ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Ngày nhận hồ sơ	
(Do CQ quản lý ghi)	

THUYẾT MINH
ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP SINH VIÊN 2022

A. THÔNG TIN CHUNG

A1. Tên đề tài

- Tên tiếng Việt: PHÂN TÍCH THÁI ĐỘ KHÁCH HÀNG THEO TỪNG KHÍA CẠNH QUA CÁC PHẢN HỒI TIẾNG VIỆT TRÊN CÁC TRANG THƯƠNG MẠI ĐIỆN TỬ MIỀN MỞ

- Tên tiếng Anh: ASPECT-BASED COMPLIMENT ANALYSIS OF VIETNAMESE OPEN-DOMAIN FEEDBACK ON E-COMMERCE SITES

A2. Loại hình nghiên cứu

(Tham khảo tiêu chuẩn đề tài đối với từng loại hình NC, chọn 01 trong 03 loại hình)

☒ Nghiên cứu cơ bản

□ Nghiên cứu ứng dụng

□ Nghiên cứu triển khai

A3. Thời gian thực hiện

06 tháng (kể từ khi được duyệt).

A4. Tổng kinh phí

(Lưu ý tính nhất quán giữa mục này và mục B8. Tổng hợp kinh phí đề nghị cấp)

Tổng kinh phí: **6** triệu đồng, gồm

- Kinh phí từ Trường Đại học Công nghệ Thông tin: **6** triệu đồng

A5. Chủ nhiệm

Họ và tên: **Phạm Tiến Dương**

Ngày, tháng, năm sinh: 11/05/2002; Giới tính (Nam/Nữ): Nam.

Số CCCD: 974202006690; Ngày cấp: 10/05/2021; Nơi cấp: Cục Cảnh sát quản lý hành chính về trật tự xã hội.

Mã số sinh viên: 20521222**.**

Số điện thoại liên lạc: 0357421330**.**

Đơn vị (Khoa hoặc BM): KH&KTTT**.**

A6. Nhân lực nghiên cứu

TT	Họ tên	MSSV	Khoa/ Bộ Môn
1	Phạm Tiến Dương	20521222	KH&KTTT
2	Trần Huỳnh Quốc An	20520955	KH&KTTT
3	Đặng Thị Thúy Hồng	20520523	KH&KTTT

B. MÔ TẢ NGHIÊN CỨU

B1. Giới thiệu về đề tài

Giới thiệu chung

Trong bối cảnh nội dung số phát triển 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. Tùy vào nhu cầu và mục đích mà người dùng có những tiêu chuẩn để phân tích các phản hồi đánh giá cũng khác nhau. Vì vậy chúng tôi quyết định thực hiện đề tài nghiên cứu "Phân tích thái độ của khách hàng theo từng khía cạnh qua các phản hồi tiếng việt trên các trang thương mại điện tử miền mở". Để giải quyết vấn đề này chúng tôi tiến hành nghiên cứu trên bộ dữ liệu UIT-ViOCD [5] gồm 5.485 đánh giá trên 4 miền dữ liệu: ***Thời trang, mỹ phẩm, phần mềm ứng dụng, điện thoại thông minh***.

B1.1. Tình hình nghiên cứu trên thế giới

Bài toán Aspect-Based Sentiment Analysis (ABSA) là một mảng lĩnh vực nghiên cứu trong Sentiment Analysis và được quan tâm nhiều trong lĩnh vực xử lý ngôn ngữ tự nhiên. Nhiều bộ dữ liệu về ABSA đã được giới thiệu trong các cuộc thi học thuật uy tín như: SemEval 2014 [1], SemEval 2015 [2] và SemEval 2016 [3]. Các bộ dữ liệu từ các cuộc thi chủ yếu trên ngôn ngữ tiếng Anh, tiếng Trung, tiếng Pháp,...

B1.2. Tình hình nghiên cứu tại Việt Nam

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 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.

