GeniL (Generalization in Language)

Data Card Authors: Aida Davani*, Sagar Gubbi*, Sunipa Dev, Shachi Dave, Vinodkumar Prabhakaran

The dataset is created as an effort for detecting generalization in language. This multilingual dataset covers sentences in English, French, Spanish, Portuguese, Arabic, Hindi, Bengali, Malay, and Indonesian and is annotated by native speakers of each language (note: spanish and portuguese items are labeled by annotators from Mexico and Brazil respectively).

Each sentence is collected from a public corpora of language (Common Crawl) and contains at least one identity group name and an attribute.

Data Card

DATASET TEAM(S)

Technology, AI, Society, and Culture (TASC) team, BiNDI, RAI-HCT

DATASET CONTACT

- Aida Davani: aidamd@google.com
- Sagar Gubbi: gubbi@google.com
- Sunipa Dev: sunipadev@google.com
- Shachi Dave: shachi@google.com
- Vinodkumar Prabhakaran: vinodkpg@google.com

DATASET AUTHORS

- Aida Davani, Research Scientist, Google
- Sagar Gubbi, Software Engineer, Google
- Sunipa Dev, Research Scientist, Google
- Shachi Dave, Software Engineer, Google
- Vinodkumar Prabhakaran, Research Scientist, Google

PRIMARY DATA MODALITY

Image Data
Text Data
Tabular Data
Audio Data
Video Data
Time Series
Graph Data
Geospatial Data
Multimodal (Please specify)
Others (please specify)
Unknown

DATASET SNAPSHOT

Size of dataset	
Number of Instances	57000
Number of Fields	57000
Field 1. Sentence	A sentence extracted from the public mc4 dataset (https://huggingface.co/datasets/mc4)
Field 2. Identity term	The identity term mentioned in the sentence
Field 3. attribute	The attribute mentioned in the sentence
Field 4. Generalizing	Whether or not the majority of the annotators labeled the text as generalizing
Field 5. Promoting	Whether or not the majority of annotators labeled the text as

DESCRIPTION OF CONTENT

The dataset contains sentences from a public corpora of text commonly used as a part of the training material for large language models.

Sentences are in nine different languages and each contain at least one identity term and attribute. The sentences are then labeled by a set of annotators who are asked whether the sentence is making a generalization about a group of people. If yes, annotators are asked to decide whether the sentence is promoting a generalization or only mentioning it and also highlighting the part of the sentence that contains this generalization.

	promoting a generalization	
DATASET SUBJECT	**EXAMPLE: DATA POINT**	**DATA FIELDS**

Sensitive Data about people
Non-Sensitive Data about people
Data about natural phenomena
Data about places and objects
Synthetically generated data
Data about systems or products and their behaviors
Unknown
Others
	This example is an actual data point from the data. E.g. of Data Point:
	
	`Sentence` | `Sierra Leoneans are the least respected people in the world, the world referred to us as subhumans, stupids and mere beggars.`
	`Identity term` | `Sierra Leonean`
	`Attribute` | `Beggars`
	`Generalizing` | `1`
	`Promoting` | `1`
		• ***Field 1.*** `sentence`
		• ***Field 2.*** `Identity term`
		• ***Field 3.*** `Attribute`
		• ***Field 4.*** `Generalizing`
		• ***Field 5.*** `Promoting`

DATASET PURPOSE(S)	**KEY DOMAINS OR APPLICATION(S)**	**PRIMARY MOTIVATION(S)**
Monitoring **Research** Production Others (please specify)	Domains Natural Language Processing, Computational Social Science Problem Space Capturing stereotyping and generalizing language.	This multilingual dataset is collected to capture the various ways in which generalizing language can be used for describing different social groups. Such a dataset can be instrumental in developing more effective methods for detecting stereotyping harms in language technologies.
DATASET USAGE	**INTENDED AND/OR SUITABLE USE CASE(S)**	**UNSUITABLE USE CASE(S)**
Safe for production use **Safe for research use** Conditional use- some unsafe applications Only approved use Others (please specify)	• To training classifiers for generalizations and stereotypes in different languages	1. As a comprehensive benchmark of all types of stereotype harms targeting specific social groups. 2. As a stand-alone benchmark for assessing safety of LLMs. 3. To fine-tune models to generate potentially stereotypical language.
SAFETY OF USE WITH OTHER DATA	**ACCEPTABLE TRANSFORMATIONS**	**BEST PRACTICES FOR JOINING OR AGGREGATING WITH DATASET**

Safe to use with other data Conditionally safe to use with other data Should not be used with other data Unknown Others* (Please specify)	**Joining with other datasets** **Subsampling and splitting** **Filtering** **Joining input sources** **Cleaning missing values** **Anomaly detection** **Grouping and summarizing** **Scaling and reducing** **Statistical transformations** **Redaction or Anonymization**	This dataset can be used along with the Multilingual SeeGULL dataset which includes stereotype labels for pairs of identity terms and attributes available in this dataset.

