University of Glasgow | School of Computing Science

Assessed Coursework

Course Name	**Text-as-Data**			
Coursework Number	**1**			
Deadline	**Time:**	**4:30pm**	**Date:**	**13th March 2025**
% Contribution to final course mark	**18**			
Solo or Group ✔	**Solo**	**✔**	**Group**	
Anticipated Hours	**30 hours**			
Submission Instructions	**As per specification below.**			
Please Note: This Coursework cannot be Re-Assessed				

Code of Assessment Rules for Coursework Submission

Deadlines for the submission of coursework which is to be formally assessed will be published in course documentation, and work which is submitted later than the deadline will be subject to penalty as set out below.

The primary grade and secondary band awarded for coursework which is submitted after the published deadline will be calculated as follows:

(i) in respect of work submitted not more than five working days after the deadline
   a. the work will be assessed in the usual way;
   b. the primary grade and secondary band so determined will then be reduced by two secondary bands for each working day (or part of a working day) the work was submitted late.

(ii) work submitted more than five working days after the deadline will be awarded Grade H.

Penalties for late submission of coursework will not be imposed if good cause is established for the late submission. You should submit documents supporting good cause via MyCampus.

Penalty for non-adherence to Submission Instructions is 2 bands

You must complete an "Own Work" form via https://studentltc.dcs.gla.ac.uk/ for all coursework

Text-as-Data Coursework

Introduction

The TaD coding coursework aims to assess your abilities to perform text processing techniques as applied to a multi-class classification problem.

Your work will be submitted through a Moodle quiz. For each question, you should submit your text answer (providing the required information) separately to your code.

The task: A museum records organisation has a large set of records that need to be assigned to one of five institutions (in the table below). They have provided a small set of data to be used as the training, validation and test set. Our goal is to build a classifier that could assign an unseen record to the correct institution.

Class Index	Institution
0	National Maritime Museum
1	National Railway Museum
2	Royal Botanic Gardens, Kew
3	Royal College of Physicians of London
4	Shakespeare Birthplace Trust

The dataset can be downloaded with the link: https://tinyurl.com/tadarchives

Generative AI usage policy: You are free to use generative AI in any way during this assessed exercise. Please note your usage in the final question. Material from this coursework may appear on the final exam.

Q1 - Training Data Cleaning [9 marks]

Download and load the dataset. There are some issues with the **training split** of the data that would stop it being used to train a classifier. Report all issues and how you fixed them

Q2: Exploration [5 marks]

Once the training set has been fixed, report the following:

- The sample counts for the training, validation and test sets
- The percentage splits for training, validation and test sets
- The minimum and maximum length (in characters) of the texts. Report separately for the training, validation and test sets
- The most frequent five tokens in each class (after tokenizing with `text_pipeline_spacy` from Lab 2)

Q3: Prompting with a large language model [10 marks]

A colleague has tried prompting a large language model (Llama-3.1-8B-Instruct) to classify each of the records in the training set. They evaluated three different prompt templates and saved the results to the provided files. Calculate the accuracy, macro precision, macro recall and macro F1 for each prompt template and comment on the result. Consider any invalid output from the LLM as predicting a sixth hypothetical class.

Q4: Fine-tune a transformer [10 marks]

Fine-tune a ‘bert-base-uncased’ transformer model on the model using the training set. You should use an AutoModelForSequenceClassification and the HuggingFace trainer. Use 8 epochs, a learning_rate of 5e-5 and a batch size of 8.

Evaluate on the validation set. Report the per-class precision, recall and F1 score as well as the accuracy, macro precision, macro recall and macro F1 score. Don’t be surprised with poor performance (see the next question)

Q5: A problem with the validation set [5 marks]

There is an issue with the validation set which causes poor performance. Provide the confusion matrix. Describe the problem, how you identified it and how you fixed it.

Q6: Hyperparameter tuning [12 marks]

Train and evaluate several fine-tuned transformer models using the corrected training and validation sets. Try the four base models listed below. Use 8 epochs, a learning_rate of 5e-5 and a batch size of 8. Ideally, you would try different base_models/learning_rates/batch_sizes/etc, but we will limit this to evaluating four different base models and keep the remaining hyperparameters static.

Base models to try: 'bert-base-uncased','roberta-base','distilbert-base-uncased', and 'microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract'

This time, we want the best model found during the training process for each base model and to save it for the final analysis. For example, if the model after 3 epochs is the best performing on the validation set (by macro-F1), we want to keep that. You should investigate the load_best_model_at_end parameter for the Trainer (which does require other parameters).

Evaluate each fine-tuned model on the validation set. Report the per-class precision, recall and F1 score as well as the accuracy, macro precision, macro recall and macro F1 score. Comment on the performance of each model.

Q7: Final evaluation and deployment [6 marks]

Load the best model (based on macro-F1 on the validation set) that was saved in the previous question using a ‘text-classification’ pipeline. Evaluate the best of the four fine-tuned models on the testing set.

State which model you used and report the per-class precision, recall and F1 score as well as the accuracy, macro precision, macro recall and macro F1 score. Comment on the performance and discuss whether the quality is high enough to be deployed for the client.

Q8: Generative AI usage [1 mark]

Report on whether and how you use generative AI in this assignment