**CHƯƠNG 1. TỔNG QUAN VỀ HỆ THỐNG TRUY VẤN THÔNG TIN**

**1.1. Cơ sở lý thuyết về xử lý ngôn ngữ tự nhiên**

**1.1.1. Định nghĩa về xử lý ngôn ngữ tự nhiên**

Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) là một lĩnh vực nghiên cứu của khoa học máy tính nhằm ứng dụng trí tuệ nhận tạo thông qua một hệ thống các kỹ thuật, phương pháp, thuật toán nhằm điều khiển các hệ thống máy tính có thể nhận biết, xử lý ngôn ngữ một cách tự nhiên nhất có thể tương tự như con người bằng nhiều ngôn ngữ khác nhau.

Mục tiêu chính của NLP là giúp máy tính hiểu, diễn giải và tạo ra ngôn ngữ tự nhiên một cách tự động. NLP bao gồm nhiều phần tử và công nghệ khác nhau, bao gồm:

- Xử lý ngôn ngữ tự nhiên: Đây là quá trình xử lý và phân tích ngôn ngữ tự nhiên. Điều này bao gồm việc tách từ, phân tích cú pháp, phân loại ngữ nghĩa, trích xuất thông tin và truy vấn ngôn ngữ tự nhiên.

- Hiểu ngôn ngữ tự nhiên: Mục tiêu là giúp máy tính hiểu ý nghĩa của ngôn ngữ tự nhiên. Điều này bao gồm hiểu ý nghĩa của câu, xác định ngữ cảnh, nhận dạng tác giả hoặc người nói, và hiểu ý đồ của người dùng.

- Tạo ngôn ngữ tự nhiên: Đây là quá trình tạo ra ngôn ngữ tự nhiên từ dữ liệu không phải ngôn ngữ tự nhiên. Ví dụ, tạo ra câu mô tả từ dữ liệu số hoặc tạo ra bài viết tự động từ dữ liệu cấu trúc.

- Dịch máy: Dịch máy là quá trình chuyển đổi văn bản từ một ngôn ngữ sang ngôn ngữ khác một cách tự động.

- Học máy trong NLP: Học máy đóng vai trò quan trọng trong NLP. Các phương pháp học máy, bao gồm học có giám sát và học không có giám sát.

**1.1.2. Các ứng dụng của NLP đang được sử dụng hiện nay**

Nhìn chung, NLP đang được ứng dụng rất rộng rãi trong mọi mặt của đời sống xã hội, tuy nhiên để có cái nhìn tổng quan, chúng ta có thể phân loại qua 3 lĩnh vực sau

1.1.2.1. Ứng dụng trong lĩnh vực xử lý tiếng nói và hình ảnh

- Nhận dạng và tổng hợp giọng nói

**[cần phân tích thêm]**

- Nhận dạng chữ viết

**[cần phân tích thêm]**

1.1.2.2. Ứng dụng trong lĩnh vực xử lý văn bản

- Dịch máy

**[cần phân tích thêm]**

- Tóm tắt văn bản

**[cần phân tích thêm]**

1.1.2.3. Ứng dụng trong lĩnh vực phân tích văn bản

- Tìm kiếm và truy vấn thông tin (chatbot): Chatbot là một hệ thống truy vấn thông tin và trả lời tự động dựa trên khuôn mẫu cho sẵn kết hợp với các thuật toàn xử lý ngôn ngữ tự nhiên nhằm đưa ra phản hồi cho người dùng thông qua văn bản hoặc âm thanh. Đối với chatbot, người dùng có thể yêu cầu bằng câu hỏi hoặc ra lệnh và chatbot sẽ phân tích, hiểu sau đó đưa ra phản hồi theo yêu cầu. Mức độ chính xác và tự nhiên của câu trả lời phụ thuộc vào khả năng xử lý dữ liệu đầu vào cũng như độ phức tạo của thuật toán xử lý.

**[cần phân tích thêm]**

- Tổng hợp văn bản

**[cần phân tích thêm]**

Đối với từng lĩnh vực khác nhau thì NLP được sử dụng theo các hướng tiếp cận khác nhau, từ đó làm đa dạng thêm các hướng xử lý riêng cho từng lĩnh vực. Trong đề án này, tôi tập trung vào nghiên cứu ứng dụng NLP vào mục đích tìm kiếm và truy vấn thông tin (chatbot)

