Q1. Multilabel Emotion Classification

Problem statement: The goal of the task is to classify the given text into appropriate emotion categories. The dataset is collected from various sources for multiple languages.

- Read the dataset from the datalink provided. The dataset is available in multiple languages. Try English and one other language of your choice.
- You are free to choose a model architecture of your choice. Starting with pre-trained models is allowed. **The usage of models explicitly trained for this/similar tasks is discouraged**. Models can be customized as desired. For example, adding modifications to layers, changing architecture, etc.
- The expected output is a vector of 0/1, with each index corresponding to a different emotion. The emotions considered in order are: [anger, fear, joy, sadness, surprise]. Non-English languages have disgust as an additional emotion label.
    - Ex:
        - Input: “I can't move, my hand is stuck, I'm making weird noises, and my mom is screaming.”
        - Output: [0,1,0,1,1] (corresponds to the presence of ‘fear’, ‘sadness’, and ‘surprise’ in the text)
        - Zero stands for the absence of emotion, and One stands for the presence of emotion.
- For evaluation: F1-macro, Precision-macro, and Recall-macro metrics can be used.
- Once you have trained the model, perform some interpretability analysis to understand what your model has learned.
    - Ex: Highlight if the model has learned any undesirable shortcuts to classify and how to address them - Faithfulness).
    - Some techniques for the same:
        - LIME
        - Attention scores based (if using transformer-based models) Ex: BERT Viz
        - You can use any other tools/techniques for this part as well
- Bonus
    - Experiment with different loss functions. Try to modify them based on your ideas to improve classification scores.
    - Use knowledge bases, such as COMET, to provide additional context for the model.

Assignment evaluation points

- **Model architecture:** Finetuning of the base model. Modifications to architecture/layers.
- **Interpretability analysis:** A case study using the analysis results to see which words are essential for the classification. Check if the words highlighted make sense or are noisy.

- **Evaluation Scores**: Report F1-macro, Precision-macro and Recall-macro. Additional metrics of your choice can be added. Scores should show your model is performing at least better than the dataset baseline which has F1-macro=0.3714

Resources

- Multilabel Emotion Classification Dataset link, dataset paper
- COMET-paper   COMET GitHub page
- LIME GitHub link
- BERT Viz

Q2. Pedagogical Ability Assessment of AI-powered Tutors

Introduction

This task aims to assess the pedagogical effectiveness of AI-powered tutors in educational settings. Given textual interactions between AI tutors and students, the objective is to evaluate whether the AI tutors demonstrate appropriate pedagogical abilities. The dataset contains real-world or simulated tutor-student dialogues across various subjects.

Dataset

- The dataset consists of 300 dialogues from MathDial and Bridge datasets, including the context of several prior turns from both the tutor and the student, the last utterance from the student containing a mistake, and a set of responses to the previous student's utterance from 7 LLM-based tutors and human tutors.
- The following fields are included in JSON:
  - conversation_id: a unique identifier for the instance
  - conversation history: the context of several prior turns from the tutor and the student extracted from the original datasets
  - tutor_responses: the set of human tutor responses extracted from the original datasets, as well as responses generated by 7 LLMs-as-tutors, each with a unique identifier
  - response: the response from a particular tutor
  - annotation: the set of annotations
- The dataset can be accessed here.

Tasks

This assignment will include four tasks:

1. **Task 1 - Mistake Identification**: To detect whether tutors' responses recognise mistakes in students' responses. The following categories are included:
   a. `Yes`: the mistake is identified/recognised in the tutor's response
   b. `To some extent`: the tutor's response suggests that there may be a mistake, but it sounds as if the tutor is not certain
   c. `No`: the tutor does not recognise the error (e.g., they proceed to provide the answer to the asked question simply)

2. **Task 2 - Mistake Location**: To assess whether tutors' responses accurately point to a genuine mistake and its location in the students' responses. The following categories are included:

    a. `Yes`: the tutor points to the exact location of a genuine error in the student's solution
    b. `To some exten`t: the response demonstrates some awareness of the exact mistake, but is vague, unclear, or easy to misunderstand
    c. `No`: the response does not provide any details related to the mistake.

