TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÙI HẢI DƯƠNG – 521H0220**

**TRẦN QUANG LUÂN -**

**TÊN ĐỀ TÀI**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**KHOA HỌC MÁY TÍNH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÙI HẢI DƯƠNG – 521H0220**

**TRẦN QUANG LUÂN - MSSV**

**TÊN ĐỀ TÀI**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN**

**HOA HỌC MÁY TÍNH**

Người hướng dẫn

**PGS. TS. Lê Anh Cường**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn …………………………………… …………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

*TP. Hồ Chí Minh, ngày 19 tháng 01 năm 2025.*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của PGS. TS. Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 19 tháng 01 năm 2025*

*Tác giả*

*(Ký tên và ghi rõ họ tên)*

**TÊN ĐỀ TÀI**

**TÓM TẮT**

(Time New Romans – 13)

**TITLE**

**ABSTRACT**

(Time New Romans – 13)

**MỤC LỤC**

[DANH MỤC HÌNH VẼ v](#_Toc143173056)

[DANH MỤC BẢNG BIỂU vii](#_Toc143173057)

[DANH MỤC CÁC CHỮ VIẾT TẮT viii](#_Toc143173058)

[CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI 1](#_Toc143173059)

[1.1 Lý do chọn đề tài 1](#_Toc143173060)

[1.2 Mục tiêu thực hiện đề tài 1](#_Toc143173061)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 2](#_Toc143173062)

[2.1 Mạng neural hồi quy 2](#_Toc143173063)

[2.1.1 Recurrent Neural Network (RNN) 2](#_Toc143173064)

[2.1.2 Long Short-term Memory (LSTM) 2](#_Toc143173065)

[2.2 Mô hình Transformer 3](#_Toc143173066)

[2.2.1 Encoder và Decoder 3](#_Toc143173067)

[2.2.2 Attention 3](#_Toc143173068)

[CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT 4](#_Toc143173069)

[CHƯƠNG 4. THỰC NGHIỆM 4](#_Toc143173070)

[4.1 Dữ liệu thực nghiệm 4](#_Toc143173071)

[4.2 Cài đặt thực nghiệm 4](#_Toc143173072)

[CHƯƠNG 5. KẾT LUẬN 5](#_Toc143173073)

[5.1 Kết luận 5](#_Toc143173074)

[5.2 Hướng phát triển 5](#_Toc143173075)

[TÀI LIỆU THAM KHẢO 6](#_Toc143173076)

# DANH MỤC HÌNH VẼ

[Hình 2.1: Scaled Dot-Product Attention 3](#_Toc142677545)

# DANH MỤC BẢNG BIỂU

[Bảng 4.1: Thống kê kiểu thực thể trong tập VLSP 2016 4](#_Toc142677565)

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |
| --- | --- |
| BERT | Bidirectional Encoder Representations from Transformers |
| GEC | Grammatical Error Correction |
| MLM | Masked Language Model |
| NLP | Natural Language Processing |
| NSP | Next Sentence Prediction |

# MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

Trong năm 2025, luật giao thông đã có nhiều thay đổi lớn và đến đời sống của người dân rất nhiều đặc biệt là các hình phạt được tăng lên được quy định trong Nghị định 168 có hiệu lực vào 1/1/2025. Mặc dù giao thông đường bộ gắn liền với mọi người hàng ngày nhưng việc người dân nắm bắt được hết luật giao thông đường bộ là rất khó do có rất nhiều quy định được ban hành trong các Luât, thông tư và nghị định và không phải ai cũng có thể bỏ thời gian, công sức để học và tiếp thu được hết những tài liệu đó. Chính vì vậy việc xây dựng và áp dụng chatbot về luật giao thông đường bộ là một giải pháp tốt, mọi người dễ tiếp cận và tiết kiệm được thời gian, công sức, giúp nâng cao kiến thức và ý thức của mọi người khi tham gia giao thông.

