BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG

PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI

BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC

BỘ MÔN: MACHINE LEARNING

TÊN ĐỀ TÀI:

Trích xuất thông tin chuyên sâu từ văn bản pháp luật dựa trên mô hình Transform và kỹ thuật Few – shot Learning

Họ và tên sinh viên	**:**	**NGUYỄN THANH HOÀNG**
Mã sinh viên	**:**	**2351267263**
Khóa - Lớp	**:**	**S26_65TTNT**
Giảng viên hướng dẫn		**Cô Vũ Thị Hạnh**

TP. Hồ Chí Minh, ngày … tháng … năm 2026

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày … tháng … năm 2026

Chữ ký của giảng viên

LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin gửi lời cảm ơn chân thành đến Khoa Công nghệ Thông tin đã đưa môn học Khai phá dữ liệu vào chương trình giảng dạy. Đây là một môn học có ý nghĩa thiết thực, giúp em tiếp cận và vận dụng các kiến thức về xử lý dữ liệu, phân tích dữ liệu và ứng dụng trí tuệ nhân tạo vào các bài toán thực tế.

Đặc biệt, em xin bày tỏ lòng biết ơn sâu sắc đến Cô Vũ Thị Hạnh, người đã trực tiếp hướng dẫn, định hướng và hỗ trợ nhóm em trong suốt quá trình thực hiện đề tài. Những góp ý và chỉ dẫn quý báu của Cô đã giúp nhóm em hiểu rõ hơn về bài toán xử lý ngôn ngữ tự nhiên, quy trình tiền xử lý và gán nhãn dữ liệu văn bản pháp luật, cũng như cách xây dựng, huấn luyện và đánh giá các mô hình ngôn ngữ dựa trên kiến trúc Transformer trong bối cảnh học tập với số lượng dữ liệu hạn chế (Few-shot Learning), Nhờ sự hướng dẫn tận tình của Cô, em đã từng bước tiếp cận được phương pháp nghiên cứu khoa học, nâng cao tư duy phân tích và khả năng triển khai các mô hình trí tuệ nhân tạo vào các bài toán thực tiễn liên quan đến văn bản pháp lý.

Trong quá trình thực hiện đề tài, do kiến thức và kinh nghiệm thực tế của em còn hạn chế, bài làm khó tránh khỏi những thiếu sót. Tuy nhiên, em đã cố gắng vận dụng tối đa những kiến thức đã học để xây dựng mô hình, thực nghiệm và phân tích kết quả một cách nghiêm túc. em rất mong nhận được những nhận xét và góp ý từ Cô để bài báo cáo được hoàn thiện hơn và trở thành nền tảng kiến thức hữu ích cho quá trình học tập và công việc sau này.

Em xin chân thành cảm ơn

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

Chữ viết tắt	**Tên đầy đủ (Tiếng Anh)**	**Giải thích (Tiếng Việt)**
AI	Artificial Intelligence	Trí tuệ nhân tạo
ML	Machine Learning	Học máy
DL	Deep Learning	Học sâu
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
NER	Named Entity Recognition	Nhận dạng thực thể có tên
LLM	Large Language Model	Mô hình ngôn ngữ lớn
EDA	Exploratory Data Analysis	Phân tích dữ liệu khám phá
BIO	Begin – Inside – Outside	Chuẩn gán nhãn chuỗi trong NER
PhoBERT	Pre-trained BERT for Vietnamese	Mô hình BERT tiền huấn luyện cho tiếng Việt
BERT	BidirectionalEncoder Representations from Transformers	Mô hình Transformer hai chiều
Transformer	Transformer Architecture	Kiến trúc học sâu dựa trên cơ chế Attention
Attention	Attention Mechanism	Cơ chế chú ý trong mô hình NLP
Tokenization	Tokenization	Quá trình tách văn bản thành token
Subword	Subword Token	Đơn vị từ con trong tokenization

Chữ viết tắt	Tên đầy đủ (Tiếng Anh)	Giải thích (Tiếng Việt)
Fine-tuning	Fine-tuning	Tinh chỉnh mô hình tiền huấn luyện
Rule-based	Rule-based System	Hệ thống suy luận dựa trên luật
Legal Reasoning	Legal Reasoning	Suy luận logic pháp lý
Relation Inference	Relation Inference	Suy luận mối quan hệ giữa thực thể
Summarization	Text Summarization	Tóm tắt văn bản
Clause	Legal Clause	Điều khoản pháp lý
Penalty	Penalty Clause	Điều khoản chế tài
Obligation	Legal Obligation	Nghĩa vụ pháp lý
Party	Contracting Party	Chủ thể trong hợp đồng
Amount	Monetary Amount	Khoản tiền
Deadline	Deadline / Time Limit	Thời hạn
JSON	JavaScript Object Notation	Định dạng dữ liệu JSON
API	Application Programming Interface	Giao diện lập trình ứng dụng
CPU	Central Processing Unit	Bộ xử lý trung tâm
GPU	Graphics Processing Unit	Bộ xử lý đồ họa
RAM	Random Access Memory	Bộ nhớ truy cập ngẫu nhiên

MỞ ĐẦU

Trong bối cảnh công nghệ thông tin và trí tuệ nhân tạo phát triển mạnh mẽ, khối lượng dữ liệu văn bản được tạo ra ngày càng lớn và đa dạng, đặc biệt trong lĩnh vực pháp lý. Các văn bản pháp luật như hợp đồng, nghị định, thông tư hay điều lệ thường có cấu trúc phức tạp, ngôn ngữ mang tính chuyên ngành cao và chứa nhiều thông tin quan trọng liên quan đến quyền, nghĩa vụ và trách nhiệm của các bên. Việc khai thác và xử lý hiệu quả các nguồn dữ liệu này là một thách thức lớn đối với con người khi thực hiện theo phương pháp thủ công.

Trong thực tế, việc đọc hiểu và phân tích các văn bản pháp lý đòi hỏi nhiều thời gian, kiến thức chuyên môn cũng như kinh nghiệm thực tiễn. Các tác vụ như xác định chủ thể tham gia, nghĩa vụ pháp lý, chế tài vi phạm, thời hạn hay các khoản tiền liên quan thường được thực hiện thủ công, dễ dẫn đến sai sót và thiếu nhất quán. Do đó, việc ứng dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), đặc biệt là bài toán nhận dạng thực thể có tên (Named Entity Recognition – NER), đã mở ra hướng tiếp cận mới nhằm tự động hóa quá trình trích xuất thông tin từ văn bản pháp luật.

Xuất phát từ nhu cầu đó, đề tài **"Trích xuất thông tin và tóm tắt điều khoản pháp lý từ hợp đồng dựa trên mô hình Transformer và kỹ thuật Few-shot Learning"** được thực hiện trong khuôn khổ môn học *Machine Learning*. Đề tài tập trung vào việc xây dựng mô hình NER dựa trên PhoBERT để nhận dạng các thực thể pháp lý quan trọng như chủ thể, nghĩa vụ, chế tài, thời hạn và khoản tiền; đồng thời kết hợp với phương pháp suy luận dựa trên luật (Rule-based Legal Reasoning) nhằm tóm tắt nội dung điều khoản một cách ngắn gọn và có ý nghĩa pháp lý. Qua đó, đề tài không chỉ giúp củng cố kiến thức về học máy và xử lý ngôn ngữ tự nhiên mà còn tạo tiền đề cho việc ứng dụng trí tuệ nhân tạo trong lĩnh vực pháp lý và quản lý văn bản pháp luật.

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1.1. Lý do chọn đề tài

Pháp luật là một lĩnh vực quan trọng trong đời sống kinh tế – xã hội, đóng vai trò điều chỉnh các mối quan hệ dân sự, kinh tế và thương mại. Tuy nhiên, các văn bản pháp lý như hợp đồng, điều khoản, nghị định hay thông tư thường có nội dung dài, cấu trúc phức tạp và sử dụng ngôn ngữ chuyên ngành, gây khó khăn cho người đọc trong việc nhanh chóng nắm bắt thông tin cốt lõi. Nếu không được phân tích và xử lý hiệu quả, các điều khoản quan trọng liên quan đến quyền, nghĩa vụ và chế tài có thể bị bỏ sót, dẫn đến rủi ro pháp lý cho các bên liên quan.

Đối với các văn bản hợp đồng, đặc biệt là hợp đồng dân sự và thương mại, những thông tin quan trọng thường tập trung ở các điều khoản như nghĩa vụ của các bên, điều kiện vi phạm, mức phạt, thời hạn thực hiện và các khoản tiền liên quan. Hiện nay, việc phân tích các nội dung này chủ yếu được thực hiện thủ công bởi con người, dựa trên kiến thức pháp lý và kinh nghiệm cá nhân. Cách tiếp cận này không chỉ tốn nhiều thời gian mà còn dễ xảy ra sai sót, đặc biệt khi phải xử lý số lượng lớn văn bản hoặc các hợp đồng có nội dung tương tự nhau.

Trong bối cảnh đó, việc ứng dụng trí tuệ nhân tạo, đặc biệt là các kỹ thuật xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), để tự động trích xuất và phân tích thông tin từ văn bản pháp luật là một hướng nghiên cứu có tính thực tiễn cao. Với sự phát triển của các mô hình học sâu dựa trên kiến trúc Transformer, các bài toán như nhận dạng thực thể có tên (Named Entity Recognition – NER) đã đạt được nhiều kết quả khả quan. Đặc biệt, các mô hình ngôn ngữ tiền huấn luyện cho tiếng Việt như PhoBERT cho phép tận dụng tri thức ngôn ngữ đã được học từ tập dữ liệu lớn, từ đó nâng cao hiệu quả khi áp dụng cho các tập dữ liệu pháp lý có quy mô vừa và nhỏ.

Từ những phân tích trên, nhóm quyết định lựa chọn đề tài "**xây dựng hệ thống trích xuất thông tin và tóm tắt điều khoản pháp lý từ văn bản hợp đồng"** dựa trên mô hình Transformer kết hợp với phương pháp suy luận dựa trên luật (Rule-based Legal Reasoning). Đề tài tập trung vào việc nhận dạng các thực thể pháp lý quan trọng như chủ thể, nghĩa vụ, chế tài, thời hạn và khoản tiền, đồng thời tóm tắt nội dung điều khoản theo logic pháp lý nhằm hỗ trợ người dùng nhanh chóng nắm bắt ý nghĩa của văn bản. Qua đó, hệ thống có thể được ứng dụng trong việc hỗ trợ đọc hiểu hợp đồng, quản lý văn bản pháp luật và giảm thiểu rủi ro trong các hoạt động pháp lý thực tế.

1.2. Mục tiêu của đề tài

Mục tiêu chung đề tài là xây dựng một hệ thống hoàn chỉnh có khả năng tự động trích xuất thông tin và tóm tắt nội dung điều khoản pháp lý từ văn bản hợp đồng dựa trên các kỹ thuật xử lý ngôn ngữ tự nhiên và học sâu.Để đạt được mục tiêu chung đó, đề tài hướng đến các mục tiêu cụ thể sau:

– Nghiên cứu và áp dụng các kỹ thuật xử lý ảnh và tiền xử lý dữ liệu nhằm đảm bảo dữ liệu đầu vào có chất lượng tốt và phù hợp cho việc huấn luyện mô hình học sâu.

– Xây dựng và huấn luyện mô hình nhận dạng thực thể có tên (Named Entity Recognition – NER) dựa trên mô hình ngôn ngữ tiền huấn luyện PhoBERT, nhằm trích xuất các thực thể pháp lý quan trọng như chủ thể (PARTY), nghĩa vụ (OBLIGATION), chế tài (PENALTY), thời hạn (DEADLINE) và khoản tiền (AMOUNT).

– Đánh giá hiệu quả của mô hình NER thông qua các chỉ số phù hợp với bài toán xử lý ngôn ngữ tự nhiên như Precision, Recall và F1-score ở mức thực thể (entity-level), đồng thời phân tích kết quả theo từng loại nhãn để đánh giá độ chính xác của mô hình.

– Xây dựng cơ chế suy luận dựa trên luật (Rule-based Legal Reasoning) nhằm xác định mối quan hệ giữa các thực thể đã trích xuất, từ đó chuyển đổi thông tin rời rạc thành các kết luận pháp lý có logic, phục vụ cho việc tóm tắt điều khoản.

