UNIVERSITY OF SOUTHERN MAINE

Text Mining and Analytics, Spring 2024, Assignment 4

Instructor: Behrooz Mansouri (behrooz.mansouri@maine.edu)

Due: April 8, 2024

Students will continue working with their previous teams. The goal is to continue the task of song classification based on genre using lyrics. In Assignment 3, you have built a large-scale dataset of songs, which you will reuse for training BERT-based classifiers.

Step 1: Pick up a pre-trained model from HuggingFace for text (sequence) classification. In your report, provide a paragraph explaining how this model works and how it differs from $BERT_{base}$/$BERT_{large}$.

Step 2: You will then fine-tune the pre-trained model with your data. Each student will select one hyperparameter to investigate. You should fine-tune the model and study the effect of that hyperparameter on your model. You will provide a table of results for different values for the hyperparameter, along with a paragraph describing the effect of it on the model's effectiveness.

Step 3: As a team, you will decide on the final hyperparameters for fine-tuning your model. Fine-tune the model and then use it to classify the instances in the test set (provided for you in the previous assignment) and report your results. You will do the same for the pre-trained model (with no fine-tuning) as well.

Step 4: Provide a discussion on your results, between the baseline (with no fine-tuning) and fine-tuned models. For each model, provide an example where one worked better than the other.

Deliverables:

- .py files for fine-tuning and running the models and getting the evaluation results
- .pdf file answering questions for each step.
- README file for running your code

On Brightspace, one student will submit the files (.py, .pdf, .md/.txt). Please make sure to include your teammates' names and the tasks they have done at the top of the .pdf file. Note that your code will be run on the previous data files that you gathered from assignment 3. Your README file should explain how to use the data and run your code.

Important note: Using a screenshot of the console as a table of results will result in 0 for the whole team. You should create tables, use meaningful captions for your tables, and refer to the table in your text.

Grading:

For grading, each task will have 5 points individually for each student. 4 points are considered for putting the steps together and providing the final report. A major part of your grade is on the effectiveness of your model. It is important to make sure your code can be run easily and reproduce the results.