Span Detection and Explanation Generation for Aspect-based Compliment Analysis on Multi-domain Feedback

1st Pham Tien Duong
University of Information Technology
Vietnam National University
Ho Chi Minh City, Vietnam
20521222@gm.uit.edu.vn

2nd Dang Thi Thuy Hong
University of Information Technology
Vietnam National University
Ho Chi Minh City, Vietnam
20520523@gm.uit.edu.vn

3rd Nguyen Van Kiet
University of Information Technology
Vietnam National University
Ho Chi Minh City, Vietnam
kietnv@gm.uit.vn

***Tóm tắt nội dung*—Aspect-based sentiment analysis là một bài toán đầy thách thức đóng vai trò thiết yếu trong xử lý ngôn ngữ tự nhiên (NLP) và trí tuệ nhân tạo. Bài toán này tập trung vào phát hiện khía cạnh, phân loại cảm xúc và phát hiện ý kiến người dùng, có tiềm năng to lớn trong các ứng dụng thực tế. Để giải quyết bài toán này chúng tôi tiến hành nghiên cứu trên bộ dữ liệu gồm 3953 cặp giải thích và 4483 nhận xét phản hồi trên 2 domain: Fashion(Thời trang), Book(Sách). Bên cạnh đó, chúng tôi áp dụng các mô hình cho các tác vụ. Kết quả đạt được 91,13% trên F1-Score với mô hình Support Vector Machine cho tác vụ Domain Classification Task, 47,29% trên F1-Score cho tác vụ Span Detection trên miền dữ liệu Fashion và 54,78% trên Bleu1 cho tác vụ Explanation Generation. Trong tương lai, chúng tôi sẽ mở rộng bài toán để giải quyết nhiều task trong NLP như khai thác ý kiến, nhận dạng cảm xúc, phân tích khiếu nại và phát hiện mang tính xây dựng.**

***Index Terms*—Span Detection, NLP, Comment Annalysis**

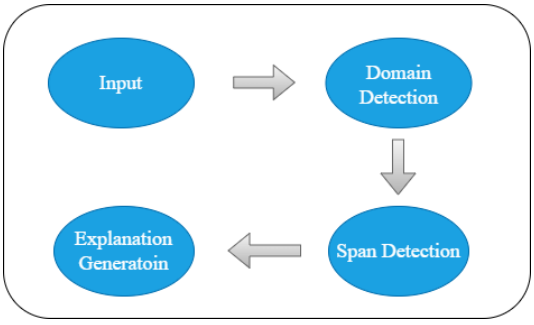
I. Introduction

Trong bối cảnh nội dung số phát triển bùng nổ như hiện nay, việc mua bán online đã không còn quá xa lạ với nhiều người. Chỉ với 1 chiếc điện thoại thông minh, chúng ta có thể mua sắm nhiều thứ tại nhà mà không cần di chuyển. Chính vì nhu cầu đó mà ngày càng có nhiều loại mặt hàng được bán trên các sàn thương mại điện tử. Đối với phái đẹp, họ thường tìm kiếm những loại mỹ phẩm có chất lượng tốt, phù hợp với da, hay những món đồ thời trang đẹp mắt khi khoác lên. Ngược lại, phái mạnh có sở thích chọn cho mình những chiếc điện thoại có bộ xử lý mạnh, có thể chơi được nhiều loại game khác nhau, màn hình to và sắc nét.

Tuy nhiên, việc mua bán online khiến khách hàng khó có thể trải nghiệm trực tiếp nên họ thường có xu hướng dựa vào phản hồi đánh giá của người mua trước. Đó như là một kênh thông tin để quyết định có nên chọn mua mặt hàng mà họ muốn hay không. Thêm vào đó vấn đề họ gặp phải là có quá nhiều đánh giá dẫn đến mất nhiều thời gian và công sức để tìm kiếm và phân tích. Chưa kể người dùng thường có xu hướng sử dụng teencode, viết tắt hay dùng những ẩn ý gây khó khăn cho người đọc đặc biệt là người lớn tuổi hoặc những người ngoại quốc. Vì vậy chúng tôi quyết định thực hiện đề tài nghiên cứu **"Span Detection and Explanation Generation for Aspect-based Compliment Analysis on Multi-domain Feedback"**.

