HEARTS Localisation to Chinese Gender Stereotype Detection (Binary Classification)

Coursework: Replication + Local contextual adaptation (China) using the HEARTS methodology

Task: Detect whether a Chinese sentence contains a **gender stereotype** (`stereotype=1`) or **non-stereotype** (`non-stereotype=0`)

Main metrics: Accuracy, Macro-F1 (with Bootstrap 95% CI)

Statistical testing: McNemar (paired significance)

Extra analysis: Training data size ablation + failure case analysis

Compute note: Experiments run on MacBook Air (M2, 2022), 8GB RAM, macOS Sequoia 15.3.2 (per system screenshot)

Abstract

This project replicates the baseline AI methodology described in HEARTS and adapts it to a Chinese local context: sentence-level gender stereotype detection as a binary classification problem. We (1) replicate a baseline BERT pipeline on the open stereotype dataset used in the HEARTS ecosystem (MGSD/EMGSD family) and verify reproducibility against reported metrics, then (2) curate a Chinese dataset derived from CORGI-PM resources, fix a deterministic preprocessing pipeline and train/validate/test split, and (3) evaluate three models: a TF-IDF + Logistic Regression baseline, a Chinese BERT fine-tuning model, and a stronger Chinese transformer baseline (MacBERT). On the Chinese test set (balanced), TF-IDF performs poorly (Macro-F1 $\approx$ 0.496), while Chinese BERT achieves strong performance (Macro-F1 $\approx$ 0.833). MacBERT provides a small improvement (Macro-F1 $\approx$ 0.840), but paired significance tests show this gain is not statistically significant. Ablations across training set sizes show diminishing returns beyond $\sim$1k samples. Failure case analysis reveals that most errors come from **lexical gender cue over-triggering** (false positives) and **implicit stereotype framing** (false negatives). We discuss ethics, limitations, sustainability/scalability trade-offs, and align the local challenge with UN SDGs, especially **SDG 5 (Gender Equality)**, **SDG 10 (Reduced Inequalities)**, and compute-cost considerations linked to **SDG 12/13**.

1 Introduction: Local Context + SDG Alignment

Gender stereotypes in Chinese text appear across social media, news reporting, advertisements, and recommendation systems. Automatic detection can support:

- safer content moderation and reporting workflows,
- bias monitoring for recommender systems,
- bias-aware auditing for NLP products used in education/employment contexts.

SDG alignment

- **SDG 5 (Gender Equality):** stereotype identification supports efforts to reduce discrimination and harmful norms in public discourse.
- **SDG 10 (Reduced Inequalities):** stereotype detection contributes to reducing discriminatory social outcomes and supports inclusive policies and systems.
- **SDG 12 & SDG 13 (Responsible consumption/production & Climate action):** model choice and training strategy matter—lighter baselines and diminishing-return ablations inform responsible compute use.

2 Background: HEARTS Methodology and the Original Study

HEARTS proposes a holistic framing for stereotype detection: **Explainability, Sustainability, and Robustness**, using curated stereotype datasets and baseline transformer models to quantify performance and trade-offs. The paper reports strong binary detection baselines (e.g., transformer fine-tuning) and emphasizes that stereotype detection is culturally dependent and benefits from careful evaluation and analysis.

Key implication for localisation: because stereotypes are **context-dependent**, transferring a pipeline into Chinese requires (i) a suitable Chinese dataset, (ii) careful preprocessing/splitting, (iii) Chinese-appropriate transformer backbones (e.g., MacBERT), and (iv) clear reporting on limitations and ethics.

3 Part A1 —Baseline Replication (Open Dataset)

3.1 Goal

Replicate the baseline AI methodology using an open dataset in the HEARTS ecosystem (MGSD/EMGSD family), using a BERT-style sentence classifier, and verify that results are within reasonable tolerance of reported baselines.

3.2 Implementation Summary

- **Code:** `baseline/BERT_MGSD_baseline.py` (provided in submission package)
- **Approach:** standard transformer fine-tuning for binary classification
- **Output evidence:** script + exported classification report and saved full results CSV (for auditability)

(Your replication evidence is in the submitted code + produced metrics artifacts; the report focuses on the reproducible process and alignment with HEARTS baseline methodology.)

4 Part A2–A3 —Local Challenge + Alternative Dataset (Chinese)

4.1 Local challenge definition

Problem: Chinese gender stereotype detection (binary).

Labels: `stereotype` vs `non-stereotype` mapped to 1/0.

Group column: `gender` retained for schema consistency; fairness metrics are not the primary analysis.