– Thiết kế và triển khai chức năng tóm tắt điều khoản pháp lý, giúp rút gọn nội dung văn bản dài thành các câu ngắn gọn, thể hiện rõ nghĩa vụ của các bên và hậu quả pháp lý trong trường hợp vi phạm.

–Xây dựng chương trình demo hoặc ứng dụng minh họa, cho phép người dùng nhập văn bản hợp đồng và nhận kết quả trích xuất thực thể cũng như bản tóm tắt điều khoản một cách trực quan.

Các mục tiêu trên không chỉ nhằm xây dựng một hệ thống hoạt động đúng về mặt kỹ thuật, mà còn hướng đến việc tạo ra một công cụ có khả năng mở rộng và ứng dụng trong thực tế, hỗ trợ người dùng trong việc đọc hiểu hợp đồng, quản lý văn bản pháp luật và giảm thiểu rủi ro pháp lý.

1.3. Phạm vi và đối tượng nghiên cứu

Đề tài tập trung nghiên cứu bài toán trích xuất thông tin và tóm tắt điều khoản pháp lý từ văn bản hợp đồng, với mục tiêu xây dựng một hệ thống có khả năng tự động nhận diện các thông tin quan trọng và rút gọn nội dung điều khoản một cách logic, dễ hiểu.

Đối tượng nghiên cứu cứu của đề tài là các văn bản hợp đồng dân sự và thương mại, trong đó tập trung vào các điều khoản phổ biến liên quan đến quyền, nghĩa vụ và trách nhiệm pháp lý của các bên tham gia hợp đồng.

Cụ thể, đề tài hướng đến việc nhận dạng và xử lý các nhóm thông tin pháp lý chính sau:

– Chủ thể hợp đồng (PARTY): các bên tham gia ký kết hợp đồng như Bên A, Bên B.
– Nghĩa vụ (OBLIGATION): các nghĩa vụ pháp lý mà các bên phải thực hiện theo hợp đồng.
– Chế tài (PENALTY): các hậu quả pháp lý hoặc hình thức xử lý khi một bên vi phạm nghĩa vụ.

–Khoản tiền (AMOUNT): các khoản tiền liên quan đến hợp đồng như tiền đặt cọc, tiền phạt, tiền bồi thường.

–Thời hạn (DEADLINE): các mốc thời gian thực hiện nghĩa vụ hoặc xử lý hợp đồng.

Việc lựa chọn các nhóm thực thể trên nhằm phản ánh các yếu tố cốt lõi thường xuất hiện trong điều khoản hợp đồng và phù hợp với mục tiêu xây dựng hệ thống trích xuất và tóm tắt nội dung pháp lý.

Phạm vi nghiên cứu của đề tài là:

– Thu thập, kiểm tra và phân tích tập dữ liệu văn bản hợp đồng pháp lý.
– Áp dụng các kỹ thuật tiền xử lý văn bản, bao gồm làm sạch dữ liệu, tách điều khoản và chuyển đổi dữ liệu sang định dạng BIO phục vụ cho huấn luyện mô hình.
– Xây dựng và huấn luyện mô hình nhận dạng thực thể có tên (NER) dựa trên mô hình ngôn ngữ tiền huấn luyện PhoBERT để trích xuất các thực thể pháp lý quan trọng.
– Đánh giá hiệu quả của mô hình NER thông qua các chỉ số Precision, Recall và F1-score ở mức thực thể.
– Xây dựng cơ chế suy luận dựa trên luật (Rule-based Legal Reasoning) để xác định mối quan hệ giữa các thực thể đã trích xuất.

–Thiết kế và triển khai chức năng tóm tắt điều khoản pháp lý dựa trên kết quả trích xuất và suy luận.

–Xây dựng chương trình demo hoặc ứng dụng minh họa cho phép người dùng nhập văn bản hợp đồng và nhận kết quả trích xuất cũng như tóm tắt điều khoản.

Đề tài chỉ tập trung vào việc tự động trích xuất thông tin và hỗ trợ tóm tắt điều khoản, nhằm giúp người dùng tiếp cận nội dung hợp đồng một cách nhanh chóng và thuận tiện hơn.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT LIÊN QUAN

2.1. Tổng quan về xử lý ngôn ngữ tự nhiên trong lĩnh vực pháp lý

Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) là một lĩnh vực thuộc trí tuệ nhân tạo, nghiên cứu các phương pháp giúp máy tính có khả năng hiểu, phân tích và xử lý ngôn ngữ của con người dưới dạng văn bản hoặc lời nói. Trong những năm gần đây, NLP đã đạt được nhiều bước tiến quan trọng nhờ sự phát triển của học sâu và các mô hình ngôn ngữ lớn.

Trong lĩnh vực pháp lý, NLP được ứng dụng để hỗ trợ các công việc như phân tích hợp đồng, trích xuất thông tin pháp lý, tìm kiếm điều khoản, phát hiện rủi ro, và tóm tắt văn bản pháp luật. Văn bản pháp lý thường có đặc điểm dài, cấu trúc phức tạp, nhiều điều kiện ràng buộc và sử dụng ngôn ngữ mang tính chính xác cao. Điều này khiến việc xử lý thủ công trở nên tốn nhiều thời gian và dễ xảy ra sai sót.

Do đó, việc ứng dụng NLP vào lĩnh vực pháp lý không chỉ giúp giảm tải công việc cho con người mà còn nâng cao hiệu quả, tính nhất quán và khả năng mở rộng trong việc xử lý số lượng lớn hợp đồng và tài liệu pháp luật.

2.2. Mô hình ngôn ngữ và học sâu trong xử lý văn bản

Mô hình ngôn ngữ là nền tảng quan trọng trong các bài toán xử lý ngôn ngữ tự nhiên. Mục tiêu của mô hình ngôn ngữ là học được cách biểu diễn và dự đoán chuỗi từ trong ngôn ngữ tự nhiên dựa trên ngữ cảnh.

Trước đây, các phương pháp truyền thống như Bag-of-Words hoặc TF-IDF chỉ xem văn bản như tập hợp các từ rời rạc, không thể nắm bắt được ngữ nghĩa và mối quan hệ ngữ cảnh giữa các từ. Sự ra đời của học sâu đã giúp giải quyết hạn chế này thông qua các biểu diễn từ (word embeddings) và các mạng nơ-ron nhiều lớp.

Học sâu trong xử lý văn bản cho phép mô hình tự động học các đặc trưng ngôn ngữ ở nhiều mức độ khác nhau, từ cú pháp đến ngữ nghĩa. Các mô hình hiện đại có khả

năng hiểu ngữ cảnh dài, quan hệ giữa các thực thể và ý nghĩa tổng thể của câu hoặc đoạn văn.

Trong đề tài này, học sâu đóng vai trò cốt lõi trong việc xây dựng mô hình trích xuất thông tin và làm nền tảng cho các bước suy luận và tóm tắt điều khoản pháp lý.

2.3. Mô hình Transformer và kiến trúc Attention

Transformer là một kiến trúc mạng nơ-ron được giới thiệu nhằm khắc phục những hạn chế của các mô hình tuần tự như RNN hay LSTM trong việc xử lý chuỗi dài. Điểm nổi bật của Transformer là cơ chế Attention, cho phép mô hình tập trung vào những phần quan trọng của chuỗi đầu vào khi xử lý một từ hoặc một vị trí cụ thể.

Cơ chế Self-Attention cho phép mỗi từ trong câu có thể "chú ý" đến tất cả các từ khác trong cùng câu, từ đó học được mối quan hệ ngữ nghĩa toàn cục. Điều này đặc biệt quan trọng đối với văn bản pháp lý, nơi các điều kiện và chế tài có thể liên quan đến các cụm từ xuất hiện ở xa nhau trong câu hoặc đoạn văn.

Transformer thường được cấu thành từ các khối Encoder và Decoder. Trong các bài toán trích xuất thông tin và phân loại chuỗi, kiến trúc Encoder đóng vai trò chính trong việc mã hóa ngữ cảnh của văn bản đầu vào.

Nhờ khả năng xử lý song song và học ngữ cảnh hiệu quả, Transformer đã trở thành kiến trúc chủ đạo trong các mô hình NLP hiện đại.

2.4. Bài toán trích xuất thực thể (Named Entity Recognition) trong văn bản pháp lý

Trích xuất thực thể có tên (Named Entity Recognition – NER) là bài toán xác định và phân loại các thực thể quan trọng trong văn bản như tên chủ thể, nghĩa vụ, khoản tiền, thời hạn hoặc chế tài.

Phương pháp này mang lại nhiều lợi ích như giảm thời gian huấn luyện, hạn chế yêu cầu về dung lượng dữ liệu và giúp mô hình đạt độ chính xác cao hơn so với việc huấn luyện từ đầu. Do đó, học chuyển giao đặc biệt phù hợp với các bài toán có tập dữ liệu vừa và nhỏ như trong đề tài này.

Trích xuất thực thể có tên (Named Entity Recognition – NER) là bài toán xác định và phân loại các thực thể quan trọng trong văn bản như tên chủ thể, nghĩa vụ, khoản tiền, thời hạn hoặc chế tài.

NER thường được triển khai dưới dạng bài toán gán nhãn chuỗi, trong đó mỗi từ được gán một nhãn theo chuẩn BIO (Begin – Inside – Outside). Cách biểu diễn này giúp mô hình nhận biết được ranh giới của các thực thể nhiều từ.

Trong đề tài, NER được sử dụng để trích xuất các nhóm thực thể pháp lý chính như PARTY, OBLIGATION, AMOUNT, DEADLINE và PENALTY, làm cơ sở cho bước suy luận và tóm tắt điều khoản.

2.5. Kỹ thuật Few-shot Learning trong xử lý ngôn ngữ tự nhiên

Few-shot Learning là kỹ thuật cho phép mô hình học và tổng quát hóa từ một số lượng rất nhỏ dữ liệu huấn luyện. Đây là hướng tiếp cận đặc biệt phù hợp với lĩnh vực pháp lý, nơi việc thu thập và gán nhãn dữ liệu thường tốn nhiều chi phí và yêu cầu kiến thức chuyên môn.

Thay vì huấn luyện mô hình từ đầu với lượng dữ liệu lớn, Few-shot Learning tận dụng các mô hình ngôn ngữ đã được tiền huấn luyện trên tập dữ liệu quy mô lớn, sau đó tinh chỉnh với một số ít mẫu đại diện cho bài toán cụ thể.

Trong đề tài này, kỹ thuật Few-shot Learning được áp dụng khi huấn luyện mô hình NER với tập dữ liệu BIO có kích thước nhỏ. Nhờ kiến thức ngôn ngữ đã học trước, mô hình vẫn có khả năng nhận diện các thực thể pháp lý với độ chính xác tương đối tốt, dù số lượng mẫu huấn luyện hạn chế.

2.6. Mô hình Transformer tiền huấn luyện cho tiếng Việt và văn bản pháp lý

Đối với tiếng Việt, các mô hình Transformer tiền huấn luyện như PhoBERT đã chứng minh hiệu quả cao trong nhiều bài toán NLP. PhoBERT được huấn luyện trên tập dữ liệu lớn gồm các văn bản tiếng Việt, giúp mô hình học được đặc trưng ngôn ngữ và ngữ cảnh một cách toàn diện.

Trong đề tài, PhoBERT được sử dụng làm mô hình nền (backbone) cho bài toán trích xuất thực thể. Quá trình fine-tuning được thực hiện bằng cách bổ sung lớp phân loại token và huấn luyện lại trên dữ liệu pháp lý đã được gán nhãn.

Việc sử dụng mô hình tiền huấn luyện giúp giảm thời gian huấn luyện, cải thiện độ chính xác và tăng khả năng tổng quát hóa của hệ thống khi áp dụng cho các điều khoản pháp lý khác nhau.

2.7. Các chỉ số đánh giá cho bài toán trích xuất và tóm tắt văn bản

Để đánh giá hiệu quả của hệ thống trích xuất thông tin, đề tài sử dụng các chỉ số phổ biến trong bài toán NER như Precision, Recall và F1-score ở mức thực thể.

Precision phản ánh mức độ chính xác của các thực thể được mô hình dự đoán, trong khi Recall thể hiện khả năng phát hiện đầy đủ các thực thể có trong văn bản. F1-score là trung bình điều hòa của Precision và Recall, giúp đánh giá sự cân bằng giữa hai chỉ số này.

