REI603M - Assignment 2
Dataset Exploration and Analysis

Submission Instructions: Please submit this assignment via Canvas as a `.zip` file containing a PDF report, your presentation slides (either in PDF or PPT format) and code. Please name your submission file as

`REI603M_Assignment2_YourName_YourPartnerName.zip`

The deadline for online submission is at 18:00 on the 24th of January.

Introduction

This assignment involves a thorough exploration and analysis of a dataset relevant to your selected project. You should present information on the dataset in a concise report and your analysis in a clear, engaging presentation.

Submission Details

- Submit a report (max. 4 pages) and a presentation (max. 10 slides) on Canvas.
- You should submit your code for the analysis as a separate file.
- Peer evaluation according to the rubric below.
- Remember to include a disclaimer on LLM use!

Assignment Tasks

Dataset Analysis

- Acquire or create a dataset suitable for your project.
- Perform exploratory analysis and document your findings in a presentation.
- Study performance measurement of your task in case you are building something with zero-shot or in-context learning.

Report Content

In your report, you should include a datasheet covering your dataset, see instructions here. There is an example in the appendix of the linked document you can start to have a look at. Your report should not include your analysis as that will be a part of your presentation.

Presentation Content

Please refer to the rubric below for how your presentation will be scored (remember to address all items in the rubric). However, you may want to tailor your presentation according to the points below:

- Start by briefly reminding us what project you are working on.
- Introduce the dataset and its creation process.
- Explain the dataset's relevance to your project.
- Explain in sufficient detail how you plan to measure the performance of your system.
- Describe the type of ML task it supports (e.g., classification, regression).
- Discuss potential for further annotation.
- Present data format and summary statistics.
- Report state-of-the-art results on the dataset.
- **Make sure to include your names on the first slide of your presentation.**
- **Be mindful of the time limit (max. 7 minutes).**

In case you are working on a project that is based on zero-shot learning or in-context learning you might not need a dataset for training. However, you will always need data to measure the performance of your system. Focus on that data in your presentation. You might need to reflect deeply on how you will measure the performance of the system. Please look up information on performance metrics beforehand so you can refer to the appropriate metric(s) in your presentation.

Dataset Selection Criteria

- **Size:** Choose a manageable dataset that is neither too small nor too large.
- **Features:** Select a dataset with sufficient features for a robust analysis.
- **Relevance:** Ensure the dataset aligns with your project.
- If you are unsure if a specific dataset is suitable or not, please check with your instructor first.

Online Dataset Resources

The following websites might be helpful when looking for datasets.

- Huggingface datasets
- The Big Bad NLP Database
- Kaggle
- Google dataset search

Peer Evaluation

For peer evaluation, we will use the rubric below. Please note that the grade 5 is reserved for exceptional performance. Please refrain from using it unless you believe that the student(s) exceeded expectations by a large margin.

1. **Dataset Selection and Relevance (0-5 points):**

- **0 points:** No dataset selected or completely irrelevant to the project.
- **1-2 points:** Dataset selected is somewhat relevant, but lacks clear connection to the project goals.
- **3-4 points:** Dataset is appropriate for the project with a good explanation of its relevance.
- **5 points:** Dataset selection is thoroughly justified with deep insights into its relevance and potential impacts on the project.

2. **Dataset Exploration (0-5 points):**
   - **0 points:** No dataset exploration presented.
   - **1-2 points:** Dataset exploration is minimal, with little evidence of thorough analysis.
   - **3-4 points:** Dataset is explored with attention to detail, showing a good understanding of its characteristics and relevance to the project.
   - **5 points:** Exceptional dataset exploration, demonstrating deep insights, comprehensive analysis, and potential implications for the project.
3. **Methodology for Performance Measurement (0-5 points):**
   - **0 points:** No methodology for measuring performance or inappropriate metrics chosen.
   - **1-2 points:** Methodology is present but lacks clarity or proper alignment with the project goals and dataset.
   - **3-4 points:** Methodology is well-chosen with good justifications for the selected metrics and their relevance to the task at hand.
   - **5 points:** Methodology for performance measurement demonstrates exceptional understanding, with well-justified, innovative, and suitable metrics that enhance the project's value.
4. **Potential for Annotation and Further Work (0-5 points):**
   - **0 points:** No consideration of further annotation or improvement.
   - **1-2 points:** Some mention of annotation possibilities, but lacking in depth or feasibility.
   - **3-4 points:** Adequate discussion of potential for further annotation and enhancement, including realistic and practical suggestions.
   - **5 points:** Exceptional identification of potential for further work, including annotation and dataset improvements, with innovative and comprehensive strategies outlined.
5. **Presentation Quality (0-5 points):**
   - **0 points:** Inadequate presentation, lacking in clarity, and coherence.
   - **1-2 points:** Presentation is somewhat clear but lacks engagement or is poorly structured.
   - **3-4 points:** Presentation is well-structured, clear, and logically sequenced, with reasonable engagement.
   - **5 points:** Outstanding presentation: highly engaging, clear, and structured, making sophisticated use of visual aids and enhancing audience understanding.

Please keep the following in mind.

- When evaluating peers, provide honest and constructive feedback.
- Provide a numerical score for each of the criteria above.
- Where possible, include comments to justify your scoring and suggest areas for improvement.
- Be considerate and respectful in the tone and content of your feedback.
- Understand that the objective of peer review is to facilitate growth and learning, not just to assign a score.

The peer evaluation form will be distributed via Google Forms prior to the presentations. You will also use that form to grade your partner where you simply need to state how you split the workload. Your grade will be adjusted based on how you split the work. Both partners must report their workload split, and the resulting grades will be averaged.

- 40-60 and 50-50 splits result in the same grade for both participants.
- 20-80 to 30-70 splits result in the higher achiever improving their grade by 0.5 (if possible) and the lower achiever lowering their grade by 0.5.
- 10-90 splits mean that the higher achiever receives a 1 higher grade (if possible) and the lower achiever lowers their grade by 1.
- More imbalanced splits will mean that only one person receives a grade for the assignment.

Please accurately report the division of your workload. If you cannot work with your partner, then you can decide to split up. You are not stuck with each other.

Presentation Schedule

- In-class presentations: Thursday, January 25th.
- Submission deadline: 18:00 on January 24th.

Support

- For exploratory analysis examples, see this Kaggle notebook.
- For visualization best practices, watch this talk.
- Questions? Post on Ed or email `hafsteinne@hi.is`.