Nghiên cứu đánh giá các kỹ thuật tăng cường dữ liệu cho bài toán phân tích cảm xúc Tiếng Việt

Nguyen Thi Thuy[1,2] and Dang Van Thin[1,2]

[1] Trường Đại học Công Nghệ Thông tin - ĐHQG TPHCM
[2] Đại học Quốc Gia Thành phố Hồ Chí Minh
21521514@gm.uit.edu.vn, thindv@uit.edu.vn

Tóm tắt nội dung Bài báo này tập trung vào việc nghiên cứu đánh giá các kỹ thuật tăng cường dữ liệu cho bài toán phân tích cảm xúc trong tiếng Việt, một trong các lĩnh vực quan trọng của Xử lý Ngôn ngữ Tự nhiên (NLP). Phân tích cảm xúc là việc hiểu và phân tích ý kiến và cảm xúc của con người trong văn bản, đóng vai trò quan trọng trong việc đánh giá sự hài lòng của khách hàng, theo dõi tình hình thị trường, và nhiều ứng dụng khác. Tuy nhiên, trong tiếng Việt, việc hiện có một lượng lớn dữ liệu chất lượng để huấn luyện các mô hình phân tích cảm xúc vẫn còn là một thách thức lớn. Gán nhãn cảm xúc cho văn bản tiếng Việt đòi hỏi sự hiểu biết sâu về ngôn ngữ và ngữ cảnh cụ thể, và quá trình thu thập dữ liệu thường tốn nhiều thời gian và công sức. Vì vậy, nghiên cứu về các phương pháp để tăng cường dữ liệu cho phân tích cảm xúc tiếng Việt là điều cần thiết. Chúng tôi thực hiện đề tài "Nghiên cứu đánh giá các kỹ thuật tăng cường dữ liệu cho bài toán phân tích cảm xúc tiếng Việt" nhằm giải quyết sự hạn chế của dữ liệu bằng cách tạo dữ liệu mới và đa dạng thông qua các phương pháp tăng cường. Kết quả thí nghiệm cho thấy phương pháp tăng cường dữ liệu nâng cao hiệu quả của mô hình phân loại ở các bộ dữ liệu thử nghiệm.

Keywords: tăng cường dữ liệu · phân tích cảm xúc · xử lý ngôn ngữ tự nhiên.

1 Giới thiệu

Phân tích cảm xúc là một trong những chủ đề nghiên cứu ở lĩnh vực Xử lý Ngôn ngữ Tự nhiên (NLP) dành riêng cho việc hiểu và phân tích cảm xúc và ý kiến trong văn bản. Hiện nay, với sự bùng nổ của dữ liệu văn bản trực tuyến và các nền tảng truyền thông xã hội, bài toán phân tích cảm xúc đã trở thành một phần quan trọng của nghiên cứu NLP và có nhiều ứng dụng thực tế. Nó có thể được sử dụng để theo dõi ý kiến của người dùng về sản phẩm hoặc dịch vụ, đánh giá tình hình thị trường và nắm bắt tâm trạng của đám đông đối với một sự kiện cụ thể.

Hiện nay hầu hết các nghiên cứu đối với bài toán phân tích cảm xúc đều tiếp cận theo hướng học có giám sát dựa trên các mô hình học máy (Nassr et al. 2020) [1], học sâu (Jyothis Joseph et al. 2022) [2] hay các mô hình ngôn ngữ

(Sayyida el al. 2022) [3]. Tuy nhiên, điều này dẫn đến một thách thức lớn đối mặt là thiếu các tập dữ liệu lớn và đa dạng để huấn luyện các mô hình hoạt động tốt, điều này đặc biệt đúng đối với các ngôn ngữ ít tài nguyên như tiếng Việt [6]. Để giải quyết vấn đề này, tăng cường dữ liệu là một kỹ thuật nhằm tạo dữ liệu mẫu tổng hợp từ một phần các mẫu dữ liệu sẵn có. Các phương pháp tăng cường dữ liệu có tiềm năng giải quyết vấn đề khan hiếm dữ liệu và cải thiện hiệu suất của các mô hình phân tích cảm xúc [4]. Ngoài thách thức thiếu dữ liệu, những năm gần đây việc phân tích cảm xúc trở nên khó khăn hơn khi dữ liệu bị nhiễu khi được thu thập từ mạng xã hội và các ứng dụng trực tuyến (P.Awatramani et al. 2021) [5].

Trong bài báo này, chúng tôi tập trung vào việc nghiên cứu và đánh giá các kỹ thuật tăng cường dữ liệu cho bài toán phân tích cảm xúc trên dữ liệu tiếng Việt và dữ liệu hỗn hợp giữa tiếng Việt và tiếng Anh. Chúng tôi thử nghiệm và phân tích các phương pháp tăng cường dữ liệu từ hai mức từng từ (token-level) và toàn câu (sentence-level). Các kết quả nghiên cứu được thử nghiệm trên các bộ dữ liệu chuẩn cho bài toán phân tích cảm xúc trên tiếng Việt. Kết quả thí nghiệm cho thấy sự hiệu quả của các phương pháp tăng cường dữ liệu trong việc cải thiện hiệu suất của các mô hình phân loại cảm xúc khác nhau.

