Báo cáo kỹ thuật: Tối ưu hóa hiệu suất phân loại văn bản bằng phương pháp Ensemble và tinh chỉnh các mô hình ngôn ngữ lớn

Nhóm 3MoTB
Khoa Khoa học Máy tính, Đại học Công nghệ Thông tin, ĐHQG-HCM

Ngày 7 tháng 10 năm 2025

Tóm tắt nội dung

Trong bối cảnh các Mô hình Ngôn ngữ Lớn (LLMs) đang định hình lại lĩnh vực xử lý ngôn ngữ tự nhiên, báo cáo này giới thiệu một phương pháp luận đột phá, kết hợp nhiều giai đoạn chiến lược nhằm tối đa hóa độ chính xác. Phương pháp của chúng tôi khởi đầu bằng việc áp dụng kỹ thuật **Tinh chỉnh có Giám sát (Supervised Fine-Tuning - SFT)**[2] trên bộ ba mô hình tiên tiến thuộc họ Qwen[4] — **Qwen3-4B**, **Qwen2.5-7B**, và **Qwen2.5-7B-Instruct** — để chuyên biệt hóa sâu sắc từng mô hình cho bộ dữ liệu đặc thù của cuộc thi. Để khuếch đại sức mạnh dự đoán, chúng tôi tiếp tục xây dựng một mô hình **Ensemble**[3], tổng hợp trọng số xác suất từ ba mô hình đã tinh chỉnh nhằm tạo ra một kết quả tổng thể mạnh mẽ và ổn định hơn. Tuy nhiên, điểm sáng tạo cốt lõi của nghiên cứu nằm ở cơ chế **Hậu xử lý Thích ứng**: chúng tôi chủ động xác định các trường hợp dự đoán bất đồng giữa mô hình Qwen2.5-7B và mô hình Ensemble, sau đó triển khai một mô hình Qwen3-4B được đào tạo chuyên biệt để đóng vai trò như một **bộ phân giải chuyên gia (expert resolver)**, đưa ra quyết định cuối cùng nhằm hiệu chỉnh và hoàn thiện kết quả. Phương pháp tiếp cận phân cấp và đa tầng này đã chứng minh hiệu quả vượt trội, mang lại một bước nhảy vọt về độ chính xác và độ tin cậy của hệ thống, qua đó khẳng định tiềm năng của việc kết hợp chiến lược nhiều mô hình để giải quyết các bài toán phức tạp.

1 Giới thiệu

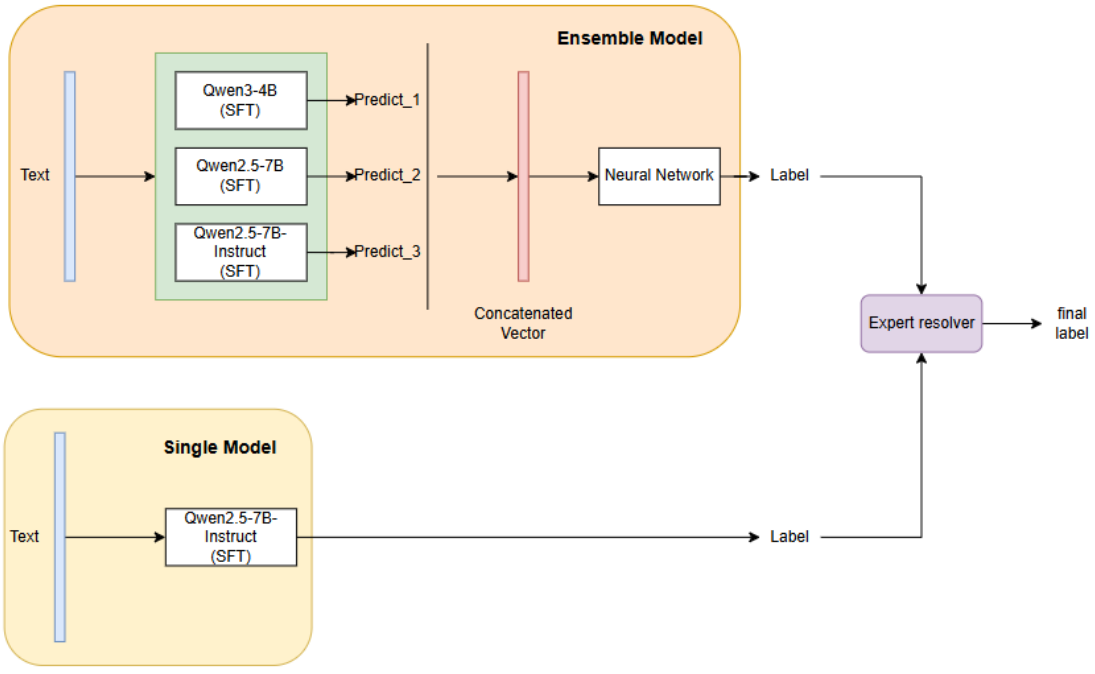
Trong thập kỷ vừa qua, sự phát triển vượt bậc của các *Mô hình Ngôn ngữ Lớn (Large Language Models - LLMs)* đã mở ra nhiều hướng tiếp cận mới trong lĩnh vực *Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing - NLP)*. Các mô hình như GPT, LLaMA, hay Qwen không chỉ đạt được những cột mốc quan trọng trong việc sinh ngôn ngữ, mà còn chứng minh khả năng thích ứng linh hoạt trong nhiều tác vụ phức tạp như dịch máy, tóm tắt văn bản, phân tích cảm xúc, và đặc biệt là phân loại văn bản.

