

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN ĐHQG TP.HCM

Năm học 2025-2026



# AUTOMATIC LABELLING ENGINE WITH LLM

Presentation

PHÙNG DŨNG QUÂN - 22280073  
NGUYỄN HỒ TUYÊN - 22280103

GIẢNG VIÊN HƯỚNG DẪN : ThS. ĐOÀN THỊ TRÂM

# Overview

- I.Bối cảnh và bài toán nghiên cứu
- II.Cách tiếp cận và kiến trúc hệ thống
- III.Tối ưu hóa tài nguyên
- IV.Dữ liệu, thực nghiệm và đánh giá
- V.Kết luận và hướng phát triển

# I.Bối cảnh & vấn đề nghiên cứu

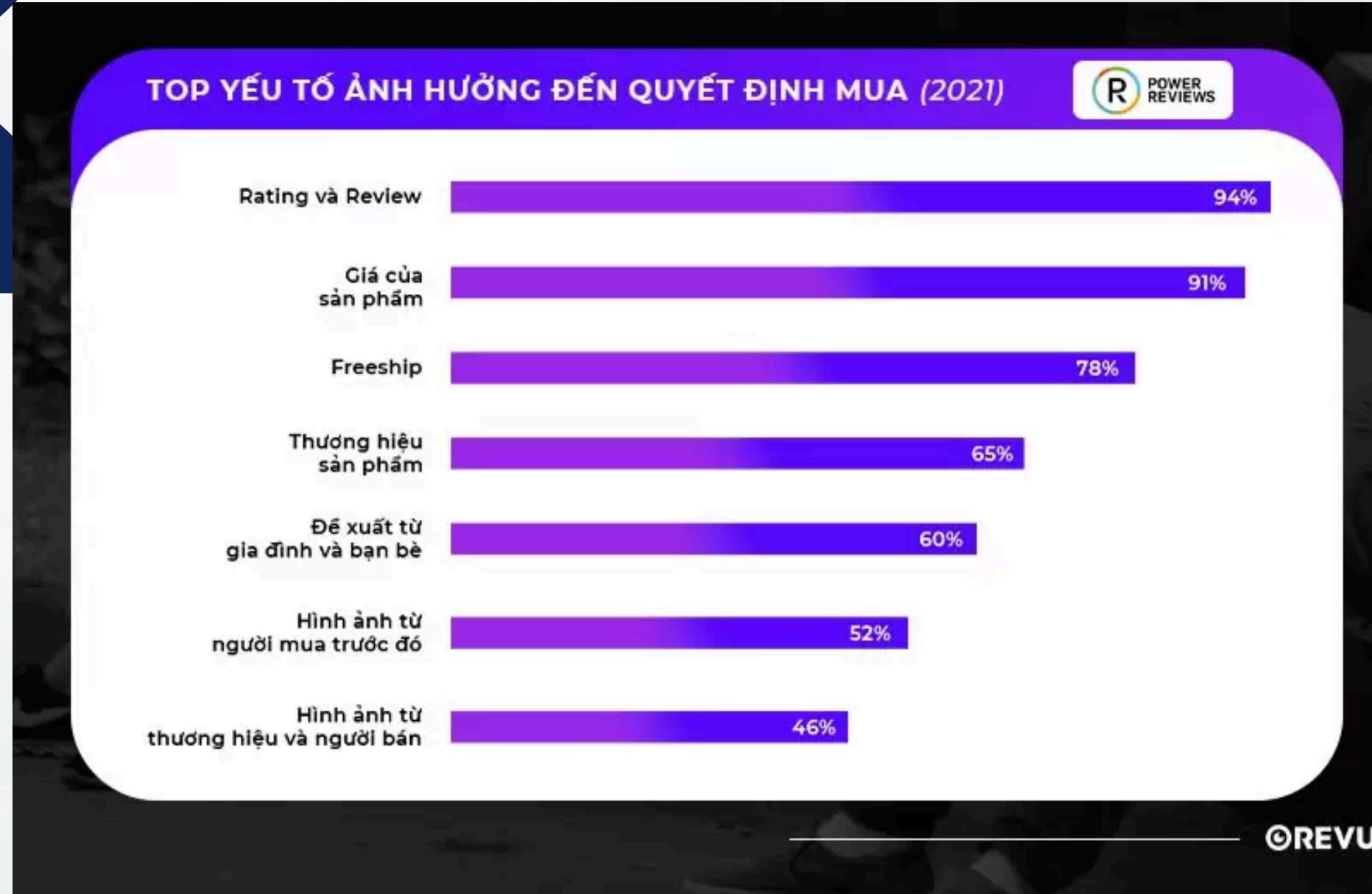
Thực trạng:

- Lượng review khách sạn trên các nền tảng trực tuyến tăng rất nhanh

Vấn đề hiện tại:

- Gán nhãn thủ công:
  - Tốn nhiều thời gian
  - Chi phí cao
  - Khó mở rộng với dữ liệu lớn
- Phân tích cảm xúc truyền thống:
  - Chỉ cho một nhãn tổng
  - Không phản ánh khách đang khen/chê điều gì





