**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

**NGỤY HỒNG LONG**



**NGỤY HỒNG LONG**

**KHOA HỌC MÁY TÍNH ỨNG DỤNG**

ĐỒ ÁN TỐT NGHIỆP

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**TÊN ĐỀ TÀI:**

**ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG HỆ THỐNG GIÁM SÁT DẦU ONLINE CHO CÁC TRẠM BIẾN ÁP**

**KHÓA 65**

**Application of Artificial Intelligence in Online Oil Monitoring Systems for Substations**

**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

ĐỒ ÁN TỐT NGHIỆP

**CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH ỨNG DỤNG**

GIẢNG VIÊN HƯỚNG DẪN

**PGS. TS. LÊ HỒNG ANH**

SINH VIÊN THỰC HIỆN

**NGỤY HỒNG LONG**

# MỤC LỤC

[MỤC LỤC I](#_Toc183787919)

[DANH MỤC TỪ VIẾT TẮT III](#_Toc183787920)

[DANH MỤC HÌNH VẼ IV](#_Toc183787921)

[LỜI CẢM ƠN VI](#_Toc183787922)

[CHƯƠNG 1 1](#_Toc183787923)

[TỔNG QUAN VỀ ĐỀ TÀI 1](#_Toc183787924)

[1.1 . Lý do chọn đề tài 1](#_Toc183787925)

[1.2 . Mục tiêu của đề tài 2](#_Toc183787926)

[1.3 . Nội dung nghiên cứu 2](#_Toc183787927)

[1.4 . Phạm vi nghiên cứu 3](#_Toc183787928)

[1.5 . Bố cục của đồ án 5](#_Toc183787929)

[CHƯƠNG 2 6](#_Toc183787930)

[CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ 6](#_Toc183787931)

[2.1. Tổng quan về máy biến áp 6](#_Toc183787932)

[2.1.1. Khái niệm về máy biến áp 6](#_Toc183787933)

[2.1.2. Cấu tạo và nguyên lý hoạt động của máy biến áp 7](#_Toc183787934)

[2.2. Giới thiệu các chỉ số trong máy biến áp 10](#_Toc183787935)

[2.1.1. Các chỉ số trong máy biến áp 10](#_Toc183787936)

[2.1.2. TDCG - Total Dissolved Combustible Gases (Tổng các khí hòa tan) 13](#_Toc183787937)

[2.1.3. Hàm lượng nước và nhiệt độ 14](#_Toc183787938)

[2.3. Tổng quan mô hình 16](#_Toc183787939)

[2.3.1. Tìm hiểu về Transformer 16](#_Toc183787940)

[2.3.2. Biểu diễn dữ liệu đầu vào cho mô hình Transformer 17](#_Toc183787941)

[2.3.3. Mô hình hồi quy 18](#_Toc183787942)

[2.4 Môi trường phát triển 20](#_Toc183787943)

[2.5.1. Ngôn ngữ lập trình python 20](#_Toc183787944)

[2.5.2. Các thư viện hỗ trợ 20](#_Toc183787945)

[2.5.3. Các phần mềm và công cụ 23](#_Toc183787946)

[CHƯƠNG 3 26](#_Toc183787947)

[PHÂN LỚP NHẬN XÉT VỚI HỌC SÂU 26](#_Toc183787948)

[3.1. Phát biểu bài toán 26](#_Toc183787949)

[3.3. Chuẩn bị dữ liệu 27](#_Toc183787950)

[3.3.1. Phân tích dữ liệu 27](#_Toc183787951)

[3.3.2. Tiền xử lý dữ liệu 29](#_Toc183787952)

[3.3.3. Phân tách dữ liệu tập đào tạo và tập kiểm thử 29](#_Toc183787953)

[3.4. Xây dựng mô hình 30](#_Toc183787954)

[3.4.1. Thông số cấu hình máy tính để đào tạo mô hình 30](#_Toc183787955)

[3.4.2. Đào tạo mô hình 31](#_Toc183787956)

[3.5. Đánh giá mô hình 37](#_Toc183787957)

[3.6. Dự đoán với mô hình đã xây dựng được 39](#_Toc183787958)

[CHƯƠNG 4 42](#_Toc183787959)

[ỨNG DỤNG MÔ HÌNH 42](#_Toc183787960)

[4.1. Xây dựng API 42](#_Toc183787961)

[4.2. Tích hợp và triển khai API phân loại nhận xét. 44](#_Toc183787962)

[4.2.1. Phân tích Yêu cầu Hệ thống 44](#_Toc183787963)

[4.2.2. Kiến trúc Hệ thống 45](#_Toc183787964)

[4.2.3. Triển khai Backend 47](#_Toc183787965)

[4.2.4. Triển khai Frontend 48](#_Toc183787966)

[4.2.5. Tối ưu và Monitoring 49](#_Toc183787967)

[4.3. Thực nghiệm kết quả 49](#_Toc183787968)

[KẾT LUẬN 52](#_Toc183787969)

[1. Đánh giá kết quả đạt được 52](#_Toc183787970)

[2. Hướng phát triển của đề tài 53](#_Toc183787971)

[TÀI LIỆU THAM KHẢO 55](#_Toc183787972)

# DANH MỤC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Diễn giải** |
| BERT | Bidirectional Encoder Representation from Transformers, biểu diễn bộ mã hóa hai chiều từ Transformers |
| CAPTCHA | “Completely Automated Public Turing test to tell Computers and Humans Apart", phép thử Turing công cộng hoàn toàn tự động để phân biệt máy tính với người |
| CPU | Central Processing Unit, bộ xử lý trung tâm |
| DDoS | Distributed Denial-of-Service, từ chối dịch vụ phân tán |
| DOM | Document Object Model, mô hình đối tượng tài liệu |
| GPU | Graphics Processing Unit, bộ xử lý đồ họa |
| HTML | HyperText Markup Language, ngôn ngữ đánh dấu siêu văn bản |
| HTTP | Hypertext Transfer Protocol, giao thức truyền tải siêu văn bản |
| IP | Internet Protocol, giao thức Internet |
| KPI | Key Performance Indicators, chỉ số hiệu suất chính |
| MLM | Masked Language Model, mô hình ngôn ngữ che giấu |
| MultinomialNB | Multinomial Naive Bayes, thuật toán phân loại theo xác suất dựa trên định lý Bayes |
| NER | Named Entity Recognition, nhận dạng thực thể được đặt tên |
| NFC | Near-Field Communications, kết nối trường gần |
| NLP | Natural Language Processing, xử lý ngôn ngữ tự nhiên |
| NSP | Next Sentence Prediction, dự đoán câu tiếp theo |
| RAM | Random Access Memory, bộ nhớ truy cập ngẫu nhiên |
| Regex | Regular expression, biểu thức chính quy |
| URL | Uniform Resource Locator, định vị tài nguyên thống nhất |
| Wi-Fi | Wireless Fidelity, truyền tín hiệu bằng sóng vô tuyến thông qua kết nối không dây |

# DANH MỤC HÌNH VẼ

[Hình 2. 1. Cấu tạo của máy biến áp 6](#_Toc183768989)

[Hình 2. 2. Hình minh họa vai trò của máy biến áp trong lưới điện 7](#_Toc183768990)

[Hình 2. 3 Quy trình xây dựng hệ thống phân loại văn bản. **Error! Bookmark not defined.**](#_Toc183768991)

[Hình 2. 4 Kiến trúc cấp cao của BERT. 14](#_Toc183768992)

[Hình 2. 5 Mã hóa từ. 15](#_Toc183768993)

[Hình 2. 6 Phân đoạn nhúng. 16](#_Toc183768994)

[Hình 2. 7 Vị trí nhúng. 16](#_Toc183768995)

[Hình 2. 8 Biểu diễn mã hóa cuối cùng. 16](#_Toc183768996)

[Hình 2. 9 Mô hình ngôn ngữ che giấu. 20](#_Toc183768997)

[Hình 2. 10 Mô hình dự đoán mã che giấu. 21](#_Toc183768998)

[Hình 2. 11 Tinh chỉnh mô hình BERT cho bài toán phân loại văn bản. 23](#_Toc183768999)

[Hình 2. 12 Unnormalized confusion matrix và normalized confusion matrix. 24](#_Toc183769000)

[Hình 2. 13 Cách tính precision và recall cho phân loại đa lớp. 25](#_Toc183769001)

[Hình 2. 14 Công thức tính F1-score 25](#_Toc183769002)

[Hình 2. 15 Cách giao tiếp giữa máy chủ và máy khách 27](#_Toc183769003)

[Hình 2. 16Công cụ nhà phát triển trên trình duyệt Chrome. 28](#_Toc183769004)

[Hình 2. 17 Tiêu đề yêu cầu. 28](#_Toc183769005)

[Hình 2. 18 Tiêu đề phản hồi. 29](#_Toc183769006)

[Hình 2. 19 Dữ liệu cookie. 29](#_Toc183769007)

[Hình 2. 20 Mã HTML. 30](#_Toc183769008)

[Hình 2. 21 Nhúng mã JavaScript vào HTML. 31](#_Toc183769009)

[Hình 2. 22 Nguồn trang YouTube. 32](#_Toc183769010)

[Hình 2. 23 Kỹ thuật lấy phần tử. 33](#_Toc183769011)

[Hình 2. 24 Tệp robots.txt trang web Điện máy xanh. 34](#_Toc183769012)

[Hình 2. 25 Một số CAPTCHA. 34](#_Toc183769013)

[Hình 3. 1 Giao diện trang web Điện Máy Xanh 44](#_Toc180962630)

[Hình 3. 2 Thống kê lượt truy cập của website theo Similarweb 45](#_Toc180962631)

[Hình 3. 3 Thống kê nhân khẩu học của website. 46](#_Toc180962632)

[Hình 3. 4 Xác thực CAPTCHA của Cloudflare. 46](#_Toc180962633)

[Hình 3. 5 Khởi tạo selenium với hồ sơ trình duyệt. 47](#_Toc180962634)

[Hình 3. 6 Thu thập các bình luận và đánh giá cho một sản phẩm. 47](#_Toc180962635)

[Hình 3. 7 Hàm cào dữ liệu. 48](#_Toc180962636)

[Hình 3. 8 Kết quả dữ liệu được cào. 49](#_Toc180962637)

[Hình 3. 9 Cài đặt môi trường Label Studio. 52](#_Toc180962638)

[Hình 3. 10 Giao diện cấu hình tùy chỉnh với Label Studio. 52](#_Toc180962639)

[Hình 3. 11 Giao diện gán nhãn. 53](#_Toc180962640)

[Hình 3. 12 Tệp dữ liệu đã được gán nhãn. 53](#_Toc180962641)

[Hình 3. 13 Mô tả tập dữ liệu thô. 54](#_Toc180962642)

[Hình 3. 14 Dữ liệu được gán nhãn. 54](#_Toc180962643)

[Hình 3. 15 Biểu đồ phân tích độ dài của các nhận xét. 55](#_Toc180962644)

[Hình 3. 16 Thống kê số lượng từ xuất hiện nhiều nhất. 55](#_Toc180962645)

[Hình 3. 17 WordCloud của dữu liệu văn bản. 56](#_Toc180962646)

[Hình 3. 18 Kiểm tra cân bằng tập dữ liệu đã gán nhãn. 56](#_Toc180962647)

[Hình 3. 19 Tiền xử lý văn bản bình luận. 57](#_Toc180962648)

[Hình 3. 20 Tách tập dữ liệu ra thành tập đào tạo và tập kiểm thử. 57](#_Toc180962649)

[Hình 3. 21 Thông tin tập đào tạo và tập kiểm thử. 58](#_Toc180962650)

[Hình 3. 22 Kết nối Google Colab với Google Drive. 59](#_Toc180962651)

[Hình 3. 23 Sử dụng GPU để đào tạo mô hình. 59](#_Toc180962652)

[Hình 3. 24 Lấy dữ liệu từ tệp. 60](#_Toc180962653)

[Hình 3. 25 Sử dụng bộ mã hóa phoBERT. 60](#_Toc180962654)

[Hình 3. 26 Chuẩn bị dữ liệu đầu vào cho mô hình học sâu. 61](#_Toc180962655)

[Hình 3. 27 Xây dựng cấu trúc mô hình học sâu. 61](#_Toc180962656)

[Hình 3. 28 Hàm đào tạo mô hình học sâu. 62](#_Toc180962657)

[Hình 3. 29 Hàm kiểm thử mô hình. 63](#_Toc180962658)

[Hình 3. 30 Đào tạo mô hình học sâu. 63](#_Toc180962659)

[Hình 3. 31 Kết quả đào tạo mô hình học sâu từ một K-fold. 64](#_Toc180962660)

[Hình 3. 32 Kết quả đánh giá mô hình học sâu. 65](#_Toc180962661)

[Hình 3. 33 Trực quan hóa kết quả đánh giá mô hình. 65](#_Toc180962662)

[Hình 3. 34 Kết quả kiểm thử mô hình với 20 mẫu. 67](#_Toc180962663)

[Hình 3. 35 Hiệu suất của mô hình. 67](#_Toc180962664)

[Hình 3. 36 Trực quan tỷ lệ dự đoán của mô hình. 68](#_Toc180962665)

[Hình 4. 1 Định nghĩa lại mô hình. 69](#_Toc180962666)

[Hình 4. 2 Khởi tạo API Phân loại nhận xét. 70](#_Toc180962667)

[Hình 4. 3 API Dự đoán và Xử lý Tokenize. 70](#_Toc180962668)

[Hình 4. 4 Giao diện Server FastAPI trên local host. 71](#_Toc180962669)

[Hình 4. 5 Cấu trúc thư mục dự án. 72](#_Toc180962670)

[Hình 4. 6 Mô hình dữ liệu. 73](#_Toc180962671)

[Hình 4. 7 Quy trình xử lý. 73](#_Toc180962672)

[Hình 4. 8 API endpoint. 74](#_Toc180962673)

[Hình 4. 9 Cấu trúc respose 75](#_Toc180962674)

[Hình 4. 10 Giao diện trang Web TechTronic. 77](#_Toc180962675)

[Hình 4. 11 Biểu đồ thống kê. 77](#_Toc180962676)

[Hình 4. 12 Giao diện quản lý những nhận xét, đánh giá. 78](#_Toc180962677)

[Hình 4. 13 Form đánh giá, nhận xét. 78](#_Toc180962678)

# LỜI CẢM ƠN

Trước hết, tôi xin bày tỏ lòng biết ơn sâu sắc đến Trường Đại học Mỏ - Địa chất, Khoa Công nghệ thông tin và Bộ môn Khoa học máy tính đã tạo điều kiện thuận lợi để tôi thực hiện và hoàn thành đồ án tốt nghiệp với đề tài: *"Ứng dụng trí tuệ nhân tạo trong hệ thống giám sát dầu online cho các trạm biến áp."*

Tôi xin gửi lời tri ân chân thành đến các thầy cô trong Trường, trong Khoa và Bộ môn Khoa học máy tính, những người đã tận tụy giảng dạy, truyền đạt kiến thức và rèn luyện cho tôi những kỹ năng cần thiết trong suốt quá trình học tập. Những kiến thức nền tảng mà các thầy cô cung cấp chính là kim chỉ nam quan trọng, giúp tôi từng bước vượt qua các khó khăn trong quá trình nghiên cứu và hoàn thiện đồ án này.

Đặc biệt, tôi xin gửi lời cảm ơn sâu sắc đến PGS. Lê Hồng Anh, giảng viên hướng dẫn, người đã tận tâm chỉ dẫn, cung cấp tài liệu và chia sẻ những kinh nghiệm quý báu trong suốt quá trình thực hiện đồ án. Sự hướng dẫn tận tình và định hướng khoa học từ thầy không chỉ giúp tôi hoàn thành tốt nhiệm vụ đề ra mà còn mang lại cho tôi những bài học quan trọng, mở ra định hướng rõ ràng cho con đường nghiên cứu và phát triển sự nghiệp tương lai.

Tôi cũng không thể không nhắc đến sự ủng hộ, động viên to lớn từ gia đình, bạn bè và các thành viên lớp Khoa học máy tính ứng dụng 65B. Chính sự đồng hành quý giá này đã tiếp thêm động lực để tôi nỗ lực hết mình trong quá trình học tập và thực hiện đồ án tốt nghiệp.

Mặc dù đã cố gắng hết sức, nhưng do thời gian và kinh nghiệm còn hạn chế, chắc chắn đồ án này không tránh khỏi những thiếu sót. Tôi hy vọng sẽ nhận được sự góp ý quý báu từ các thầy cô và những người có chuyên môn để tiếp tục hoàn thiện hơn trong những nghiên cứu và dự án sắp tới.

Tôi xin chân thành cảm ơn!

Sinh viên thực hiện

Ngụy Hồng Long

# CHƯƠNG 1

# TỔNG QUAN VỀ ĐỀ TÀI

* 1. . Lý do chọn đề tài

Hiện nay, các trạm biến áp đóng vai trò quan trọng trong hệ thống điện quốc gia, đảm bảo cung cấp điện năng ổn định và liên tục cho các hoạt động kinh tế, công nghiệp, và sinh hoạt của con người. Trong quá trình vận hành, dầu cách điện trong máy biến áp là một trong những yếu tố quan trọng cần được giám sát thường xuyên, vì nó ảnh hưởng trực tiếp đến hiệu suất làm việc và tuổi thọ của thiết bị. Tuy nhiên, các hệ thống giám sát dầu hiện tại đang bộc lộ nhiều hạn chế cả về mặt công nghệ lẫn khả năng ứng dụng trong thực tế.

Một trong những vấn đề lớn nhất của các hệ thống giám sát dầu truyền thống là khả năng xử lý dữ liệu còn hạn chế. Trong bối cảnh lượng dữ liệu được ghi nhận từ các cảm biến ngày càng tăng lên, hệ thống cũ không thể đáp ứng tốc độ xử lý dữ liệu kịp thời. Điều này dẫn đến việc giám sát tình trạng dầu và thiết bị trở nên chậm trễ, thiếu tính hiệu quả. Hơn nữa, các hệ thống này chủ yếu cung cấp dữ liệu dạng bảng và biểu đồ, đòi hỏi người vận hành phải tự phân tích và đánh giá thủ công. Điều này không chỉ làm mất nhiều thời gian mà còn dễ dẫn đến sai sót, đặc biệt khi xử lý các trường hợp phức tạp.

Ngoài ra, một điểm yếu khác của các hệ thống hiện có là chúng không được tích hợp các chức năng cảnh báo sớm. Khi xuất hiện các biến động bất thường trong chỉ số dầu hoặc thiết bị, hệ thống không thể đưa ra cảnh báo kịp thời để người vận hành có biện pháp xử lý, dẫn đến nguy cơ xảy ra các sự cố nghiêm trọng. Hơn nữa, các hệ thống này không có khả năng phân tích dữ liệu chuyên sâu, điều này hạn chế việc phát hiện các xu hướng ẩn hoặc những dấu hiệu cảnh báo tiềm ẩn trong các thông số vận hành.

