HW06 – ISIS 4221

Natural Language Processing 2023

Due date: 28-11-2023

Coding rules: Use jupyter notebooks and be sure that the notebook is executed and contains the results before submitting. All classes, methods, functions and free-code MUST contain docstrings with a detailed explanation. **Build a notebook for each model selected**.

Report: Together with the notebooks, you must submit a written report (please use pdf format) with the answers to the questions and a short summary of the implementation.

Submission: individually or in a group of two. Make a zip that includes all the files.

Models: You must provide a link to your tunned models (Dropbox, OneDrive, etc).

Datasets

- **Books from Gutenberg project https://www.gutenberg.org/**

You will use the **same** classification dataset built for HW04 to identify the most likely author for a set of input lines of text.

I. Compose a summary text explaining the distinctions among the encoder families: BERT, ALBERT, DistilBERT, ELECTRA, RoBERTA, and MPnet. You will compare at least five different pretrained models belonging to any of the families. You may not select two models from the same family.
II. Fine-tune the pretrained models for the classification task. Implement callbacks to ensure the best model is retained. Plot training and validation loss (be sure that the model is not overfitting). For each model, specify the process for identifying the optimal hyperparameters. Document these hyperparameters in a summary table.
III. Create another table summarizing the classification results in testing dataset, including the confusion matrix and classification report (macro/micro for precision, recall, and F1).
IV. Compare the outcomes of this assignment with those from HW04. Which types of models performed better, and what is the underlying reason?