Empirical Study of Text Augmentation for Span Detection for Aspect-Based Sentiment Analysis in Vietnamese

Nguyễn T.Thiện Quá[1], Nguyễn Hiếu Nghĩa[2], Ngô Thị Phúc[3], Huỳnh L. Phương Vy[4]

[1]20521783, [2]20521654, [3]20521765, [4]20520951 - KHDL2020

Ngày 27 tháng 12 năm 2022

Tóm tắt nội dung

Những bình luận phản hồi trong các hệ thống trực tuyến là một nguồn dữ liệu mang nhiều thông tin, cảm xúc của khách hàng về những sản phẩm hoặc dịch vụ. Những thông tin này được khai thác nhằm mang lại lợi ích trong việc hoạch định chiến lược, quản trị khách hàng. Bên cạnh đó, phân tích tình cảm dựa trên khía cạnh và phát hiện ý kiến người dùng là một nhiệm vụ đầy thách thức đóng vai trò quan trọng trong xử lý ngôn ngữ tự nhiên. Các công trình gần đây như nghiên cứu của Thanh et al. (2021) với bộ dữ liệu UIT-ViSD4SA về những phản hồi liên quan đến điện thoại thông minh trên các trang thương mại điện tử đã được công bố nhưng kết quả chỉ ở mức khá và bộ dữ liệu đã gán nhãn đang bị mất cân bằng nghiệm trọng. Trong đồ án này, chúng tôi sẽ tiến hành tăng cường dữ liệu cho bộ dữ liệu trên bằng phương pháp tăng cường dữ liệu đơn giản, thu thập thêm dữ liệu để giải quyết vấn đề mất cân bằng dữ liệu và xây dựng mô hình phoBERT nhằm tăng hiệu suất mô hình. Kết quả thực nghiệm đã chứng minh phương pháp của chúng tôi tăng 3.1% độ đo f1-score macro trên hiệu suất của mô hình.

1 Giới thiệu

Thông thường, trước khi mua một món hàng hay quyết định sử dụng dịch vụ online nào đó, mọi người thường có xu hướng tìm kiếm các lời khuyên, review từ những người đã mua món hàng hoặc sử dụng dịch vụ đó. Kèm theo sự phát triển vượt trội của các trang thương mại điện tử, các diễn đàn đánh giá sản phẩm. Do đó, số lượng đánh giá ngày càng tăng và trở thành nguồn tài nguyên quý giá cho khách hàng và cho doanh nghiệp. Đối với khách hàng, nguồn dữ liệu này cung cấp thông tin về sản phẩm và những lời khuyên hữu ích giúp họ tránh mua các sản phẩm hoặc đăng ký dịch vụ không phù hợp. Mặc khác, đánh giá của người dùng cũng là thông tin quý giá cho doanh nghiệp, nếu sử dụng hiệu quả, nguồn dữ liệu này có thể giúp cho doanh nghiệp nâng cao chất lượng sản phẩm, xác định chính xác nhu cầu khách hàng.

Phân tích cảm xúc dựa trên khía cạnh có thể được xem là một bài toán trong khai thác văn bản thuộc lĩnh vực xử lý ngôn ngữ tự nhiên. Do đó phải hiểu được ngữ nghĩa trong bối cảnh nhất định, cho nên việc phân tích trên những đoạn văn bản ngắn khó khăn hơn so với những đoạn văn bản dài. Dựa trên mục đích của việc phân loại tình cảm, cảm xúc của một khía cạnh có thể được phân ra thành: tích cực, tiêu cực và trung lập. Như vậy, việc thu thập một lượng lớn dữ liệu không có nhãn từ các hệ thống mạng xã hội là tương đối đơn giản nhưng việc gán nhãn đầy đủ và chính xác tốn rất nhiều thời gian và chi phí. Kết quả mô hình phần lớn dựa vào dữ liệu được gán nhãn, đồng thời yêu cầu số lượng dữ liệu đủ lớn. Phương pháp tăng cường dữ liệu đầu vào cho một mô hình là một trong những phương pháp ít tốn kém nhưng hiệu quả để giải quyết vấn đề này. Việc làm tăng thêm dữ liệu này được áp dụng rộng rãi trong các bài toán thị giác máy tính (Perez and Wang (2017)) bằng cách sử dụng những kỹ thuật đơn giản như lật hình, xoay hình, cắt hình, thay đổi tỷ lệ ảnh hoặc biến đổi màu sắc (Duong and Hoang (2019)) nhằm thay đổi hình ảnh ban đầu. Do sự phức tạp về mặt ngữ nghĩa, sự đa dạng về mặt ngữ pháp và ngữ cảnh của ngôn ngữ, cho nên phương pháp làm tăng thêm dữ liệu đối với bài toán sử dụng dữ liệu văn bản vẫn còn là vấn đề nhiều thách thức.

Bộ dữ liệu UIT-ViSD4SA (Thanh et al. (2021)) bao gồm 35.396 khoảng vị trí trong câu đã được gán nhãn từ 11.122 nhận xét từ các trang thương mại điện tử để phục vụ bài toán Span Detection cho phân tích tình cảm dựa trên khía cạnh. Ví dụ như câu: *"Sp ổn, mỗi tội vân tay lúc*

nhận lúc không, nhân viên nhiệt tình, pin trâu, cả đêm tụt 1%" sẽ được gán nhãn là: [[0, 5, "GENERAL#POSITIVE"], [15, 41, "FEATURES#NEGATIVE"], [43, 63, "SER&ACC#POSITIVE"], [65, 88, "BATTERY#POSITIVE"]]. Bộ dữ liệu gồm 10 khía cạnh và 3 loại cảm xúc. Tuy nhiên, sự phân bố của các nhãn và cảm xúc này lại không đồng đều và vô vùng mất cân bằng. Vì thế, chúng tôi tiến hành thu thập thêm dữ liệu, gán nhãn chúng, đồng thời chúng tôi cũng sử dụng các kĩ thuật tăng cường dữ liệu để tạo ra các câu bình luận mới thuộc nhóm thiểu số trong tập dữ liệu gốc để giải quyết vấn đề này. Chúng tôi cũng tiến hành thử nghiệm trên tập dữ liệu tăng cường và so sánh với tập dữ liệu gốc để chỉ ra sự hiệu quả của phương pháp này. Những kĩ thuật tăng cường dữ liệu này bao gồm thay thế từ đồng nghĩa, chèn từ ngẫu nhiên vào câu, hoán đổi ngẫu nhiên giữa các từ trong câu và xóa ngẫu nhiên từ trong câu (Wei and Zou (2019)).