<https://blog.vn.revu.net/tac-dong-cua-so-luong-review-den-chuyen-doi/>

### Nhu cầu doanh nghiệp tăng cao:

- Phân tích theo từng khía cạnh dịch vụ
- Hiểu rõ điểm mạnh – điểm yếu cụ thể

### Từ đó đặt ra yêu cầu:

- Hệ thống gán nhãn tự động
- Chi tiết
- Hiệu quả trên dữ liệu lớn

Dựa vào thống kê của Power Reviews vào năm 2018, nhận xét:

- Review có tác động mạnh đến hành vi và quyết định mua
- Là nguồn dữ liệu quan trọng để doanh nghiệp cải thiện dịch vụ

=> Review hiện nay có tác động mạnh mẽ đến hành vi và quyết định mua hàng của người tiêu dùng, đồng thời là nguồn dữ liệu quan trọng mà doanh nghiệp có thể khai thác để cải thiện và phát triển dịch vụ.

# Bài toán nghiên cứu: Aspect-Based Sentiment Analysis (ABSA)

Thay vì chỉ gán nhãn tích cực hay tiêu cực cho cả review, ABSA cho phép phân tích chi tiết theo từng khía cạnh dịch vụ.

Mỗi review được phân tích thành:

- Đơn vị ngữ nghĩa:
  - Clause: mệnh đề mang ý nghĩa độc lập
- Nội dung được nhắc đến:
  - Aspect Term: đối tượng (room, staff, location, ...)
  - Opinion Phrase: ý kiến đánh giá (clean, expensive, friendly, ...)
- Ngữ cảnh đánh giá:
  - Aspect Category: nhóm khía cạnh (6 nhóm)
  - Polarity: cảm xúc (Positive, Negative, Neutral)

Mục tiêu: biến review tự do thành dữ liệu có cấu trúc, có thể phân tích và khai thác.



# Mục tiêu đề tài



Xây dựng pipeline gán nhãn tự động hoàn chỉnh

Lợi ích đạt được

- Giảm thời gian xử lý dữ liệu
- Giảm chi phí gán nhãn thủ công

Đảm bảo hệ thống:

- Độ chính xác cao
- Khả năng mở rộng
- Vận hành được trên tài nguyên phần cứng hạn chế
- Định hướng: hướng đến ứng dụng thực tế.

## II. Cách tiếp cận và kiến trúc hệ thống

### Data Collection

Mục tiêu: xây dựng tập dữ liệu review khách sạn đa nguồn, đủ lớn cho huấn luyện và đánh giá mô hình

Sử dụng Selenium và BeautifulSoup để thu thập review khách sạn từ các nền tảng:

- Traveloka
- Agoda
- Trip.com
- Booking.com
- Trivago



# Data Preprocessing

Nhằm đảm bảo dữ liệu sạch và đồng nhất trước khi đưa vào pipeline phân tích.

Các bước làm sạch dữ liệu trước khi cho vào phân tích:

- Loại bỏ ký tự đặc biệt và nhiễu
- Loại bỏ bình luận rác hoặc quá ngắn (không phải tiếng Anh)
- Loại bỏ các review trùng lặp



# Hybrid Pipeline Architecture

Thay vì dùng LLM cho toàn bộ pipeline, chúng tôi chọn kiến trúc lai.

## Kết hợp hai hướng tiếp cận:

Large Language Models (LLMs)

- Phi-3
- Nhiệm vụ:
  - Tách mệnh đề theo ngữ nghĩa
  - Trích xuất Aspect Term & Opinion

Transformer Encoder Models

- RoBERTa, DistilBERT
- Nhiệm vụ:
  - Phân loại Aspect Category
  - Phân loại Polarity

=>Tận dụng khả năng suy luận và hiểu ngữ nghĩa mạnh của LLM

=>Kết hợp tốc độ, độ ổn định và chi phí thấp của mô hình nhỏ



## Lý do chọn mô hình:

- **Large Language Models (LLMs):** Phi-3-mini-4k-Instruct:

Mô hình có quy mô và kiến trúc (LLM ~4B, decoder-only, instruction-tuned), tập trung tác vụ—giúp đánh giá ảnh hưởng của inductive bias lên bài toán trích xuất term và opinion.

