Checkpoint 2: Model Training for Fake Review Detection

What is Model Training?

Model training is the process of feeding preprocessed data into a machine learning algorithm to enable it to learn patterns and make predictions. This stage involves selecting a suitable algorithm, splitting data into training and testing sets, training the model, evaluating its performance, and fine-tuning hyperparameters.

Basic Components of Model Training:

1. **Dataset Preparation:**
   - Load the preprocessed dataset created during Checkpoint 1.
   - Split the dataset into training and testing sets to assess the model's performance.
2. **Model Selection:**
   - Experiment with different classifiers such as Random Forest, Support Vector Machine (SVM), and Logistic Regression to identify the best-performing model for the given data.
3. **Pipeline Creation:**
   - Construct pipelines to streamline the process of vectorization, transformation, and model application.
4. **Model Training:**
   - Train each model using the training dataset.
5. **Model Evaluation:**
   - Use the testing dataset to evaluate the model's performance based on accuracy, precision, recall, and F1 score.
   - Compare the results of different models to choose the optimal one.

6. **Model Serialization:**
    - Save the trained models using serialization techniques (e.g., joblib) for future use.

Tasks:

1. Load the dataset created in Checkpoint 1 and inspect its structure.
2. Split the dataset into training and testing sets.
3. Train models such as Random Forest, SVM, and Logistic Regression using pipelines for preprocessing and classification.
4. Evaluate each model's performance using metrics such as accuracy, precision, recall, and F1 score and identify the best-performing model.
5. Save each trained model using joblib with appropriate filenames (e.g., random_forest_model.pkl, svc_model.pkl, logistic_regression_model.pkl).
6. Perform test predictions on sample data to validate the functionality of saved models.
7. Upload the code and the pkl files to GitHub with the repository titled **"Project_WoC_7.0_Fake_Review_Detection"** and inside it create folder for "**checkpoint 2"**.

Deadline:

19th January, 2025, 11:59 PM

Deliverables:

1. GitHub repository with:
    - Model training code.
    - README file describing the model training process, evaluation results, and saved models.
    - Serialized model files.

We've intentionally kept the task manageable, allowing you to enjoy the festival of Kites🪁, Uttarayan, while still making progress. Remember, learning should be both enjoyable and rewarding