Coursework Administrative Details

Module/Lecture Course:	Natural Language Processing
Deadline for submission:	Individual reports (1500 words):
Submission instructions:	Submit all files via blackboard
Submission file type(s) required:	Pdf + Jupyter notebook(s)
Format:	Report as a pdf document. Accompanying data analysis for individual report as a Jupyter notebook. Do not put your name on your report, just your username.
Contribution:	The report contributes 100% to the final mark for the module.

In accordance with university procedures, **submissions that are up to 5 working days late will be subject to a cap of the module pass mark**, and **later submissions will receive a mark of zero**.

Stance Detection

General Requirements

Students are expected to work on the coursework individually.
The students will work on the task of fake news detection. The task focuses on solving fake news detection by predicting the stance associated to every new article using the FNC dataset (https://github.com/FakeNewsChallenge/fnc-1). FNC dataset consists of 49,973 pairs of headlines and article bodies. The body text is annotated by the following classes: Agree, Disagree, Discuss, or Unrelated to the headline.

Input: A headline and a body text
Output: Classify the stance of the body text relative to the claim made in the headline into one of four categories:

- Agrees: The body text agrees with the headline.
- Disagrees: The body text disagrees with the headline.
- Discusses: The body text discusses the same topic as the headline, but does not take a position
- Unrelated: The body text discusses a different topic than the headline

The data suffers from an imbalance problem where around 70% of the articles are unrelated.

Students are expected to:

1- Implement word embeddings using standard semantic based techniques and neural techniques.
2- Understand the challenges of the provided problem and suggest solutions to handle the imbalance nature of the data.
3- Implement natural language processing models and classifiers to predict the right category of a given test example.
4- Develop hierarchical Deep Learning models.

Individual Report [100%]

Each student should separately develop their own NLP models to classify news articles into one of the four categories. Write a report (max 1,500 words) on the ***challenges*** the dataset present, the ***solutions***, and your ***findings*** which will be assessed as follows:

1) Apply the following feature extraction techniques and explain how they work and discuss their advantages and disadvantages
    a) Term Frequency-Inverse Document Frequency (TF-IDF) **[5%]**
    b) A Transformer of your choice (e.g., BERT, GPT) **[10%]**

2) Two step Classification:
    a) **Related/Unrelated classification**:
        i) Use the features extracted using TF-IDF (1.a) and your chosen transformer (1.b) to train a standard Machine Learning method e.g., SVM, Logistic Regression, Random Forest, and discuss its performance on the testing set to classify whether the article body is *related* or *unrelated* given the headline. [**10%**]
        ii) Train one Deep Learning model (e.g., LSTM, RNN, CNN, hybrid model) using TF-IDF (1.a) and transformer embedding (1.b). Explain the architecture of the Deep Learning model, the hyper-parameters used, and the loss function. Discuss the performance on the testing set to classify whether the article body is *related* or *unrelated* given the headline. [**15%**]
        iii) Analyse and compare the performance results for both ML and DL models. [**5%**]

    b) **Agree/Disagree/Discuss classification**: Build a new Deep Learning model on top of the best performing models you implemented in a)
        i) Build a Deep Learning model of your choice to classify articles into the remaining three categories (Agree/Disagree/Discuss) [**15%**].
        ii) Analyse the performance of your model and report the results. **[10%]**

    c) Combine the two models in a) and b) to ***test*** your model end to end, report and discuss the overall performance of your solution. **[15%]**

3) What are the ***ethical implications*** of your proposed solutions? What are the potential biases and future misuse cases? **[10%]**

4) Academic English writing, with good use of technical vocabulary, correct grammar, appropriate document structure and referencing where relevant. [**5%**]

The summative submission deadline is 02/05/2023 at 14:00

The coursework aims at evaluating the students' knowledge and their understanding of the fundamentals and advances in NLP and not their programming skills. Therefore, we ask you to implement the solutions using any Python libraries you are most comfortable with, this includes and is not limited to, PyTorch, Keras, TensorFlow, SpaCy, HuggingFace, Gensim, NLTK…

The report should include the following sections: Introduction, Problem Definition, Proposed Solutions, Analysis of Results, Discussion, Ethical Implications, and Conclusion. The report should use diagrams, figures, and tables to demonstrate the results and analysis.
You should submit your 1,500-word report and the associated Jupyter notebook used to produce your analysis and graphs. **Jupyter notebook file should be saved along with all the produced outputs, results, and figures.**

We understand that not every student has access to the same equipment and therefore this could introduce bias in model performance regardless of the of the quality of proposed solutions. Therefore, using high spec GPUs that can accelerate the performance with longer runs (e.g., epochs) will not grant the student extra marks.

The report word count should:

- *Include* all the text, including title, preface, introduction, in-text citations, quotations, footnotes, and any other item not specifically excluded below.
- *Exclude* diagrams, tables (including tables/lists of contents and figures), equations, executive summary/abstract, acknowledgements, declaration, bibliography/list of references and appendices

Examiners will stop reading once the word limit has been reached, and work beyond this point will not be assessed. Checks of word counts will be carried out on submitted work. Checks may take place manually and/or with the aid of the word count provided via an electronic submission. Students are strongly advised to use Arial font size 12 for their assignments.

PLAGIARISM and COLLUSION

Your assignment will be put through the plagiarism detection service on Ultra.
Students suspected of plagiarism, either of published work or work from unpublished sources, including the work of other students, or of collusion will be dealt with according to Computer Science and University guidelines.

FAQ

1. Can I use existing github code as part of my solution?
   Yes, under the condition that you reference the code and acknowledge the authors and that the code you borrow from github is not the main part of your solution.

2. How many models do I need to build for step 2.a.i?
   You need to implement and train two ML models. ML model with tf.idf features and ML model with transformer-based features.

3. How many models do I need to build for step 2.a.ii?
   You need to implement and train two DL models. DL model with tf.idf features and DL model with transformer-based features.

4. In step 2.b.i, what do you mean by building DL model on top of the best performing model from the previous step?
   This means that you need to build a hierarchical architecture, where the first model makes the prediction on whether a sample is related of unrelated. If the sample is predicted as related, it will be passed to the second model where it will be used in the prediction of Agree/Disagree/Discuss.

5. Do you expect us to build and train a new model for step 2.c?
   No! The question only asks to test the two models you already built and trained in step 2.a and 2.b in a hierarchical manner. You will notice that the over-all performance will not be able to outperform the best model in the hierarchy as the over-all error will be accumulated.

6. Why are we asked to write a report and not just a list of answers for the questions raised in the assignment?
   Writing reports is a standard practice in both academia and industry and at your seniority as L4 students, we try here to prepare you to work at a high level that matches your experience and expertise.

7. The test set available online is not labelled, what should I do?
   Please split the training set into training, validation, and testing sets.

8. Do captions count in the word limit?
   No! they do not count in the word limit. However, it is not appropriate to use diagrams or tables merely as a way of circumventing the word limit. Please be reasonable ☺.

9. What is the logic behind choosing this topic for a coursework?
   This topic was chosen to help you strengthen your resume and career potential. Stance detection is a challenging task in NLP and has a big social impact. Talking about this task in an NLP job interview will guarantee drawing interest and questions.