Ứng dụng AI kiểm duyệt nội dung văn bản tiếng Việt

# Danh mục các ký hiệu, các chữ viết tắt

| STT | Từ viết tắt | Từ gốc | Tiếng Việt |
| --- | --- | --- | --- |
| 1 | AI | Artificial Intelligence | Trí tuệ nhân tạo |
| 2 | API | Application Programing Interface | Giao diện lập trình ứng dụng |
| 3 | NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên |
| 4 | SOTA | State Of The Art | Chỉ mức độ tiên tiến nhất của một công nghệ |
| 5 | BERT | Bidirectional Encoder Representations from Transformers | Biểu diễn Thể hiện Mã hóa Hai chiều từ Transformer |
|  |  |  |  |

# Mở đầu

Trong bối cảnh số hóa và toàn cầu hóa hiện nay, việc kiểm duyệt nội dung trên các nền tảng trực tuyến đã trở thành một nhu cầu cấp thiết đối với các tổ chức chính trị - xã hội. Với sự phát triển mạnh mẽ của công nghệ trí tuệ nhân tạo (AI), việc áp dụng các giải pháp AI vào kiểm duyệt nội dung văn bản tiếng Việt không chỉ giúp tối ưu hóa quy trình kiểm duyệt mà còn đảm bảo tính chính xác và nhanh chóng. Điều này đặc biệt quan trọng trong việc duy trì môi trường trao đổi thông tin lành mạnh, bảo vệ quyền lợi và an ninh của các hội viên.

Tính năng AI kiểm duyệt nội dung văn bản tiếng Việt được phát triển nhằm hỗ trợ các tổ chức trong việc tự động hóa quá trình kiểm duyệt, giúp nhận diện và loại bỏ những nội dung không phù hợp một cách hiệu quả. Hệ thống sẽ tích hợp nhiều chức năng tiện ích như phát hiện ngôn từ phản cảm, đánh giá tính phù hợp của nội dung, lọc bỏ các thông tin sai lệch hoặc có hại. Bên cạnh đó, AI cũng sẽ hỗ trợ phân loại và gắn nhãn các nội dung dựa trên mức độ nhạy cảm, giúp quản lý và điều hành các hoạt động truyền thông một cách chặt chẽ và hiệu quả hơn.

Tài liệu này sẽ cung cấp cái nhìn toàn diện về cách thức hệ thống AI kiểm duyệt nội dung văn bản tiếng Việt được thiết kế, triển khai và vận hành nhằm đáp ứng các yêu cầu và mục tiêu đề ra. Từ cấu trúc phần mềm, các dịch vụ API, đến thiết kế chức năng và quy trình thực hiện, tất cả đều được trình bày rõ ràng và cụ thể. Thông qua tài liệu này, đội ngũ phát triển sẽ có đầy đủ thông tin cần thiết để triển khai tính năng AI kiểm duyệt một cách hiệu quả, đồng thời các bên liên quan cũng có thể dễ dàng theo dõi và hiểu rõ tiến trình của dự án.

# Chương 1.Các quyết định về Công nghệ và Kiến trúc sản phẩm

## 1.1.Nguyên tắc thiết kế

## 1.2.Lựa chọn công nghệ

### 1.2.1.Công nghệ phân đoạn từ

Phân loại nội dung văn bản tiếng Việt là một nhiệm vụ quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Để thực hiện hiệu quả nhiệm vụ này, việc phân đoạn từ (Word Segmentation) là một bước thiết yếu và không thể bỏ qua vì các lí do sau:

* Đặc điểm của tiếng Việt: Tiếng Việt là ngôn ngữ không phân tách từ bằng khoảng trắng như tiếng Anh. Một câu tiếng Việt thường được viết liên tục và các từ có thể ghép lại mà không có dấu hiệu rõ ràng để phân biệt. Ví dụ: "học sinh" (student) hay "giáo viên" (teacher) đều là các cụm từ không có khoảng trắng phân tách giữa các từ thành phần. Điều này dẫn đến khó khăn trong việc nhận diện từ và phân tích câu mà không có bước phân đoạn từ.
* Hỗ trợ mô hình phân loại: Các từ điển ngôn ngữ và mô hình ngôn ngữ (như Word2Vec, BERT) đều yêu cầu đầu vào là các từ đã được phân đoạn chính xác. Nếu không có bước phân đoạn từ, các mô hình này sẽ không thể học được ngữ nghĩa chính xác của các từ và cụm từ, từ đó ảnh hưởng đến hiệu quả phân loại nội dung.
* Giảm bớt sự nhầm lẫn ngữ nghĩa: Phân đoạn từ giúp giảm bớt sự nhầm lẫn ngữ nghĩa trong văn bản. Ví dụ, không phân đoạn từ có thể dẫn đến việc "công lý" (justice) bị hiểu nhầm thành "công" và "lý" (public reason), hai từ có ý nghĩa hoàn toàn khác. Điều này sẽ làm cho quá trình phân loại trở nên kém chính xác.