- **Transformer Encoder Models:**

### 1. RoBERTa

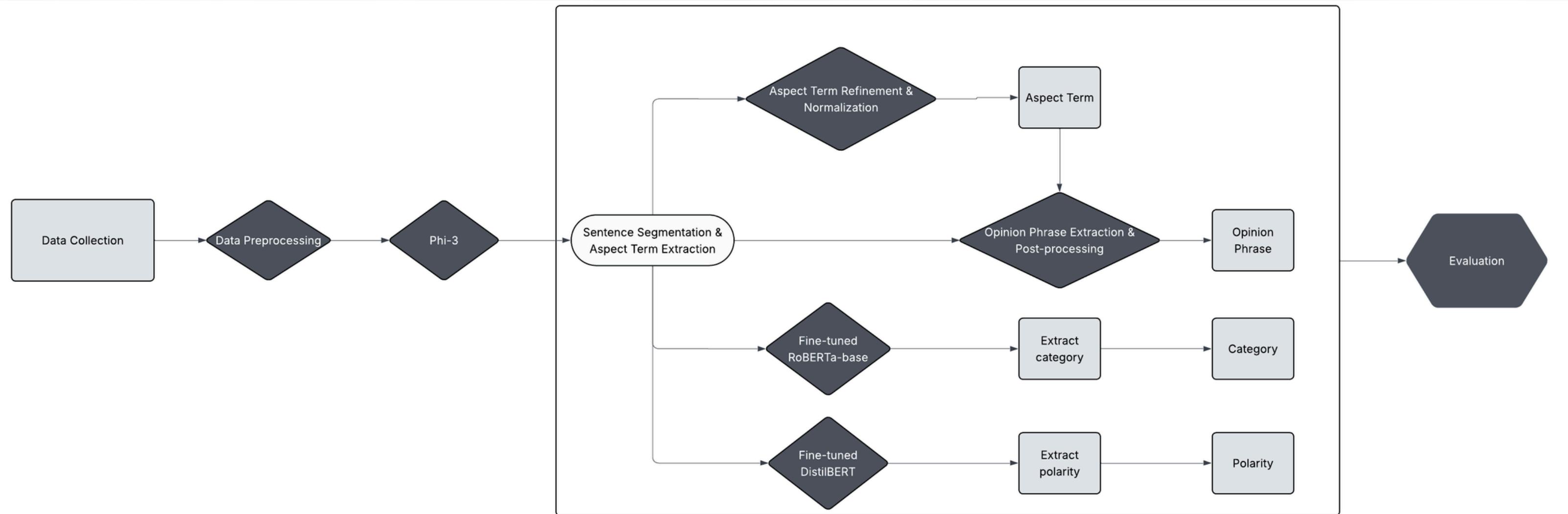
RoBERTa được lựa chọn vì mô hình cho độ chính xác cao hơn trong việc học biểu diễn ngữ nghĩa và ngữ cảnh sâu của văn bản, phù hợp với các bài toán NLP, yêu cầu độ tin cậy cao. Nhờ được huấn luyện trên tập dữ liệu lớn với Dynamic Masking và loại bỏ Next Sentence Prediction, RoBERTa cho khả năng tổng quát hóa tốt phù hợp với bài toán nâng cao hơn là trích xuất category.

### 2. Distilbert:

DistilBERT được lựa chọn do đáp ứng tốt yêu cầu cân bằng giữa hiệu năng, chi phí tính toán và khả năng triển khai thực tế. DistilBERT giữ lại phần lớn khả năng hiểu ngữ nghĩa của BERT thông qua kỹ thuật knowledge distillation, phù hợp với bài toán trích xuất polarity đơn giản cho clause



# Tổng quan kiến trúc hệ thống



Pipeline được thiết kế theo hướng chia nhỏ thành các module độc lập.