2 Nghiên cứu liên quan

Tăng cường dữ liệu là kỹ thuật tạo thêm dữ liệu huấn luyện từ dữ liệu gốc ban đầu mà không cần tạo thêm dữ liệu mới một cách tường tận [6]. Mục tiêu của tăng cường dữ liệu là làm cho tập dữ liệu huấn luyện trở nên đa dạng hơn bằng cách thêm vào các biến thể của dữ liệu ban đầu mà vẫn giữ nguyên ý nghĩa hoặc thông tin quan trọng. Mặc dù được áp dụng rộng rãi và thành công trong lĩnh vực thị giác máy tính, các kỹ thuật DA được thiết kế cho các tác vụ NLP được cho là tiến bộ chậm hơn nhiều và thành công hạn chế trong việc đạt được hiệu suất. Kết quả là, ngoại trừ các ứng dụng dịch ngược cho các tác vụ dịch máy, các kỹ thuật DA vẫn chưa được khai thác kỹ trong cộng đồng NLP (Pellicer et al. 2022) [7].

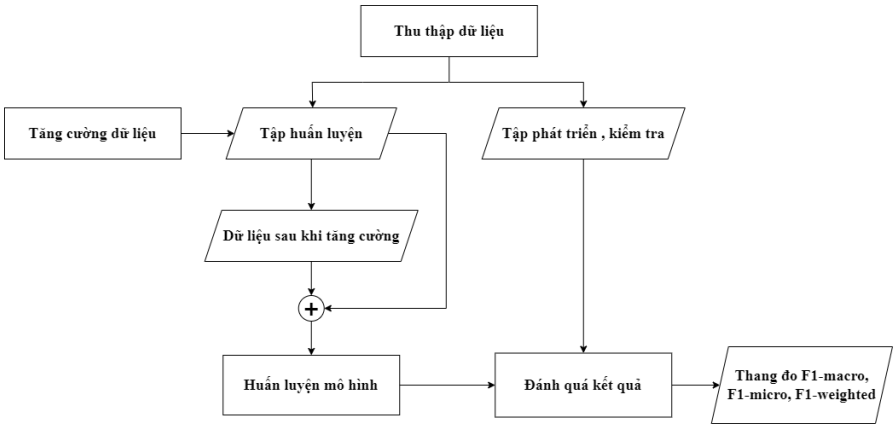
Thực tế DA cho dữ liệu văn bản là một nhiệm vụ khó khăn bởi vì nó có thể gây ra sự không chính xác về ngữ nghĩa trong quá trình tăng cường. Đó đó các phuong pháp DA cho văn bản thường liên quan đến việc thay thế một từ bằng từ đồng nghĩa, xóa một từ hoặc thay đổi một từ (Wei et al.2019) [8]. Ngoài ra, với những tiến bộ gần đây của các mô hình ngôn ngữ các phương pháp DA dùng mô hình ngôn ngữ cũng được sử dụng trong bài toán phân loại văn bản (Anaby-Tavor et al. 2019) [9] và một số nhóm phương pháp DA khác như Token-Level, Sentence-Level, Adversarial Data, Hidden-Space cũng đã được đề cập trong bài khảo sát thực nghiệm bởi (Chen et al. 2021) [4].

Với sự phát triển của mạng xã hội và các ứng dựng trực tuyến, vấn đề đa ngôn ngữ cũng đã gây ra nhiều thách thức cho bài toán phân tích cảm xúc. Mọi người thường có có xu hướng kết hợp giữa ngôn ngữ bản địa và một ngôn ngữ khác mà cộng động của họ thân thuộc (P.Awatramani et al. 2021) [5], với người Việt ngôn ngữ thứ hai mà họ thường sử dụng đó là tiếng Anh. Vấn đề đa ngôn

ngữ này cũng có thể tạo ra nhiều sự khác biệt về hiệu quả của các phương pháp tăng cường khi áp dụng các bộ phân khác nhau. Trong nghiên cứu của (Van et al. 2022) [10] đã công bố hai bộ dữ liệu Nhận xét hỗn hợp giữa tiếng Việt và tiếng Anh trong lĩnh vực nhận xét nhà hàng, đồng thời nhóm tác giả cũng đã thực hiện việc đánh giá hiệu quả của việc áp dụng các phương pháp học máy, học sâu, mô hình ngôn ngữ trong bài bào này.

3 Phương pháp

3.1 Quy trình thực hiện



Hình 1. Quy trình thực hiện tăng cường dữ liệu và thực nghiệm

Các bước thực hiện tăng cường dữ liệu và thực nghiệm (Hình 1):

- Thu thập dữ liệu: các bộ dữ liệu được chia thành 3 tập khác nhau bao gồm tập huấn luyện, pháp triển và thử nghiệm để phục vụ cho quá trình thực nghiệm.
- Tăng cường dữ liệu: ở bước này các kỹ thuật tăng cường dữ liệu được áp dụng trên tập huấn luyện để tạo ra dữ liệu tăng cường.
- Huấn luyện mô hình: sử dụng dữ liệu huấn luyện gốc kết hợp với dữ liệu sau khi tăng cường theo từng kỹ thuật tăng cường để huấn luyện các bộ phân loại cảm xúc.
- Đánh giá kết quả: để đánh giá hiệu quả của các phương pháp tăng cường dữ liệu, chúng tôi tiến hành thực nghiệm bằng cách đánh giá các bộ phân loại đã được huấn luyện bằng dữ liệu tăng cường trên tập kiểm tra, sử dụng các thang đo F1-macro, F1-micro và F1-weighted.