Trong đề tài này, chúng tôi tiến hành xây dựng bộ dữ liệu UIT-ViSE4AM gồm 4483 nhận xét phản hồi trên 2 domain: **Fashion(Thời trang), Book(Sách)** và 3953 cặp giải thích. Bên cạnh đó, chúng tôi tập trung áp dụng Machine Learning và Transfer Learning để giải quyết các tác vụ Domain Classification, Span Detection và Explanation Generation:

- **Task 1: Domain Classification** là quá trình phân loại domain, bởi mỗi domain có mỗi bộ aspect khác nhau, do đó cần phải tiến hành phân loại domain của bình luận đánh giá đầu vào để có thể thực hiện các task sau.
- **Task 2: Span Detection** là quá trình gán nhãn chỉ ra các Aspect(Khía cạnh), Sentiment(Cảm xúc) và vị trí của các nhãn này trong bình luận.
- **Task 3: Explanation Generation** nhằm mục đích giải thích các nhãn khía cạnh cảm xúc và đoạn văn được chỉ ra ở Task 2 bởi vì phản hồi của khách hàng thường viết tắt, sử dụng teencode gây nhầm lẫn, khó hiểu.



Hình 1: Mô tả bài toán

II. Related Work

Bài toán Aspect-Based Sentiment Analysis (ABSA) là một lĩnh vực nghiên cứu được quan tâm nhiều trong xử lý ngôn ngữ tự nhiên, đa có nhiều bộ dữ liệu chất lượng được công bô trong các cuộc thi học thuật uy tín. Có thể kể đến, cuộc

thi SemEval đã công bố nhiều bộ dữ liệu bao gồm đánh giá của người dùng từ các trang web thương mại điện tử, tạo nền tảng cho nhiều nghiên cứu liên quan. Trong đó, bộ dư liệu SE-ABSA14 ở Task 4 [1]của SemEval 2014 đa được công bố bộ bao gồm các đánh giá về nhà hàng và máy tính xách tay, đánh giá về nhà hàng bao gồm năm khía cạnh (Food, Service, Price, Ambience and Anecdotes/Miscellaneous) và bốn nhãn cảm xúc (Positive, Negative, Conflict and Neutral). Tiếp đó, bộ dữ liệu SE-ABSA15 ở Task12 của SemEval 2015 [2]được xây dựng dựa trên SE-ABSA14 mô tả khía cạnh và thuộc tính của khía cạnh đó (ví dụ: FoodStyle). Đến bộ bộ dữ liệu SEABSA16 ở Task5 của SemEval 2016 [3] đã mở rộng SE-ABSA15 thành các miền dữ liệu mới như Khách sạn, Điện tử tiêu dùng, Viễn thông, Bảo tàng ở các ngôn ngữ khác như tiếng Hà Lan, Tiếng Pháp, tiếng Nga, tiếng Tây Ban Nha, tiếng Thổ Nhĩ Kỳ và tiếng Ả Rập.

Cùng với đó, hiện tại, lĩnh vực xử lý ngôn ngữ tự nhiên tại Việt Nam đang ngày càng phát triển với các bộ dữ liệu chất lượng. Một trong số đó có những bộ dữ liệu về ABSA trên tiếng Việt như là UIT-ViSFD [4] gồm 11.122 câu đánh giá trên miền dữ liệu điện thoại thông minh được gán nhãn với kết quả F1 84,48% cho việc phân tích khía cạnh và 63,06% cho việc phân tích cảm xúc trên các mô hình máy học và học sâu. Bên cạnh đó, có thể kể đến bộ dataset UIT-ViSD4SA [5] bao gồm 35,396 nhãn span gán bởi con người trên 11,122 đánh giá Tiếng Việt trên miền dữ liệu điện thoại thông minh với kết quả F1 62,76% trên F1 macro khi sử dụng BiLSTM-CRF cho tác vụ span detetion.