B1.3. Thách thức

Phần lớn các tài liệu liên quan đến bài toán Aspect-Based Sentiment Analysis (ABSA) đều được viết bằng tiếng Anh nên trở ngại về mặt ngôn ngữ. Bộ dữ liệu của chúng tôi được thu thập từ các đánh giá mang xu hướng là ngôn ngữ mạng xã hội, không đúng cấu trúc ngữ pháp của tiếng Việt, chứa nhiều teencode và biểu tượng cảm xúc. Bên cạnh đó, ngôn ngữ trong câu không thống nhất, có nhiều loại ngôn ngữ được viết đan xen nhau. Gần đây, giới trẻ thường hay sử dụng

những từ ngữ bị biến đổi đi ít nhiều so với ngôn ngữ chuẩn. Với sự biến đối này khiến cho đội ngũ gán nhãn gặp nhiều khó khăn trong việc đọc hiểu nội dung của các phản hồi.

Hiện nay, mặc dù các cộng đồng xử lý ngôn ngữ tự nhiên trên thế giới đã có rất nhiều công trình nghiên cứu với kết quả nổi bật có thể tham khảo và học hỏi, nhưng về tiếng Việt thì còn hạn chế. Do đó, chúng tôi kế thừa các công trình từ ngôn ngữ khác để học hỏi các phương pháp xây dựng bài toán phù hợp cho tiếng Việt. Vì vậy, trong đề tài nghiên cứu này, chúng tôi mong muốn xây dựng bộ dữ liệu đủ lớn và chất lượng trên nhiều miền dữ liệu để phục vụ cho bài toán Phân tích thái độ của khách hàng theo từng khía cạnh qua các phản hồi tiếng việt trên các trang thương mại điện tử miền mở.

B2. Mục tiêu, nội dung, kế hoạch nghiên cứu

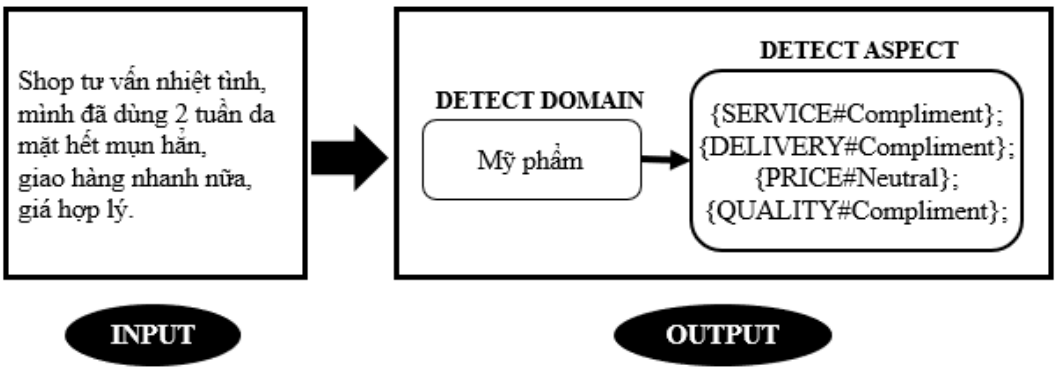
B2.1 Mục tiêu

Chúng tôi kế thừa bộ dữ liệu UIT-ViOCD [5] với 5.485 bình luận gồm 4 lĩnh vực: Cosmetic, App, Mobile, Fashion,và xem xét mở rộng bộ dữ liệu thêm 3 lĩnh vực khoảng 10.000 bình luận. Sau đó, hướng dẫn gán nhãn, huấn luyện gán nhãn và thực hiện quá trình gán nhãn đồng thời thực hiện kiểm tra chất lượng dữ liệu. Tiếp theo, chúng tôi tiến hành nghiên cứu thực nghiệm các mô hình máy học truyền thống, học sâu hiện đại và học chuyển tiếp. Cuối cùng, kết quả thu thập được sẽ đem so sánh và đánh giá hiệu suất của các mô hình.

B2.2 Nội dung và phương pháp nghiên cứu

Nội dung 1: Xác định bài toán

- Bài toán của chúng tôi gồm 3 nhiệm vụ chính:
  - + Nhận diện miền dữ liệu (Domain Detection).
  - + Nhận diện khía cạnh (Aspect Detection).
  - + Phân tích thái độ (Compliment Analysis).
- Đầu vào: Đánh giá của khách hàng về sản phẩm.
- Đầu ra:
  - + Lĩnh vực của sản phẩm, bao gồm 7 lĩnh vực đã được đề cập: **Cosmetic** (mỹ phẩm), **Fashion** (thời trang), **App** (phần mềm ứng dụng), **Mobile** (điện thoại di động), **Book** (sách), **Toy** (đồ chơi) và **Food** (đồ ăn vặt).
  - + Các khía cạnh mà bình luận đề cập: Mỗi miền dữ liệu sẽ có tập khía cạnh khác nhau.
  - + Thái độ của khía cạnh: **Compliment** (khen), **Complaint** (chê) và **Neutral** (trung tính).