## Đặc điểm:

- Kiến trúc modular
- Dễ mở rộng
- Tối ưu tài nguyên



## Input

- Câu review gốc

## Output

- Các clause
- Aspect term tương ứng ứng trong từng clause

## Phương pháp

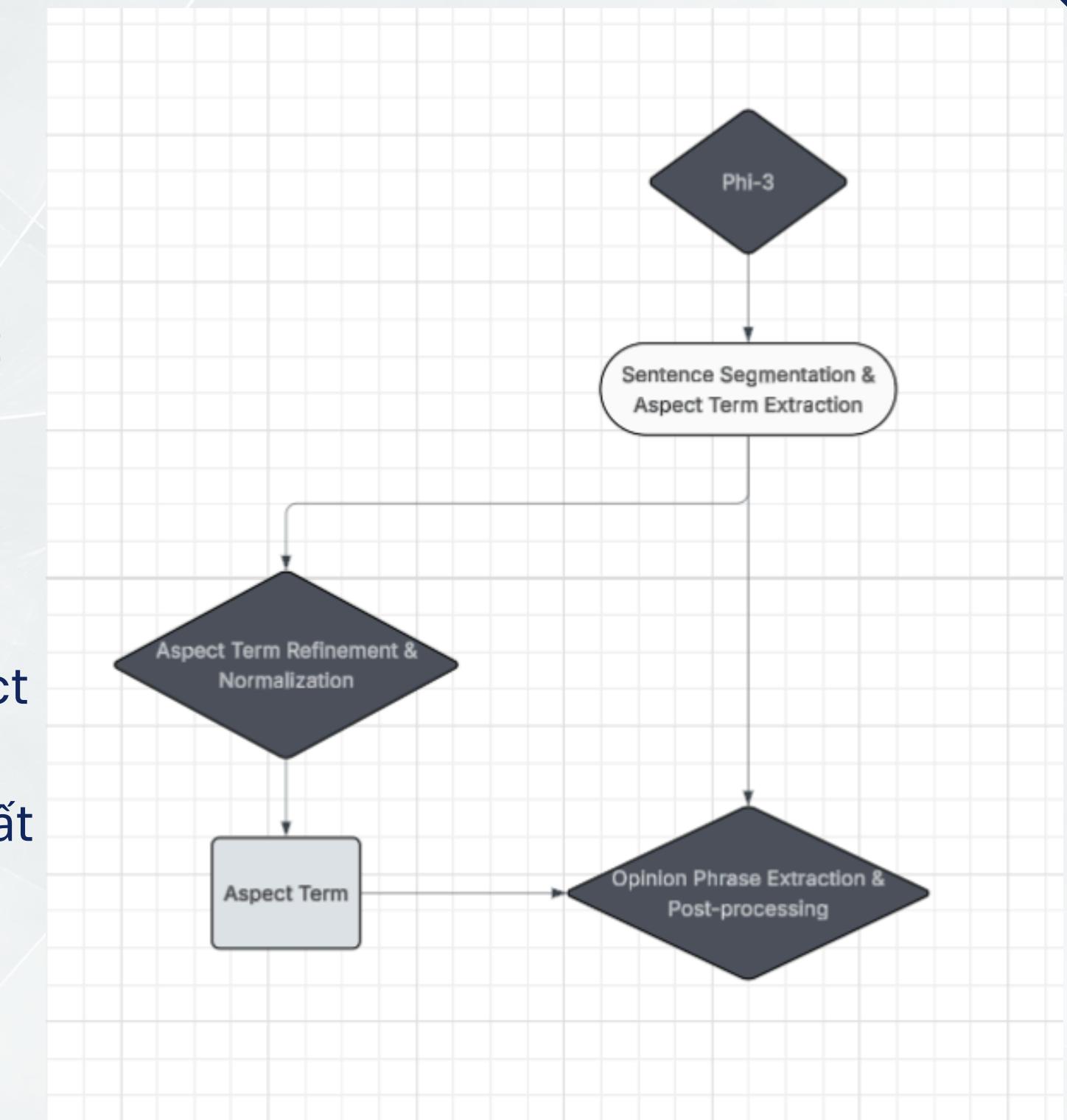
- Sử dụng LLM với chiến lược prompt few-shot để:
  - Tách văn bản thành các clause độc lập
  - Trích xuất aspect term trong mỗi clause

## Chuẩn hóa

- Áp dụng khối tinh chỉnh và chuẩn hóa aspect term
- Đưa các aspect về dạng biểu diễn thống nhất

## Lợi ích

- Giảm nhiễu trong dữ liệu
- Đảm bảo tính nhất quán của nhãn
- Tạo đầu vào sạch và ổn định cho các bước phân loại phía sau



## Input

- Clause + Aspect Term

## Output

- Clause + Aspect Term + Opinion Phrase

## Nhiệm vụ của khối này

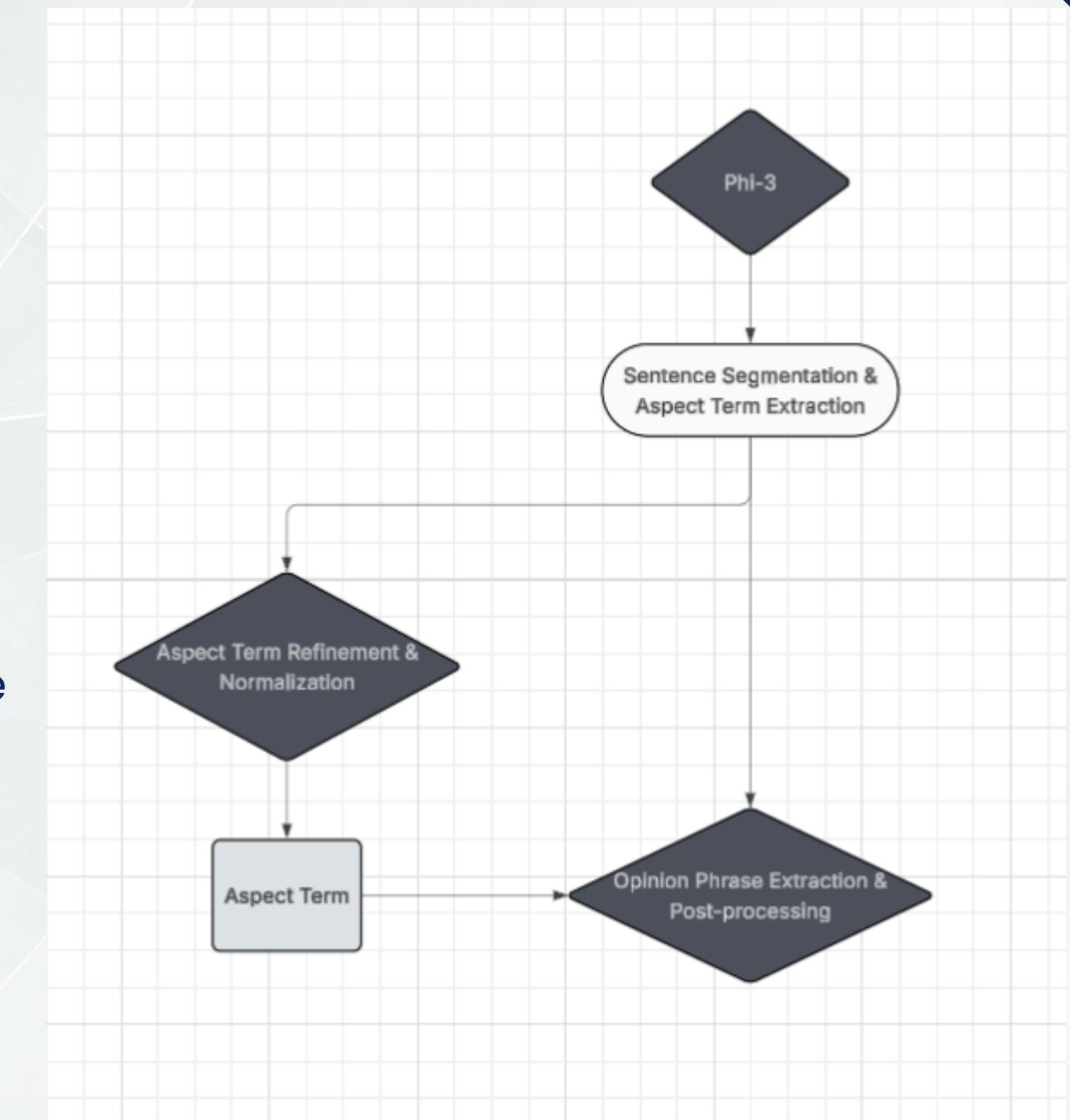
- Sử dụng LLM với prompt few-shot để trích xuất opinion phrase tương ứng với từng aspect trong clause