III. Dataset Creation

A. Data Collection

Chúng tôi tiến hành thu thập các phản hồi đánh giá của khách hàng trên các miền dữ liệu bao gồm **Fashion(Thời trang), Book(Sách)** từ trang thương mại điện tử Shopee. Đây là nguồn dữ liệu vô cùng đa dạng và chất lượng bởi nó chứa số lượng đánh giá từ rất nhiều người dùng trong nước. Các bước thu thập và tiền xử lý dữ liệu sẽ được trình bày chi tiết dưới đây:

Đầu tiên, chúng tôi tiến hành liệt kê và thu thập đánh giá từng sản phẩm của từng miền dữ liệu trên sàn thương mại điện tử Shopee. Với mỗi miền dữ liệu chúng tôi chọn những sản phẩm theo hai tiêu chí. Thứ nhất sản phẩm có lượng mua và lượng đánh giá cao. Thứ hai những sản phẩm phải đa dạng mẫu mã, loại sản phẩm, độ tuổi, giới tính,... Dữ liệu thô thu thập được bao gồm nội dung của đánh giá, mức độ đánh giá sản phẩm trên thang điểm từ 1 đến 5 và đường dẫn tới trang mua sản phẩm.

Tiếp theo, chúng tôi tiến hành xử lý dữ liệu thô bằng cách loại bỏ cột chứa đường dẫn của sản phẩm vì cột này không cần thiết trong quá trình huấn luyện. Cùng với đó là loại bỏ những dữ liệu trùng, những đánh giá spam và sắp xếp dữ liệu theo mức độ đánh giá sản phẩm.

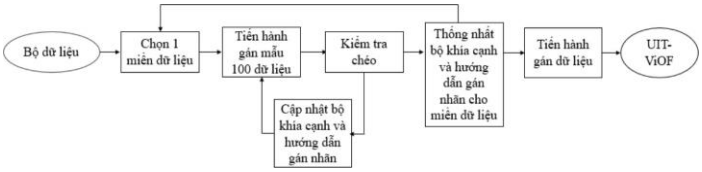
Sau khi tiền xử lý dữ liệu, chúng tôi thu được số lượng dữ liệu được khá lớn, từ 10,000 cho đến 20,000 bình luận ở mỗi miền dữ liệu, chúng tôi đã tiến hành chọn lọc dữ liệu của mỗi

Hình 2: Quy trình thu thập và tiền xử lý dữ liệu

miền dữ liệu nhằm chọn ra những điểm dữ liệu chất lượng, không trùng lặp ý.

B. Annotation Process

Sau khi thu thập dữ liệu, chúng tôi tiến hành gán nhãn bộ dữ liệu gồm 5000 bình luận trên 2 miền dữ liệu khác nhau trên 3 khía cạnh cảm xúc là COMPLIMENT, COMPLAINT, NEUTRAL. Đoạn bình luận được gán nhãn được định nghĩa là đoạn bình luận ngắn nhất chỉ rõ ý kiến của người dùng về khía cạnh đó. Chi tiết được miêu tả dưới đây.



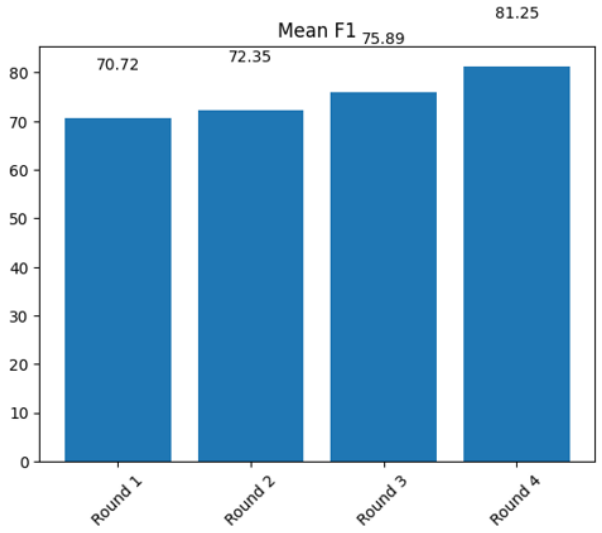
Hình 3: Quy trình gán nhãn dữ liệu

Đầu tiên, chúng tôi chọn ra một miền của bộ dữ liệu và 100 bình luận từ đó gán độc lập và kiểm tra chéo, điều này nhằm mục đích tìm ra các khía cạnh khác của bình luận mà chúng tôi chưa đề cập tới hoặc cần chỉnh sửa sao cho phù hợp với bài toán. Tiếp đó chúng tôi lại tiếp tục chọn ra 100 bình luận cho đến khi thống nhất được bộ khía cạnh hoàn chỉnh nhất cho mỗi miền dữ liệu

Sau đó, chúng tôi tiến hành gán chung 75 dữ liệu và tính độ đồng thuận. Sau 4 round chúng tôi đã đạt được kết quả là 81,25% trên F1-Score. Sau cùng, chúng tôi đã trực quan số lượng nhãn của mỗi miền dữ liệu ở hình bên dưới.