## Mục tiêu thực hiện đề tài

Mục tiêu của đề tài này là nghiên cứu, thiết kế và xây dựng một chatbot thông minh, có khả năng tự động trả lời các câu hỏi liên quan đến luật giao thông, đồng thời hỗ trợ giải đáp những tình huống cụ thể mà người dùng có thể gặp phải trong quá trình tham gia giao thông. Chatbot này không chỉ giúp nâng cao nhận thức pháp luật giao thông mà còn cung cấp thông tin nhanh chóng, chính xác và dễ hiểu, góp phần hỗ trợ người dùng tuân thủ các quy định và giảm thiểu các vi phạm giao thông.

# PHƯƠNG PHÁP SỬ DỤNG

## Phương pháp tiếp cận

Thông thường sẽ có hai cách tiếp cận cho bài toán về chatbot này đó chính là Finetune mô hình có sẵn hoặc sử dụng Retrieval Augmented Generation. Cả hai cách tiếp cận này đều sử dụng dữ liệu có sẵn của chúng ta để trả lời những câu hỏi của người dung khi tương tác với chatbot nhưng cách thực hiện của chúng lại khác nhau.

### Finetune mô hình

Finetune mô hình có sẵn là việc sử dụng các mô hình đã được huấn luyện (pre-trained model) dựa trên bộ dữ liệu rất lớn. Những mô hình đó sẽ được sử dụng và được huấn luyện tiếp trên dữ liệu của chúng ta và từ đó mô hình có thể được sử dụng trong task cụ thể của chúng ta.

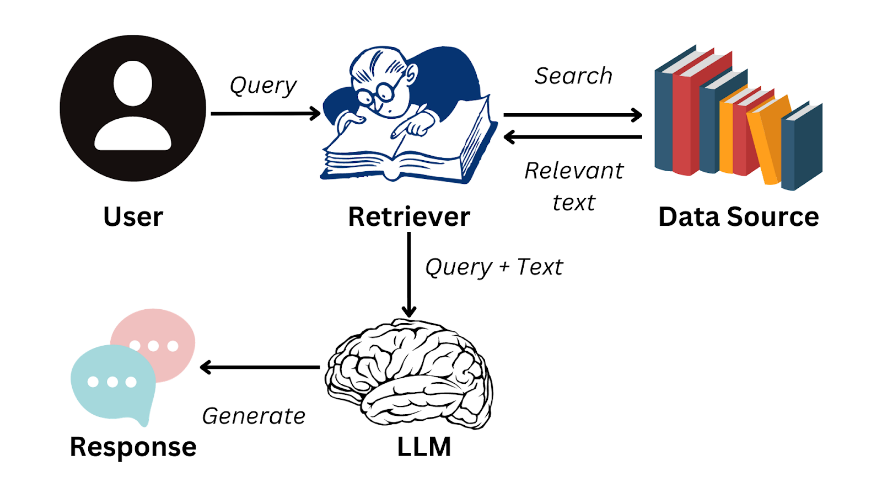
A diagram of a function

Description automatically generated

Hình trên đây thể hiện cho quá trình Finetune mô hình. Từ những pre-train model, chúng ta sẽ huấn luyện tiếp trên dữ liệu của chúng ta. Thông qua quá trình backpropagation, trọng số và bias của mô hình trong các lớp sẽ được cập nhật, từ đó sẽ phù hợp trong tác vụ mà chúng ta thực hiện.

### Retrieval Augmented Generation (RAG)

Cách hoạt động của RAG thì khác so với việc Finetune mô hình. Về cách thức hoạt động, RAG sẽ đưa vào mô hình ngôn ngữ lớn (LLM) câu hỏi của người dùng và các tài liệu liên quan để LLM có thể trả lời.



