ITESM Campus Monterrey

**Proof of Concept: Fine-Tuning BERT for Medical Text Classification**

Luis Alberto Portilla López

Research Stay - Going beyond Artificial Intelligence: Artificial Emotions

TC3073 | Group 573

## Introduction

Pre-trained language models, such as BERT, have transformed natural language processing (NLP) by enabling fine-tuning for domain-specific tasks. Fine-tuning involves adapting these models to specialized datasets, allowing them to handle the unique vocabulary and context of specific domains. This Proof of Concept (PoC) focuses on fine-tuning BERT for medical text classification, a critical application in healthcare analytics.

## Business Problem

The medical field produces vast amounts of unstructured text data, including patient records, research articles, and clinical notes. However, using this data effectively for tasks like text classification is challenging due to:

* Terminology Complexity: Specialized vocabulary and abbreviations in medical texts.
* Data Scarcity: Limited availability of labeled datasets for supervised learning.
* Generic Models: Difficulty in achieving high accuracy with models not tailored to medical contexts.

## Proposed Solution

This PoC proposes fine-tuning a pre-trained BERT model on a medical dataset to perform text classification tasks. The approach includes:

* Dataset: Use labeled datasets such as MIMIC-III or PubMed for fine-tuning and evaluation.
* Fine-Tuning: Adapt the BERT model using domain-specific data to improve understanding of medical terminology.
* Evaluation Metrics: Assess performance using metrics like accuracy, precision, recall, and F1-score.
* Comparison: Benchmark the fine-tuned model against generic BERT and traditional classifiers.

## Expected Outcomes

The implementation is expected to achieve:

* Improved Classification Accuracy: Higher performance in categorizing medical text due to fine-tuning.
* Domain Adaptability: Demonstrate the effectiveness of fine-tuning for handling domain-specific tasks.
* Insights into Model Performance: Highlight the importance of adapting pre-trained models for specialized NLP applications.

## Conclusion

This PoC demonstrates the potential of fine-tuning pre-trained models like BERT for domain-specific tasks, focusing on medical text classification. By addressing challenges such as terminology complexity and data scarcity, fine-tuned models provide robust solutions for specialized NLP applications. The findings will inform future research and deployment strategies for domain-specific tasks.