Bảng I: Bảng chi tiết bộ khía cạnh cho từng miền dữ liệu.

Miền dữ liệu	Tập khía cạnh
Book	*Quality (Chất lượng), Service (Dịch vụ), Packing (Đóng gói), Delivery (Vận chuyển), Price (Giá thành), Advice(Lời khuyên), Feeling(Cảm nhận), Promotion (Khuyến mãi), Design (Thiết kế), Content (Nội dung), Other (các khía cạnh khác).*
Fashion	*Quality (Chất lượng), Service (Dịch vụ), Packing (Đóng gói), Delivery (Vận chuyển), Price (Giá thành), Advice (Lời khuyên), Feeling (Cảm nhận), Promotion (Khuyến mãi), Design (Thiết kế), Other (các khía cạnh khác).*



Hình 4: Độ đồng thuận được đánh giá trên F1-Score.

IV. Our Approach

Trong phần này, chúng tôi giới thiệu các mô hình được sử dụng cho mỗi tác vụ.

A. Task 1: Domain Classification

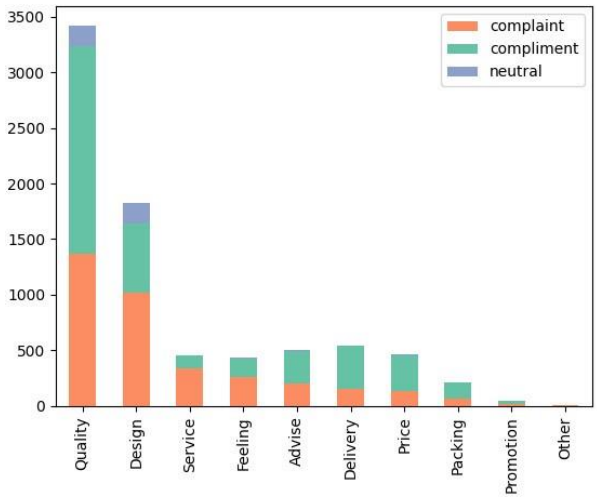
LOGISTIC REGRESSION (LR) [6] là một thuật toán phân loại có thể được áp dụng cho nhiệm vụ phân loại văn bản bởi vì nó thể hiện mối quan hệ giữa các biến nhị phân phụ thuộc và biến độc lập.

SUPPORT VECTOR MACHINE (SVM) [7] được sử dụng trong nghiên cứu này vì mô hình cung cấp một phương pháp ưu việt để phân loại văn bản bằng hàm nhân. Trong tập dữ liệu có một số nhận xét đánh giá chứa các từ khá phức tạp khó phân biệt đâu là miền dữ liệu nên thuật toán SVM trong trường hợp này có thể cho kết quả mô phỏng đoán tốt hơn.

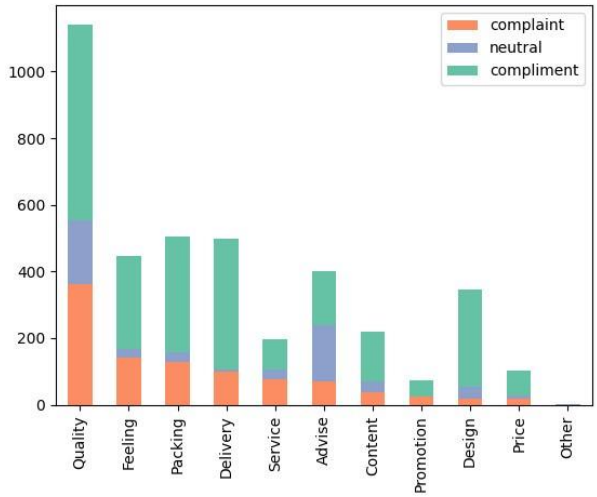
NAIVE BAYES [8] được ứng dụng rộng rãi để phân loại văn bản, trong nghiên cứu này, mô hình được sử dụng để phân loại miền dữ liệu của các bình luận đánh giá.