Nội dung của các phần tiếp theo được trình bày như sau: Phần 2 sẽ trình bày các công trình liên quan, phần 3 sẽ giới thiệu bộ dữ liệu UIT-ViSD4SA, phần 4 sẽ trình bày các phương pháp tăng cường dữ liệu, phần 5 sẽ xây dựng mô hình phoBERT và kết quả thực nghiệm, cuối cùng sẽ là kết luận rút ra sau khi hoàn thành đồ án.

2 Các công trình nghiên cứu liên quan

2.1 Tăng cường dữ liệu văn bản

Tăng cường dữ liệu văn bản đề cập đến các phương pháp làm xáo trộn không gian ngôn ngữ mà không làm thay đổi nhãn để cải thiện tính mạnh mẽ và khả năng khái quát hóa của mô hình trong NLP. So với các ngôn ngữ giàu tài nguyên như tiếng Anh, tiếng Hoa, số lượng bộ dữ liệu tiếng Việt chất lượng cao vẫn còn hạn chế. Tăng cường dữ liệu đã được nghiên cứu rộng rãi trong bối cảnh hiện nay. Các kỹ thuật tăng cường dữ liệu đưa sự nhiễu nhỏ cục bộ và hợp lý vào không gian ngôn ngữ (từ hoặc cụm từ), với hy vọng rằng các nhiễu loạn tạo ra các mẫu có thể chấp nhận được về mặt ngôn ngữ trong khi vẫn duy trì được sự nhất quán của nhãn Luu et al. (2020)

Một loại kỹ thuật tăng cường khác sử dụng các mô hình ngôn ngữ bên ngoài để cải thiện sự gắn kết và nhất quán toàn cục. Phương pháp dịch ngược khai thác tính nhất quán ngữ nghĩa trong các cặp ngôn ngữ dịch để tạo ra các cách diễn đạt mới lạ Fadaee et al. (2017). Gần đây hơn, các mô hình ngôn ngữ được đào tạo trước, chẳng hạn như BERT Devlin et al. (2018) hoặc biến thể BART theo trình tự Lewis et al. (2019) được sử dụng để đạt được sự đa dạng hơn về mặt ngôn ngữ mẫu tăng cường chính xác. Ví dụ: BART đã được chứng minh là có hiệu quả trong việc điền các mẫu văn bản cho các nhãn khan hiếm dữ liệu và không cân bằng Kumar et al. (2020). Một số nhà nghiên cứu khác đã đi theo hướng làm xáo trộn các không gian tiềm ẩn, tùy ý bằng cách đưa ra suy luận biến thể trong kiến trúc (Chen et al. (2020), Xie et al. (2019)).

Duyen et al. (2014) đã thực hiện một nghiên cứu thực nghiệm về phân tích tình cảm cho các văn bản tiếng Việt dựa trên học máy để nghiên cứu ảnh hưởng đến độ chính xác của mô hình. Tuy nhiên, bên cạnh ảnh hưởng từ khả năng của mô hình và việc lựa chọn các tính năng như dựa trên từ, dựa trên âm tiết và trích xuất các từ cần thiết, thì sự mất cân bằng trong tập dữ liệu cũng ảnh hưởng đến kết quả. Sự mất cân bằng trong phân bổ nhãn hiệu diễn ra thường xuyên Ali et al. (2015) khi mô hình tập trung vào một nhãn hơn các nhãn còn lại. Ví dụ: trong các mạng truyền thông xã hội, các bình luận lăng mạ và thù hận thường được người dùng hoặc quản trị viên ẩn đi vì các bình luận trong sạch chiếm đa số. Bộ dữ liệu ngôn từ kích động thù địch VLSP2019 (Vu et al. (2020)) và bộ dữ liệu UIT-VSFC (Nguyen et al. (2018)) cũng chịu sự mất cân bằng trong phân bổ các lớp. Bên cạnh đó, Ibrahim et al. (2018) đã trình bày các kỹ thuật tăng cường dữ liệu khác nhau để giải quyết vấn đề mất cân bằng trong bộ dữ liệu Wikipedia và một phương pháp tập hợp được sử dụng cho mô hình đào tạo. Cuối cùng, (Wei and Zou (2019)) đã cung cấp các kỹ thuật EDA (Easy Data Augmentation) được sử dụng để cải thiện dữ liệu và tăng hiệu suất đối với nhiệm vụ phân loại văn bản. Nó chứa bốn thao tác: thay thế từ đồng nghĩa, chèn ngẫu nhiên, hoán đổi ngẫu nhiên và xóa ngẫu nhiên. Trong đồ án này, chúng tôi sẽ áp dụng kỹ thuật này cho bộ dữ liệu UIT-ViSD4SA để cải thiện hiệu suất.

2.2 Phân tích tìm cảm dựa trên khía cạnh

Aspect-based Sentiment Analysis (ABSA) là một loại phân tích văn bản phân loại ý kiến theo khía cạnh và xác định tình cảm liên quan đến từng khía cạnh. Lưu ý rằng trong cùng một câu, các khía cạnh khác nhau có thể có những tình cảm khác nhau. Theo nghĩa này, đầu ra của ABSA không phải

là một dự đoán chung về tình cảm được thể hiện trong câu mà nhằm mục đích cung cấp mức độ thông tin chi tiết hơn. Gần đây, một nhiệm vụ chia sẻ về phân tích cảm xúc dựa trên khía cạnh tiếng Việt đã được giới thiệu tại hội thảo quốc tế lần thứ năm về Xử lý giọng nói và ngôn ngữ tiếng Việt (VLSP 2018). Trong bối cảnh của bài toán được chia sẽ, một số nghiên cứu đã được trình bày bằng cách sử dụng thuật toán học có giám sát (Dang et al. (2019)), công việc của chúng tôi khác với những nghiên cứu đó ở chỗ sẽ chỉ ra vị trí cụ thể của khía cạnh tồn tại trong câu, chúng tôi tin rằng nhiệm vụ này sẽ thực tế hơn và có thể áp dụng trong các ứng dụng trong thế giới thực.

