BiasTestGPT: Datasheets for Datasets

Rafal Kocielnik

June 2023

1 Introduction

This Datasheet follows the structure introduced in Gebru'21. We provide answers to questions about the dataset creation, preprocessing, intended use, and maintenance plan. The various components of the **BiasTestGPT** framework are hosted on publicly available platforms. The codebase, the scripts to reproduce the benchmarks, and the data from the paper are hosted as a public BiasTestGPT Github repository. The framework offers an Open-source Bias Testing Tool hosted on HuggingFace Spaces. This tool is connected to a dataset that we describe in this datasheet - Test Sentences Dataset. The datset is hosted on HuggingFace in common CSV format. The bias testing tool loads the dataset and also adds new entries to it based on user interaction. Specifically new test sentences generated in the tool are added to the **Test Sentences Dataset** dataset.

2 Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable the testing of Pre-Trained Language Models (PLMs) for the presence of social bias. Existing datasets rely on simplistic templates, or on crowd-sourced data. A limited number of fixed templates have been shown to produce misleading results due to lack of grammatical structure and unnatural context of use. Crowd-sourced datasets, on the other hand, are expensive to collects and update.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Rafal Kocielnik, Shrimai Prabhumoye, and Vivian Zhang using open-sourced gpt-3.5-turbo generative PLM and commercial MT-NLP-530b generative PLM.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from several sources: the National Science Foundation (Computing Innovation Fellows Program), Caltech, and Activision-Blizzard-King.

Any other comments? None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, pho-

tos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The **Test Sentences Dataset** consists of **generated test sentences** along with **social group term** and **attribute term** used for controlling the generation. It also includes **template** version of the sentence as well as social group terms (**grp_term1** and **grp_term2**) for filling-in "stereotyped" and "anti-stereotyped" version of the sentence. The two label fields - **label_1** and **label_2** indicate which version of the sentence is counted as "stereotyped" and which as an "anti-stereotyped". Additional **type** and **gen_model** columns indicate whether the sentence was generated by the authors or came from the associated HuggingFace tool. The precise listing of the data fields, their meaning as well as a preview of the dataset is available on Test Sentences Dataset HuggingFace hosting.

How many instances are there in total (of each type, if appropriate)?

There are **15332 sentences** in **Test Sentences Dataset**, but the dataset is growing as new sentences can be created using our HuggingFace tool and they are automatically saved to this dataset.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample in a way as the automated generations were controlled by social bias specification. I.e. the generator models (gpt-3.5-turbo and MT-NLP-530b) was not asked to generate free text but rather to generate test sentences connecting particular social group and attribute terms. This was done to: 1) control the behavior of the model so that that test sentences as usable for testing other PLMs, 2) To trigger expressions of particular biases following provided specification. We will release all the data that we have generated.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

In **Test Sentences Dataset**, the raw data in each instance consists of text fields *sentence*, *org_grp_term*, *att_term*, *template*, *grp_term1*, *grp_term2*, *label_1*, *label_2*, and *bias_spec*. Both *org_group_term* and *att_term* appear as tokens in the sentence. The *template* represents version of the sentence where *org_grp_term* is replaced by a *"[T]"* mask token. Each instance also includes *type* and *gen_model* columns defining id the sentence was included in original analysis or not and what generator model produced the sentence. The precise listing of the data fields, their meaning as well as a preview of the dataset is available on Test Sentences Dataset HuggingFace hosting.

Is there a label or target associated with each instance? If so, please provide a description.