Hình trên thể hiện quy trình hoạt động của RAG. Dữ liệu của chúng ta đầu tiên sẽ được lưu trữ trong một hoặc nhiều database. Khi người dùng thực hiện nhập vào một câu hỏi, chúng ta sẽ thực hiện tìm kiếm các tài liệu mà chúng ta đã lưu trước đó trong database có liên quan đến câu hỏi của người dùng. Câu hỏi của người dùng và các tài liệu đã được truy vấn sẽ được truyền trực tiếp vào trong mô hình để mô hình có thể tự trả lời dựa trên những thông tin đã tìm được.

### Điểm mạnh và điểm yếu của hai phương pháp

Điểm mạnh:

* Finetune pre-trained model:
  + Mô hình được finetune sẽ phù hợp với tác vụ chuyên biệt.
  + Tận dụng được những dữ liệu đã được huấn luyện trước đó.
  + Vì không cần lấy từ các dữ liệu bên ngoài và chỉ dựa và những gì đã được huấn luyện để thực hiện tác vụ nên tốc độ phản hồi rất nhanh.
* RAG:
  + Dữ liệu mới sẽ được cập nhật một cách dễ dàng bằng cách đưa vào trong database mà không cần cập nhật lại mô hình.
  + Không phải mất thời gian huấn luyện mô hình nên sẽ tiết kiệm được thời gian và chi phí.
  + Kết quả được dựa vào trực tiếp từ nguồn dữ liệu nên sẽ đưa ra kết quả chính xác và không gây ra hiện tượng ảo giác.

Điểm yếu:

* Finetune pre-trained model:
  + Việc huấn luyện lại mô hình tốn rất nhiều thời gian và chi phí.
  + Muốn cập nhật dữ liệu mới cho mô hình thì phải huấn luyện mô hình trên dữ liệu mới đó.
  + Khi thực hiện tác vụ thì mô hình có thể sẽ bị ảo tưởng từ những kiến thức mình đã học và đưa ra kết quả thiếu chính xác.
* RAG:
  + Việc đưa ra kết quả sẽ chậm, phụ thuộc vào quá trình truy xuất thông tin từ câu hỏi.
  + Nếu thông tin truy xuất kém thì có thể dẫn tới kết quả trả về thiếu xót hoặc không chính xác.
  + Vì quy trình của RAG phức tạp do có nhiều thành phần kết hợp lại với nhau.

### Phương pháp tiếp cận phù hợp

Vì tác vụ cần làm là về luật giao thông đường bộ nên các sẽ cần đáp ứng các tiêu chí sau:

* Dữ liệu pháp luật phải là dữ liệu mới nhất và dễ dàng cập nhật.
* Kết quả trả về phải chính xác dựa theo dữ liệu pháp luật.

Việc finetune mô hình cũng có thể cập nhật được dữ liệu mới những sẽ tốn rất nhiều thời gian và chi phí. Bên cạnh đó mô hình được finetune có khả năng sẽ trả lời không chính xác do mô hình có thể bị hiện tượng ảo tưởng.

Với RAG thì việc cập nhật dữ liệu này sẽ rất dễ dàng và cũng rất nhanh chóng và việc đưa trực tiếp tài liệu lưu trữ vào cho mô hình trả lời sẽ chính xác hơn và không gây ra hiện tượng ảo giác. Chính vì vậy RAG sẽ là giải pháp tiếp cận phù hợp.

## Mô hình sử dụng

### Mô hình embedding

Chức năng chính của mô hình embedding là chuyển đổi các token như ảnh, video, chữ cái hoặc âm thành sang biểu diễn vector, từ đó sẽ giúp việc xử lí dữ liệu trong học máy một cách dễ dàng hơn.

Trong RAG thì việc xử lý embedding này sẽ được thực hiện trong khâu truy xuất thông tin, bằng cách đưa các chunk được chia nhỏ từ các văn bản chính, từ đó embedding chúng thành dạng vector và lưu trữ trong database. Khi các câu hỏi của người dùng được đưa vào thì chúng sẽ được chuyển qua dạng vector và thực hiện so sánh với các vector đã được lưu trong database và truy xuất ra những dữ liệu liên quan.