Ngoài ra, kết quả tóm tắt điều khoản được đánh giá định tính thông qua các ví dụ thực tế, so sánh nội dung tóm tắt với văn bản gốc nhằm kiểm tra tính đúng ngữ nghĩa và logic pháp lý.

Việc kết hợp đánh giá định lượng và định tính giúp đưa ra cái nhìn toàn diện về chất lượng của hệ thống trích xuất và tóm tắt điều khoản pháp lý.

F1-Score là trung bình điều hòa của Precision và Recall, giúp đánh giá sự cân bằng giữa hai chỉ số này.

Việc sử dụng đồng thời nhiều chỉ số giúp đánh giá mô hình một cách khách quan và toàn diện hơn, tránh phụ thuộc vào một thước đo duy nhất.

CHƯƠNG 3: MÔ TẢ DỮ LIỆU VÀ TIỀN XỬ LÝ

3.1. Nguồn và đặc điểm tập dữ liệu

Tập dữ liệu được sử dụng trong đề tài là dữ liệu văn bản pháp lý, cụ thể là các điều khoản được trích xuất từ hợp đồng dân sự và hợp đồng thuê nhà. Dữ liệu ban đầu ở dạng thô được thu thập từ các tệp văn bản và tài liệu định dạng .docx, sau đó được chuyển đổi sang văn bản thuần (.txt) để phục vụ cho quá trình xử lý và gán nhãn.

Theo Sau bước tiền xử lý và phân tách điều khoản, dữ liệu được gán nhãn thủ công theo chuẩn BIO nhằm phục vụ bài toán nhận dạng thực thể có tên (Named Entity Recognition – NER). Các thực thể được gán nhãn trong đề tài bao gồm:

– PARTY: Chủ thể tham gia hợp đồng (Bên A, Bên B).
– OBLIGATION: Nghĩa vụ pháp lý.
– AMOUNT: Số tiền, giá trị tài chính.
– DEADLINE: Thời hạn, mốc thời gian.
– PENALTY: Chế tài, hậu quả pháp lý.

Dữ liệu sau khi gán nhãn được lưu dưới dạng tệp train.bio, trong đó mỗi dòng tương ứng với một token và nhãn BIO đi kèm, các câu được phân tách bằng dòng trống. Đây là định dạng tiêu chuẩn thường được sử dụng để huấn luyện các mô hình NER dựa trên Transformer như PhoBERT.

Theo thống kê từ tệp train.bio, tập dữ liệu huấn luyện bao gồm nhiều điều khoản pháp lý khác nhau, với số lượng thực thể được phân bố trên các nhóm nhãn. Mặc dù số lượng dữ liệu không lớn, nhưng các thực thể đều thuộc cùng miền pháp lý, giúp mô hình học được các mẫu ngữ nghĩa đặc trưng của hợp đồng. Do quy mô dữ liệu hạn chế, đề tài lựa chọn hướng tiếp cận fine-tuning mô hình PhoBERT kết hợp với chiến lược Few-shot Learning, nhằm tận dụng tri thức đã được học từ dữ liệu tiếng Việt quy mô lớn.

Việc sử dụng dữ liệu văn bản đã được gán nhãn thủ công giúp đảm bảo độ chính xác của nhãn, đồng thời tạo nền tảng cho việc đánh giá khả năng trích xuất thông tin và tóm tắt điều khoản pháp lý của hệ thống.

3.2. Cấu trúc thư mục dữ liệu

Dữ liệu trong đề tài được tổ chức theo cấu trúc thư mục phân cấp nhằm phục vụ cho bài toán trích xuất thông tin và tóm tắt điều khoản pháp lý từ hợp đồng, cụ thể như sau:

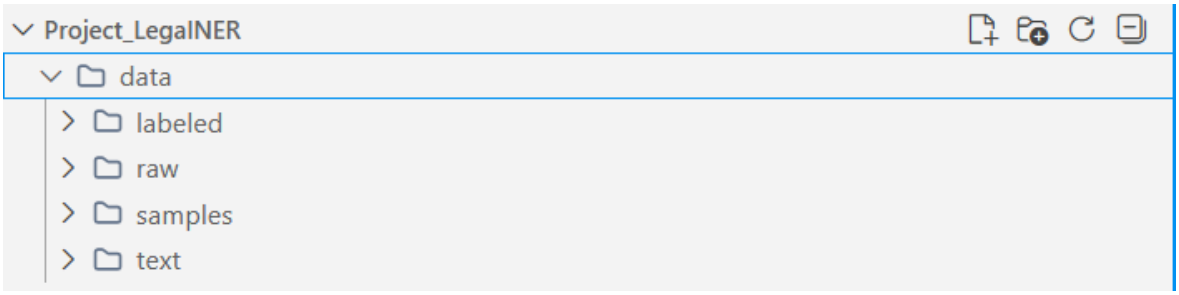
– Thư mục data/raw/ Lưu trữ các tệp văn bản thuần (.txt) sau khi được chuyển đổi từ dữ liệu thô. Dữ liệu trong thư mục này đã được chuẩn hóa định dạng, loại bỏ các ký tự không cần thiết và sẵn sàng cho các bước xử lý tiếp theo.

– Thư mục data/samples/ chứa các đoạn điều khoản hoặc câu văn pháp lý đã được tách ra từ văn bản gốc. Các mẫu dữ liệu này được sử dụng cho việc kiểm tra, minh họa, và gán nhãn thử nghiệm trong quá trình xây dựng mô hình.

– Thư mục data/labeled Chứa dữ liệu đã được gán nhãn thủ công theo chuẩn BIO, phục vụ cho việc huấn luyện mô hình NER. Tệp quan trọng nhất trong thư mục này là train.bio, trong đó mỗi dòng tương ứng với một token và nhãn thực thể đi kèm, các câu được phân tách bằng dòng trống.

Cấu Cấu trúc dữ liệu này được sử dụng xuyên suốt trong toàn bộ project, phù hợp với pipeline xử lý dữ liệu gồm các bước: tiền xử lý văn bản → phân tách điều khoản → gán nhãn thủ công → chuyển đổi sang định dạng BIO → huấn luyện mô hình PhoBERT.

Việc tổ chức dữ liệu rõ ràng theo từng giai đoạn giúp quá trình phát triển, huấn luyện và mở rộng hệ thống trở nên thuận tiện và dễ quản lý hơn.



Hình 3.1 Cấu trúc thư mục dữ liệu

3.3. Tiền xử lý văn bản pháp lý

Trước khi tiến hành huấn luyện mô hình trích xuất thực thể (NER), nhóm thực hiện bước kiểm tra và làm sạch dữ liệu nhằm đảm bảo chất lượng dữ liệu đầu vào, đặc biệt là dữ liệu văn bản và dữ liệu gán nhãn theo chuẩn BIO.

Quá trình kiểm tra dữ liệu được triển khai trong giai đoạn tiền xử lý, thông qua các script xử lý văn bản trong module preprocessing.py. Các bước chính bao gồm:
– Loại Kiểm tra tính hợp lệ của dữ liệu văn bản, đảm bảo các tệp văn bản được đọc đúng định dạng UTF-8, không bị lỗi mã hóa hoặc mất nội dung trong quá trình chuyển đổi từ file gốc.
– Loại bỏ các ký tự không cần thiết, chẳng hạn như ký tự đặc biệt, khoảng trắng dư thừa, ký hiệu định dạng hoặc các thành phần không mang ý nghĩa ngữ nghĩa trong văn bản pháp lý.
– Chuẩn hóa văn bản, bao gồm chuẩn hóa dấu câu, chữ hoa – chữ thường và cách viết các thuật ngữ pháp lý phổ biến nhằm giảm nhiễu cho mô hình học máy.

–Chuẩn hóa văn bản, bao gồm chuẩn hóa dấu câu, chữ hoa – chữ thường và cách viết các thuật ngữ pháp lý phổ biến nhằm giảm nhiễu cho mô hình học máy.

–Kiểm tra dữ liệu gán nhãn BIO, đảm bảo mỗi token đều có nhãn tương ứng, các nhãn tuân thủ đúng định dạng BIO (B-, *I*-, O) và không xuất hiện nhãn không hợp lệ.

Sau quá trình kiểm tra, các dòng dữ liệu không hợp lệ hoặc không nhất quán được loại bỏ hoặc chỉnh sửa thủ công. Dữ liệu sau khi làm sạch được lưu lại dưới dạng tệp train.bio trong thư mục data/labeled/, sẵn sàng cho bước huấn luyện mô hình PhoBERT.

Việc kiểm tra và làm sạch dữ liệu giúp đảm bảo mô hình không bị ảnh hưởng bởi nhiễu hoặc lỗi gán nhãn, đồng thời nâng cao độ ổn định và độ chính xác trong quá trình huấn luyện và suy luận.

3.4. Phân tích dữ liệu khám phá (EDA)

Phân tích dữ liệu khám phá (Exploratory Data Analysis – EDA) được thực hiện nhằm hiểu rõ hơn về đặc điểm của tập dữ liệu văn bản pháp lý trước khi tiến hành huấn luyện mô hình PhoBERT cho bài toán nhận dạng thực thể (Named Entity Recognition – NER). Các bước EDA được cài đặt trong file eda.py và tập trung vào hai khía cạnh chính: phân bố thực thể pháp lý và độ dài câu trong tập dữ liệu.

Hình đầu tiên minh họa phân bố số lần xuất hiện của các thực thể pháp lý trong tập dữ liệu, không bao gồm nhãn O (Outside). Các thực thể được xem xét gồm: PARTY, OBLIGATION, AMOUNT, DEADLINE và PENALTY.

– Thực thể OBLIGATION xuất hiện nhiều nhất, phản ánh đặc trưng của văn bản hợp đồng, trong đó các nghĩa vụ pháp lý được mô tả với tần suất cao.
– Thực thể AMOUNT và PARTY cũng xuất hiện với số lượng lớn, cho thấy các điều khoản thường xuyên đề cập đến chủ thể tham gia hợp đồng và các khoản tiền liên quan.
– Thực thể DEADLINE có tần suất thấp hơn nhưng vẫn xuất hiện đều đặn, thể hiện các mốc thời gian quan trọng trong hợp đồng.

–Thực thể PENALTY có số lần xuất hiện thấp nhất, do chế tài thường chỉ được quy định trong một số điều khoản cụ thể.

Nhìn chung, phân bố các thực thể tương đối hợp lý và phản ánh đúng cấu trúc của văn bản hợp đồng pháp lý. Tuy nhiên, vẫn tồn tại sự chênh lệch nhất định giữa các loại thực thể, điều này có thể ảnh hưởng đến khả năng học của mô hình đối với các nhãn ít xuất hiện như PENALTY và DEADLINE.

Hình thứ hai thể hiện **phân bố độ dài câu trong tập dữ liệu**, được đo bằng số lượng token sau khi tiền xử lý và tách từ. Trục hoành biểu diễn độ dài câu, trục tung biểu diễn số lượng câu tương ứng. Đường nét đứt màu đỏ biểu thị **giá trị trung bình**, khoảng **52.8 token/câu.**

Từ biểu đồ có thể nhận thấy:
– Phần lớn các câu có độ dài dao động trong khoảng từ 20 đến 80 token.
– Một số câu có độ dài lớn hơn 90 token, thường là các điều khoản pháp lý dài, chứa nhiều mệnh đề điều kiện và chế tài.
– Độ dài câu trung bình ở mức vừa phải, phù hợp với giới hạn độ dài đầu vào (128 token) được sử dụng trong quá trình huấn luyện mô hình PhoBERT.

Phân bố độ dài câu này cho thấy việc lựa chọn ngưỡng cắt/padding ở mức 128 token là hợp lý, giúp mô hình giữ được đầy đủ thông tin ngữ nghĩa quan trọng trong hầu hết các điều khoản mà không gây lãng phí tài nguyên tính toán.

