


# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://www.youtube.com/watch?v=QzBv7oK5sbE>
- Link slides (dạng .pdf đặt trên Github):  
<https://github.com/TranVanTinhUIT/CS2205.APR2023/blob/main/T%E1%BB%8Bnh%20Tr%E1%BA%A7n%20V%C4%83n%20-%20xCS2205.DeCuong.FinalReport.APR2023.pdf>
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

<ul style="list-style-type: none"><li>● Họ và Tên: Trần Văn Tịnh</li><li>● MSHV: 220101039</li></ul> 	<ul style="list-style-type: none"><li>● Lớp: CS2205.APR2023</li><li>● Tự đánh giá (điểm tổng kết môn): 8/10</li><li>● Số buổi vắng: 1</li><li>● Link Github: <a href="https://github.com/TranVanTinhUIT/CS2205.APR2023">https://github.com/TranVanTinhUIT/CS2205.APR2023</a></li></ul>
---	--

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

PHÂN TÍCH Ý KIẾN KHÁCH HÀNG VĂN BẢN TIẾNG VIỆT DỰA TRÊN MÔ HÌNH PHOBERT

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

PHOBERT FOR SENTIMENT ANALYSIS OF CUSTOMER FEEDBACK IN VIETNAMESE TEXT

## TÓM TẮT (Tối đa 400 từ)

Nghiên cứu này tập trung vào ứng dụng mô hình PhoBERT để xây dựng mô hình phân loại ý kiến của khách hàng văn bản tiếng việt. Chúng tôi thu thập một lượng lớn các ý kiến đánh giá của khách hàng bằng tiếng việt và gán nhãn tương thích cho mỗi ý kiến theo mức độ hài lòng. Sau đó sử dụng dữ liệu thu thập để xây dựng bộ dữ liệu cho tác vụ phân loại ý kiến và tiếp tục huấn luyện lại mô hình PhoBERT cho tác vụ phân loại ý kiến. Thiết lập một lớp đầu ra để phân loại ý kiến vào các mức độ hài lòng tương ứng. Sử dụng các kỹ thuật fine-tuning để tối ưu hóa mô hình. Đánh giá hiệu suất mô hình được bằng các độ đo như độ chính xác, độ phủ và F1-score.

## GIỚI THIỆU (Tối đa 1 trang A4)

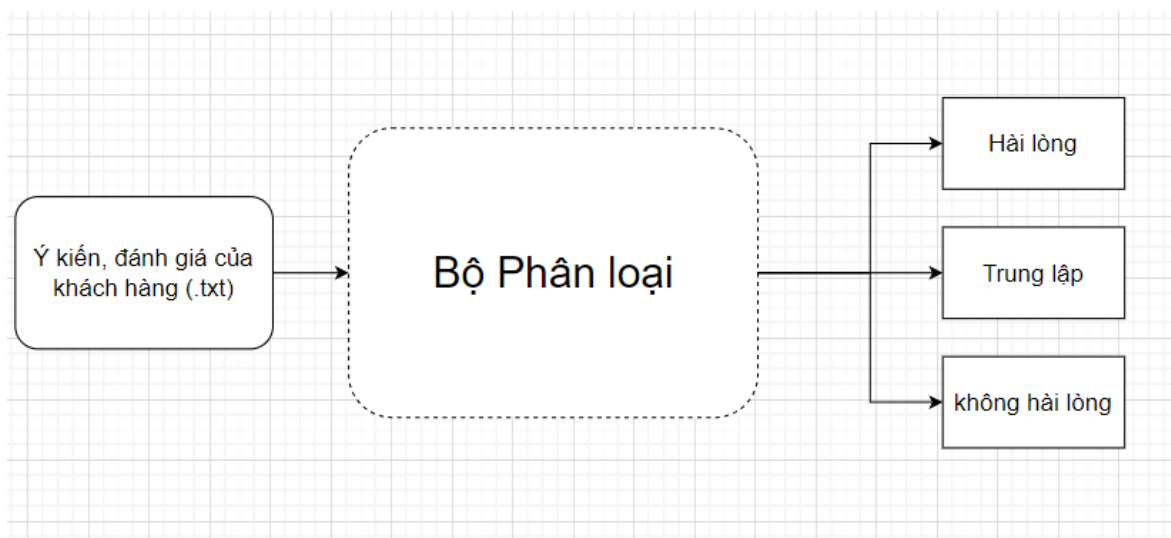
Trong kinh doanh lĩnh vực dịch vụ như dịch vụ lưu trú khách sạn, ý kiến khách hàng đóng vai trò thiết yếu cho việc cải thiện chất lượng sản phẩm, dịch vụ và nâng cao sự hài lòng của khách hàng. Việc phân loại các ý kiến, đánh giá của khách hàng vào từng nhóm cụ thể là điều cần thiết để kịp thời phát hiện các thiếu sót và đưa ra những cải tiến thích hợp đảm bảo sự hài lòng của khách hàng. Tuy nhiên, sử dụng các cách thủ công áp dụng cho việc phân tích các ý kiến, đóng góp này sẽ mất rất nhiều thời gian và việc tổng quát hóa các kết quả cũng rất khó khăn.

