# *Title: Context-Aware Entity Recognition and Sensitivity Masking*

## *Objective:*

**Problem Statement:** Traditional pattern matching often fails when digital identifiers (User IDs, URLs) mimic standard alphanumeric strings. This assignment requires you to implement a **Deep Learning-based NER pipeline** capable of distinguishing between general strings and PII based on surrounding linguistic context.

**Challenge Scenario:** Given an input string such as:

"Technical support is available via UID: 8829-X or help.desk@corp.net. Visit our portal at https://user-profile.com/john_doe."

Your trained neural network must classify these tokens correctly under a **BIO tagging scheme** and generate a sanitized output:

"Technical support is available via UID: [MASK] or [MASK]. Visit our portal at [MASK]."

### PART A: [7 MARKS]

You can use any publicly available dataset, or alternatively use the attached dataset.

a. Import the dataset and pre-process it.
b. Use appropriate embedding techniques
c. Use appropriate Neural Network / Deep Learning / ML models to train the provided dataset
d. You can also use open source models like BERT, huggingface or SpaCy pre-trained models, but you have to Fine-tune the said models using the provided dataset.
e. Extract Named Entities from the dataset and print a few examples.
f. Build a simple web page which will take text as input and provide masked data as output. (you can also provide appropriate provisions in Jupyter notebook to take text input and print masked output)
g. Evaluate the appropriate identification of entities using appropriate metrics and plot a confusion matrix using test data.

### PART B: [3 MARKS]

h. Build a user-friendly interface using a framework of your choice (HTMl, React etc)
i. Provide feature for users to Upload Text Files for batch processing of entities and events
j. View Results in Real-Time with color-coded highlights for different entity types and events

### PART C: [5 MARKS]

a. Provide a design document discussing your model architecture and training approach.

b. Conduct a literature survey to understand the evolution of Named Entity Recognition and Relation Extraction techniques and approaches.
c. Explore limitations of the approach you have used to implement and suggest improvements.
d. Discuss how LLM can be used to approach this problem and provide suggestion in the design document.

## *Deliverables:*

PART - A

a. A well-documented code (Python (or Jupyter notebook) and frontend web page as applicable) for the application.
b. Design and research Document
c. Instructions for running the application locally.
d. A set of screen shots that shows your test cases, input and output example cases using which you have run your application.

**For any queries, please email to : ajay.naik@wilp.bits-pilani.ac.in**