VERSION STATUS	DATASET VERSION	MAINTENANCE PLAN
Regularly Updated New versions of the dataset have been or will continue to be made available. Actively Maintained No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data. Limited Maintenance The data will not be updated, but any technical issues will be addressed. Deprecated This dataset is obsolete or is no longer being maintained.	**Current Version** 1.0 **Last Updated** 05/2024 **Release Date** 05/2025	

ACCESS POLICY	RETENTION POLICY	WIPEOUT POLICY
The data will be accessible under the Apache License 2.0	N/A	NA

DATA COLLECTION METHODS	DATA SOURCES	DATA COLLECTION

API Artificially Generated **Crowdsourced - Paid** Crowdsourced - Volunteer Vendor Collection Efforts Scraped or Crawled Survey, forms or polls **Taken from other existing datasets** Unknown To be determined Others (please specify)	Common crawl dataset (https://huggingface.co/datasets/mc4)	**Taken from other existing datasets** We extracted sentences from the public multilingual Common Crawl dataset. **Crowdsourced - Paid** We then had the items annotated through **Date of Collection:** Oct 2023 - Jan 2024 **Instrumentation:** Google's proprietary crowd work platform **Data Modality**: Text Data
INCLUSION CRITERIA	EXCLUSION CRITERIA	DATA PROCESSING
Sentences for annotation: Taken from existing datasets Items from the multilingual Common Crawl dataset were extracted by querying for sentences that mention specific pairs of identity terms and attributes. For each pair of identity terms and attributes we randomly selected at most 20 sentences. The participants were selected by the crowd compute team and they were native speakers of each language.		
SENSITIVE DATA	FIELDS WITH SENSITIVE DATA	SECURITY AND PRIVACY HANDLING

User Content User Metadata User Activity Data Identifiable Data S/PII Business Data Employee Data Pseudonymous Data Anonymous Data Health Data Children's Data **None** Others* (*please specify)	NA	NA

SENSITIVE HUMAN ATTRIBUTES	SOURCE(S) OF HUMAN ATTRIBUTES	RATIONALE FOR COLLECTING HUMAN ATTRIBUTES
Race Gender Ethnicity Socio-economic status **Geography** **Language** Sexual Orientation Religion Age Culture Disability Experience or Seniority Others (please specify)	**[Language]** and **[Geography]**: we know where annotators of each item are from and that their first language matches the language of the sentence.	We needed annotators to label items in their first language.

TRANSFORMATIONS APPLIED		LIBRARIES AND METHODS USED

Anomaly Detection Cleaning Mismatched Values Cleaning Missing Values Converting Data Types Data Aggregation Dimensionality Reduction Joining Input Sources Redaction or Anonymization **Others***	○	

SAMPLING METHOD(S)	SAMPLING CHARACTERISTIC(S)	• SAMPLING CRITERIA
Cluster Sampling Haphazard Sampling Multi-stage Sampling **Random Sampling** Retrospective Sampling Stratified Sampling Systematic Sampling Weighted Sampling Unknown Unsampled Others		For each pair of identity terms and attributes, we extracted all sentences in the corpus that mention both words. We then randomly selected 20 sentences.

ANNOTATION WORKFORCE TYPE	ANNOTATION CHARACTERISTICS		ANNOTATION DESCRIPTION
Annotation Target in Data Machine-generated Annotations Human Annotations - Expert **Human Annotations - Non-expert** Human Annotations - Employees Human Annotations - Contractors Human Annotations - Crowdsourcing Human Annotations - Outsourced / Managed Teams Unlabeled Others* (*Please specify)	Number of annotators per example	3	• Each sentence is shown to three native speakers of the language. • Annotators are asked whether the sentence makes a generalization. If the annotator selects yes, they are then asked what type of generalization they capture in the text. They can select either “promoting”, “mentioning”, or “other”. They are also asked who the generalization is about, and what word (or words) are being used to describe that group.

	ANNOTATOR BREAKDOWN		ANNOTATOR DESCRIPTION
	Annotator type	Paid - Non-expert	• We recruited ~50 annotators across all regions for annotating stereotypes. We needed each item in a language to be annotated by at least 3 people and we conducted the study in 9 languages. • To test their understanding of the task, we conducted a pilot annotation. The annotators were consistently in touch with the task supervisors to share their questions.
	Total unique annotators	50	
	Payment per annotator	Payment varied across different locales as per local market rates, but always above minimum wages where applicable.	
	Expertise of annotators	No previous training	

VALIDATION METHOD(S)	VALIDATION BREAKDOWN	DESCRIPTION OF VALIDATION
Data Type Validation Range and Constraint Validation Code/cross-reference Validation Structured Validation Consistency Validation **Not Validated** Others* (*Please specify)	N/A	

	VALIDATORS CHARACTERISTIC(S)	VALIDATORS DESCRIPTION(S)
	N/A (automatic validation)	N/A (automatic validation)

ML APPLICATION(S)		
N/A The dataset was not used for any applications. No training or fine-tuning of systems was performed. The data was only used for diagnostic analysis of existing models and not used to create any new systems		

Terms of Art

Concepts and Definitions referenced in this Data Card

Identity terms	Attribute Tokens (or tokens for short)
Definition: These are words used to describe a group of people with a common trait or identity. In the context of this data we focus on identity terms that pertain to regional identity, specifically demonyms. For eg: Croatians is a term used to describe the people of Croatia, Hawaiians is a term used to describe people who are from the US state of Hawaii.	Definition: These are characteristics or attributes for which we aim to identify stereotypical associations. These span categories like profession, adjectives, socio-economic status, subjects of study and so on. For eg: doctor, teacher (profession), poor, powerful (socio-economic status), smart, handsome, ugly (adjectives), computer science, mathematics (subjects of study) and so on.
Generalization	
Generalization is a broad statement made about a group of people, suggesting that certain characteristics or behaviors apply to all members of that group.	

Reflections on Data

Languages are selected to have representations from different continents and also cover languages with a high number of speakers across the globe. We cover languages spoken by North America, Europe, and Australia (English and French), Latin America (Spanish and Portuguese), Middle East and North Africa (Arabic), Southern Asia (Hindi and Bengali), and East Asia (Malay and Indonesian). Our choice of languages also were impacted by feasibility and cost of recruiting annotators from different regions.	