Giải pháp được đưa ra để xử lý các vấn đề trên là sử dụng VNCoreNLP - một công cụ xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) dành riêng cho tiếng Việt, cung cấp nhiều tính năng hữu ích trong việc phân tích và xử lý văn bản tiếng Việt.

VnCoreNLP hỗ trợ các nhiệm vụ xử lý ngôn ngữ tự nhiên quan trọng bao gồm phân đoạn từ, gán nhãn từ loại, nhận diện thực thể có tên và phân tích cú pháp phụ thuộc, và đạt được kết quả tiên tiến nhất (SOTA) cho các nhiệm vụ này.

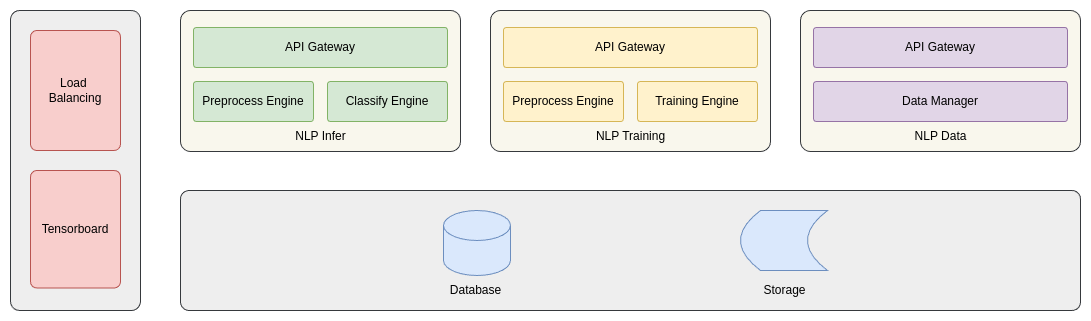
### 1.2.3.Công nghệ phân loại văn bản

PhoBERT là một mô hình ngôn ngữ dựa trên kiến trúc BERT (Bidirectional Encoder Representations from Transformers) được phát triển đặc biệt cho tiếng Việt với các ưu điểm:

* Độ chính xác cao: PhoBERT vượt trội hơn các phương pháp tiếp cận đơn ngữ và đa ngôn ngữ trước đây, đạt được hiệu suất tiên tiến nhất trên bốn nhiệm vụ NLP tiếng Việt bao gồm gắn nhãn một phần giọng nói, phân tích cú pháp phụ thuộc, nhận dạng thực thể có tên và suy luận ngôn ngữ tự nhiên.
* Hiểu ngữ cảnh tốt: Nhờ vào kiến trúc BERT, PhoBERT có khả năng hiểu ngữ cảnh của từ trong câu tốt hơn các mô hình truyền thống. Điều này đặc biệt quan trọng trong tiếng Việt, nơi ngữ nghĩa của từ có thể thay đổi tùy thuộc vào ngữ cảnh.
* Tiền huấn luyện trên dữ liệu tiếng Việt: PhoBERT được huấn luyện trên một lượng lớn dữ liệu tiếng Việt, giúp mô hình nắm bắt được đặc trưng ngôn ngữ và cú pháp của tiếng Việt một cách chính xác hơn. Điều này làm cho PhoBERT trở thành lựa chọn lý tưởng cho các tác vụ NLP tiếng Việt.
* Hỗ trợ và Tài liệu phong phú: PhoBERT có tài liệu hướng dẫn và cộng đồng người dùng phong phú, giúp việc triển khai và tùy chỉnh mô hình trở nên dễ dàng hơn.
* Mở rộng và tùy chỉnh: PhoBERT cho phép dễ dàng mở rộng và tùy chỉnh mô hình cho các tác vụ cụ thể. Người dùng có thể sử dụng PhoBERT như một mô hình cơ sở và tinh chỉnh nó cho các nhiệm vụ phân loại văn bản theo yêu cầu của mình.

