Documentation

1. Dataset:

- `pavlick-formality-scores` - dataset consisting of 11,274 sentences collected from news, emails, blogs and forums. Annotated by humans on Amazon Mechanical Turk. Label is a value equal to average score given to the sentence from −3 to 3 with lower scores meaning less formal sentences

2. Models:

- `roberta-base-formality-ranker` - 125M parameters model, fine-tuned on GYAFC dataset, with RoBERTa base being the base model
- `deberta-large-formality-ranker` - 406M parameters model, also fine-tuned on GYAFC dataset, and the base model is DeBERTa large

3. Metrics:

- F1, Recall, Precision - the goal was to focus on the False Positive and False Negative predictions to analyze the behavior of the model on the sentences that are hard to label as some particular class, as I explained in more detail in report

4. Tools:

- `datasets` - Hugging Face library which I used to load dataset from Hugging Face
- `transformers` - another Hugging Face library that is needed to work with pre-trained transformer models
- `torch` - library that I used to perform operations with tensors while evaluation
- `sklearn` - library that I used in order to calculate evaluation metrics