DECISION TREE (DT) [9]là mô hình phân loại biểu diễn kết quả khả thi bằng phương pháp phân nhánh, mô hình này đạt được hiệu suất khá cao trong việc phân loại đưa ra quyết định.

RANDOM FOREST [10] một thuật toán học có giám sát được sử dụng cho các tác vụ phân loại, thuật toán xây dựng nhiều cây quyết định dựa trên phép đo độ tương đồng giữa các mẫu. Khi cho trước một bộ dữ liệu mới, thuật toán sử dụng mỗi cây để dự đoán đầu ra và trung bình giá trị của các dự đoán này được lấy làm giá trị đầu ra cuối cùng



Hình 5: Thống kê số lượng nhãn miền dữ liệu Fashion.



Hình 6: Thống kê số lượng nhãn miền dữ liệu Book.

B. Task 2: Span Detection

PhoBERT [11] là một mô hình monolingual language được được pre-train trên bộ dữ liệu Tiếng Việt 20 GB và có cấu trúc giống như RoBERTa. PhoBERT là một phương pháp tiên tiến nhất trong nhiều tác vụ NLP dành riêng Tiếng Việt như Part-Of-Speech Tagging, Dependency Parsing, and NER.

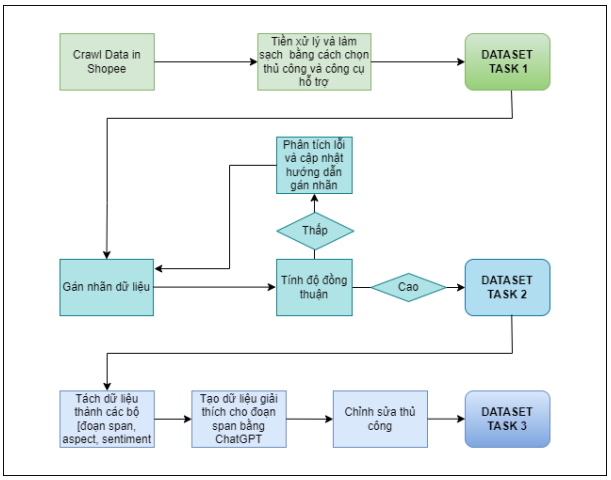
XLM-ROBERTA [12] là một mô hình multilingual language và là một biến thể của RoBERTa, được pre-train trên 2,5T dữ liệu trên 100 ngôn ngữ bao gồm 137GB là văn bản bằng Tiếng Việt.

BERT [13] là một model pre-train biểu diễn ngôn ngữ (Language Model- LM) được google giới thiệu vào năm 2018. BERT là một mô hình đạt sự đột phá lớn trong Machine Learning bởi vì khả năng ứng dụng của nó vào nhiều bài toán

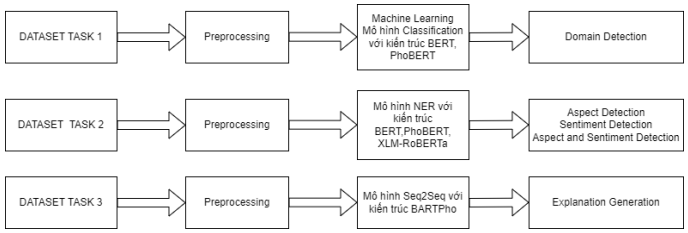
NLP khác nhau như Question Answering, Natural Language Inference,. . . với kết quả rất tốt. BERT là một mô hình pre-train, học ra các vector đại diện theo ngữ cảnh 2 chiều của từ.

C. Task 3: Explaination Generation

BARTPho là một biến thể của mô hình BART (Bidirectional and Auto-Regressive Transformers) được điều chỉnh và tối ưu hóa đặc biệt cho các bài toán sequence-to-sequence. BARTPho sử dụng kiến trúc transformer và đã được huấn luyện trên một lượng lớn dữ liệu để có khả năng tạo ra các đầu ra tự nhiên và chất lượng.



Hình 7: Tổng quan phương pháp tiếp cận.



Hình 8: Mô tả phương pháp tiếp cận từng tác vụ.

