NLP for Healthcare - CL3.411
Assignment 1

De-identification in Medical Report

Deadline: 28th August, 2025 11:59 PM

General Instructions:

- Submit your code as either a Jupyter Notebook (.ipynb) or a Python script (.py).
- De-identified output must be saved in a CSV file with the column header named de-identified.
- Clarity and depth of explanation will be considered during grading.

De-identification refers to the process of removing personal identifiers from protected health information so that individuals cannot be readily identified. In this assignment, you will explore different approaches to de-identification and evaluate their effectiveness.

NOTE: The assignment will be released in 2 parts, and each part will have its own deadline.

- Part 1: Literature Review & Report
- Part 2: Experimentation & Comparative Analysis

1 Literature Review & Report: [25 Marks]

Deadline: 22nd August, 2025 11:59 PM

In this part, you are expected to submit a detailed report covering different ways to de-identify personal information in healthcare datasets.
Your report should include:

1. **Basic approaches to de-identification**
   - Example: Rule-based approaches such as Regex for names, dates, emails, phone numbers, etc.
2. **Other approaches to explore**
   - Statistical, ML-based, or hybrid methods.
3. **Model-based methods**
   - If any models are available (domain-specific or general), provide a small description of their architecture, datasets used for training, and their typical performance.
4. **Proposed ideas/pipelines**
   - If you have any new ideas or workflows for de-identification, explain them.
   - Can this be done with the existing dataset?
   - If not, can data be created synthetically?

References: You must cite any papers, tools, or healthcare information policies you refer to.

Submission: A written report in PDF format. Please keep the report concise (2–4 pages recommended).

2 Part 2: Experimentation & Comparative Analysis: [55 Marks]

Deadline: 28th August, 2025 11:59 PM

In this part, you are given a set of healthcare medical discharge reports containing personal information. Your task is to perform de-identification using three different methods and then analyze the results.

2.1 Tasks: [40 Marks]

Method 1: Basic NER + Regex

- Use any open-source Named Entity Recognition (NER) model to identify names and organizations.
- Use regular expressions (Regex) or other rule-based approaches for dates, emails, phone numbers, and other similar data types.
- Submit the pipeline description and output.

Method 2: Medical Domain NER

- Use any existing medical domain-specific NER model.
- Use any other pipeline, like any machine learning method or any method used in Method 1.

Method 3: Large Language Model (LLM)

- Use the 3B LLM model (without any fine-tuning) for de-identification.
- Compare its output with Methods 1 and 2.

2.2 Report: [15 Marks]

Along with your code/outputs, you must submit a detailed report that includes:

- **Method 1 pipeline:** How you implemented it, where it works well, where it fails, and why.
- **Method 2 model details:** Architecture, dataset, and training information. Discuss the feasibility of training such models. Suggest possible improvements or alternative approaches.
- **Method 3 analysis:** How the LLM performs compared to the other methods.
  - Is it better or worse?
  - If better, how?
  - If worse, how could it be improved? Suggest methods.

Submission: A written report in PDF format. Please keep the report concise (4–6 pages recommended).

3 Viva: [20 Marks]

Good Luck