**1.2. Các công trình nghiên cứu trước đây về việc sử dụng hệ thống truy vấn thông tin**

1.2.1. Tình hình nghiên cứu ngoài nước

*1.2.1.1. Các nghiên cứu về lý thuyết*

*- Word2Vector*

*- Deep Learning*

*- z*

1. **Tên Công trình/Bài báo:** Sequence to Sequence Learning with Neural Networks
   * **Tác giả chính:** Ilya Sutskever, Oriol Vinyals, Quoc V. Le
   * **Nguồn công bố:** NeurIPS 2014
   * **Năm:** 2014
   * **Điểm nổi bật:** Giới thiệu kiến trúc sequence-to-sequence (Seq2Seq) sử dụng mạng nơ-ron hồi quy (RNN), đặt nền móng cho việc áp dụng học sâu vào các nhiệm vụ tạo văn bản, bao gồm cả đối thoại.
2. **Tên Công trình/Bài báo:** A Neural Conversational Model
   * **Tác giả chính:** Oriol Vinyals, Quoc V. Le
   * **Nguồn công bố:** arXiv preprint arXiv:1506.05869 (cũng trình bày tại ICML 2015 Deep Learning Workshop)
   * **Năm:** 2015
   * **Điểm nổi bật:** Một trong những công trình đầu tiên áp dụng trực tiếp mô hình Seq2Seq cho việc xây dựng chatbot miền mở, cho thấy khả năng tạo ra các phản hồi mạch lạc và phù hợp ngữ cảnh ở mức độ nhất định.
3. **Tên Công trình/Bài báo:** Attention Is All You Need
   * **Tác giả chính:** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin
   * **Nguồn công bố:** NeurIPS 2017
   * **Năm:** 2017
   * **Điểm nổi bật:** Công trình cách mạng giới thiệu kiến trúc **Transformer**, hoàn toàn dựa trên cơ chế "attention" thay vì RNN. Đây là nền tảng cho hầu hết các mô hình ngôn ngữ lớn hiện đại như BERT và GPT, và do đó là nền tảng của nhiều chatbot AI tiên tiến ngày nay.

*1.2.1.2. Các công trình ứng dụng đã thực hiện*

**Các Công Trình Nền Tảng và Kiến Trúc:**

**Các Mô Hình Ngôn Ngữ Lớn và Chatbot Hiện Đại:**

1. **Tên Công trình/Bài báo:** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
   * **Tác giả chính:** Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
   * **Nguồn công bố:** NAACL 2019
   * **Năm:** 2018 (arXiv), 2019 (Hội nghị)
   * **Điểm nổi bật:** Mặc dù BERT chủ yếu dùng cho *hiểu* ngôn ngữ, các kỹ thuật huấn luyện trước (pre-training) và kiến trúc Transformer của nó đã ảnh hưởng sâu sắc đến việc phát triển các mô hình *sinh* ngôn ngữ cho chatbot.
2. **Tên Công trình/Bài báo:** Language Models are Few-Shot Learners (Giới thiệu GPT-3)
   * **Tác giả chính:** Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. (OpenAI)
   * **Nguồn công bố:** NeurIPS 2020
   * **Năm:** 2020
   * **Điểm nổi bật:** Trình bày mô hình GPT-3 với quy mô khổng lồ (175 tỷ tham số) và khả năng học "few-shot" ấn tượng, cho thấy tiềm năng của việc tăng quy mô mô hình để cải thiện đáng kể hiệu năng trên nhiều tác vụ NLP, bao gồm cả đối thoại. GPT-3 và các phiên bản sau này là cốt lõi của ChatGPT.
3. **Tên Công trình/Bài báo:** Towards a Human-like Open-Domain Chatbot (Giới thiệu Meena)
   * **Tác giả chính:** Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, et al. (Google)
   * **Nguồn công bố:** arXiv preprint arXiv:2001.09977
   * **Năm:** 2020
   * **Điểm nổi bật:** Giới thiệu Meena, một chatbot miền mở được huấn luyện trên dữ liệu đối thoại lớn và đề xuất một độ đo mới (Sensibleness and Specificity Average - SSA) để đánh giá chất lượng hội thoại gần với con người hơn.
4. **Tên Công trình/Bài báo:** Recipes for Building an Open-Domain Chatbot (Giới thiệu BlenderBot)
   * **Tác giả chính:** Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Jason Weston, et al. (Facebook AI/Meta AI)
   * **Nguồn công bố:** EACL 2021
   * **Năm:** 2020 (arXiv), 2021 (Hội nghị)
   * **Điểm nổi bật:** Mô tả chi tiết các "công thức" (kỹ thuật huấn luyện, kiến trúc, dữ liệu) để xây dựng BlenderBot, một chatbot miền mở mạnh mẽ, tập trung vào các kỹ năng như thể hiện tính cách, kiến thức và sự đồng cảm. Có các phiên bản BlenderBot 2.0, 3.0 sau đó.
5. **Tên Công trình/Bài báo:** LaMDA: Language Models for Dialog Applications
   * **Tác giả chính:** Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, et al. (Google)
   * **Nguồn công bố:** arXiv preprint arXiv:2201.08239
   * **Năm:** 2022
   * **Điểm nổi bật:** Giới thiệu LaMDA, mô hình ngôn ngữ được tinh chỉnh đặc biệt cho ứng dụng đối thoại, tập trung vào các tiêu chí như tính hợp lý (sensibleness), tính cụ thể (specificity), và tính thú vị (interestingness), cùng với các nỗ lực về an toàn và giảm thiểu thông tin sai lệch.