3.2 Các phương pháp tăng cường dữ liệu

Tăng cường dữ liệu ở cấp từ Tăng cường cấp từng từ bao gồm việc biến đổi một hoặc một số từ trong một câu để tạo ra văn bản tăng cường. Mục tiêu là tạo ra các biến thể của câu trong khi vẫn giữ nguyên ý nghĩa ngữ nghĩa và nhãn cảm xúc ban đầu của văn bản. Chúng tôi đề xuất các kỹ thuật tăng cường cấp từ sau:

- Xóa ngẫu nhiên (Random delete): Kỹ thuật này loại bỏ một hoặc một số từ ngẫu nhiên khỏi văn bản đầu vào.
- Đổi chỗ ngẫu nhiên (Random swap): Trong kỹ thuật này, một hoặc một số cặp từ được chọn ngẫu nhiên trong câu để đổi chỗ chỗ
- Thay thế ngẫu nhiên (Random replace): Trong phương pháp này, các từ được chọn ngẫu nhiên và thay thế. Chiến lược bao gồm việc thay thế bằng từ đồng nghĩa (Synonym), sử dụng các vectơ nhúng Word2Vec, thay thế dựa trên TF-IDF và thay thế dựa trên mô hình ngôn ngữ – BERT.
- Chèn ngẫu nhiên (Random insert): Kỹ thuật tăng cường này liên quan đến việc chèn các từ vào các vị trí ngẫu nhiên trong câu. Tương tự như thay thế, việc chèn có thể đạt được bằng cách sử dụng từ đồng nghĩa, nhúng Word2Vec, điểm TF-IDF hoặc BERT.

Các kỹ thuật DA nói trên là những kỹ thuật tăng cường phổ biến được sử dụng trong lĩnh vực NLP hiện nay. Tuy nhiên, tiếng Việt là một ngôn ngữ ít tài nguyên và phức tạp, không có khoảng trắng giữa các từ mà chỉ có khoảng trắng giữa các âm tiết, một từ có thể có một hay nhiều âm tiết. Do tính phức tạp và hạn chế về tài nguyên và công cụ trong tiếng Việt, khi thực hiện DA trên tiếng Việt chúng ta cần phải áp dụng các kỹ thuật một cách sáng tạo và thực hiện một số thao tác bổ sung để có thể đạt được hiệu quả cao nhất.

- Thực hiện tách từ cho tiếng Việt: trước khi áp dụng kỹ thuật DA chúng tôi đã giới thiệu (ngoại trừ kỹ thuật chèn và thay thế bằng chiến lược BERT) và huấn luyện các mô hình nhúng Word2Vec, TF-TF chúng tôi đều thực hiện tách từ. Chúng tôi sử dụng thư viện VnCoreNLP được công bố bởi (Vu et al. 2018) [11].
- Tăng cường từ đồng nghĩa: quá trình tìm kiếm từ đồng nghĩa trong tiếng Việt vẫn còn là một thách thức do thiếu cơ sở dữ liệu từ đồng nghĩa ổn định và toàn diện. Để giải quyết vấn đề này, chúng tôi áp dụng cách tiếp cận nhiều bước. Đối với từ mục tiêu là từ tiếng Việt, đầu tiên chúng tôi dịch từ mục tiêu từ tiếng Việt sang tiếng Anh. Tiếp theo, chúng tôi sử dụng WordNet (George A et al. 1995) [12] , một nguồn tài nguyên đáng tin cậy về từ đồng nghĩa tiếng Anh, để tìm từ đồng nghĩa phù hợp. Sau đó, chúng tôi dịch các từ đồng nghĩa tiếng Anh này sang tiếng Việt, tạo ra một quy trình thay thế từ đồng nghĩa tiếng Việt hợp lý. Đối với từ mục tiêu là tiếng Anh thì chúng tôi sử dụng WordNet để tìm từ đồng nghĩa mà không cần phải trải qua quá trình dịch. Chiến lược chuyển ngôn ngữ này giúp giải quyết khó khăn trong viễ tìm từ đồng nghĩa tiếng Việt.

- Tăng cường dựa trên BERT: thử nghiệm của chúng tôi cho thấy các mô hình ngôn ngữ lớn hiện tại cho tiếng Việt có thể biểu hiện sự không ổn định khi áp dụng cho phương pháp tăng cường này. Để giải quyết vấn đề này, chúng tôi thực hiện phương pháp chuyển ngôn ngữ sang tiếng Anh để thực hiên tăng cường dựa trên BERT, quá trình thực hiên gồm ba bước. Chúng tôi dịch câu tiếng Việt hoặc câu hỗn hợp Việt-Anh (một câu bao gồm các từ tiếng Việt và tiếng Anh) sang tiếng Anh, áp dụng các kỹ thuật tăng cường chèn và thay thế bằng cách mô hình ngôn ngữ BERT, sau đó dịch câu tiếng Anh tăng cường trở lại tiếng Việt. Trình tự dịch và tăng cường này thu hẹp khoảng cách giữa những hạn chế của mô hình ngôn ngữ và việc nâng cao dữ liệu phân tích cảm xúc.