## Hậu xử lý

- Loại bỏ nhiễu
- Chuẩn hóa cách diễn đạt
- Ràng buộc rõ cặp Aspect Term – Opinion Phrase

## Mục tiêu

- Mỗi clause trở thành một đơn vị đánh giá hoàn chỉnh:
- “Aspect – Opinion – Ngữ cảnh”



# QA GUIDELINE

Khía cạnh	Mô tả
Cơ sở vật chất (Facility)	Bao gồm các yếu tố như thiết bị, nội thất phòng, trang trí khách sạn, thiết kế nội thất, ban công, và khu vực hồ bơi,...
Tiện ích (Amenity)	Bao gồm các dịch vụ công cộng như bãi đậu xe, spa, nhà hàng, quà lưu niệm, các điểm đến lân cận, các tuỳ chọn thanh toán, sự an ninh.
Dịch vụ (Service)	Liên quan đến chất lượng dịch vụ của khách sạn, chất lượng món ăn, thái độ của nhân viên, ...
Trải nghiệm (Experience)	Liên quan đến trải nghiệm của khách hàng về khách sạn như bầu không khí, cái nhìn tổng thể và mức độ thư giãn mà khách sạn mang lại.
Thương hiệu (Branding)	Phản ánh mức độ hài lòng chung của khách hàng so với thông tin được cung cấp.
Mức độ trung thành (Loyalty)	Cho biết khả năng khách quay lại và giới thiệu khách sạn cho người khác.

Bảng 2.4: Các nhóm phân cực cảm xúc

Nhóm cảm xúc	Mô tả
Tích cực (Positive)	Bao gồm các cảm xúc như hài lòng, vui vẻ, hạnh phúc.
Tiêu cực (Negative)	Bao gồm các cảm xúc như không hài lòng, thất vọng, phàn nàn, và khó chịu.
Trung tính (Neutral)	Bao gồm các cảm xúc không mạnh mẽ, không thuần túy tích cực hoặc tiêu cực, cảm xúc bình thường, không có cảm xúc đặc biệt.