4.2 Dataset source: CORGI-PM resources

We derive our Chinese dataset from CORGI-PM, a Chinese corpus designed for gender bias probing/mitigation, introduced to address the scarcity of high-quality Chinese gender bias resources.

4.3 Curated balanced pool and dataset size constraint

We build a **balanced pool of 5,000 Chinese sentences** (2,500 per class), then create fixed splits and enforce a maximum training size of **3,500** after balancing (1:1).

Rationale: ensures stable evaluation, controls label imbalance, and makes ablation results interpretable.

Note (quality caveat): because this is a curated subset rather than the full CORGI-PM corpus, results reflect this specific sampling and should not be over-generalised to all Chinese stereotype phenomena.

5 Preprocessing and Fixed Splits (Deterministic, Reproducible)

5.1 Preprocessing script (core contribution)

We implement a minimal, fully reproducible preprocessing pipeline in `local_context/preprocessing.py`:

1. Load raw CSV: `local_context/data/chinese_stereotypes.csv`
2. Clean text: cast to string + `strip()` (no aggressive normalization)
3. Encode labels:
   - `stereotype` → 1
   - `non-stereotype` → 0
4. Stratified split with fixed seed `random_state=42`:
   - Train 70%
   - Val 15%
   - Test 15%
5. Save to `local_context/data/processed/{train,val,test}.csv`

Why minimal cleaning? We intentionally avoid heavy language-specific heuristics so failure modes remain visible and comparable across models.

5.2 Resulting split sizes (from fixed pool)

Because the pool is 5,000 balanced samples, the deterministic split yields:

- **Train:** 3,500
- **Val:** 750
- **Test:** 750

6 Models and Training Pipeline (Local Adaptation)

6.1 Baselines

1. **TF-IDF + Logistic Regression** (weak baseline)
   - **Script:** `local_context/baseline_tfidf_lr.py`

- **Motivation:** cheap, interpretable lexical baseline.

2. **Chinese BERT fine-tuning** (main adapted model)
   - **Script:** `local_context/train_zh_bert.py`
   - **Model:** Chinese BERT/transformer backbone fine-tuned for binary classification.
3. **MacBERT fine-tuning** (stronger baseline)
   - **Script:** `local_context/train_zh_macbert.py`
   - **Model:** `hfl/chinese-macbert-base` (Chinese-optimized pretraining)

6.2 Hyperparameter tuning

We tune the transformer models with a small grid:

- epochs $\in \{3, 5\}$
- lr $\in \{2e{-}5, 3e{-}5\}$

Selected configuration for main comparisons: `epochs=3, lr=2e-5`

Reason: best overall validation behavior and stable test performance, then reused for ablation + significance testing (to keep comparisons fair).

7 Evaluation Protocol

7.1 Metrics (main)

- **Accuracy**
- **Macro-F1** (primary, robust under class balance)
- Also report precision/recall per class via confusion matrices.

7.2 Statistical testing

- **McNemar test** for paired significance on the same test set.
- **Bootstrap 95% CI** for Macro-F1 to quantify uncertainty.

7.3 Ablation: Training data size

We evaluate performance with stratified 1:1 subsets of the training set:

- sizes: 500 / 1000 / 2000 / 3000 / 3500
- seeds: 42 / 43 / 44 (controls subset sampling; training seed fixed)

Command used (as provided):

```
python local_context/ablation_data_size.py \
  --sizes 500,1000,2000,3000,3500 \
  --seeds 42,43,44 \
  --epochs 3 \
  --lr 2e-5
```

8 Results: Baseline Comparison (Chinese Local Dataset)

8.1 Main test performance (test n=750, balanced)

From the saved full-results evaluation summary (confusion matrices and derived metrics):

Model	Accuracy	Macro-F1	Predicted stereotype rate
TF-IDF + LR	0.561	0.497	0.859
Chinese BERT	0.833	0.833	0.504
MacBERT	0.840	0.840	0.551

Interpretation

- TF-IDF strongly over-predicts the positive class (≈86% predicted as stereotype), yielding poor Macro-F1 because class-0 recall collapses.
- Chinese BERT yields a large, meaningful jump in Macro-F1 (∼0.83).
- MacBERT is slightly higher, but the gain is small (see significance + CI below).

8.2 Confusion matrices (Chinese test set)

- **BERT:** TN=311, FP=64, FN=61, TP=314
- **MacBERT:** TN=296, FP=79, FN=41, TP=334

- **TF-IDF:** TN=76, FP=299, FN=30, TP=345

Notably, TF-IDF has **299 false positives**, consistent with lexical over-triggering.