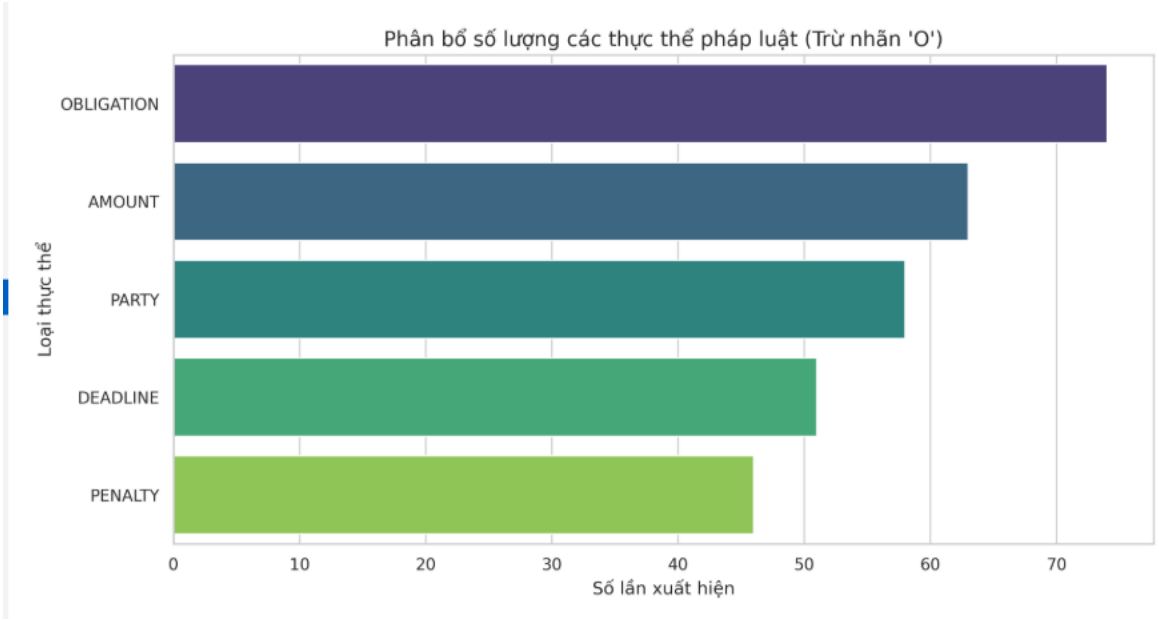
Thông qua bước phân tích dữ liệu khám phá, có thể rút ra một số nhận xét quan trọng:

–Tập dữ liệu có cấu trúc phù hợp cho bài toán NER trong lĩnh vực pháp lý.

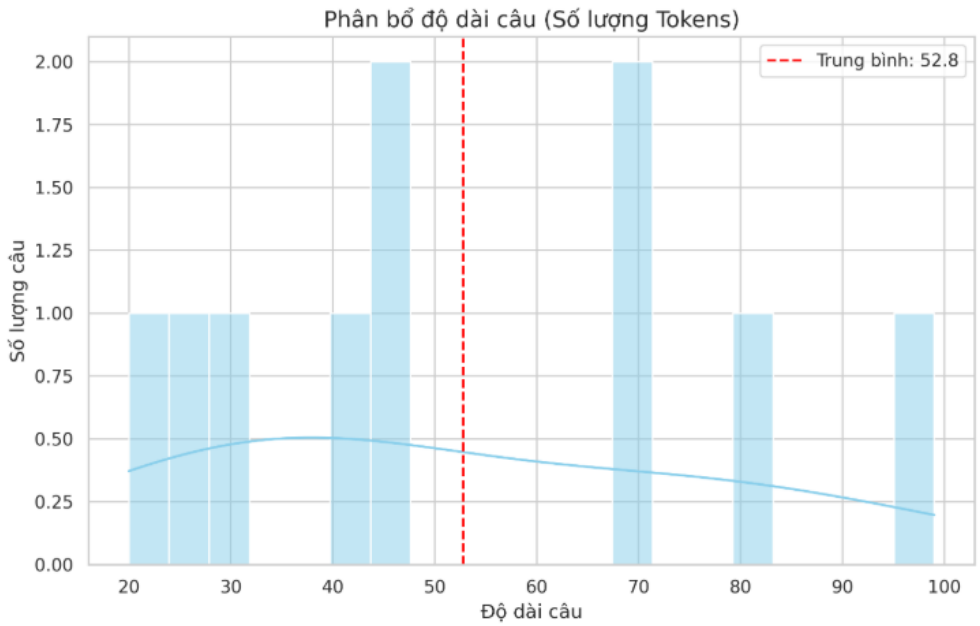
–Các thực thể quan trọng như chủ thể, nghĩa vụ, khoản tiền và chế tài đều xuất hiện với tần suất đủ lớn để mô hình học được đặc trưng.

–Độ dài câu tương đối đồng đều và nằm trong phạm vi xử lý hiệu quả của mô hình Transformer.

Những kết quả từ EDA là cơ sở quan trọng để thiết kế chiến lược tiền xử lý, lựa chọn cấu hình mô hình và xây dựng các luật suy luận pháp lý ở các bước tiếp theo.



Hình 3.2 Phân bổ số lượng các thực thể pháp luật



Hình 3 3 Phân bổ độ dài câu

3.5. Phân tách điều khoản và chuẩn hóa văn bản

Trước khi tiến hành gán nhãn và huấn luyện mô hình nhận dạng thực thể pháp lý, dữ liệu văn bản hợp đồng cần được phân tách hợp lý theo điều khoản và chuẩn hóa về mặt hình thức nhằm đảm bảo tính nhất quán và giảm nhiễu cho mô hình học sâu.

Dữ liệu văn bản thô trong đề tài được thu thập từ các hợp đồng ở định dạng văn bản (hoặc chuyển đổi từ file DOCX/PDF). Các hợp đồng này thường bao gồm nhiều điều khoản dài, mỗi điều khoản có thể chứa nhiều câu và mệnh đề pháp lý phức tạp.

Do đó, nhóm thực hiện bước phân tách hợp đồng thành các điều khoản riêng biệt, trong đó mỗi điều khoản được xem như một đơn vị dữ liệu độc lập. Việc phân tách dựa trên các đặc trưng hình thức phổ biến trong văn bản pháp lý như:

– Số thứ tự điều khoản (ví dụ: “Điều 4.”, “4.2.”).
– Dấu xuống dòng và cấu trúc đoạn văn.
– Các từ khóa mở đầu điều khoản như “Nếu”, “Trong trường hợp”, “Bên A”, “Bên B”.
Quá trình này được cài đặt trong module tiền xử lý, giúp đảm bảo mỗi mẫu dữ liệu đầu vào cho mô hình PhoBERT tương ứng với một điều khoản pháp lý hoàn chỉnh, thuận lợi cho cả bài toán NER và tóm tắt.

Sau khi phân tách điều khoản, văn bản tiếp tục được **chuẩn hóa** để giảm sự đa dạng không cần thiết trong cách biểu diễn ngôn ngữ, từ đó giúp mô hình học tốt hơn.

Các bước chuẩn hóa chính bao gồm:

–Chuẩn hóa khoảng trắng và xuống dòng: loại bỏ khoảng trắng thừa, gộp các dòng bị ngắt không cần thiết.

–Chuẩn hóa chữ hoa – chữ thường: giữ nguyên chữ hoa trong các thực thể pháp lý quan trọng (ví dụ: “Bên A”, “Bên B”), đồng thời chuẩn hóa các từ viết thường để đảm bảo tính nhất quán.

–Loại bỏ ký tự đặc biệt không cần thiết: như ký hiệu trang, dấu chấm kéo dài, hoặc các ký tự trang trí không mang ý nghĩa pháp lý.

–Giữ nguyên dấu câu quan trọng: các dấu như dấu phẩy, dấu chấm, dấu chấm phẩy được giữ lại vì có vai trò trong việc phân tách mệnh đề và suy luận ngữ nghĩa pháp lý.

Giữ nguyên dấu câu quan trọng: các dấu như dấu phẩy, dấu chấm, dấu chấm phẩy được giữ lại vì có vai trò trong việc phân tách mệnh đề và suy luận ngữ nghĩa pháp lý.

Giữ nguyên dấu câu quan trọng: các dấu như dấu phẩy, dấu chấm, dấu chấm phẩy được giữ lại vì có vai trò trong việc phân tách mệnh đề và suy luận ngữ nghĩa pháp lý.

–Giúp dữ liệu đầu vào có cấu trúc rõ ràng, phù hợp với cách tiếp cận token-level của mô hình Transformer.

–Giảm sự đa dạng không cần thiết trong biểu diễn văn bản, từ đó cải thiện khả năng tổng quát hóa của mô hình.

–Giảm sự đa dạng không cần thiết trong biểu diễn văn bản, từ đó cải thiện khả năng tổng quát hóa của mô hình.

3.6. Gán nhãn thực thể theo chuẩn BIO

Sau khi văn bản hợp đồng được phân tách thành các điều khoản độc lập và chuẩn hóa về mặt hình thức, bước tiếp theo là **gán nhãn thực thể pháp lý theo chuẩn BIO**

nhằm tạo dữ liệu huấn luyện cho mô hình nhận dạng thực thể (Named Entity Recognition – NER).

Dựa trên mục tiêu của đề tài và đặc trưng của văn bản hợp đồng, nhóm xác định năm loại thực thể pháp lý chính cần trích xuất:

–**PARTY**: Chủ thể tham gia hợp đồng (ví dụ: Bên A, Bên B).

–**OBLIGATION**: Nghĩa vụ pháp lý mà các bên phải thực hiện.

–**AMOUNT**: Các khoản tiền hoặc giá trị tài chính được đề cập trong hợp đồng.

–**DEADLINE**: Thời hạn hoặc mốc thời gian liên quan đến nghĩa vụ hoặc quyền lợi.

–**PENALTY**: Chế tài, hậu quả pháp lý khi một bên vi phạm nghĩa vụ.

Những thực thể này đóng vai trò cốt lõi trong việc phân tích và tóm tắt nội dung điều khoản pháp lý.

Đề tài sử dụng chuẩn gán nhãn **BIO (Begin – Inside – Outside)**, là chuẩn phổ biến trong các bài toán NER. Theo chuẩn này:

–**B-XXX**: token bắt đầu của một thực thể loại XXX.

–**I-XXX**: token nằm bên trong thực thể loại XXX.

–**O**: token không thuộc bất kỳ thực thể nào.

3.7. Xây dựng tập dữ liệu huấn luyện cho mô hình NER

Sau khi hoàn tất quá trình gán nhãn thực thể theo chuẩn BIO, dữ liệu văn bản được sử dụng để xây dựng **tập dữ liệu huấn luyện cho mô hình nhận dạng thực thể pháp lý (NER)** dựa trên kiến trúc Transformer.

Dữ liệu huấn luyện được lưu dưới dạng file BIO (train.bio), trong đó mỗi dòng tương ứng với một token và nhãn BIO của token đó. Các câu hoặc điều khoản được phân tách bằng một dòng trống.

```
64    Bên B-PARTY
65    A   I-PARTY
66    sẽ  O
67    không   B-PENALTY
68    phải    I-PENALTY
69    hoàn    I-PENALTY
70    trả I-PENALTY
71    lại I-PENALTY
72    Bên B-PARTY
```

Mô hình PhoBERT sử dụng cơ chế **SentencePiece subword tokenization**, trong đó một từ có thể bị tách thành nhiều subword. Do đó, cần thực hiện bước **căn chỉnh nhãn (label alignment)** giữa token gốc và các subword.

Do mô hình Transformer có giới hạn về độ dài đầu vào, mỗi câu/điều khoản được:

- Cắt (truncate) hoặc đệm (padding) về độ dài cố định (ví dụ: 128 token).

- Sử dụng attention_mask để phân biệt token thật và token padding.

Việc chuẩn hóa độ dài chuỗi giúp:

- Ổn định quá trình huấn luyện.

- Tối ưu bộ nhớ và thời gian tính toán.

- Phù hợp với kiến trúc PhoBERT.

3.8. Nhận xét và đánh giá dữ liệu

Nhóm xây dựng bộ sinh dữ liệu (Data Generator) chuyên biệt cho bài toán trích xuất thông tin thực thể (NER) từ văn bản pháp luật, nhằm phục vụ quá trình huấn luyện và đánh giá mô hình Transformer trong điều kiện dữ liệu gán nhãn còn hạn chế.