Đặc biệt, các hệ thống giám sát hiện tại chưa tận dụng được sức mạnh của trí tuệ nhân tạo (AI) để nâng cao hiệu quả giám sát và dự đoán. AI có khả năng vượt trội trong việc xử lý dữ liệu lớn, nhận diện mẫu và đưa ra dự đoán chính xác. Tuy nhiên, các hệ thống cũ chỉ tập trung vào việc lưu trữ và hiển thị dữ liệu đơn thuần mà không ứng dụng các mô hình AI để phát cảnh báo sớm, dự đoán sức khỏe thiết bị, hoặc ước tính tuổi thọ của máy biến áp dựa trên các chỉ số hiện tại. Việc thiếu vắng các tính năng này khiến cho việc bảo trì thường xuyên và dự phòng trở nên kém hiệu quả, làm tăng nguy cơ gián đoạn hệ thống điện.

Trước những hạn chế trên, đề tài “Ứng dụng trí tuệ nhân tạo trong hệ thống giám sát dầu online cho các trạm biến áp” được lựa chọn với mục tiêu cải thiện hiệu quả giám sát và bảo trì thiết bị. Đề tài hướng đến việc tích hợp các công nghệ AI hiện đại vào hệ thống giám sát dầu, giúp cải thiện tốc độ xử lý dữ liệu, cung cấp cảnh báo sớm và phân tích dữ liệu chuyên sâu. Đặc biệt, việc ứng dụng các mô hình AI sẽ giúp hệ thống không chỉ dừng lại ở việc giám sát mà còn có khả năng dự đoán tình trạng sức khỏe của máy biến áp, dự đoán tuổi thọ và hiệu suất thiết bị, từ đó hỗ trợ xây dựng chiến lược bảo trì chủ động và hiệu quả hơn.

Hệ thống đề xuất không chỉ cung cấp các cảnh báo thời gian thực mà còn có khả năng phân tích các xu hướng dài hạn, từ đó giúp người vận hành đưa ra các quyết định chính xác và kịp thời. Điều này không chỉ góp phần giảm thiểu rủi ro xảy ra sự cố mà còn nâng cao hiệu quả vận hành của toàn bộ hệ thống điện. Đồng thời, việc ứng dụng AI trong hệ thống giám sát dầu online còn mở ra cơ hội hiện đại hóa các hệ thống quản lý thiết bị, nâng cao năng lực cạnh tranh của ngành điện trong thời đại chuyển đổi số.

* 1. . Mục tiêu của đề tài

Mục tiêu của đề tài “**Ứng dụng trí tuệ nhân tạo trong hệ thống giám sát dầu online cho các trạm biến áp**” là nghiên cứu và phát triển mô hình dự đoán các chỉ số vận hành dầu trong những ngày tiếp theo, tính toán chỉ số sức khỏe (Health Index) và tuổi thọ dự kiến (Life Expectation) của thiết bị. Hệ thống đề xuất sẽ tích hợp AI để xử lý dữ liệu lớn, phát hiện các xu hướng tiềm ẩn, cung cấp cảnh báo thời gian thực và hỗ trợ người vận hành đưa ra các quyết định bảo trì chủ động. Qua đó, đề tài không chỉ hướng đến việc giảm thiểu rủi ro sự cố, tối ưu hóa hiệu suất thiết bị, mà còn góp phần xây dựng hệ thống giám sát hiện đại, nâng cao năng lực quản lý và cạnh tranh của ngành điện trong thời kỳ chuyển đổi số.

* 1. . Nội dung nghiên cứu

Để đạt được các mục tiêu đã nêu trên, đồ án có những nội dung sau:

***Về mặt lý thuyết:***

* Nghiên cứu và tổng hợp lý thuyết về các mô hình học máy, học sâu trong bài toán dự đoán chỉ số giám sát dầu.
* Bài toán dự đoán chuỗi thời gian và hồi quy.
* Các ứng dụng của bài toán dự đoán chuỗi thời gian và hồi quy.
* Mô tả một số phương pháp và mô tả từng bước triển khai bước xử lý và xây dựng mô hình học sâu, học máy hoàn chỉnh.
* Nghiên cứu xây dựng, triển khai một API từ các mô hình đã được xây dựng.
* Tích hợp API với hệ thống có sẵn để mở rộng chức năng.

***Về mặt thực hành:***

* Tạo môi trường thực hành để ứng dụng các lý thuyết vào bài toán cụ thể: Cài đặt ngôn ngữ lập trình Python, các công cụ, phần mềm và thư viện hỗ trợ để giải quyết bài toán.
* Xây dựng mô hình học máy có khả năng dự đoán chỉ số Health Index và Life Expectation, mô hình học sâu để dự đoán chỉ số các ngày tiếp theo dựa trên chỉ số các ngày trước đó.
* Chạy thử nghiệm mô hình đã xây dựng được.
* Triển khai mô hình vào API phục vụ cho hệ thống.
  1. . Phạm vi nghiên cứu

*Bài toán dự đoán chuỗi thời gian*

Bài toán dự đoán chuỗi thời gianlà một ứng dụng của học sâu trong lĩnh vực phân tích và dự đoán các dữ liệu thời gian liên tục. Mục tiêu chính của bài toán là dự đoán giá trị tương lai của một chuỗi dữ liệu thời gian dựa trên các giá trị đã biết trong quá khứ.

***Bài toán hồi quy***

Bài toán hồi quy là một vấn đề trong học máy nhằm dự đoán một giá trị liên tục từ các dữ liệu đầu vào. Mục tiêu chính của bài toán là xây dựng một mô hình học máy có thể học được mối quan hệ giữa các đặc trưng đầu vào và giá trị đầu ra, từ đó dự đoán giá trị của một biến số chưa biết dựa trên các yếu tố đã biết. Hồi quy có thể áp dụng trong nhiều lĩnh vực như dự đoán giá cổ phiếu, phân tích xu hướng thị trường, dự đoán nhu cầu tiêu thụ, hay dự báo các chỉ số sức khỏe.

***Đối tượng phân tích****:*

Các chỉ số đo lường khí và nước liên quan đến vận hành của thiết bị theo thời gian. Dữ liệu bao gồm các cột chính như sau:

Bảng 1 Bảng chỉ số đo lường khí và nước

|  |  |
| --- | --- |
| Timestamp | Thời điểm ghi nhận các chỉ số, được dùng để sắp xếp và phân tích theo chuỗi thời gian. |
| Hydrogen | Nồng độ khí hydro thường được sử dụng để phát hiện các hiện tượng phóng điện cục bộ hoặc sự cố trong dầu cách điện. |
| Methane | Nồng độ khí methane, chỉ báo cho các hiện tượng nhiệt bất thường bên trong thiết bị. |
| Acetylene | Nồng độ khí acetylene, thường liên quan đến các sự cố hồ quang hoặc nhiệt độ rất cao. |
| Ethylene | Nồng độ khí ethylene, chỉ thị sự gia nhiệt trong dầu cách điện. |
| Ethane | Nồng độ khí ethane, có liên quan đến quá trình phân hủy nhiệt của dầu cách điện. |
| Carbon Monoxide | Nồng độ khí CO, chỉ báo cho sự phân hủy nhiệt của vật liệu cách điện rắn. |
| Carbon Dioxide | Nồng độ khí CO₂, cung cấp thông tin về mức độ suy giảm của cách điện cellulose. |
| Oxygen | Nồng độ khí oxy, ảnh hưởng đến tốc độ oxy hóa dầu và độ bền của hệ thống cách điện. |
| TDCG | Tổng các khí hòa tan, là chỉ số tổng quát đánh giá tình trạng dầu cách điện. |
| Water | Hàm lượng nước trong dầu, ảnh hưởng lớn đến tính chất cách điện và tuổi thọ của thiết bị. |

***Nguồn dữ liệu:*** Dữ liệu mã nguồn mở gồm 1 năm 3 tháng dữ liệu chỉ số khí theo ngày của một trạm biến áp.

1.5 . Bố cục của đồ án

Bố cục của đồ án được trình bày với các nội dung chính như sau:

**Chương 1.** Tổng quan về đề tài

Chương này giới thiệu về mục tiêu, ý nghĩa, phạm vi và tóm lược những nội dung của đồ án.

**Chương 2.** Cơ sở lý thuyết và công nghệ

Chương này giới thiệu về xử lý ngôn ngữ tự nhiên, bài toán dự đoán chuỗi thời gian, bài toán hồi quy, xây dựng mô hình, tạo API cho mô hình, tích hợp mô hình, các công cụ, thư viện lập trình để triển khai trong bài toán cụ thể.

**Chương 3.** Tính toán chỉ số sức khỏe, vòng đời dự kiến và chỉ số tiếp theo

Chương này trình bày từng bước cụ thể để giải quyết bài toán dự đoán chuỗi thời gian, bài toán hồi quy. Bao gồm, tiền xử lý dữ liệu, khám phá dữ liệu, xây dựng mô hình, đào tạo mô hình và sử dụng mô hình để dự đoán các các chỉ số.

**Chương 4.** Ứng dụng mô hình

Chương này trình bày từng bước cách để ứng dụng mô hình vào hệ thống. Bao gồm xây dựng API cho mô hình huấn luyện.

**Kết luận**

Phần này tôi tổng kết lại các kết quả và những đóng góp mà việc thực hiện đề tài này đem lại. Ngoài ra, tổng kết những việc chưa làm được cần khắc phục, đề xuất các phương hướng nghiên cứu tiếp theo, làm cho đề tài trở lên hoàn thiện và hữu ích hơn.

# CHƯƠNG 2

# CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ

## Tổng quan về máy biến áp

### Khái niệm về máy biến áp

Máy biến áp (Transformer) là thiết bị điện từ tĩnh, được sử dụng để truyền tải và chuyển đổi năng lượng điện từ một mức điện áp này sang một mức điện áp khác thông qua hiện tượng cảm ứng điện từ. Máy biến áp hoạt động dựa trên nguyên lý truyền tải năng lượng giữa các cuộn dây thông qua từ trường biến thiên mà không cần bất kỳ bộ phận chuyển động nào.

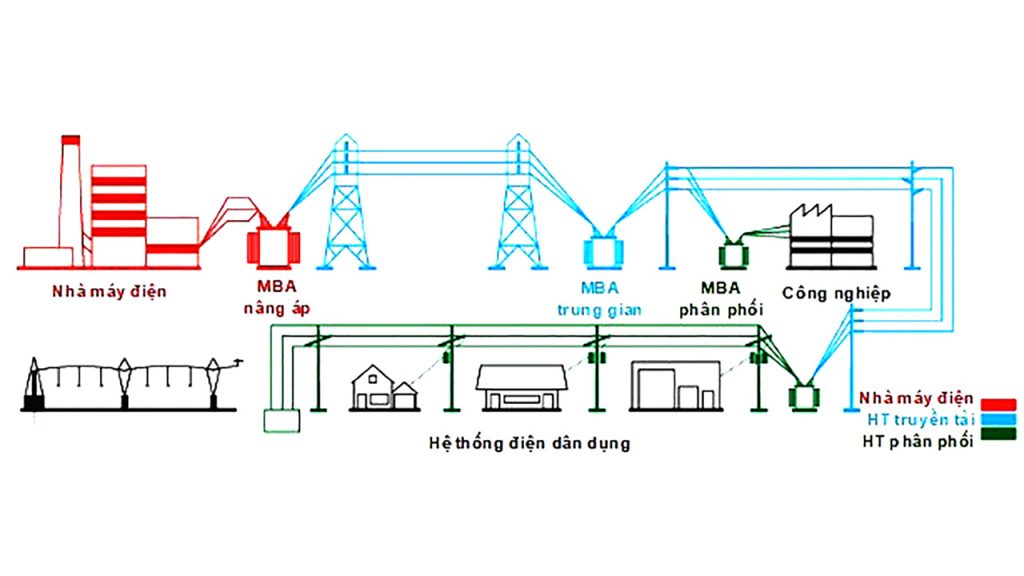


Hình 2. 1. Cấu tạo của máy biến áp

Vai trò của máy biến áp trong truyền tải và phân phối điện năng

Máy biến áp đóng vai trò quan trọng trong hệ thống điện, giúp:

1. **Truyền tải điện năng hiệu quả**: Trong lưới điện cao áp, máy biến áp tăng điện áp của dòng điện để giảm hao phí trong quá trình truyền tải.
2. **Phân phối điện năng đến các khu vực tiêu thụ**: Ở đầu ra, máy biến áp hạ điện áp về mức an toàn, phù hợp với nhu cầu sử dụng tại các khu dân cư, nhà máy, hoặc các thiết bị điện.
3. **Ổn định và bảo vệ hệ thống**: Máy biến áp giúp ngăn chặn hiện tượng ngắn mạch giữa các phần của lưới điện và cải thiện hiệu suất vận hành.



Hình 2. 2. Hình minh họa vai trò của máy biến áp trong lưới điện

### Cấu tạo và nguyên lý hoạt động của máy biến áp

**Cấu tạo cơ bản của máy biến áp**

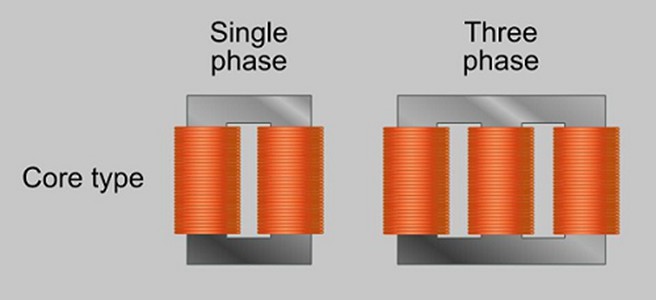
Máy biến áp được cấu tạo từ các bộ phận chính sau:

1. **Lõi thép (Core)**:

Là bộ phận quan trọng trong máy biến áp, được chế tạo từ các lá thép kỹ thuật điện có độ từ tính cao. Những lá thép này được ghép lại với nhau theo dạng mỏng và cách điện nhằm giảm tổn hao năng lượng do dòng điện xoáy (Eddy Current) gây ra. Dòng điện xoáy thường xuất hiện trong các vật liệu dẫn từ khi có từ trường biến thiên, và việc sử dụng các lá thép mỏng giúp giảm thiểu hiệu ứng này, cải thiện hiệu suất của máy biến áp.Lõi thép đóng vai trò dẫn từ, tạo ra đường từ thông giữa các cuộn dây.

Ngoài ra, lõi thép còn được thiết kế để tối ưu hóa khả năng dẫn từ, đảm bảo từ thông sinh ra từ cuộn dây sơ cấp được truyền tải hiệu quả sang cuộn dây thứ cấp. Vai trò chính của lõi thép là tạo ra đường dẫn từ thông ổn định và mạnh mẽ giữa các cuộn dây, nhờ đó giảm tổn thất từ trường ra môi trường xung quanh, đảm bảo quá trình biến đổi điện áp diễn ra hiệu quả và chính xác.

Để đáp ứng các yêu cầu vận hành, lõi thép thường được chế tạo từ các vật liệu chất lượng cao như thép silic (silicon steel), có khả năng chịu được từ trường mạnh mà không bị bão hòa từ. Thiết kế lõi thép cũng đa dạng, từ dạng hình chữ nhật cho máy biến áp nhỏ đến dạng tròn hoặc hình xuyến cho các loại máy biến áp lớn, tùy thuộc vào ứng dụng cụ thể trong truyền tải hoặc phân phối điện năng.



1. **Dây quấn (Windings)**:

là thành phần quan trọng trong cấu tạo máy biến áp, được chế tạo từ vật liệu dẫn điện như đồng hoặc nhôm. Các dây này được quấn chặt chẽ xung quanh lõi thép để tạo ra khả năng cảm ứng điện từ, qua đó truyền tải và biến đổi điện áp giữa các mạch điện. Đồng thường được sử dụng phổ biến hơn do có độ dẫn điện cao và khả năng chịu nhiệt tốt hơn nhôm, nhưng nhôm lại được lựa chọn trong một số trường hợp vì chi phí thấp hơn và trọng lượng nhẹ.

Máy biến áp bao gồm hai cuộn dây chính: cuộn sơ cấp (Primary winding) và cuộn thứ cấp (Secondary winding). Cuộn sơ cấp là nơi nhận năng lượng điện từ nguồn cấp vào, trong khi cuộn thứ cấp là nơi chuyển năng lượng điện ra để cấp cho tải. Số vòng dây quấn trong mỗi cuộn được thiết kế theo tỷ lệ biến áp mong muốn, quyết định mức độ tăng hoặc giảm điện áp của máy.

Để đảm bảo an toàn và tránh hiện tượng chạm chập giữa các vòng dây hoặc giữa các cuộn dây với lõi thép, dây quấn được bọc cách điện cẩn thận. Lớp cách điện này thường làm từ vật liệu chịu nhiệt cao như nhựa epoxy, giấy cách điện, hoặc sơn cách điện, giúp máy biến áp hoạt động ổn định trong môi trường nhiệt độ và điện áp cao.

Cách bố trí dây quấn cũng đóng vai trò quan trọng trong hiệu suất hoạt động của máy biến áp. Dây quấn có thể được quấn theo kiểu đồng tâm hoặc kiểu xen kẽ, tùy thuộc vào thiết kế cụ thể của máy. Để giảm tổn thất năng lượng và tăng độ bền cơ học, dây quấn thường được quấn theo các kỹ thuật đặc biệt nhằm tối ưu hóa phân bố từ trường và dòng điện trong máy biến áp.

Ngoài hai cuộn dây chính, một số máy biến áp còn được trang bị các cuộn dây phụ hoặc cuộn dây điều chỉnh để đáp ứng các yêu cầu kỹ thuật cụ thể, như điều chỉnh điện áp hoặc cung cấp nguồn điện phụ.

