Author Identification in Persian Literature using Language Models

Part 1: Dataset Creation

In the first part of this mini-project, your task is to create a dataset of at least 10 authors writing in the same genre in Persian literature. Include at least 30 documents per author and exactly 500 words per document in your dataset. You can use web scraping techniques and the Beautiful Soup library in Python for this purpose or create your dataset from different authors through their books or texts, etc. Ensure that your dataset is diverse and representative of the chosen genre. Include metadata such as author name, text content, and other relevant information.

Part 2: Author Identification Task using Large Language Models

For the second segment, concentrate on the author identification task using BERT architecture models available on Hugging Face. Specifically, you are required to fine-tune different BERT model for your specific task.

Detailed Report

Write a comprehensive report detailing your mini-project, primarily focusing on the insights gained and lessons learned. The report should be well-structured and hold more importance than the code itself. Here are the key components your report should cover:

- Introduction
  - Briefly introduce the problem of author identification in Persian literature.
  - Clearly state the objectives of your mini-project.

- Dataset Construction Techniques
  - Describe the process of dataset creation.
  - Provide details on the chosen genre, the number of authors, and the rationale behind your selection.
  - Discuss any challenges encountered during dataset construction and how they were addressed.
- Model Selection and Fine-Tuning
  - Explain your choice of model and the reasoning behind it.
  - Detail the fine-tuning process, including modifications made to the model.
  - Discuss the architecture and parameters chosen.
- Experiments and Results with 5-Fold Cross Validation
  - Present the results of your author identification experiments using 5-fold cross-validation.
  - Include performance metrics for different approaches including accuracy, F1 Score, percision and recall.
  - Include confusion matrix.
  - Analyze how changes in fine-tuning parameters (learning rate) impacted the model's performance.
  - Analyze how omitting the stopwords can impacted the model's performance..
  - Analyze how document length can impact your model's performance.
  - Include any other interesting experiment that is intuitive to this report.
- Comparison with Traditional ML Approaches
  - If applicable, compare the performance of your model with traditional machine learning methods.
  - Discuss the advantages and limitations of each approach.
- Conclusion and Future Work
  - Summarize your findings and the overall success of your mini-project.
  - Suggest potential areas for future improvement or expansion.

Ensure your report adheres to proper academic writing standards and includes citations where necessary. The report's clarity, depth of analysis, and coherence will be essential in evaluating your work.

Remember, the quality of your report holds significant weight in the assessment, so dedicate ample time to its development. Good luck!