Chinese Stereotype Detection: HEARTS Framework Adaptation

📋 Project Overview

This project adapts the HEARTS framework for stereotype detection from English to Chinese contexts. We construct a culturally-adapted Chinese dataset and fine-tune Chinese pretrained language models (RoBERTa-wwm-ext and MacBERT) to detect stereotypes across multiple social dimensions.

🎯 Key Features

- **Culturally-Adapted Dataset**: 4,000 Chinese samples across 6 dimensions (Gender, Profession, Nationality, Region, Education, Age)
- **Dual Construction Strategy**:
  - EMGSD translation for universal dimensions
  - LLM-based data augmentation for Chinese-specific dimensions
- **State-of-the-art Models**: Fine-tuned RoBERTa and MacBERT for Chinese stereotype detection
- **Explainability Analysis**: SHAP and LIME interpretations for model predictions

📊 Dataset Composition

Dimension	Target Size	Construction Strategy
Gender	800	✅ English → Chinese translation from EMGSD

Dimension	Target Size	Construction Strategy
Profession	2000	✅ English → Chinese translation from EMGSD
Nationality	400	✅ English → Chinese translation from EMGSD
Region	1000	✅ LLM-based data augmentation from manual seeds
Age	400	✅ LLM-based data augmentation from manual seeds
Education	600	✅ LLM-based data augmentation from manual seeds
Total	**4000**	✅ Mixed translated + LLM-constructed

🗂️ Project Structure

```
Chinese_model/
├── model_outputs/              # RoBERTa model checkpoints
│   ├── checkpoint-488/
│   └── checkpoint-732/
├── model_outputs_macbert/      # MacBERT model checkpoints
│   ├── checkpoint-244/
│   └── checkpoint-732/
├── train_dev_test/             # Train/Dev/Test splits
│   ├── train.json
│   ├── dev.json
│   └── test.json
├── Data/                       # Dataset files
│   ├── emgsd_selected_en_2000.csv      # Original English
EMGSD subset
│   ├── emgsd_selected_zh_2000.csv      # Chinese translation
│   ├── final_emgsd_zh.csv              # Final combined
dataset (CSV)
```

```
|   ├── final_emgsd_zh.json              # Final combined
dataset (JSON)
|   ├── generated_age.csv                # LLM-generated age
stereotypes
|   ├── generated_education.csv          # LLM-generated
education stereotypes
|   ├── generated_region.csv             # LLM-generated region
stereotypes
|   ├── llm_seeds_zh.json                # Manual seeds for LLM
generation
|   ├── data_create.ipynb                # Dataset construction
pipeline
|   └── data_process.ipynb               # Data preprocessing
scripts
├── train_model.ipynb            # Model training notebook
├── SHAP-LIME.ipynb             # Explainability analysis
└── macbert_shap_lime_bar.png   # Visualization output
```

🚀 Quick Start

1. Environment Setup

Install dependencies:

```
pip install -r requirements.txt
```

Or install manually:

```
pip install torch transformers datasets scikit-learn shap lime
matplotlib numpy pandas jupyter
```

Requirements:

- Python >= 3.8
- CUDA (optional, for GPU acceleration)

2. Dataset Construction

Run the data construction pipeline:

```
jupyter notebook Data/data_create.ipynb
```

This will:

- Translate EMGSD samples to Chinese
- Generate Chinese-specific stereotypes using LLM
- Create train/dev/test splits (70%/15%/15%)

3. Model Training

Train RoBERTa or MacBERT models:

```
jupyter notebook train_model.ipynb
```

Training Configuration:

- **Models**: `hfl/chinese-roberta-wwm-ext` or `hfl/chinese-macbert-base`
- **Optimizer**: AdamW (lr=2e-5)
- **Loss Function**: Cross-Entropy Loss
- **Batch Size**: 8
- **Epochs**: 3
- **Max Length**: 128 tokens

4. Model Evaluation

The training script automatically evaluates on the test set and provides:

- Overall accuracy and Macro F1-Score
- Dimension-wise performance breakdown
- Classification report

5. Explainability Analysis

Run SHAP and LIME interpretations:

```
jupyter notebook SHAP-LIME.ipynb
```

This generates visual explanations showing which tokens contribute most to stereotype predictions.

📈 Results

Model Performance

Model	Accuracy	Macro F1-Score
RoBERTa (pretrained)	0.3785	0.2746
MacBERT (pretrained)	0.4064	0.3354
RoBERTa (fine-tuned)	0.7131	0.7045
MacBERT (fine-tuned)	**0.7291**	**0.7243**

Baseline Comparison

- **Original ALBERT-V2** (English EMGSD): 81.50% Macro F1
- **Our Replication** (ALBERT-V2): 86.45% Macro F1

🔍 Explainability

We provide model interpretations using:

- **SHAP** (SHapley Additive exPlanations): Global feature importance
- **LIME** (Local Interpretable Model-agnostic Explanations): Local instance-level explanations

Example visualization saved in `macbert_shap_lime_bar.png`

🌍 SDG Alignment

This project supports:

- **SDG 5**: Gender Equality
- **SDG 10**: Reduced Inequalities
- **SDG 16**: Peace, Justice and Strong Institutions

By detecting language-embedded stereotypes in Chinese contexts, this work provides a diagnostic tool for bias monitoring and mitigation.

⚠️ Limitations

- **Artificial Bias**: LLM-generated data may introduce artificial biases
- **Generalization Gap**: Benchmark performance doesn't guarantee real-world fairness
- **Misuse Potential**: Automated detection risks being exploited for surveillance without proper governance

📧 Contact

- **Author**: Qingqing Liu
- **Email**: zczqq26@ucl.ac.uk
- **Institution**: University College London, MSc AI for Sustainable Development