Each sentence has two label fields: *label_1* and *label_2*. The value in these indicates whether terms from *grp_term1*, *grp_term2* are considered *"stereotyped"* or *"anti-stereotyped"* when plugged into the *template*.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Data is artificially generated and nothing is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Not explicitly, but generated individual test sentence instances from Test Sentences Dataset are related to bias specifications in column “bias_spec”. They are indeed linked dynamically in the HuggingFace Bias Test Tool. The link is in the form of sentence generated for particular social group term (e.g., “man”) and attribute term e.g., “math” such: “The man is really good at math” can be used for testing bias specification that defines these social groups and attributes.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no specific train/test splits. Train split is not applicable for our use case because we are releasing a dataset to explicitly only test the models. However, the datasets contain two categories of instance due to the fact that it is live and able to grow. This types are mean to delineate novel data from the data used during the analysis.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Yes, there are aspects in the test sentences that could be considered noise. We specifically talk about the categories of noise present in our paper based on manual examination of some 1.3k generated test sentences. We estimate the noise prevalence in different categories as:

- **I1:** Related group references - Additional terms (e.g.,“her”, “his”) that reveal social group (12.8%)
- **I2:** Additional attributes - Attributes additional to the tested ones (3.7%)
- **I3:** No group - attribute link - Does not directly link group and attribute terms (3.3%)
- **I4:** Negative framing - The group and the attribute connected via negation (3.0 %)
- **I5:** Unrelated group references - Incoherent or non-grammatical (2.9%)
- **I6:** Different meaning - Terms referring to social groups others than tested (2.6%)
- **I7:** Incoherent/non-grammatical - Different interpretation of tested terms (2.5%)
- **I8:** Incomplete sentence - Generation does not form a complete sentence (1.9%)

We subsequently remove sentences with these issues and replace them with new generations without issues to estimate the impact. We find that the mean bias score across all Tested PLMs changed by 1.6% (from 55.2 to 54.3). We investigate the impact of each individual issue category as well. Removing only Related group references leads to the highest mean change in individual bias score of 2.89% followed by Additional attributes with 1.56% and Negative framing with 1.48%.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

No, it does not link to anything and it it self-sufficient.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, the dataset is synthetic, generated using PLMs so it should not contain any private or identifiable information.

Does the dataset contain data that, if

viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Some of the data (especially generated test sentences) can indeed be considered offensive, insulting or toxic. This is because the dataset is meant to test social bias in existing PLMs and some associations of social groups and negative attributes can indeed be considered offensive. We perform analysis of **sentiment** and **toxicity** of our dataset and find that our generations actually have lower toxicity scores than crowd-sourced bias testing dataset we compare to. Still, we found 60 sentences that received high toxicity score (> 0.5) using ToxicBERT model. We examined these sentences, a representative example is: *"Cindy was a horrible person to be around."* generated for requested terms: *"Cindy"* and *"Horrible"*. We note that the toxicity in these sentences stem from the requested terms provided in the bias specification and hence are unavoidable to be able to use the sentences for social bias testing.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No, the sentences were not directly written by people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

It can be considered sensitive as it relates to social stereotypes, but these stereotypes are not collected directly from any people. Some of the bias specifications might be considered sensitive as they relate to race, gender, and intersectional biases.

Any other comments? No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data has been automatically generated using large language models - gpt-3.5-turbo and MT-NLP-530b. The sentences were generated using bias specification terms provided by prior work. The model was instructed to write sentences connecting these terms.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We used ChatGPT API for gpt-3.5-turbo and an API for MT-NLP-530b model.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

We used temperature=0.8 for generation. The exact details are provided on Section 2. BiasTestGPT Generation Framework of the paper as well as in Appendix B. We included all attribute terms from bias specification in our generation, but we samples randomly uniformly the social group terms.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process involved only the researchers working on the paper.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The gpt-3.5-turbo-based sentences have been created between 1st May 2023 and mid-June 2023. The data from the MT-NLP-530b in the dataset has been created earlier, between May and December 2022.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

There was no process as no human subjects were involved and not human data was collected.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The dataset relates to social biases expected in language use and reflected in the Pretrained Large Language models. This dataset however has not been collected by people, does not contain any personal or identifying information.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Not applicable.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Not applicable.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Not applicable.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Not applicable.

Any other comments?

No additional comments.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The generations from gpt-3.5-turbo and MT-NLP-530b that did not contain the terms requested were removed (rejection sampling). The sentences were also sanitized by removing unusual symbols in generations (such as #_ × ˜) and ensured to end with a ”.” in case the model did not generate a punctuation symbol. The multi-word phrases for social group and attribute terms were ensured to be separated by ” ” in the sentence.