Tóm lại, đặc thù của ngôn ngữ tiếng Việt và hiện trạng của các mô hình ngôn ngữ đòi hỏi những cách tiếp cận sáng tạo. Bằng cách sử dụng các phương pháp này, chúng tôi đã khai thác lợi ích của việc WordNet để tìm từ đồng nghĩa và mô hình ngôn ngữ BERT bằng cách dịch sang tiếng Anh. Điều này giúp giảm thiểu những hạn chế vốn và khó khăn có trong nhiệm vụ DA cho bài phân tích cảm xúc cho văn bản tiếng Việt.

Tăng cường dữ liệu ở cấp câu - Dịch ngược Dịch ngược (back-translation) là một phương pháp mạnh mẽ để tăng cường cấp câu trong phân tích cảm xúc. Bằng cách dịch các câu tiếng Việt sang tiếng Anh và sau đó dịch ngược lại, ta tạo ra các câu mới có nghĩa tương tự nhưng có thể mang cảm xúc khác nhau, làm phong phú thêm tập dữ liệu với cấu trúc câu và từ vựng đa dạng.

3.3 Các bộ phân loại

Chúng tôi đánh giá các kỹ thuật tăng cường dữ liệu đã đề xuất của mình bằng cách sử dụng một loạt các mô hình thuộc các nhóm phương pháp khác nhau như tinh chỉnh các mô hình ngôn ngữ lớn, học máy, học sâu:

- **Tinh chỉnh các mô hình ngôn ngữ**: Chúng tôi đã thực hiện việc tinh chỉnh hai mô hình ngôn ngữ lớn, trong đó có một mô hình dành riêng cho tiếng Việt - PhoBERT (Dat et al. 2020) [14] và một mô hình đa ngôn ngữ XLM- RoBERTa (Alexis Conneau at al. 2020) [15]. PhoBERT được đào tạo đặc biệt cho tiếng Việt, điều này làm cho PhoBERT trở thành một công cụ mạnh mẽ trong các tác vụ liên quan đến tiếng Việt. Trong khi đó, XLM-RoBERTa là một biến thể khác của mô hình RoBERTa có khả năng biểu diễn văn bản tự nhiên mạnh mẽ. Chúng tôi đã thực hiên tinh chỉnh hai mô hình này với tốc độ học leanring_rate = 2e-5 giảm dần tuyến tính (lr_scheduler_type = linear), số vòng lặp epochs = 8, và batch_size = 16.
- **Mô hình máy học truyền thống**: Thay vì sử dụng mô hình TF – IDF, một phương pháp thường được sử dụng trong các tác cụ biểu diễn văn bản tiếng Việt, chúng tôi đã sử dụng laBSE – một mô hình biểu diễn văn bản đa ngôn ngữ được cho là cho kết quả tốt nhất trên bộ dữ liệu hỗn hợp Việt – Anh (Van et al. 2022) [10] để biểu diễn các văn bản thành các vector nhúng.

Bảng 1. Phân bố các nhãn trên bộ dữ liệu thí nghiệm bao gồm bộ dữ liệu đơn ngữ tiếng Việt và bộ dữ liệu hỗn hợp Việt - Anh.

Thông tin dữ liệu		Tiêu cực	Trung tính	Tích cực
Bộ dữ liệu đơn ngữ	Tập huấn luyện	242	200	258
	Tập kiểm tra	73	71	56
	Tập phát triển	35	29	36
Bộ dữ liệu hỗn hợp	Tập huấn luyện	249	207	253
	Tập kiểm tra	72	62	66
	Tập phát triển	38	31	31

Sau đó sử dụng các các vector này để huấn luyện các thuật toán Support Vector Machines - SVM, Nayve Bayes - NB, Multilayer Perceptron - MLP.

- Support Vector Machine: là thuật toán học máy truyền thống đạt hiệu quả cao trong các bài toán phân loại. Chúng tôi huấn luyện SVM với các tham số C (C – Support Vector Classification) = 1.0, hàm kernel = "rbf", ngưỡng dừng tối ưu hóa tol = 1e-5.
- Naive Bayes: đây cũng là một phương pháp phân loại được sử dụng rộng rãi cho dữ liệu văn bản. Trong nghiên cứu này, chúng tôi đã sử dụng mô hình Naive Bayes với tham số smoothing laplace $\alpha = 1.0$.
- Multi-layer Perceptron: MLP là một kiểu mạng nơ-ron nhiều lớp, trong đó thông tin được truyền qua nhiều lớp ẩn trung gian giúp học các biểu diễn phức tạp từ dữ liệu đầu vào. Chúng tôi huấn luyện mô hình có 2 lớp ẩn với số nốt của hai lớp lần lượt là 1024 và 512, tốc đô học learning_rate = 0.0001.

