D3DECODE (Disentangling Disagreements in Data)

(Cross Cultural Disagreements in Data)
Data Card Authors: Aida Davani, Mark Diaz, Vinodkumar Prabhakaran

The dataset is collected for an empirical experiment on cross-cultural and social psychological differences in annotating offensiveness with a globally diverse set of participants.

Each participant is asked to label offensiveness in 35 textual items (items collected from a public research dataset from Jigsaw), and answer a questionnaire that evaluates their moral concerns (how much they value Care, Equality, Proportionality, Authority, Loyalty, and Purity).

Data Card

DATASET TEAM(S)

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

DATASET CONTACT

- Aida Davani: aidamd@google.com
- Mark Díaz: markdiaz@google.com
- Vinodkumar Prabhakaran: vinodkpg@google.com

DATASET AUTHORS

- Aida Davani, Research Scientist, Google
- Mark Díaz, Research Scientist, Google
- Dylan Baker, DAIR institute
- 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

Number of Labels	171800
Number of Items	4554
Number of Participants	4309
Average Labels Per Item	29.8
Algorithmic Labels	0%
Human Labels	100%

DESCRIPTION OF CONTENT

The dataset is collected for an empirical experiment with a globally diverse set of participants.

Each of the 4309 participants is asked to label offensiveness in 35 textual items (from 4554 items collected from a public research dataset from Jigsaw), and answer a questionnaire that evaluates their moral concerns (how much they value Care, Equality, Proportionality, Authority, Loyalty, and Purity).

Each row of the dataset shows the offensiveness label assigned by an annotator to a piece of text, along with information about the annotator.

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:

rater_id	R_37MEbYW3CbdvIyt
Gender	man
Age	18-24
Ses	6
Region	Sinosphere
Country	Vietnam
care	4.17
equality	2.83
proportionality	4.00
loyalty	4.33
authority	3.67
purity	3.33
item_id	123
text	Interesting reasoning. No doubt you feel we should apply it in cases of human suffering as well. When disasters strike, when famines hit, when wars are afoot... if there's too many victims to treat them all humanely, we should euthanize them. Right?
item_category	random
item_subcategory	random
rating_raw	1 (slightly offensive)
rating_binary	0

raters.csv:

- `rater_id`
  A combination of numbers and characters that is assigned as the ID to each participant
- `Gender`
  It can be either be man, woman,, non-binary / third gender, or other
- `Age`
  Can be any 5-year range from 18 to 100, e.g., 24-29
- `SES`
  Self-reported social and economic standing of the participant relative to their community. It is a number in the range of 1 to 10.
- `Region`
  One of the eight regions: Arab Culture, Indian Cultural Sphere, Latin America, North America, Oceania, Sinosphere, Sub-Saharan Africa, or Western Europe
- `Country`
  A country within the region that the participant lives in
- Participant moral value scores
  Six values in the range of 1 to 5, that shows participants response to questions regarding their moral concerns for Care, Equality, Proportionality, Loyalty, Authority, and Purity

items.csv:

- text
  A textual instance from Jigsaw's datasets collected for either the Toxic Comments Classification or Unintended Bias challenges.

rating.csv

- rating_raw
  a value from 0 to 4, 0 being not offensive at all and 5 being extremely offensive, a value of -1 in this column means that the rater did not understand the message
- rating_birany
  a binary value, with 0 showing a raw rating of 0 or 1, and 1 showing a raw rating of 2 or higher. A na value is equal to a raw rating of -1

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 Subjectivity in offensive language annotation	This dataset is collected to study the cross-cultural and social psychological differences in annotators that lead to their disagreements in annotating offensiveness in language
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 demonstrate cross-cultural, and psychological differences in detecting offensiveness	1. To training models for identifying annotators moral concerns, demographics, or origins based on their annotations 2. Creating benchmarks for offensive language detection by aggregating the labels 3. Training offensive language detection based on perspective that are rooted in annotators demographics
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**	N/A (we have not attempted to use this dataset with other datasets, but we do not anticipate any issues)
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** 08/2023 **Release Date** 08/2023	

ACCESS POLICY	RETENTION POLICY	WIPEOUT POLICY
The data will be accessible under the Apache License 2.0		There is no raw participant information stored on our end, and the research vendor is required to delete raw data from their equipment and drive within 3 months of data collection.