Hình 1: Mô tả đầu vào và đầu ra của bài toán.

Nội dung 2: Xây dựng và đánh giá bộ dữ liệu

Định nghĩa nhãn của dữ liệu

- **Compliment (Khen):** Là những đánh giá của khách hàng mang thái độ tích cực, nhằm bày tỏ sự hài lòng với chất lượng, dịch vụ của sản phẩm.

- **Complaint (Chê):** Là những đánh giá của khách hàng mang thái độ tiêu cực, nhằm bày tỏ sự không vừa ý, thậm chí bức xúc với chất lượng, dịch vụ của sản phẩm.

- **Neutral (Trung tính):** Là những đánh giá không khen cũng không chê; có thể có ý vừa khen vừa chê nhưng vẫn xét vào trung tính (những trường hợp này sẽ được phân tích thêm trong những nghiên cứu tiếp theo về nhận diện cảm xúc theo đoạn).

Tập khía cạnh của mỗi lĩnh vực

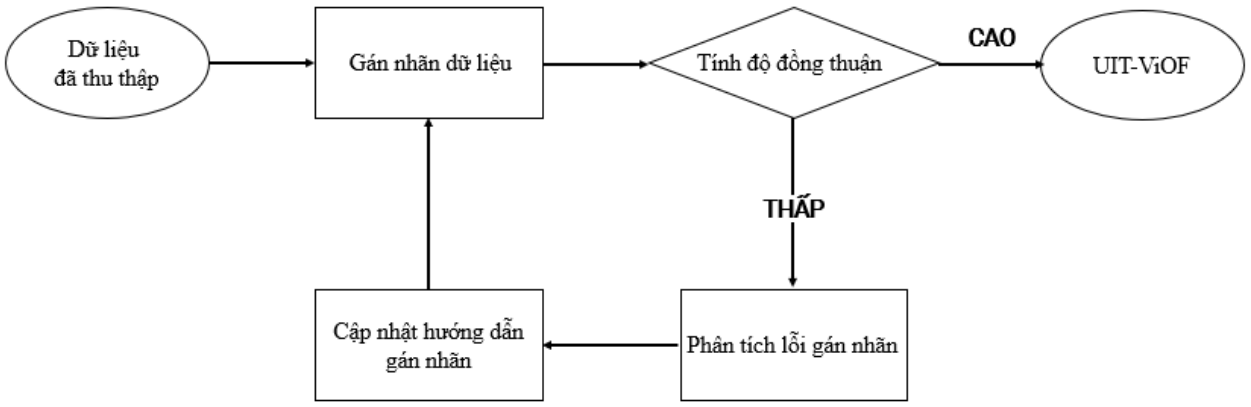
- **Fashion:** Service (dịch vụ), Packing (đóng gói), Price (giá thành), Delivery (vận chuyển), Quality (chất lượng), Advise (lời khuyên), Model (mẫu mã sản phẩm), Texture (kết cấu), Colour (màu sắc) và Material (chất liệu).
- **Cosmetic:** Service (dịch vụ), Packing (đóng gói), Price (giá thành), Delivery (vận chuyển), Quality (chất lượng), Advise (lời khuyên), Model (mẫu mã sản phẩm), Texture (kết cấu), Colour (màu sắc) và Promotion (khuyến mãi).
- **App:** Service (dịch vụ), Price (giá thành), Quality (chất lượng), Advise (lời khuyên), Graphic (đồ họa), Delay (độ trễ), Advertising (quảng cáo), Capacity (dung lượng) và Download(sự tải xuống).
- **Mobile:** Service (dịch vụ), Packing (đóng gói), Price (giá thành), Delivery (vận chuyển), Quality (chất lượng), Advise (lời khuyên), Model (mẫu mã sản phẩm) và Configuration (cấu hình).
- **Toy:** Service (dịch vụ), Packing (đóng gói), Price (giá thành), Delivery (vận chuyển), Quality (chất lượng), Advise (lời khuyên), Model (mẫu mã sản phẩm), Texture (kết cấu) và Colour (màu sắc).
- **Food:** Service (dịch vụ), Packing (đóng gói), Price (giá thành), Delivery (vận chuyển), Quality (chất lượng), Advise (lời khuyên) và Flavor (mùi vị).
- **Book:** Service (dịch vụ), Packing (đóng gói), Price (giá thành), Delivery (vận chuyển), Quality (chất lượng), Advise (lời khuyên), Model (mẫu mã sản phẩm), Texture (kết cấu), Colour (màu sắc) và Content (nội dung).