1.2.2. Tình hình nghiên cứu trong nước

1. **Công trình nền tảng (Mô hình ngôn ngữ cho Tiếng Việt):**
   * **Tên công trình/Bài báo:** PhoBERT: Pre-trained language models for Vietnamese
   * **Tác giả chính:** Dat Quoc Nguyen, Anh Tuan Nguyen (VinAI Research)
   * **Nguồn công bố:** Findings of EMNLP 2020
   * **Năm:** 2020
   * **Điểm nổi bật:** Mặc dù không phải là về chatbot, PhoBERT là một mô hình Transformer được huấn luyện trước cực kỳ quan trọng cho tiếng Việt, làm nền tảng cho nhiều ứng dụng NLP tiếng Việt sau này, bao gồm cả các chatbot tiên tiến. Nó cho phép hiểu ngữ nghĩa tiếng Việt tốt hơn nhiều.
2. **Công trình nền tảng (Mô hình sinh ngôn ngữ cho Tiếng Việt):**
   * **Tên công trình/Bài báo:** BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese
   * **Tác giả chính:** Nguyen Luong Tran, Duong Minh Le, Dat Quoc Nguyen (VinAI Research)
   * **Nguồn công bố:** Findings of ACL: EMNLP 2021
   * **Năm:** 2021
   * **Điểm nổi bật:** Tương tự PhoBERT nhưng tập trung vào mô hình sequence-to-sequence (như BART), rất phù hợp cho các tác vụ sinh văn bản, bao gồm cả sinh phản hồi trong chatbot tiếng Việt.

**1.3. Tổng quan về**

**. Kết luận Chương 1**

**CHƯƠNG 2. THỰC TRẠNG HOẠT ĐỘNG GIẢNG DẠY MÔN LUẬT BIỂN VÀ QUY TẮC TRÁNH VA QUỐC TẾ TẠI TRƯỜNG CĐKT HẢI QUÂN**

**2.1. Giới thiệu về Trường CĐKT Hải quân**

Trường CĐKT Hải quân được thành lập ngày 12/02/1979 trên cơ sở Trường Công nhân Kỹ thuật Ba Son, Trường Sơ cấp CMKT Hải quân. Quá trình xây dựng và phát triển trường CĐKT Hải quân đã trải qua nhiều lần đổi tên gọi: Trường TCKT tàu Hải quân (năm 1990); Trường THKT Hải quân (năm 1995); Trường TCKT Hải quân (năm 2008). Năm 2017, Nhà trường được nâng cấp lên thành Trường CĐKT Hải quân. Địa điểm đóng quân của Nhà trường tại 1295A, Nguyễn Thị Định, Phường Cát Lái, TP. Thủ Đức, TP. Hồ Chí Minh.

Trải qua hơn 45 năm xây dựng và trưởng thành, Nhà trường đã hoàn thành tốt nhiệm vụ đào tạo, cung cấp nguồn nhân lực có chất lượng cho Quân chủng Hải quân và nhiều đơn vị khác trong toàn quân (Nhà trường đã đào tạo được hơn 10 ngàn quân nhân chuyên nhân chuyên nghiệp môn kỹ thuật trình độ trung cấp, cao đẳng; 30 ngàn hạ sĩ quan- chiến sĩ chuyên môn kỹ thuật sơ cấp, đào tạo nghề được hơn 20 ngàn người …ra trường công tác trên các con tàu, đài trạm, nhà máy, xí nghiệp… xứng đáng là lực lượng nòng cốt, “xương sống” xây dựng Quân chủng Hải quân hiện đại). Các thế hệ cán bộ, giảng viên, QNCN, CNVQP, Học viên, HSQ – BS Trường CĐKT Hải quân đã nỗ lực, bền bỉ, kiên trì phấn đấu hoàn thành tốt nhiệm vụ giáo dục, đào tạo và xây dựng Nhà trường vững mạnh toàn diện; Đảng bộ trường TSVM, xây đắp nên truyền thống vẻ vang: *“Đoàn kết, sáng tạo. Khắc phục khó khăn. Kỷ luật, chính quy. Dạy tốt, học tốt”.*

\* Hiện nay Trường CĐKT Hải quân có nhiệm vụ:

- Đào tạo nhân viên chuyên môn kỹ thuật theo 03 trình độ: Cao đẳng, trung cấp, sơ cấp kỹ thuật.