2.3 Thách thức của phân tích tình cảm dựa trên khía cạnh

Từ góc độ ngôn ngữ học, nhiệm vụ tự động xác định các khía cạnh và tình cảm liên quan đặt ra nhiều thách thức, bởi vì chúng ta thường thấy các hiện tượng ngôn ngữ phức tạp không dễ diễn giải và hiểu. Đặc biệt, xác định đúng khía cạnh có thể được coi là nhiệm vụ khó nhất của phân tích này, bởi vì khách hàng có thể bày tỏ ý kiến của họ về rất nhiều khía cạnh (ví dụ: đối với một sản phẩm: giá cả, chất lượng, hiệu suất hoặc thiết kế), tùy thuộc vào tình huống, trong khi thực hiện phân tích cảm tính dựa trên khía cạnh, chúng tôi gặp phải những thách thức sau:

- **Khi tình cảm phụ thuộc vào mục tiêu**

Trong khi mô tả sản phẩm hoặc dịch vụ, khách hàng cũng có thể tham chiếu đến các thực thể khác. Ví dụ: "*Sản phẩm này có giá tốt; nhưng cái mà anh tôi mua có thiết kế đẹp hơn.*" Trong ví dụ, nó không đủ để mô hình có thể truy xuất tất cả các khía cạnh được đề cập trong văn bản và cảm xúc liên quan của chúng (PRICE#POSITIVE và DESIGN#POSITIVE). Điều quan trọng là mô hình chỉ có thể chọn cặp price#positive, vì đó là cặp liên quan đến sản phẩm đang được đề cập, Không nên chọn thiết kế cặp design#positive vì nó đang đề cập đến một thực thể khác. Từ quan điểm tổng quát hơn, mô hình cần có khả năng phân biệt "mục tiêu" (khía cạnh chính của phân tích hiện tại) với các thực thể khác.

- **Khi tình cảm không mang nghĩa đen**

Cho đến nay, chúng tôi đã thấy các ví dụ về văn bản phản hồi trong đó cảm xúc được truyền đạt ngầm (chỉ hiểu được nhờ vào hiểu biết về sản phẩm), được truyền đạt rõ ràng (nhưng đôi khi không đề cập đến mục tiêu) không nên hiểu theo nghĩa đen. "*100% pin về 0 sau 3h sử dụng, thật sự tuyệt vời!*" Mặc dù bản thân cụm từ *"thực sự tuyệt vời!"* rõ ràng là tích cực , nhưng ngữ cảnh cho thấy khách hàng muốn truyền đạt cảm xúc ngược lại, do đó đưa ra phản hồi tiêu cực. Những trường hợp này phụ thuộc nhiều vào ngữ cảnh và sự giải thích của con người.

3 Bộ dữ liệu

Bộ dữ liệu UIT-ViSD4SA cung cấp bởi Thanh et al. (2021) phát triển Task Span Detection bằng tiếng Việt dựa trên benchmark dataset do Phan et al. (2021) đề xuất, thu thập từ các bình luận sản phẩm trên các trang thương mại điện tử về điện thoại thông minh ở Việt Nam. Bộ dữ liệu bao gồm 11,122 câu bình luận chứa một hoặc nhiều khía cạnh rõ ràng hoặc ẩn ý liên quan về các đặc điểm của điện thoại thông minh đã được gán nhãn (chỉ số bắt đầu khía cạnh (aspect), chỉ số kết thúc khía cạnh, aspect#polarity). Bảng 1 mô tả tóm tắt 10 khía cạnh, với mỗi khía cạnh có một trong ba cảm xúc tích cực (positive), tiêu cực (negative) và trung lập (neutral). Tổng quan về tập huấn luyện của bộ dữ liệu được mô tả chi tiết trong Bảng 2

4 Phương pháp tăng cường dữ liệu

4.1 Dữ liệu tự thu thập và gán nhãn

Chúng tôi tiến hành thu thập thêm một số câu bình luận về điện thoại thông minh trên các diễn đàn thương mại điện tử như tiki, lazada, shoppe,... với điều kiện tất cả các dữ liệu phải được người dùng nhận xét từ đầu năm 2022 đến hiện tại, điều này nhằm mục đích hạn chế tối đa việc trùng lặp dữ liệu với bộ UIT-ViSD4SA cũng thu thập từ các nguồn tương tự từ 2021 trở về trước. Và điều đặc biệt là chúng tôi chỉ sưu tầm dữ liệu nghiên về các khía cạnh mà bộ dữ liệu trên đang khan hiếm như

Khía cạnh	Mô tả
SCREEN	Nhận xét về chất lượng màn hình, kích thước, màu sắc, công nghệ hiển thị.
CAMERA	Nhận xét về chất lượng của máy ảnh, độ rung, độ trễ, lấy nét.
FEATURES	Các tính năng, cảm biến vân tay, kết nối wifi hay nhận diện khuôn mặt.
BATTERY	Các ý kiến mô tả dung lượng pin hay chất lượng pin.
PERFORMANCE	Dung lượng RAM, chip xử lý, hiệu năng hay độ mượt của điện thoại
STORAGE	Nhận xét về khả năng lưu trữ, khả năng mở rộng dung lượng qua thẻ nhớ.
DESIGN	Các đánh giá đề cập đến kiểu dáng, thiết kế hoặc vỏ máy.
PRICE	Các ý kiến trình bày giá của điện thoại.
GENERAL	Nhận xét chung chung về điện thoại
SER&ACC	Các bình luận đề cập đến dịch vụ bán hàng, bảo hành, hoặc phụ kiện.