- **Mô hình học sâu**: Đối với phương pháp học sâu, chúng tôi đã xây dựng một mô hình mạng nơ-ron tích chập Convolutional Neural Network - CNN được đề xuất bởi tác giả [13] với số kernel_size = 3, số bộ lọc filters = 64, hàm kích hoạt activation = "relu", số vòng lặp epoch = 15, và batch_size = 32.

4 Kết quả thực nghiệm

4.1 Mô tả dữ liệu

Trong bài báo nay, chúng tôi thực hiện phân tích cảm xúc trong lĩnh vực khách sạn trên hai bộ dữ liệu. Trong đó có một bộ dữ liệu Nhận xét tiếng Việt (bao gồm câu nhận xét bằng tiếng Việt) và bộ dữ liệu Nhận xét hỗn hợp Việt – Anh (bao gồm các câu nhận xét chứa cả tiếng Anh và tiếng Việt) [10] để thực nghiệm các kỹ thuật tăng cường dữ liệu. Mỗi bộ dữ liệu sẽ có 3 nhãn: tích cực, trung tính, tiêu cực. Phân bố của các nhãn trong mỗi bộ dữ liêu được trình bày ở Bảng 1. Với mỗi tập dữ liệu, chúng tôi trích ra một 1000 mẫu tiến hành thực nghiệm.

4.2 Đánh giá kết quả

Nghiên cứu đã tiến hành các thử nghiệm với nhiều mô hình và kỹ thuật tăng cường dữ liệu khác nhau và đánh giá trên các thang đo F1 - Macro, F1 - Micro,

Bảng 2. Kết quả thực nghiệm các kỹ thuật tăng cường dữ liệu trên các Bộ dữ liệu đơn ngữ và Nhận xét hỗn hợp Việt - Anh bằng phương pháp tinh chỉnh các mô hình ngôn ngữ lớn trên các thang đo F1.

	PhoBERT			XLM-RoBERTa		
	Macro	Micro	Weighted	Macro	Micro	Weighted
Bộ dữ liệu đơn ngữ						
Origin	80.18	80.00	79.78	77.69	77.50	77.30
Swap	79.80	79.50	79.44	80.99	81.00	80.95
Delete	79.27	79.00	78.87	78.62	78.50	78.40
Insert Synonym	80.24	80.00	79.79	80.00	80.00	79.79
Insert Word2Vec	78.21	78.00	77.77	77.44	77.50	77.32
Insert TF-IDF	79.66	79.50	79.26	76.69	76.50	76.23
Insert BERT	79.68	79.50	79.17	78.16	78.00	77.79
Replace Synonym	79.76	79.50	79.25	79.54	79.50	79.30
Replace Word2Vec	80.12	80.00	79.71	78.65	78.50	78.20
Replace TF-IDF	81.67	81.50	81.47	77.61	77.50	77.36
Replace BERT	80.80	80.50	80.31	76.40	76.50	75.79
Back-translation	**82.09**	**82.00**	**81.68**	**81.92**	**81.50**	**81.23**
Bộ dữ liệu hỗn hợp						
Origin	74.35	74.50	74.67	74.91	75.50	75.27
Swap	71.58	71.50	71.93	74.37	74.50	74.73
Delete	75.53	75.50	75.86	75.33	75.50	75.58
Insert Synonym	74.95	75.00	75.19	74.32	74.50	74.66
Insert Word2Vec	74.87	75.00	75.17	74.09	74.50	74.43
Insert TF-IDF	75.62	76.00	75.98	75.31	76.00	75.81
Insert BERT	73.58	74.00	73.88	74.45	75.00	74.90
Replace Synonym	71.43	71.50	71.79	**77.38**	**77.50**	**77.72**
Replace Word2Vec	76.76	77.00	77.19	75.61	76.00	75.99
Replace TF-IDF	74.23	74.00	74.48	75.21	75.50	75.58
Replace BERT	**76.93**	**77.00**	**77.23**	75.17	75.50	75.57
Back-translation	74.86	75.00	75.29	75.16	75.50	75.53

F1 - Weighted. Kết quả thí nghiệm trên các mô hình được trình bày ở các Bảng số liệu (xem Bảng 2, 3, 4). Sau đây nghiên cứu sẽ nhận xét chi tiết các kỹ thuật tăng cường dữ liệu.

Kỹ thuật đổi chỗ (swap), xóa (delete) ngẫu nhiên: trên hầu hết các mô hình phân loại và hai bộ dữ liệu thực nghiệm, hai kỹ thuật tăng cường đổi chỗ và xóa ngẫu nhiên cái thiện hiệu suất không đáng kể khi không đạt hiệu suất tốt nhất trong tất cả các thí nghiệm. Bên cạnh đó, các phương pháp này còn làm giảm hiệu suất so sánh với việc chỉ sử dụng dữ liệu huấn luyện gốc để huấn luyện trong một số trường hợp. Lý do là bởi vì phương pháp xóa hay đổi chỗ ngẫu nhiên sẽ dẫn đến việc mất mát thông tin khi xử lý. Ví dụ như phương pháp này có thể làm mất đi những từ khóa quan trong mang nhiều ý nghĩa cảm xúc gây thay đổi ý nghĩa của câu. Còn đối với phương pháp đổi chỗ ngẫu nhiên không làm mất mát thông tin nhưng việc đổi chỗ các cặp từ trong câu làm phá vỡ cấu trúc ngữ pháp và thay đổi ngữ cảnh của câu. Do đó kỹ thuật đổi chỗ ngẫu nhiên

Bảng 3. Kết quả thực nghiệm các kỹ thuật tăng cường dữ liệu trên các Bộ dữ liệu đơn ngữ và Nhận xét hỗn hợp Việt - Anh bằng các phương pháp học máy truyền thống trên các thang đo F1.