- Bồi dưỡng nâng cao trình độ kỹ năng nghề cho đội ngũ nhân viên chuyên môn kỹ thuật Hải quân.

- Nghiên cứu, ứng dụng, phát triển và chuyển giao kỹ thuật - công nghệ nâng cao chất lượng, hiệu quả đào tạo; hợp tác tổ chức đào tạo theo quy định của pháp luật.

- Thực hiện các nhiệm vụ khác do Quân chủng Hải quân và Bộ Quốc phòng giao theo quy định của pháp luật.

\* Về cơ cấu tổ chức, Trường CĐKT Hải quân có: Ban Giám hiệu; 04 Phòng, 03 Ban chức năng; 09 khoa giảng viên; 05 Tiểu đoàn quản lý học viên; Hải đội 10. Hiện nay Nhà trường có 04 Tiến sỹ, 81 Thạc sỹ.

\* Hệ thống cơ sở vật chất, giảng đường, nhà ở được đầu tư xây dựng khang trang với đầy đủ tiện nghi, hiện đại.

\* Lưu lượng đào tạo của Nhà trường mỗi năm hơn 1.000 học viên, với nhiều ngành, nhiều trình độ khác nhau.

\* Để đáp ứng yêu cầu xây dựng phát triển Quân đội, Quân chủng Hải quân Việt Nam; trong những năm tới, Trường CĐKT Hải quân đã và đang:

- Tiếp tục đổi mới mạnh mẽ nội dung, chương trình, phương pháp dạy học; “chuẩn hóa”, cập nhật đầy đủ kiến thức về vũ khí trang bị mới và tiếp thu thành tựu KHCN của các nước tiên tiến.

- Tăng cường cơ sở vật chất, trang thiết bị kỹ thuật hiện đại phục vụ cho nhiệm vụ giáo dục, đào tạo; đáp ứng yêu cầu của chiến tranh công nghệ cao.

- Bồi dưỡng đội ngũ cán bộ, giảng viên, học viên phát triển toàn diện có đủ trình độ, năng lực, ngoại ngữ đáp ứng yêu cầu nhiệm vụ.

- Nỗ lực xây dựng Nhà trường chính quy, hiện đại; xứng đáng là địa chỉ giáo dục, đào tạo có uy tín của quân đội.

**2.2. Các yếu tố ảnh hưởng đến quá trình giảng dạy tại Nhà trường**

Một số trang thiết bị kỹ thuật phục vụ giáo dục đào tạo đã cũ, xuống cấp chưa đáp ứng yêu cầu đào tạo; hệ thống hạ tầng công nghệ thông tin chưa đáp ứng được quá trình dạy và học trực tuyến; thao trường, bãi tập chật hẹp ảnh hưởng đến quá trình học tập các môn học kỹ chiến thuật bộ binh.

- Phương pháp tự học, tự rèn của một số học viên còn hạn chế. Chất lượng giờ tự ôn của học viên có thời điểm chất lượng chưa cao.

- Vẫn còn một số học viên không an tâm học tập, viết đơn xin thôi học. Lễ tiết tác phong, xưng hô, chào hỏi của một số đồng chí học viên chưa tốt, còn để cán bộ chấn chỉnh, nhắc nhở.

+ Chất lượng đầu vào học viên không đồng đều, số lượng thí sinh tham gia dự tuyển ngày càng thấp, Nhà trường không được tham gia sơ tuyển đầu vào. Ngoài ra đơn vị đóng quân trên địa bàn thành phố Hồ Chí Minh chịu ảnh hưởng tác động đời sống xã hội bên ngoài dẫn đến ảnh hưởng quá trình học tập của học viên.

+ Tinh thần thái độ, động cơ học tập của một số học viên chưa cao, một số học viên còn ngại học, ngại rèn luyện.

**2.3. Thực trạng giảng dạy môn Luật biển và Quy tắc tránh va quốc tế tại Trường CĐKT Hải quân**

- Theo chức năng nhiệm vụ, Khoa Hàng hải được phân công giảng dạy môn Luật biển và Quy tắc tránh va quốc tế.