V. Experiments and Results

A. Experiment Settings

Theo định dạng IOB (Inside, Outside, Beginning), tập dữ liệu của chúng tôi được chuyển đổi với dữ liệu chứa các nhãn khía cạnh (QUALITY, SERVICE, PACKING,v.v.), các nhãn cảm tính (COMPLIMENT, COMPLAIN, NEUTRAL) và dữ liệu chứa cả nhãn khía cạnh và cảm tính (QUALITY#POSITIVE, SERVICE#NEGATIVE,vân vân) để đánh giá các mô hình một cách toàn diện.

B. Evaluation Metrics

Chúng tôi sử dụng 3 độ đo để đánh giá là Precision, Recall và F1-Score. Nhãn dự đoán được đánh giá là đúng chỉ khi trùng khớp với nhãn tiêu chuẩn. Chúng tôi tiến hành đánh giá trên cả hai miền dữ liệu, trên nhãn khía cạnh, nhãn cảm xúc và nhãn khía cạnh cảm xúc.

C. Experimental Results

Chúng tôi đánh giá các mô hình máy học cho cả hai phương pháp tokenize trong Bảng 2. Kết quả cho thấy Support Vector Machine đạt kết quả tốt nhất với phương pháp TF-IDF. Trong khi đó, với phương pháp COUNTER VECTO, Logistic Regression đạt kết quả cao nhất so với phương pháp học máy truyền thống khác. Nhìn chung, ở cả 2 phương pháp thì TF-IDF đạt kết quả khả quan hơn so với COUNTER VECTO.

Ở Bảng 3, so sánh tất cả các mô hình Machine Learning và Transfer Learning, mô hình PhoBERT đạt được kết quả tốt nhất với 91,59% trên Precision trên tập dữ liệu. Giải thích cho điều này là bởi vì mô hình PhoBERT là một mô hình pre-train trên ngôn ngữ Tiếng Việt, do đó khi huấn luyện mô hình này với tập dữ liệu chúng tôi thu được kết quả khá khả quan.

Ở bảng 4, mô hình Seq2Seq Transfomers đạt được mức độ tương đồng tương đối tốt với các câu tham chiếu trong tập dữ liệu. Điểm BLEU-1 đạt 54.78% cho thấy mô hình có khả năng tạo ra các từ hoặc cụm từ phù hợp với các câu tham chiếu. Tuy nhiên, điểm BLEU score giảm dần từ BLEU-1 đến BLEU-4, cho thấy mô hình chưa tạo ra các câu dài và phức tạp một cách chính xác.

Ở bảng 5, so sánh tất cả kết quả của các mô hình pretrain, chúng tôi nhận thấy kiến trúc XLM-RoBERTa đạt F1 - Score tốt nhất cho tất cả các tác vụ và khía cạnh. Chỉ riêng với tác vụ nhận diện Aspect và Domain trên miền Fashion của BERT đạt kết quả cao hơn phần còn lại(47,29%)

Cuối cùng là bảng mình họa Input, Output của các Task trong đề tài này.

VI. Conclusion and Future Work

Trong nghiên cứu này, chúng tôi đã tiến hành xây dựng bộ dữ liệu UIT-ViSE4AM gồm 4483 nhận xét phản hồi trên 2 domain: Fashion(Thời trang), Book(Sách) và 3953 cặp giải thích, mỗi bình luận được gán nhãn với 3 nhãn cảm xúc. Cùng với đó, chúng tôi áp dụng các mô hình Machine Learning và Transfer Learning cho các tác vụ. Kết quả đạt được 91,48\% trên F1-Score cho mô hình Support Vector Machine với phương pháp TF-IDF ở Domain Classification Task, 47,29\% trên F1-Score ở Span Detection Task trên miền dữ liệu Fashion và 54,78\% trên Bleu-1 cho tác vụ Explanation Generation. Nhìn chung, hiệu suất đạt được tương đối khả quan ở cả 3 tác vụ.

Trong tương lai, chúng tôi đưa ra một số hướng phát triển như sau: (1) Mở rộng thêm bộ dữ liệu và tập khía cạnh trên nhiều miền dữ liệu khác. (2) Áp dụng thêm các pre-train model đa ngôn ngữ để tăng độ chính xác cho các tác vụ. (3) Cải thiện hiệu suất của nghiên cứu này để có thể áp dụng vào nhiều mảng nghiên cứu khác như phát triển bài toán Machine Comprehension Reading cho Tiếng Việt, khai thác ý kiến và phân tích khiếu nại người dùng.