Bảng 1: Danh sách đầy đủ mười khía cạnh và định nghĩa ngắn gọn của chúng

Bình luận	Trung bình khía cạnh trên câu	Positive	Negative	Neutral	Tổng Span
11,122	3.2	21,732	11,206	2,214	35,396

Bảng 2: Tổng quan về tập huấn luyện bộ dữ liệu UIT-ViSD4SA

storage, screen, price, v.v. Kế thừa guideline có sẵn từ bộ dữ liệu trên, chúng tôi tiến hành đào tạo người gán nhãn, dựa trên sự thống nhất giữa các thành viên về quy trình gán nhãn, chúng tôi lấy ngẫu nhiên 40 câu bình luận trong bộ dữ liệu mới thu thập gán nhãn độc lập và dùng độ đo Cohen's Kappa để đánh giá mức độ đồng thuận giữa những thành viên gán nhãn. Kết quả cuối cùng chúng tôi sưu tầm được 2000 câu bình luận đã gán nhãn với độ đồng thuận 54.2%.

4.2 Các kỹ thuật tăng cường

Trong đồ án này, chúng tôi triển khai các kỹ thuật EDA được giới thiệu bởi Wei and Zou (2019) trên các bộ dữ liệu tiếng Anh và nghiên cứu trên bộ dữ liệu tiếng Việt của Luu et al. (2020). Những kỹ thuật đó sẽ lấy một câu làm đầu vào và thực hiện một trong các thao tác sau để tạo ra các câu mới:

- **Thay thế từ đồng nghĩa**: Thao tác này tạo một câu mới bằng cách chọn ngẫu nhiên n từ, từ câu đầu vào và thay thế chúng bằng các từ đồng nghĩa của nó, không tính stop words. Đã sử dụng wordnet tiếng Việt của Nguyen et al. (2016) để thay từ đồng nghĩa và Từ điển tiếng Việt để loại bỏ stop words trong câu.
- **Chèn từ ngẫu nhiên**: Thao tác này tạo dữ liệu mới bằng cách trước tiên tìm một từ ngẫu nhiên trong câu đầu vào, không phải là stop word, sau đó lấy từ đồng nghĩa của từ này và đặt từ đó vào vị trí ngẫu nhiên của câu. Các từ đồng nghĩa được lấy từ wordnet tiếng Việt.
- **Hoán đổi ngẫu nhiên**: Kỹ thuật này tạo một câu mới bằng cách chọn hai từ ngẫu nhiên trong câu đầu vào và hoán đổi vị trí của chúng.
- **Xóa ngẫu nhiên**: Kỹ thuật này tạo ra một câu mới bằng cách ngẫu nhiên xóa p từ trong câu (p là xác suất được người dùng xác định trước đó).

Theo Wei and Zou (2019), n biểu thị số lượng từ bị thay đổi đối với các phương pháp thay từ đồng nghĩa, chèn ngẫu nhiên và hoán đổi ngẫu nhiên, được tính như sau n = *alpha**l, trong đó *alpha* là tỷ lệ phần trăm của từ thay thế trong câu và l là độ dài của câu. Đối với phương pháp xóa ngẫu nhiên, xác suất xóa từ p bằng *alpha*. *alpha* được xác định bởi người dùng. Bảng 3 cho thấy các ví dụ về dữ liệu gốc và dữ liệu được tạo ra bằng các kỹ thuật EDA trong bộ dữ liệu UIT-ViSD4SA.

4.3 Kết quả tăng cường dữ liệu

Chúng tôi áp dụng kỹ thuật EDA cho bộ dữ liệu UIT-ViSD4SA để tăng cường dữ liệu 6 nhãn khan hiếm như STORAGE, PRICE, SCREEN, DESIGN, CAMARA VÀ SER&ACC trong tập dữ liệu huấn

Kỹ thuật	Câu bình luận
Câu gốc	Nhân viên phục vụ nhiệt tình và vô cùng lịch sự
Thay thế từ	Nhân viên **cung cấp** nhiệt tình và vô cùng lịch sự
Chèn từ ngẫu nhiên	Nhân viên phục vụ **giữ** nhiệt tình và vô cùng lịch sự
Hoán đổi từ	**và** viên phục vụ nhiệt tình **Nhân** vô cùng lịch sự
Xóa từ ngẫu nhiên	Nhân viên phục vụ nhiệt tình và lịch sự

Bảng 3: Ví dụ câu được tạo ra từ các kỹ thuật EDA

luyện và giữ nguyên tập kiểm thử để đánh giá kết quả. Sau khi dữ liệu được tạo ra, chúng tôi tiến hành lọc lại và xóa đi một số câu vô nghĩa. Có thể nhận thấy thông qua Bảng 4 rằng sau khi áp dụng kỹ thuật EDA thì số lượng dữ liệu ở các nhãn khan hiếm đã tăng đáng kể. Hình 1 minh họa sự phân phối dữ liệu ở các nhãn trước và sau khi tăng cường dữ liệu. Ở bộ dữ liệu sau cùng, 10 khía cạnh đã có sự phân phối tốt hơn, sự mất cân bằng đã được cải thiện đáng kể.

Tuy nhiên, sự phân phối trên chỉ cải thiện giữa các khía cạnh, còn trong mỗi khía cạnh, sự phân phối cảm xúc giữa 3 nhãn cảm xúc vẫn chưa được cải thiện, phần lớn các khía cạnh đều mang cảm xúc tích cực (POSITIVE), đối với bình luận trung lập (NEUTRAL) vẫn chỉ chiếm phần khá ít trên mỗi khía cạnh.