- Kết quả giảng dạy môn **Luật biển và Quy tắc tránh va quốc tế** của Khoa Hàng hải trong 2 năm học 2022 – 2023 và 2022 – 2024

- Đánh giá

**Kết luận chương 2**

**[BRO XỬ LÝ TỪ ĐÂY, CÁI TRÊN CÓ THÔNG TIN GÌ QUAN TRỌNG CÓ THỂ NOTE LẠI GIÚP MÌNH NHÉ]**

**CHƯƠNG 3: XÂY DỰNG MÔ HÌNH TÌM KIẾM VÀ TRUY VẤN THÔNG TIN LIÊN QUAN ĐẾN MÔN HỌC LUẬT BIỂN VÀ QUY TẮC TRÁNH VA QUỐC TẾ CỦA TRƯỜNG CAO ĐẲNG KỸ THUẬT HẢI QUÂN**

**3.1. Giới thiệu tổng quan thành phần của hệ thống**

**3.2. Xây dựng chi tiết các thành phần của mô hình**

**3.3 Cài đặt hệ thống**

**3.4. Đánh giá kết quả qua thực nghiệm và khảo sát người dùng**

**3.5. Giao diện người dùng**

**3.6 Các vấn đề cần chú ý trong việc ứng dụng thực tế**

**Kết luận chương 3**

**CHƯƠNG 3: XÂY DỰNG MÔ HÌNH TÌM KIẾM VÀ TRUY VẤN THÔNG TIN LIÊN QUAN ĐẾN MÔN HỌC LUẬT BIỂN VÀ QUY TẮC TRÁCH VA QUỐC TẾ CỦA TRƯỜNG CAO ĐẲNG KỸ THUẬT HẢI QUÂN**

**3.1. Giới thiệu tổng quan hệ thống**

Hệ thống được xây dựng nhằm mục đích hỗ trợ truy vấn thông tin liên quan đến môn học Luật biển và các quy tắc tránh va quốc tế. Mô hình sử dụng phương pháp tích hợp giữa tìm kiếm theo vector (vector search) và truy vấn ngữ nghĩa (semantic retrieval) thông qua kỹ thuật Retrieval Augmented Generation (RAG). Các thành phần chính của hệ thống bao gồm:

* **Cơ sở dữ liệu vector:** Dùng để lưu trữ các embedding được tạo ra từ dữ liệu văn bản (ví dụ như các câu hỏi – đáp án về luật biển)
* **Quy trình tạo embedding:** Áp dụng các mô hình chuyển đổi ngôn ngữ (như sentence-transformers) để chuyển đổi văn bản thành các vector đại diện
* **API backend:** Được triển khai bằng FastAPI, xử lý yêu cầu từ phía client, thực hiện truy vấn dựa trên hai phương án: (a) tìm kiếm chính xác (exact match) và (b) tìm kiếm ngữ nghĩa (semantic search) với Qdrant, sau đó gọi OpenAI để tạo phản hồi nếu cần
* **Giao diện người dùng (frontend):** Xây dựng bằng React với các trang như HomePage, ChatBot, FAQPage và IssuePage. Ngoài ra, việc quản lý người dùng, xác thực và định tuyến (ProtectedRoute) được xử lý qua Context API để đảm bảo bảo mật và trải nghiệm người dùng mượt mà.

**3.1.2. Phân tích các thành phần chính**

* **Xử lý dữ liệu:**  
   Dữ liệu đầu vào bao gồm file Excel chứa cặp câu hỏi – đáp án và các file văn bản (txt) được chia nhỏ thành các đoạn. Qua đó, dữ liệu được tạo thành các đối tượng Document có cấu trúc gồm nội dung trang (page\_content) và thông tin metadata (nguồn, câu hỏi, …).
* **Mô hình embedding:**  
   Sử dụng mô hình “sentence-transformers/paraphrase-multilingual-mpnet-base-v2” để tạo embedding cho từng tài liệu. Embedding giúp chuyển văn bản sang dạng vector, cho phép tính toán khoảng cách và điểm tương đồng giữa các văn bản.
* **Vector Database với Qdrant:**  
   Qdrant được chọn làm nền tảng lưu trữ vector nhờ khả năng truy vấn nhanh và linh hoạt. Hệ thống tạo collection lưu trữ các vector tương ứng với nội dung và metadata của từng tài liệu.
* **Mô hình API và xử lý truy vấn:**  
   API backend có chức năng:
  + **Tìm kiếm chính xác:** Sử dụng filter để xác định các câu hỏi trùng khớp hoàn toàn với truy vấn của người dùng.
  + **Tìm kiếm ngữ nghĩa:** Khi không có kết quả chính xác, hệ thống tính toán embedding cho câu hỏi của người dùng và truy vấn trong Qdrant để tìm các tài liệu có độ tương đồng cao. Nếu điểm số tương đồng vượt qua ngưỡng đặt trước, hệ thống có thể trả về kết quả trực tiếp hoặc kết hợp với LLM (OpenAI) để tạo ra phản hồi tự động.
* **Giao diện người dùng:**  
   Frontend được xây dựng theo triết lý component-based:
  + **NavBar.jsx:** Điều hướng giữa các trang chính với chức năng chuyển đổi theme.
  + **ProtectedRoute.jsx:** Bảo đảm các trang cần đăng nhập mới có thể truy cập được.
  + **ChatBot.jsx:** Giao diện trò chuyện với người dùng, hiển thị lịch sử hội thoại và xử lý việc gửi yêu cầu/nhận phản hồi.
  + **FAQPage.jsx, HomePage.jsx và IssuePage.jsx:** Cung cấp thông tin, trợ giúp và thu thập phản hồi của người dùng.

