Research Article Classification

Objective

Classify research articles into predefined categories based on their content, such as biology, chemistry, physics, computer science, and social sciences.

Dataset

Use public arxiv dataset -
https://www.kaggle.com/datasets/Cornell-University/arxiv
Use abstracts for classification and subject categories for class association.

Instructions:

1. Perform exploratory data analysis to understand the structure and characteristics of the dataset.
2. Preprocess the data - clean the text, remove special characters, lowercase the words, remove stop words, perform stemming or lemmatization, and vectorize the text.
3. Train a suitable NLP model for text classification (use hugging face for that - https://huggingface.co/models?library=tf&language=en&license=license:apache-2.0&sort=downloads&search=bert-base)
4. Evaluate the model's performance on the validation set using appropriate metrics such as accuracy, precision, recall, and F1-score.
5. Perform hyperparameter tuning or model selection to improve the model's performance.
6. Implement a simple API interface to input research article abstracts and return the predicted categories using Django rest framework.

Suitable libraries for the tasks could be scikit-learn, nltk, TensorFlow, Keras and Spacy. When choosing models please use TensorFlow or Keras based.