Research:Detox/Data Release

< Research:Detox

All data we have collected and generated for the Wikipedia Detox project is available under free licenses on the Wikipedia Talk Corpus on Figshare (https://figshare.com/projects/Wikipedia_Talk/16731), per our open access policy. There are currently two distinct types of data included:

1. A corpus of all 95 million user and article talk diffs made between 2001–2015 which can be scored by our personal attacks model.
2. An annotated dataset of 1m crowd-sourced annotations that cover 100k talk page diffs (with 10 judgements per diff) for personal attacks, aggression, and toxicity.

For details on data collection methodology and modeling, please refer to our research paper: Ex Machina: Personal Attacks Seen at Scale (https://arxiv.org/abs/1610.08914). For a quick demo of how to use the data for model building and analysis, see the ipython notebook in our Github project (https://github.com/ewulczyn/wiki-detox/blob/master/src/figshare/Wikipedia%20Talk%20Data%20-%20Getting%20Started.ipynb).

Contents

Overview of the Datasets

Wikipedia Comments Corpus

Comments from English Wikipedia talk pages. Comments are grouped into different files by year, and by the user or article talk-namespace. Comments are generated by computing diffs over the full revision history and extracting the content (https://github.com/ewulczyn/wiki-detox/blob/master/src/data_generation/diff_utils.py) added for each revision.

- Figshare dataset (https://figshare.com/articles/Wikipedia_Talk_Corpus/4264973).
- Files: *comments_{ns}_{year}.tar.gz*: Each file contains all comments posted in talk page discussions in {year} and in namespace *{ns} containing at least 3 words and at least 20 characters*. The data for each folder is broken into several files with the following schema.

Annotations Corpora

We have annotated selected fragments of the wikipedia comments corpus for personal attacks, aggression, and toxicity.

For each annotated corpora (one of *attack*, *aggression*, or *toxicity*) there are three files:

- *{attack/aggression/toxicity}_annotated_comments.tsv*: the raw revisions and derived comments that were labelled by crowdworkers.
- *{attack/aggression/toxicity}_annotations.tsv*: the annotations labeled by several crowdworkers for each comment in *{attack/aggression/toxicity}_annotated_comments.tsv*.
- *{attack/aggression/toxicity}_worker_demographics.tsv*: To help understand the generality of the crowd-worker labels, we conducted a survey to get some basic anonymized demographic information about on the crowdworkers who provided the labels.

These files can be joined as follows:

- *{attack/aggression/toxicity}_annotations.tsv* and *{attack/aggression/toxicity}_annotated_comments.tsv* can be joined by **rev_id**.
- *{attack/aggression/toxicity}_annotations.tsv* and *{attack/aggression/toxicity}_worker_demographics.tsv* can be joined by **worker_id**.

Personal Attacks

100k labeled comments from English Wikipedia by approximately of 10 annotators via Crowdflower on whether it contains a personal attack. We also include some demographic data for each crowd-worker.

- The questionnaire (https://github.com/ewulczyn/wiki-detox/blob/master/src/modeling/attack_question.png).
- The Personal Attacks Figshare dataset (https://figshare.com/articles/Wikipedia_Talk_Labels_Personal_Attacks/4054689).

Aggression

100k labeled comments from English Wikipedia by approximately 10 annotators via Crowdflower on how aggressive the comment was perceived to be. We also include some demographic data for each crowd-worker.

- The questionnaire (https://github.com/ewulczyn/wiki-detox/blob/master/src/modeling/aggression_question.png).
- The Aggression Figshare dataset (https://figshare.com/articles/Wikipedia_Talk_Labels_Aggression/4267550).

Toxicity

160k labeled comments from English Wikipedia by approximately 10 annotators via Crowdflower on a spectrum of how toxic the comment is (perceived as likely to make people want to leave the discussion) to how healthy to conversation the contribution is.

- The questionnaire (https://github.com/ewulczyn/wiki-detox/blob/master/src/modeling/toxicity_question.png).
- The Toxicity Figshare dataset (https://figshare.com/articles/Wikipedia_Talk_Labels_Toxicity/4563973).

Schemas for Dataset Files

Schema for *comments_{ns}_{year}.tar.gz*

Wikipedia talk page comments for namespace {ns} and year {year}.

- **rev_id:** MediaWiki revision id of the edit that added the comment to a talk page (i.e. discussion).
- **comment:** Comment text. Consists of the concatenation of content added during a revision/edit of a talk page. MediaWiki markup and HTML have been stripped out. To simplify tsv parsing, \n has been mapped to NEWLINE_TOKEN, \t has been mapped to TAB_TOKEN and " has been mapped to `.