# Chương 2: Mô hình kiến trúc hệ thống

## 2.1.Kiến trúc tổng thể



## 2.2.Core service

NLP Infer: cung cấp chức năng gán nhãn văn bản phục vụ kiểm duyệt nội dung tin tức.

* Preprocess Engine: Chịu trách nhiệm tiền xử lý văn bản đầu vào bao gồm: phân đoạn từ, tách đoạn.
* Classify Engine: Nhận văn bản đã qua tiền xử lý từ Preprocess Engine và sử dụng mô hình AI để thực hiện suy luận, từ đó gắn nhãn cho văn bản.

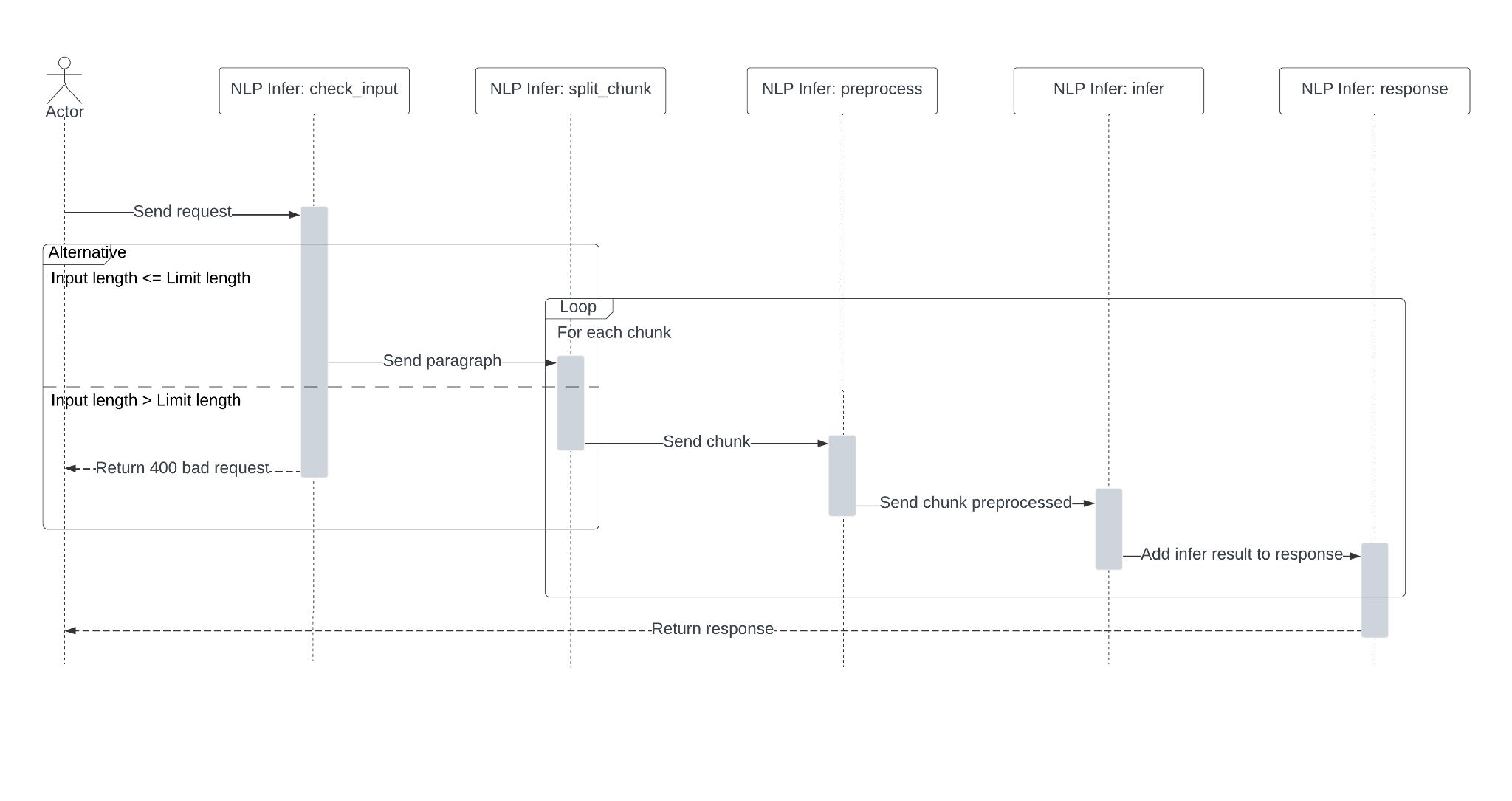
NLP Training: cung cấp chức năng huấn luyện model AI.