1. **Dầu cách điện (Transformer Oil)**:

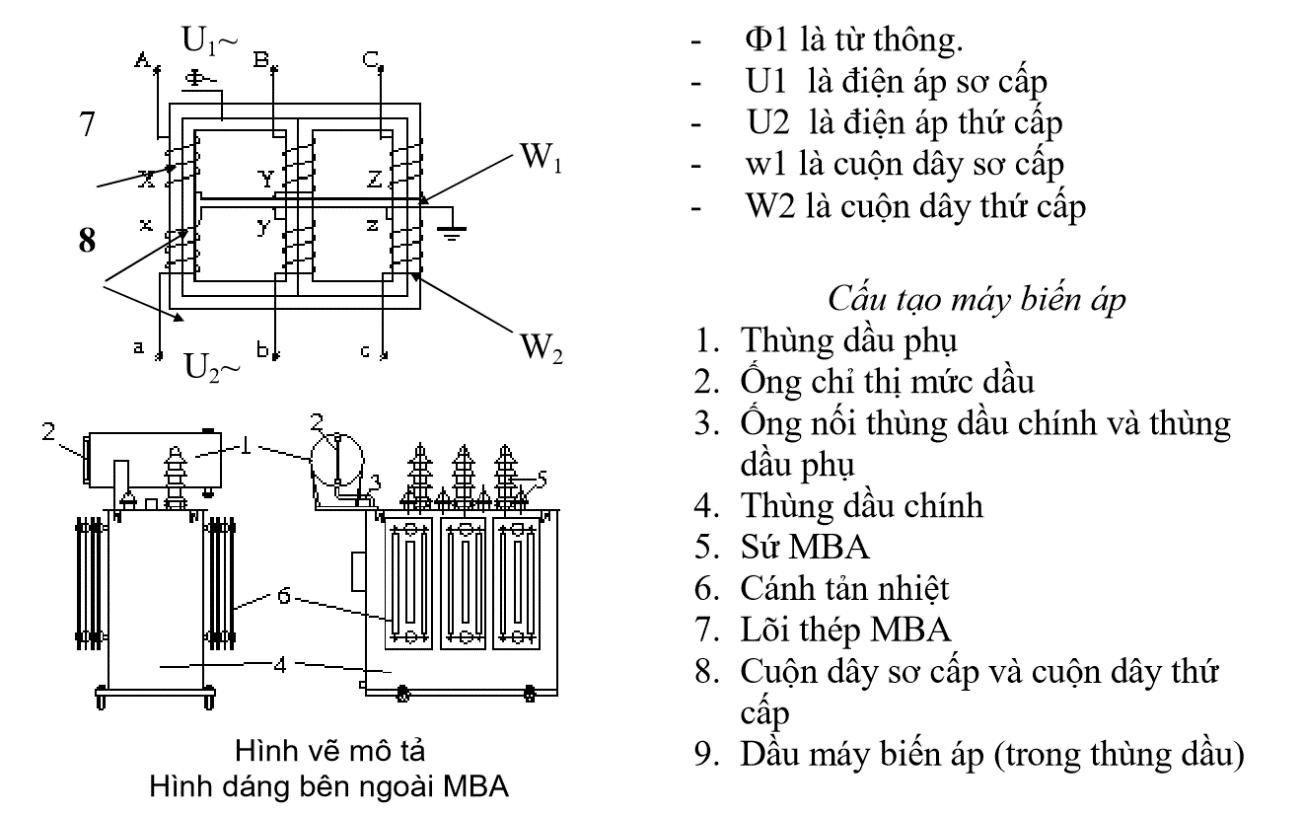
 **Dầu cách điện** là thành phần quan trọng trong máy biến áp, có chức năng chính là **làm mát** và **cách điện** cho các bộ phận bên trong. Nó truyền nhiệt từ lõi thép và cuộn dây ra bên ngoài, ngăn ngừa quá nhiệt và đảm bảo hoạt động ổn định.

 Dầu cũng đóng vai trò **cách điện**, giúp ngăn chặn phóng điện giữa các bộ phận dẫn điện như cuộn dây và vỏ máy. Dầu chất lượng cao phải đảm bảo độ tinh khiết và không chứa nước hoặc tạp chất để duy trì hiệu quả cách điện.

 Một trong những khía cạnh quan trọng của dầu cách điện là khả năng cung cấp thông tin dự đoán về tình trạng máy biến áp. **Các chỉ số hòa tan trong dầu**, như hàm lượng khí Hydro, Methane, Acetylene và CO₂, có thể được phân tích để xác định:

* **Hiện tượng phóng điện cục bộ** thông qua sự gia tăng Hydro.
* **Quá nhiệt hoặc sự cố nhiệt độ cao** qua Methane và Ethylene.
* **Sự phân hủy vật liệu cách điện** qua Carbon Monoxide và Carbon Dioxide.

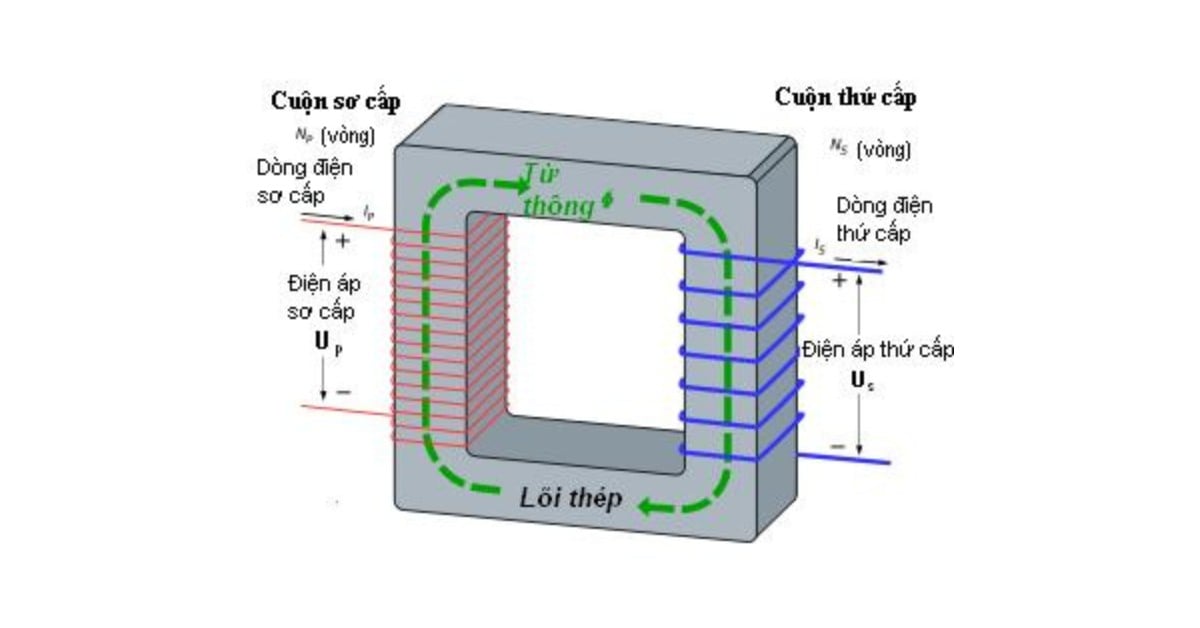
 Những dữ liệu này giúp dự đoán các sự cố tiềm ẩn, như quá nhiệt, phóng điện, hay suy giảm cách điện, từ đó đưa ra các biện pháp bảo dưỡng kịp thời để kéo dài tuổi thọ thiết bị và giảm nguy cơ hỏng hóc nghiêm trọng.



**Nguyên lý hoạt động của máy biến áp**

Máy biến áp hoạt động dựa trên hiện tượng cảm ứng điện từ, với nguyên lý cơ bản như sau:

1. **Hiện tượng cảm ứng điện từ**:
   * Khi dòng điện xoay chiều chạy qua cuộn sơ cấp, nó tạo ra từ trường biến thiên.
   * Từ trường này được dẫn qua lõi thép, tạo ra từ thông biến thiên trong cuộn dây thứ cấp.
   * Theo định luật cảm ứng điện từ Faraday, từ thông biến thiên này sẽ tạo ra một suất điện động (điện áp) trong cuộn dây thứ cấp.



1. **Quan hệ giữa số vòng dây và điện áp**:

Điện áp đầu vào và đầu ra của máy biến áp tỷ lệ thuận với số vòng dây của cuộn sơ cấp và thứ cấp, được biểu diễn bằng công thức:

Trong đó:

****  :Điện áp ở cuộn sơ cấp và thứ cấp.

 ​: Số vòng dây ở cuộn sơ cấp và thứ cấp.

## Giới thiệu các chỉ số trong máy biến áp

### Các chỉ số trong máy biến áp

Trong hệ thống máy biến áp, các chỉ số đo lường và theo dõi tình trạng của thiết bị đóng vai trò hết sức quan trọng không chỉ để đảm bảo hiệu suất vận hành liên tục và ổn định mà còn để phát hiện và dự đoán các sự cố có thể xảy ra trong quá trình sử dụng, từ đó giúp kịp thời đưa ra các biện pháp bảo trì, sửa chữa hoặc thay thế các bộ phận bị hư hỏng, nhằm nâng cao độ bền, giảm thiểu rủi ro và đảm bảo an toàn cho toàn bộ hệ thống điện.

Giám sát khí hòa tan trong dầu cách điện là một phương pháp quan trọng để đánh giá tình trạng vận hành và phát hiện sớm các sự cố tiềm ẩn trong máy biến áp. Việc này mang lại nhiều ý nghĩa quan trọng, bao gồm:

1. Phát hiện sớm các hiện tượng bất thường

Các loại khí hòa tan, như Hydrogen, Methane, Acetylene, và Carbon Monoxide, được sinh ra trong quá trình vận hành của máy biến áp khi xảy ra các hiện tượng bất thường như:

Phóng điện cục bộ (Partial Discharge).

Quá nhiệt cục bộ hoặc nhiệt độ cao bất thường.

Sự phân hủy vật liệu cách điện do nhiệt hoặc hồ quang.

Phân tích khí hòa tan giúp phát hiện sớm những hiện tượng này trước khi chúng gây ra hư hỏng nghiêm trọng.

2. Chẩn đoán nguyên nhân và loại sự cố

Mỗi loại khí có thể đại diện cho một nguyên nhân cụ thể:

Hydrogen (H₂): Dấu hiệu của phóng điện hoặc quá nhiệt nhẹ.

Methane (CH₄) và Ethylene (C₂H₄): Cho thấy sự quá nhiệt ở các mức độ khác nhau.

Acetylene (C₂H₂): Báo hiệu sự cố hồ quang hoặc nhiệt độ rất cao.

Carbon Monoxide (CO) và Carbon Dioxide (CO₂): Liên quan đến sự phân hủy cellulose, cho thấy cách điện rắn (như giấy cách điện) bị hư hại.

Phân tích các tỷ lệ khí (ví dụ: tỷ lệ CH₄/C₂H₄) giúp xác định loại sự cố một cách chính xác.

3. Dự đoán tuổi thọ và lên kế hoạch bảo trì

Giám sát khí hòa tan cho phép đánh giá mức độ suy giảm của dầu cách điện và cách điện rắn. Từ đó:

Dự đoán được tuổi thọ còn lại của thiết bị.

Xác định thời điểm và phương pháp bảo trì định kỳ hoặc thay thế các bộ phận.

4. Giảm nguy cơ sự cố nghiêm trọng và tối ưu chi phí

Phát hiện và xử lý sự cố sớm giúp ngăn chặn các hư hỏng nghiêm trọng, như chập điện hoặc cháy nổ, từ đó bảo vệ tài sản và an toàn cho hệ thống điện.

Tối ưu hóa chi phí bằng cách giảm thiểu thời gian ngừng hoạt động không mong muốn và tránh phải thay thế thiết bị trước thời hạn.

5. Đảm bảo hiệu suất và độ tin cậy của hệ thống điện

Máy biến áp là thành phần quan trọng trong hệ thống truyền tải và phân phối điện. Việc giám sát khí hòa tan đảm bảo máy biến áp hoạt động ổn định, giảm nguy cơ mất điện và nâng cao độ tin cậy của toàn bộ hệ thống.

|  |  |  |
| --- | --- | --- |
| **Chỉ số** | **Khái niệm** | **Tầm qsuan trọng** |
| **Hydrogen** | Nồng độ khí Hydro trong dầu cách điện có thể chỉ ra các hiện tượng phóng điện cục bộ (Partial Discharge) trong máy biến áp, làm giảm hiệu quả của dầu cách điện. | Tăng nồng độ hydro là dấu hiệu của sự cố phóng điện cục bộ hoặc hỏng hóc trong các bộ phận của máy biến áp, cần được giám sát để tránh sự cố nghiêm trọng. |
| **Methane** | Methane là khí gây ra từ sự phân hủy nhiệt của dầu cách điện, thường xuất hiện khi có sự cố nhiệt độ cao hoặc quá tải trong hệ thống máy biến áp có dầu. | Methane có thể chỉ ra sự cố như quá tải hoặc sự cố trong bộ cách điện, cần giám sát để phát hiện sớm các vấn đề có thể gây tổn hại cho thiết bị. |
| **Acetylene** | Acetylene là khí liên quan đến các sự cố hồ quang hoặc nhiệt độ cực kỳ cao trong máy biến áp, xuất hiện khi có sự phân hủy các vật liệu cách điện. | Nồng độ acetylene cao có thể chỉ ra sự cố nghiêm trọng trong máy biến áp, cần ngừng vận hành để tránh hư hỏng lớn. |
| **Ethylene** | Ethylene là khí sinh ra từ quá trình phân hủy dầu cách điện do gia nhiệt hoặc phản ứng hóa học, là chỉ báo cho sự gia nhiệt trong máy biến áp. | Tăng nồng độ ethylene có thể là dấu hiệu của quá nhiệt, cần kiểm tra hệ thống làm mát hoặc các bộ phận liên quan. |
| **Ethane** | Ethane là khí xuất hiện trong quá trình phân hủy nhiệt của dầu cách điện nhưng không nguy hiểm như ethylene. | Sự hiện diện của ethane có thể chỉ ra một sự cố nhiệt, cần được kiểm tra để phát hiện dấu hiệu của quá nhiệt. |
| **Carbon Monoxide** | Carbon Monoxide là sản phẩm phân hủy của vật liệu cách điện, đặc biệt là cellulose, khi bị đốt nóng. | Nồng độ CO cao có thể chỉ ra sự phân hủy vật liệu cách điện, đe dọa an toàn của thiết bị và cần được giám sát để ngăn ngừa sự cố. |
| **Carbon Dioxide** | CO₂ sinh ra từ sự phân hủy vật liệu cách điện cellulose khi chúng bị phân hủy trong máy biến áp. | Mức CO₂ cao có thể chỉ ra quá trình phân hủy vật liệu cách điện, ảnh hưởng đến tuổi thọ thiết bị, cần theo dõi để đưa ra cảnh báo sớm. |
| **Oxygen** | Oxy trong máy biến áp ảnh hưởng đến quá trình oxy hóa của dầu cách điện, làm giảm khả năng cách điện của dầu và hiệu suất của thiết bị. | Nồng độ oxy cao có thể chỉ ra quá trình oxy hóa trong dầu, làm giảm tuổi thọ hệ thống cách điện và hiệu suất của máy biến áp. Giám sát nồng độ oxy rất quan trọng. |
| **TDCG** | TDCG là tổng hợp các khí hòa tan trong dầu cách điện, dùng để đánh giá tình trạng tổng thể của dầu và máy biến áp. | TDCG là chỉ số tổng thể cho thấy tình trạng dầu, sự thay đổi đột ngột trong TDCG có thể là dấu hiệu của sự cố hoặc thay đổi trong điều kiện hoạt động của máy biến áp. |
| **Water** | Nước trong dầu cách điện có thể làm giảm tính chất cách điện của dầu, tạo điều kiện cho sự phân hủy hóa học hoặc hình thành các vết nứt. | Hàm lượng nước cao làm giảm khả năng cách điện và tăng nguy cơ sự cố như phóng điện cục bộ hoặc cháy nổ. Việc giám sát lượng nước là rất quan trọng để duy trì an toàn cho thiết bị. |

Các chỉ số này cung cấp thông tin quan trọng để:

* **Theo dõi tình trạng thiết bị theo thời gian**: Giúp phát hiện sớm các sự cố như phóng điện, gia nhiệt bất thường hoặc phân hủy vật liệu cách điện.
* **Dự đoán xu hướng trong tương lai**: Từ đó hỗ trợ lập kế hoạch bảo trì, giảm thiểu rủi ro hỏng hóc đột ngột.
* **Tính toán chỉ số sức khỏe (Health Index)**: Đánh giá mức độ suy giảm tổng thể của thiết bị.
* **Ước tính tuổi thọ (Life Expectation)**: Xác định thời gian còn lại trước khi thiết bị cần sửa chữa hoặc thay thế.

### TDCG - Total Dissolved Combustible Gases (Tổng các khí hòa tan)

TDCG (Total Dissolved Combustible Gases) là chỉ số tổng hợp của tất cả các loại khí cháy được hòa tan trong dầu cách điện, bao gồm Hydrogen (H₂), Methane (CH₄), Ethylene (C₂H₄), Acetylene (C₂H₂), Ethane (C₂H₆), và các khí khác. Đây là những khí sinh ra trong quá trình vận hành máy biến áp khi có các hiện tượng bất thường như phóng điện, quá nhiệt, hoặc phân hủy vật liệu cách điện.

TDCG không chỉ phản ánh tình trạng của dầu cách điện mà còn cung cấp thông tin tổng quan về sức khỏe của máy biến áp. Chỉ số này thường được đo lường thông qua phương pháp phân tích khí hòa tan (Dissolved Gas Analysis - DGA), một công cụ quan trọng trong công tác bảo trì và giám sát thiết bị.Quy trình các bước xây dựng hệ thống phân loại văn bản

Tầm quan trọng trong việc đánh giá tình trạng máy biến áp

Phát hiện sớm các sự cố tiềm ẩn

TDCG tăng cao bất thường là dấu hiệu cảnh báo các hiện tượng không mong muốn, chẳng hạn:

Phóng điện cục bộ khi sinh ra nhiều Hydrogen (H₂).

Quá nhiệt khi khí Methane (CH₄) hoặc Ethylene (C₂H₄) xuất hiện ở nồng độ lớn.

Sự cố hồ quang với sự hiện diện của Acetylene (C₂H₂).

Giúp phát hiện sự cố ngay từ giai đoạn đầu, tránh hư hỏng nghiêm trọng.

Đánh giá tình trạng tổng thể của máy biến áp

TDCG phản ánh tình trạng chung của dầu cách điện và hệ thống cách điện rắn (cellulose).

Sự tăng nồng độ khí hòa tan có thể chỉ ra hiện tượng suy giảm chất lượng dầu, lỗi trong cuộn dây, hoặc sự cố liên quan đến lõi thép.

Hỗ trợ đưa ra các quyết định bảo trì và vận hành

Chỉ số TDCG giúp các kỹ sư xác định mức độ ưu tiên cho việc kiểm tra và bảo trì máy biến áp.

Dựa vào dữ liệu TDCG, có thể lên kế hoạch bảo trì phù hợp, tránh các hoạt động sửa chữa khẩn cấp không cần thiết.

Dự đoán tuổi thọ và quản lý rủi ro

Khi chỉ số TDCG đạt đến một mức ngưỡng nhất định, điều đó báo hiệu các sự cố có thể xảy ra trong tương lai gần, giúp chủ động thay thế hoặc sửa chữa thiết bị.

Việc giám sát TDCG thường xuyên giúp kéo dài tuổi thọ của máy biến áp và giảm thiểu rủi ro vận hành.

Tăng cường hiệu quả kinh tế

Phát hiện và xử lý sự cố thông qua TDCG giúp giảm thời gian ngừng hoạt động, tối ưu hóa chi phí bảo trì và vận hành.

Ngăn chặn các hư hỏng lớn, tránh phải thay thế máy biến áp toàn bộ.