Bình luận	Trung bình khía cạnh trên câu	Positive	Negative	Neutral	Tổng Span
34530	1.9	36,951	20,905	4,865	62,721

Bảng 4: Tổng quan về bộ dữ liệu UIT-ViSD4SA sau khi áp dụng kỹ thuật EDA

Hình 1: Kết quả tăng cường dữ liệu. Hình bên trên là dữ liệu ban đầu, hình bên dưới là kết quả của quá trình tăng cường dữ liệu

5 Triển khai mô hình

5.1 BERT và RoBERTa

BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019) và RoBERTa (Robustly Optimized BERT Pretraining Approach) Liu et al. (2019) cả hai đều lấy Transformers Vaswani et al. (2017) làm kiến trúc xương sống, được phát triển để lập sequence-to-sequence modeling để giải quyết vấn đề phụ thuộc tầm xa. Transformer models gồm 3 phần: tokenizer, transformers và heads. Tokenizer biến đổi câu đầu vào thành sparse index encodings sau đó transformers sẽ tiến hành nhúng theo ngữ cảnh để đào tạo sâu hơn, cuối cùng heads được triển khai để bao bọc mô hình transformer sao cho có thể sử dụng tính năng nhúng theo ngữ cảnh cho các tác vụ downtream.

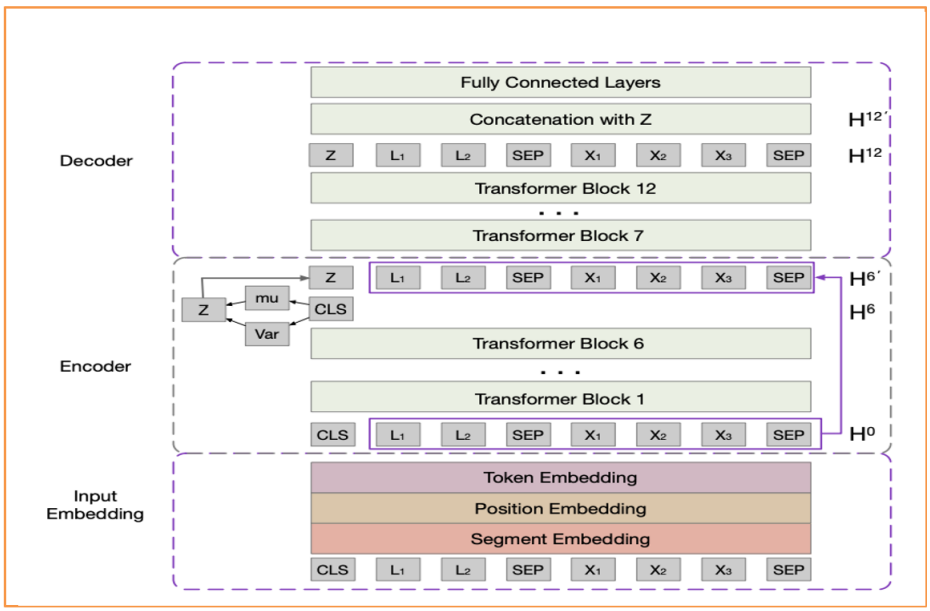
BERT hơi khác so với các mô hình ngôn ngữ hiện có ở chỗ nó có thể học cách thể hiện theo ngữ cảnh từ cả hai đầu câu. Đối với phần tokenization BERT dùng 30 nghìn từ vựng cho character level Byte-Pair Encoding, ngược lại RoBERTa dùng byte-level Byte-Pair Encoding với tập từ vựng lớn chứa 50 nghìn subword units. Ngoài ra, mô hình RoBERTa tinh chỉnh mô hình BERT bằng cách đào tạo trên nhiều dữ liệu hơn, các câu dài hơn, thời gian lâu hơn và dynamically changing masking được áp dụng cho dữ liệu huấn luyện.

Nói chung, cả hai có thể được xây dựng như các phương trình sau:

$$\hat{h}^l = LN(h^{l-1} + MHAtt(h^{l-1})) \tag{1}$$

$$h^l = LN(\hat{h}^l + FFN(\hat{h}^l)) \tag{2}$$

Trong đó h^0 biểu diễn đầu vào BERT/RoBERTa, được hình thành bởi tổng số lần token embedding, position embedding và segment embedding. LN là layer normalization layer, MHAtt là multi-head self-attention; FFN chứa 3 lớp, lớp đầu tiên là lớp linear projection, sau đó là lớp activation, và cuối cùng là lớp linear projection khác; l là độ sâu của lớp transformer, base và large của BERT và RoBERTa có 12 và 24 lớp transformer. Cấu trúc tổng quát được thể hiện ở sơ đồ hình 2.



Hình 2: Sơ đồ mô hình phoBERT

5.2 Thiết kế thí nghiệm

Trong phần này sẽ tiến hành thí nghiệm như sau: Đầu tiên chúng tôi sẽ tiến hành xây dựng mô hình mới so với bộ dữ liệu gốc nhằm cố gắng cải thiện hiệu suất hiện tại bằng mô hình được đào tạo trước phoBERT Nguyen and Tuan Nguyen (2020) trên bộ dữ liệu gốc, sau đó chạy lại mô hình trên

bộ dữ liệu tăng cường và so sánh kết quả đạt được. Bản chất của phoBERT chính là RoBERTa trên corpus tiếng Việt.

Theo BIO format (short for Begin, Inside, Outside), tập dữ liệu huấn luyện được chuyển đổi chứa nhãn khía cạnh và nhãn cảm xúc kết hợp lại với nhau (PRICE#POSITIVE, SCREEN#NEUTRAL,...) đưa vào mô hình phoBERT-base gồm 12 tầng con, kích thước nhúng 768 và số lượng head attention là 12.

Chúng tôi dựa trên 3 chỉ số đánh giá theo Thanh et al. (2021): Precision, Recall và f1-score, tính toán các chỉ số đánh giá này trên cả mức average micro và average macro để có được cái nhìn bao quát nhất. Với đầu vào và đầu ra được xác định như sau:

- **Input**: Một câu bình luận của khách hàng S về điện thoại di động bao gồm n kí tự.
- **Output**: Cảm xúc của khách hàng về một hay nhiều khía cạnh được trích xuất trực tiếp từ câu S. Mỗi ý kiến được trích xuất từ vị trí i đến j với điều kiện $0 \leq i, j \leq n$ và $i \leq j$.