3. **Task 3 - Pedagogical Guidance**: To evaluate whether tutors' responses offer correct and relevant guidance, such as an explanation, elaboration, hint, examples, etc. The following categories are included:

    a. `Yes`: the tutor provides guidance that is correct and relevant to the student's mistake
    b. `To some extent`: guidance is provided but it is wholly or partially incorrect, incomplete, or somewhat misleading
    c. `No`: the tutor's response does not include any guidance, or the guidance provided is irrelevant to the question or factually incorrect

4. **Task 4 - Actionability**: To assess whether tutors' feedback is actionable, i.e., it clarifies what the student should do next. The following categories are included:

    a. `Yes`: the response provides straightforward suggestions on what the student should do next
    b. `To some extent`: the response indicates that something needs to be done, but it is not clear what precisely that is
    c. `No`: the response does not suggest any action on the part of the student (e.g., it simply reveals the final answer)

Evaluation

All tasks should use accuracy and macro F1 as the primary metrics. These will be used in two settings in which evaluation should be performed:

- **Exact evaluation**: Predictions should be evaluated for the precise prediction of the three classes ("`Yes`", "`To some extent`", and "`No`")

- **Lenient evaluation**: Since for these dimensions, tutor responses annotated as "`Yes`" and "`To some extent`" share a certain amount of qualitative value, consider "`Yes`" and "`To some extent`" as a single class and evaluate predictions under the 2-class setting ("`Yes`" + "`To some extent`" vs "`No`")

Model Selection & Training

- Choose any model architecture to perform the classification task. Pre-trained models are allowed, but models explicitly trained for this or similar tasks should be avoided.
- Fine-tuning is encouraged. Modifications to model layers or architectures to enhance performance are allowed.
- The expected output is classification labels indicating whether the AI tutor response meets the pedagogical criteria.
- Experiment with different loss functions and modify them based on your ideas to boost classification scores.
- Use other knowledge bases to generate and provide additional context for the model.

References

- Kaushal Kumar Maurya, KV Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors. In Proceedings of NAACL 2025 (main).

Q3. Cross-Lingual and Cross-Cultural Word Embedding Alignment

Background

Cross-lingual and Cross-Cultural word embeddings are crucial for various multilingual NLP tasks. This assignment focuses on aligning monolingual word embeddings of 2 Indian Languages to create a shared cross-lingual embedding space. Additionally, creation of cross-cultural dictionaries of two cultures among your group members belonging to different cultures. Cross-cultural alignment refers to aligning cultural knowledge from different cultures. For e.g. Bihu from Assamese culture is similar to Ugadi from Andhra Pradesh culture. Here, the proxy of culture can be considered as a permanent resident state.

Assignment Question

Implement and evaluate a supervised cross-lingual word embedding alignment system for 2 Indian languages using the Procrustes method. And perform cross-cultural alignment on the created dataset.

Evaluation

a. Perform word translation from English to Hindi using the aligned embeddings.
b. Evaluate the accuracy of the translation using the MUSE test dictionary
c. Report Precision@1 and Precision@5 metrics for the word translation task
d. Compute and analyze cosine similarities between word pairs to assess cross-lingual semantic similarity

e. Conduct an ablation study to assess the impact of bilingual lexicon size on alignment quality. Experiment with different training dictionary sizes (e.g., 5k, 10k, 20k word pairs).

Resources

- MUSE dataset and pre-trained embeddings: https://github.com/facebookresearch/MUSE
- GATITOS dataset: https://huggingface.co/datasets/google/smol/blob/main/README.md
- FastText: https://fasttext.cc/
- Procrustes alignment method: "Word Translation Without Parallel Data" by Conneau et al. (2017)

Note: The quality of the cross-cultural dictionary created will be considered as a criterion for a grade.