	SVM			NB			MLP		
	Macro	Micro	Weighted	Macro	Micro	Weighted	Macro	Micro	Weighted
Bộ dữ liệu đơn ngữ									
Origin	71.86	72.00	71.51	**74.00**	**73.50**	73.57	73.53	73.50	73.38
Swap	71.40	71.50	71.05	72.94	72.50	72.61	73.18	73.50	73.05
Delete	69.90	70.00	69.40	73.92	**73.50**	**73.58**	66.3	66.00	65.73
Insert Synonym	72.14	72.50	71.78	72.93	72.50	72.45	71.85	72.00	71.73
Insert Word2Vec	69.61	69.50	68.89	72.24	72.00	72.03	73.27	73.00	72.84
Insert TF-IDF	73.14	73.00	72.78	73.42	73.00	73.05	71.91	71.50	71.69
Insert BERT	72.92	73.00	72.47	73.84	**73.50**	73.39	71.13	71.00	70.66
Replace Synonym	71.47	71.50	70.98	72.49	72.00	71.97	**74.17**	**74.00**	**73.65**
Replace Word2Vec	**73.41**	**73.50**	**72.95**	73.40	73.00	73.07	71.83	71.50	71.33
Replace TF-IDF	72.21	72.00	71.69	68.86	68.50	68.61	72.37	72.00	72.00
Replace BERT	66.38	67.00	65.60	71.68	71.50	71.36	71.60	71.50	71.25
Back-translation	71.36	71.50	70.97	72.38	72.00	72.06	71.71	72.00	71.57
Bộ dữ liệu hỗn hợp									
Origin	72.50	72.50	72.41	72.00	72.00	72.28	71.50	71.50	70.63
Swap	72.50	72.50	72.40	72.00	72.00	72.10	70.00	70.00	69.81
Delete	71.00	71.00	70.81	72.00	72.00	72.13	71.00	71.00	71.14
Insert Synonym	72.00	72.00	71.89	71.00	71.00	71.16	**75.00**	**75.00**	**74.91**
Insert Word2Vec	72.50	72.50	72.07	70.50	70.50	70.60	71.50	71.50	71.54
Insert TF-IDF	**74.50**	**74.50**	**74.50**	72.50	72.50	72.59	72.00	72.00	71.98
Insert BERT	72.00	72.00	71.87	72.50	72.50	72.57	69.00	69.00	68.94
Replace Synonym	74.00	74.00	73.95	70.5	70.5	70.54	71.00	71.00	70.78
Replace Word2Vec	71.00	71.00	70.66	**73.00**	**73.00**	**72.97**	72.50	72.50	72.49
Replace TF-IDF	71.50	71.50	71.23	71.58	71.58	71.58	72.50	72.50	72.42
Replace BERT	70.00	70.00	69.27	72.50	72.50	72.51	71.00	71.00	70.03
Back-translation	73.00	73.00	72.79	72.50	72.50	72.75	70.50	70.50	70.45

cũng không giúp cải thiện hiệu suất so với dữ liệu huấn luyện gốc, nhưng hầu các trường hợp đều cho hiệu suất tốt hơn kỹ thuật xóa ngẫu nhiên.

Nhóm kỹ thuật chèn (insert) ngẫu nhiên: các phương phăp tăng cường chèn ngẫu nhiên bằng các chiến lược từ đồng nghĩa (synonym), Word2Vec, TF-IDF, BERT có tiềm năng cải thuật hiệu xuất, tuy nhiên có phần không ổn định, phụ thuộc vào bộ dữ liệu và mô hình phân loại:

- **Trên bộ dữ liệu đơn ngữ tiếng Việt**: Ngoài phương pháp chèn ngẫu nhiên bằng chiến lược sử dụng từ đồng nghĩa thì các chiến lược còn lại hầu hết đều không cải thiện hiệu suất so với việc không áp dụng việc tăng cường dữ liệu. Mặc dù việc chèn từ ngẫu nhiên giúp giúp bổ sung thêm thông tin cho câu. Tuy nhiên việc chèn ngẫu nhiên các từ cũng sẽ làm phá vỡ cấu trúc của câu. Đôi khi các chiến lược chèn tự động như Word2Vec,TF-IDF, BERT đôi khi sẽ gây sao nhãn thông tin, ảnh hưởng đến nhãn của câu.

Bảng 4. Kết quả thực nghiệm các kỹ thuật tăng cường dữ liệu trên các Bộ dữ liệu đơn ngữ và Nhận xét hỗn hợp Việt - Anh bằng phương pháp học sâu - CNN trên các thang đo F1.

