AP2: Datasheet

In this deliverable, you will create a *partial* datasheet for the annotated dataset that you are building. Review Gebru et al.'s "Datasheets for Datasets" paper and answer the following questions about your annotated dataset. This is an opportunity to reflect on and improve the design process for your dataset. The questions for this deliverable are taken directly from the first five sections of a datasheet (Gebru et al. 2021).

Answer each of these questions to the best of your ability. If a question does not apply to your dataset, answer with "Not applicable" followed by a short explanation. Be sure to cite your sources.

Submission: Submit (as a group) a PDF of your completed datasheet to Gradescope by ~~**Friday, November 4th at 11:59pm**~~ **Monday, November 7th at 11:59pm**.

Motivation

Note: If your dataset is *derived* from an existing dataset. Make sure to answer the following questions for **both your dataset and the one it's derived from.**

1. 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 social media platform Reddit largely relies on users and moderators to regulate content, but this can be inconsistent and their tolerance level of various inappropriate contents varies. Moreover, there is no in-platform tool to gauge how family-friendly an individual post is. This is a big gap since Reddit allows users who are between 13 and 18 years old to register, and this is also not strictly enforced. We hope that this dataset could be used to get an idea of whether posts are appropriate or not.

Citations: https://www.redditinc.com/policies/user-agreement-september-12-2021
https://www.reddit.com/r/modnews/comments/i5nc5/moderators_you_can_now_mark_and_unmark_posts_as/

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

This dataset was created by an undergraduate team from University of Massachusetts Amherst, as part of an annotation project for a NLP course.

3. 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.

No funds were received.

4. Any other comments?

We define family-friendliness of text as something that everyone in the family would be comfortable reading or being aware that other family members are reading. To view what makes something family-friendly or not, consult the annotation guidelines of the dataset.

Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, 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 instances of the dataset consists of documents and labels; the documents are Reddit comments with character lengths between 300 and 500, with the links shortened to their domain names, extracted from the list of 200 subreddits that we have compiled. Our list of subreddits were the top 200 subreddits by subs in 2015. We tried to get a more recent list, but no reliable information could be found.

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

1000.

3. 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 of Reddit posts from the 200 subreddit list we compiled, with the latest posts created in December 2021. It is definitely not representative of all Reddit posts, since we picked posts from our list of subreddits and didn't randomly sample from posts across a long period of time. There are a few reasons for our approach. Firstly, it would be unreasonable to represent the entirety of Reddit within our dataset. Additionally, from an annotation perspective, older posts are harder to annotate since cultural values might have been different, and the contexts of the posts are harder to understand, since everything referenced will be much older. Also, because of limitations in manpower and time, we decided to focus on subreddits that are more popular. A model trained on this dataset would do better on more posts because we are sampling from a diverse group of posters.

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

The data consists of raw text from individual posts with the links truncated to domain names. This is for readability as well as keeping the character length consistent. Also, it also enables models to use the domains as features.

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

There is a label for each comment, which is an integer on a Likert scale, where 1 represents completely not family-friendly and 5 is completely family-friendly.

6. 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.

No information is missing.

7. 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.

There are no relations between instances.

8. 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.

None.

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

The posts pulled are unique, and the number of comments from a single thread(subreddit) have been limited to diversify the dataset. We chose posts with character lengths of 300 to 500 because short posts are noisy or, in other words, hard to score because they will be innately missing context.

10. 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.

All data was taken from public forums where users willingly submitted the information publicly.

11. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Yes, while the data presumably abides by Reddit's current content guidelines, there was no effort taken to ensure that the data taken is appropriate for all audiences. In fact, the very purpose of this dataset is to include posts that may not be suitable for all audiences, and score them accordingly.

12. Does the dataset relate to people? If not, you may skip questions 13-15 of this section.

Yes, included are comments from real people on Reddit.

13. 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.

While we do use the author's username during processing to ensure that we limit the number of comments from any one user, this is discarded immediately after processing and is not included in the dataset. As such, all entries have been anonymized unless the post itself contains personally identifiable information.

14. 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.

See above, data is anonymized unless commenters willingly divulge personally identifiable information within their comment.

15. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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.

See above, data is anonymized unless commenters willingly divulge personally identifiable information within their comment. There is no attempt to strip sensitive information from comments if the author willingly includes them, nor is there any attempt to capture this information.

16. Any other comments?

None.

Collection Process

Note: If your dataset is *derived* from an existing dataset. Make sure to answer the following questions for **both your dataset and the one it's derived from.**

1. 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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Directly observable raw text was gathered from comments using the Reddit API.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

Raw text was gathered from comments using the PushshiftAPI wrapped by the PMAW library using our Reddit API token imported using PRAW.