5.3 Kết quả thực nghiệm

Khía cạnh	**Precision**			**Recall**			**f1-score**		
	CRF	phoB	phoB*	CRF	phoB	phoB*	CRF	phoB	phoB*
NEGATIVE	47.05	54.00^{+}	60.40	47.56	55.79^{+}	61.38	47.30	54.8^{+}	60.89
NEUTRAL	36.57	47.04^{+}	49.50	35.97	41.57^{+}	45.43	36.26	44.14^{+}	45.44
POSITIVE	63.52	72.78^{+}	75.11	68.5	74.66^{+}	77.73	65.92	73.71^{+}	**76.40**

Bảng 5: Kết quả trên mỗi nhãn cảm xúc

Khía cạnh	**Precision**			**Recall**			**f1-score**		
	CRF	phoB	phoB*	CRF	phoB	phoB*	CRF	phoB	phoB*
BATTERY	71.04	75.11^{+}	78.43	73.58	76.12^{+}	79.62	72.29	75.61^{+}	79.02
CAMERA	75.09	75.88^{+}	79.97	77.82	76.50	82.17	76.43	76.19	**81.06**
DESIGN	68.13	70.59^{+}	71.16	70.66	70.24	74.63	69.37	70.42^{+}	72.86
FEATURES	58.76	56.00	60.20	59.34	57.84	64.52	59.05	56.90	62.29
GENERAL	64.74	65.37^{+}	68.41	68.90	67.43	70.40	66.76	66.38	69.39
PERFORMANCE	62.37	63.81^{+}	66.22	63.11	65.35^{+}	69.08	62.74	64.57^{+}	67.62
PRICE	46.72	46.40	54.15	47.98	49.81^{+}	57.92	47.35	48.04^{+}	55.97
SCREEN	65.83	67.35^{+}	69.37	68.70	71.74^{+}	73.19	67.23	69.47^{+}	69.66
SER&ACC	65.18	56.72	66.36	61.83	65.27	69.27	63.46	60.69	67.79
STORAGE	45.16	55.17^{+}	57.24	46.67	47.06^{+}	48.08	45.90	50.79^{+}	52.26

Bảng 6: Kết quả trên mỗi nhãn khía cạnh

Khía cạnh	**NEGATIVE**			**NEUTRAL**			**POSITIVE**		
	CRF	phoB	phoB*	CRF	phoB	phoB*	CRF	phoB	phoB*
BATTERY	54.64	59.44^{+}	67.33	44.07	52.31^{+}	46.14	78.40	81.21^{+}	80.96
CAMERA	58.97	58.36	70.20	55.65	58.23^{+}	54.55	77.54	80.62^{+}	**83.54**
DESIGN	46.15	41.88	54.00	00.00	28.57^{+}	33.33	75.75	76.00^{+}	78.51
FEATURES	50.73	48.94	54.30	22.22	45.45^{+}	41.79	68.11	66.67	69.71
GENERAL	52.12	54.78^{+}	56.76	52.73	46.85	46.02	67.87	67.05	70.60
PERFORMANCE	45.87	50.61^{+}	54.73	24.19	27.14^{+}	30.86	70.84	71.65^{+}	74.29
PRICE	32.69	39.37^{+}	47.76	15.05	29.63^{+}	22.41	52.63	49.07	57.97
SCREEN	48.62	50.67^{+}	55.20	46.15	35.29	40.00	71.13	77.48^{+}	78.15
SER&ACC	22.56	29.39^{+}	41.38	00.00	28.57^{+}	40.00	72.17	67.27	72.45
STORAGE	15.38	34.78^{+}	45.45	00.00	44.44^{+}	41.11	57.14	60.00^{+}	58.06

Bảng 7: Kết quả f1-score trên từng nhãn aspect#sentiment

System	$\mathbf{P}_{Micro}$	$\mathbf{R}_{Micro}$	$\mathbf{F1}_{Micro}$	$\mathbf{P}_{Macro}$	$\mathbf{R}_{Macro}$	$\mathbf{F1}_{Macro}$
Aspect (XLM-R_{Base})	65.63	65.15	65.39	62.88	61.62	62.17
Aspect (XLM-R_{Large})	64.96	66.85	65.89	62.00	63.56	62.76
phoBERT	65.28^{+}	67.5^{+}	66.37^{+}	63.24^{+}	64.74^{+}	63.91^{+}
phoBERT + data augmentation	68.76	71.65	**70.17**	65.35	68.79	**67.01**
Polarity (XLM-R_{Base})	54.88	55.91	55.39	46.87	46.39	46.57
Polarity (XLM-R_{Large})	56.89	59.78	58.30	49.00	50.60	49.77
phoBERT	65.26^{+}	66.55^{+}	65.90^{+}	57.94^{+}	57.34^{+}	57.57^{+}
phoBERT + data augmentation	68.50	70.13	**69.30**	59.01	59.51	**59.24**
Aspect-polarity (XLM-R_{Base})	60.71	61.62	61.16	46.18	43.42	44.37
Aspect-polarity (XLM-R_{Large})	61.78	62.99	62.38	46.84	45.46	45.70
phoBERT	62.13^{+}	64.16^{+}	63.13^{+}	53.91^{+}	51.95^{+}	52.06^{+}
phoBERT + data augmentation	65.25	68.25	**66.72**	54.16	54.93	**54.22**

Bảng 8: Tổng quan kết quả và so sánh giữa các mô hình

Bảng 8 trình bày kết quả hiệu suất của mô hình phoBERT trên bộ dữ liệu UIT-ViSD4SA (màu vàng nhạt) và trên bộ dữ liệu tăng cường (màu vàng đậm), bên cạnh đó cũng trình bày kết quả nghiên cứu của Thanh et al. (2021) bằng mô hình BiLSTM-CRF (màu trắng). Trong đó CRF là mô hình BiLSTM-CRF theo nghiên cứu của tác giả, phoB là kết quả của mô hình phoBERT trên bộ dữ liệu UIT-ViSD4SA, phoB* là kết quả của mô hình phoBERT trên tập dữ liệu tăng cường. Đối với những số có dấu (+) đánh dấu nhận biết mô hình phoBERT trên bộ UIT-ViSD4SA có hiệu suất tốt hơn so với mô hình BiLSTM-CRF, còn các số màu xanh dương thể hiện hiệu suất tốt hơn sau khi tăng cường dữ liệu trên cùng mô hình phoBERT. Số liệu in đậm là hiệu suất tốt nhất của mô hình.