Dưới đây là thống kê về kết quả của các mô hình embedding dựa trên 20% của tập huấn luyện luật Zalo, chưa được huấn luyện trên các mô hình dưới đây. Kết quả thể hiện đánh giá về Acurracy, Precision và Recall:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy@1 | Accuracy@3 | Accuracy@5 | Accuracy@10 |
| vietnamese-bi-encoder | 0.8169 | 0.9108 | 0.9437 | 0.9640 |
| sup-SimCSE-VietNamese-phobert-base | 0.5540 | 0.7308 | 0.7981 | 0.8748 |
| halong\_embedding  (768) | 0.8294 | 0.9233 | 0.9437 | 0.9687 |
| halong\_embedding (512) | 0.8138 | 0.9233 | 0.9390 | 0.9703 |
| halong\_embedding (256) | 0.7934 | 0.8967 | 0.9280 | 0.9593 |
| halong\_embedding (128) | 0.7840 | 0.8951 | 0.9264 | 0.9515 |
| halong\_embedding (64) | 0.6980 | 0.8435 | 0.8920 | 0.9358 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Precision@1 | Precision@3 | Precision@5 | Precision@10 |
| vietnamese-bi-encoder | 0.8169 | 0.3099 | 0.1931 | 0.0987 |
| sup-SimCSE-VietNamese-phobert-base | 0.5540 | 0.2473 | 0.1621 | 0.0892 |
| halong\_embedding  (768) | 0.8294 | 0.3146 | 0.1931 | 0.0991 |
| halong\_embedding (512) | 0.8138 | 0.3146 | 0.1922 | 0.0992 |
| halong\_embedding (256) | 0.7934 | 0.3062 | 0.1900 | 0.0981 |
| halong\_embedding (128) | 0.7840 | 0.3046 | 0.1894 | 0.0975 |
| halong\_embedding (64) | 0.6980 | 0.2864 | 0.1815 | 0.0954 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Recall@1 | Recall@3 | Recall@5 | Recall@10 |
| vietnamese-bi-encoder | 0.8020 | 0.9045 | 0.9390 | 0.9601 |
| sup-SimCSE-VietNamese-phobert-base | 0.5446 | 0.7246 | 0.7903 | 0.8693 |
| halong\_embedding  (768) | 0.8146 | 0.9178 | 0.9390 | 0.9640 |
| halong\_embedding (512) | 0.7989 | 0.9178 | 0.9343 | 0.9656 |
| halong\_embedding (256) | 0.7786 | 0.8920 | 0.9233 | 0.9546 |
| halong\_embedding (128) | 0.7707 | 0.8889 | 0.9210 | 0.9476 |
| halong\_embedding (64) | 0.6854 | 0.8365 | 0.8842 | 0.9311 |

Dựa vào thống kê ở trên, ta có thể thấy được mô hình embedding cho ra kết quả cao nhất đó là halong\_embedding (768). Vậy nên halong\_embedding (768) sẽ là lựa chọn để sử dụng làm mô hình embedding.

### Mô hình tạo sinh

Mô hình tạo sinh là các mô hình ngôn ngữ lớn có khả năng tạo ra dữ liệu mới từ những dữ liệu mà nó đã học thông qua dữ liệu đầu vào. Đây là các mô hình đã được huấn luyện trên các tập dữ liệu cực kì lớn với các chủ đề khác nhau.

Các mô hình tạo sinh thường được nhắc đến khi áp dụng vào RAG thường là mô hình của OpenAI, Google, Meta

Mô hình gemini của Meta hiện đang miễn phí nên sẽ được sử dụng cho các dự án nhỏ, thiếu kinh phí.

## Hệ thống và các thành phần

## Chi tiết các thành phần

# THỰC NGHIỆM

## Dữ liệu sử dụng

# TÀI LIỆU THAM KHẢO

Tiếng Việt

…

Tiếng Anh

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation*, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762