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

We crawled data from our list of 200 popular subreddits. The comments were retrieved as most recent, with a limit on the number of comments queried. As there is no prescriber order in which Reddit comments are made, there is no reason to assume this data to be significantly biased in any way towards family friendliness.

4. 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 acquisition script was written by a student in python for class credit. The Postgres database to store the data and a python script for storing and accessing the data was written by another student for class credit. Open-source libraries were used for each.

5. 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 data was collected in a recent crawl using the pushshiftAPI in November 2022.

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

No. We do not have access to a university review board, nor do we find it particularly pertinent as the data is not personally identifying the authors.

7. Does the dataset relate to people? If not, you may skip questions 8-12 of this section.

Yes, it relates to people.

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

We obtained the data from Reddit.

9. 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 by us, but by Reddit when they agreed to Reddit's user agreement and privacy policy.

10. 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.

Yes. The users agreed to the Reddit user agreement:

> You retain any ownership rights you have in Your Content, but you grant Reddit the following license to use that Content:
>
> When Your Content is created with or submitted to the Services, you grant us a worldwide, royalty-free, perpetual, irrevocable, non-exclusive, transferable, and sublicensable license to use, copy, modify, adapt, prepare derivative works of, distribute, store, perform, and display Your Content and any name, username, voice, or likeness provided in connection with Your Content in all media formats and channels now known or later developed anywhere in the world. This license includes the right for us to make Your Content available for syndication, broadcast, distribution, or publication by other companies, organizations, or individuals who partner with Reddit. You also agree that we may remove metadata associated with Your Content, and you irrevocably waive any claims and assertions of moral rights or attribution with respect to Your Content.

https://www.redditinc.com/policies/user-agreement-september-12-2021

And we in turn use the data in agreement with Reddit API terms:

> d. User Content. Reddit user photos, text and videos ("User Content") are owned by the users and not by Reddit. Subject to the terms and conditions of these Terms, Reddit grants You a non-exclusive, non-transferable, non-sublicensable, and revocable license to copy and display the User Content using the Reddit API through your application, website, or service to end users. You may not modify the User Content except to format it for such display. You will comply with any requirements or restrictions imposed on usage of User Content by their respective owners, which may include "all rights reserved" notices, Creative Commons licenses or other terms and conditions that may be agreed upon between you and the owners.

https://www.reddit.com/wiki/api-terms/

11. 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).

We include an email address to contact us, so anyone can request we remove their content from display.

12. 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.

No. We do not have the resources. Moreover, the data subjects are not identifiable by the data in our set.

13. Any other comments?

We restricted the number of documents from one subreddit in the dataset, as well as the number of posts from a single author.

Preprocessing/Cleaning/Labeling

Note: If your dataset is *derived* from an existing dataset. Make sure to answer the following questions for **both your dataset and the one it's derived from.**

1. 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, process of missing values)? If so, please provide a description.

The only processing done is the extraction of raw text, the cleaning of HTML tags, Unicode emojis, etc. from the raw text using ftfy.fix_text(). Links were truncated to their domains using urlparse and urlextract python packages. There is no original dataset as we are crawling Reddit directly, so this pertains just to our dataset.

2. Any other comments?

None.

Uses

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

Really, this set is just a collection of unrelated Reddit comments. Some effort was made to ensure the independence of the elements in the dataset, so it would not be useful for tasks that try to find some relation between comments. However, it would be useful for tasks that wish to understand something about Reddit comments as a whole. Ideas might be how positive or partisan posts are, identifying sarcasm, etc.

2. 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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

All contextual information about the post is lost. If the author clarified something in the thread, or the response is given with some implicit or explicit understanding by all those in the thread, we fail to capture it. All metadata is also lost during processing, if you are trying to understand the sentiment of Reddit through these posts, you won't have the context of how supported or downvoted the post was. This data was collected from posts in late 2021, and as such, they generally reflect the social values, Reddit rules, and applicable laws of that time. None of the content present here is endorsed by the members of this team. Any unsavory, hateful, harmful, or inappropriate post is included as it is essential to the task we are performing. If these posts are not needed for your application, you may find it appropriate to discard these less savory comments, and not give them yet another platform to reach more people.

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

As mentioned in part 1, there was some effort taken to ensure that these comments are independent. If you're working on a project that involves some relation between the posts, this is likely not a useful dataset.

4. Any other comments?

We chose popular subreddits because the more popular the subreddits that we include in our dataset, the more likely the posts from those subreddits will be fed into our trained model. This should give us better performance on a higher number of posts.