Bảng 1: Ví dụ minh họa về dữ liệu của mỗi lĩnh vực.

STT	Bình luận	Miền dữ liệu	Khía cạnh và thái độ của khía cạnh
1	Áo đẹp xịn xò lắm, có thư cảm ơn nữa. Nên mua nha mn	Fashion	{SERVICE#Compliment}; {QUALITY#Compliment}; {ADVISE#Compliment};
2	Son xinh lắm ạ, giao hàng hơi lâu vì dịch nên không sao ạ, lần sau sẽ ủng hộ shop nhiều hơn… Hình ảnh chỉ mang tính chất nhận xu thôi hihihiih	Cosmetic	{DELIVERY#Neutral}; {QUALITY#Compliment}; {COLOUR#Compliment};
3	Nhiều quảng cáo bực cả mình	App	{ADVERTISING#Complaint};
4	Máy tốt, loa hơi nhỏ tí	Mobile	{QUALITY#Compliment}; {CONFIGURATION#Complaint};
5	Đồ chơi đẹp, cháu nhà mình thích lắm	Toy	{QUALITY#Compliment};
6	Thơm, đóng gói cẩn thận, cho shop 5 sao	Food	{FLAVOR#Compliment}; {PACKING#Compliment};
7	Sách hay nhưng giấy mỏng dễ rách	Book	{CONTENT#Compliment}; {QUALITY#Complaint};

Quy trình xây dựng:

Chúng tôi dự kiến thu thập dữ liệu từ các đánh giá của người mua hàng trên sàn thương mại điện tử nổi tiếng Shopee. Sau khi thu thập dữ liệu, chúng tôi thực hiện quá trình gán nhãn dữ liệu. Với quy trình này, chúng tôi sẽ cung cấp và cập nhật bản hướng dẫn gán nhãn đến khi đội ngũ gán nhãn nắm rõ bài toán và đảm bảo độ đồng thuận giữa các thành viên khi gán nhãn.



Hình 2: Quy trình xây dựng bộ dữ liệu.

Nội dung 3: Các phương pháp tiếp cận

Sau khi xây dựng bộ dữ liệu, chúng tôi tiến hành nghiên cứu với các phương pháp máy học truyền thống, học sâu, học chuyển tiếp, tìm ra mô hình hiệu quả nhất cho bài toán. Dưới đây là những phương pháp chúng tôi sẽ sử dụng trong bài toán. Cả ba phương pháp đều phổ biến trong nhiệm vụ phân tích cảm xúc.

Logistic Regression: Đây là mô hình máy học truyền thống và cơ bản. Khác với Linear Regression, mặc dù trong tên có chứa Regression, nhưng Logistic Regression thường được sử dụng phổ biến trong các bài toán phân lớp, đặc biệt là bài toán phân lớp nhị phân. Đầu ra của phương này có thể được thể hiện dưới dạng xác suất.

Bidirectional Long Short Term Memory (Bi-LSTM): Đây là mô hình học sâu, với Bi-LSTM ta có thể lọc ra thông tin không cần thiết cũng như chọn lựa và lưu trữ các thông tin quan trọng. Chúng tôi, sử dụng Bi-LSTM cùng với bộ biểu diễn từ fastTest. Vì dữ liệu đầu vào là bình luận thuộc ngôn ngữ xã hội nên chúng tôi chọn fastTest bởi khả năng huấn luyện nhanh và có thể mã hóa các từ hiếm gặp hoặc các từ không xuất hiện trong quá trình huấn luyện.