Bảng II: Đánh giá F1-Score trên các mô hình Machine Learning của task Domain Classification.

MODELS	TF IDF			COUNTER VECTO		
	Book	Fashion	Average	Book	Fashion	Average
SVM	**90.78%**	**91.48%**	**91.13%**	89.43%	90.42%	89.92%
NAIVE BAYES	75.77%	62.39%	69.08%	74.85%	57.50%	66.18%
DECISION TREE	87.62%	87.68%	87.65%	88.57%	88.61%	88.59%
RANDOM FOREST	90.34%	90.57%	90.46%	89.70%	90.03%	89.86%
LOGISTIC REGRESSION	90.75%	91.34%	91.05%	**90.43%**	**91.15%**	**90.79%**

Bảng III: Thử nghiệm các mô hình trên tập dữ liệu

MODELS		Precision	Recall	F1-Score
MACHINE LEARNING	SVM	91.40%	91.14%	91.13%
	NAIVE BAYES	75.26%	70.53%	69.08%
	DECISION TREE	88.59%	88.59%	88.59%
	RANDOM FOREST	90.34%	90.29%	90.29%
	LOGISTIC REGRESSION	91.24%	91.06%	91.05%
TRANSFER LEARNING	PHOBERT	**91.59%**	**91.57%**	**91.57%**
	BERT	90.75%	90.72%	90.72%

Bảng IV: Đánh giá mô hình Seq2Seq Transfomers sử dụng kiến trúc BARTPho.

Bleu 1	Bleu 2	Bleu 3	Bleu 4
54,78%	50,80%	47,53%	44,59%

Bảng V: Đánh giá F1-Score trên các mô hình ở tác vụ Span Detection

Mô hình	Aspect		Sentiment		Aspect#Sentiment	
	Fashion	Book	Fashion	Book	Fashion	Book
PhoBERT	46,61%	41,83%	50,88%	45,44%	42,72%	32,86%
XLM-ROBERTa	**50,34%**	**48,08%**	**52,21%**	**48,78%**	46,36%	**40,46%**
BERT	46,30%	42,59%	48,21%	42,20%	**47,29%**	34,43%

TÀI LIỆU

[1] D. Kirange, R. R. Deshmukh, and M. Kirange, "Aspect based sentiment analysis semeval-2014 task 4," *Asian Journal of Computer Science and Information Technology (AJCSIT) Vol*, vol. 4, 2014.

[2] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "Semeval-2015 task 12: Aspect based sentiment analysis," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 486–495.

[3] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis," in *ProWorkshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, 2016, pp. 19–30.

[4] B. H. Le, H. M. Nguyen, N. K.-P. Nguyen, and B. T. Nguyen, "A new approach for vietnamese aspect-based sentiment analysis," in *2022 14th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2022, pp. 1–6.

[5] K. T.-T. Nguyen, S. K. Huynh, L. L. Phan, P. H. Pham, D.-V. Nguyen, and K. Van Nguyen, "Span detection for aspect-based sentiment analysis in vietnamese," *arXiv preprint arXiv:2110.07833*, 2021.

[6] G. Ifrim, G. Bakir, and G. Weikum, "Fast logistic regression for text categorization with variable-length n-grams," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 354–362.

[7] E. Leopold and J. Kindermann, "Text categorization with support vector machines. how to represent texts in input space?" *Machine Learning*, vol. 46, no. 1-3, pp. 423–444, 2002.

[8] H. Zhao and F. Kamareddine, "Advance gender prediction tool of first names and its use in analysing gender disparity in computer science in the uk, malaysia and china," in *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2017, pp. 222–227.

[9] M. A. Farooqui and D. Sheetlani, "Different classification technique for data mining in insurance industry using weka," *IOSR J. Comput. Eng.*, vol. 19, no. 01, pp. 11–18, 2017.

[10] S. Maruf, K. Javed, and H. A. Babri, "Improving text classification performance with random forests-based feature selection," *Arabian Journal for Science and Engineering*, vol. 41, pp. 951–964, 2016.

[11] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese," *arXiv preprint arXiv:2003.00744*, 2020.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.