# **THÔNG TIN CHUNG CỦA BÁO CÁO**

* Link YouTube video của báo cáo (tối đa 5 phút):   
  *https://www.youtube.com/watch?v=nrgrmTeVq8c*
* Link slides (dạng .pdf đặt trên Github):   
  *https://github.com/huynguyen261219/CS2205/blob/main/Huy%20Nguye%CC%82%CC%83n%20Quang%20-%20xCS2205.DeCuong.FinalReport.Template.Slide.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*

| * Họ và Tên: Nguyễn Quang Huy * MSSV: 230201014 | * Lớp: CS2205.CH181 * Tự đánh giá (điểm tổng kết môn): 8.0/10 * Số buổi vắng: 2 * Số câu hỏi QT cá nhân: 0 * Link Github: https://github.com/huynguyen261219/CS2205 |
| --- | --- |

# **ĐỀ CƯƠNG NGHIÊN CỨU**

| **TÊN ĐỀ TÀI (IN HOA)**  PHÁT TRIỂN HỆ THỐNG Q-A VỀ LUẬT VI PHẠM LIÊN QUAN ĐẾN XE MÁY TẠI VIỆT NAM |
| --- |
| **TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**  DEVELOPMENT OF A QUESTION ANSWERING SYSTEM ON MOTORCYCLE-RELATED TRAFFIC VIOLATIONS IN VIETNAM |
| **TÓM TẮT** *(Tối đa 400 từ)*  Xe máy là một phương tiện giao thông chính tại Việt Nam đối với tất cả mọi người dân, vì thế các điều luật được ban hành ra các quy định xử phạt hành chính đối với xe máy gây rối trật tự an toàn giao thông. Tuy nhiên có một bộ phận không nhỏ người dân vẫn chưa hiểu và nắm hết về luật này. Chính vì vậy, cần phải có một ứng dụng hỏi đáp để có thể cung cấp một cách thuận tiện nhất cho người dân hỏi đáp các quy định về luật giao thông mà có liên quan đến xe máy, các chế tài xử phạt và mức phạt khi bị cảnh sát giao thông bắt giữ là gì.  Ứng dụng này sẽ thuần về các câu hỏi tiếng Việt cho nên sẽ sử dụng model PhoBERT large và fine-tune cho tác vụ luật vi phạm liên quan đến xe máy. Ứng dụng này phải đáp ứng được yêu cầu độ trễ thấp (tốc độ response ít hơn 5s), độ chính xác cao (vì là thông tin quan trọng) và giao diện phải thân thiện đáp ứng UI/UX cho người dùng. |
| **GIỚI THIỆU** *(Tối đa 1 trang A4)*  Thông tin về luật giao thông hiện nay đã được public trên các trang mạng điện tử bằng các văn bản. Tuy nhiên việc người dân phải tra cứu trên mạng Internet những thông tin liên quan đến luật vi phạm trên xe máy trên các văn bản đó đôi khi khiến mọi người mất rất nhiều thời gian và việc lọc lại thông tin đôi khi có những thiếu sót.  Do đó việc xây dựng ứng dụng để dễ dàng hỏi đáp là rất cần thiết và với sự phát triển của các mô hình học sâu, ta có thể tận dụng PhoBERT, một mô hình được pre-train dành riêng cho tiếng Việt dựa trên mô hình BERT nổi tiếng của Google để fine-tune và sử dụng. Mô hình này đạt được độ chính xác và hiệu suất cao trên các tác vụ NLP về tiếng Việt, ta cũng có thể tuỳ chỉnh cho các tác vụ khác nhau.  PhoBERT có thể hiểu được các ngữ nghĩa, ngữ pháp và sắc thái của tiếng Việt theo câu hỏi của người dùng vì nó đã được huấn luyện trên một tập dữ liệu văn bản tiếng Việt lớn, bao gồm nhiều loại văn bản khác nhau như: tin tức, văn học, các tài liệu trực tuyến,... và với việc được huấn luyện hai chiều thì mô hình này có thể bắt ngữ cảnh của từ trong cả hai chiều (trái và phải) trong một câu, do đó mô hình có thể hiểu sâu hơn về cấu trúc câu và ngữ nghĩa, giúp cung cấp các câu trả lời chính xác hơn trong các hệ thống QA.  Ứng dụng xây dựng trên app điện thoại trên cả 2 nên tảng là Android và IOS với giao diện dễ sử dụng và thân thiện với người dùng, đáp ứng đầy đủ yếu tố UI/UX. |