* Preprocess Engine: Chịu trách nhiệm tiền xử lý dữ liệu đầu vào, bao gồm phân đoạn từ, tách đoạn, kiểm tra nhãn, và tự động chia dataset.
* Training Engine: Nhận dữ liệu đã qua tiền xử lý từ Preprocess Engine và tiến hành huấn luyện model AI.

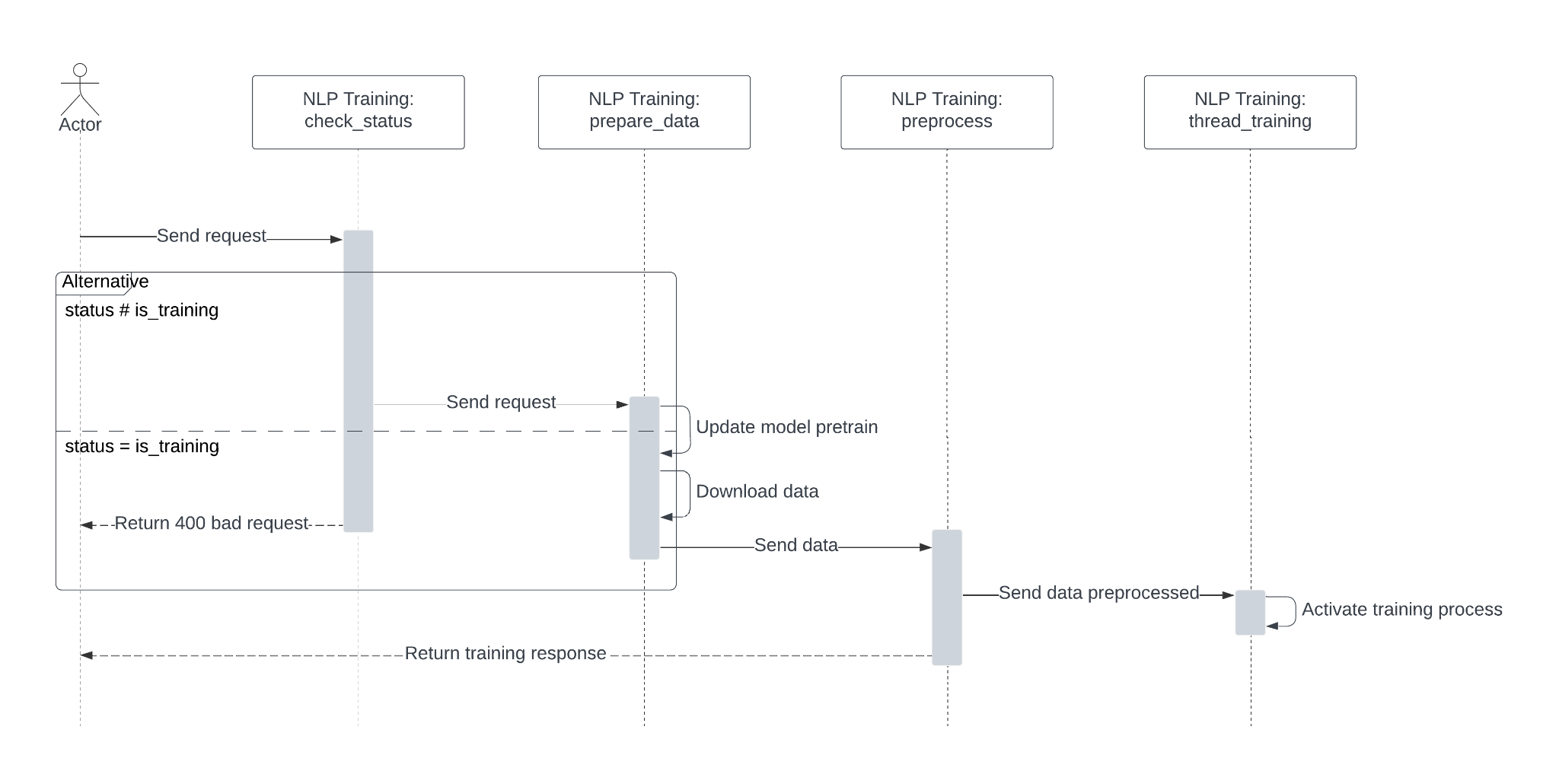
NLP Data: cung cấp chức năng bổ sung dữ liệu huấn luyện.