9 Uncertainty and Significance

9.1 Bootstrap 95% CI (Macro-F1) —reported from `bootstrap_macro_f1.csv`

- **TF-IDF:** mean 0.49588, CI [0.46239, 0.52891]
- **BERT:** mean 0.83388, CI [0.80769, 0.86125]
- **MacBERT:** mean 0.83999, CI [0.81431, 0.86870]

Key takeaways

- TF-IDF vs BERT: CIs do **not** overlap → very strong improvement.
- BERT vs MacBERT: CIs overlap heavily → improvement is marginal.

9.2 McNemar test —reported from `mcnemar_results.csv`

TF-IDF vs BERT

- TF-IDF wrong / BERT correct: 266
- TF-IDF correct / BERT wrong: 62
- statistic ≈ 125.637, p ≈ 0 → significant ($p<0.05$)

BERT vs MacBERT

- BERT wrong / MacBERT correct: 31
- BERT correct / MacBERT wrong: 26
- statistic ≈ 0.2807, p ≈ 0.596 → not significant

Conclusion Chinese BERT substantially and significantly improves over TF-IDF; MacBERT's gain over BERT is small and not statistically significant under this test set.

10 Ablation: Training Data Size (Must-report Result)

From `ablation_data_size_agg.csv` (mean ± std over subset sampling seeds 42/43/44):

Train size	Accuracy (mean ± std)	Macro-F1 (mean ± std)
500	0.804 ± 0.02195	0.80377 ± 0.02215
1000	0.82222 ± 0.00505	0.82196 ± 0.00530
2000	0.82400 ± 0.01041	0.82304 ± 0.01074
3000	0.83067 ± 0.02771	0.82952 ± 0.02910
3500	0.83111 ± 0.01602	0.82977 ± 0.01784

Interpretation (important narrative)

- Performance improves as data increases, but the **marginal gains decay after ∼1,000 examples** (plateau behavior).
- The variance mainly reflects **which examples are selected** (boundary/ambiguous samples), not training randomness (training seed fixed).

11 Failure Case Analysis (10–15 examples + storytelling)

We analyze 15 representative misclassified test cases and categorize them into 3 main buckets. (All examples below are verified as misclassified by checking the saved full-results predictions used to construct the failure case table.)

11.1 Category A —Lexical gender cues trigger false positives (non-stereotype → predicted stereotype)

Most models over-rely on gender-coded words (“女/男/母亲/阿姨/老公/美女...”) even when the sentence is neutral description rather than a stereotype claim.

Examples (true=0, BERT=1, MacBERT=1, TF-IDF often=1):

- “放假了，平时在派出所里做饭的阿姨回家了。”
- “孩子牙口好不好，还与母亲的最高学历有很大关系。”
- “好几次法院集中执行，女法官都不忍心对两老采取强制腾退。”
- “女演员愣住了一会儿，眼泪便哗啦哗啦地掉下。”
- “几天之后，我又可以活蹦乱跳地玩耍了，但妈妈却因为过度劳累生病了。”

Why this happens: These are *gender-mention* sentences, but not *stereotype-generalization* sentences. Models (especially TF-IDF) conflate "mentions gender" with "expresses stereotype".

Practical implication: For deployment, false positives may over-flag benign content involving women/family roles, which can introduce secondary harms (over-moderation of women-related narratives). Human-in-the-loop review is recommended for borderline cases.

11.2 Category B —Implicit stereotype framing (stereotype → predicted non-stereotype)

Some stereotypes are expressed indirectly via framing, presuppositions, or institutional context. BERT sometimes misses these when no obvious stereotype keywords exist.

Verified cases (true=1, BERT=0):

- "平台还通过综合算法模式，为女性用户寻找更合适的同路人。" (BERT=0, MacBERT=1, TF-IDF=1)
- "哎，妈妈忘了给自己做心理建设了！" (BERT=0, MacBERT=1, TF-IDF=1)

Interpretation: These may require recognizing a broader stereotype implication (e.g., assumptions about women's preferences/needs or socially loaded framing). MacBERT appears slightly more sensitive to these patterns, consistent with its pretraining adaptations for Chinese.

11.3 Category C —Annotation subjectivity / borderline semantics

Stereotype labeling is inherently subjective and context-dependent (also emphasized by HEARTS).

Some sentences are near the boundary between stereotype and narrative description; model disagreement is expected.

Example:

- "账号中近 90 条视频记录了 45 岁的男校长带着孩子们做各种美食..." (true=1, BERT=0, MacBERT=1)

Actionable improvement: Future dataset improvements could add guideline clarifications or multi-annotator disagreement flags for borderline items.