**3.2. Xây dựng các thành phần của mô hình**

**a. Tiền xử lý và tạo embedding**

* **Đọc dữ liệu:**  
   Sử dụng Python và Pandas để đọc dữ liệu từ file Excel (các cột “Câu hỏi” và “Đáp án”) và tách nội dung của các file text thành các đoạn dựa trên ký tự xuống dòng đôi.
* **Tạo Document objects:**  
   Mỗi tài liệu được biến đổi thành đối tượng có cấu trúc JSON với trường page\_content và metadata. Đây là bước quan trọng giúp sau này dễ dàng truy xuất thông tin khi truy vấn.
* **Embedding generation:**  
   Sử dụng thư viện langchain\_community.embeddings với mô hình HuggingFace để chuyển đổi văn bản thành các vector có kích thước cố định. Vector được tạo sẽ được lưu vào Qdrant qua phương thức from\_documents.

**b. Xây dựng vector database với Qdrant**

* **Cấu hình kết nối:**  
   Sử dụng các biến môi trường (như QDRANT\_URL, QDRANT\_API\_KEY và COLLECTION\_NAME) để thiết lập kết nối đến Qdrant.
* **Thao tác dữ liệu:**  
   Thực hiện các thao tác như kiểm tra sự tồn tại của collection, thêm vector document và thực hiện truy vấn thông qua API của Qdrant.
* **Test kết nối:**  
   Một file test (test\_qdrant\_connection.py) được xây dựng để xác nhận rằng kết nối và thao tác với Qdrant được thực hiện thành công.

**c. Xây dựng API backend bằng FastAPI**

* **Khởi tạo dịch vụ:**  
   Ứng dụng FastAPI được khởi tạo với các endpoint:
  + /v1/chat/completions: Xử lý các yêu cầu chat. Trước tiên, thực hiện tìm kiếm chính xác. Nếu không có kết quả, sử dụng semantic search rồi gọi OpenAI để sinh phản hồi.
  + /health: Endpoint kiểm tra tình trạng hoạt động của các thành phần (embeddings, Qdrant, OpenAI).
* **Xử lý truy vấn:**  
   Các hàm search\_exact và search\_semantic được triển khai để thực hiện các loại truy vấn khác nhau. Ngoài ra, quy trình xử lý và truyền dữ liệu giữa các thành phần được log bằng logger để dễ dàng theo dõi và gỡ lỗi.
* **Tích hợp LLM:**  
   Nếu không có kết quả từ truy vấn nội bộ, OpenAI API (hoặc các LLM khác) được gọi với ngữ cảnh được tổng hợp từ các tài liệu liên quan.

**d. Xây dựng giao diện người dùng với React**

* **Routing và cấu trúc trang:**  
   Sử dụng React Router, các trang chính (HomePage, ChatBot, FAQPage, IssuePage, Login) được định tuyến hợp lý.
* **Định hướng bảo mật:**  
   ProtectedRoute.jsx kiểm tra trạng thái đăng nhập của người dùng thông qua Context API và cho phép truy cập hoặc điều hướng đến trang đăng nhập nếu chưa đăng nhập.
* **Giao diện chat và lịch sử:**  
   ChatBot.jsx cung cấp giao diện chat tương tác với API backend, xử lý việc gửi yêu cầu, hiển thị phản hồi từ hệ thống và quản lý lịch sử chat. Các thành phần khác như NavBar.jsx cung cấp tính năng chuyển đổi theme, điều hướng giữa các trang và hiển thị thông tin người dùng (bao gồm cả avatar và email).
* **Tính năng feedback và báo lỗi:**  
   IssuePage.jsx cung cấp một giao diện để người dùng gửi góp ý hoặc báo lỗi, với chức năng hiển thị modal xác nhận (sử dụng EmailJS có thể được kích hoạt để gửi email thông báo).

**3.3. Đánh giá kết quả thực hiện thông qua thực nghiệm**

**3.3.1. Phương pháp đánh giá**

* **Chất lượng truy xuất thông tin:**  
   Thực hiện các truy vấn cụ thể và so sánh kết quả trả về giữa tìm kiếm chính xác và tìm kiếm ngữ nghĩa. Qua đó đánh giá độ chính xác và tính liên quan của các tài liệu được truy xuất.
* **Thời gian phản hồi:**  
   Ghi nhận thời gian phản hồi của hệ thống với các truy vấn đơn giản và truy vấn cần suy luận phức tạp. Ví dụ, một số truy vấn đơn giản trả về kết quả trong khoảng 500ms, trong khi các truy vấn phức tạp có thể mất từ 8 đến 12 giây.
* **Phản hồi người dùng:**  
   Sử dụng feedback từ người dùng thông qua giao diện IssuePage và khảo sát nội bộ để đánh giá mức độ hài lòng với câu trả lời cũng như trải nghiệm tương tác của hệ thống.