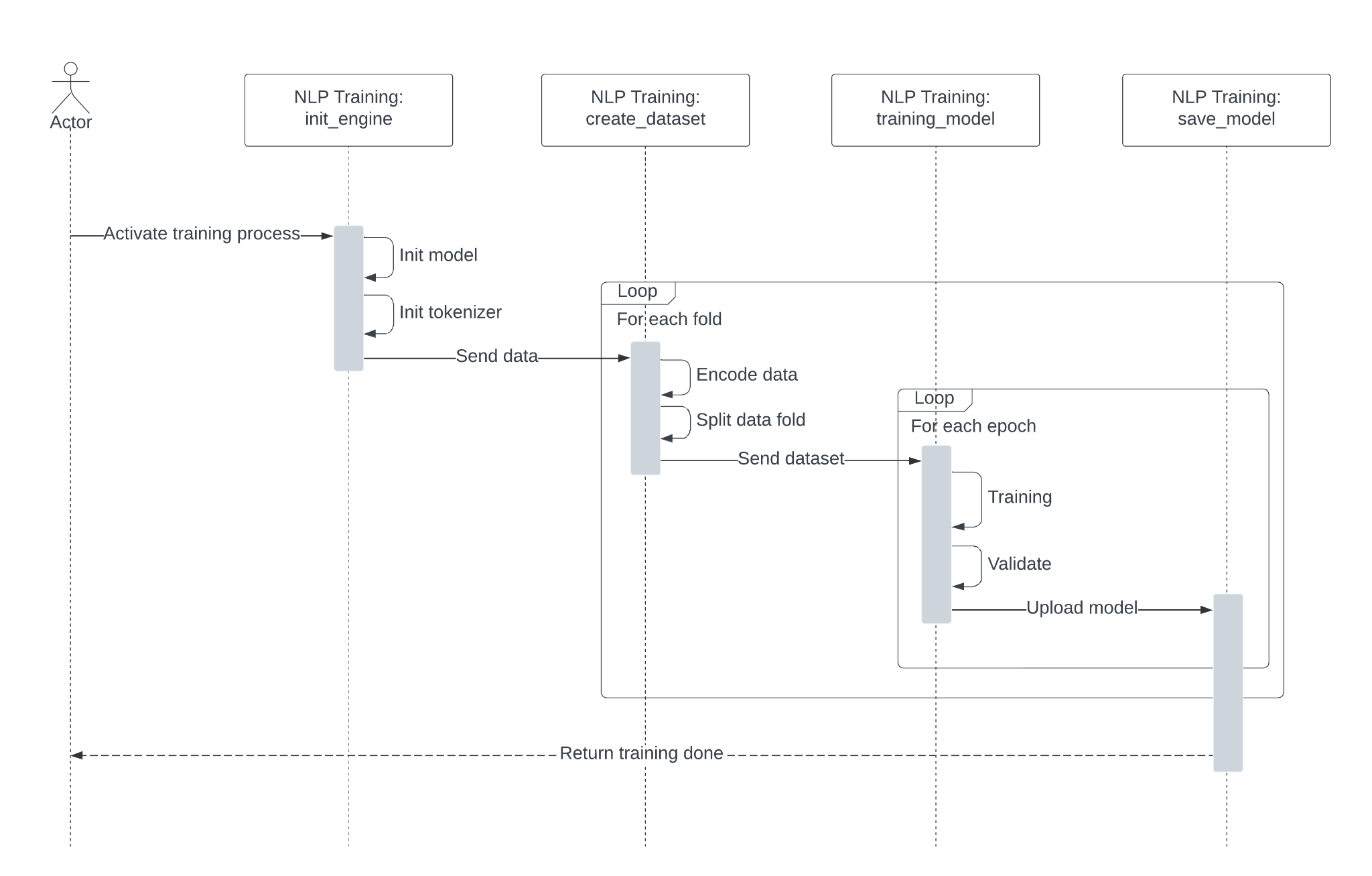
* Data Manager: Chịu trách nhiệm khởi tạo bảng trong database và storage buckets khi lần đầu khởi chạy hệ thống. Cung cấp phương thức thêm dữ liệu vào database.

