LLATMU at #SMM4H-HeaRD 2025: Finetuned-LLM vs BERT

Hsuan-Lei Shao[1], Chih-Chuan Fan[1], Wei-Hsin Wang[2], Sieh-Chuen Huang[2]*

[1]Graduate Institute of Health and Biotechnology Law, Taipei Medical University
[2]College of Law, National Taiwan University
hlshao@tmu.edu.tw, gingerfanfan@gmail.com, wangweihsin07@gmail.com, schhuang@ntu.edu.tw

Abstract

This paper presents LLATMU's participation in the #SMM4H-HeaRD 2025 Task 3, which focused on identifying tweets that indicate caregiving for family members with dementia. We compare two approaches: a supervised BERTweet classifier and a generative Meta-LLaMA-3-8B model fine-tuned by Low-Rank Adaptation (LoRA). The BERTweet model achieved a higher F1-score (0.948) than the LLaMA-3 model (0.931), but the LLM model demonstrated strong contextual understanding, especially for vague or implicit caregiving references. By reformatting the task into a Yes/No prompt-based classification and incorporating a final verification layer using the OpenAI API, the LLaMA-3 system exhibited promising semantic inference capabilities with limited data.

While traditional transformer models remain strong baselines, our findings suggest that fine-tuned LLMs are viable and scalable alternatives in low-resource settings. The use of parameter-efficient tuning and model quantization enabled the adaptation of a large LLM within constrained computational environments. Moreover, LLMs' multilingual capacity offers potential for broader public health surveillance. This study highlights the strengths and limitations of generative LLMs in health-related NLP tasks and points toward future directions in data augmentation, multilingual modeling, and real-time health signal detection.

Introduction

Social media platforms have emerged as essential resources for individuals to share personal health experiences, seek medical advice, and discuss healthcare concerns publicly. The rich volume of user-generated content (UGC) available on these platforms offers valuable opportunities for researchers to understand public perceptions of various health conditions, contributing significantly to advancements in digital health surveillance and epidemiological research. Among these platforms, Twitter stands out due to its real-time and open nature, becoming an instrumental source for studies on disease surveillance, mental health, and medication usage behaviors.

In this report, we address Task 3 ("Detection of Dementia Family Caregivers on Twitter") from the 2024 Social Media Mining for Health/Health Real-World Data (#SMM4H-HeaRD) shared tasks (Klein et al. 2025). To address the task's challenges, we experimented with fine-tuning a large language model (LLM)—Meta-LLaMA-3-8B—with 4-bit quantization and Low-Rank Adaptation (LoRA). Our results indicated that the fine-tuned LLM achieved a competitive F1-score of 0.931, though this performance fell short of both our own baseline method, a specialized BERTweet-based model (F1-score = 0.948), and the task's official benchmark (F1-score = 0.96). The strong performance of existing benchmarks highlights the primary challenge of this classification task: surpassing well-established and highly effective baseline models.

Related Work

In recent years, the SMM4H shared tasks have reflected a growing convergence between traditional machine learning architectures and large language models (LLMs) in health-related natural language processing. A comparative analysis of participating teams' approaches in Task reveals three key insights. First, domain-specific pretrained models such as **BERTweet** consistently outperformed generic transformer models like RoBERTa in tweet classification tasks. For instance, the LT4SG team achieved an **F1-score of 0.938** using an ensemble of BERTweet-large models, surpassing their RoBERTa-large variant (F1 = 0.925) and Golden_Duck's RoBERTa-based system (F1 = 0.928) (see Table:1). (Fan, Yang, and Cao 2024; Li et al. 2024; Mia et al. 2024; Athukoralage et al. 2024)

Table 1: Performance Comparison of Related Articles

Team	Model	Method Description	F1 Score (Test)
Golden_Duck	RoBERTa-large (layer concat + pooling)	Weighted concatenation of last 4 hidden layers, followed by pooling	0.928
LT4SG	RoBERTa-large	Single fine-tuned RoBERTa-large model	0.925
LT4SG	BERTweet-large (Ensemble)	Ensemble of 3 BERTweet-large models with hard voting	0.938
CTYUN-AI	Qwen-72B-Chat + cross-task training	LLM fine-tuning with Task 6 labeled/unlabeled data via cross-task strategy	0.956

Research Design

Based on these findings, our study design incorporates **both BERTweet and LLMs** into our system architecture. However, we differentiate ourselves by **fine-tuning LLMs using machine learning principles**, rather than relying on zero-shot prompting alone. Specifically, we apply **parameter-efficient adaptation via LoRA (Low-Rank Adaptation)** to

*Corresponding author, orcid: 0000-0003-3571-5236.

large-scale LLMs like Meta-LLaMA-3-8B, enabling them to specialize in detecting tweets that report family caregiving for dementia. This approach aims to bridge the performance gap between zero-shot LLMs and fully supervised transformers by combining the generalizability of LLMs with domain-specific fine-tuning.

