Mining Media Data II
Summer Semester 25
Assignment 2

Prof. Dr. Rafet Sifa, Maren Pielka, Lorenz Sparenberg
University of Bonn

Material for this task can be found under: `https://www.statmt.org/wmt21/quality-estimation-task.html`

Introduction

In this assignment, we focus on detecting critical translation errors in machine translation systems. This task requires predicting sentence-level binary scores indicating whether a translation contains at least one critical error. Such errors may have significant implications in health, safety, legal, and other sensitive contexts. This assignment consists of 2 main parts, totaling 50 points (pts.), with a mix of theoretical explanations, data exploration, and model implementation.

In case of computational bottlenecks, you can use free Cloud Computing providers such as Google Colab, etc.

1 Error Detection Categories (10 Pts.)

In this section, briefly describe each type of critical translation error outlined in the task. Also, explain why they would be considered critical in specific application contexts. Include examples where relevant.

2 Model Development and Analysis (30 Pts.)

2.1 Implementation (25 Pts.)

In this task, you should train and evaluate models for the Critical Error detection task. Consider only the binary (*Critical Error* vs. *No critical Error*) classification problem.

- Fine-tune a transformer-based model (e.g. DistilBERT[1]), OR a feature-

[1] `https://huggingface.co/distilbert/distilbert-base-multilingual-cased`

based classifier (e.g. a model from scikit-learn[2] in combination with a Bag of Words/tf-idf featurizer) on the dataset, and evaluate its performance on the development set.

- Use a generative model (e.g. LLama3[3], Phi3[4], Mixtral[5]) in a zero- or few-shot setup for comparison. You are NOT expected to fine-tune this model, just to devise an intelligent prompt.

Provide the code and document each step, including preprocessing, training, and evaluation.

2.2 Result Analysis (5 Pts.)

Compare the results of the models you trained/prompted in the previous exercise. What are strengths and weaknesses of the respective models? Which one would you use in an application scenario, and under which prerequisites?

3 Error Analysis and Trustworthiness (20 Pts.)

3.1 Error Categorization (10 Pts.)

Analyze errors in the test set predictions. What error types do the models specifically struggle with? What could be a explanation for that, and how could the performance be improved?

3.2 Trustworthiness Discussion (10 Pts.)

Critically evaluate the trustworthiness of the developed models:

- Strengths and weaknesses of the approach.
- Potential biases in the data and their implications.
- Suggestions for improving the model and evaluation framework.

4 Submission

All submissions will be made electronically by sending a single .zip file (including your Python code and PDF or Jupyter Notebook) to amllab@bit.uni-bonn.de by the submission deadline with the title **MMD SS2025 Assignment 2 [GroupID]**, where [GroupID] refers to your group id (name). Submissions sent after the deadline and not following the title convention will not be evaluated. The submission deadline for this assignment is on **2.7.25 at 11:59 pm**

[2] `https://scikit-learn.org/stable/`
[3] `https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`
[4] `https://huggingface.co/microsoft/Phi-3.5-mini-instruct`
[5] `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1`

Germany time. Further updates about this assignment will be published on the course website.

Important

The results have to be presented in class by each group. No presentation of the assignment will lead to zero points being awarded to the entire group.

All results should be present in the submitted file without having to execute the code. Missing results will lead to zero points being awarded to the entire group.

You are required to present the final results of each task in the form of a PowerPoint/Beamer presentation.