### Hàm lượng nước và nhiệt độ

**Tác động của nước đến tính chất cách điện của dầu**  
Nước là một yếu tố không mong muốn trong dầu cách điện của máy biến áp, bởi vì nó có thể gây ra những tác động tiêu cực lớn đối với hiệu suất hoạt động và độ bền của thiết bị. Dầu cách điện đóng vai trò rất quan trọng trong việc duy trì sự cách điện và làm mát cho máy biến áp. Khi nước hòa tan trong dầu, nó làm giảm đáng kể khả năng cách điện của dầu, tạo ra môi trường thuận lợi cho sự hình thành các vết nứt và phóng điện cục bộ. Điều này có thể dẫn đến sự phân tán dòng điện, gia tăng tổn thất năng lượng và làm hư hỏng các bộ phận quan trọng của máy biến áp.

Sự hiện diện của nước trong dầu có thể dẫn đến một số phản ứng hóa học và vật lý không mong muốn. Một trong những hậu quả chính là sự phân hủy của các vật liệu cách điện như cellulose, thường có mặt trong các cuộn dây hoặc các bộ phận khác của máy biến áp. Khi nước tiếp xúc với cellulose, nó tạo ra các sản phẩm phân hủy, chẳng hạn như axit acetic và axit formic, làm giảm chất lượng và độ bền của các vật liệu cách điện. Điều này có thể dẫn đến sự cố phóng điện hoặc hỏng hóc nặng trong máy biến áp.

Khi hàm lượng nước trong dầu vượt quá ngưỡng an toàn, khả năng chống lại hiện tượng phóng điện cục bộ của dầu giảm sút rõ rệt. Phóng điện cục bộ là một hiện tượng nguy hiểm xảy ra khi có sự tích tụ năng lượng điện trong một khu vực nhỏ của thiết bị, gây ra sự phát nổ cục bộ hoặc tổn hại nghiêm trọng. Nếu không được phát hiện và xử lý kịp thời, phóng điện cục bộ có thể lan rộng và gây hư hỏng nghiêm trọng cho các bộ phận khác của máy biến áp, dẫn đến các sự cố như cháy nổ, làm gián đoạn nguồn điện cung cấp và thậm chí là nguy cơ gây nguy hiểm cho con người.

Ngoài ra, nước trong dầu cũng có thể làm tăng nguy cơ hình thành các bọt khí trong quá trình vận hành. Các bọt khí này có thể dẫn đến sự giảm hiệu quả của việc làm mát, làm tăng nhiệt độ của máy biến áp và tạo ra điều kiện thuận lợi cho sự phân hủy của dầu cách điện. Điều này càng làm tăng thêm sự mài mòn và giảm tuổi thọ của máy biến áp.

Đặc biệt, khi hàm lượng nước trong dầu vượt mức cho phép, quá trình ăn mòn và phân hủy có thể diễn ra nhanh chóng, dẫn đến việc giảm chất lượng cách điện của toàn bộ hệ thống. Điều này khiến cho máy biến áp dễ gặp sự cố hơn và có thể làm giảm đáng kể tuổi thọ của thiết bị, gây tổn thất lớn về mặt chi phí bảo trì và sửa chữa.

**Ý nghĩa của việc giám sát nhiệt độ**

Nhiệt độ vận hành của máy biến áp là một yếu tố quyết định đến hiệu suất và độ bền của thiết bị trong suốt quá trình hoạt động. Máy biến áp phải hoạt động trong một dải nhiệt độ nhất định để đảm bảo các bộ phận của nó không bị hư hỏng. Nếu nhiệt độ vượt quá giới hạn cho phép, sẽ gây ra sự phân hủy các vật liệu cách điện, dầu cách điện và các bộ phận khác trong máy biến áp. Điều này không chỉ làm giảm khả năng cách điện của dầu mà còn có thể dẫn đến sự mất ổn định trong các hoạt động của máy biến áp, gây ra tình trạng quá nhiệt, giảm hiệu suất và thậm chí là sự cố nặng.

Việc giám sát nhiệt độ giúp phát hiện sớm các dấu hiệu của sự cố hoặc tình trạng hoạt động không bình thường. Các hệ thống giám sát nhiệt độ có thể cảnh báo khi nhiệt độ vượt quá ngưỡng an toàn, từ đó cho phép các kỹ thuật viên hoặc người vận hành can thiệp kịp thời, như giảm tải cho máy biến áp, điều chỉnh hệ thống làm mát hoặc tạm ngừng hoạt động của máy biến áp để tránh các hư hỏng nghiêm trọng. Khi có sự gia tăng nhiệt độ, hệ thống cảnh báo sẽ thông báo để người vận hành có thể đưa ra các quyết định kịp thời, giảm thiểu nguy cơ xảy ra sự cố.

**Mối quan hệ giữa nhiệt độ và hiệu suất vận hành**

Nhiệt độ có một mối quan hệ trực tiếp và chặt chẽ với hiệu suất vận hành của máy biến áp. Khi máy biến áp hoạt động, năng lượng điện được chuyển đổi và phần lớn năng lượng này chuyển hóa thành nhiệt. Nếu nhiệt không được tản đi kịp thời, sẽ làm tăng nhiệt độ của các bộ phận bên trong máy biến áp, đặc biệt là cuộn dây và lõi thép. Khi nhiệt độ của cuộn dây tăng lên, nó có thể làm giảm điện trở của dây quấn và làm tăng dòng điện dẫn qua đó, dẫn đến tình trạng quá nhiệt. Điều này không chỉ gây ra sự phân hủy các vật liệu cách điện mà còn có thể làm giảm chất lượng cách điện của dầu cách điện, khiến cho nguy cơ phóng điện cục bộ hoặc sự cố trở nên cao hơn.

Hơn nữa, nhiệt độ cao cũng ảnh hưởng trực tiếp đến khả năng làm mát của máy biến áp. Khi nhiệt độ vượt quá mức thiết kế, khả năng trao đổi nhiệt giữa các bộ phận của máy biến áp và môi trường bên ngoài sẽ bị giảm sút. Điều này làm giảm hiệu quả của hệ thống làm mát, từ đó tạo ra một vòng lặp khép kín, khi nhiệt độ ngày càng cao hơn sẽ dẫn đến sự suy giảm thêm trong khả năng làm mát của máy biến áp. Điều này càng làm cho máy biến áp dễ gặp sự cố hơn và giảm tuổi thọ của thiết bị.

Ngược lại, khi nhiệt độ duy trì ở mức thấp và ổn định, máy biến áp sẽ hoạt động hiệu quả hơn. Nhiệt độ thấp giúp giảm thiểu các tổn thất năng lượng dưới dạng nhiệt và giữ cho các vật liệu cách điện và dầu cách điện không bị phân hủy nhanh chóng. Hệ thống làm mát của máy biến áp sẽ hoạt động hiệu quả hơn khi nhiệt độ được duy trì trong phạm vi an toàn, giúp tối ưu hóa hiệu suất và kéo dài tuổi thọ của thiết bị. Việc duy trì nhiệt độ ổn định không chỉ giúp giảm chi phí bảo trì mà còn giảm thiểu nguy cơ hư hỏng và sự cố nghiêm trọng, đảm bảo hệ thống điện luôn hoạt động một cách ổn định và đáng tin cậy.

**Tầm quan trọng của việc duy trì nhiệt độ ổn định**

Duy trì nhiệt độ ổn định trong suốt quá trình vận hành máy biến áp là rất quan trọng. Một hệ thống giám sát nhiệt độ chính xác và hiệu quả sẽ giúp đảm bảo rằng máy biến áp luôn hoạt động trong phạm vi an toàn. Điều này có thể đạt được thông qua việc điều chỉnh và kiểm soát nhiệt độ của máy biến áp bằng các phương pháp làm mát như làm mát bằng không khí hoặc làm mát bằng dầu, sử dụng các quạt tản nhiệt, hệ thống làm lạnh bằng nước hoặc các thiết bị làm mát chuyên dụng.

Hệ thống giám sát nhiệt độ cũng giúp xác định các khu vực nóng trong máy biến áp, từ đó đưa ra các giải pháp làm mát hợp lý, tối ưu hóa quá trình làm mát và tránh tình trạng quá nhiệt. Việc duy trì một nhiệt độ vận hành ổn định không chỉ giúp tăng hiệu suất làm việc của máy biến áp mà còn bảo vệ hệ thống điện, giảm thiểu chi phí bảo trì và nâng cao độ tin cậy của máy biến áp trong việc cung cấp điện cho các phụ tải quan trọng.

## Tổng quan mô hình

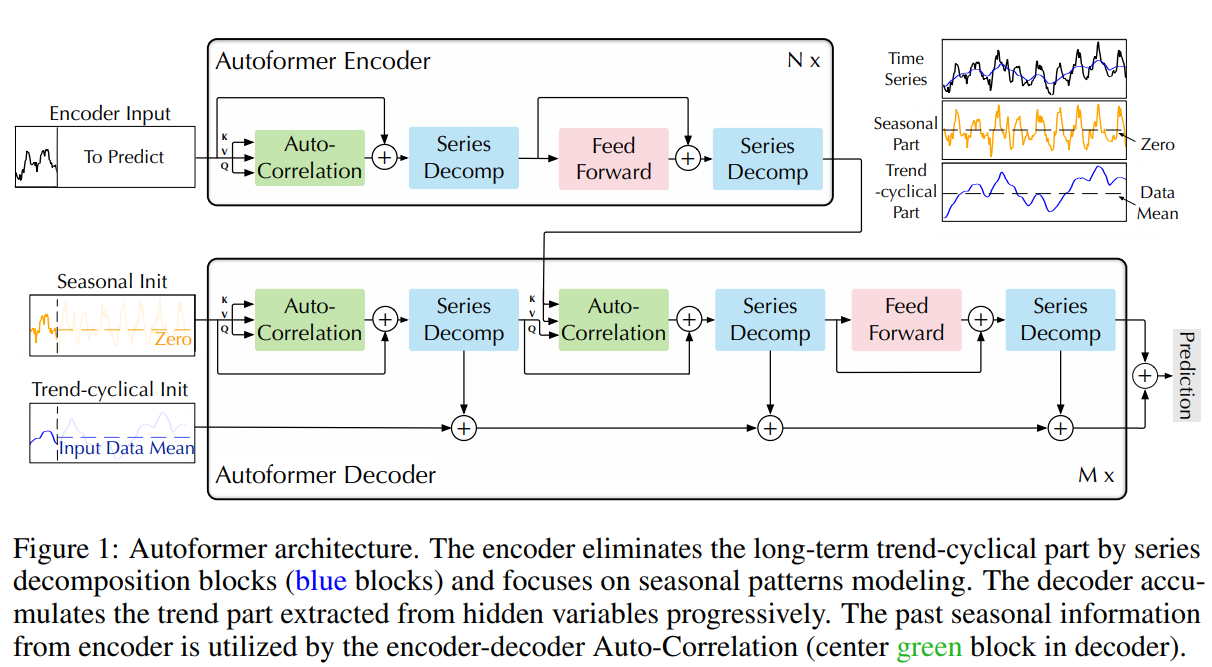
### Tìm hiểu về Transformer

Transformer là một kiến trúc mạng nơ-ron mạnh mẽ được giới thiệu trong bài báo "Attention is All You Need" của Vaswani et al. vào năm 2017. Mô hình này đã cách mạng hóa lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và nhiều lĩnh vực khác nhờ vào cơ chế **attention** (chú ý) giúp mạng học được mối quan hệ giữa các phần tử trong chuỗi dữ liệu mà không cần sử dụng cấu trúc tuần tự như các mô hình RNN hay LSTM.

Các mô hình Transformer có thể được chia thành hai loại chính: **bộ mã hóa** (Encoder) và **bộ giải mã** (Decoder), hoặc kết hợp cả hai. Một điểm mạnh của Transformer là khả năng xử lý đồng thời toàn bộ đầu vào, giúp tăng tốc độ huấn luyện và cải thiện hiệu quả.

**Cơ sở toán học**: Transformer là một kiến trúc dựa trên **Attention Mechanism**, cụ thể là **Self-Attention**, cho phép mô hình học các mối quan hệ giữa các phần tử trong chuỗi dữ liệu. Điều này đặc biệt hữu ích khi xử lý dữ liệu thời gian, như các chỉ số dầu.

**Ứng dụng trong dự đoán chuỗi thời gian**: Với khả năng học các phụ thuộc dài hạn trong dữ liệu, Transformer vượt trội hơn các mô hình truyền thống như ARIMA, LSTM trong việc dự đoán các chỉ số vận hành phức tạp và phi tuyến.



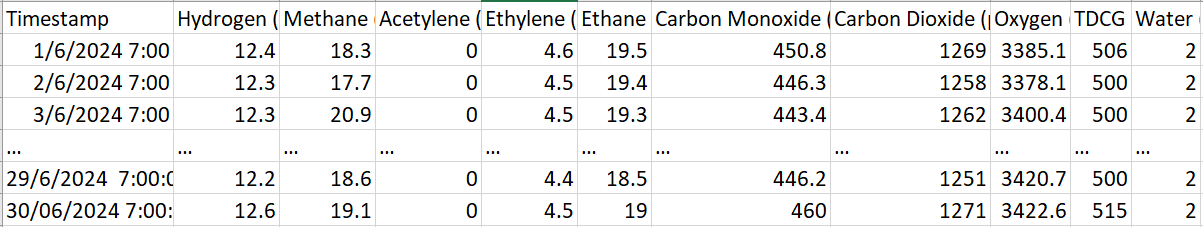
### Biểu diễn dữ liệu đầu vào cho mô hình Transformer

Khi áp dụng mô hình **Transformer** để dự đoán các chỉ số tiếp theo, ta sẽ áp dụng mô hình vào bài toán **dự báo chuỗi thời gian (Time Series Forecasting)**, dữ liệu đầu vào được chỉ định là 30 ngày dữ liệu chỉ số Timestamp, Hydrogen, Methane, Acetylene, Ethylene, Ethane, Carbon Monoxide, Carbon Dioxide, Oxygen và Water sẽ được chuyển đổi thành các nhúng đầu vào mà mô hình có thể xử lý hiệu quả. Dưới đây là cách biểu diễn dữ liệu đầu vào cho Transformer với mục đích dự báo trong bài toán chuỗi thời gian.

**Dữ liệu đầu vào**

Giả sử dữ liệu chuỗi thời gian từ 30 ngày với các chỉ số về nồng độ khí và các thông số khác của trạm biến áp, bao gồm:

|  |
| --- |
|  |
|  |
|  |
|  |

****

Trong đó, mỗi ngày dữ liệu bao gồm các giá trị của nhiều loại khí và các thông số quan trọng khác (như TDCG và Water) được ghi nhận tại một thời điểm cụ thể.

**Biểu diễn đầu vào cho Transformer**

Để đưa dữ liệu vào Transformer, ta cần chuẩn bị dữ liệu theo định dạng phù hợp để mô hình có thể học các mối quan hệ giữa các giá trị trong chuỗi thời gian. Các bước dưới đây sẽ giúp chuyển đổi dữ liệu của bạn thành đầu vào cho Transformer.

a. **Mã hóa các chỉ số thành các vector nhúng**

Mỗi chỉ số khí và thông số sẽ được mã hóa thành một vector nhúng (embedding) để biểu diễn trong không gian tính toán. Ví dụ, đối với từng loại khí (Hydrogen, Methane, etc.), ta có thể áp dụng một lớp nhúng cho từng chỉ số, tạo ra các vector đặc trưng cho mỗi giá trị của chúng.

Giả sử, ta có một cửa sổ chuỗi thời gian (ví dụ, 30 ngày) cho từng chỉ số khí, mô hình sẽ nhận đầu vào như sau:

* **Ngày 1** (Hydrogen = 12.4, Methane = 18.3, Acetylene = 0, ...).
* **Ngày 2** (Hydrogen = 12.3, Methane = 17.7, Acetylene = 0, ...).
* ...
* **Ngày 30** (Hydrogen = 12.6, Methane = 19, Acetylene = 0, ...).

Đầu vào cho mỗi ngày sẽ được chuyển thành một vector nhúng cho mỗi chỉ số khí. Cả 10 chỉ số sẽ được mã hóa thành 10 vector riêng biệt cho mỗi ngày.

b. **Vị trí nhúng (Position Embedding)**

Vì Transformer không có kiến thức về thứ tự các điểm trong chuỗi thời gian, ta sử dụng vị trí nhúng (position embedding) để cung cấp thông tin về thứ tự của các ngày. Vị trí nhúng sẽ giúp mô hình nhận biết rằng "Ngày 1" khác với "Ngày 30".

Ví dụ, đối với một chuỗi thời gian 30 ngày, ta sẽ có 30 vị trí nhúng, một cho mỗi ngày:

**P[1], P[2], P[3], ..., P[30]**

Mỗi giá trị trong vị trí nhúng sẽ được học trong quá trình huấn luyện và cung cấp thông tin về thứ tự của các dữ liệu.

c. **Tổng hợp đầu vào**

Tổng hợp các đầu vào sẽ là sự cộng của **token embedding** (vector nhúng của các chỉ số khí), **position embedding** (vector vị trí của mỗi ngày trong chuỗi thời gian), và **segment embedding** (nếu cần phân biệt các đoạn trong chuỗi). Ví dụ, tổng hợp đầu vào cho ngày 1 có thể là:

**Input\_day\_1 = E[Hydrogen] + P[1] + E[Methane] + P[1] + ... + E[Water] + P[1]**

Khi mô hình Transformer nhận các đầu vào này, nó sẽ học các mối quan hệ giữa các chỉ số khí và các thông số khác theo thời gian. Việc sử dụng các thành phần như **multi-head attention** giúp mô hình học được các mối quan hệ phức tạp và sự tương tác giữa các giá trị trong chuỗi thời gian, từ đó dự báo các giá trị tiếp theo.

### Mô hình hồi quy

Hồi quy là một nhóm các phương pháp học máy được thiết kế để dự đoán giá trị liên tục dựa trên các biến đầu vào. Trong lĩnh vực vận hành và bảo trì hệ thống điện, các mô hình hồi quy được ứng dụng rộng rãi để dự đoán chỉ số sức khỏe (Health Index - HI) và tuổi thọ dự kiến của các thiết bị điện, đặc biệt là máy biến áp. Việc đánh giá chính xác tình trạng máy biến áp không chỉ giúp phát hiện sớm các rủi ro tiềm ẩn mà còn tối ưu hóa chi phí bảo trì và đảm bảo sự ổn định của hệ thống điện.

Nghiên cứu này áp dụng nhiều mô hình hồi quy khác nhau để dự đoán chỉ số sức khỏe và tuổi thọ còn lại của máy biến áp dựa trên các thông số quan trọng được thu thập qua thời gian, bao gồm:

* **Hydrogen (ppm)**: Nồng độ khí Hydro, biểu hiện sự suy thoái của vật liệu cách điện.
* **Methane ,** **Ethylene ,** **Acetylene (ppm)**: Các khí sinh ra từ sự phân hủy nhiệt hoặc phóng điện cục bộ.
* **Carbon Monoxide (ppm)** và **Carbon Dioxide (ppm)**: Các chỉ số liên quan đến sự phân hủy của giấy cách điện.
* **Water Content (ppm)**: Hàm lượng nước trong dầu cách điện, ảnh hưởng lớn đến khả năng cách điện.
* **Oxygen (ppm)**: Hàm lượng oxy liên quan đến quá trình oxy hóa dầu.
* **TDCG (ppm)**: Tổng lượng khí hòa tan, chỉ số quan trọng trong đánh giá sức khỏe tổng thể.
* **Nhiệt độ**: Dữ liệu nhiệt độ ảnh hưởng trực tiếp đến tuổi thọ và hiệu suất máy biến áp.