Cụ thể, dữ liệu được tổ chức theo định dạng BIO, trong đó mỗi dòng tương ứng với một token và nhãn thực thể đi kèm. Bộ sinh dữ liệu có các nhiệm vụ chính như sau:

– Đọc và duyệt qua các tệp dữ liệu văn bản đã được gán nhãn theo chuẩn BIO để thu thập chuỗi token và nhãn tương ứng.
– Thực hiện tiền xử lý văn bản, bao gồm chuẩn hóa ký tự, tách từ, mã hóa token bằng tokenizer của mô hình Transformer (ví dụ: PhoBERT), đồng thời xử lý các token đặc biệt như [CLS], [SEP].
– Chuyển đổi nhãn thực thể sang chỉ số số nguyên hoặc dạng one-hot phù hợp với bài toán phân loại chuỗi (sequence labeling).
– Thực hiện padding và masking nhằm đảm bảo các câu trong cùng một batch có độ dài thống nhất, đồng thời tránh ảnh hưởng của các token đệm trong quá trình huấn luyện.
– Chia dữ liệu thành các batch để đưa vào mô hình, giúp tối ưu bộ nhớ và tăng hiệu quả huấn luyện.

Nhìn chung, cách tổ chức và xây dựng bộ sinh dữ liệu phù hợp với đặc thù của văn bản pháp luật, giúp mô hình học được mối quan hệ ngữ nghĩa giữa các thực thể, đồng thời hỗ trợ hiệu quả cho việc áp dụng kỹ thuật Few-shot Learning trong bối cảnh dữ liệu huấn luyện hạn chế.

.

3.9. Kết luận chương

Trong chương này, đề tài đã trình bày chi tiết về tập dữ liệu văn bản pháp luật được sử dụng cho bài toán trích xuất thông tin thực thể, bao gồm cấu trúc dữ liệu, định dạng gán nhãn theo chuẩn BIO, cũng như kết quả thống kê và phân tích dữ liệu khám phá ban đầu. Bên cạnh đó, quy trình kiểm tra, làm sạch và tiền xử lý dữ liệu văn bản cũng được mô tả cụ thể, bám sát các module triển khai thực tế trong project.

Các bước tiền xử lý như chuẩn hóa văn bản, tách từ, mã hóa token bằng tokenizer của mô hình Transformer, padding và masking đóng vai trò quan trọng trong việc đảm bảo dữ liệu đầu vào phù hợp với kiến trúc mô hình và giữ được ngữ nghĩa của văn bản pháp luật. Việc xây dựng bộ sinh dữ liệu chuyên biệt cho bài toán NER giúp quá trình huấn luyện diễn ra ổn định, tiết kiệm tài nguyên tính toán và đảm bảo tính nhất quán giữa giai đoạn huấn luyện và đánh giá.

Những nội dung được trình bày trong chương này đóng vai trò là nền tảng quan trọng cho các chương tiếp theo, nơi các mô hình Transformer kết hợp với kỹ thuật Few-shot Learning sẽ được xây dựng, huấn luyện và đánh giá nhằm nâng cao hiệu quả trích xuất thông tin chuyên sâu từ văn bản pháp luật trong điều kiện dữ liệu gán nhãn còn hạn chế.

CHƯƠNG 4: PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU VÀ MÔ HÌNH ML

4.1. Tổng quan quy trình xây dựng mô hình

Dựa trên toàn bộ mã nguồn trong project, quá trình xây dựng mô hình được thiết kế thống nhất cho bài toán trích xuất thông tin thực thể (NER) từ văn bản pháp luật, sử dụng kiến trúc Transformer làm nền tảng và kết hợp với kỹ thuật Few-shot Learning nhằm giải quyết bài toán thiếu dữ liệu gán nhãn.

Tất cả các mô hình đều sử dụng chung quy trình tiền xử lý và bộ sinh dữ liệu được xây dựng trong tệp preprocessing, nhằm đảm bảo tính nhất quán giữa các giai đoạn huấn luyện, đánh giá và suy luận.

Quy trình chung của mỗi mô hình gồm các bước:

– Tự động xác định và tải các tệp dữ liệu huấn luyện (train) và ở định dạng BIO.
– Tiến hành tiền xử lý văn bản, bao gồm chuẩn hóa dữ liệu, mã hóa token bằng tokenizer của mô hình Transformer, padding và masking.
– Xây dựng kiến trúc mô hình Transformer cho bài toán gán nhãn chuỗi (sequence labeling), trong đó tầng đầu ra được thiết kế phù hợp với số lượng nhãn thực thể.
– Thiết lập hàm mất mát (loss), bộ tối ưu (optimizer) và các callback cần thiết như early stopping và điều chỉnh tốc độ học.
– Huấn luyện mô hình theo từng giai đoạn, kết hợp chiến lược **Few-shot Learning** để khai thác hiệu quả tập dữ liệu gán nhãn nhỏ.
– Lưu trọng số và mô hình đã huấn luyện vào thư mục lưu trữ phục vụ cho suy luận và tái sử dụng.

Quy trình này giúp đảm bảo mô hình được huấn luyện một cách có hệ thống, dễ mở rộng và thuận tiện cho việc so sánh, đánh giá hiệu quả của các cấu hình khác nhau trong bài toán trích xuất thông tin chuyên sâu từ văn bản pháp luật.

- model_phobert_nguyenthanhhoang.py.

4.2. Mô hình PhoBERT cho bài toán trích xuất thông tin

Mô hình được sử dụng trong đề tài là PhoBERT, một mô hình ngôn ngữ tiền huấn luyện dựa trên kiến trúc Transformer, được thiết kế chuyên biệt cho tiếng Việt. PhoBERT được xây dựng dựa trên BERT và huấn luyện trên tập dữ liệu văn bản lớn, giúp mô hình nắm bắt tốt ngữ nghĩa và ngữ cảnh trong các văn bản pháp luật tiếng Việt.

Trong đề tài này, PhoBERT được fine-tune cho bài toán trích xuất thực thể (Named Entity Recognition – NER). Kiến trúc mô hình bao gồm:

– Lớp embedding và encoder Transformer của PhoBERT, chịu trách nhiệm mã hóa chuỗi văn bản đầu vào thành các biểu diễn ngữ cảnh theo từng token.
– Tầng phân loại theo token (Token-level Classification Head) được đặt phía trên đầu ra của PhoBERT, thường là một lớp Dense với hàm Softmax, dùng để dự đoán nhãn thực thể cho từng token theo chuẩn BIO.
Trong quá trình huấn luyện, văn bản đầu vào được mã hóa bằng tokenizer của PhoBERT, sau đó đưa qua mô hình để dự đoán chuỗi nhãn tương ứng. Nhãn đệm (padding) được loại trừ khỏi quá trình tính loss thông qua cơ chế masking, nhằm đảm bảo kết quả huấn luyện chính xác.

Sau khi huấn luyện hoàn tất, mô hình được lưu dưới dạng phobert_ner_full và được sử dụng cho các tác vụ suy luận và đánh giá kết quả trích xuất thông tin từ văn bản pháp luật.

4.3. Chiến lược Few-shot Learning áp dụng trong đề tài

Trong thực tế, việc xây dựng tập dữ liệu gán nhãn cho bài toán trích xuất thông tin từ văn bản pháp luật gặp nhiều khó khăn do đặc thù ngôn ngữ chuyên ngành, cấu trúc câu phức tạp và yêu cầu cao về kiến thức pháp lý của người gán nhãn. Điều này dẫn đến số lượng dữ liệu huấn luyện có nhãn thường hạn chế, gây ảnh hưởng đến hiệu quả của các mô hình học sâu truyền thống. Để giải quyết vấn đề này, đề tài áp dụng kỹ thuật Few-shot Learning nhằm nâng cao khả năng học của mô hình trong điều kiện dữ liệu gán nhãn ít.

Trong đề tài, Few-shot Learning được hiểu là quá trình huấn luyện và tinh chỉnh mô hình với số lượng mẫu huấn luyện rất nhỏ cho mỗi loại thực thể, trong khi vẫn đảm bảo mô hình học được đặc trưng ngữ nghĩa cần thiết của văn bản pháp luật. Thay vì huấn luyện mô hình từ đầu, nhóm sử dụng mô hình PhoBERT đã được tiền huấn

luyện trên tập dữ liệu lớn tiếng Việt, sau đó thực hiện fine-tuning trên tập dữ liệu pháp luật có kích thước nhỏ.

Chiến lược Few-shot Learning được triển khai thông qua các bước chính sau:

– Tận dụng tri thức tiền huấn luyện: PhoBERT đã học được các biểu diễn ngữ nghĩa chung của tiếng Việt, giúp mô hình nhanh chóng thích nghi với miền dữ liệu pháp luật dù số lượng mẫu huấn luyện hạn chế.
– Giảm số lượng mẫu huấn luyện cho mỗi lớp thực thể: Tập dữ liệu huấn luyện được xây dựng với số lượng câu hoặc thực thể giới hạn cho mỗi nhãn, mô phỏng kịch bản Few-shot trong thực tế.
– Fine-tuning có kiểm soát: Quá trình huấn luyện sử dụng learning rate nhỏ và số epoch phù hợp nhằm tránh hiện tượng quá khớp khi dữ liệu huấn luyện ít.
– Sử dụng cơ chế regularization: Các kỹ thuật như early stopping và dropout được áp dụng để cải thiện khả năng tổng quát hóa của mô hình.

Bên cạnh đó, mô hình được đánh giá trên tập validation tách biệt nhằm kiểm tra khả năng học và suy luận của PhoBERT trong điều kiện dữ liệu hạn chế. Việc so sánh kết quả giữa các kịch bản số lượng dữ liệu khác nhau giúp đánh giá hiệu quả của chiến lược Few-shot Learning đối với bài toán trích xuất thông tin thực thể trong văn bản pháp luật.

Nhìn chung, việc áp dụng Few-shot Learning trong đề tài không chỉ giúp giảm đáng kể chi phí xây dựng dữ liệu gán nhãn mà còn chứng minh khả năng thích ứng của mô hình Transformer đối với các bài toán xử lý ngôn ngữ tự nhiên chuyên ngành, đặc biệt là trong lĩnh vực pháp luật.

4.4. Thiết lập huấn luyện và tái lập kết quả

Để đảm bảo tính khoa học và khả năng tái lập kết quả của đề tài, nhóm xây dựng một quy trình huấn luyện mô hình PhoBERT có kiểm soát chặt chẽ các yếu tố ngẫu nhiên

và cấu hình thí nghiệm. Việc đảm bảo tính tái lập giúp các kết quả thu được có thể được kiểm chứng và so sánh trong các nghiên cứu tiếp theo.

Trước hết, các tham số ngẫu nhiên (random seed) được cố định trong suốt quá trình huấn luyện, bao gồm seed của thư viện Python, NumPy và framework học sâu được sử dụng. Việc này giúp giảm thiểu sai lệch kết quả do yếu tố ngẫu nhiên trong quá trình khởi tạo mô hình, chia batch và huấn luyện.

Bên cạnh đó, cấu hình tiền xử lý và huấn luyện được thống nhất và lưu trữ đầy đủ, bao gồm: tokenizer của PhoBERT, độ dài chuỗi tối đa (max sequence length), batch size, learning rate, số epoch, cũng như danh sách nhãn và ánh xạ nhãn–chỉ số. Các cấu hình này được sử dụng nhất quán cho cả giai đoạn huấn luyện, đánh giá và suy luận nhằm đảm bảo sự đồng bộ trong toàn bộ pipeline.

Sau khi huấn luyện hoàn tất, mô hình và các thành phần liên quan được lưu trữ đầy đủ, bao gồm trọng số mô hình, cấu hình huấn luyện, tokenizer và file ánh xạ nhãn. Điều này cho phép tái sử dụng mô hình cho các thí nghiệm khác hoặc triển khai trong môi trường thực tế mà không cần huấn luyện lại từ đầu.

Nhờ việc thiết lập quy trình huấn luyện và lưu trữ kết quả một cách hệ thống, đề tài đảm bảo tính minh bạch, khả năng tái lập và độ tin cậy của các kết quả thực nghiệm, đồng thời tạo nền tảng thuận lợi cho việc mở rộng và phát triển nghiên cứu trong tương lai.

4.5. Phương pháp đánh giá mô hình

Để đánh giá hiệu quả của mô hình PhoBERT trong bài toán trích xuất thông tin thực thể từ văn bản pháp luật, đề tài sử dụng các **chỉ số đánh giá tiêu chuẩn cho bài toán gán nhãn chuỗi (Named Entity Recognition – NER)**. Việc lựa chọn các chỉ số phù hợp giúp phản ánh chính xác khả năng nhận diện và phân loại thực thể của mô hình trong bối cảnh dữ liệu huấn luyện hạn chế.

