

Dataset for BEAD (BEAD Dataset)

June 5, 2024

1 Datasheet

We follow the datasheet documentation framework in [3].

1.1 Motivation

- What is the purpose of the dataset?

This dataset aims to provide a comprehensive resource for bias detection and evaluation across multiple NLP tasks. The goal is to identify biases in data before they are employed to build AI applications or to analyze outputs in LLMs generations. This research endeavor is designed to help stakeholders better understand the capabilities, limitations, and potential risks associated with deploying different language models into applications. This project is organized around the following five areas:

1. Text classification for bias, toxicity, and sentiments.
2. Token classification to find indicators of biases in language.
3. Bias quantification in demographic groups through stereotype analysis and bias measurement.
4. Language generation for creating benign language.

Additionally, the dataset can be used for fairness studies, aiming to improve AI system equity and inclusivity.

- Who created the dataset and who funded its creation?

This dataset is created by Shaina Raza in Vector Institute for Artificial Intelligence [Vector Institute](#).

1.2 Composition

- What do the instances that comprise the dataset represent?
The instances in the dataset represent labeled text segments aimed at identifying various types of biases in language. Each instance consists of a text snippet annotated for the presence of bias based on specific categories such as gender, ethnicity, or age or demographic mention and is labeled for bias, toxicity and sentiment. The dataset also contains portion of data for benign language generation to debias the biased outputs.
- How many instances are there in total?
The total number of instances in the broader dataset is 3,674,927. For gold labels, which represent the second portion of annotated data, there are different counts for each task:
 - Text Classification: 50,000 instances
 - Token Classification: 27,800 instances (including 139,200 biased tokens)
 - Aspect Categories: 50,000 instances
 - Bias Quantification:
 - * Token prediction: 29,850 instances
 - * Stereotype Assessment: 352 instances
 - * Demographic Alignment: 520 instances
 - Language Generation: 8,300 instances
- What data does each instance consist of?
Each instance in the dataset consists of a text snippet and corresponding annotations specific to the task it belongs to. For Text Classification, instances include the text and labels for bias, toxicity, or sentiment. Token Classification instances include text and annotated named entities or biased tokens. Aspect Categories instances consist of text and aspect-related labels. Bias Quantification instances contain text with annotations for token prediction, stereotype assessment, or demographic alignment. Language Generation instances involve text snippets used for generating benign language annotations. The metadata of the dataset is comprehensively described following the Croissant metadata standard and can be accessed at [this link](#).

- Is there any missing data? If so, why?
There may be few instances in the main dataset where data is missing. This can occur due to various reasons such as incomplete text segments, ambiguities during annotation, or difficulties in obtaining certain annotations. Any missing data in the gold labels portion has been addressed and filled through additional verification and relabeling processes.
- Are there any errors, sources of noise, or redundancies in the data?
Efforts have been made to minimize errors, sources of noise, and redundancies in the dataset. However, due to the complex nature of language and human subjectivity in review process of LLM based annotations, some instances may still contain noise or minor redundancies. These issues are addressed through verification and a second relabeling process (gold data is released) to ensure the quality and reliability of the data.

1.3 Collection Process

- How was the data collected?
To ensure our dataset covers a wide range of bias aspects (political, hate speech in social media, ageism, social prejudices) and multiple NLP tasks, we gathered data from various existing datasets by complying with licensing and permissions and re-labeled them to support multiple NLP tasks. Existing datasets include MBIC (1,700 records), Hyperpartisan (755,000 records), and Multi-dimensional News (2,000 records) for political biases; Toxic Comments (159,571 records) and Jigsaw Unintended Bias (1,800,000 records) for hate speech and toxicity; Ageism (887,656 records) for health, ability, and body image biases; and Social Biases (44,000 records) for various social prejudices. They are credited in [website](#) and [paper](#).

For underrepresented aspects like occupational, technological, and environmental biases, we curated approximately 25,000 news articles via Google RSS from January 1, 2023, to May 30, 2023, to identify and label these biases. We have secured the necessary permissions and appropriately cited the sources utilized for our research.

- Who was involved in the data collection process?
Shaina Raza.

- Over what timeframe was the data collected?
January 2023-June 2023
- How was the data associated with each instance acquired?
Each instance in the dataset was carefully associated with its annotations through a rigorous process. Initially, data from various existing datasets were collected and re-labeled to ensure consistency and support for multiple NLP tasks. For political biases, hate speech, and toxicity, data from datasets such as MBIC, Hyperpartisan, and Jigsaw Unintended Bias were utilized. Health, ability, and body image biases were sourced from the Ageism dataset, while social prejudices were covered using the Social Biases dataset. Underrepresented aspects were addressed by curating news articles via Google RSS. Each text snippet was annotated using a combination of GPT-3.5 and active learning techniques. Subsequently, a refined gold dataset was created using GPT-4 with intensive manual review, ensuring high-quality annotations for approximately 50,000 records.
- If the dataset is a sample, then what is the population? What sampling strategy was used?
Both the main dataset and the gold portion are constructed with samples from various sources, covering a broad range of bias dimensions such as political biases, hate speech, ageism, and social prejudices. The main dataset was sampled based on these bias dimensions from sources (existing dataset with proper permissions and licence usage) and through scrapping Google RSS to ensure diverse and representative coverage. The gold portion was further sampled using a mix of aspects and balanced labels to create a high-quality, refined subset for more precise annotations.