	CNN					
	Bộ dữ liệu đơn ngữ			Bộ dữ liệu hỗn hợp		
	Macro	Micro	Weighted	Macro	Micro	Weighted
Origin	66.35	66.00	65.88	60.88	61.50	61.07
Swap	66.06	66.00	65.23	62.80	63.00	63.11
Delete	64.78	65.00	64.33	61.70	62.00	61.93
Insert Synonym	67.38	67.50	66.68	63.71	64.00	64.00
Insert Word2Vec	63.2	63.50	62.43	62.3	63.00	62.66
Insert TF-IDF	60.2	60.50	59.23	61.78	62.00	61.92
Insert BERT	61.08	61.50	60.27	64.33	64.50	64.37
Replace Synonym	**68.96**	**69.00**	**68.09**	63.83	64.00	64.19
Replace Word2Vec	64.38	65.00	63.73	63.27	63.50	63.39
Replace TF-IDF	68.32	68.00	67.66	62.24	62.50	62.44
Replace BERT	65.63	66.00	65.19	**65.84**	**66.00**	**66.00**
Back-translation	67.40	67.00	66.84	61.57	62.00	61.85

- **Trên bộ dữ liệu hỗn hợp Việt – Anh**: Tuy không ổn định, nhưng khi áp dụng nhóm kỹ thuật chèn trên bộ dữ liệu cũng có cái thiện hiệu suất. Đặc biệt là kỹ thuật chèn còn cho hiệu suất tốt nhất trong mô hình SVM và MLP. Chúng tôi thấy rằng, sở dĩ nhóm kỹ thuật chèn này gây giảm hiệu suất trên tập dữ liệu Nhận xét tiếng Việt nhưng lại giúp cải thiên hiệu xuất trên bộ dữ liệu Nhận xét hỗn hợp Việt – Anh là bởi vì một số nguyên nhân. Đầu tiên, bộ dữ liệu Nhận xét hỗn hợp Việt – Anh gốc về bản chất nó đã không tuân theo cấu trúc ngữ pháp của một văn bản đơn ngôn ngữ. Nên có thể nói việc chèn thêm từ trên bộ hỗn hợp không gây ảnh hưởng xấu về mặt cấu trúc câu như trên bộ tiếng Việt. Ngoài ra vì là hỗn hợp nên lượng thông tin sẽ được chia ra cho 2 ngôn ngữ, nên thông tin ở mỗi ngôn ngữ là tương đối ít. Việc sử dụng các chiến lược chèn từ trên bộ dữ liệu này sẽ giúp tăng thông tin ở mỗi ngôn ngữ, giúp cải thiện hiệu xuất phân loại.

Nhóm kỹ thuật thay thế (replace) ngẫu nhiên: đây có thể nói là nhóm kỹ thuật tốt, ổn định và cho hiệu suất tốt nhất trong nhiều trường hợp nhất trong các kỹ thuật tăng cường ở cấp từ. Vì nhóm kỹ thuật là đa dạng thêm thông tin nhưng không gây mất thông tin như kỹ thuật xóa ngẫu nhiên, cũng như không làm vỡ cấu trúc câu như nhóm kỹ thuật đổi chỗ và chèn ngẫu nhiên. Còn đối với kỹ thuật dịch ngược (back-translation) thì đây là một kỹ thuật đơn giản, dễ tiếp cận nhưng lại cho hiệu xuất ổn định nhất, dịch ngược giúp cải thiện hiệu suất trên hầu hết các mô hình và bộ dữ liệu. Đặc biệt là nhóm mô hình tinh chỉnh trên bộ dữ liệu Nhận xét tiếng Việt, dịch ngược cho hiệu suất tố nhất ở cả hai mô hình PhoBERT và XLM-RoBERTa. Tuy nhiên, nhóm chúng tôi thấy rằng đây là một kỹ thuật tăng cường an toàn, bởi vì nó giúp cải thiện tuy nhiên lại không có sự cải thiện vượt bật nhiều như các kỹ thuật khác.

5 Kết luận

Trong bài báo này, chúng tôi đã tiến hành nghiên cứu và đánh giá hiệu quả của các kỹ thuật tăng cường dữ liệu cho bài toán phân tích cảm xúc tiếng Việt. Qua việc sử dụng một loạt các phương pháp tăng cường dữ liệu ở cấp từng từ và toàn câu, chúng tôi đã thực hiện các phương pháp xóa từ ngẫu nhiên, đổi chỗ từ ngẫu nhiên, thay thế và chèn từ ngẫu nhiên bằng các chiến lược sử dụng từ đồng nghĩa, nhúng Word2Vec, TF-IDF và BERT, và dịch ngược để tạo dữ liệu mới và đa dạng.