**3.3.2. Kết quả và phân tích**

* **Trích xuất kết quả tìm kiếm:**  
   Qua các thực nghiệm, các kết quả trả về từ tìm kiếm chính xác cho thấy sự khớp hoàn toàn với câu hỏi ban đầu, trong khi tìm kiếm ngữ nghĩa cung cấp các tài liệu có độ liên quan cao trong trường hợp không có kết quả chính xác.
* **Hiệu năng xử lý:**  
   Đánh giá chỉ số thời gian phản hồi cho thấy hệ thống có hiệu năng xử lý tốt đối với các truy vấn đơn giản, nhưng cũng cần tối ưu thêm ở phần truy vấn ngữ nghĩa và gọi API LLM để giảm thời gian chờ của người dùng.
* **Những bất cập và góp ý cải tiến:**  
   Một số truy vấn có thể không trả về tài liệu liên quan do dữ liệu đầu vào có tính chất đa dạng. Việc tinh chỉnh ngưỡng điểm tương đồng và cải tiến quy trình tiền xử lý dữ liệu có thể giúp nâng cao chất lượng phản hồi.

**3.4. Các vấn đề cần chú ý trong việc ứng dụng thực tế**

**3.4.1. Quản lý và tối ưu hóa dữ liệu**

* **Độ chính xác của dữ liệu:**  
   Đảm bảo rằng dữ liệu đầu vào (các cặp câu hỏi – đáp án, file văn bản) được tiền xử lý sạch sẽ, chuẩn hóa ngôn ngữ, và không có lỗi chính tả để mô hình embedding có thể hoạt động hiệu quả.
* **Cập nhật và bảo trì cơ sở dữ liệu vector:**  
   Các vector document cần được cập nhật định kỳ để phản ánh những thay đổi và bổ sung thông tin mới liên quan đến môn học.

**3.4.2. Hiệu năng và tối ưu hóa hệ thống**

* **Tối ưu hóa thời gian phản hồi:**  
   Tối ưu hóa việc tính toán embedding và truy vấn Qdrant bằng cách điều chỉnh cấu hình API và áp dụng caching đối với các truy vấn thường gặp.
* **Tối ưu hóa các cuộc gọi tới LLM:**  
   Giảm thiểu số lần gọi API từ OpenAI thông qua việc xác định rõ ràng ngưỡng trả về kết quả từ vector search trước khi gọi mô hình sinh câu trả lời.

