CITS4012 Natural Language Processing
Individual Assignment Specification

Due: 5 September 2025, 11:59PM AWST

1 Objective

In this individual assignment, you will complete a **Medical Abstract Classification** task to classify patient conditions based on medical abstracts. Using the provided dataset, you will experiment with word embedding methods and recurrent neural network models to evaluate their effectiveness on the classification task.

2 Dataset

You will be provided with:

1. A training set (`medical_tc_train.csv`)
2. A testing set (`medical_tc_test.csv`)
3. The list of class labels (`medical_tc_labels.csv`)

Each instance in the dataset consists of a medical abstract and its corresponding condition label. The dataset download link will be shared on LMS.

3 Task Requirements and Marking Scheme

You are expected to complete the following tasks:

3.1 Preprocess the Dataset (2 marks)

Clean and prepare the text data appropriately for model training. Proper text processing methods should be included.

3.2 Word Embedding Construction (4 marks total)

- Train a word embedding model (e.g., Word2Vec or FastText) on the provided dataset. (2 marks)
- Load a pretrained biomedical word embedding (e.g., BioWordVec or similar) and extract vectors for your dataset's vocabulary. (2 marks)

Notes: the pretrained BioWordVec embedding can be found in the shared dataset folder, in `.bin` *format.*

3.3 Visualization (2 marks)

- Select 10 medical terms (e.g., *aspirin*, *diabetes*—you may choose others).
- Visualize their embeddings from both your trained model and the pretrained model using t-SNE plots.
- Briefly interpret and compare the visualizations.

3.4 Build RNN-Based Model Training (4 marks)

Build two RNN-based models, one using your trained embedding and one using the pretrained embedding, and train them on the classification task.

3.5 Performance Evaluation (3 marks)

Evaluate and compare the two models on the test set using appropriate classification metrics (e.g., accuracy, F1-score). Provide a short analysis of the results.

3.6 Interactive Inference Form (5 marks)

Develop a simple interactive interface that allows a user to:

- **Choose** one of the available word embedding methods.
- **Select** a medical abstract from the test set.
- **Run** the prediction and **display** the selected abstract, the ground truth label, and the predicted label using your trained model.

The interface should be independently runnable in a Google Colab notebook.

4 Submission Notes

- The submission must be made via **LMS**.
- You **must submit a single `.ipynb` file**.
- Name your notebook file as: `CITS4012_YourStudentID.ipynb`.
- The submitted notebook should contain your full implementation for the assignment. Ensure that it is **well-documented and clearly structured**.
- You MUST use the given code template for implementation.
- You may **optionally submit a `.zip` file** that includes:
  - A `README` file with instructions on how to run your code (if not already clear from the notebook)
  - Any trained model files or additional resources required to run your program
- Add **brief explanations of your implementation choices** throughout the notebook where appropriate.
- Ensure that **all relevant output logs are included and saved** in the notebook.
  Marks will not be awarded for parts of the assignment that lack corresponding output logs.

5 Late Submission of Assignment

A penalty of 5 per cent of the total mark allocated for the assessment item is deducted per day for the first 7 days (including weekends and public holidays) after which the assignment is not accepted. For the first two days there will be a penalty waiver, which means that if you submit within that 48 hours period, the assessment will be marked late but the late penalty will not be applied. After 48 hours, the accrued penalty will apply, i.e. a 15% deduction will be applied on Day 3 (after 48 hours), with an additional 5% per day after that (to day 7).