School of Systems Science and Industrial Engineering

SHIFTING THE PARADIGM: A GENERATIVE AI FRAMEWORK FOR LARGE LANGUAGE MODELS INTEGRATION IN HEALTHCARE TEXT CLASSIFICATION

PH.D. DISSERTATION DEFENSE

Hajar Sakai

Thursday, May 1, 2025, 10:00 a.m. to 12:00 p.m.

Location: Zoom Link

ABSTRACT

Large Language Models (LLMs) continuously transform how multiple Natural Language Processing (NLP) tasks are handled, and text classification is no exception. Their impact has reached various domains and industries, which include high-stakes and regulated ones such as healthcare. Healthcare text classification confronts unique challenges that stem from complex medical terminology, intricate clinical relationships, strict regulatory requirements, limited labeled textual data, and computational resource constraints. The conventional paradigm based on a sequential pipeline that consists of text preprocessing, generation of embeddings, and document categorization, traditionally adopted for text classification, often struggles with these aspects, which lead to suboptimal performance. Current literature focuses on developing approaches where LLMs are leveraged to advance healthcare text classification. However, as these models are both relatively recent and undergoing continuous development, there is still room for more exploration, especially in healthcare settings. Moreover, this research area is rapidly gaining momentum due to its transformative potential to revolutionize healthcare through data-driven decision-making. On one hand, healthcare textual data is incessantly generated yet rarely investigated. On the other hand, LLM-based approaches are introducing more efficient solutions.

This dissertation contributes to the growing research that explores healthcare text classification using LLMs. It emphasizes the importance of shifting the paradigm deployed for this particular task, given the previously mentioned challenges the conventional paradigm faces and the status quo of the literature. The ultimate goal is to contribute to the transformation of how healthcare professionals interact with and derive insights from healthcare narratives, potentially advancing healthcare information management and decision-making processes. For this purpose, three LLM-based approaches that tackle one or more of the future directions identified from a systematic review of LLMs for healthcare text classification are proposed. The first, QUAD-LLM-MLTC, uses four distinct LLMs in a prompt engineering-based ensemble learning setting to perform Multi-Label Text Classification (MLTC). This approach outperforms existing approaches. The results, which achieve an F1 score of 79.28% ± 0.93%, demonstrate the importance of providing rich context inputs in prompt engineering and the value of developing a hybrid approach that leverages the strengths of different LLMs. The second approach, KDH-MLTC, proposes a knowledge distillation-based fine-tuning sequential training approach for robust MLTC that can be run locally. Reaching an F1 score of 82.70% ± 0.89%, this approach outperforms not only existing models but also achieves marginally better results than QUAD-LLM-MLTC; however, the latter remains the preferred choice for small datasets or scenarios with limited annotated data. The third approach, HAMLET, uses an LLM for topic generation, SBERT-BERT Hybridization for embeddings, Semantic-Geometric Similarity Method for similarity computation, and Graph Neural Networks (GNNs) for topic refinement, which achieves a topic modeling composite score that varies between 70.50% and 75.90% across six English and French datasets. The proposed novel framework improves healthcare text classification and balances accuracy and computational efficiency; this research can be extended to text classification tasks in other domains.