**3.4.3. Vấn đề bảo mật và xác thực**

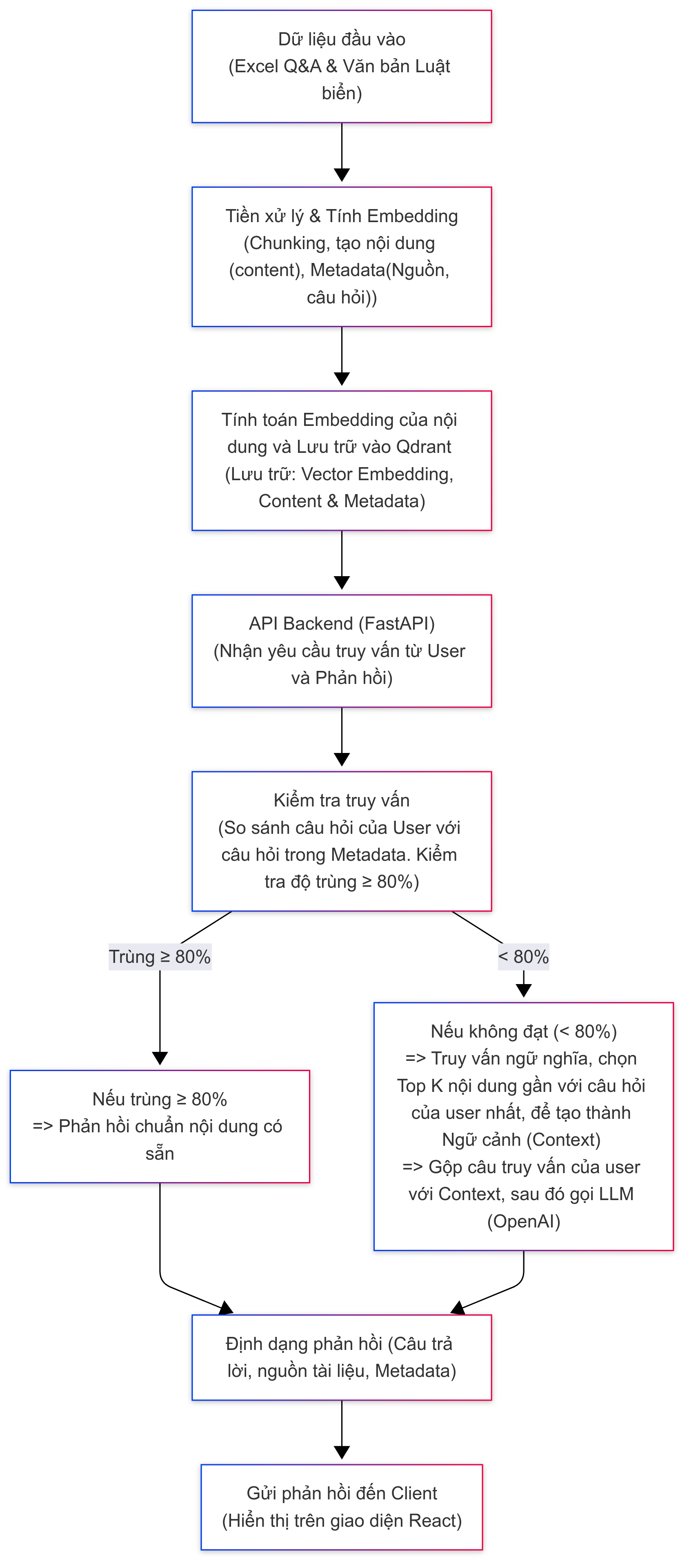
* **Quản lý thông tin người dùng:**  
   Sử dụng các giao thức bảo mật khi xử lý thông tin đăng nhập, yêu cầu bảo vệ các khóa API và thông tin nhạy cảm khác.
* **Đảm bảo an toàn khi truy vấn dữ liệu:**  
   Xây dựng các biện pháp kiểm soát truy cập tới API backend, nhất là với các endpoints quan trọng như /v1/chat/completions.

**3.4.4. Giao diện người dùng và trải nghiệm**

* **Tối ưu giao diện người dùng:**  
   Giao diện cần tương thích đa thiết bị, dễ sử dụng và phản hồi nhanh. Việc chuyển đổi theme và điều hướng trực quan góp phần nâng cao trải nghiệm người dùng.
* **Phản hồi người dùng:**  
   Hệ thống cần có cơ chế thu thập góp ý, báo lỗi nhanh chóng thông qua IssuePage, đảm bảo người dùng có thể liên hệ hỗ trợ và gửi phản hồi trực tiếp.

**Kết luận**

Chương 3 đã trình bày chi tiết quy trình xây dựng mô hình tìm kiếm và truy vấn thông tin liên quan đến môn học Luật biển và quy tắc tránh va quốc tế. Từ quá trình tiền xử lý dữ liệu, tạo embedding, xây dựng vector database trên Qdrant, tích hợp API backend với FastAPI đến xây dựng giao diện người dùng hiện đại bằng React, hệ thống đã cho thấy tiềm năng trong việc cung cấp thông tin chính xác và hỗ trợ ra quyết định cho người dùng.  
 Tuy nhiên, để ứng dụng thực tế, cần chú ý tối ưu hóa hiệu năng, cải thiện chất lượng dữ liệu và đảm bảo vấn đề bảo mật. Qua các thí nghiệm và phản hồi ban đầu, mô hình sẽ được tiếp tục cải tiến để phục vụ hiệu quả hơn cho nhu cầu tra cứu thông tin và tư vấn pháp lý của sinh viên cũng như người dùng ngoài trường.



**KẾT LUẬN – KIẾN NGHỊ**

Trong đề án này tôi đã tìm hiểu tổng quan về NLP, hệ thống chatbot, các thành phần cấu trúc và nhiệm vụ của các thành phần trong chatbot, tìm hiểu mô hình môn ngữ để ứng dụng vào xây dựng chatbot giải quyết được vấn đề đặt ra. Chatbot được xây dựng trên RAG và có thể tích hợp triển khai sử dụng trong môi trường nội bộ.

Định hướng phát triển

1. Hoàn thiện chatbot để có thể có khả năng phân tích càm xúc, nhằm đánh giá và đưa ra được câu trả lời phù hợp hơn đối với từng hoàn cảnh.

2. Tích hợp khả năng xử lý hình ảnh, giúp quá trình tra cứu của người dùng thêm trực quan và sinh động hơn.

3. Mở rộng nghiên cứu sang ứng dụng cho các môn học khác đang được giảng dạy tại nhà trường.

**TÀI LIỆU THAM KHẢO**

**PHỤ LỤC**