| **MỤC TIÊU**  *(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)*   * Thu thập dataset về các văn bản luật giao thông liên quan đến xe máy, các mức xử phạt, các tài liệu hướng dẫn và các giải thích rõ ràng các quy định * Đưa mô hình PhoBERT vào và tiến hành fine-tune theo tác vụ Question-Answering với dữ liệu được thu thập hoàn chỉnh bằng tiếng Việt. Tiến hành các bước tiền xử lý dữ liệu, sau đó đưa vào huấn luyện với dữ liệu mới * Đánh giá mô hình sau khi huấn luyện (độ chính xác và tốc độ câu trả lời) bằng các phương pháp EM (Exact Match), F1 Score và deploy và đưa vào trong ứng dụng di động được lập trình. Tiến hành update thêm thông tin sau khi luật sửa đổi |
| **NỘI DUNG VÀ PHƯƠNG PHÁP**  *(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)*  Nội dung 1: Thu thập các dữ liệu văn bản pháp luật về giao thông xe máy   * Thu thập trên mạng Internet các văn bản liên quan trên nhiều trang thông tin chính thống và tin cậy để thực hiện tiền xử lý dữ liệu * Loại bỏ các thông tin gây nhiều như: chấm câu, ký tự đặc biệt, icon * Tokenize để chuyển input text thành token mà mô hình có thể hiểu được * Chia thành các tập train, val và test để thực hiện huấn luyện, đánh giá và test mô hình sau khi fine-tune PhoBERT   Nội dung 2: Fine-tune PhoBERT   * Chọn mô hình PhoBERT-large với 370M tham số * Thêm các lớp Attention vào cuối PhoBERT để có thể fine-tune tập trung vào tác vụ của câu hỏi về văn bản luật giao thông liên quan đến xe máy làm cho mô hình cải thiện được độ chính xác. Hơn nữa điều này còn giúp xử lý được các văn bản dài, giảm nhiều và nắm bắt được các phụ thuộc phức tạp. * Sau khi mô hình được fine-tune theo tác vụ thì tiến hành đánh giá bằng các phương pháp EM (Exact match), F1 Score và test các câu hỏi và câu trả lời trên mô hình. Cải thiện thêm để tăng tốc độ trả lời của mô hình với độ trễ thấp nhằm đáp ứng cho nhu cầu của người dùng nhưng vẫn giữ được độ chính xác của câu trả lời.   Nội dung 3: Deploy   * Lưu mô hình PhoBERT đã được fine-tuned, sau đó chuyển đổi mô hình thành các định dạng tương thích như: TorchScript, ONNX * Import mô hình đã được chuyển đổi vào ứng dụng di động Android và IOS |
| **KẾT QUẢ MONG ĐỢI**  *(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)*   * Thiết kế dataset hoàn chỉnh, đầy đủ và chính xác nhất cho mô hình để train * Fine-tune mô hình PhoBERT cho tác vụ với độ chính xác trên 97% và độ trễ thấp dưới 10s * Dùng hạn chế tài nguyên của mô hình trên GPU, CPU vì được ứng dụng trên di động với tài nguyên thấp * Các biến thể của câu hỏi từ user được xử lý tốt như: viết tắt, đa nghĩa, ngữ cảnh * Tối ưu kích thước mô hình cho ứng dụng di động |
| **TÀI LIỆU THAM KHẢO** *(Định dạng DBLP)*  [1]. Dat Quoc Nguyen and Anh Tuan Nguyen (2020). PhoBERT: Pre-trained language models for Vietnamese. ArXiv preprint arXiv:2003.00744.  [2]. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv preprint arXiv:1810.04805  [3]. Thi-Thanh Ha, Van-Nha Nguyen, Kiem-Hieu Nguyen, Kim-Anh Nguyen and Tien-Thanh Nguyen (2020). Utilizing BERT for Question Retrieval in Vietnamese E-commerce Sites. ACL |