TIDAL

(Textual Identity Detection and Augmentation Lexicon)
Data Card Authors: Sameer Sethi and Emmanuel Klu

TIDAL (Textual Identity Detection and Augmentation Lexicon) is created as part of the TIDE research effort (https://arxiv.org/abs/2309.04027) by the SCOUTS team (https://sites.research.google/scouts/). It consists of a lexicon of identity terms and associated context and is formatted in XML using the Lexical Markup Framework ISO standard (https://www.iso.org/standard/82014.html). This dataset has been created to aid annotation of textual datasets with identity tokens to help with fairness evaluation and remediation of datasets and models. It supports only the English language, but has broad coverage of identity terms and grammatical variants across three IdentityGroups: Race, Nationality or Ethnicity (RNE), Sexual Orientation, Gender Identity, Gender Expression and Sex Characteristics (SOGIESC) and Religion.

Data Card

DATASET TEAM(S)

Societal Context Understanding Tools and Solutions (SCOUTS), Responsible AI and Human Centered Technology (RAI-HCT)
Google Research

DATASET CONTACT

- Sameer Sethi: sethis@google.com
- Emmanuel Klu: eklu@google.com

PRIMARY DATA MODALITY

Image Data
Text Data
Tabular Data
Audio Data
Video Data
Time Series
Graph Data
Geospatial Data
Multimodal (Please specify)
Others: XML Lexicon dataset based on LMF ISO standard
Unknown

DATASET SNAPSHOT

Size of dataset	23 MB
Number of Instances	15123
Number of Fields	12

DESCRIPTION OF CONTENT

The TIDAL dataset consists of lexical entries and their related forms (e.g. black, gay, trans, hindus) that are associated with identity groups. Each head and related form is associated with grammatical properties (e.g. part-of-speech, grammatical gender) and context (or "sense") entries (e.g. identity groups/subgroups, connotation).

TIDAL has 1565 English language identity lexical entries, with over 14148 related lexical forms and 15123 context/sense entries.

DATASET SUBJECT

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*
(*Data about social phenomena)

EXAMPLE: DATA POINT

This example is a sample data point from the data, showing some key fields in the dataset.

Term	Language	partOfSpeech	Connotation	IdentityGroup	IdentitySubgroup
trans	en	NOUN	NEUTRAL	SOGIESC	Gender Identity > Transgender

DATA FIELDS

- ***Field 1.*** *Term: Word or a phrase describing, associated with or targeting an Identity Group*
- ***Field 2.*** *Language: Language of the identity term*
- ***Field 3.*** *Group: Top-level Identity Group associated with the identity term*
- ***Field 4.*** *Subgroup: Specific Identity Subgroup associated with the identity term*
- ***Field 5.*** *Connotation: Whether the term can have Neutral or Pejorative or both meanings in different contexts*
- ***Field 6.*** *HasNonIdentityMeaning: A boolean value for whether the Term has a non-identity meaning in some context or not*
- ***Field 7.*** *partOfSpeech: Part-of-speech of the term*
- ***Field 8.*** *grammaticalGender: Grammatical gender of the term*
- ***Field 9.*** *grammaticalNumber: Grammatical number of the term*
- ***Field 10.*** *grammaticalCase: Grammatical case of the term*
- ***Field 11.*** *relatedForm_relType: If the term is a grammatical variant, then what root term is it related to (VariantOf, PersonNounCombinationOf)*
- ***Field 12.*** *Provenance: This field is available across all above field to identify the provenance of where this data came from (if it's from human computation, then the value of this field is HCOMP)*

DATASET PURPOSE(S)

Monitoring
Research
Production
Others (please specify)

KEY DOMAINS OR APPLICATION(S)

Domains

Natural Language Processing, Algorithmic Fairness

Problem Space

ML Fairness evaluation and remediation of text classifiers and generative models

PRIMARY MOTIVATION(S)

This dataset was created to aid in the development of a textual identity detection annotator. The annotator is then used to improve human-in-the-loop processes and fairness evaluations of text classifiers and language models.

DATASET USAGE

Safe for production use
Safe for research use
Conditional use- some unsafe applications
Only approved use
Others (please specify)

INTENDED AND/OR SUITABLE USE CASE(S)

- Identity token(s) detection in textual datasets to aid in fairness evaluations and remediations of models and datasets.

