DEPARTMENT OF COMPUTING

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

NLP Coureswork

Author: Anton Zhitomirsky

A Mark Scheme

A.1 Data analysis of the training data (15 marks)

For a written description of the training data. This should include

1. **5 marks**: Analysis of the class labels: how frequent these are and how they correlate with any feature of the data, e.g. input length.

2. **10 marks**: Qualitative assessment of the dataset, considering either how hard or how subjective the task is, providing examples in your report.

A.2 Modelling (40 marks)

For the successful implementation of a classifier model (this could be a transformer or any other ML model of your choice. Do give justification for your choice.):

1. **10 marks**: Successful implementation of a model (train and produce predictions which outperform the F1 score for the RoBERTa-base baseline provided). 7 marks for outperforming the baseline model on the official dev set (0.48) and 3 marks for outperforming the baseline model on the test set (0.49).

2. **5 marks**: Choice of model hyper-parameters and description of your model setup. This should include choosing an appropriate learning rate and checking whether implementing a learning schedule improves performance. Also consider whether your model is cased or uncased. You should mention how many epochs you train the model for, whether you are using any early-stopping, and how you are using the training labels.

3. **10 marks**: Further model improvements (beyond using a bigger transformer model), for example pre-processing, data sampling, data augmentation, ensembling, etc. Two main improvements, with a third less explored improvement is sufficient. For example: try several different data sampling approaches, try several data augmentation strategies by perturbing observations in different ways, and then see if incorporating one of the categorical columns improves performance.

4. **10 marks**: Compare your model performance to two simple baselines (e.g. a BoW model). Share some of the features that one of your baseline models used, and highlight an example misclassified with a suggestion of why the baseline may have made the misclassification.

5. **5 marks**: Description of the model results and your hyper-parameter tuning (some evidence of this is required in your report). Your results should show how the different strategies you have tried impacted the model performance. For any results presented in your paper, you should be clear if these are from your own internal dev set or the official dev set.

A.3   Analysis (15 marks)

Analysis questions to be answered (these questions can be answered without training any additional models): Your report should state the analysis questions so that this can be read as a self-contained report, rather than referring to 'analysis question 1' etc.

1. **5 marks**: To what extent is the model better at predicting examples with a higher level of patronising content? Justify your answer.

2. **5 marks**: How does the length of the input sequence impact the model performance? If there is any difference, speculate why.

3. **5 marks**: To what extent does model performance depend on the data categories? E.g. Observations for homeless vs poor-families, etc.

A.4   Written report (30 marks)

Marks are awarded for the quality of your written report:

1. **5 marks**: Introduction, with an explanation of the task and the dataset. You may want to read/cite the task paper (and any other paper of your choosing).

2. **10 marks**: Readability of the report (language, coherence, clarity of results etc.).

3. **10 marks**: Good use of graphs or results tables that address the analysis questions. Make sure any text on your graphs is clearly readable.

4. **5 marks**: Conclusion, with a summary of your results, and your key findings from the analysis questions. You should suggest at least one further experiment as a next step.