Meta-LLaMA-3 Quantization and LoRA Fine-tuning

The Meta-LLaMA-3 8B model was fine-tuned using **4-bit quantization** to reduce memory usage and computational cost. Quantization was implemented with load_in_4bit=True, along with **double quantization** enabled (bnb_4bit_use_double_quant=True) and the **NF4 quantization type** (bnb_4bit_quant_type="nf4"). The training precision was set to **torch.bfloat16** to maintain numerical stability during low-precision computation.

For parameter-efficient adaptation, we employed **LoRA (Low-Rank Adaptation)** across multiple transformer components, including q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj, and lm_head. The LoRA configuration included a **rank of 32,** α **(alpha) value of 64**, and a **dropout rate of 0.05**.

Tokenizer Configuration and Sequence Length

Analysis We manually added a [PAD] token to the LLaMA tokenizer, and enabled add_bos_token and add_eos_token for proper sequence framing. The model's token embedding matrix was resized accordingly. We performed a sequence length analysis using Matplotlib, and found that a maximum length of **60 tokens** was sufficient.

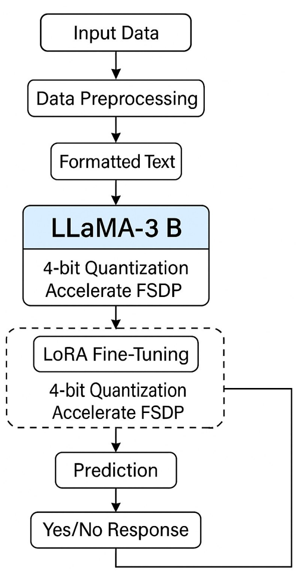

Figure 1: Overview of the Proposed Model Architecture

Inference and Strategy Integration

The inference process consisted of two pipelines:

- **BERTweet**: Texts were directly fed into the fine-tuned classification model to generate Yes/No predictions.
- **LLaMA-3 (LoRA)**: Preprocessed Q&A format was used to prompt the model, and the probability of generating "Yes" was computed. A threshold of 0.4 was used: if P("Yes") $>$ **0.4**, the prediction was treated as positive ("1"), otherwise negative ("0").

We selected the threshold of 0.4 based on empirical evaluation on the validation set. After fine-tuning, the LLM was prompted in a question-answering format, and we extracted the softmax probability of generating the token "Yes" as the decision score.

To determine the optimal threshold, we evaluated F1 scores across a range of thresholds from 0.0 to 1.0 in increments of 0.05. For each threshold t, a tweet was classified as positive if $P(Yes) > t$, and negative otherwise. As shown in the Figure below, the threshold of 0.4 yielded the highest accuracy score on the validation set. This suggests that 0.4 provides the best balance between precision and recall under the current calibration of our model.

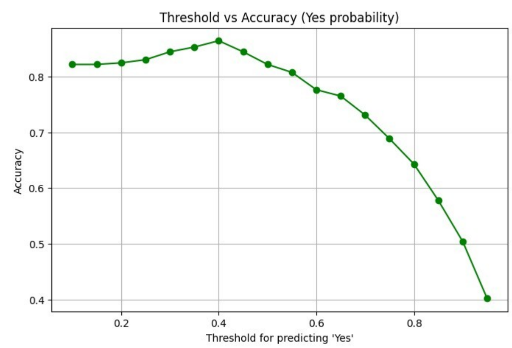

Figure 2: Overview of Threshold Selection

OpenAI-based Answer Verification

To enhance the interpretability and reliability of our predictions, especially for ambiguous or borderline cases, we introduced a final verification layer using the OpenAI GPT-4o API with default parameters: temperature = 1.0, top_p = 1.0, and the prompt is:

*The model predicted: **Yes/No**. Do you think this is correct?*

If the OpenAI response confirmed the prediction (e.g., by replying "That is correct" or similar), we accepted the result as validated. This final step adds a lightweight human-aligned verification layer, especially useful in high-risk decision scenarios.