### 2.2.1.Luồng hoạt động NLP Infer

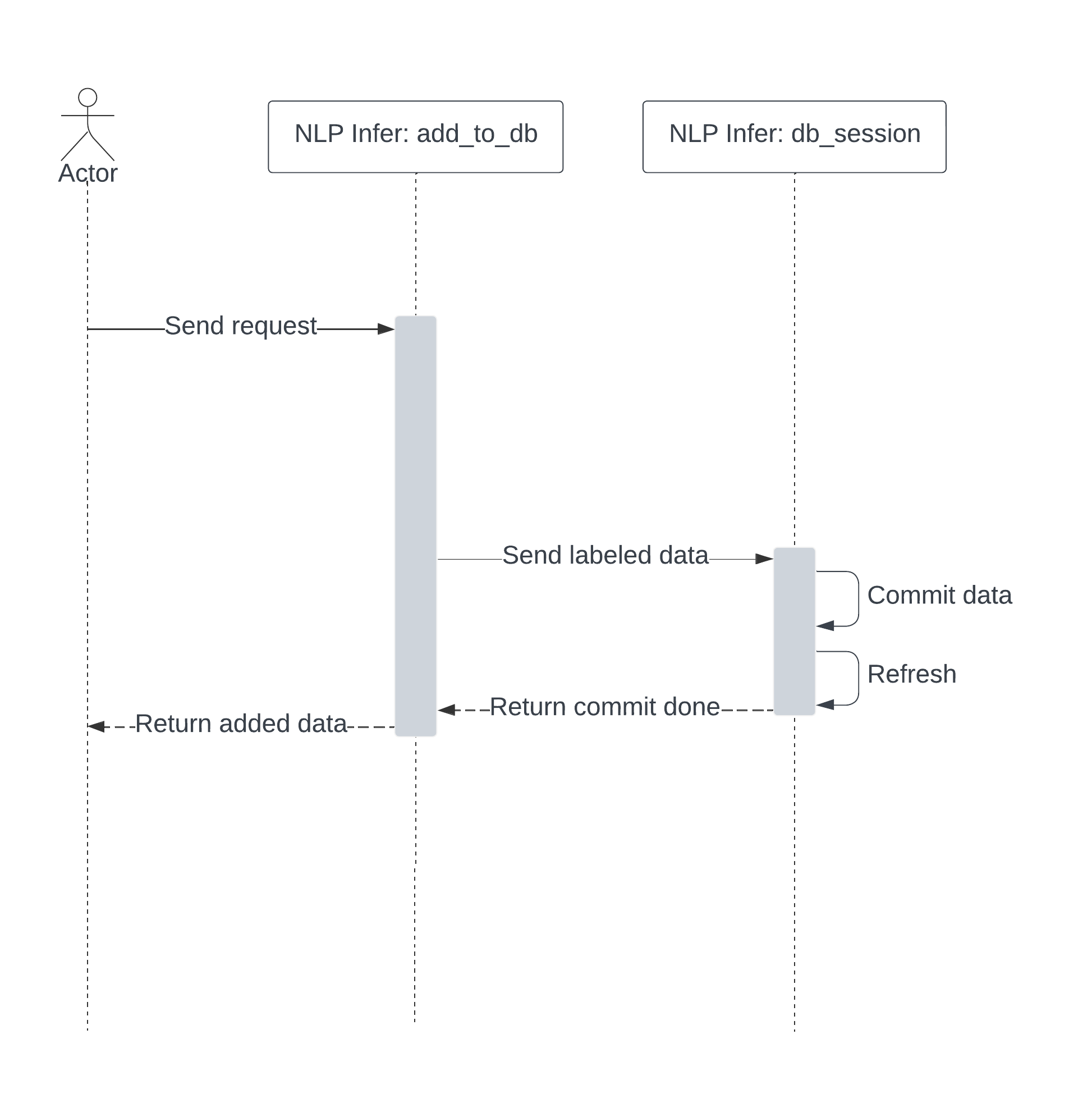


### 2.2.2.Luồng hoạt động NLP Training





### 2.2.3. Luồng hoạt động NLP Data



# Kết luận

# Tài liệu tham khảo

[1] Thành Vũ, Đạt Quốc Nguyễn, Đại Quốc Nguyễn, Mark Dras và Mark Johnson. 2018 . [VnCoreNLP: Bộ công cụ xử lý ngôn ngữ tự nhiên tiếng Việt](http://aclweb.org/anthology/N18-5012) . Trong *Kỷ yếu Hội nghị năm 2018 của Hiệp hội Ngôn ngữ học tính toán Bắc Mỹ: Trình diễn* , [NAACL 2018](http://naacl2018.org/) , trang 56-60. [[.bib]](http://aclweb.org/anthology/N18-5012.bib)

[2] Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. [PhoBERT: Pre-trained language models for Vietnamese](https://aclanthology.org/2020.findings-emnlp.92). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.