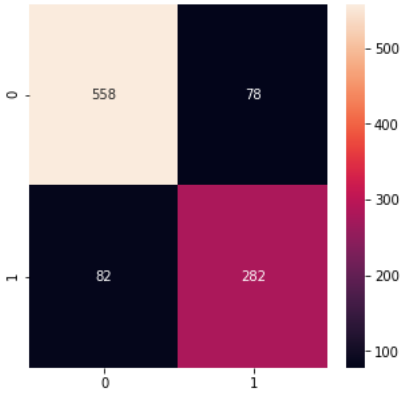


Hình 7: Confusion matrix UIT-VSMEC với PhoBERT không bao gồm tiền xử lí

Với Others tương ứng với 0, Disgust tương ứng với 1, Enjoyment tương ứng với 2, Anger tương ứng với 3, Surprise tương ứng với 4, Sadness tương ứng với 5 và Fear tương ứng với 6
Dựa vào Hình 7 ta thấy được, dự đoán sai cao nhất đối với Anger với 101 nhãn dự đoán sai trên tổng 167 nhãn trong tập test. Dự đoán sai thấp nhất đối với Enjoyment với 52 dự đoán sai trên tổng số 193 nhãn trong tập test.

C. Đối với UIT-ViCTSD

- Trên tính xây dựng

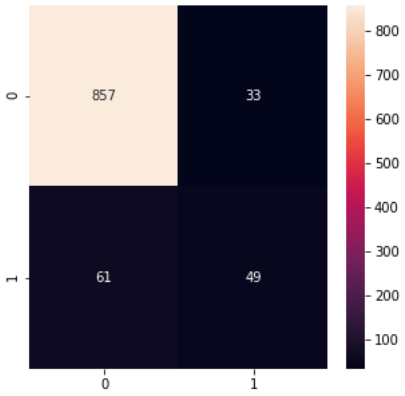


Hình 8: Confusion matrix UIT-ViCTSD với PhoBERT bao gồm tiền xử lí với khía cạnh Tính xây dựng

Với Non-constructive tương ứng với 0 và Constructive tương ứng với 1.

Dựa vào Hình 8 ta thấy được, dự đoán sai cao nhất đối với Constructive với 82 nhãn dự đoán sai trên tổng 364 nhãn trong tập test.

- Trên tính tiêu cực



Hình 9: Confusion matrix UIT-ViCTSD với PhoBERT bao gồm tiền xử lí với khía cạnh Tính tiêu cực

Với Non-toxic tương ứng với 0 và Toxic tương ứng với 1.

Dựa vào Hình 9 ta thấy được, dự đoán sai cao nhất đối với Toxic với 61 nhãn dự đoán sai trên tổng 101 nhãn trong tập test. Dự đoán sai thấp nhất đối với Lecturer với 126 dự đoán sai trên tổng số 2290 nhãn trong tập test.

VIII. Kết luận và hướng phát triển

Sau quá trình thực hiện 5 phương pháp trên 3 bộ dữ liệu thu được kết quả phương pháp PhoBERT là phương pháp phù hợp nhất trên cả ba bộ dữ liệu. Đồng thời, nhóm đã tạo ra một ứng dụng để phân tích một câu khi người dùng nhập vào. Trong tương lai, nhóm dự định sẽ cải thiện mô hình bằng BERT multilingual và thay thế các lớp Embedding như fastText, Multi Embedding nhằm tăng độ chính xác, hiệu suất phân loại.

Tài liệu

[1] Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. Emotion recognition based on eeg using lstm recurrent neural network. *Emotion*, 8(10):355–358, 2017.

[2] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Transferring naive bayes classifiers for text classification. In *AAAI*, volume 7, pages 540–545, 2007.

[3] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.

[4] Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Emotion recognition for vietnamese social media text. In *International Conference of the Pacific Association for Computational Linguistics*, pages 319–333. Springer, 2019.

[5] Huy Duc Huynh, Hang Thi-Thuy Do, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. A simple and efficient ensemble classifier combining multiple neural network models on social media datasets in vietnamese. *arXiv preprint arXiv:2009.13060*, 2020.

[6] Michael E Mavroforakis and Sergios Theodoridis. A geometric approach to support vector machine (svm) classification. *IEEE transactions on neural networks*, 17(3):671–682, 2006.

[7] Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*, 2020.

[8] Khang Phuoc-Quy Nguyen and Kiet Van Nguyen. Exploiting vietnamese social media characteristics for textual emotion recognition in vietnamese. In *2020 International Conference on Asian Language Processing (IALP)*, pages 276–281. IEEE, 2020.

[9] Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. A vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, 2020.

[10] Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Constructive and toxic speech detection for open-domain social media comments in vietnamese. *arXiv preprint arXiv:2103.10069*, 2021.

[11] Phu XV Nguyen, Tham TT Hong, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Deep learning versus traditional classifiers on vietnamese students' feedback corpus. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 75–80. IEEE, 2018.

[12] Vu Duc Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Variants of long short-term memory for sentiment analysis on vietnamese students' feedback corpus. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 306–311. IEEE, 2018.

[13] Hong Qiu, Chongdi Fan, Jie Yao, and Xiaohan Ye. Chinese microblog sentiment detection based on cnn-bigru and multihead attention mechanism. *Scientific Programming*, 2020, 2020.

[14] Nguyễn Duy Sim. Phân lớp bằng random forests trong python. https://viblo.asia/p/phan-lop-bang-random-forests-trong-python-djeZ1D2QKWz, 2018.

[15] The Economic Times. Definition of 'Stratified Sampling'. https://economictimes.indiatimes.com/definition/stratified-sampling, 2021.

[16] Kiet Van Nguyen, Vu Duc Nguyen, Phu XV Nguyen, Tham TH Truong, and Ngan Luu-Thuy Nguyen. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24. IEEE, 2018.

[17] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015.

[18] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212, 2016.