## ANNOTATION AGREEMENT

- Cohen's Kappa: 0.88
- Weighted Kappa: 0.9

=> Độ đồng thuận cao giữa  
các annotator

RESEARCH

## Phân loại danh mục khía cạnh (Aspect Category Classification)

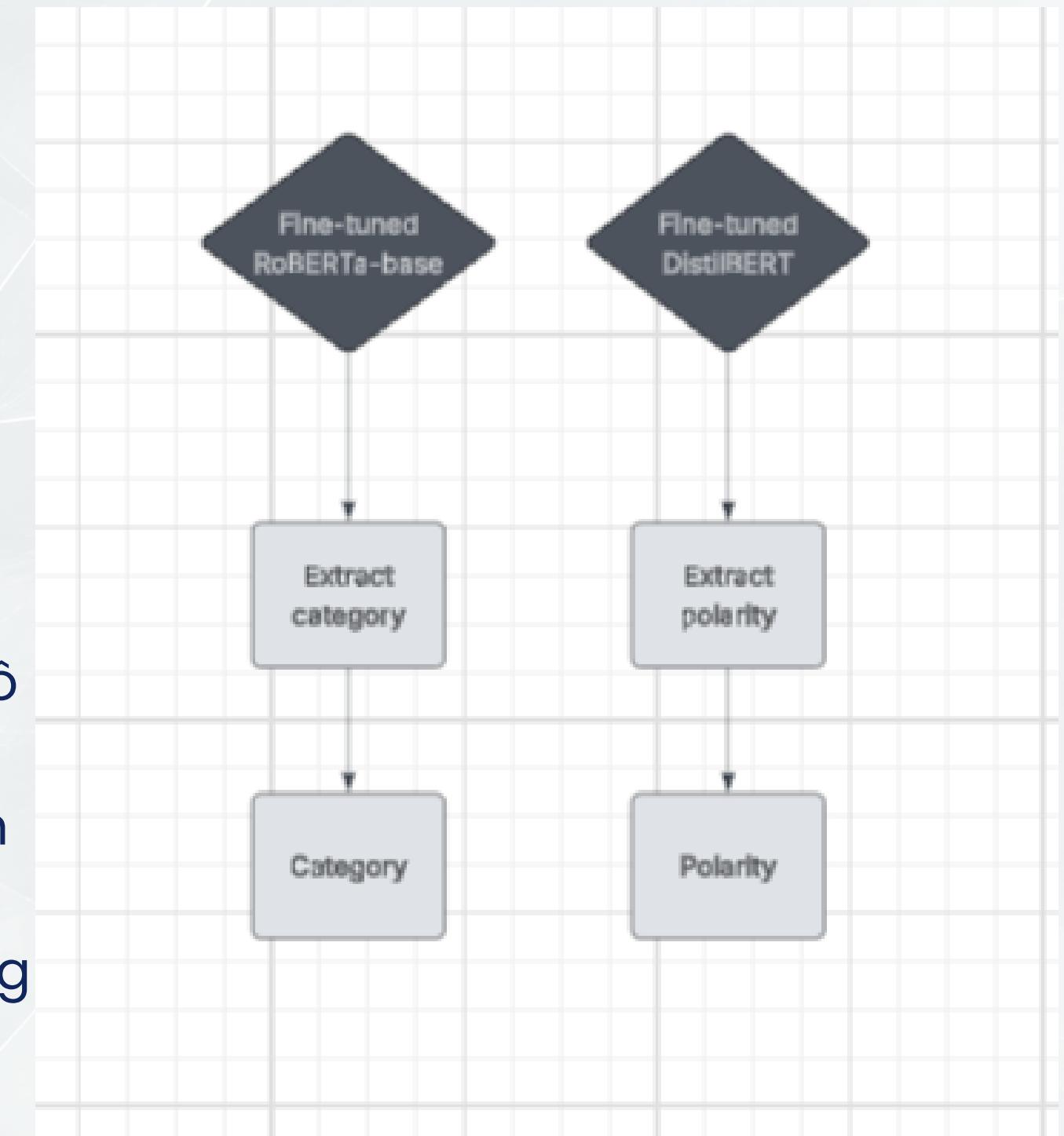
Hệ thống sử dụng mô hình ngôn ngữ tiền huấn luyện (Pre-trained Transformer Encoder), được fine-tune cho bài toán phân loại đa lớp ở mức Clause.

Trong quá trình huấn luyện:

- Mỗi Clause được gán một nhãn Aspect Category duy nhất
- Clause được mã hóa bằng tokenizer của mô hình
- Dữ liệu được cắt hoặc đệm về độ dài cố định để phù hợp với kiến trúc mô hình
- Hàm mất mát phân loại đa lớp được sử dụng để tối ưu mô hình (Cross Entropy Loss)

Cách tiếp cận này giúp:

- Đảm bảo tính nhất quán của nhãn

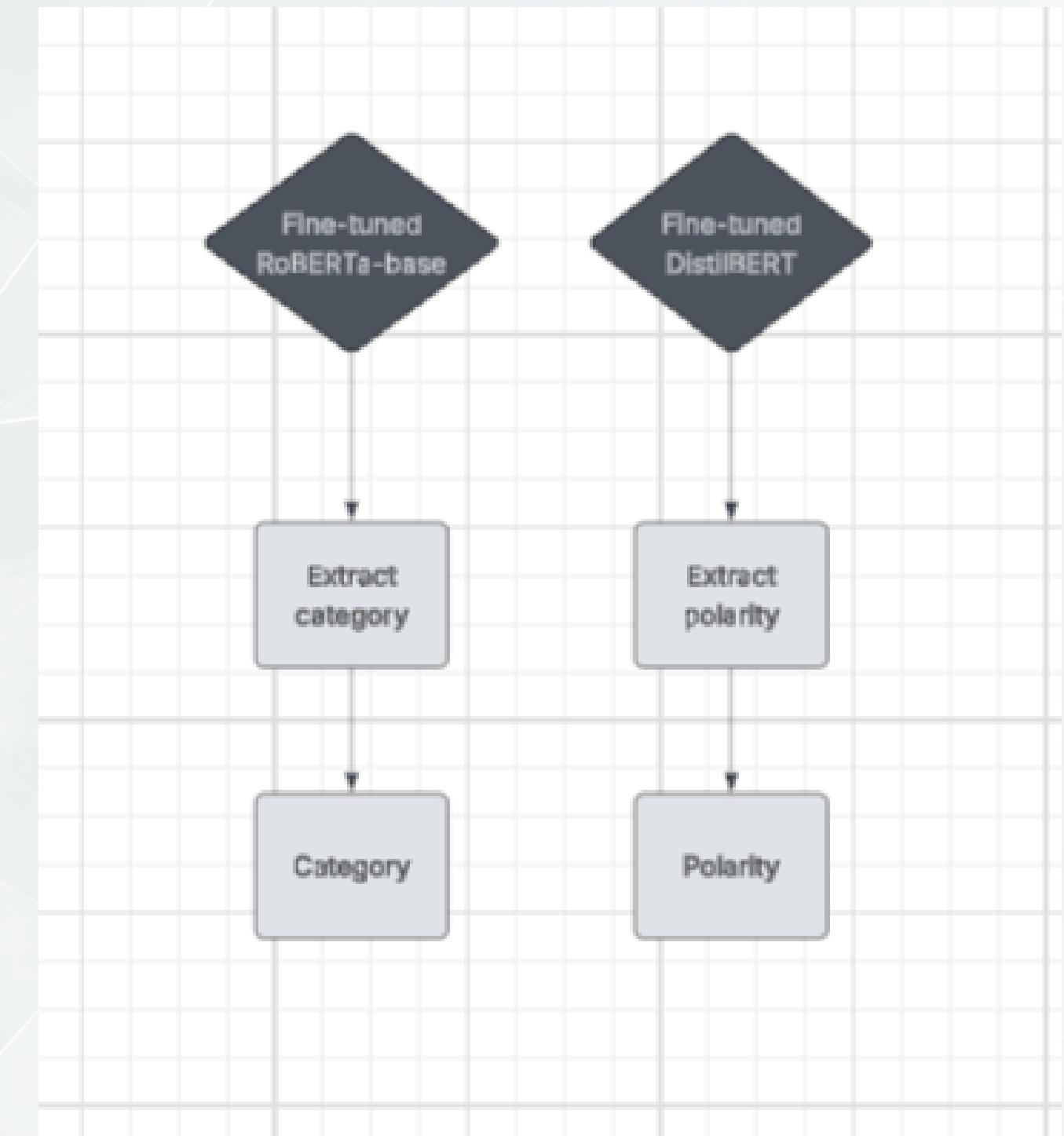


## Phân loại cảm xúc (Polarity Classification)

Hệ thống sử dụng mô hình Transformer Encoder được tinh chỉnh cho bài toán phân loại cảm xúc ở mức Clause.

Trong quá trình huấn luyện:

- Clause được mã hóa bằng tokenizer
- Hàm mất mát phân loại đa lớp được sử dụng (CE Loss)
- Mô hình học cách ánh xạ biểu đạt cảm xúc trong Clause sang nhãn Polarity tương ứng



Kết quả cuối cùng của hệ thống được tổng hợp thành một bảng dữ liệu có cấu trúc, trong đó mỗi bản ghi bao gồm các trường thông tin:

- Clause
- Aspect Term
- Opinion Phrase
- Aspect Category
- Polarity

Định dạng dữ liệu đầu ra (JSON hoặc DataFrame) cho phép hệ thống dễ dàng tích hợp với các bài toán downstream như:

- Phân tích trải nghiệm khách hàng
- Hệ thống gợi ý
- Dashboard phân tích kinh doanh

# III. Tối ưu hóa tài nguyên

## LARGE LANGUAGE MODEL

- Phi-3 Mini (FP16)
  - Quantization 4-bit (NF4)
- => Giảm VRAM từ 8GB->3,5GB

## TRANSFORMER ENCODER MODELS

- Fine-tune với LoRA
  - Giảm mạnh số tham số cần huấn luyện:
    - RoBERTa: ~0.35%
    - DistilBERT: ~1.09%
- => Phù hợp môi trường tài nguyên hạn chế



# IV. Dữ liệu, thực nghiệm và đánh giá

## DATASET DESCRIPTION

- Dữ liệu review khách sạn
- Được xử lý ở mức clause-level
- ~30,000 mẫu dữ liệu

Đặc điểm dữ liệu:

- Mất cân bằng nhãn Category
- Positive chiếm đa số

Chất lượng dữ liệu:

- Có quy trình Data Quality Assurance
- Mỗi mẫu được gán nhãn bởi 2 annotators
- Đánh giá độ tin cậy giữa các annotators



# DATA QUALITY & EVALUATION

## DATA QUALITY

- Làm sạch dữ liệu
- Chuẩn hóa văn bản
- Loại bỏ trùng lặp

## EVALUATION METRICS

### Không ròi rạc (term,opinion):

- Exact Match F1-score
- Token-level F1-score
- ROUGE-L
- Embedding Similarity

### Ròi rạc (category,polarity):

- Accuracy
- Precision
- Recall
- Macro-F1



# Kết quả trích xuất term và opinion

## Phi-3

Task	Exact Match F1	Token-level F1	ROUGE-L	Embedding Similarity
Term Extraction	0.92	0.99	0.97	0.98
Opinion Extraction	0.94	0.99	0.98	0.98

## Qwen-3

Task	ExactMatch F1	Token-level F1	ROUGE-L	Embedding Sim
Term Extraction	0.78	0.91	0.87	0.90
Opinion Extraction	0.87	0.90	0.90	0.94

- Phi-3 đạt Exact Match 92% và Token F1 99%, khắc phục hoàn toàn điểm yếu về độ chính xác của Qwen-3 (~78%).
- Chỉ số Embedding Similarity đạt 0.98 khẳng định khả năng nắm bắt ngữ cảnh tài chính phức tạp và chính xác.
- Mô hình duy trì phong độ cao đồng đều ở cả hai tác vụ trích xuất Term và Opinion, không bị lệch pha (bias).

