Dataset

We are solving the 2025 PAN shared task on Multi-Author Writing Style Analysis. The dataset consists of Reddit comments. The training set has 158280 labels, and we test on 33654 labels from another set.
There are three difficulty levels:

- **Easy**: The sentences in the document cover various topics, which allows for the use of topic information to detect authorship changes.
- **Medium**: The topic changes in the document are subtle, which forces the method to focus more on style to effectively solve the detection task.
- **Hard**: All sentences in the document involve the same topic.