UNSUITABLE USE CASE(S)

1. As a benchmark for assessing fairness or ensuring lack of fairness
2. As a resource for any bias mitigation in production systems
3. To train demographic predictors using lists of proxy identity terms

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** Others (please specify)	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** 06/2023 **Release Date** 09/2023	• We will address any issues that people might face in the dataset usage.

ACCESS POLICY	RETENTION POLICY	WIPEOUT POLICY
The data will be accessible under the CC-BY 4.0	N/A	N/A

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)
	Seed sources for Terms
	Date of Collection: Jan 2020 - Dec 2022
	Data Modality: Text Data
	Process:
	• Sources:
	  ◦ **UN Data Ethnic Groups**
	  ◦ **UN Data Religious Groups**
	  ◦ **CAMEO Event Data Codebook**
	  ◦ **Wikipedia list of demonyms**
	  ◦ **GLAAD Glossary**
	  ◦ **HRC Glossary**
	• Processing: Seed terms were parsed from above data sources, cleaned (remove hyphens, slashes etc.) and then lowercased.
	
	Annotations: Crowdsourced - Paid
	Appen Data Annotation Platform: Annotation Data Annotation Platform and Managed Services were used to grammatically expand the seed terms
	Date of Collection: Jan 2020 - Dec 2022
	Instrumentation: Annotation Platform
	Data Modality: Text Data
		Annotations: Crowdsourced - Paid
		Collected and included
		• Term: word or phrase
		• IdentityGroup: Groups the term is associated with
		• NonIdentityMeaning: whether the term has a non-identity meaning
		• PartOfSpeech, Grammatical gender, number and case: part of speech properties of the term
		• Connotation: whether the term is NEUTRAL or PEJORATIVE
		Collected and excluded
		• Country names were used as a filter to remove mentions of countries from the seed terms

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) **None**	**N/A**	N/A

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*	○	• Term filtering: pandas dataframe basic functions • Annotation aggregation: pandas dataframe basic functions

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* **None**		
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)	Context annotations Number of expert annotators per example: 1 Number of non-expert annotators per example: 3	Expansion • Expand the term to all grammatical variants • Expand the term to Person Noun Combinations (e.g. indian man) Context annotations • What are all the applicable connotations for the term (NEUTRAL, PEJORATIVE or BOTH) • What IdentityGroups does the term describe, target or associate with (RNE, Religion, SOGIESC)
	ANNOTATOR BREAKDOWN	
	Annotator type: Paid - Expert Total unique annotators: 1 Average cost/tuple: 2 USD (avg) Expertise of annotators: Linguistic Expertise Annotator type: Paid - Non-Expert (if applicable) Total unique annotators: 3 Average cost/tuple: 0.5-1 USD Expertise of annotators: Trained on task	
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	Data Type Validation Check for null values, connotation values (NEUTRAL, PEJORATIVE), IdentityGroup values (RNE, SOGIESC, Religion), Citation values (URL) using basic python dataframe functions.
	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.		

Reflections on Data

Limitations due to human annotation	Annotation about identity terms context (like connotation) and their prevalence in society is subjective. While we attempt to capture diversity in our annotator pool, we recognize that it still does not capture all different opinions and perspectives. Future iterations of such data collection should take more participatory approaches and involve communities with lived experiences on the harms of bias in society.
No ground truth on "Identity Terms" or associated context	We recognize that there is no "ground-truth" on labeling something as an "identity term" or having "NEUTRAL" or "PEJORATIVE" connotation, and that the context varies based on languages, locales, cultures, lived experience, etc. This is an inherently subjective opinion that is influenced by socio-cultural factors and personal experiences. Thus, we caution against using the data in this dataset to classify text as "identity" vs "non-identity" in production.
Identity terms or context not captured by this dataset	We generate candidate identity terms using seeds which could influence what is generated. Diversity of our human annotators available also limits our annotations, impacting what gets annotated as neutral or pejorative. Additionally, identity terms and context not captured by our dataset may exist.

Caution against calling models “fair” based on evaluation on this dataset	This dataset is insufficient to capture all identity mentions for all languages, locales, cultures, etc. The dataset reflects the judgments of a small number of annotators. Hence, it should be used only for diagnostic and research purposes, and not as benchmarks to prove the lack of bias.