Results

While our BERTweet-based model did not outperform the benchmark results, we observed notable potential in the fine-tuned LLM for detecting references to dementia diagnoses among family members. By reformulating the task as a generative classification problem, we prompted the LLaMA-3 model using a structured Yes/No format and used the model's confidence in generating "Yes" as a binary decision rule. Specifically, a response probability greater than **0.4** was treated as a positive classification. On the validation set, which consisted of **353 labeled tweets**, the validation and test dataset performance of our BERT and LLM model, and other contestants as follows table:

Table 2: Qualitative Comparison of Misclassified Examples

Model	F1-score	Precision	Recall
BERT (validation)	0.947	0.925	0.970
Finetuned LLM (validation)	0.860	0.860	0.860
BERT (test)	0.948	0.911	0.987
Finetuned LLM (test)	0.931	0.896	0.969
Mean of all contestants	0.885	0.925	0.892
Median of all contestants	0.953	0.946	0.969

These results indicate that while the LLM-based system did not surpass traditional discriminative models in overall performance, it demonstrated strong semantic comprehension and sensitivity to implicit health disclosures. With further tuning or assembly, such models could play a valuable role in enhancing signals. For example, we can compare the mistaken predictions of the two models:

Table 3: Compared Examples LLM and BERT Errors

Sentence (excerpt)	Label	LLM	BERT
Darkness overwhelms my mother's memory. And yet dementia is not too dark...	1	0	1
You never really see that dark side of hearing your grandmother with dementia calling her son...	0	1	0
My grandfather had dementia Alzheimerś disease. John shows he has cognitive decline too.	1	0	1
I thank my sister everytime I fill up. I love her but I cannot believe she voted for this dementia fool in there now.	0	0	1
Irish Alzheimer? you remember nothing but the grudges. My grandma from Castlerea used to say that.	0	0	1
Irish my mum suggesting i gaslight the grandparents and say ive always been theo cos they have dementia...	1	1	0

We observed complementary strengths and weaknesses in how the fine-tuned LLM and the BERT-based classifier handle complex or ambiguous expressions related to caregiving. The LLM tends to overinterpret emotionally expressive language or the co-occurrence of dementia-related terms with family references (e.g., "mum", "hug", "my dad"), especially in structurally ambiguous sentences where the caregiver role is not explicit. It also misclassifies tweets containing indirect mentions, such as prayers, eulogies, or narrative reflections, as caregiving. These tendencies lead to false positives in emotionally charged but non-caregiving contexts.

In contrast, the BERT model demonstrates a more conservative classification strategy. Its reliance on token-level co-occurrences between family-related terms and medical keywords helps it correctly reject non-caregiving tweets when language is casual or contextually unclear. However, this conservatism also causes BERT to miss tweets that imply caregiving through subtle cues, such as indirect involvement, poetic metaphors, or long-term emotional labor. In such cases, the LLM's deeper contextual understanding and sensitivity to speaker intent allow it to succeed where BERT fails.

These contrasting behaviors suggest that LLMs and BERT offer distinct advantages: BERT performs well in structurally clear instances using lexical heuristics, while LLMs provide stronger performance in context-rich or pragmatically complex tweets where surface tokens alone are insufficient.

This study presents a novel architecture that integrates large-scale generative models with instruction-style fine-tuning for a real-world public health classification task. Our results suggest that such LLM-based frameworks are viable in **low-resource environments**, where labeled data is scarce, and provide a scalable alternative to traditional supervised learning in health-related NLP applications.

References

Athukoralage, D.; Atapattu, T.; Thilakaratne, M.; and Falkner, K. E. 2024. LT4SG@SMM4H'24: Tweets Classification for Digital Epidemiology of Childhood Health Outcomes Using Pre-Trained Language Models. In *Proceedings of the 9th Social Media Mining for Health Applications (SMM4H) Workshop and Shared Tasks*, 38–41. Association for Computational Linguistics.

Fan, Y.; Yang, D.; and Cao, L. 2024. CTYUN-AI@SMM4H-2024: Knowledge Extension Makes Expert Models. In *Proceedings of the 9th Social Media Mining for Health Applications (SMM4H) Workshop and Shared Tasks*, 5–9. Association for Computational Linguistics.

Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Gryboski, L.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Mazzotti, D.; Onishi, T.; Powell, J.; Raithel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.

Li, H.; Zhang, Y.; Zhang, Y.; Jiang, S.; and Dong, B. 2024. SRCB at #SMM4H 2024: Making Full Use of LLM-based Data Augmentation in Adverse Drug Event Extraction and Normalization. In *Proceedings of the 9th Social Media Mining for Health Applications (SMM4H) Workshop and Shared Tasks*, 32–37. Association for Computational Linguistics.

Mia, M. A.; Yahan, M.; Murad, H.; and Khan, M. I. 2024. Golden_Duck at #SMM4H 2024: A Transformer-based Approach to Social Media Text Classification. In *Proceedings of the 9th Social Media Mining for Health Applications (SMM4H) Workshop and Shared Tasks*, 28–31. Association for Computational Linguistics.