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)	[1] Jigsaw. Toxic comment classification challenge (2018) Accessed: 2021-05-01 [2] Jigsaw. Unintended bias in toxicity classification (2019) Accessed: 2021-05-01	**Taken from other existing datasets** We first selected textual items from the two Jigsaw datasets. **Vendor Collection Efforts** We then had the items annotated through **Date of Collection:** Dec 2021 - Aug 2022 **Instrumentation** **Data Modality**: Text Data **Survey, forms or polls** The same data collection also asked participants to fill out a survey **Date of Collection:** Dec 2021 - Aug 2022 **Instrumentation** **Data Modality**: Text Data

INCLUSION CRITERIA	EXCLUSION CRITERIA	DATA PROCESSING

Our selected set items include three categories of items: 1. Random (50%): items selected randomly from the Jigsaw dataset. We assigned more weight to items with high inter-annotator disagreement in the original dataset; 2. Moral Sentiment (10%): items that include moral sentiment. Moral sentiment of items were identified by a moral language tagger; 3. Social Group Mentions (40%): items that mention specific social group identities related to gender, sexual orientation, or religion. The participants were selected based on their demographics and language: 1. Region of residence: we recruited at least 450 participants from each of the eight regions, which also included at least 100 participants per country within the regions, 2. Gender: within regions, we controlled each binary genders (Collecting non-binary gender information was not safe across the countries, so we limited this study to binary gender.) to describe at most 60% of the participants, 3. Age: in each region at most 60% of participants are 18 to 30 years old and at least 15% are 50 years old or older, 4. English fluency: we only selected participants with high self-reported proficiency in reading and writing English.	Participants were excluded if: • They did not self report high fluency in writing and reading English • They did not agree with the consent form • They were younger than 18 • They did not finish the survey • They took less than 5 minutes to finish the survey • They labeled every question with the same answer • They failed one or several of the test items* *Test items These items (listed below) were unanimously labeled as “not offensive at all” by authors and were used as an attention check to remove participants who annotated them otherwise. 1. Could you explain your point in a little more detail please. 2. The NY Times yesterday had a rebuttal by the Iranian UN ambassador. 3. Ditto to everything said here. 4. Just looked at the link. Impressive and missed opportunity. 5. Don't be so hard on yourself. Your life will go on.	Offensive labels are converted from categorical options (not offensive at all to extremely offensive) to their respective numeric values.
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)	**[Gender]:** self-reported **[Age]:** self-reported **[Socio-economic status]:** self-reported **[Geography]**: The self-reported country of residence is collected in the survey.	Previous research on annotating offensiveness suggests that demographic differences can lead to subjectivity in annotations, for instance, women who have been exposed to harassment are more likely to recognize abusive language toward women. We collected these data to control for their impacts on annotations.

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		

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	24	• Each item is annotated by at least 3 participants in each region (leading to at least 24 total annotations for each item) • Each annotators were asked to annotate offensiveness of 35 of text items • For half of the participants we used the following definition for "extremely offensive language" from prior literature as: *profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group.* • The other half of the annotators were not provided with a definition to create a control setting. They were asked to annotate based on their definition of offensiveness.

	ANNOTATOR BREAKDOWN		ANNOTATOR DESCRIPTION
	Annotator type	Paid - Non-expert	We recruited 450 annotators from each of the eight regions, with at least 100 annotators from specific counties. Annotators are not typically involved in crowdsourcing tasks and use the platform to take part in social surveys.
	Total unique annotators	4309	
	Payment per annotator	Annotators were paid in line with company policy and above market rate.	
	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		