Medical Text Classification Using Large Language Models

1) Background of the Area of Your Project Problem and the Important Problems to Be Solved in That Area

Medical text classification is a critical task in the healthcare industry, particularly in areas like pathology reporting and medical review analysis. Healthcare professionals generate large amounts of unstructured data in the form of medical reports, notes, and findings, which often need to be categorized for better management, analysis, and decision-making. Medical reports are usually complex, contain domain-specific terminology, and vary in format and quality. As a result, automating the classification and extraction of useful information from these reports can greatly improve efficiency, reduce human error, and enhance clinical decision-making.

Despite the significant advances in Natural Language Processing (NLP) over the years, challenges remain in effectively classifying complex medical texts. Traditional NLP techniques like TF-IDF and machine learning models provide baseline solutions, but they often struggle with handling noisy, unstructured, and domain-specific medical texts. Moreover, these methods typically lack the ability to capture contextual meaning, which is essential for understanding complex medical terminology.

Recent advancements in deep learning and large language models (LLMs), such as BioBERT and ClinicalBERT, have shown promise in addressing these challenges. These models are pre-trained on medical texts and exhibit the ability to capture intricate relationships within medical language. However, the adaptation of these models to specific medical datasets, as well as the application of techniques like few-shot learning, remain relatively unexplored in the context of medical text classification.

The project aims to bridge this gap by exploring how LLMs can improve the classification of pathology reports, particularly in the context of cancer classification, and compare their performance with traditional models.

2) Identify the Project Problem to Be Solved

The problem to be solved is the accurate classification of complex medical texts, specifically pathology reports, using both traditional text classification methods (e.g., TF-IDF and machine learning) and advanced LLMs. The project will focus on overcoming the challenges of handling noisy, unstructured medical text, and adapting large language models to effectively classify medical data related to cancer types.

3) Justify Why This Project Problem is Important and Worth Your Investigation

Accurately classifying pathology reports and other medical texts is crucial for improving clinical decision-making and patient outcomes. With the vast amount of unstructured medical data available, it is impractical to rely solely on manual methods for categorizing and analyzing reports. Automated classification systems can significantly reduce the burden on healthcare professionals, enhance data accessibility, and ensure better management of patient information.

While traditional methods have been used in the past, the advent of LLMs, particularly domain-specific models like BioBERT and ClinicalBERT, offers an exciting opportunity to enhance classification accuracy. These models have demonstrated impressive results in medical text understanding, but their application in the context of pathology report classification, specifically for cancer-related datasets, is not well-explored.

This project is particularly timely given the increasing volume of electronic health records (EHR) and the need for effective systems to handle this data. By comparing traditional methods with LLMs, this project aims to contribute valuable insights into the effectiveness of modern NLP techniques in healthcare, with the potential to improve clinical workflows and support decision-making processes.

4) Describe the Expected Result from This Project and the New Features of Your Project Compared with Existing Similar Products

The expected results from this project are as follows:

1. **Baseline Model**: A traditional model for medical text classification, using TF-IDF combined with machine learning models, will be built to establish a benchmark for comparison.
2. **LLM-Based Model**: A large language model, such as BioBERT or ClinicalBERT, will be fine-tuned or adapted to classify pathology reports based on cancer type. This model is expected to outperform the traditional baseline model in terms of classification accuracy and F1 score.
3. **Few-Shot Learning Techniques**: The project will also explore few-shot learning using the LLMs, which can potentially offer strong performance with a smaller amount of labeled data, an essential factor for many real-world applications where labeled data is limited.

Compared to existing products, this project will contribute new insights into how domain-specific LLMs can be adapted to medical datasets for better text classification performance. It will also provide a comparison between traditional and cutting-edge models, offering a clearer understanding of their respective strengths and weaknesses in medical contexts.

5) Present the Feasibility of the Project

This project is feasible within the given time constraints, primarily due to the use of pre-trained large language models, which eliminate the need to train a model from scratch. The process of fine-tuning a pre-trained LLM on a medical dataset, such as TCGA Pathology Reports, is manageable within the available time, particularly using LoRA or Adapter methods for domain adaptation. These techniques are computationally efficient, allowing for the adaptation of LLMs to specific datasets without requiring extensive computational resources.

Furthermore, the use of TF-IDF and machine learning models as a baseline will allow for a quicker comparison and evaluation of the LLM-based approach. The expected evaluation of performance, including accuracy and F1 score, will be straightforward, and the exploration of few-shot learning is an additional task that is feasible within the timeframe.

6) References

1. Lee, J., Yoon, W., Kim, S., Kim, D., & So, C. H. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics (ACL)*.
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., & Naumann, T. (2019). Publicly available clinical BERT embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.
3. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., & Hovy, D. (2016). Hierarchical attention networks for document classification. *Proceedings of NAACL-HLT 2016*.