1.4 Preprocessing/cleaning/labeling

- Was any preprocessing, cleaning, or labeling of the data done?
Yes, extensive preprocessing, cleaning, and labeling were done to ensure the quality and consistency and easy usage of the dataset.
- How did the preprocessing, cleaning, or labeling process work?
The preprocessing involved several steps. First, data from various existing datasets were collected and combined. Any duplicates or irrelevant

data were removed. The text data were then normalized, which included lowercasing, removing special characters, and tokenization. For labeling, a combination of automated and manual processes was used. Initially, automated labeling was done using GPT-3.5, followed by an active learning approach to iteratively improve the labels. Finally, the gold portion of the dataset underwent a thorough manual review and relabeling using GPT-4 to ensure high-quality annotations.

- Who was responsible for these processes?

The preprocessing, cleaning, and labeling processes were carried out by Shaina Raza and her team with expertise in natural language processing. The automated labeling was overseen by the team, while the manual review and refinement of the gold dataset were conducted by qualified students under the supervision of senior researchers.

- What tools were used to process the data?

The data processing utilized several tools and technologies. For text normalization and tokenization, standard NLP libraries such as NLTK and SpaCy were used. Automated labeling and active learning employed OpenAI’s GPT-3.5 and GPT4 models, and manual review and relabeling were done using GPT-4. Additionally, custom scripts in Python were developed for data cleaning and preprocessing tasks. The models built are evaluated using ML methods and some LLM based scoring like DeepEval [2].

1.5 Uses

- Has the dataset been used for any tasks already?

Yes, the dataset is used for multiple NLP tasks, including text classification for detecting bias, toxicity, and sentiment, as well as token classification for identifying biased tokens and named entities. The tasks are used for benchmarking the tasks.

- Has it been used for training, validation, or testing?

The dataset has been used for training, validation, and testing purposes across various NLP models. Specifically, it has been used in training bias detection models, validating the performance of these models, and testing their generalization capabilities.

- What other tasks could the dataset be used for?
The dataset could be used for several other tasks, including but not limited to:
 - Stereotype and bias quantification in demographic groups
 - Language generation for creating benign and unbiased text
 - Fairness and ethical AI research
 - Sentiment analysis and hate speech detection
 - Developing tools for automated bias detection in media and social platforms
- Is there a repository that links to any or all papers or systems that use the dataset?
Yes, there is a repository that links to papers and systems utilizing the dataset. The repository is available at [Huggingface Link](#) and [website](#) where users can find references to details about the dataset.

1.6 Distribution

- How is the dataset distributed?
The dataset is distributed online through the Hugging Face platform. It is available for download and use by the research community and other interested parties. The metadata of the dataset is comprehensively described following the Croissant metadata standard and can be accessed at [this link](#).
- What license, if any, is it distributed under?
The dataset is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) [1] license. This allows users to share and adapt the dataset, provided appropriate credit is given.
- Are there any fees for access?
No, there are no fees for accessing the dataset. It is freely available to the public to encourage research and development in bias detection and evaluation.
- Is there a digital object identifier (DOI) for the dataset?
Yes, the dataset has been assigned a digital object identifier (DOI) to

facilitate citation and reference. The DOI for the dataset is associated with [Huggingface](#).

1.7 Maintenance

- How is the dataset maintained?
The dataset is maintained by Shaina Raza. Regular checks are conducted with her team to ensure the dataset remains up-to-date and any issues are promptly addressed.
- Who is supporting, hosting, and maintaining the dataset?
The dataset is supported and maintained by Shaina Raza at [email here](#). The dataset is hosted and distributed on the Hugging Face platform.
- Will the dataset be updated? How often and what is the process?
Yes, the dataset will be periodically updated to incorporate new data, improve existing annotations, and correct any identified errors. Updates are planned on a need basis. The process involves re-collecting data, re-annotating it as necessary, and verifying the quality of the updates.
- Is there a point of contact for the dataset?
Yes. Please contact **Shaina Raza** at shaina.raza@utoronto.ca for any questions or support related to the dataset. For more information or to provide feedback, please visit [BEADs Feedback](#).

1.8 Legal and Ethical Considerations

- Discuss any legal and ethical considerations in the collection and use of the dataset.
Full permissions are taken from the sources of existing datasets. Google RSS usage was validate. The dataset underwent legal review at the Vector Institute and was approved following a thorough consultation and review process by the legal team.
- Are there privacy or security concerns?
No, there are not any privacy or security concerns involved. The dataset may contain indicators of some social media personalities but those that are available open and no security concerns associated.

- Is there any sensitive information involved?
The dataset may include sensitive information related to individuals' demographics, opinions, or behaviors. Such information is handled with utmost care to prevent any misuse or unauthorized access, in compliance with data protection regulations.

References

- [1] CC BY-NC 4.0 Deed — Attribution-NonCommercial 4.0 International — Creative Commons. <https://creativecommons.org/licenses/by-nc/4.0/>. Accessed on 05/30/2024.
- [2] Confident AI. Deepeval. GitHub repository, 2024.
- [3] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.