Để đảm bảo tính toàn diện, nghiên cứu sử dụng các mô hình hồi quy từ tuyến tính cơ bản đến phi tuyến phức tạp, bao gồm:

* **Linear Regression, Ridge, Lasso, ElasticNet**: Các mô hình tuyến tính có tính giải thích cao, phù hợp với dữ liệu không quá phức tạp.
* **Random Forest, Gradient Boosting, AdaBoost, XGBoost, LightGBM**: Các mô hình dựa trên cây quyết định mạnh mẽ, tối ưu trong việc học các quan hệ phi tuyến giữa các chỉ số.
* **SVM (Support Vector Machine)**: Khả năng học các mô hình hồi quy phi tuyến với kernel thích hợp.
* **Neural Network**: Mô hình học sâu có khả năng học các biểu diễn phức tạp và phi tuyến từ dữ liệu.
* **K-Nearest Neighbors (KNN)**: Phương pháp dựa trên khoảng cách, thích hợp với dữ liệu nhỏ và dễ hiểu.

Ứng dụng trong dự đoán chỉ số sức khỏe và tuổi thọ của máy biến áp

Mục tiêu của nghiên cứu là xây dựng các mô hình dự đoán chỉ số sức khỏe và tuổi thọ dự kiến dựa trên dữ liệu chỉ số hàng ngày của máy biến áp. Việc áp dụng nhiều mô hình khác nhau không chỉ giúp hiểu rõ mối quan hệ giữa các thông số đầu vào và kết quả đầu ra, mà còn hỗ trợ việc đánh giá hiệu quả của từng mô hình.

Kết quả sẽ được đánh giá dựa trên các tiêu chí như:

* Độ chính xác (Mean Absolute Error - MAE, Root Mean Squared Error - RMSE).
* Khả năng tổng quát hóa trên dữ liệu chưa từng thấy (generalization).
* Tốc độ huấn luyện và triển khai.

## Môi trường phát triển

### Ngôn ngữ lập trình python

Python là một ngôn ngữ lập trình bậc cao, đa năng, được thiết kế để dễ đọc, dễ học và dễ sử dụng. Nó được phát triển bởi Guido van Rossum vào năm 1991 và đã trở thành một trong những ngôn ngữ lập trình phổ biến nhất trên thế giới.

Dưới đây là một số đặc điểm cơ bản về Python:

• Mã nguồn mở: Python là ngôn ngữ mã nguồn mở, nghĩa là nó miễn phí và có thể được sử dụng bởi bất kỳ ai.

• Dễ hiểu: Cú pháp của Python sử dụng khoảng trắng (thụt lề) để định nghĩa các khối mã, không sử dụng dấu ngoặc như nhiều ngôn ngữ lập trình khác.

• Đa năng: Python có thể được sử dụng cho nhiều mục đích khác nhau, bao gồm phát triển ứng dụng web, phân tích dữ liệu, trí tuệ nhân tạo, học máy, tự động hóa công việc, và nhiều ứng dụng khác.

• Thư viện và khung lập trình: Python có một hệ sinh thái mạnh mẽ của thư viện và khung lập trình. Django và Flask cho phát triển web, NumPy và Pandas cho xử lý dữ liệu, TensorFlow và PyTorch cho học máy và trí tuệ nhân tạo, và nhiều thư viện khác.

• Cộng đồng lớn: Python có một cộng đồng rất lớn và đa dạng, với hàng triệu lập trình viên trên khắp thế giới. Nghĩa là có rất nhiều tài liệu, và cộng đồng để hỗ trợ người mới học và những người đã có kinh nghiệm.

• Sự phát triển liên tục: Python liên tục phát triển, với việc ra mắt các phiên bản mới với nâng cấp và cải tiến định kỳ. Python 3.x là phiên bản được khuyến nghị cho mọi dự án mới, trong khi Python 2.x đã bị ngừng hỗ trợ.

### Các thư viện hỗ trợ

**Thư viện numpy**

NumPy là gói cơ bản dành cho tính toán khoa học bằng Python. Nó là một thư viện Python cung cấp một đối tượng mảng đa chiều và một loạt các thao tác nhanh trên mảng, bao gồm toán học, logic, thao tác hình dạng, sắp xếp, chọn, đại số tuyến tính, các phép toán thống kê.

Dưới đây là một số tính năng của numpy:

• Tốc độ xử lý: NumPy xử lý mảng đa chiều có hiệu suất cao, giúp thực hiện các phép toán số học nhanh chóng và dễ dàng trên dữ liệu đa chiều, như vectơ, ma trận và tensor.

• Phân tích dữ liệu: NumPy thường được sử dụng trong phân tích dữ liệu và khoa học dữ liệu để thực hiện các phép toán thống kê, tính toán trung bình, độ lệch chuẩn, và phương sai trên dữ liệu.

• Thực hiện nhân tích chập: NumPy là một công cụ mạnh mẽ để thực hiện tích chập trong mạng nơ-ron.

• Tính toán đại số tuyến tính: NumPy cung cấp các hàm để giải các vấn đề đại số tuyến tính như giải phương trình tuyến tính và tính toán giá trị riêng, ma trận nghịch đảo và nhiều phép toán khác.

• Tính toán đại số đạo hàm: NumPy có thể được sử dụng để tính đạo hàm và tích phân, đặc biệt trong các vấn đề tối ưu hóa và học máy.

**Thư viện pandas**

Pandas là một thư viện mã nguồn mở theo giấy phép BSD cung cấp các công cụ phân tích dữ liệu và cấu trúc dữ liệu hiệu suất cao, dễ sử dụng cho ngôn ngữ lập trình Python.

Dưới đây là một số ứng dụng của pandas:

• Đọc và ghi dữ liệu: Pandas cho phép đọc dữ liệu từ nhiều nguồn khác nhau như tệp CSV, Excel, SQL và nhiều định dạng dữ liệu khác. Nó cũng cho phép ghi dữ liệu vào các định dạng này.

• Khám phá và tiền xử lý dữ liệu: xem dữ liệu, kiểm tra giá trị trống, loại bỏ trùng lặp, và các thao tác khác để chuẩn bị dữ liệu cho việc phân tích hoặc học máy.

• Lọc và truy vấn dữ liệu: lọc dữ liệu dựa trên các điều kiện, thực hiện truy vấn và tổng hợp dữ liệu theo nhiều cách khác nhau.

• Thao tác với dữ liệu đa chiều: hỗ trợ các phép biến đổi dữ liệu đa chiều để cấu trúc lại dữ liệu cho mục đích phân tích cụ thể.

• Thống kê và tính toán: tính toán các thống kê cơ bản như trung bình, trung vị, mốt và các phép toán thống kê phức tạp hơn trên dữ liệu.

• Kết hợp dữ liệu: kết hợp dữ liệu từ nhiều nguồn thông qua các phép toán như merge, join và concat.

**Thư viện matplotlib**

Matplotlib là một thư viện tạo ra các biểu đồ và đồ thị đa dạng. Nó được sử dụng rộng rãi trong lĩnh vực khoa học dữ liệu, phân tích số liệu, và trực quan hóa dữ liệu.

Dưới đây là một số đặc điểm và ứng dụng của matplotlib:

• Tạo biểu đồ: biểu diễn dữ liệu số học bằng cách tạo các biểu đồ dựa trên dữ liệu, ví dụ như biểu đồ đường, biểu đồ cột, biểu đồ phân phối, biểu đồ điểm, và nhiều loại biểu đồ khác.

• Trực quan hóa khoa học dữ liệu: biểu đồ dùng để biểu thị dữ liệu đo lường từ các thí nghiệm khoa học, dữ liệu về tài chính, dữ liệu địa lý, và nhiều loại dữ liệu khác.

• Hỗ trợ nhiều định dạng đầu ra: lưu biểu đồ được tạo ra bằng matplotlib dưới nhiều định dạng khác nhau.

• Biểu đồ tùy chỉnh: cho phép tùy chỉnh hầu hết các khía cạnh của biểu đồ, bao gồm tiêu đề, nhãn, màu sắc, kích thước, chú thích, vùng đánh dấu, và nhiều thuộc tính khác.

**Thư viện seaborn**

Seaborn là thư viện trực quan hóa dữ liệu dựa trên matplotlib. Nó cung cấp một giao diện cấp cao để vẽ đồ họa thống kê hấp dẫn và giàu thông tin.

Dưới đây là một số đặc điểm của seaborn:

• Dễ sử dụng: tạo ra các biểu đồ phổ biến như biểu đồ đường, biểu đồ điểm, biểu đồ hộp, và heatmap một cách dễ dàng và nhanh chóng.

• Tùy chỉnh dễ dàng: có thể tùy chỉnh nhiều khía cạnh của biểu đồ, bao gồm màu sắc, kiểu dáng, tiêu đề và chú thích để làm cho biểu đồ phù hợp với nhu cầu.

• Hỗ trợ đa biểu đồ: hỗ trợ việc tạo ra các biểu đồ kết hợp như Pair Plots, Joint Plots và Facet Grids để hiển thị mối tương quan giữa các biến trong dữ liệu.

**Thư viện scikit-learn**

Scikit learn (hoặc sklearn) bao gồm các công cụ đơn giản và hiệu quả để phân tích dự đoán dữ liệu. Mọi người đều có thể truy cập và có thể sử dụng lại trong nhiều bối cảnh khác nhau. Nó được xây dựng trên NumPy, SciPy và matplotlib, và có thể sử dụng về mặt thương mại - giấy phép BSD.

Dưới đây là một số đặc điểm của scikit-learn:

• Hỗ trợ cho nhiều loại mô hình: cung cấp một loạt các thuật toán học máy và khám phá dữ liệu, bao gồm hồi quy tuyến tính, phân loại, gom cụm, giảm chiều dữ liệu và nhiều thuật toán khác.

• Hỗ trợ cho việc tiền xử lý dữ liệu: cung cấp nhiều công cụ để tiền xử lý dữ liệu như chuẩn hóa, mã hóa biến phân loại, xử lý dữ liệu thiếu, và trích xuất đặc trưng.

• Hiệu suất tốt: Scikit-learn được xây dựng trên thư viện NumPy và SciPy, giúp tối ưu hóa hiệu suất và khả năng mở rộng.

• Cộng đồng lớn và tài liệu phong phú: Scikit-learn có một cộng đồng sử dụng rộng lớn và nhiều tài liệu giúp người dùng học và làm việc với nó.

**Thư viện pytorch**

PyTorch một thư viện mã nguồn mở và mạnh mẽ được phát triển bởi Facebook, để hỗ trợ việc xây dựng và đào tạo các mạng nơ-ron.

Dưới đây là một số tính năng của pytorch:

• Hỗ trợ bộ xử lý đồ họa (GPU): được tối ưu hóa để hoạt động trên GPU, giúp tăng tốc quá trình đào tạo mô hình. Nó cung cấp các phép toán được tối ưu hóa để sử dụng GPU hiệu quả và hỗ trợ nhiều kiến trúc GPU khác nhau.

• Cộng đồng lớn và phong phú: có một cộng đồng người dùng và phát triển đông đảo, với nhiều tài liệu và thư viện bổ sung cho việc xây dựng các ứng dụng học máy và trí tuệ nhân tạo.

• Mạnh mẽ trong việc xây dựng mô hình mạng và học sâu: PyTorch cung cấp các công cụ mạnh mẽ để xây dựng và tinh chỉnh các mạng phức tạp. Ta có thể dễ dàng tạo ra các mô hình mạng tuỳ chỉnh và kiểm soát chúng một cách linh hoạt.

### Các phần mềm và công cụ

**Phần mềm visual studio code**

Visual Studio Code là trình soạn thảo mã phổ biến được phát triển bởi Microsoft. Nó nhẹ nhưng mạnh mẽ cho cả Windows, macOS và Linux. Nó đi kèm với hỗ trợ tích hợp cho JavaScript, TypeScript và Node.js, đồng thời có kho phần mở rộng phong phú cho các ngôn ngữ khác, chẳng hạn như C++, C#, Java, Python.

Dưới đây là các tính năng nổi bật về Visual Studio Code:

• Hoạt động như một môi trường phát triển tích hợp: Cung cấp hoàn thành mã thông minh, đề xuất và định dạng mã tự động dựa trên ngữ cảnh cụ thể của ngôn ngữ.

• Khả năng tùy chỉnh: Tùy biến cao để phù hợp với sở thích và nhu cầu của từng cá nhân. Nó cho phép người dùng thay đổi chủ đề, tổ hợp phím và cài đặt. Ngoài ra, các nhà phát triển có thể tạo các tiện ích mở rộng hoặc cài đặt các tiện ích mở rộng được phát triển để thêm các chức năng.

• Tích hợp ứng dụng đầu cuối: Tính năng này cho phép chạy các lệnh, xây dựng dự án và thực hiện các tác vụ khác nhau mà không cần chuyển sang một ứng dụng đầu cuối riêng biệt.

• Gỡ lỗi: Cung cấp trải nghiệm gỡ lỗi mạnh mẽ với sự hỗ trợ cho các ngôn ngữ khác nhau. Nó cho phép thiết lập các điểm ngắt, duyệt qua mã, kiểm tra các biến và xử lý các ngoại lệ, giúp xác định và khắc phục sự cố trong mã dễ dàng hơn.

• Kiểm soát phiên bản: Tích hợp git cho phép quản lý mã nguồn trực tiếp. Nó cung cấp các tính năng như quản lý nhánh và giải quyết xung đột, giúp làm việc thuận tiện với các hệ thống kiểm soát phiên bản.

• Chia sẻ trực tiếp: Chia sẻ phiên viết mã với những người khác, cho phép cộng tác theo thời gian thực, đánh giá mã.

• Cộng đồng và tài liệu: Có một cộng đồng lớn và tích cực các nhà phát triển. Tài liệu chính thức cung cấp thông tin chi tiết về các tính năng khác nhau, tùy chọn cấu hình và phát triển tiện ích mở rộng, cùng với các hướng dẫn để giúp người dùng tận dụng tối đa trình chỉnh sửa.

**Framework FastAPI**

FastAPI là một framework hiện đại và nhanh cho việc xây dựng các ứng dụng API trong Python. Nó được thiết kế để dễ sử dụng, với cú pháp đơn giản và khả năng tự động tạo tài liệu API, giúp tăng tốc độ phát triển ứng dụng.

FastAPI hỗ trợ việc tạo ra các RESTful API một cách nhanh chóng nhờ vào các tính năng như:

• Hiệu suất cao: Được xây dựng dựa trên ASGI và Starlette, FastAPI có tốc độ gần bằng với các framework như NodeJS và Go.

• Tự động tạo tài liệu API: Tích hợp Swagger UI và ReDoc, giúp việc kiểm thử và tương tác với API trở nên đơn giản.

• Hỗ trợ type hints của Python: Sử dụng các tính năng type hints để tự động kiểm tra đầu vào và đầu ra của API, giúp mã nguồn rõ ràng và dễ bảo trì.

FastAPI là một lựa chọn tốt cho việc xây dựng api cho mô hình để ứng dụng vào website.

# CHƯƠNG 3

# PHÂN LỚP NHẬN XÉT VỚI HỌC SÂU

## 3.1. Phát biểu bài toán

Trong nội dung của nghiên cứu này, tôi tập trung nghiên cứu, tìm hiểu và ứng dụng các cơ sở lý thuyết vào bài toán dự đoán các chỉ số hoạt động của máy biến áp. Tập dữ liệu thực tế bao gồm các chỉ số đo lường liên tục được thu thập hàng ngày tại các trạm biến áp, như Hydrogen , Methane , Acetylene , Ethylene , Carbon Monoxide , Carbon Dioxide , Water Content , TDCG , Oxygen và nhiệt độ.

Mục tiêu của bài toán bao gồm:

* **Phát hiện sớm các dấu hiệu bất thường** thông qua dự đoán chính xác các chỉ số hoạt động của máy biến áp.
* **Hỗ trợ công tác bảo trì dự đoán (predictive maintenance)**, giúp kéo dài tuổi thọ thiết bị, tối ưu hóa hiệu suất vận hành và giảm thiểu chi phí bảo trì.
* **Xây dựng công cụ hỗ trợ kỹ sư vận hành**, cung cấp các dự báo đáng tin cậy về các chỉ số vận hành trong tương lai, từ đó cải thiện hiệu quả hoạt động của hệ thống điện.

Để đạt được các mục tiêu trên, bài toán được chia thành các nhiệm vụ nhỏ như sau:

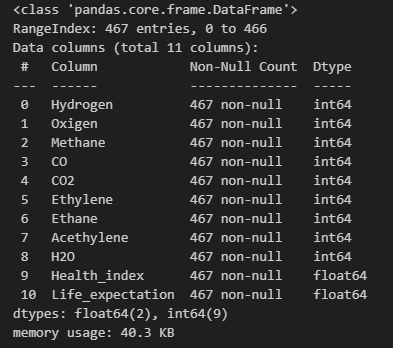
1. **Thu thập dữ liệu:** Tập hợp các chỉ số đo lường của máy biến áp từ Tổng công ty truyền tải điện 1, và dữ liệu mã nguồn mở trên nền tảng Kaggle
2. **Tiền xử lý và phân tích dữ liệu:** Làm sạch, chuẩn hóa dữ liệu và phân tích các đặc điểm quan trọng, bao gồm xu hướng, mối quan hệ giữa các chỉ số và các yếu tố ngoại cảnh (nếu có).
3. **Xây dựng mô hình dự đoán:** Sử dụng các mô hình học máy và học sâu để dự đoán các chỉ số hoạt động của máy biến áp trong tương lai.
4. **Đánh giá và tối ưu hóa mô hình:** So sánh hiệu suất của nhiều mô hình để chọn ra phương pháp dự đoán chính xác và hiệu quả nhất.
5. **Triển khai mô hình:** Áp dụng mô hình được xây dựng để dự đoán các chỉ số mới, phục vụ công tác vận hành và bảo trì.

Phát biểu này cung cấp một khuôn khổ rõ ràng cho việc tiếp cận bài toán dự đoán chỉ số của máy biến áp, đồng thời đảm bảo các mục tiêu nghiên cứu và ứng dụng thực tiễn được thực hiện hiệu quả.

## 3.3. Chuẩn bị dữ liệu

### 3.3.1. Phân tích dữ liệu

Mô tả dữ liệu : Bộ dữ liệu gồm **1 năm 3 tháng** thu thập các chỉ số từ một trạm biến áp, như **Hydrogen (ppm)**, **Methane (ppm)**, **Acetylene (ppm)**, và các thông số khác, được sử dụng để xây dựng các mô hình dự đoán chỉ số sức khỏe, vòng đời ước tính và dự báo dữ liệu ngày tiếp theo.