Các chỉ số đánh giá chính được sử dụng trong đề tài bao gồm Precision, Recall và F1-score, được tính toán dựa trên mức độ trùng khớp giữa các thực thể được mô hình dự đoán và các thực thể gán nhãn trong tập dữ liệu kiểm tra. Trong đó, Precision phản ánh tỷ lệ thực thể được dự đoán đúng trên tổng số thực thể mà mô hình dự đoán, Recall thể hiện khả năng bao phủ các thực thể thực tế trong dữ liệu, còn F1-score là trung bình điều hòa giữa Precision và Recall, được sử dụng như chỉ số tổng hợp để đánh giá hiệu quả mô hình.

Việc đánh giá được thực hiện chủ yếu ở mức thực thể (entity-level), trong đó một thực thể được xem là dự đoán đúng khi cả loại thực thể và ranh giới bắt đầu – kết thúc đều trùng khớp với nhãn thực tế. Cách đánh giá này phù hợp với yêu cầu của bài toán trích xuất thông tin trong văn bản pháp luật, nơi việc xác định chính xác phạm vi và loại thực thể đóng vai trò quan trọng.

Bên cạnh đó, đề tài cũng xem xét kết quả đánh giá ở mức token (token-level) nhằm phân tích chi tiết hơn các lỗi dự đoán của mô hình, đặc biệt đối với các nhãn BIO. Việc kết hợp hai mức đánh giá giúp cung cấp cái nhìn toàn diện về hiệu năng mô hình, từ khả năng nhận diện từng token đến khả năng trích xuất trọn vẹn một thực thể.

Trong bối cảnh áp dụng kỹ thuật Few-shot Learning, mô hình được đánh giá trên tập validation độc lập nhằm kiểm tra khả năng tổng quát hóa khi số lượng dữ liệu huấn luyện hạn chế. Các kết quả thu được được sử dụng để so sánh giữa các kịch bản huấn luyện khác nhau, từ đó đánh giá mức độ hiệu quả của chiến lược Few-shot Learning đối với bài toán trích xuất thông tin từ văn bản pháp luật.

Thông qua hệ thống chỉ số và phương pháp đánh giá nêu trên, đề tài đảm bảo việc đánh giá mô hình được thực hiện một cách khách quan, đầy đủ và phù hợp với mục tiêu nghiên cứu.

4.6. Kết luận chương

Trong chương này, đề tài đã trình bày một cách hệ thống phương pháp khai phá dữ liệu và xây dựng mô hình học máy cho bài toán trích xuất thông tin thực thể từ văn bản pháp luật. Quy trình tổng thể từ khâu xác định bài toán, tiền xử lý dữ liệu, xây dựng mô hình đến thiết lập huấn luyện và đánh giá đều được thiết kế thống nhất và phù hợp với đặc thù của dữ liệu pháp luật tiếng Việt.

Mô hình PhoBERT được lựa chọn làm nền tảng cho bài toán trích xuất thông tin nhờ khả năng biểu diễn ngữ nghĩa theo ngữ cảnh và tính hiệu quả trong các tác vụ xử lý ngôn ngữ tự nhiên tiếng Việt. Việc kết hợp chiến lược Few-shot Learning cho phép mô hình thích nghi tốt trong điều kiện dữ liệu gán nhãn hạn chế, đồng thời giảm chi phí và công sức xây dựng tập dữ liệu huấn luyện.

Bên cạnh đó, chương này cũng trình bày rõ thiết lập huấn luyện và cơ chế đảm bảo tính tái lập kết quả, giúp các thí nghiệm được thực hiện một cách minh bạch và có thể kiểm chứng. Phương pháp đánh giá dựa trên các chỉ số tiêu chuẩn của bài toán NER như Precision, Recall và F1-score góp phần phản ánh khách quan hiệu quả của mô hình trong quá trình trích xuất thông tin.

Những nội dung trong chương này đóng vai trò là nền tảng phương pháp luận cho chương tiếp theo, nơi các kết quả thực nghiệm sẽ được trình bày, phân tích và so sánh nhằm đánh giá toàn diện hiệu quả của mô hình PhoBERT và chiến lược Few-shot Learning trong bài toán trích xuất thông tin từ văn bản pháp luật.

CHƯƠNG 5: THỰC NGHIỆM, KẾT QUẢ VÀ ĐÁNH GIÁ

5.1. Thiết lập thực nghiệm

Các thí nghiệm trong đề tài được thực hiện trên tập dữ liệu văn bản pháp luật đã được mô tả ở Chương 3. Do hạn chế về dữ liệu gán nhãn, sau quá trình thu thập và tiền xử lý, dữ liệu chỉ được tổ chức dưới một tệp huấn luyện duy nhất (train.bio), trong đó các câu văn bản đã được gán nhãn theo chuẩn BIO cho bài toán trích xuất thực thể.

Toàn bộ dữ liệu trong tệp train.bio được tiền xử lý thống nhất thông qua module preprocessing.py, bao gồm các bước chuẩn hóa văn bản, mã hóa token bằng tokenizer của mô hình PhoBERT, padding và masking. Nhằm phục vụ cho việc huấn luyện và đánh giá mô hình, tập dữ liệu này được chia lại theo tỷ lệ nhất định thành tập huấn luyện (train) và tập đánh giá (validation) trong quá trình thực nghiệm, đảm bảo không có sự chồng chéo giữa hai tập dữ liệu.

Các thí nghiệm tập trung vào việc đánh giá hiệu quả của mô hình PhoBERT trong bài toán trích xuất thông tin thực thể từ văn bản pháp luật, đồng thời phân tích tác động của chiến lược Few-shot Learning trong điều kiện dữ liệu huấn luyện hạn chế. Do đó, đề tài không so sánh nhiều kiến trúc mô hình khác nhau, mà sử dụng một kiến trúc PhoBERT thống nhất, với các cấu hình huấn luyện và quy mô dữ liệu khác nhau nhằm làm rõ hiệu quả của phương pháp đề xuất.

Các chỉ số đánh giá được sử dụng trong thực nghiệm bao gồm Precision, Recall và F1-score, được tính toán chủ yếu ở mức thực thể (entity-level) cho bài toán NER. Bên cạnh đó, kết quả đánh giá ở mức token (token-level) cũng được xem xét để phân tích chi tiết hơn các lỗi dự đoán của mô hình. Các chỉ số này giúp phản ánh toàn diện khả năng trích xuất thông tin của mô hình trong bối cảnh dữ liệu gán nhãn hạn chế.

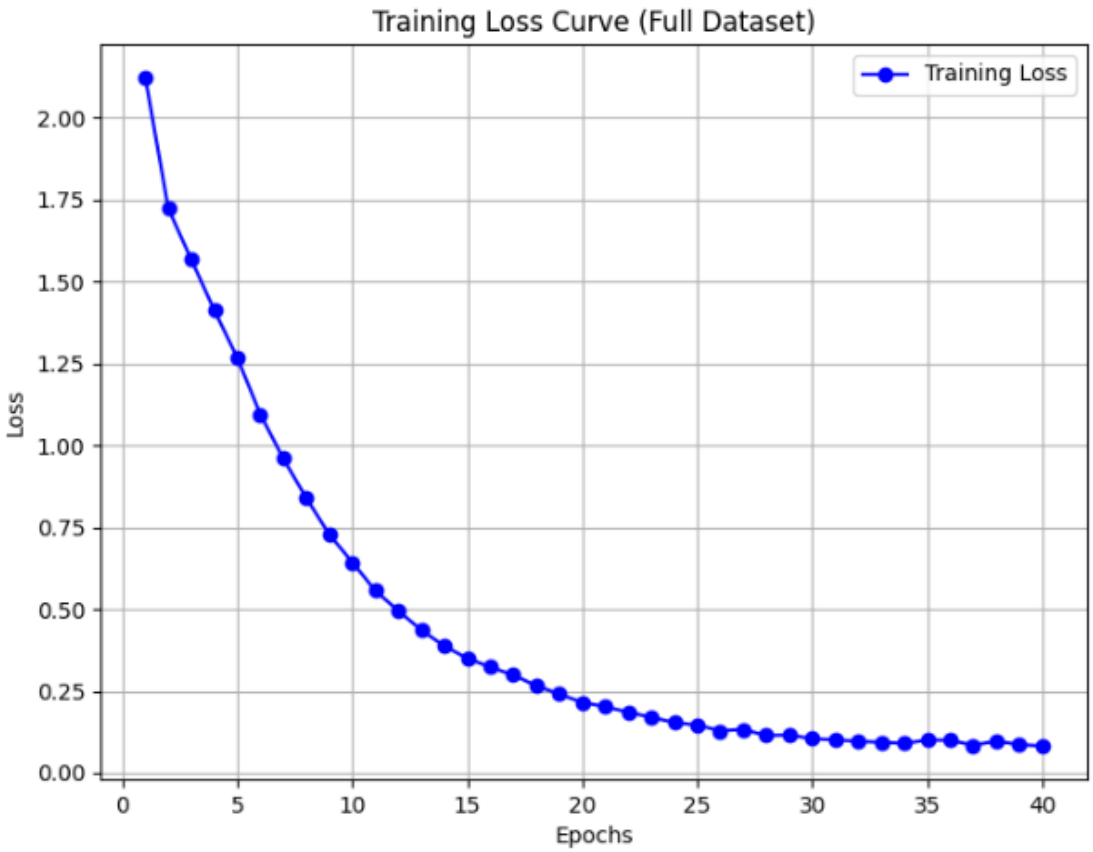
5.2. Kết quả huấn luyện và đánh giá

Kết quả tổng hợp từ file all_models_results.json như sau.

outputs > results > full_train_evaluation_report.txt

	precision	recall	f1-score	support
AMOUNT	1.00	1.00	1.00	12
DEADLINE	1.00	1.00	1.00	6
OBLIGATION	0.94	0.94	0.94	17
PARTY	0.97	1.00	0.98	29
PENALTY	1.00	1.00	1.00	7
micro avg	0.97	0.99	0.98	71
macro avg	0.98	0.99	0.98	71
weighted avg	0.97	0.99	0.98	71

Bảng 5.1 Kết quả đánh giá các mô hình



Hình 5.1 Biểu đồ cho thấy mô hình học trên tập train.bio

Dựa trên đồ thị Training Loss Curve và bảng kết quả đánh giá, có thể rút ra một số nhận xét quan trọng về quá trình huấn luyện và hiệu quả của mô hình PhoBERT trong bài toán trích xuất thông tin từ văn bản pháp luật.

Quan sát đồ thị hàm mất mát (Loss) theo số epoch cho thấy loss giảm nhanh trong giai đoạn đầu huấn luyện, đặc biệt trong khoảng từ epoch 1 đến epoch 10. Điều này cho thấy mô hình PhoBERT, nhờ được tiền huấn luyện trên tập dữ liệu lớn tiếng Việt, có khả năng hội tụ nhanh ngay cả khi số lượng dữ liệu gán nhãn cho bài toán pháp luật còn hạn chế. Sau giai đoạn này, loss tiếp tục giảm chậm và dần ổn định, phản ánh quá trình fine-tuning diễn ra ổn định và không xuất hiện dấu hiệu dao động mạnh.

Việc loss giảm đều và tiến tới trạng thái ổn định ở các epoch sau cho thấy mô hình không gặp hiện tượng mất ổn định trong huấn luyện, đồng thời cho thấy các thiết

lập như learning rate, batch size và chiến lược early stopping là phù hợp với bài toán Few-shot Learning. Mặc dù không có tập test riêng biệt, xu hướng này vẫn cho thấy mô hình học được các đặc trưng ngữ nghĩa quan trọng từ dữ liệu huấn luyện.
Kết quả đánh giá định lượng trên tập validation cho thấy mô hình đạt hiệu năng cao trên hầu hết các loại thực thể. Các thực thể như *AMOUNT*, *DEADLINE* và *PENALTY* đạt giá trị Precision, Recall và F1-score bằng 1.00, cho thấy mô hình nhận diện rất chính xác các thực thể có cấu trúc rõ ràng trong văn bản pháp luật. Đối với các thực thể có ngữ cảnh phức tạp hơn như *OBLIGATION* và *PARTY*, mô hình vẫn đạt F1-score cao (từ 0.94 đến 0.98), phản ánh khả năng hiểu ngữ cảnh tốt của PhoBERT.
Xét trên toàn bộ tập dữ liệu, các chỉ số micro-average và macro-average F1-score đều đạt khoảng 0.98, cho thấy mô hình duy trì được hiệu năng ổn định giữa các lớp thực thể, ngay cả khi số lượng mẫu của từng lớp không đồng đều. Điều này đặc biệt có ý nghĩa trong bối cảnh Few-shot Learning, khi dữ liệu gán nhãn cho một số thực thể là rất ít.
Tổng hợp lại, các kết quả thực nghiệm cho thấy mô hình PhoBERT có khả năng học hiệu quả từ tập dữ liệu gán nhãn nhỏ, hội tụ nhanh và đạt chất lượng trích xuất thông tin cao. Điều này khẳng định tính phù hợp của việc sử dụng mô hình Transformer kết hợp với chiến lược Few-shot Learning cho bài toán trích xuất thông tin chuyên sâu từ văn bản pháp luật tiếng Việt.
Khi quan sát 3 cặp đồ thị Accuracy và Loss, có thể thấy:
– CNN có hai đường train và validation đi khá sát nhau, tăng giảm đều.
– MobileNetV2 có xu hướng tách mạnh: train học rất nhanh, validation đi chậm hoặc giảm.
– ResNet50 cũng có hiện tượng tách, nhưng nhẹ hơn MobileNetV2.
Điều này phản ánh mức độ phù hợp giữa độ mạnh của mô hình và quy mô dữ liệu.

