UK-HEARTS: Culturally-grounded stereotype detection

Refining EMGSD for detection of stereotypes in the UK context

Grace Y. Lin

Original Paper

The original paper that this research is based on introduces a framework for detecting stereotypes in text, with a focus on sustainability, performance, and transparency. King et al. [1] expanded on the data in MGSD (an existing stereotype dataset) [2], by transforming and appending data from WinoQueer [3] and SeeGULL [4], to better capture LGBTQ+ and nationality stereotypes. The combined dataset was then used to train and compare different models in the task of detecting bias, with a fine-tuned ALBERT-V2 [5] architecture chosen as the final model due to its comparable performance to larger models with much larger carbon footprints.

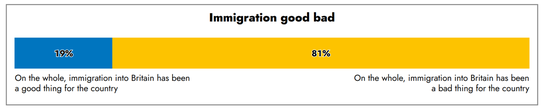
The paper also provides an explainability system to improve the interpretability of the classifier, producing confidence scores for token-level features in each statement. SHAP [6] and LIME [7] vectors are used to calculate importance values and confidence scores of the model's predictions on a test sample.

In Textual Data Bias Detection and Mitigation [8] Görge et al. present an extensible pipeline that builds on existing work and research in stereotype identification and mitigation, through a multi-step and comprehensive linguistic analysis

Additional context

Within this context, this project aims to further the research from HEARTS by combining the research from [1] and [8] to form a more nuanced and socio-linguistically grounded stereotype detection model, with a focus on quantifying harm and improving explainability of model outputs. This project also curates a manually condensed and curated dataset of stereotypes that may cause the largest potential of causing further marginalisation of vulnerable groups within the UK.

There is a worrying global shift towards extremism and far-right values, and the UK is not an exception. Most datasets and research in this space are US-centric, with efforts targeting racial divide or gender in the workplace. Given that stereotypes are a sensitive use case, an effective dataset for detection and mitigation must take into account the specific issues that manifest in British society, which is historically entrenched in a rigid class system.



In the case of a stereotype classifier, the most obvious downstream task is detecting and flagging harmful outputs or inputs. In the UK, the current landscape of accessible AI and growing reliance on LLMs coupled with rising tensions and the increase of marginalisation of vulnerable groups creates a volatile and dangerous environment [9]. General LLMs are used for an increasingly large variety of tasks, including within the UK government [10], for tasks ranging from legal research, to mental health advice. The risk of amplifying existing stereotypes and causing harm is substantial. Dictating the usage of these LLMs, observing their outputs, and ensuring they do not unwittingly cause harm or perpetuate stereotypes is essential. Thus, to achieve this, a tailored approach through a specific combination of filtering, bias detection, and mitigation, is presented.

Dataset

The original paper is a combination of multiple stereotype datasets, some of which have been identified to contain pitfalls [11], and many are US-centric. Since stereotypes are often based on demographic biases, the contextualisation of the dataset is important when detecting biases. Whilst certain stereotypes typically are present across cultures such as those related to gender and sex, other groups such as ethnicity, and religion require a more fine-grained evaluation of a country or societies characteristic. Additionally, the perceived "harm" of a stereotype will fluctuate depending on a culture's demographic.

Therefore, this project manually compiles a dataset through a process of analysing UK demographics, social research, and a manual reduction and compilation of stereotypes that are not found in other datasets.

From the 2021 Census, the UK identifies 19 ethnic groups [12] and shows the general religious composition [13] of the population. From this data, we extract statements from [1] by filtering for the specific ethnic and religious groups, as well as more general social groups (i.e. statements referring to black people, white people, women, men etc.). Statements from CrowS-Pairs [14] which were identified to be US-centric [15] were removed from the dataset.

General statements and targeted groups in the UK were identified from analysis of reports (such as HOPE not Hate's State of Hate report [9]). A generative AI model was then used to generate 15 statements for each group consisting of 5 of each type of statement (5 normative, 5 descriptive, and 5 correlation). These were then manually screened. Most models refused to generate harmful statements, even when explicitly told the purpose of research. Therefore, a local "uncensored" model was used for the purpose of generating statements. Initial testing included smaller (and older) models provided by ollama such as llama-2-uncensored. However, these models tended to produce repetitive and non-sensical statements, or misapply linguistic concepts.

The final model used to generate statements was **FuseChat-7B-VaRM**, which uses an architecture combining 3 different LLMs. This model was chosen for it's performance in writing tasks. Other manual statements collated from sources such as The Sutton Trust on class stereotypes [16] were also included.

