Dataset Report 22

Mario Reda Matricola: 1937116

1 Dataset Description

The task I had to work on is Task 22, which primarily focuses on analyzing the gender of the author based on a collection of texts. The task is formulated as a binary classification task, where we have two genders, male and female. The objective is to predict the gender of the author of the texts based on the provided text contents. The dataset is a balanced dataset, with a gender distribution of 50/50. This ensures a fair representation of both genders in the dataset.

2 Dataset Format

This dataset is structured into two parts:

- Training data
- Test data

Each sample in the dataset must be represented as a JSON object, with the following fields:

- id: a unique identifier for the sample
- text: the text of the sample, which is the content provided by the author
- choices: a list of choices
- label: a label indicating the gender of the author associated with the sample, represented as an integer value

The dataset is divided into training and test data, with the labels of the test data present in the "gold" file.

3 Methodology for Dataset Reframing

In this task, I used regular expressions, which helped me extract relevant information from the text files, such as the document identifier, gender, and text. Subsequently, this data is organized into the expected JSON format as described earlier.

4 List of Suitable Prompts

- "Confronta il testo 'text' seguente e prova a individuare il genere dell'autore tra le categorie 'choices'."
- "Analizzando il testo 'text' fornito, cerca di capire il genere dell'autore tra le categorie 'choices'."
- "Considerando il seguente testo 'text', tenta di ipotizzare il genere dell'autore tra le categorie 'choices'."
- "Esplora il testo 'text' e cerca di determinare il genere dell'autore tra le categorie 'choices'."
- "Osserva attentamente il seguente testo 'text' e prova a individuare il genere dell'autore tra le categorie 'choices'."

5 Instructions for Running Code

To run the code, simply run the files named "creator test" and "creator training", which will subsequently generate all the required JSON files.