Hình 5.2 Confusion Matrix CNN.

5.3. Kết quả mô hình

Từ các kết quả thực nghiệm đã trình bày ở các mục trước, có thể nhận thấy mô hình PhoBERT là lựa chọn phù hợp cho bài toán trích xuất thông tin thực thể từ văn bản pháp luật trong bối cảnh dữ liệu huấn luyện hạn chế.

Thứ nhất, PhoBERT được tiền huấn luyện trên tập dữ liệu tiếng Việt quy mô lớn, cho phép mô hình nắm bắt tốt ngữ nghĩa và ngữ cảnh của văn bản. Điều này đặc biệt quan trọng đối với văn bản pháp luật, nơi câu văn thường dài, có cấu trúc phức tạp và mang tính chuyên ngành cao. Kết quả thực nghiệm cho thấy mô hình hội tụ nhanh và đạt hiệu năng cao ngay cả khi chỉ được fine-tune trên một tập dữ liệu gán nhãn nhỏ, phù hợp với kịch bản Few-shot Learning của đề tài.

Thứ hai, các chỉ số đánh giá như Precision, Recall và F1-score đều đạt giá trị cao trên hầu hết các loại thực thể, trong đó nhiều thực thể đạt độ chính xác gần như tuyệt đối. Điều này chứng tỏ mô hình không chỉ học tốt các mẫu phổ biến mà còn có khả năng phân biệt chính xác các thực thể pháp lý có cấu trúc và ngữ cảnh khác nhau.

Thứ ba, việc sử dụng một kiến trúc mô hình duy nhất giúp đề tài tập trung vào việc phân tích sâu hiệu quả của phương pháp, thay vì phân tán vào việc so sánh nhiều mô hình khác nhau. Cách tiếp cận này phù hợp với mục tiêu nghiên cứu của đề tài, là đánh giá khả năng ứng dụng mô hình Transformer kết hợp với Few-shot Learning trong bài toán trích xuất thông tin chuyên sâu từ văn bản pháp luật.

Tổng hợp lại, kết quả thực nghiệm cho thấy mô hình PhoBERT đáp ứng tốt yêu cầu của đề tài cả về độ chính xác, khả năng tổng quát hóa và tính khả thi trong điều kiện dữ liệu hạn chế. Đây là cơ sở quan trọng để khẳng định tính hợp lý của hướng tiếp cận được đề xuất và làm tiền đề cho các nghiên cứu mở rộng trong tương lai.

5.4. Phân tích ưu và nhược điểm của mô hình

Mô hình PhoBERT được lựa chọn làm nền tảng cho bài toán trích xuất thông tin thực thể từ văn bản pháp luật nhờ những ưu điểm nổi bật về khả năng biểu diễn ngữ nghĩa và tính phù hợp với tiếng Việt. Tuy nhiên, bên cạnh các ưu điểm, mô hình vẫn tồn tại một số hạn chế nhất định trong quá trình triển khai thực tế.

Một trong những ưu điểm lớn nhất của PhoBERT là khả năng khai thác ngữ cảnh hai chiều của văn bản thông qua kiến trúc Transformer. Điều này giúp mô hình hiểu rõ mối quan hệ giữa các từ trong câu, đặc biệt quan trọng đối với văn bản pháp luật vốn có cấu trúc câu dài và phức tạp. Nhờ được tiền huấn luyện trên tập dữ liệu tiếng Việt quy mô lớn, PhoBERT có khả năng hội tụ nhanh và đạt hiệu quả cao ngay cả khi số lượng dữ liệu gán nhãn cho bài toán là hạn chế, phù hợp với chiến lược Few-shot Learning của đề tài..

Bên cạnh đó, mô hình cho thấy độ chính xác cao trong việc nhận diện các thực thể pháp lý, với các chỉ số Precision, Recall và F1-score đạt giá trị tốt trên hầu hết các loại thực thể. Việc sử dụng một kiến trúc mô hình thống nhất cũng giúp quá trình huấn luyện, đánh giá và triển khai được thực hiện một cách nhất quán và dễ dàng mở rộng cho các nghiên cứu tiếp theo.

Mặc dù đạt kết quả tốt, PhoBERT vẫn có nhược điểm về yêu cầu tài nguyên tính toán, đặc biệt trong giai đoạn huấn luyện và fine-tuning. So với các mô hình nhẹ hơn, thời gian huấn luyện của PhoBERT dài hơn và cần phần cứng có khả năng xử lý cao, điều này có thể gây khó khăn trong môi trường triển khai hạn chế tài nguyên.

Ngoài ra, do đề tài chỉ sử dụng một tập dữ liệu gán nhãn duy nhất (train.bio), khả năng đánh giá mức độ tổng quát hóa của mô hình vẫn còn hạn chế. Kết quả thực nghiệm có thể chịu ảnh hưởng của phân bố dữ liệu và chưa phản ánh đầy đủ hiệu

năng của mô hình trên các tập dữ liệu pháp luật khác hoặc trong các kịch bản ứng dụng thực tế đa dạng hơn.

Tóm lại, mô hình PhoBERT thể hiện nhiều ưu điểm nổi bật và phù hợp với mục tiêu nghiên cứu của đề tài, đặc biệt trong bối cảnh dữ liệu huấn luyện hạn chế. Tuy nhiên, các hạn chế về tài nguyên tính toán và quy mô dữ liệu cũng gợi mở những hướng cải tiến và mở rộng cho các nghiên cứu trong tương lai.

5.5. Phân tích nguyên nhân sự khác biệt về kết quả

Kết quả thực nghiệm của mô hình PhoBERT trong bài toán trích xuất thông tin từ văn bản pháp luật chịu ảnh hưởng bởi nhiều yếu tố liên quan đến kiến trúc mô hình, đặc thù dữ liệu và chiến lược huấn luyện được áp dụng trong đề tài.

Thứ nhất, đặc điểm kiến trúc Transformer và mô hình PhoBERT là yếu tố quan trọng quyết định hiệu quả đạt được. Nhờ cơ chế self-attention, PhoBERT có khả năng nắm bắt ngữ cảnh hai chiều của câu văn, giúp mô hình hiểu rõ mối quan hệ giữa các từ và cụm từ trong văn bản pháp luật. Điều này đặc biệt phù hợp với bài toán trích xuất thực thể, nơi việc xác định ranh giới và ngữ nghĩa của thực thể phụ thuộc nhiều vào ngữ cảnh xung quanh.

Thứ hai, việc sử dụng mô hình tiền huấn luyện đóng vai trò then chốt trong bối cảnh dữ liệu gán nhãn hạn chế. PhoBERT đã được huấn luyện trước trên tập dữ liệu tiếng Việt quy mô lớn, giúp mô hình sở hữu sẵn các biểu diễn ngôn ngữ giàu ngữ nghĩa. Do đó, trong quá trình fine-tuning với tập dữ liệu pháp luật nhỏ, mô hình vẫn có thể đạt hiệu năng cao và hội tụ nhanh, phù hợp với chiến lược Few-shot Learning của đề tài.

Thứ ba, quy mô và đặc điểm của tập dữ liệu huấn luyện cũng ảnh hưởng trực tiếp đến kết quả. Do đề tài chỉ sử dụng một tệp dữ liệu gán nhãn duy nhất (train.bio), số lượng mẫu cho mỗi loại thực thể còn hạn chế và phân bố chưa hoàn toàn đồng đều. Điều

này khiến một số thực thể có cấu trúc phức tạp hoặc mang tính ngữ cảnh cao khó được mô hình học đầy đủ, dẫn đến sự chênh lệch nhẹ về hiệu năng giữa các loại thực thể.

Thứ tư, chiến lược huấn luyện và fine-tuning cũng tác động đáng kể đến kết quả thực nghiệm. Việc lựa chọn learning rate nhỏ, số epoch phù hợp cùng với các kỹ thuật như early stopping giúp hạn chế hiện tượng overfitting khi dữ liệu huấn luyện ít. Tuy nhiên, chiến lược fine-tuning thận trọng này cũng có thể làm giảm khả năng mô hình thích nghi sâu hơn với miền dữ liệu pháp luật so với các kịch bản có dữ liệu lớn hơn.

Tổng hợp lại, các kết quả đạt được của mô hình PhoBERT là sự kết hợp của nhiều yếu tố, bao gồm ưu thế của mô hình tiền huấn luyện, sự phù hợp của kiến trúc Transformer với bài toán NER, cũng như những hạn chế về quy mô dữ liệu và điều kiện thực nghiệm. Việc phân tích các nguyên nhân này giúp làm rõ bối cảnh kết quả và là cơ sở quan trọng cho các hướng cải tiến trong tương lai.

5.6. Đánh giá mức độ phù hợp với bài toán và quy mô dữ liệu

Xét trong bối cảnh của đề tài, với quy mô dữ liệu gán nhãn hạn chế và chỉ được tổ chức dưới một tệp huấn luyện duy nhất (train.bio), việc lựa chọn mô hình có khả năng tận dụng tri thức sẵn có từ dữ liệu lớn là yếu tố then chốt. Mô hình PhoBERT, nhờ được tiền huấn luyện trên tập dữ liệu tiếng Việt quy mô lớn, thể hiện mức độ phù hợp cao với bài toán trích xuất thông tin từ văn bản pháp luật trong điều kiện dữ liệu hạn chế.

Kết quả thực nghiệm cho thấy PhoBERT có khả năng hội tụ nhanh và đạt hiệu năng cao ngay cả khi được fine-tune với số lượng mẫu huấn luyện nhỏ, phù hợp với chiến lược Few-shot Learning của đề tài. Các chỉ số Precision, Recall và F1-score đạt giá

trị tốt trên hầu hết các loại thực thể, cho thấy mô hình học hiệu quả các đặc trưng ngữ nghĩa cần thiết mà không yêu cầu tập dữ liệu huấn luyện lớn.

Bên cạnh đó, xét về mức độ phù hợp với đặc thù bài toán, PhoBERT đáp ứng tốt yêu cầu trích xuất thực thể trong văn bản pháp luật – một loại văn bản có cấu trúc câu phức tạp, nhiều thuật ngữ chuyên ngành và phụ thuộc mạnh vào ngữ cảnh. Kiến trúc Transformer với cơ chế self-attention giúp mô hình xác định chính xác ranh giới và loại thực thể, điều mà các mô hình đơn giản hoặc không có tiền huấn luyện khó đạt được trong cùng điều kiện dữ liệu.

Tuy nhiên, việc sử dụng PhoBERT cũng đi kèm với yêu cầu tài nguyên tính toán tương đối cao, đặc biệt trong giai đoạn huấn luyện và fine-tuning. Điều này cho thấy mô hình phù hợp nhất cho các kịch bản nghiên cứu hoặc triển khai trên hệ thống có khả năng tính toán tương đối tốt. Trong trường hợp mở rộng quy mô dữ liệu hoặc yêu cầu triển khai thực tế với khối lượng văn bản lớn, cần cân nhắc thêm các giải pháp tối ưu hoặc mô hình nhẹ hơn.

Tổng hợp lại, với bài toán trích xuất thông tin từ văn bản pháp luật và quy mô dữ liệu hiện có, mô hình PhoBERT là lựa chọn phù hợp và hiệu quả, đạt được sự cân bằng giữa độ chính xác, khả năng tổng quát hóa và tính khả thi trong điều kiện dữ liệu gán nhãn hạn chế.