Tuy nhiên, việc áp dụng trực tiếp các mô hình tiền huấn luyện (pre-trained models) vào các bài toán đặc thù thường chưa đạt hiệu suất tối ưu do sự chênh lệch giữa dữ liệu huấn luyện gốc và dữ liệu thực tế. Do đó, các kỹ thuật tinh chỉnh như *Supervised Fine-Tuning (SFT)* đã trở thành một giải pháp hiệu quả nhằm giúp mô hình thích ứng tốt hơn với miền dữ liệu cụ thể. Song song đó, các phương pháp kết hợp mô hình (*Ensemble Learning*) cũng được chứng minh là công cụ mạnh mẽ để gia tăng độ chính xác và tính ổn định của hệ thống, đặc biệt trong bối cảnh dự đoán trên dữ liệu đa dạng và phức tạp.

Báo cáo này tập trung vào việc xây dựng một hệ thống phân loại văn bản hiệu suất cao, dựa trên sự kết hợp chiến lược giữa **SFT**, **Ensemble Learning**, và một cơ chế **Hậu xử lý Thích ứng** mới. Cụ thể, chúng tôi tiến hành tinh chỉnh có giám sát trên ba biến thể của họ Qwen: Qwen3-4B, Qwen2.5-7B, và Qwen2.5-7B-Instruct. Sau đó, các mô hình này được tích hợp thông qua một khung Ensemble nhằm khai thác sức mạnh cộng hưởng. Đặc biệt, chúng tôi đề xuất cơ chế phân giải chuyên gia (expert resolver) sử dụng mô hình Qwen3-4B đã huấn luyện chuyên biệt, nhằm xử lý các trường hợp bất đồng trong dự đoán giữa mô hình Ensemble và các thành phần riêng lẻ.

Bằng việc áp dụng cách tiếp cận đa tầng này, nghiên cứu hướng tới việc nâng cao đáng kể độ chính xác và độ tin cậy trong phân loại văn bản, đồng thời mở ra triển vọng áp dụng rộng rãi cho các bài toán NLP phức tạp trong thực tiễn.

2 Phương pháp đề xuất



Hình 1: Sơ đồ tổng quan phương pháp đề xuất.

2.1 Giai đoạn 1: Supervised Fine-Tuning các mô hình LLM

Trong giai đoạn đầu, nhóm nghiên cứu lựa chọn ba mô hình thuộc họ Qwen để tiến hành tinh chỉnh. Sự lựa chọn này dựa trên hai cơ sở chính: thứ nhất, Qwen3 là một trong những kiến trúc mô hình ngôn ngữ tiên tiến nhất tại thời điểm thực hiện, hứa hẹn mang lại hiệu suất cao. Thứ hai, họ mô hình Qwen được biết đến với khả năng hỗ trợ và xử lý tiếng Việt hiệu quả, một yếu tố quan trọng phù hợp với dữ liệu của bài toán.
Các mô hình cụ thể được sử dụng bao gồm:

- **Qwen3-4B**
- **Qwen2.5-7B**
- **Qwen2.5-7B-Instruct**

Quá trình tinh chỉnh có giám sát (*Supervised Fine-Tuning - SFT*) được thực hiện với sự hỗ trợ của thư viện **Unsloth**[1], nhằm tối ưu hóa việc sử dụng bộ nhớ và tăng tốc độ huấn luyện. Để phù hợp với bài toán phân loại ba nhãn, lớp `llm_head` của mỗi mô hình được điều chỉnh thành ba đầu ra tương ứng.

2.1.1 Cấu hình huấn luyện

Để đảm bảo hiệu quả huấn luyện, chúng tôi áp dụng phương pháp **LoRA (Low-Rank Adaptation)**[5] với bộ siêu tham số được thiết kế cẩn trọng. Các mô-đun được chọn để huấn luyện bao gồm: `lm_head`, `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, và `down_proj`. Trong đó, các mô-đun `lm_head` có kích thước nhỏ, giúp rút ngắn thời gian huấn luyện mà vẫn đảm bảo tính hiệu quả.