Each sentence was turned into its templated version, by replacing the social group term with ”[T]” using a regular expression in the form of f“$(|[]+)grp\_term.lower()[.,!]+$”, which accounts for multi-word phrases and terms being a substring of longer words, e.g., “he” in the “there”.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

Yes, the dataset contains raw data in the form of original generated sentences as well as the templates version. The dataset, however, does not contain generations not including the requested terms.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the processing scripts at all steps are available in the github repository.

Any other comments?

None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

At the time of writing the dataset has been used in the end-user open-sourced tool available on HuggingFace HuggingFace Bias Testing Tool and for evaluations presented in the paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

As the dataset is publicly hosted on HuggingFace platform, any downloads and uses in HuggingFace spaces are automatically tracked and displayed on the dataset landing site Sentence Dataset.

What (other) tasks could the dataset be used for?

The dataset's primary use is for social bias testing on open-sourced PLMs. For this purposes it can also be used to compare several different metrics of bias on natural sentences. It can also be used as a benchmark for controllable text generation.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

We note that there is some noise in the data. We descibed the categories of noise and quantified its impact, but it may still impact future uses with other types of PLMs than the ones we tested on. There is not risk of harm in terms of privacy. There is a minimal risk of harm in relation to exposure to toxic content. As we noted on the limitations, despite control over the generaiton, there is a possibility that the generator model (i.e., ChatGPT) still introduces some biases of its own. We note that this is also on issue with any available crowdsourced or generated dataset.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset is collected to evaluate internal representation of social bias in pretrained language models. It is hard to say how well this internal bias translates to downstream tasks aftet the evaluated model is fine-tuned. Therefore social bias evaluation using this dataset should not be the sole means of estimating model bias.

Any other comments? None

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset is publicly available on the internet and hosted on the HuggingFace platform Sentences Dataset.

How will the dataset be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is hosted on Hugging Face hub platform Sentences Dataset. It does not currently have a DOI, but it is very esy to generate one using this plaform.

When will the dataset be distributed?

The dataset was released on June 10th.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The datast is distributed under Apache-2.0 license. There is no license, but there is a request to cite the corresponding paper of the dataset is used.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown to authors of the datasheet.

Any other comments?

None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of the paper will be maintaining the dataset as well as members of AnimaLab group on HuggingFace. The tool is hosted on the organization page - AnimaLab, which ensures that the tool and the dataset will be maintained, even if the individual authors will leave the group.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The curator of the dataset can be contacted on rafalko@caltech.edu.

Is there an erratum? If so, please provide a link or other access point.

There is no erratum at this point.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Updates to the dataset are happening constantly via the HuggingFace tool which saves the new generations to the dataset. To separate these updates from the evaluated portion of the dataset as described in the paper, there is a *“type”* field present in the dataset which contains value *“paper”* for instances used in the original evaluation in the paper, and value *“tool”* for instances added via interaction with the tool. In the future, we are planning to add a field to designate different verified versions of the data. The maintenance plan includes updates to the sentences if major challenges are found as well as additional of additional bias specifications to the dataset in a manner that is well-marked and separated from the initial biases we introduced. Again, we would like to emphasize that the main purpose of the tool is to support the discovery and documentation of novel social bias via the HuggingFace tool. We added this maintenance plan to the supplementary materials.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No, the dataset does not relate to people.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes, the older version will be kept under the same dataset hosting location and separated by a *"version"* field.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

The contributions to the dataset are made via interaction with the HuggingFace tool, which is also publicly available, so anyone can contribute to the dataset. The contribution is, however, controlled as the sentences are generated via gpt-3.5-turbo. As there is no specific validation/verification mechanism, but the contributions beyond the orignal dataset introduced in the paper are designated with different value in the *"type"* field: *"tool"* - for public contributions via the tool interface, *"paper"* - for original sentences generated by the authors and used in the evaluation in the paper.

Any other comments?

None.