- **Đầu tiên, chúng tôi sẽ so sánh hiệu suất giữa mô hình BiLSTM-CRF của tác giả với mô hình mà chúng tôi đã xây dựng**

Dựa vào kết quả tổng quan ở bảng 8 kết quả f1-score macro tôt nhất của mô hình $XLM - R_{Large}$ là 62.76% trên nhận diện khía cạnh, con số tương tự ở mô hình của chúng tôi là 63.91%, hiệu suất có cải thiện 1.15%. Bên cạnh đó thì kết quả dự đoán trên nhãn cảm xúc và nhãn khía cạnh kết hợp cảm xúc cũng tăng lần lượt là 7.8% và 6.36%. Tuy có sự cải thiện đôi chút nhưng kết quả f1-macro vẫn ở mức đương đối thấp, đặc biệt nhãn aspect#polarity chỉ ở mức 52.06% có hiệu suất thấp nhất cho thấy mô hình phân loại không được mạnh mẽ.

Kết quả chi tiết được chúng tôi trình bày trong bảng 5, 6, 7 (kết quả nhãn aspect#polarity chúng tôi chỉ trình bày f1-score). Đối với task nhận nhận cảm xúc (bảng 5), mô hình phoBERT đều có sự cải thiện so với mô hình cũ với nhãn POSITIVE đạt hiệu suất cao nhất các số liệu thống kê đều trên 70%, f1-score POSITIVE đạt 73.71%. Hai nhãn còn lại thì hiệu suất khá thấp, nhãn NEUTRAL có các chỉ số thống kê đều dưới 50%. Đối với task nhận diện khía cạnh (bảng 6), vẫn có sự gia tăng hiệu suất so với trước và chỉ có các nhãn BATTERY, CAMERA, và DESIGN là có f1-score trên 70%, còn lại hiệu suất vẫn khá thấp, đặc biệt nhãn PRICE vẫn dưới 50%. Và cuối cùng là task nhận diện aspect#polarity (bảng 7), có thể xem đây là sự kết hợp của 2 task trước, mô ình của chúng tôi vẫn có sự cải thiện ở một số nhãn và tệ hơn ở các nhãn còn lại, với sự kết hợp của khía cạnh và tình cảm thì đối với những khía cạnh mang cảm xúc tích cực có hiệu suất f1-score cao hơn, đa số đều trên 60%. Tệ nhất là đối với các khía cạnh mang cảm xúc trung tính, hầu hết đều dưới rất thấp, dưới 50%. Đều này cũng giải thích tập dữ liệu UIT-ViSD4SA hiện đang mất cân bằng (nhãn NEUTRAL chỉ chiếm 6.25% toàn bộ dữ liệu).

Nhìn chung ta thấy mô hình phoBERT có sự cải thiện đôi chút về mặt hiệu suất phân loại.

- **So sánh kết quả sau khi tăng cường dữ liệu trong cùng mô hình phoBERT**

Quan sát bảng 8, sau khi tăng cường dữ liệu, về tổng quan, tất cả các phép đo đánh giá đều có sự cải thiện, f1-score macro đạt cao nhất là 67.01% ở task nhận diện khía cạnh, tăng 3.1% so với khi chưa tăng cường dữ liệu, đối với task nhân diện cảm xúc và kết hợp aspect#polarity đều tăng ít nhất 2% nâng hiệu suất vượt trên 50% mỗi loại.

Chi tiết kết quả trình bày trong các bảng trên, đối với task nhận diện cảm xúc bảng 5, f1-score đều có sự cải thiện đôi chút, cao nhất vẫn là nhãn POSITIVE đạt 76.40%, riêng nhãn NEUTRAL đã

có tăng đôi chút nhưng không nhiều, hiệu suất vẫn dưới 50%. Đối với task phân loại khía cạnh, tất cả các phép đo đều cho thấy hiệu suất có sự cải thiện, f1-score đạt cao nhất là 81.06% ở khía cạnh CAMERA, các khía cạnh còn lại cũng đạt trên 50%. Đối với task phân loại aspect#polarity bảng 7, cao nhất vẫn thuộc về các khía cạnh mang cảm xúc tích cực, hiệu suất của BATTERY và CAMERA tích cực đạt trên 80%, thấp nhất vẫn là các khía cạnh trên cảm xúc trung tính, dù có sự cải thiện nhưng hiệu suất vẫn không trên 50%.

Nhìn chung, sau khi tăng cường dữ liệu, mặc dù ở một số khía cạnh, tình cảm, hiệu suất còn thấp nhưng đã có sự cải thiện so với tập dữ liệu nguyên mẫu. Đối với kết quả tổng thể trong có vẻ tạm chấp nhận được, nhưng khi nhìn vào hiệu suất cá nhân thì khá tệ, nguyên nhân do bộ dữ liệu mất cân bằng, dù đã giải quyết phần nào sự mất cân bằng này nhưng kết quả vẫn không tăng đáng kể.