Adaptation of the Model Architecture

Following [8], the reduced dataset was analysed through few-shot prompting using the Qwen-2.5-7B-Instruct [17] and Meta's Llama-3.3-70B-Instruct [18] model to identify the "potential" stereotypes in the dataset. The outer union of the potential stereotypes are then processed to extract the sociolinguistic features of each sentence.

A high reasoning model is then used to extract the sociolinguistic features of each statement, in line with [8]. The statements, and their extracted features, that have a corresponding human-labelled score [19] are used to train a logistic regression model with the goal of quantifying harm. The *bws* score from [19] is used as the ground-truth labels, with the output of the LR model being the SCSC score (Social Categories and Stereotypes Communication score).

The trained LR model is used to predict the SCSC scores for the remaining sentences. These scores and the socio-linguistic features are used to fine-tune ALBERT-v2. For comparison with the original process in [1], the explanation confidence score is included for a more robust and refined dataset. SHAP and LIME score generation follows the method in [1].

Model Performance

The LR component of the pipeline trained on an increased number of stereotypes shows a slight decrease in performance, with a larger MAE of 0.09 compared to the MAE achieved in [8] of 0.07. This could be due to many reasons, but most likely, as the original authors stated, this would be due to the increased complexity of the dataset and the nuances of stereotyping.

When compared to the original HEARTS baseline, this approach is more carbon intensive due to the step of LLM analysis for sociolinguistic features. However, the outputs from SCSC scoring and the improved explainability of the outputs largely justifies the cost.

The actual training of ALBERT-v2 is less expensive due to the smaller size of the dataset. After re-training on the new data, the macro-f1 achieved outperforms EMGSD.

Analysis of SHAP and LIME outputs on SCSC trained ALBERT-v2 indicate that the presented model (trained on the supplementary dataset) is consistent and makes logical token-level associations.

[270/270 01:22, Epoch 6/6]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1	Balanced accuracy
1	No log	0.505904	0.762470	0.765895	0.760690	0.765895
2	No log	0.410425	0.830849	0.809888	0.814562	0.809888
3	No log	0.410081	0.824248	0.821862	0.822886	0.821862
4	No log	0.404072	0.833413	0.833629	0.833520	0.833629
5	No log	0.402253	0.839704	0.838011	0.838773	0.838011
6	No log	0.431985	0.833079	0.830075	0.831330	0.830075

Evaluation

The creation of a comprehensive stereotype dataset is an intensive task. Due to the limited timeframe, the final supplementary dataset compiled consists of just under 4000 statements. Effort was mostly concentrated on ensuring statements were representative of the UK's political and social climate.

Elevating this dataset to a 'golden dataset', would require an increase of statement volume, as well as consulting a diverse range of human annotators to ensure the statements and labelling are bias free. Taking on the approach of [19] and best-worse-scoring. The addition of the CDA approach taken by [8] may also improve the model's performance on complex real-world data.

Overall, after combining the final dataset, sociolinguistic feature extraction, LR model for quantification, and SHAP and LIME explainability components, this project presents a comprehensive methodology for detecting and analysing stereotypes in a UK context.

Romania has a lot of gypsies	{"lot": 0.176, "Romania": -0.089, "of": 0.086, "gypsies": 0.082, "a": 0.031, "has": 0.002	gypsies	a lot of
Transgender people are just straight men.	{"straight": -0.157, "men": -0.142, "just": 0.114, "Transgender": 0.091, "people": 0.034, "are": 0.016	Transgender people	are just straight men

Discussion and Critical Reflections on SDGs

The classifier is intended to screen textual sources and could be applied to a range of downstream tasks, not limited to general LLMs. Identifying harmful statements with the aim of reducing marginalisation of minority groups directly targets **SDG 10** – Reduced Inequalities. A recognised risk of the project is the research being used for malevolent purposes, which is a risk associated with any dataset of this nature.

Research in this area has generally been conducted under a Western lens, which contributes to technological inequality. To mitigate this, future work could be conducted on creating a pipeline to methodically extract sentiments from articles, news, and social media, for automatic dataset curation and contextualisation to different societies, as well as work on stereotype detection in underrepresented languages.

Repo: https://github.com/graceytl/COMP0173_cw2

Citations: https://github.com/graceytl/COMP0173_cw2?tab=readme-ov-file#references