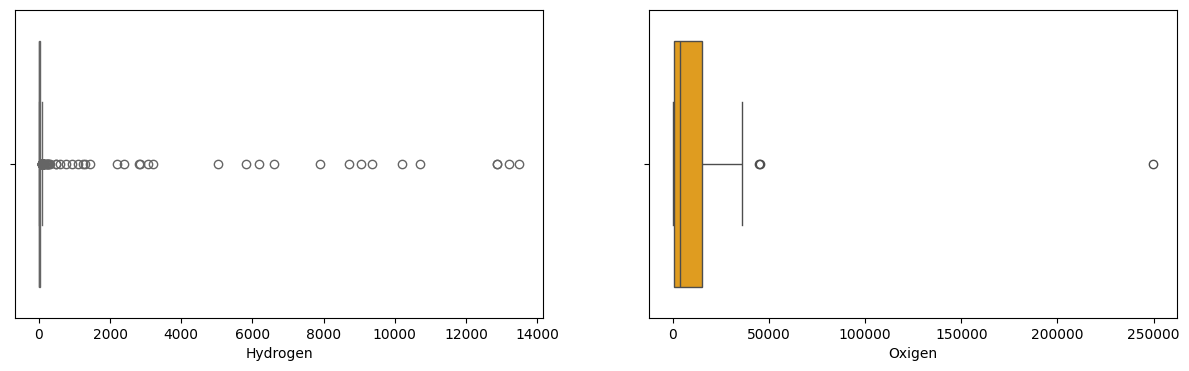
**Đánh giá dữ liệu về khối lượng và chất lượng**

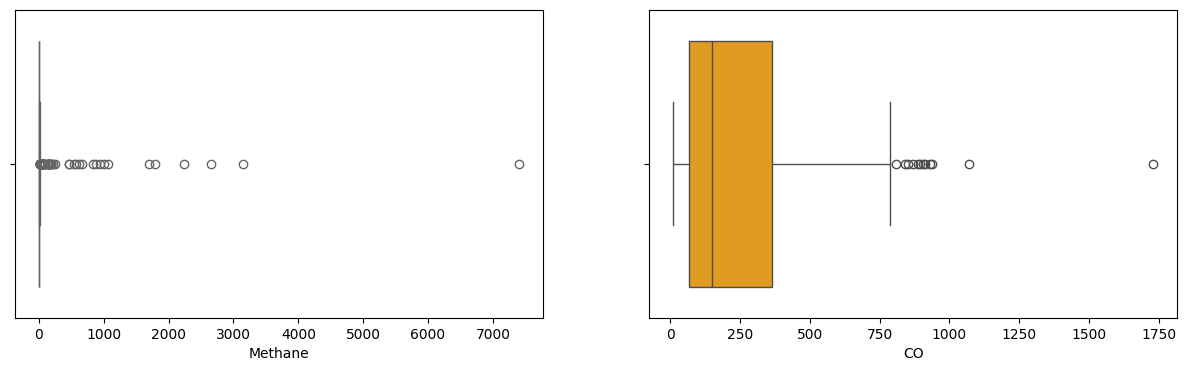
* **Khối lượng dữ liệu:**
  + Với 467 ngày dữ liệu, đây là một bộ dữ liệu có kích thước trung bình. Tuy nhiên, đối với bài toán dự đoán chuỗi thời gian, dữ liệu này có thể hạn chế nếu chứa các xu hướng phức tạp hoặc sự bất thường không phổ biến.
  + Đối với các mô hình học sâu như transformer, yêu cầu dữ liệu lớn hơn để tận dụng được toàn bộ khả năng của mô hình. Tuy nhiên, với các thiết lập nhỏ gọn (giảm số lớp và tham số), bộ dữ liệu này vẫn có thể được sử dụng hiệu quả.
* **Chất lượng dữ liệu:**
  + Dữ liệu cần được kiểm tra các yếu tố như **nhiễu**, **các giá trị bất thường**.
  + Cần xác định mối quan hệ giữa các chỉ số, loại bỏ các chỉ số không liên quan hoặc dư thừa, nhằm tránh hiện tượng đa cộng tuyến (multicollinearity).

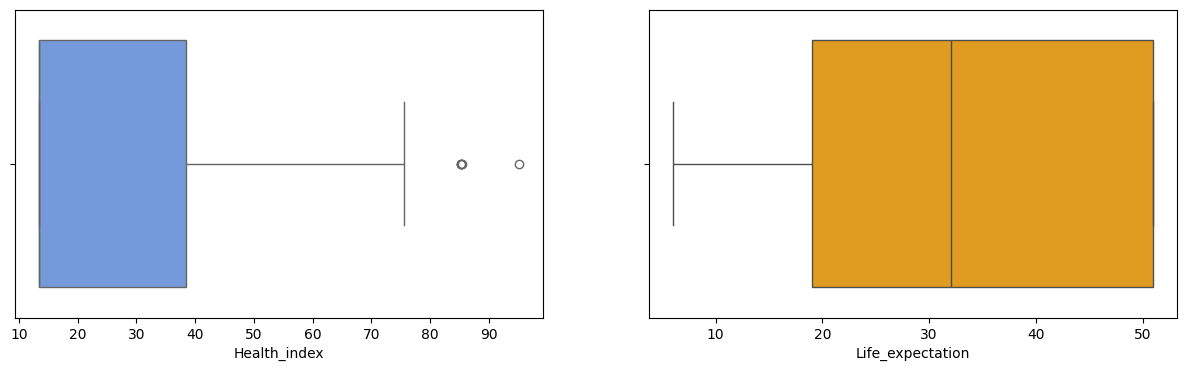
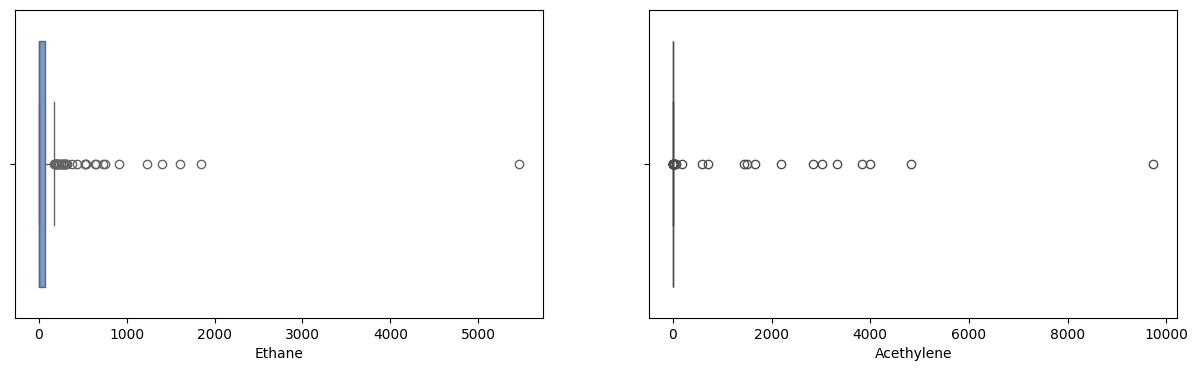
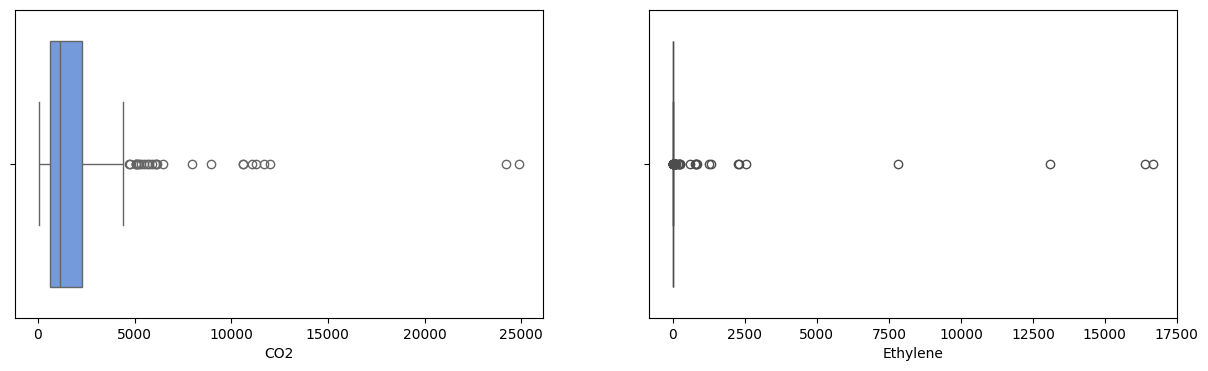
**Độ phù hợp cho các bài toán dự đoán**

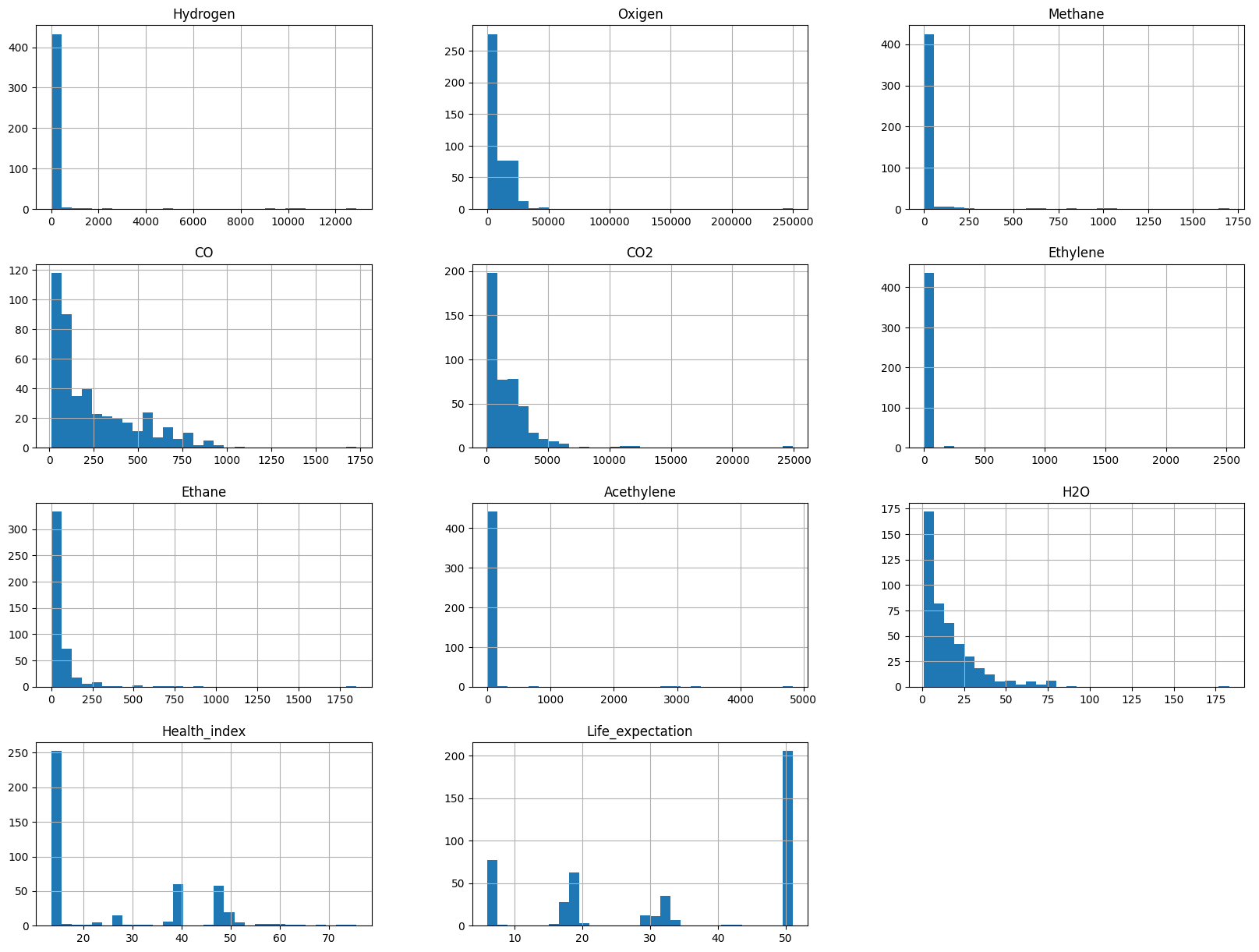
* **Dự đoán chỉ số sức khỏe và vòng đời ước tính:**
  + **Ưu điểm:**
    - Dữ liệu bao gồm các chỉ số hóa học và vật lý có mối liên hệ trực tiếp đến tình trạng sức khỏe và tuổi thọ của máy biến áp. Điều này giúp mô hình học máy dễ dàng tìm ra các đặc trưng quan trọng và mối quan hệ giữa các yếu tố.
    - Với số lượng biến đầu vào đa dạng, các mô hình như Random Forest, Gradient Boosting hoặc XGBoost có thể hoạt động tốt nhờ khả năng xử lý dữ liệu phi tuyến và xác định các đặc trưng quan trọng.
  + **Thách thức:**
    - Dữ liệu chỉ thu thập từ một trạm biến áp, nên có thể thiếu tính đa dạng trong các mẫu dữ liệu, hạn chế khả năng tổng quát hóa của mô hình khi áp dụng cho các trạm khác.
    - Chu kỳ thay đổi của các chỉ số có thể bị ảnh hưởng bởi các yếu tố bên ngoài không được ghi lại, gây khó khăn trong việc dự đoán vòng đời chính xác.
* **Dự đoán dữ liệu ngày tiếp theo bằng transformer:**
  + **Ưu điểm:**
    - Mô hình transformer có khả năng xử lý tốt các chuỗi thời gian, đặc biệt khi dữ liệu có tính phụ thuộc dài hạn. Nó có thể nắm bắt mối quan hệ phức tạp giữa các chỉ số và xu hướng thời gian.
  + **Thách thức:**
    - Dữ liệu 467 ngày có thể không đủ lớn để tận dụng toàn bộ sức mạnh của transformer. Cần áp dụng kỹ thuật nâng cao để mở rộng lượng mẫu cho mô hình học.
    - Tính liên tục và thời vụ trong dữ liệu cần được phân tích kỹ để đảm bảo rằng các mẫu huấn luyện mang tính đại diện và không gây nhiễu cho mô hình.

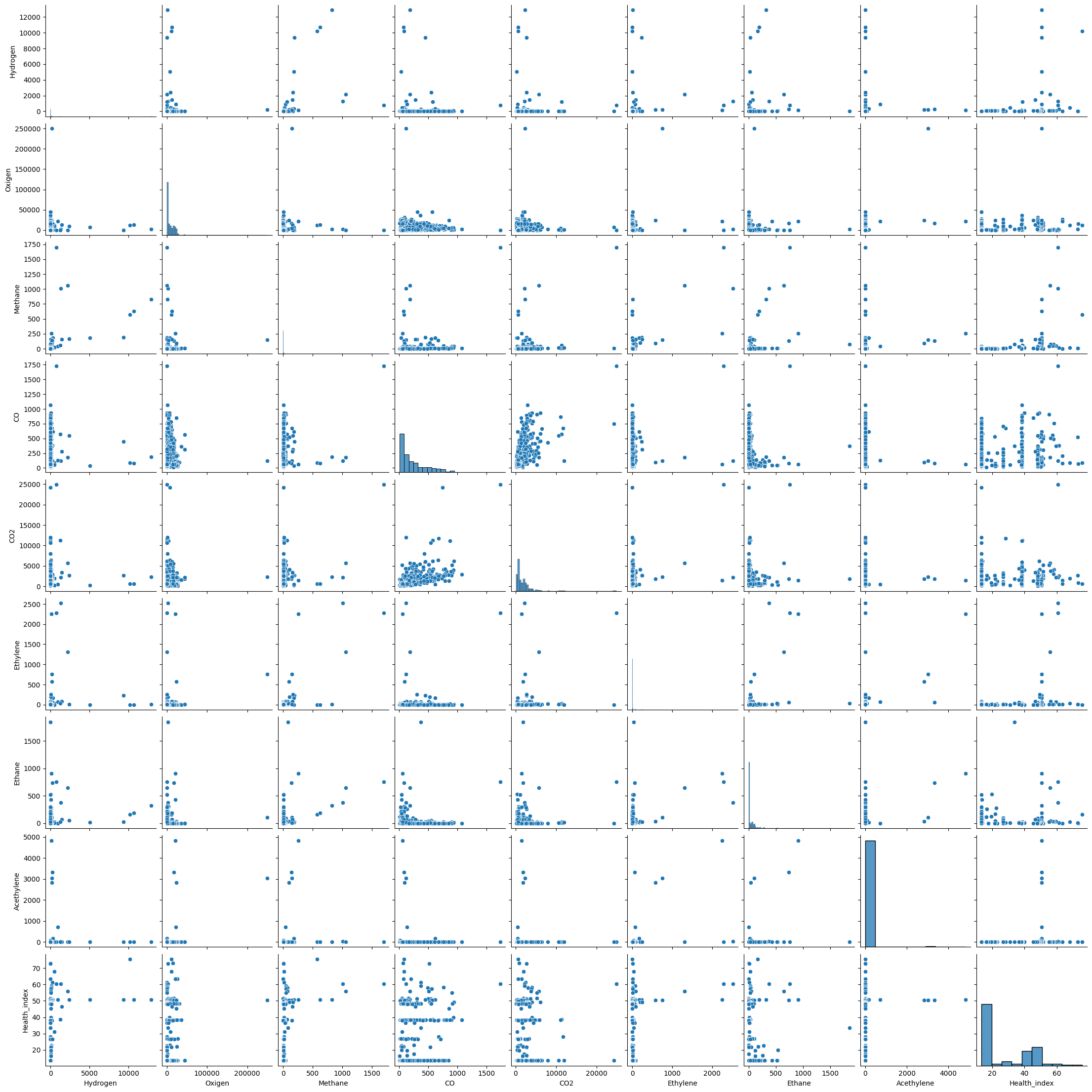
### 3.3.2. Tiền xử lý dữ liệu











## 3.4. Xây dựng mô hình

### 3.4.1. Thông số cấu hình máy tính để đào tạo mô hình

Mô hình học sâu yêu cầu tài nguyên tính toán lớn, Google Colab cung cấp phiên bản GPU miễn phí chạy trên web cho phép người nghiên cứu thử nghiệm mô hình.

Thông số được sử dụng để chạy mô hình:

* RAM hệ thống: 12.7GB.
* RAM GPU: 15GB.
* GPU T4.
* Hệ điều hành: Ubuntu 22.04.2 LTS.

### 3.4.2. Đào tạo mô hình

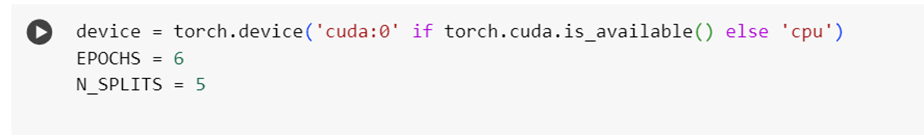
**Kết nối môi trường:**

Kết nối Google Colab với Google Drive để thuận tiện cho việc nhập dữ liệu:



Hình 3. 22 Kết nối Google Colab với Google Drive.