Các siêu tham số chính được sử dụng trong giai đoạn SFT bao gồm:

- **LoRA rank**: 16
- **Batch size**: 8
- **Gradient accumulation steps**: 4
- **Warmup steps**: 10
- **Learning rate**: 1×10^{-4}
- **Optimizer**: AdamW (8-bit)[6]
- **Weight decay**: 0.01
- **LR scheduler type**: Cosine with Restarts
- **Seed**: 0
- **Num train epochs**: 3

2.2 Giai đoạn 2: Xây dựng mô hình Ensemble bằng Neural Network

Đầu ra từ mỗi mô hình là một vector xác suất kích thước 3. Ba vector này được nối lại (concatenate) tạo thành vector đầu vào kích thước 9 cho mạng NN.

```python
import torch.nn as nn

class EnsembleNN(nn.Module):
    def __init__(self, input_dim: int, num_classes: int, dropout_p: float = 0.5):
        super().__init__()
        self.net = nn.Sequential(
            nn.Linear(input_dim, 32),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(32, 16),
            nn.ReLU(),
            nn.Dropout(dropout_p),
            nn.Linear(16, num_classes)
        )
    def forward(self, x):
        return self.net(x)
```

Listing 1: Mã nguồn mô hình EnsembleNN

Kiến trúc mạng Neural Network này được lấy cảm hứng từ công trình của Lai và cộng sự [3]. Mô hình Ensemble được huấn luyện bằng CrossEntropy Loss. Các lớp Dropout giúp giảm overfitting và tăng tính ổn định.

2.2.1 Cấu hình huấn luyện

Quá trình huấn luyện mô hình Ensemble được thiết lập với các siêu tham số sau:

- **Số epochs**: 50
- **Batch size**: 128
- **Learning rate**: 0.005

2.3 Giai đoạn 3: Hậu xử lý với các Bộ phân giải chuyên gia (Expert Resolvers)

2.3.1 Cơ sở toán học cho Giả thuyết Hậu xử lý

Giả thuyết của chúng tôi dựa trên cơ sở rằng cả hai mô hình, Ensemble (M_E) và Qwen2.5-7B (M_Q), đều là các bộ phân loại mạnh với độ chính xác cao. Gọi x là một mẫu dữ liệu đầu vào, y_{true} là nhãn thật, y_E và y_Q lần lượt là các nhãn dự đoán bởi M_E và M_Q.

Từ thực nghiệm, chúng ta quan sát thấy xác suất dự đoán đúng của mỗi mô hình là rất cao:

$$P(y_E = y_{true}) \geq \alpha \quad \text{và} \quad P(y_Q = y_{true}) \geq \alpha \tag{1}$$

trong đó α là một ngưỡng độ chính xác cao (ví dụ: $\alpha \approx 0.85$).

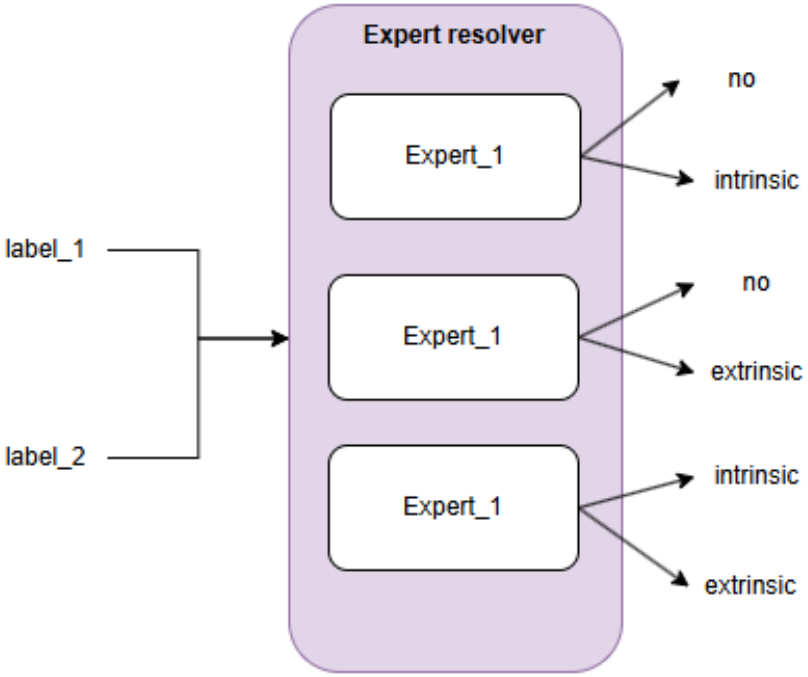
Giả thuyết chính của bước hậu xử lý có thể được phát biểu dưới dạng xác suất có điều kiện: *Xác suất để nhãn đúng nằm trong tập hợp các dự đoán của hai mô hình, với điều kiện hai mô hình đó bất đồng, là rất cao.* Công thức hóa, chúng tôi đặt giả thuyết rằng:

$$P(y_{true} \in \{y_E, y_Q\} \mid y_E \neq y_Q) \approx 1 \tag{2}$$

Điều này tương đương với việc nói rằng xác suất để cả hai mô hình cùng dự đoán sai trên cùng một mẫu dữ liệu (khi chúng đang bất đồng) là cực kỳ thấp:

$$P(y_{true} \notin \{y_E, y_Q\} \mid y_E \neq y_Q) \approx 0 \tag{3}$$

Sự hợp lý của giả thuyết (2) đến từ việc M_E và M_Q có kiến trúc và quá trình học khác nhau, dẫn đến các *điểm mù (failure modes)* của chúng có thể khác nhau. Do đó, trường hợp cả hai cùng sai trên một mẫu là rất hiếm. Dựa trên cơ sở này, việc xây dựng các bộ phân giải chuyên gia (expert resolvers) để phân xử giữa $\{y_E, y_Q\}$ là một chiến lược hợp lý và có cơ sở.



Hình 2: Bộ phân giải chuyên gia (Expert Resolvers)

Trong quá trình đánh giá, nhóm nghiên cứu nhận thấy hai mô hình có hiệu suất cao nhất là mô hình Ensemble và mô hình Qwen2.5-7B, với độ chênh lệch về độ chính xác (Accuracy) trên tập kiểm thử là rất nhỏ (≤ 0.002). Từ quan sát này, nhóm đưa ra giả

thuyết: đối với những mẫu dữ liệu mà hai mô hình này đưa ra dự đoán khác nhau, nhãn đúng có khả năng cao là một trong hai dự đoán đó. Để giải quyết các trường hợp bất đồng này, nhóm đã xây dựng một tập hợp các mô hình chuyên biệt, được gọi là các **bộ phân giải chuyên gia (expert resolvers)**. Mỗi bộ phân giải này là một mô hình Qwen3-4B được tinh chỉnh (fine-tuned) cho một tác vụ phân loại nhị phân cụ thể trên từng cặp nhãn:

- Một bộ phân giải chuyên biệt cho cặp lớp (0, 1).
- Một bộ phân giải chuyên biệt cho cặp lớp (0, 2).
- Một bộ phân giải chuyên biệt cho cặp lớp (1, 2).

Quy trình hoạt động như sau: khi có sự xung đột trong dự đoán giữa mô hình Ensemble và Qwen2.5-7B, mẫu dữ liệu đó sẽ được chuyển đến bộ phân giải chuyên gia tương ứng. Ví dụ, nếu Ensemble dự đoán nhãn là 1 và Qwen2.5-7B dự đoán là 2, mẫu dữ liệu sẽ được đưa vào bộ phân giải (1, 2) để đưa ra quyết định cuối cùng. Cách tiếp cận này giúp tận dụng sức mạnh của các mô hình chuyên biệt để giải quyết các trường hợp khó, qua đó tăng độ chính xác của toàn hệ thống.

3 Kết quả và phân tích

Mô hình	**Accuracy**	**F1 Macro**
Qwen2.5-7B-Instruct	0.8395	0.839
Qwen2.5-7B	0.8429	0.842
Qwen3-4B	0.8448	0.851
EnsembleNN	0.8643	0.863
Ensemble + Experts	**0.8671**	**0.866**

Bảng 1: So sánh kết quả giữa các mô hình

Phân tích cho thấy EnsembleNN giúp giảm sự thiên lệch giữa các mô hình riêng lẻ. Post-processing đóng vai trò quan trọng trong việc sửa lỗi biên, nơi các mô hình có xác suất gần nhau.

Tài liệu

[1] Unsloth Team, “Memory-efficient Fine-tuning for LLMs,” 2024. [Online]. Available: https://github.com/unslothai/unsloth

[2] L. Ouyang, J. Wu, X. Jiang, et al., "Training language models to follow instructions with human feedback,"*Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 27730-27744, 2022.

[3] Z. Lai, X. Zhang, and S. Chen, "Adaptive Ensembles of Fine-Tuned Transformers for LLM-Generated Text Detection,"*arXiv preprint arXiv:2405.16104*, 2024.

[4] An Yang *et al.*, "Qwen3 Technical Report,"*arXiv preprint arXiv:2505.09388*, 2025.

[5] E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models,"in *International Conference on Learning Representations (ICLR)*, 2022.

[6] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization,"in *International Conference on Learning Representations (ICLR)*, 2019.