# Kết quả đánh nhän Category

**RoBERTa-Base + LoRA**

Category	Precision	Recall	F1-score	Support
Amenity	0.87	0.89	0.88	1,312
Branding	0.51	0.38	0.43	220
Experience	0.76	0.72	0.74	1,319
Facility	0.90	0.91	0.91	2,683
Loyalty	0.91	0.93	0.92	331
Service	0.92	0.93	0.93	3,084
Overall Accuracy	-	-	<b>0.88</b>	8,949
Macro Avg	0.81	0.80	0.80	8,949

**DistilBERT + LoRA**

Category	Precision	Recall	F1-score	Support
Amenity	0.86	0.88	0.87	1312
Branding	0.56	0.32	0.40	220
Experience	0.73	0.74	0.73	1319
Facility	0.90	0.91	0.90	2683
Loyalty	0.90	0.91	0.90	331
Service	0.92	0.93	0.92	3084
Overall Accuracy	-	-	<b>0.87</b>	8949
Macro Avg	0.81	0.78	0.79	8949

- RoBERTa đạt Overall Accuracy 88% và Macro F1 80%, vượt qua DistilBERT (Accuracy 87%, Macro F1 79%).
- Quan trọng nhất, RoBERTa xử lý các lớp dữ liệu khó như Branding tốt hơn hẳn với Recall 38% (so với 32% của DistilBERT).
- Các lớp quan trọng như Service, Facility đều đạt F1 > 91%, đảm bảo độ tin cậy cao cho các tác vụ cốt lõi.
- Mức tăng về độ chính xác (đặc biệt là Recall cho lớp hiếm) xứng đáng để đánh đổi lấy chi phí tính toán cao hơn một chút so với bản Distil.

# Kết quả đánh nhận polarity

**RoBERTa-Base + LoRA**

Class	Precision	Recall	F1-score	Support
<b>Negative</b>	0.54	0.30	0.60	1552
<b>Neutral</b>	0.00	0.00	0.00	520
<b>Positive</b>	0.77	0.99	0.87	6877
<b>Overall Accuracy</b>	-	-	<b>0.77</b>	8949
<b>Macro Avg</b>	0.44	0.34	0.31	8949

**Harsha + LoRA**

Class	Precision	Recall	F1-score	Support
<b>Negative</b>	0.79	0.90	0.84	1552
<b>Neutral</b>	0.66	0.28	0.39	520
<b>Positive</b>	0.96	0.97	0.96	6877
<b>Overall Accuracy</b>	-	-	<b>0.92</b>	8949
<b>Macro Avg</b>	0.80	0.72	0.73	8949

**DistilBERT + LORA**

Polarity	Precision	Recall	F1-score	Support
<b>Positive</b>	0.97	0.97	0.97	6877
<b>Negative</b>	0.85	0.86	0.86	1552
<b>Neutral</b>	0.57	0.52	0.54	520
<b>Overall Accuracy</b>	-	-	<b>0.92</b>	8949
<b>Macro Avg</b>	0.79	0.78	0.79	8949

- DistilBERT đạt Neutral Recall 52%, vượt xa Harsha (28%) và hoàn toàn đánh bại RoBERTa (0% - model bị lỗi).
- Duy trì Overall Accuracy 92% ngang bằng Harsha nhưng có chỉ số Macro F1 cao hơn (0.79 vs 0.73), chứng tỏ model không bị bias vào lớp đa số.
- RoBERTa bị loại vì gặp hiện tượng "Model Collapse" (dự đoán toàn bộ là Positive), Harsha bị loại vì bỏ sót quá nhiều tin trung tính.
- DistilBERT vừa nhẹ vừa xử lý tốt nhất các mẫu khó (Neutral), là lựa chọn "ngon-bổ-rẻ" nhất để triển khai.

# WEB APP DEMO

**Link Code Github**

<https://github.com/DungQuanPhung/Automatic-Labelling-Engine>

# V.CONCLUSION & FUTURE WORK

## Conclusion

- Pipeline hiệu quả trong việc tự động gán nhãn dữ liệu ABSA
- Giảm đáng kể chi phí gán nhãn thủ công
- Dữ liệu đầu ra có cấu trúc, phù hợp cho các bài toán downstream

## Future Work

- Human-in-the-loop để cải thiện chất lượng nhãn
- Fine-tune LLM theo domain khách sạn
- Data Quality Improvement
- Model Optimization: sử dụng các mô hình kiến trúc sâu hơn





**THANK  
YOU**