Sử dụng GPU, đặt epochs là 6 và chia tập đào tạo thành 5 phần:



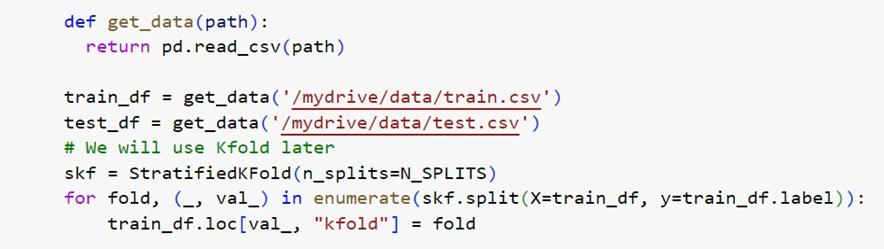
Hình 3. 23 Sử dụng GPU để đào tạo mô hình.

Trong học sâu, epoch được sử dụng để mô tả một lần chuyển toàn bộ tập dữ liệu đào tạo trong quá trình đào tạo mạng. Tăng epoch giúp mô hình hội tụ đến một giải pháp tốt hơn, nhưng nó cũng có nguy cơ quá khớp dữ liệu. Vì vậy, việc tìm đúng số lượng epoch là rất quan trọng, thường liên quan đến các kỹ thuật như dừng sớm hoặc giám sát hiệu suất trên một tập dữ liệu xác thực riêng biệt.

Tập đào tạo được áp dụng phương pháp lấy mẫu kiểm thử chéo. Kiểm thử chéo thường được sử dụng để so sánh và chọn ra mô hình tốt nhất cho một bài toán. Kỹ thuật này được áp dụng trong trường hợp dữ liệu không được nhiều.

**Lấy dữ liệu tập đào tạo và kiểm thử :**

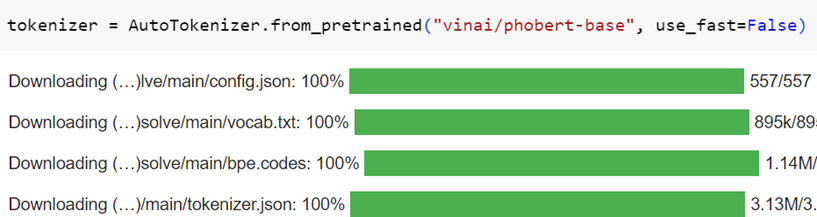
Dữ liệu được tổ chức thành hai tập riêng biệt: tập huấn luyện (train) và tập kiểm thử (test), được đọc từ các file CSV tương ứng. Để tăng tính hiệu quả trong quá trình đánh giá mô hình, chúng ta sử dụng phương pháp K-Fold Cross Validation thông qua lớp StratifiedKFold. Phương pháp này chia tập huấn luyện thành N phần (được xác định bởi tham số N\_SPLITS), đảm bảo phân phối nhãn đồng đều giữa các fold. Mỗi fold được đánh dấu trong DataFrame thông qua cột "kfold", giúp theo dõi và đánh giá hiệu suất mô hình một cách khách quan trên các phần dữ liệu khác nhau.



Hình 3. 24 Lấy dữ liệu từ tệp.

**Kết nối mô hình huấn luyện :**

Sử dụng bộ mã hóa được đào tạo trước phoBERT. PhoBERT được huấn luyện trên 20Gb dữ liệu tiếng Việt nhằm giải quyết các bài toán cho tiếng Việt. Với mô hình học sâu lần này với số lượng dữ liệu khoảng hơn 4000 bản ghi tôi quyết định chọn phiên bản PhoBERT-base: Một phiên bản nhẹ hơn với số lượng tham số ít hơn, thích hợp cho các ứng dụng yêu cầu tốc độ xử lý cao và tài nguyên tính toán hạn chế.



Hình 3. 25 Sử dụng bộ mã hóa phoBERT.

**Chuẩn bị dữ liệu đầu vào cho mô hình:**

Tham số max\_len được sử dụng để giới hạn độ dài tối đa của văn bản sau khi được mã hóa (encode). Điều này giúp chuẩn hóa dữ liệu đầu vào, trong đó các văn bản dài sẽ được cắt bớt và các văn bản ngắn sẽ được thêm padding để đạt đến độ dài thống nhất. Việc này giúp đảm bảo tính đồng nhất của dữ liệu khi đưa vào huấn luyện mô hình.

A screenshot of a computer program

Description automatically generated

Hình 3. 26 Chuẩn bị dữ liệu đầu vào cho mô hình học sâu.

**Xây dựng mô hình phân loại:**

A screenshot of a computer code

Description automatically generated

Hình 3. 27 Xây dựng cấu trúc mô hình học sâu.

Dropout xác định xác suất bỏ học là 0,3. Nó được sử dụng để ngăn chặn quá khớp dữ liệu bằng cách đặt một phần đơn vị (nút mạng) về 0 trong quá trình huấn luyện. Mô hình sử dụng PhoBERT làm nền tảng chính để trích xuất đặc trưng, kết hợp với một lớp fully-connected để thực hiện phân loại.

**Hàm đào tạo mô hình:**

Xây dựng hàm đào tạo nhằm thực hiện việc huấn luyện mô hình bằng cách chuyển dữ liệu đầu vào sang thiết bị tính toán (GPU/CPU). Trong mỗi vòng lặp, mô hình tính toán đầu ra dự đoán và so sánh với nhãn thực tế để tính loss. Dựa vào loss này, optimizer sẽ cập nhật các trọng số của mô hình. Quá trình huấn luyện được giám sát thông qua độ chính xác (accuracy) và giá trị loss.



Hình 3. 28 Hàm đào tạo mô hình học sâu.

**Hàm kiểm thử mô hình:**

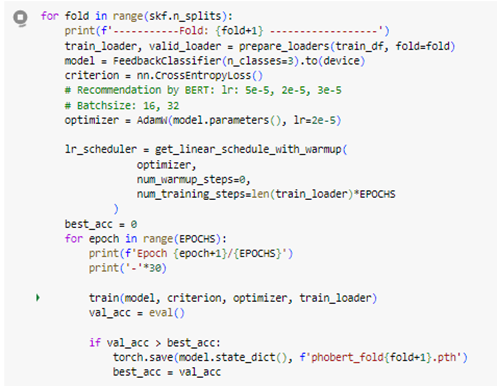
A screen shot of a computer program

Description automatically generated

Hình 3. 29 Hàm kiểm thử mô hình.

Hàm test mô hình đảm nhiệm quá trình đánh giá hiệu suất của mô hình. Quy trình thực hiện bao gồm việc tải các trọng số từ nhiều fold khác nhau để tạo ensemble model. Trong quá trình dự đoán, hàm sử dụng torch.no\_grad() để vô hiệu hóa gradient computation, tối ưu hóa bộ nhớ và thời gian tính toán. Kết quả cuối cùng là sự kết hợp dự đoán từ tất cả các mô hình, được đánh giá thông qua các metrics trong classification\_report.

**Tiến hành đào tạo mô hình:**



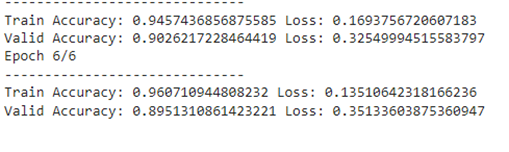
Hình 3. 30 Đào tạo mô hình học sâu.

Quá trình đào tạo mô hình được cấu hình với nhiều tham số quan trọng:

* CrossEntropyLoss được chọn làm hàm mất mát (loss function), phù hợp cho bài toán phân loại nhiều lớp. Hàm này đánh giá mức độ chênh lệch giữa phân phối xác suất dự đoán và nhãn thực tế.
* Tối ưu hóa sử dụng thuật toán Adam với learning rate 2e-5, một lựa chọn phổ biến cho việc tinh chỉnh các mô hình BERT. Learning rate scheduler được áp dụng kèm warm-up để ổn định quá trình huấn luyện ban đầu.
* Mô hình được huấn luyện qua nhiều fold, với việc lưu lại trọng số tốt nhất dựa trên độ chính xác trên tập validation (val\_acc).

## 3.5. Đánh giá mô hình

Kết quả thực nghiệm cho thấy mô hình đạt được hiệu suất tốt trong quá trình huấn luyện, cụ thể:



Hình 3. 31 Kết quả đào tạo mô hình học sâu từ một K-fold.

Tại epoch thứ 6, mô hình đạt độ chính xác trên tập huấn luyện (Train Accuracy) là 94.57% và độ mất mát (Loss) là 0.169

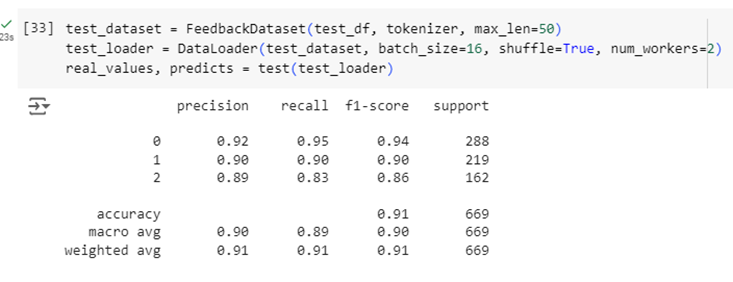
Độ chính xác trên tập kiểm thử (Valid Accuracy) đạt 90.26% với độ mất mát là 0.325

Trong lần chạy tiếp theo, mô hình còn cải thiện với độ chính xác trên tập huấn luyện đạt 95.87% và độ mất mát giảm xuống 0.135

Tập kiểm thử cũng cho kết quả khả quan với độ chính xác 89.51% và độ mất mát 0.351

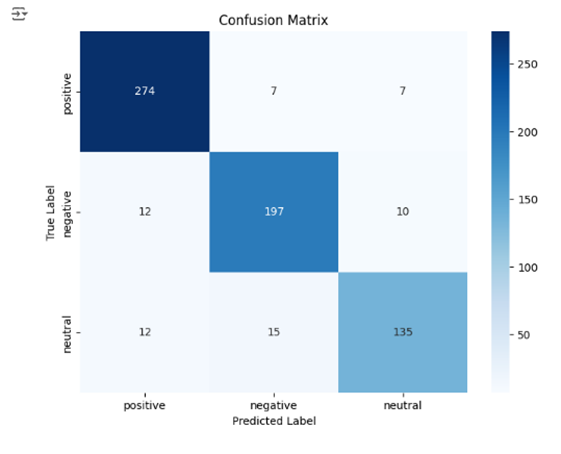
Nhận xét: Mặc dù được huấn luyện trên tập dữ liệu có kích thước khiêm tốn, mô hình vẫn thể hiện khả năng học tốt với độ chính xác cao. Không xuất hiện hiện tượng quá khớp hay dưới khớp, thể hiện qua sự chênh lệch không đáng kể giữa hiệu suất trên tập huấn luyện và tập kiểm thử. Điều này chứng tỏ mô hình có khả năng tổng quát hóa tốt trên dữ liệu mới.

*Kết quả đánh giá mô hình:*



Hình 3. 32 Kết quả đánh giá mô hình học sâu.

*Trực quan hóa kết quả đánh giá mô hình:*



Hình 3. 33 Trực quan hóa kết quả đánh giá mô hình.

**Nhận xét kết quả đánh giá mô hình:**

Lớp 0 (positive, tích cực) cho precision, recall, F1-score lần lượt là 0.92, 0.95, 0.94. Precision bằng 0.92, nghĩa là mô hình dự đoán là “positive” đúng khoảng 92%.

Recall bằng 0.95, nghĩa là mô hình chỉ phát hiện được khoảng 95% trường hợp thực sự là “positive”. F1-score là 0.94, cho biết bộ phân loại tốt.

Lớp 1 (negative, tiêu cực) cho precision, recall, F1-score lần lượt là 0.90, 0.90, 0.90. Precision bằng 0.90, nghĩa là mô hình dự đoán là “negative” đúng khoảng 90%. Recall là 0.86, nghĩa là mô hình tốt trong việc phát hiện các trường hợp thực sự là " negative " và nó đã phát hiện được khoảng 90%. F1-score là 0.90, cho thấy bộ phân loại tốt.

Tương tự, lớp 2 (neutral, tiêu cực) cho precision, recall, F1-score lần lượt là 0.89, 0.83, 0.86. Precision bằng 0.89, nghĩa là mô hình dự đoán là “neutral” đúng khoảng 89%. Recall là 0.83, nghĩa là mô hình tốt trong việc phát hiện các trường hợp thực sự là "neutral" và nó đã phát hiện được khoảng 83%. F1-score là 0.86, cho thấy bộ phân loại khá tốt.

Accuracy là 0.91, mô hình đoán đúng khoảng 91% trên toàn bộ dự đoán.

Macro avg là trung bình của các giá trị precision, recall và F1-score của cả ba lớp. Trong trường hợp này, kết quả lần lượt là 0.90, 0.89, 0.90.

Weighted avg: mỗi lớp được coi trọng dựa trên tỷ lệ số lượng mẫu của lớp đó so với tổng số lượng mẫu. Các lớp có số lượng mẫu lớn hơn sẽ có trọng số lớn hơn trong tính toán. Trong trường hợp này, kết quả lần lượt là 0.91, 0.91, 0.91.

Trong kết quả, Macro avg và weighted avg khá đồng đều, điều này nghĩa là sự không cân bằng trong số lượng mẫu giữa các lớp không ảnh hưởng đáng kể đến giá trị trung bình của các số liệu đánh giá hiệu suất của mô hình.

## 3.6. Dự đoán với mô hình đã xây dựng được

Sau khi đào tạo mô hình ta lấy 20 mẫu kiểm thử đã được gán nhãn cho chạy qua mô hình. Kết quả thu được:

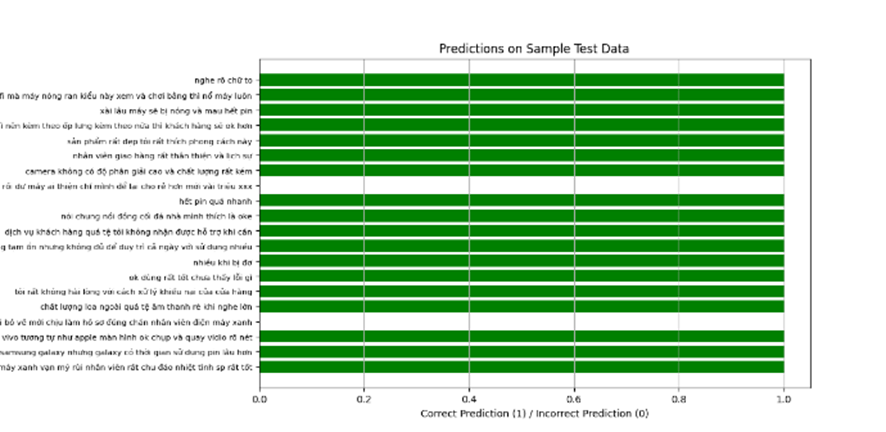
A screenshot of a computer

Description automatically generated

Hình 3. 34 Kết quả kiểm thử mô hình với 20 mẫu.

Dựa trên kết quả phân tích từ việc kiểm tra 20 mẫu được chọn ngẫu nhiên, tôi có thể đưa ra một số nhận xét quan trọng về hiệu suất của mô hình.

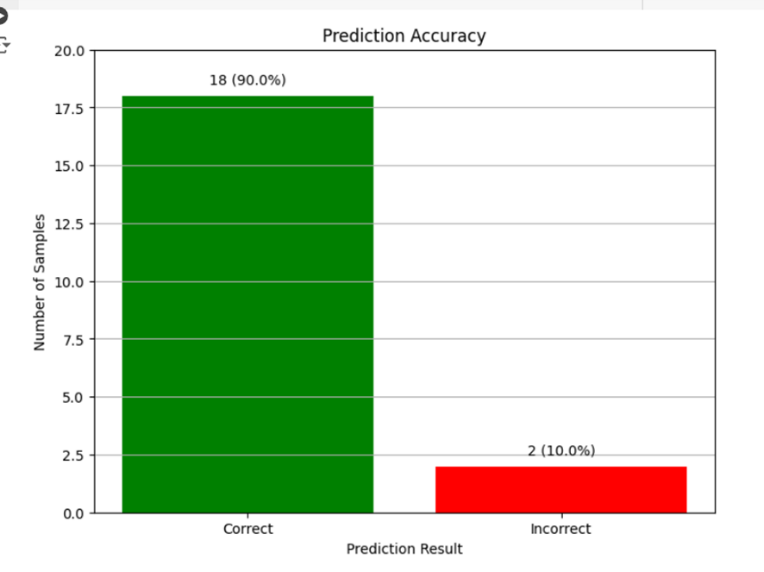
Về hiệu suất tổng thể, mô hình thể hiện kết quả rất khả quan với độ chính xác đạt 90%, tương đương với 18 mẫu dự đoán đúng trong tổng số 20 mẫu kiểm tra. Chỉ có 2 trường hợp, chiếm 10% tổng số mẫu, được dự đoán không chính xác.



Hình 3. 35 Hiệu suất của mô hình.

Khi đi sâu vào phân tích chi tiết, tôi nhận thấy phần lớn các dự đoán của mô hình đều phù hợp với nhãn thực tế. Điều đáng chú ý là mô hình thể hiện khả năng xử lý tốt với nhiều loại phản hồi khác nhau về sản phẩm và dịch vụ. Đặc biệt, mô hình cho thấy khả năng phân biệt hiệu quả giữa các mức độ đánh giá khác nhau (0, 1, 2).

Trong quá trình phân tích lỗi, tôi phát hiện hai trường hợp dự đoán sai. Trường hợp thứ nhất liên quan đến phản hồi về "nhân viên diện mày xanh", trong đó mô hình dự đoán giá trị 1 nhưng thực tế là 0. Trường hợp thứ hai liên quan đến nội dung "để lại cho rẻ hơn", mô hình dự đoán là 1 trong khi giá trị thực tế là 2.



Hình 3. 36 Trực quan tỷ lệ dự đoán của mô hình.

Kết luận, mô hình đã thể hiện hiệu suất ấn tượng trên tập dữ liệu kiểm tra. Tỷ lệ dự đoán đúng cao (90%) không chỉ cho thấy độ chính xác tốt mà còn thể hiện khả năng tổng quát hóa hiệu quả của mô hình. Những trường hợp dự đoán sai, mặc dù ít, vẫn có thể được sử dụng như những điểm tham chiếu quý giá để cải thiện hiệu suất của mô hình trong tương lai.