PhoBERT, mô hình huấn luyện trên dữ liệu tiếng tiếng dựa trên kiến trúc BERT (Bidirectional Encoder Representations from Transformers) đang được đánh giá là tốt

nhất về khả năng hiểu và biểu diễn ngôn ngữ tiếng Việt, và đã đạt được kết quả tốt trên nhiều tác vụ ngôn ngữ tiếng Việt. Trong nghiên cứu này chúng tôi tập trung ứng dụng mô hình PhoBERT để xây dựng một công cụ phân loại ý kiến của khách hàng vào các lớp cho trước.

**Input:** đoạn văn bản hoặc câu chứa ý kiến khách hàng.

**Output:** nhãn dự đoán ý kiến khách hàng dựa trên input đầu vào theo các mức độ hài lòng, trung lập và không hài lòng.



## MỤC TIÊU

*(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)*

- Xây dựng mô hình phân loại ý kiến khách hàng vào các lớp cho trước với hiệu suất và độ chính xác cao.
- Ứng dụng kết quả xây dựng ứng dụng công cụ hỗ trợ doanh nghiệp, tổ chức tự động phân loại ý kiến của khách hàng và phát hiện nhanh các hạn chế, thiếu sót.
- Đặt tiền đề cho các nghiên cứu sâu hơn, đóng góp vào quá trình nghiên cứu các ứng dụng xử lý ngôn ngữ tự nhiên vào các vấn đề đời sống, xã hội.

## NỘI DUNG VÀ PHƯƠNG PHÁP

*(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)*

## **Nội dung:**

Thu thập và tiền xử lý dữ liệu:

- Thu thập dữ liệu các đánh giá và nhận xét của khách hàng đã được gán nhãn từ các trang web như Booking.com, Agoda, Expedia,... để tạo bộ dữ liệu huấn luyện cho phân tích đánh giá ý kiến khách hàng tiếng Việt. Ngoài ra cũng thu thập một bộ dữ liệu test để đánh giá mô hình.
- Tiền xử lý: loại bỏ dấu chấm câu, các ký tự đặc biệt, icon không cần thiết.
- Mã hóa các câu tiếng Việt thành các token và thêm các special token [CLS] vào đầu câu, [SEP] vào cuối câu và token [PAD] để các câu có độ dài đồng nhất phù hợp với yêu cầu đầu vào của mô hình PhoBERT.

Huấn luyện và tinh chỉnh mô hình:

- Thêm một lớp dự đoán nhãn output và tiếp tục huấn luyện mô hình PhoBERT cho tác vụ phân loại ý kiến dựa vào dữ liệu đã thu thập.
- Áp dụng các kỹ thuật fine-tuning để tối ưu mô hình như Gradient, Regularization, Cross-Validation.

Đánh giá hiệu suất và độ chính xác của mô hình.

- Đánh giá độ chính xác và hiệu suất của mô hình trên dữ liệu test.
- So sánh kết quả của mô hình với các nghiên cứu sử dụng các phương pháp truyền thống như SVM, Naive Bayes, Random Forest và các mô hình máy học sâu như CNN, RNN, BERT.

## **KẾT QUẢ MONG ĐỢI**

*(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)*

- Mô hình tự động phân loại ý kiến, đánh giá của khách hàng có độ chính xác cao và hiệu suất tốt so sánh với các mô hình phân loại dựa trên các phương pháp khác.

- Mô hình có khả năng hiểu và xử lý ngôn ngữ tự nhiên một cách hiệu quả và chính xác.
- Mô hình phân loại có khả năng ứng dụng trong kinh doanh để tối ưu hóa chiến lược kinh doanh, chất lượng sản phẩm.

## **TÀI LIỆU THAM KHẢO** (*Định dạng DBLP*)

- [1] [Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT \(1\) 2019: 4171-4186](#)
- [2] [Dat Quoc Nguyen, Anh Tuan Nguyen: PhoBERT: Pre-trained language models for Vietnamese. EMNLP \(Findings\) 2020: 1037-1042](#)
- [3] [Nguyen Phuc Minh, Tran Hoang Vu, Vu Hoang, Ta Duc Huy, Trung Huu Bui, Steven Quoc Hung Truong: ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining. LREC 2022: 328-337](#)
- [4] [Philip Kenneweg, Sarah Schröder, Barbara Hammer: Neural Architecture Search for Sentence Classification with BERT. ESANN 2022](#)
- [5] [Jinghui Lu, Maeve Henchion, Ivan Bacher, Brian Mac Namee: A Sentence-level Hierarchical BERT Model for Document Classification with Limited Labelled Data. CoRR abs/2106.06738 \(2021\)](#)
- [6] [Philip Kenneweg, Sarah Schröder, Barbara Hammer: Neural Architecture Search for Sentence Classification with BERT. ESANN 2022](#)
- [7] [Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, Manabu Okumura: Fusing Label Embedding into BERT: An Efficient Improvement for Text Classification. ACL/IJCNLP \(Findings\) 2021: 1743-1750](#)
- [8] [Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention is All you Need. NIPS 2017: 5998-6008](#)