Age, Gender, and Personality Prediction Model Documentation

Table of Contents

1. Overview

The provided code focuses on predicting age interval, gender, and personality scores using text data, including the utilization of DistilBERT for certain tasks. It involves the use of various models, including Support Vector Machines (SVM), XGBoost, and Neural Networks. Additionally, LIWC features are experimented with to predict personality scores using different machine learning models.

2. Usage

2.1 Dependencies

Ensure that the necessary Python libraries and frameworks are installed. Key dependencies include:

- **numpy**
- **pandas**
- **scikit-learn**
- **joblib**
- **xml.etree.ElementTree**
- **os**

- **tensorflow**
- **transformers** (Hugging Face)
- **xgboost**
- **matplotlib**
- **seaborn**
- Custom **LIWC** library or feature extraction tool

2.2 Data Input

- **Profile Data**: The code expects input profile data in CSV format containing user information, including user ID, text data, and personality scores. The profile data is assumed to have columns like "userid," "text_data," "age," "gender," etc.
- **LIWC Features (for Experiments)**: LIWC features should be available in CSV format, containing linguistic features extracted from text data.

2.3 Command-line Arguments

The code for age, gender, and personality prediction expects two input paths provided as command-line arguments:

bashCopy code

```
Usage: tcss555 –input <input1> --output <input2)
```

- **<input1>**: Path to the input profile data CSV file.
- **<input2>**: Path to the output directory where XML files will be saved.

3. Models and Training

3.1 Personality Prediction

- **Models Used**: SVM (Support Vector Machines), XGBoost, Neural Networks
- **Data Preprocessing**: TF-IDF vectorization of text data
- **Personality Traits Predicted**: Openness, Neuroticism, Extraversion, Agreeableness, Conscientiousness
- **Output Format**: XML files with predicted personality scores and other user information

3.2 Age Classification

- **Model Used**: SVM (Support Vector Machines)
- **Age Groups**: "xx-24", "25-34", "35-49", "50-xx"
- **Data Preprocessing**: TF-IDF vectorization of text data

- **Output Format**: XML files with predicted age group and other user information

3.3 Gender Prediction

- **Model Used**: SVM (Support Vector Machines)
- **Gender Classes**: "male," "female"
- **Data Preprocessing**: TF-IDF vectorization of text data
- **Output Format**: XML files with predicted gender and other user information

3.4 DistilBERT Usage

- **Model Used**: DistilBERT
- **Task**: Unseen test data classification
- **Data Preprocessing**: Tokenization, padding, and attention mask creation
- **Output Format**: Predicted labels

4. LIWC Feature Experiments

4.1 Experiment Overview

- **Objective**: Predict personality scores using only LIWC features.
- **Models Used**: Linear Regression, ElasticNet, Ridge, Lasso, SVR, Random Forest, XGBoost
- **Data Preprocessing**: Standard scaling of LIWC features
- **Output Format**: Model-specific files (e.g., **.pkl**) saved for each trained model

4.2 Models Used

- Linear Regression
- ElasticNet
- Ridge
- Lasso
- SVR (Support Vector Regressor)
- Random Forest
- XGBoost

5. Conclusion

Both sets of code contribute to predicting personality scores, age groups, and gender based on textual data and LIWC features. The documentation provides insights into the models used, dependencies, and the overall structure of the code, including the integration of DistilBERT for certain tasks. Users are encouraged to customize and extend the documentation based on their specific project requirements and additional details.

Feel free to reach out if further clarification or customization is needed.