# CHƯƠNG 4

# ỨNG DỤNG MÔ HÌNH

Sau khi xây dựng thành công mô hình phân lớp nhận xét khách hàng. Để khai thác hiệu quả mô hình này, tôi đã tích hợp nó vào một website thương mại điện tử mà tôi đã phát triển trong quá trình học môn Thương mại điện tử (ở đồ án này tôi đã kế thừa các chức năng cũ của website mà mình đã xây dựng và tôi có xây dựng thêm một số chức năng để tích hợp mô hình). Website này sẽ sử dụng mô hình để tự động phân tích các nhận xét từ khách hàng, giúp hiển thị trực quan các ý kiến đánh giá, từ đó cải thiện trải nghiệm người dùng cũng như tối ưu hóa chiến lược kinh doanh của doanh nghiệp. Việc áp dụng mô hình phân lớp nhận xét không chỉ nâng cao tính chuyên nghiệp của website mà còn góp phần vào việc phát triển dịch vụ và sản phẩm dựa trên nhu cầu thực tế của khách hàng.

## 4.1. Xây dựng API

Trong quá trình phát triển hệ thống, tôi đã sử dụng FastAPI để xây dựng API phục vụ cho việc phân lớp nhận xét khách hàng. Mô hình phân lớp được xây dựng dựa trên phobert-base, một mô hình BERT được huấn luyện cho ngôn ngữ tiếng Việt. Dưới đây là các bước chính trong quá trình xây dựng API:

* Khởi tạo mô hình phân lớp



Hình 4. 1 Định nghĩa lại mô hình.

Mô hình FeedbackClassifier được xây dựng trên cơ sở lớp nn.Module của PyTorch. Mô hình này sử dụng phobert-base từ thư viện transformers, bao gồm một lớp dropout và một lớp fully connected để dự đoán.

* Khởi tạo ứng dụng FastAPI

Tôi đã khởi tạo một ứng dụng FastAPI để tạo các API endpoint. Ứng dụng này sẽ tiếp nhận các yêu cầu và trả về kết quả dự đoán.

* Tải mô hình và tokenizer

Mô hình phân lớp đã được huấn luyện trước được tải vào và đưa vào chế độ dự đoán. AutoTokenizer từ phobert-base được sử dụng để token hóa các nhận xét đầu vào.

* Định nghĩa kiểu dữ liệu đầu vào

Kiểu dữ liệu đầu vào cho API được định nghĩa bằng BaseModel của Pydantic. Dữ liệu đầu vào bao gồm trường comment chứa nhận xét từ người dùng.



Hình 4. 2 Khởi tạo API Phân loại nhận xét.

* API dự đoán

API dự đoán được định nghĩa thông qua một endpoint POST với đường dẫn /predict/. Khi nhận được một nhận xét, API sẽ token hóa văn bản và sử dụng mô hình để dự đoán lớp phù hợp.



Hình 4. 3 API Dự đoán và Xử lý Tokenize.

* Chạy server

Server FastAPI được khởi chạy bằng Uvicorn tại host và port được chỉ định, giúp API có thể tiếp nhận các yêu cầu từ bên ngoài.



Hình 4. 4 Giao diện Server FastAPI trên local host.

## 4.2. Tích hợp và triển khai API phân loại nhận xét.

### 4.2.1. Phân tích Yêu cầu Hệ thống

Trong quá trình phát triển hệ thống phân loại nhận xét, tôi đã tiến hành phân tích kỹ lưỡng các yêu cầu chức năng và phi chức năng. Các yêu cầu này được xác định dựa trên nhu cầu thực tế của người dùng và mục tiêu kinh doanh của doanh nghiệp.

**Về mặt chức năng**, hệ thống cần đảm bảo tính xác thực và bảo mật thông qua việc kiểm tra đăng nhập của người dùng trước khi cho phép họ thực hiện đánh giá. Quá trình xác thực không chỉ dừng lại ở việc kiểm tra tình trạng đăng nhập, mà còn bao gồm việc xác minh tài khoản đang hoạt động và có đủ quyền truy cập vào chức năng đánh giá.

Để đảm bảo tính chính xác và giá trị của đánh giá, hệ thống sẽ thực hiện kiểm tra kỹ lưỡng lịch sử mua hàng của người dùng. Điều này bao gồm việc xác nhận đơn hàng đã hoàn thành, thời gian giao hàng thành công, và đảm bảo người dùng đã có đủ thời gian trải nghiệm sản phẩm. Ngoài ra, hệ thống cũng sẽ kiểm tra để tránh trường hợp người dùng đánh giá trùng lặp cho cùng một sản phẩm.

Về khía cạnh xử lý dữ liệu, hệ thống tích hợp chặt chẽ với API phân loại bình luận để thực hiện việc phân tích và phân loại các đánh giá một cách tự động. Kết quả phân loại sẽ được lưu trữ có hệ thống trong cơ sở dữ liệu, đồng thời được sử dụng để tạo ra các báo cáo thống kê chi tiết.

**Về các yêu cầu phi chức năng**, hiệu năng của hệ thống được đặt lên hàng đầu. Hệ thống cần đảm bảo thời gian phản hồi nhanh chóng, với mục tiêu xử lý phản hồi trong vòng 2 giây cho mỗi thao tác người dùng. Để đạt được điều này, tôi áp dụng cơ chế xử lý bất đồng bộ cho các tác vụ nặng và tối ưu hóa các truy vấn cơ sở dữ liệu.

Tính nhất quán của dữ liệu được đảm bảo thông qua việc đồng bộ hóa theo thời gian thực và các cơ chế sao lưu tự động. Hệ thống cũng được trang bị khả năng xử lý lỗi toàn diện, bao gồm việc ghi log chi tiết và các cơ chế tự phục hồi cho các lỗi phổ biến.

Giao diện người dùng được thiết kế với tiêu chí thân thiện và dễ sử dụng, có khả năng tương thích trên đa dạng thiết bị.

### 4.2.2. Kiến trúc Hệ thống

Kiến trúc của hệ thống được thiết kế theo mô hình phân lớp, cho phép tách biệt rõ ràng giữa các thành phần và dễ dàng mở rộng trong tương lai.

**Cấu trúc Thư mục**

A screenshot of a computer code

Description automatically generated

Hình 4. 5 Cấu trúc thư mục dự án.

Cấu trúc này phân chia rõ ràng các thành phần theo chức năng, bao gồm các thư mục controllers, models, services và utils, giúp việc quản lý và phát triển code trở nên thuận tiện hơn.

**Mô hình Dữ liệu**

Hệ thống sử dụng cơ sở dữ liệu quan hệ với schema được thiết kế như sau:

A screenshot of a computer

Description automatically generated

Hình 4. 6 Mô hình dữ liệu.

Mô hình dữ liệu này được tối ưu cho việc lưu trữ và truy xuất thông tin đánh giá, với các bảng được liên kết chặt chẽ thông qua các khóa ngoại, đảm bảo tính toàn vẹn của dữ liệu.

**Luồng Xử lý**

Quy trình xử lý của hệ thống được thể hiện trong sơ đồ luồng sau:

A screenshot of a computer program

Description automatically generated

Hình 4. 7 Quy trình xử lý.

Luồng xử lý này mô tả chi tiết các bước từ khi nhận yêu cầu đánh giá từ người dùng cho đến khi hoàn thành việc phân loại và lưu trữ kết quả.

### 4.2.3. Triển khai Backend

**API Endpoint**



Hình 4. 8 API endpoint.

Các API endpoint này tuân thủ chuẩn RESTful và được thiết kế để xử lý đầy đủ các tác vụ liên quan đến đánh giá và phân loại.

**Xử lý Response**

Cấu trúc response của hệ thống được thiết kế nhất quán. Mỗi response đều bao gồm các thành phần chuẩn như status code, message và data, giúp client dễ dàng xử lý kết quả trả về.

A screen shot of a computer code

Description automatically generated

Hình 4. 9 Cấu trúc respose

**Mã Lỗi HTTP**

* 400: Bad Request - Yêu cầu không hợp lệ
* 401: Unauthorized - Chưa đăng nhập
* 403: Forbidden - Không có quyền truy cập
* 500: Internal Server Error - Lỗi phân loại

### 4.2.4. Triển khai Frontend

**Dashboard**

*Giao diện*

- Biểu đồ phân phối phản hồi (pie chart)

- Biểu đồ số lượng theo thời gian (bar chart)

- Bộ lọc thời gian

*Thống kê*

- Phân loại: Positive, Negative, Neutral

- Tỷ lệ phần trăm từng loại

- Màu sắc trực quan:

+ Positive: Xanh nhạt

+ Negative: Hồng

+ Neutral: Vàng nhạt

**Danh sách Phản hồi**

*Cấu trúc*

- Bảng hiển thị chi tiết

- Phân trang và sắp xếp

- Bộ lọc đa tiêu chí

*Thông tin Hiển thị*

- Người dùng và sản phẩm

- Nội dung bình luận

- Kết quả phân loại

- Thời gian đánh giá

### 4.2.5. Tối ưu và Monitoring

Để đảm bảo hiệu năng tối ưu, hệ thống được tích hợp nhiều giải pháp kỹ thuật tiên tiến. Việc sử dụng asyncio cho các API calls giúp tăng khả năng xử lý đồng thời và giảm thời gian chờ. Caching được áp dụng cho kết quả phân loại để tránh việc phải tính toán lại các kết quả đã có.

Batch processing được triển khai để xử lý hiệu quả các yêu cầu phân loại số lượng lớn, trong khi transaction handling đảm bảo tính nhất quán của dữ liệu trong mọi tình huống.

Hệ thống monitoring được thiết lập để theo dõi các chỉ số quan trọng như thời gian phản hồi API, tỷ lệ thành công/thất bại, độ chính xác của việc phân loại và khối lượng request. Các metric này được theo dõi và phân tích liên tục để đảm bảo hệ thống luôn hoạt động ổn định và hiệu quả.

## 4.3. Thực nghiệm kết quả

Sau khi tích hợp thành công mô hình phân lớp nhận xét của khách hàng vào trang web thương mại điện tử bán đồ công nghệ “TechTronic” tôi tiến hành thực nghiệm các quy trình đánh giá với vai trò là khách hàng.

A screenshot of a computer

Description automatically generated

Hình 4. 10 Giao diện trang Web TechTronic.

Với kết quả thực nghiệm đáng chú ý như sau:

A screenshot of a graph

Description automatically generated

Hình 4. 11 Biểu đồ thống kê.

***Hiệu Suất:***

Phân loại chính xác 31 phản hồi trong tháng 10/2024:

* 41.9% tích cực (13)
* 38.7% tiêu cực (12)
* 19.4% trung lập (6)

A screenshot of a computer

Description automatically generated

Hình 4. 12 Giao diện quản lý những nhận xét, đánh giá.

***Tính Năng Chính:***

* Xử lý ngôn ngữ linh hoạt (từ đơn giản đến biệt ngữ).
* Dashboard quản lý trực quan..
* Bộ lọc thông minh.

A screenshot of a computer

Description automatically generated

Hình 4. 13 Form đánh giá, nhận xét.

***Giao Diện Người Dùng:***

* Form phản hồi thân thiện
* Quản lý dữ liệu hiệu quả

Hệ thống đã thành công trong việc tự động hóa phân loại nhận xét cũng như các phản hồi, với độ chính xác cao và giao diện thân thiện, giúp TechTronic nắm bắt tốt hơn ý kiến khách hàng.

# KẾT LUẬN

## Đánh giá kết quả đạt được

Sau quá trình nghiên cứu và thực hiện đề tài " Nghiên cứu, ứng dụng xử lý ngôn ngữ tự nhiên (NLP) trong bài toán phân lớp nhận xét của khách hàng ", với sự đam mê, nỗ lực nghiêm túc cùng thời gian đầu tư đáng kể, em đã hoàn thành Đồ án tốt nghiệp đúng tiến độ và đạt được những kết quả đáng khích lệ. Đề tài này không chỉ là một bài toán kỹ thuật đơn thuần mà còn mang tính ứng dụng cao trong thực tiễn kinh doanh hiện đại.

Về mặt học thuật, đề tài đã giúp em nghiên cứu và nắm vững các kiến thức nền tảng về xử lý ngôn ngữ tự nhiên (NLP), đặc biệt là hiểu sâu về đặc trưng và thách thức trong việc xử lý tiếng Việt. Quá trình thực hiện đã giúp em tích lũy được kinh nghiệm quý báu trong việc xây dựng một hệ thống phân loại văn bản hoàn chỉnh, từ thu thập dữ liệu đến triển khai ứng dụng thực tế. Những kiến thức này không chỉ là nền tảng cho đề tài mà còn là hành trang quý giá cho con đường nghề nghiệp của em trong tương lai.

Về mặt kỹ thuật, đồ án đã xây dựng thành công một pipeline xử lý dữ liệu hoàn chỉnh và hiện đại, bao gồm các bước thu thập và tổ chức dữ liệu có hệ thống, tiền xử lý dữ liệu chuyên sâu, trích xuất đặc trưng từ dữ liệu văn bản, và thiết kế cùng huấn luyện mô hình học sâu. Đặc biệt, kết quả đạt được với độ chính xác trên 90% trên tập kiểm thử là một minh chứng thuyết phục cho hiệu quả của mô hình PhoBERT trong việc xử lý các bài toán chuyên ngành bằng tiếng Việt, ngay cả khi làm việc với bộ dữ liệu có kích thước khiêm tốn. Con số này không chỉ thể hiện độ tin cậy của mô hình mà còn khẳng định tính khả thi của việc ứng dụng các công nghệ học sâu vào các bài toán thực tế.

Về khía cạnh ứng dụng thực tế, đồ án đã thành công trong việc phát triển một hệ thống hoàn chỉnh có khả năng tích hợp vào môi trường trang web thương mại điện tử có sẵn. Cụ thể, em đã xây dựng được API bằng cách sử dụng FastAPI để triển khai mô hình AI phân loại tự động các nhận xét và phản hồi của khách hàng. Trong hệ thống còn được trang bị giao diện quản trị trực quan, tích hợp các công cụ trực quan hóa dữ liệu, giúp người quản lý dễ dàng theo dõi và nắm bắt xu hướng phản hồi của khách hàng một cách tổng quan và chi tiết. Hệ thống này không chỉ đáp ứng được yêu cầu về mặt kỹ thuật mà còn tạo ra giá trị thực sự cho doanh nghiệp trong việc quản lý và phân tích phản hồi khách hàng.

Tuy nhiên, với tinh thần cầu tiến và mong muốn hoàn thiện sản phẩm, em cũng nhận thấy đề tài vẫn còn một số hạn chế cần được cải thiện trong tương lai. Điển hình như kích thước bộ dữ liệu còn khiêm tốn, chưa thử nghiệm được nhiều kiến trúc mô hình khác nhau, thời gian tối ưu hóa mô hình còn hạn chế, và hệ thống hiện tại có giao diện người dùng chưa được đẹp. Đây sẽ là những thách thức đáng giá cần được giải quyết trong các giai đoạn phát triển tiếp theo.

## Hướng phát triển của đề tài

Về hướng phát triển trong tương lai, em dự kiến sẽ tập trung vào ba khía cạnh chính:

Thứ nhất, cải thiện hiệu suất mô hình thông qua việc mở rộng bộ dữ liệu huấn luyện, nghiên cứu và áp dụng các kỹ thuật tiền xử lý tiên tiến, thử nghiệm các kiến trúc mô hình học sâu mới, và tối ưu hóa các siêu tham số. Điều này sẽ giúp nâng cao độ chính xác và độ tin cậy của hệ thống, đồng thời mở rộng khả năng xử lý các trường hợp phức tạp.

Thứ hai, mở rộng tính năng của hệ thống bằng cách phát triển khả năng phân loại đa nhãn, tích hợp phân tích cảm xúc chi tiết, và xây dựng hệ thống phản hồi tự động dựa trên kết quả phân loại. Những cải tiến này sẽ giúp hệ thống trở nên toàn diện và thông minh hơn trong việc xử lý phản hồi khách hàng.

Thứ ba, nâng cấp hệ thống quản trị thông qua việc phát triển dashboard phân tích chuyên sâu, tích hợp hệ thống báo cáo tự động, và xây dựng cơ chế cảnh báo thông minh cho các phản hồi tiêu cực. Những tính năng này sẽ tạo ra một công cụ quản lý mạnh mẽ, hỗ trợ đắc lực cho việc ra quyết định trong kinh doanh.

Quá trình thực hiện đồ án không chỉ giúp em củng cố vận dụng tốt kiến thức đã học tại khoa Công nghệ thông tin trường Đại học Mỏ - Địa chất, mà còn là cơ hội quý báu để phát triển các kỹ năng thiết yếu như tư duy phân tích, xử lý dữ liệu và lập trình ứng dụng thực tế. Đề tài cũng đã giúp em làm quen với các công nghệ mới và phương pháp nghiên cứu khoa học, tạo nền tảng vững chắc cho sự phát triển nghề nghiệp trong tương lai.

Em xin chân thành cảm ơn sự hướng dẫn tận tình của quý thầy cô trong suốt quá trình thực hiện đồ án. Đặc biệt, em xin gửi lời tri ân sâu sắc đến thầy Lê Hồng Anh, người đã không quản ngại thời gian và công sức để hỗ trợ, chỉ bảo em trong từng bước thực hiện đề tài. Những ý kiến đóng góp quý báu và sự định hướng của thầy đã giúp em hoàn thiện đồ án một cách tốt nhất.

Mặc dù đề tài còn những hạn chế nhất định do thời gian và kinh nghiệm còn hạn chế, em rất mong tiếp tục nhận được những ý kiến đóng góp từ quý thầy cô để có thể hoàn thiện và phát triển dự án này tốt hơn trong tương lai. Một lần nữa, em xin bày tỏ lòng biết ơn chân thành đến quý thầy cô đã đồng hành, giúp đỡ em trong suốt quá trình học tập và thực hiện Đồ án tốt nghiệp này !

# TÀI LIỆU THAM KHẢO

[1] H. Phê, Từ điển tiếng Việt, Hà Nội: NXB Đà Nẵng, 2003.

[2] Đ. Điền, Giáo trình "Xử lý ngôn ngữ tự nhiên", NXB Đại học Quốc gia - HCM, 2006.

[3] Mai Ngọc Chừ, Vũ Đức Nghiệu, Hoàng Trọng Phiến, Cơ sở ngôn ngữ học và tiếng Việt, NXB Giáo dục.

[4] "Đặc điểm tiếng Việt," 21 11 2010. [Online]. Available: https://vnlp.net/ti%E1%BA%BFngvi%E1%BB%87tc%C6%A1b%E1%BA%A3n/d%E1%BA%B7c-di%E1%BB%83m- ti%E1%BA%BFng-vi%E1%BB%87t/.

[5] J. Singh, Natural Language Processing in the Real World, CRC Press, 2023.

[6] Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana, Practical Natural Language Processing, O’Reilly Media, Inc., 2020.

[7] T. Ganegedara, Natural Language Processing with TensorFlow, Packt Publishing, 2022.

[8] S. Ravichandiran, Getting Started with Google BERT, Packt Publishing, 2021.

[9] "Các phương pháp đánh giá một hệ thống phân lớp," [Online]. Available: https://machinelearningcoban.com/2017/08/31/evaluation/.

[10] A. Chapagain, Hands-On Web Scraping with Python, Packt Publishing.

[11] "Website DienMayXanh - Check and Analyze Any Website | Similarweb," Similarweb LTD, [Online]. Available: https://www.similarweb.com/.