5.4 Phân tích lỗi

Câu và nhãn thực tế	Dự đoán
1. Thích camera sau$_{\text{CAMERA\#POSITIVE}}$. Sài mượt chiến game lq ngon lành$_{\text{PERFORMANCE\#POSITIVE}}$. Lưng kính nhìn sang mà dễ bám mồ hôi$_{\text{DESIGN\#NEUTRAL}}$	1. Thích camera sau: CAMERA#POSITIVE 2.Sài mượt chiến game lq ngon lành: PERFORMANCE#POSITIVE
2. Wifi bắt rất kém$_{\text{FEATURES\#NEGATIVE}}$ mọi chức năng khác thi ok$_{\text{GENERAL\#POSITIVE}}$ đứng ngay cục wifi mà luc được 4 vạch lúc 3 vạch$_{\text{FEATURES\#NEGATIVE}}$	1.Wifi bắt rất kém: FEATURES#NEGATIVE 2. mọi chức năng khác thi ok: GENERAL#POSITIVE 3. đứng ngay cục wifi mà luc được 4 vạch lúc 3 vạch: PERFORMANCE#NEGATIVE

Hình 3: Kết quả lỗi giữa giá trị thực tế và giá trị dự đoán. Nhãn dự đoán màu xanh là kết quả đúng, màu đỏ là kết quả sai

Hình 3 minh họa một số câu dự đoán từ kết quả mô hình phoBERT trên tập dữ liệu tăng cường, sau khi xem xét lại vài câu kết quả, chúng tôi nhận thấy mô hình không phát hiện ra một số cảm xúc, khía cạnh ví dụ như câu 1: *Lưng kính nhìn sang mà dễ bám mồ hôi* mang nhãn DESIGN#NEUTRAL mà kết quả dự đoán không nhận diện được. Một trường hợp khác ở câu 2 gồm 3 nhãn nhưng nhãn thứ 3 bị nhận diện sai từ FEATURES sang PERFORMENCE, tuy nhiên mô hình vẫn nhận dạng đúng cảm xúc là NEUTRAL. Đặc điểm này cần có một nghiên cứu kỹ càng vì tiếng Việt là một ngôn ngữ dùng khá nhiều từ mượn, các từ này có thể khác với nghĩa gốc và trong tập dữ liệu có khá nhiều trường hợp các bình luận vừa có tiếng Việt vừa có tiếng Anh, điều này là một thách thức lớn trong bài toán hiện tại.

6 Tổng kết

Dữ liệu không cân bằng luôn là một vấn đề lớn cần phải giải quyết trong lĩnh vực xử lý ngôn ngữ tự nhiên, và nó ảnh hưởng khá nhiều đến hiệu suất mô hình. Do đó, đồ án này tập trung vào các kỹ thuật làm giảm phân phối sai lệch trong tập dữ liệu bằng cách tăng cường dữ liệu của các lớp thiểu số và xây dựng một mô hình mới cho Task Span Detection với 67.61% độ đo f1-score macro trên tập dữ liệu tăng cường, kết quả tăng 3.1%. Chúng tôi đã triển khai các kỹ thuật EDA trên bộ dữ liệu UIT-ViSD4SA, đồng thời nghiên cứu hiệu quả của việc tăng cường dữ liệu trên bộ dữ liệu mất cân bằng. Kết quả cho thấy, khi dữ liệu về các nhãn thiểu số được tăng lên, khả năng dự đoán các nhãn đó của mô hình sẽ cao hơn. Tuy nhiên, các kỹ thuật tăng cường dữ liệu làm giảm độ chính xác của một số nhãn khác. Do đó, cần xem xét việc áp dụng các kỹ thuật tăng dữ liệu trong một bài toán cụ thể có phù hợp hay không.

Tài liệu

A. Ali, S. M. H. Shamsuddin, and A. L. Ralescu. Classification with class imbalance problem: A review. In *Soft Computing Models in Industrial and Environmental Applications*, 2015.

J. Chen, Z. Yang, and D. Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *CoRR*, abs/2004.12239, 2020.

T. Dang, V. Nguyen, K. Nguyen, and N. Ngan. A transformation method for aspect-based sentiment analysis. *Journal of Computer Science and Cybernetics*, 34:323–333, 01 2019. doi: 10.15625/1813-9663/34/4/13162.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

H.-T. Duong and V. T. Hoang. Data augmentation based on color features for limited training texture classification. In *2019 4th International Conference on Information Technology (InCIT)*, pages 208–211. IEEE, 2019.

N. T. Duyen, N. X. Bach, and T. M. Phuong. An empirical study on sentiment analysis for vietnamese. In *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, pages 309–314, 2014. doi: 10.1109/ATC.2014.7043403.

M. Fadaee, A. Bisazza, and C. Monz. Data augmentation for low-resource neural machine translation. *CoRR*, abs/1705.00440, 2017.

M. Ibrahim, M. Torki, and N. El-Makky. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 875–878, 2018. doi: 10.1109/ICMLA.2018.00141.

V. Kumar, A. Choudhary, and E. Cho. Data augmentation using pre-trained transformer models. *CoRR*, abs/2003.02245, 2020.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

S. Luu, K. Nguyen, and N. Nguyen. Empirical study of text augmentation on social media text in Vietnamese. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 462–470, Hanoi, Vietnam, Oct. 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.paclic-1.53.

D. Q. Nguyen and A. Tuan Nguyen. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.92. URL https://aclanthology.org/2020.findings-emnlp.92.

K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L.-T. Nguyen. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24, 2018. doi: 10.1109/KSE.2018.8573337.

P.-T. Nguyen, V.-L. Pham, H.-H. Vu, N.-A. Tran, H. Truong Thi Thu, and A. Sakach. A two-phase approach for building vietnamese wordnet. 01 2016.

L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621, 2017. URL http://arxiv.org/abs/1712.04621.

L. L. Phan, P. H. Pham, K. T. Nguyen, T. T. Nguyen, S. K. Huynh, L. T. Nguyen, T. V. Huynh, and K. V. Nguyen. SA2SL: from aspect-based sentiment analysis to social listening system for business intelligence. *CoRR*, abs/2105.15079, 2021.

K. N. T. Thanh, S. H. Khai, P. P. Huynh, L. P. Luc, D.-V. Nguyen, and K. N. Van. Span detection for aspect-based sentiment analysis in vietnamese. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 318–328, Shanghai, China, 11 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.paclic-1.34.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

X.-S. Vu, T. Vu, M.-V. Tran, T. Le-Cong, and H. Nguyen. Hsd shared task in vlsp campaign 2019: Hate speech detection for social good. *arXiv preprint arXiv:2007.06493*, 2020.

J. W. Wei and K. Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196, 2019. URL http://arxiv.org/abs/1901.11196.

Q. Xie, Z. Dai, E. H. Hovy, M. Luong, and Q. V. Le. Unsupervised data augmentation. *CoRR*, abs/1904.12848, 2019.