***Bidirectional Encoder Representations from Transformers (PhoBERT) - PhoBERT:*:** BERT được thiết kế để có thể huấn luyện trước các biểu diễn từ (bộ biểu diễn từ pre-trained). Điểm nổi bật của BERT là nó có thể điều chỉnh cân bằng ngữ cảnh theo cả 2 chiều trái và phải. Và nó có thể giải quyết vấn đề thiếu hụt dữ liệu bằng cách transfer từ một mô hình chung được đào tạo từ một lượng lớn các dữ liệu không được gán nhãn. PhoBERT là mô hình BERT đã được VinAI nghiên cứu trên ngôn ngữ tiếng Việt. PhoBERT sử dụng RDRSegmenter của VNCoreNLP để tách các từ cho dữ liệu vào trước khi qua BPE encoder.

Nội dung 4: Kế hoạch thực hiện

Bảng 2: Kế hoạch nghiên cứu trong 6 tháng

Công việc	T1	T2	T3	T4	T5	T6
Nghiên cứu tổng quan bài toán + Nghiên cứu các phương pháp máy học, học sâu. + Nghiên cứu đặc điểm về xử lý ngôn ngữ tự nhiên trên tiếng Việt. + Nghiên cứu các công cụ hỗ trợ cho quá trình nghiên cứu.						
Xây dựng bộ dữ liệu + Thu thập dữ liệu + Gán nhãn dữ liệu. + Kiểm tra và đánh giá dữ liệu.						
Nghiên cứu phương pháp thực nghiệm + Nghiên cứu thử nghiệm các phương pháp học máy, học sâu và học chuyển tiếp trên bộ dữ liệu đã xây dựng. + Đánh giá các kết quả thử nghiệm thu thập được.						
Báo cáo tiến độ + Viết báo cáo quá trình và kết quả nghiên cứu. + Báo cáo và nghiệm thu đề tài.						

B3. Kết quả nghiên cứu

Xây dựng bộ dữ liệu với 10.000 câu đánh giá trên 7 miền dữ liệu: **Fashion, Cosmetic, App, Mobile, Toy, Food, Book** để phục vụ bài toán Phân tích thái độ của khách hàng theo từng

khía cạnh qua các phản hồi tiếng Việt trên các trang thương mại điện tử miền mở. Chúng tôi đã xây dựng hướng dẫn gán nhãn dữ liệu chi tiết, chất lượng cũng như là công cụ hỗ trợ gán nhãn dữ liệu để phục vụ cho quá trình thu thập và xử lý dữ liệu. Sau đó, tiến hành cài đặt các phương pháp máy học, học sâu và học chuyển tiếp trên bộ dữ liệu được xây dựng. Và chúng tôi hi vọng kết quả của công trình nghiên cứu này sẽ là tiền đề cho những công trình nghiên cứu tiếp theo về nhiệm vụ nhận diện khía cạnh trên tiếng Việt.

B4. Tài liệu tham khảo

[1] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.:SemEval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Work-shop on Semantic Evaluation (SemEval 2014). pp. 27–35. Association for Computational Linguis-tics, Dublin, Ireland (Aug 2014). https://doi.org/10.3115/v1/S14-2004,https://www.aclweb.org/anthology/S14-2004.

[2] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12:Aspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evalu-ation (SemEval 2015). pp. 486–495. Association for Computational Linguistics, Denver, Colorado (Jun2015). https://doi.org/10.18653/v1/S15-2082,https://www.aclweb.org/anthology/S15-2082

[3] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi,M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier,X., Loukachevitch,N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryi ̆git, G.: SemEval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 19–30. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/S16-1002,https://www.aclweb.org/anthology/S16-1002

[4] Phan, Luong Luc, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Tham Thi Nguyen, Sieu Khai Huynh, Luan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. “SA2SL: From Aspect-Based Sentiment Analysis to Social Listening System for Business Intelligence.” arXiv preprint arXiv:2015.15079(2021).

[5] Nhung Thi- Hong Nguyen, Phuong Phan- Dieu Ha, Luan Thanh Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen. “Vietnamese Complaint Detection on E-Commerce Websites”, arXiv preprint arXiv:2104.11969, 2021.

Ngày 30 tháng 04 năm 2022	*Ngày 30 tháng 04 năm 2022*
Chủ nhiệm đề tài	**Giảng viên hướng dẫn**
(Ký và ghi rõ họ tên)	(Ký và ghi rõ họ tên)
Phạm Tiến Dương	**Nguyễn Thành Luân**