Kết quả thực nghiệm đã cho thấy tính hiệu quả của các phương pháp tăng cường dữ liệu mà chúng tôi đã đề xuất. Hiệu quả này không chỉ xuất phát từ việc sáng tạo trong việc áp dụng các kỹ thuật tăng cường dữ liệu, mà còn từ sự lựa chọn mô hình phân loại phù hợp với đặc thù của ngôn ngữ tiếng Việt. Các kết quả thử nghiệm đã chứng minh rằng việc kết hợp tăng cường dữ liệu và lựa chọn mô hình phân loại thích hợp có thể cải thiện hiệu suất trong việc phân tích cảm xúc cho văn bản tiếng Việt. Đặc biệt là kỹ thuật thay thế từ ngẫu nhiên bằng các chiến lược và dịch ngược đã cho thấy tính ổn định và hiệu quả tốt nhất trên hầu hết các mô hình và hai bộ dữ liệu. Kết quả thực nghiệm cho thấy rằng các phương pháp tăng cường dữ liệu mà nhóm đã đề xuất đề có tiềm năng cải thiện hiệu suất của các bộ phân loại khác nhau. Hiệu quả này là kết quả của sự sáng tạo trong việc áp dụng các kỹ thuật tăng cường dữ liệu phù hợp với đặc thù của ngôn ngữ tiếng Việt.

Tuy nhiên, chúng ta không thể phớt lờ đi thách thức về khả năng có đủ dữ liệu chất lượng để huấn luyện các mô hình phân tích cảm xúc tiếng Việt. Điều này đặc biệt đúng đối với các ngôn ngữ ít phổ biến như tiếng Việt và xu hướng sử dụng ngôn ngữ hỗn hợp hiện nay. Bài báo này đã đóng góp quan trọng cho lĩnh vực này bằng cách giải quyết một thách thức quan trọng trong NLP tiếng Việt thông qua việc tạo dữ liệu mới và đa dạng thông qua các phương pháp tăng cường dữ liệu. Chúng tôi hy vọng rằng nghiên cứu của chúng tôi sẽ giúp thúc đẩy phát triển của lĩnh vực này và đánh dấu một bước tiến quan trọng trong việc phân tích cảm xúc tiếng Việt trong tương lai.

Tài liệu

1. Nassr, Zineb & Sael, Nawal & Benabbou, Faouzia. (2020). Machine Learning for Sentiment Analysis: A Survey. https://doi.org/10.1007/978-3-030-37629-1_6.
2. Jyothis Joseph, S. Vineetha, N.V. Sobhana,(2022). A survey on deep learning based sentiment analysis, https://doi.org/10.1016/j.matpr.2022.02.483
3. Sayyida Tabinda Kokab, Sohail Asghar, Shehneela Naz, (2022) Transformer-based deep learning models for the sentiment analysis of social media data, https://doi.org/10.1016/j.array.2022.100157
4. Chen, J., Tam, D., Raffel, C., Bansal, M., & Yang, D. (2021). An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. arXiv preprint arXiv:2106.07499 [cs.CL] https://arxiv.org/abs/2106.07499.pdf
5. P. Awatramani, R. Daware, H. Chouhan, A. Vaswani and S. Khedkar, (2021). "Sentiment Analysis of Mixed-Case Language using Natural Language Process-

ing," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 651-658,
6. Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, Eduard Hovy. (2021). A Survey of Data Augmentation Approaches for NLP. arXiv:2105.03075 [cs.CL]. https://doi.org/10.48550/arXiv.2105.03075
7. Pellicer, L. F. A. O., Ferreira, T. M., & Costa, A. H. R. (2022). Data Augmentation Techniques in Natural Language Processing. Applied Soft Computing, 2022, 109803. https://doi.org/10.1016/j.asoc.2022.109803
8. Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv preprint arXiv:1901.11196 [cs.CL]. https://doi.org/10.48550/arXiv.1901.11196
9. Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., & Zwerdling, N. (2019). Not Enough Data? Deep Learning to the Rescue! arXiv preprint arXiv:1911.03118 [cs.CL]. https://doi.org/10.48550/arXiv.1911.03118
10. Thin Dang Van, Hao Duong Ngoc, and Ngan Nguyen Luu-Thuy. (2022). Sentiment Analysis in Code-Mixed Vietnamese-English Sentence-level Hotel Reviews. https://aclanthology.org/2022.paclic-1.7.pdf
11. Vu, T., Nguyen, D. Q., Nguyen, D. Q., Dras, M., & Johnson, M. (2018). VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 56–60). New Orleans, Louisiana: Association for Computational Linguistics https://aclanthology.org/N18-5012.pdf
12. George A. Miller. (1995). WordNet: a lexical database for English. Commun. ACM 38, 11 (Nov. 1995), 39–41.https://doi.org/10.1145/219717.219748
13. Y. Kim. Convolutional neural networks for sentence classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746-17. https://aclanthology.org/D14-1181.pdf
14. Dat Quoc Nguyen, Anh Tuan Nguyen. (2020) "PhoBERT: Pre-trained language models for Vietnamese." arXiv:2003.00744 [cs.CL]. https://doi.org/10.48550/arXiv.2003.00744
15. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov. (2020) “Unsupervised Cross-lingual Representation Learning at Scale” arXiv:1911.02116 [cs.CL]. https://doi.org/10.48550/arXiv.1911.02116
16. Ton Nu Thi Sau, Do Phuoc Sang, Pham Thi Thu Trang. (2021). “ASPECT-BASED SENTIMENT ANALYSIS ON STUDENT’S FEEDBACK IN VIETNAMESE”. https://vjol.info.vn/index.php/tnu/article/view/63913