Assignment

Problem Statement:

Given an abstract of a paper, the objective of this task is to classify it into one of the seven predefined domains. The predefined domains are:

1. Computation and Language (CL)
2. Cryptography and Security (CR)
3. Distributed and Cluster Computing (DC)
4. Data Structures and Algorithms (DS)
5. Logic in Computer Science (LO)
6. Networking and Internet Architecture (NI)
7. Software Engineering (SE)

Approach:

Data Preprocessing

1. **Loading Data**:
    - The training and test datasets are loaded using pandas. The datasets consist of abstracts of papers and their corresponding target domains.
2. **Text Preprocessing**:
    - Text preprocessing includes converting text to lowercase, removing non-alphanumeric characters, single characters, and extra spaces. Additionally, stopwords are removed, and words are lemmatized using NLTK's WordNetLemmatizer.
3. **Encoding Targets**:
    - The target domains are encoded into numerical labels using LabelEncoder.
4. **Handling Class Imbalance**:
    - SMOTE (Synthetic Minority Over-sampling Technique) is used to handle class imbalance in the training data.

Feature Extraction

- **TF-IDF Vectorization**:
    - The preprocessed abstracts are transformed into TF-IDF vectors with a maximum feature limit of 5000.

Model Training and Evaluation

1. **Logistic Regression**:
    - A logistic regression model is trained on the resampled training data and evaluated on the test data.
2. **Support Vector Machine (SVM)**:

    - An SVM model with a linear kernel is trained and evaluated similarly.
3. **XGBoost**:
    - An XGBoost model is trained and evaluated.
4. **LSTM**:
    - An LSTM model is built using Keras. The abstracts are tokenized, padded, and converted to sequences. The LSTM model is then trained and evaluated.

Model Performance

- The performance of each model is evaluated using precision, recall, and F1-score metrics for each class. Confusion matrices are also plotted to visualize the performance of the models.

Cross-Validation and Learning Curves

- Cross-validation is performed for logistic regression to obtain a reliable estimate of the model's performance. Learning curves are plotted to understand the model's learning behavior.

Assumptions

1. **Text Processing**:
    - Assumed that basic text preprocessing steps (lowercasing, removing non-alphanumeric characters, etc.) would suffice for cleaning the abstracts.
2. **Feature Extraction**:
    - TF-IDF vectorization with 5000 features is assumed to be sufficient for capturing the significant terms in the abstracts.
3. **Model Selection**:
    - Chosen models (Logistic Regression, SVM, XGBoost, LSTM) based on their suitability for text classification tasks and availability of resources.
4. **Handling Class Imbalance**:
    - SMOTE is assumed to be effective in addressing class imbalance in the dataset.

Future Scope

1. **Hyperparameter Tuning**:
    - Perform hyperparameter tuning for all models using techniques like Grid Search or Random Search to potentially improve performance.
2. **Advanced Text Processing**:
    - Explore advanced text processing techniques such as n-grams, part-of-speech tagging, and named entity recognition to enhance feature extraction.
3. **Deep Learning Models**:
    - Experiment with more advanced deep learning models like BERT or transformer-based models to potentially improve classification accuracy.
4. **Ensemble Methods**:

    - Explore ensemble methods to combine the strengths of different models for better performance.
5. **More Data**:
    - Collect more data to improve the robustness and generalization of the models.

Evaluation Metrics

- The evaluation is based on the class-wise weighted F1 scores on the test data. The scores for each model are as follows:

1. **Logistic Regression**:
    - Accuracy: 0.91
    - Weighted F1 Score: 0.91
2. **Support Vector Machine (SVM)**:
    - Accuracy: 0.91
    - Weighted F1 Score: 0.91
3. **XGBoost**:
    - Accuracy: 0.90
    - Weighted F1 Score: 0.90
4. **LSTM**:
    - Accuracy: 0.88
    - Weighted F1 Score: 0.88

The best-performing model based on the weighted F1 score is the Logistic Regression model.