Exercise 1:

Q1. State the primary contributions of this work. (2 Marks)

- This work introduces a dataset labeled from 0 to 4 reflecting a balanced view of annotations of three field experts. The data points labeled from 2-4 are further annotated as PCL present. The work further aims to demonstrate the utility of the dataset for NLP models for both detecting presence of PCL and categorizing the specific type of PCL when it is present.

Q2. Identify the technical strengths that justify the paper's publication. (2 Marks)

- Relevant explanations and examples about PCL are given under " 3 Background on PCL." This section helps understand the direction of the labels.
- Annotation of the data points are well-thought and detailed in the work. Difficulties and disagreements faced are mentioned, and the way used to deal with these are detailed.
- The language is kept simple and understandable, while also adding relevant citations for further details.

Q3. Highlight the key weaknesses or areas where the authors failed to provide sufficient evidence. (2 Marks)

- While the abstract creates interest for the paper, its tone is a bit too informal and includes vague claims, such as "achieving the best result," where the notion of "best" is not well defined and may vary depending on the evaluation metric.
- The idea of proving usefulness of the dataset is directly relevant to the goals. However, the justification of the dataset is weakened by the lack of handling for the unbalanced classes during training. In particular, for the binary classification task, if a model predicts everything as non-PCL it achieves approximately 90.6% accuracy.

Overall, I believe the dataset created could be useful for future NLP research and applications, but a better structured reasoning and experiments are necessary. With the consideration of crucial details of the problem, such as unbalanced data and multi-label handling metrics, the justifications of the work could be improved. (Weak Accept)

Exercise 2:

Refer to the EDA.ipynb document in the repository.

Exercise 3:

Proposed Approach (2 marks)

To surpass the RoBERTa-base baseline, I propose a probability-level ensemble that combines:

- A fine-tuned RoBERTa-based transformer
- A classical machine learning pipeline with additional context features

The RoBERTa model available through Hugging Face is planned to be used and fine tuned for class imbalance using a WeightedTrainer.

For the ML pipeline, given categorical features (“keyword” and “country”) are planned to be incorporated through one-hot-encoding. Numerical features, such as verb noun ratio, punctuation ratios of exclamation point and question marks, and number of named entities are also planned to be added to the pipeline to enrich the context of the model without increasing the size too much. Then, the text is going to be converted into a lowercase, punctuation-free, stopword-filtered version suitable for further feature extraction of TF-IDF. With TF-IDF, we aim to find n-grams that represent text by emphasizing words that are frequent in a document but rare in the dataset to better understand the context and word choice differences. Moreover, due to the similar nature of the problem on sentiment analysis sentiment features are also planned to be explored. Given the large dimensionality resulting from these feature extraction steps, feature selection is applied to retain only the most impactful features while keeping computation at a manageable level.

Another important point we will try to take advantage of in the pipeline is the provided exact PCL categories of texts. We aim to do this by doing a multi-class classification with the highest weights to categories 1 and 5 and lowest to 2. Then, we will be using a model wrapper to binarize this using weights for binary classification tasks:

$$\text{positive} = 0.5 * \text{cat\_2\_prob} + \text{cat\_3\_prob} + 1.5 * \text{cat\_4\_prob}$$

$$\text{negative} = \text{cat\_1\_prob} + 1.5 * \text{cat\_0\_prob}$$

After each model is trained independently, we will ensemble them using:

$$\text{avg\_probs} = \text{hf\_ratio} * \text{hf\_probs} + (1 - \text{hf\_ratio}) * \text{pipeline\_probs}$$

$$\text{Decision} = (\text{avg\_probs} >= \text{threshold}).\text{astype(int)}$$

Where hf_ratio and threshold hyperparameters will be fine tuned by grid search using f1 score.

Rationale and Expected Outcome (2 marks)

RoBERTa-base excels at capturing deep contextual and semantic relationships, but it can struggle with class imbalance, rare but highly discriminative surface-level cues, and dataset-specific heuristics (punctuation, sentiment intensity, stylistic patterns, etc.). Conversely, classical ML models with engineered features are more interpretable, can emphasize explicit statistical patterns, and often perform well in edge cases where transformers are overconfident or underconfident. By ensembling the two, we aim to reduce variance and model-specific bias, exploit complementary error patterns, and improve minority-class PCL detection

We are aware that the ML-pipeline alone may not be powerful enough by itself due to the lack of contextual embeddings. Therefore, we aim to use this as part of the ensemble model. We expect the ensemble to achieve a consistent and measurable improvement in F1-score over the RoBERTa-base baseline.The weighted training strategy and probability-level ensemble are directed to improving recall for the minority positive class without a disproportionate loss in precision. The use of more interpretable engineered features and PCL category weights in the pipeline are expected to enhance robustness on borderline and ambiguous samples, leading to a more stable precision–recall trade-off across different data splits. Overall, the approach is expected to produce a more reliable and generalizable classifier than the RoBERTa-base baseline.

Exercise 5.2:

Local evaluation is done throughout the model construction steps. However, for a final local evaluation please refer to the local_evaluation.ipynb document in the repository.