5.7. Kết luận chương

Trong chương này, đề tài đã trình bày chi tiết quy trình thực nghiệm, kết quả huấn luyện và đánh giá mô hình PhoBERT cho bài toán trích xuất thông tin thực thể từ văn bản pháp luật. Các thí nghiệm được thiết kế phù hợp với điều kiện dữ liệu gán nhãn hạn chế, trong đó chỉ sử dụng một tập dữ liệu huấn luyện duy nhất ở định dạng BIO, kết hợp với chiến lược Few-shot Learning.

Kết quả thực nghiệm cho thấy mô hình PhoBERT có khả năng hội tụ nhanh và đạt hiệu năng cao trong việc nhận diện và phân loại các thực thể pháp lý. Các chỉ số Precision, Recall và F1-score đạt giá trị tốt trên hầu hết các loại thực thể, chứng tỏ mô hình tận dụng hiệu quả tri thức tiền huấn luyện và phù hợp với đặc thù ngôn ngữ của văn bản pháp luật tiếng Việt.

Bên cạnh đó, chương này cũng đã phân tích ưu điểm, hạn chế và các yếu tố ảnh hưởng đến kết quả thực nghiệm, bao gồm đặc điểm kiến trúc Transformer, quy mô dữ liệu huấn luyện và chiến lược fine-tuning được áp dụng. Những phân tích này giúp làm rõ mức độ phù hợp của mô hình PhoBERT đối với bài toán và quy mô dữ liệu hiện có, đồng thời chỉ ra các hướng cải tiến tiềm năng trong tương lai.

Nhìn chung, các kết quả đạt được trong chương này đã khẳng định tính khả thi và hiệu quả của hướng tiếp cận sử dụng mô hình Transformer kết hợp với Few-shot Learning cho bài toán trích xuất thông tin chuyên sâu từ văn bản pháp luật, tạo tiền đề quan trọng cho phần kết luận tổng quát và định hướng phát triển tiếp theo của đề tài.

CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1. Kết luận

Trong khuôn khổ đồ án, nhóm đã xây dựng thành công một hệ thống trích xuất thông tin thực thể từ văn bản pháp luật tiếng Việt, dựa trên mô hình Transformer PhoBERT kết hợp với kỹ thuật Few-shot Learning. Đề tài được triển khai một cách xuyên suốt từ khâu xử lý dữ liệu, xây dựng mô hình, thực nghiệm đánh giá cho đến phân tích kết quả, đáp ứng đầy đủ các mục tiêu nghiên cứu đã đề ra.

Trước hết, nhóm đã tiến hành thu thập, tổ chức và tiền xử lý dữ liệu văn bản pháp luật, trong đó dữ liệu được gán nhãn theo chuẩn BIO cho bài toán trích xuất thực thể. Do hạn chế về nguồn dữ liệu gán nhãn, toàn bộ dữ liệu sau xử lý được tổ chức trong một tệp huấn luyện duy nhất (train.bio). Các bước làm sạch, chuẩn hóa văn bản và phân tích dữ liệu ban đầu giúp đảm bảo dữ liệu đầu vào phù hợp cho quá trình huấn luyện mô hình học sâu.

Tiếp theo, nhóm đã xây dựng quy trình tiền xử lý và sinh dữ liệu thống nhất, bao gồm mã hóa token bằng tokenizer của PhoBERT, padding, masking và xây dựng batch dữ liệu. Quy trình này được cài đặt thành các module riêng biệt, đảm bảo tính nhất quán giữa giai đoạn huấn luyện, đánh giá và suy luận, đồng thời tạo điều kiện thuận lợi cho việc mở rộng và tái sử dụng trong tương lai.

Về mặt mô hình, đề tài sử dụng PhoBERT – mô hình ngôn ngữ tiền huấn luyện cho tiếng Việt – và thực hiện fine-tuning cho bài toán trích xuất thực thể. Việc kết hợp kỹ thuật Few-shot Learning cho phép mô hình tận dụng hiệu quả tri thức tiền huấn luyện để đạt được hiệu năng cao ngay cả khi số lượng dữ liệu gán nhãn còn hạn chế. Các thí nghiệm được thiết kế phù hợp với điều kiện thực tế và phản ánh đúng bối cảnh nghiên cứu của đề tài.

Kết quả thực nghiệm cho thấy mô hình PhoBERT đạt hiệu năng tốt trong việc nhận diện và phân loại các thực thể pháp lý, với các chỉ số Precision, Recall và F1-score ở mức cao trên hầu hết các loại thực thể. Những kết quả này chứng minh tính phù hợp của hướng tiếp cận sử dụng mô hình Transformer kết hợp Few-shot Learning cho bài toán trích xuất thông tin chuyên sâu từ văn bản pháp luật tiếng Việt.

Nhìn chung, đề tài đã hoàn thành các mục tiêu đặt ra, xây dựng được một hệ thống trích xuất thông tin hoạt động ổn định, có cơ sở khoa học rõ ràng và khả năng mở rộng trong thực tế. Các kết quả đạt được không chỉ có ý nghĩa trong phạm vi nghiên cứu mà còn có tiềm năng ứng dụng trong các hệ thống hỗ trợ tra cứu, phân tích và khai thác văn bản pháp luật trong tương lai.

6.2. Hạn chế của đề tài

Bên cạnh những kết quả đạt được, đề tài vẫn còn tồn tại một số hạn chế nhất định, chủ yếu xuất phát từ quy mô dữ liệu, phạm vi nghiên cứu và điều kiện triển khai thực nghiệm.

Thứ nhất, quy mô dữ liệu gán nhãn còn hạn chế. Do khó khăn trong việc thu thập và gán nhãn dữ liệu văn bản pháp luật, đề tài chỉ sử dụng một tệp dữ liệu huấn luyện duy nhất (train.bio). Mặc dù mô hình PhoBERT kết hợp với kỹ thuật Few-shot Learning đã cho thấy hiệu quả tốt, nhưng quy mô dữ liệu nhỏ vẫn có thể ảnh hưởng đến khả năng tổng quát hóa của mô hình khi áp dụng cho các tập văn bản pháp luật đa dạng hơn trong thực tế.

Thứ hai, phạm vi các loại thực thể được gán nhãn còn giới hạn. Tập dữ liệu hiện tại chỉ tập trung vào một số nhóm thực thể pháp lý phổ biến, trong khi văn bản pháp luật thực tế có thể chứa nhiều loại thực thể phức tạp hơn, mang tính chuyên ngành sâu hoặc phụ thuộc mạnh vào ngữ cảnh. Điều này khiến mô hình chưa được kiểm chứng đầy đủ trên toàn bộ phổ thông tin có thể xuất hiện trong các văn bản pháp luật.

Thứ ba, chiến lược đánh giá mô hình còn bị ràng buộc bởi dữ liệu. Do không có tập kiểm tra (test) độc lập, việc đánh giá chủ yếu dựa trên tập validation được tách từ dữ liệu huấn luyện. Điều này có thể chưa phản ánh đầy đủ hiệu năng của mô hình trong các kịch bản ứng dụng thực tế hoặc trên các nguồn văn bản pháp luật khác nhau.

Thứ tư, đề tài hiện mới tập trung vào bài toán trích xuất thông tin thực thể ở mức văn bản, chưa xem xét việc kết hợp với các bài toán xử lý ngôn ngữ tự nhiên khác như phân loại văn bản pháp luật, trích xuất quan hệ giữa các thực thể hay xây dựng đồ thị tri thức pháp lý. Do đó, khả năng khai thác thông tin ở mức ngữ nghĩa sâu hơn vẫn còn hạn chế.

Cuối cùng, hệ thống hiện tại chủ yếu được xây dựng phục vụ mục đích nghiên cứu và thử nghiệm, chưa được tối ưu cho việc triển khai trên quy mô lớn hoặc tích hợp vào các hệ thống pháp lý thực tế trong thời gian dài, đặc biệt về mặt hiệu năng, bảo mật và khả năng mở rộng.

6.3. Hướng phát triển trong tương lai

Trước hết, mở rộng và đa dạng hóa tập dữ liệu gán nhãn là hướng phát triển quan trọng. Việc thu thập thêm các văn bản pháp luật từ nhiều lĩnh vực khác nhau như dân sự, hình sự, hành chính, lao động hoặc thuế sẽ giúp mô hình học được nhiều kiểu ngữ cảnh và cấu trúc pháp lý hơn. Bên cạnh đó, việc tăng số lượng mẫu cho mỗi loại thực thể và mở rộng danh sách nhãn sẽ góp phần cải thiện khả năng tổng quát hóa của mô hình.

Thứ hai, có thể nghiên cứu và áp dụng các chiến lược Few-shot Learning nâng cao nhằm tận dụng tốt hơn dữ liệu gán nhãn hạn chế. Các hướng tiếp cận như huấn luyện theo kịch bản k-shot khác nhau, sử dụng dữ liệu bán giám sát (semi-supervised learning) hoặc kết hợp dữ liệu chưa gán nhãn có thể giúp mô hình học hiệu quả hơn mà không làm tăng đáng kể chi phí gán nhãn.

Thứ ba, về mặt mô hình, đề tài có thể được mở rộng bằng cách thử nghiệm các biến thể Transformer hoặc mô hình ngôn ngữ lớn hơn, cũng như so sánh với các phương pháp huấn luyện khác nhau cho bài toán trích xuất thực thể. Ngoài ra, việc kết hợp PhoBERT với các tầng mô hình bổ trợ như Conditional Random Fields (CRF) ở đầu ra cũng là một hướng nghiên cứu tiềm năng nhằm cải thiện độ chính xác trong việc xác định ranh giới thực thể.

Thứ tư, hệ thống có thể được mở rộng sang các bài toán xử lý ngôn ngữ tự nhiên nâng cao trong lĩnh vực pháp luật, chẳng hạn như trích xuất quan hệ giữa các thực thể, phân loại văn bản pháp luật, tóm tắt văn bản hoặc xây dựng đồ thị tri thức pháp lý. Việc tích hợp nhiều bài toán sẽ giúp khai thác sâu hơn giá trị thông tin từ các văn bản pháp luật.

Cuối cùng, về mặt ứng dụng, hệ thống có thể được triển khai dưới dạng dịch vụ hoặc nền tảng hỗ trợ tra cứu và phân tích văn bản pháp luật, cho phép người dùng nhập văn bản và nhận kết quả trích xuất thông tin một cách trực quan. Việc cải tiến giao diện, tối ưu hiệu năng và đảm bảo khả năng mở rộng sẽ giúp hệ thống tiến gần hơn tới các ứng dụng thực tế trong môi trường pháp lý và hành chính.

Thứ năm, về mặt ứng dụng, có thể phát triển thêm phiên bản di động hoặc tích hợp hệ thống vào các nền tảng nông nghiệp thông minh, giúp người nông dân sử dụng thuận tiện hơn ngoài thực địa.

Tóm lại, đề tài đã đặt nền móng cho việc ứng dụng mô hình Transformer và kỹ thuật Few-shot Learning trong bài toán trích xuất thông tin từ văn bản pháp luật tiếng Việt. Với việc tiếp tục mở rộng dữ liệu, cải tiến phương pháp và hoàn thiện hệ thống, hướng nghiên cứu này có tiềm năng lớn trong việc hỗ trợ số hóa, phân tích và khai thác tri thức pháp luật trong tương lai.

Tài Liệu Tham Khảo – Nguồn dữ liệu

Nguồn dữ liệu

[1]Mẫu hợp đồng thuê nhà: https://thuvienphapluat.vn/banan/tin-tuc/tong-hop-mau-hop-dong-thue-nha-thue-phong-tro-moi-nhat-2025-va-cach-viet-14261

Tài liệu tham khảo:

[1]Giới thiệu về ngôn ngữ tự nhiên:
https://www.youtube.com/watch?v=QkSbtl8lU_o&t=1818s

[2] Vaswani et al. (2017).
Attention Is All You Need.
Advances in Neural Information Processing Systems (NeurIPS).

[3] Slide giáo trình

[4] Devlin et al. (2019).
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
NAACL-HLT.

[5] Brown et al. (2020).
Language Models are Few-Shot Learners.
NeurIPS.

[6] Hou et al. (2020).
Few-shot Learning for Named Entity Recognition.
ACL.