11.4 "Story highlight": MacBERT fixes BERT (2–3 cases)

We include three narrative-friendly highlights where **BERT makes an error but MacBERT is correct**:

1. true=0: "我觉得，挤挤挨挨一大架子书，怎么也得有位女神坐镇啊。"
   - BERT predicts stereotype (over-trigger), MacBERT predicts non-stereotype.
2. true=0: "这个孩子跟母亲在一起的时间要多一些... 她是跟母亲睡的..."
   - BERT predicts stereotype; MacBERT correctly treats it as descriptive family narrative.
3. true=1: "藏枪坤包看似与普通女包无异... 价签上写着：携带、冷静..."
   - BERT misses; MacBERT detects stereotype (or stereotype-coded framing) correctly.

These cases support the result pattern: MacBERT's gains exist but are narrow and concentrated in certain linguistic contexts.

12 Ethics, Risks, and Responsible Use

1. **Harmful content exposure:** stereotype datasets necessarily contain sensitive/discriminatory language. We restrict usage to research and evaluation; avoid generating new stereotypes.
2. **False positive harm:** over-flagging women-related content can itself be discriminatory (Category A). Deployments should include thresholds, calibration, and human review.
3. **Synthetic data caution:** an early attempt using LLM-generated samples produced suspicious near-perfect accuracy (likely memorization artifacts and distribution mismatch). We therefore do not rely on synthetic data for core claims.
4. **Dataset limitations:** subset sampling, potential duplicates or near-duplicates, and cultural subjectivity mean results should be interpreted as **contextual evidence**, not a universal detector of all Chinese gender stereotypes.

13 Scalability and Sustainability (HEARTS "S" in local context)

Transformers are compute-heavy relative to bag-of-words baselines, so sustainability matters.

- **TF-IDF baseline:** minimal compute and fast iteration, but poor Macro-F1 (≈0.50).

- **Transformer fine-tuning:** substantial performance gain (Macro-F1 ≈0.83), but higher compute cost.

Diminishing returns informs sustainable choices: Ablation shows that after ∼1k samples, performance improves only slightly with more data. This provides a practical rule: **stop earlier or prioritize data quality/coverage rather than scaling training size blindly** —supporting responsible compute use aligned with **SDG 12 and SDG 13** (resource efficiency, emission reduction).

Measurement limitation: In this environment, CodeCarbon did not reliably return stable CO_2 estimates across runs, so we report sustainability primarily as **qualitative trade-offs and data-efficiency evidence** instead of unreliable absolute emission numbers.

14 Comparison to Original Study (Replicate vs Localise)

- The original HEARTS work reports strong transformer baselines for stereotype detection and emphasizes explainability, robustness, and sustainability trade-offs.
- Our MGSD replication confirms a comparable pipeline can be reproduced in code and metrics (within acceptable tolerance for a reproduction exercise).
- Our localisation demonstrates that the same methodology transfers to Chinese with:
  - a Chinese dataset derived from CORGI-PM resources (Chinese context),
  - Chinese-appropriate backbones (MacBERT),
  - statistical significance testing and uncertainty estimation.

15 Limitations and Future Work

1. **Subset representativeness:** curated pool (5k) may not cover all Chinese stereotype varieties.
2. **Potential duplicates/near-duplicates:** no explicit deduplication was performed; future work should add de-duplication and leakage checks.
3. **Explainability (HEARTS "E"):** we did not implement SHAP/LIME token attribution analysis; adding explainability would better align with HEARTS' holistic goals.
4. **Robustness (HEARTS "R"):** further stress tests could include domain shifts (news vs social media), adversarial paraphrases, or counterfactual gender swaps.

5. **Fairness metrics:** although `gender` is preserved for schema, we do not present subgroup fairness as the main claim; future work can add subgroup evaluation.

16 Conclusion

We successfully replicate a HEARTS-style transformer baseline and localise it to a Chinese gender stereotype detection task using CORGI-PM-derived data. The adapted Chinese BERT model strongly outperforms TF-IDF, with both bootstrap CIs and McNemar test confirming the gain. MacBERT yields a small but non-significant improvement over BERT, and failure analysis shows that errors are dominated by lexical gender cue over-triggering and implicit framing. Ablations indicate diminishing returns beyond $\sim$1,000 training examples, supporting sustainable modelling choices. The work supports SDG-aligned goals (gender equality and reduced discrimination) while explicitly acknowledging compute and evaluation limitations.