- **raw_comment:** Raw comment text. Consists of the concatenation of raw content added during a revision of a talk page. To simplify tsv parsing, \n has been mapped to NEWLINE_TOKEN, \t has been mapped to TAB_TOKEN and " has been mapped to `.
- **timestamp:** Timestamp in UTC.
- **page_id:** MediaWiki page id of the talk page the comment was made on.
- **page_title:** Title of the talk page the comment was made on.
- **user_id:** MediaWiki user id of the author of the comment. Is always "0" for anonymous contributions.
- **user_text:** Username of the author of the comment. Is an IP in the case of anonymous contributions.
- **bot:** Indicator of whether the comment was made by a bot based on simple heuristics (https://github.com/ewulczyn/wiki-detox/blob/master/src/data_generation/etl.txt#L112-L117).
- **admin:** Indicator of whether the comment serves and administrative purpose based on simple heuristics (https://github.com/ewulczyn/wiki-detox/blob/master/src/data_generation/etl.txt#L118-L149).

Schema for *attack_annotations.tsv*

Personal attack labels from crowd-workers for each comment in *attack_annotated_comments.tsv*. It can be joined with *attack_annotated_comments.tsv* on **rev_id**.

- **rev_id:** MediaWiki revision id of the edit that added the comment to a talk page (i.e. discussion).
- **worker_id:** Anonymized crowd-worker id.
- **quoting_attack:** Indicator for whether the worker thought the comment is quoting or reporting a personal attack that originated in a different comment..
- **recipient_attack:** Indicator for whether the worker thought the comment contains a personal attack directed at the recipient of the comment.
- **third_party_attack:** Indicator for whether the worker thought the comment contains a personal attack directed at a third party.
- **other_attack:** Indicator for whether the worker thought the comment contains a personal attack but is not quoting attack, a recipient attack or third party attack.
- **attack:** Indicator for whether the worker thought the comment contains any form of personal attack. The exact question we posed can be found . The annotation takes on value 0 if the worker selected the option "This is not an attack or harassment" and value 1 otherwise.

Schema for *aggression_annotations.tsv*

Aggression labels from several crowd-workers for each comment in *aggression_annotated_comments.tsv*. It can be joined with *aggression_annotated_comments.tsv* on **rev_id**.

- **rev_id:** MediaWiki revision id of the edit that added the comment to a talk page (i.e. discussion).
- **worker_id:** Anonymized crowd-worker id.
- **aggression_score:** Categorical variable ranging from very aggressive (-2), to neutral (0), to very friendly (2).
- **aggression:** Indicator variable for whether the worker thought the comment has an aggressive tone . The annotation takes on the value 1 if the worker considered the comment aggressive (i.e worker gave an **aggression_score** less than 0) and value 0 if the worker considered the comment neutral or friendly (i.e worker gave an **aggression_score** greater or equal to 0). Takes on values in {0, 1}.

Schema for *toxicity_annotations.tsv*

Toxicity labels from several crowd-workers for each comment in *toxicity_annotated_comments.tsv*. It can be joined with *toxicity_annotated_comments.tsv* on **rev_id**.

- **rev_id:** MediaWiki revision id of the edit that added the comment to a talk page (i.e. discussion).
- **worker_id:** Anonymized crowd-worker id.
- **toxicity_score:** Categorical variable ranging from very toxic (-2), to neutral (0), to very healthy (2).
- **toxicity:** Indicator variable for whether the worker thought the comment is toxic. The annotation takes on the value 1 if the worker considered the comment toxic (i.e worker gave a **toxicity_score** less than 0) and value 0 if the worker considered the comment neutral or healthy (i.e worker gave a **toxicity_score** greater or equal to 0). Takes on values in {0, 1}.

Schema for *{attack/aggression/toxicity}_annotated_comments.tsv*

The comment text and metadata for comments with attack/aggression/toxicity labels generated by crowd-workers. The actual labels are in the corresponding *{attack/aggression/toxicity}_annotations.tsv* since each comment was labeled multiple times.

- **rev_id:** MediaWiki revision id of the edit that added the comment to a talk page (i.e. discussion).
- **comment:** Comment text. Consists of the concatenation of content added during a revision/edit of a talk page. MediaWiki markup and HTML have been stripped out. To simplify tsv parsing, \n has been mapped to NEWLINE_TOKEN, \t has been mapped to TAB_TOKEN and " has been mapped to `.
- **year:** The year the comment was posted in.
- **logged_in:** Indicator for whether the user who made the comment was logged in. Takes on values in {0, 1}.
- **ns:** Namespace of the discussion page the comment was made in. Takes on values in {user, article}.
- **sample:** Indicates whether the comment came via random sampling of all comments, or whether it came from random sampling of the 5 comments around a block event for violating WP:npa or WP:HA. Takes on values in {random, blocked}.
- **split:** For model building in our paper (https://arxiv.org/abs/1610.08914) we split comments into train, dev and test sets. Takes on values in {train, dev, test}.

Schema for *{attack/aggression/toxicity}_worker_demographics.tsv*

Demographic information about the crowdworkers. This information was obtained by an optional demographic survey administered after the labelling task. It is meant to be joined with *{attack/aggression/toxicity}_annotations.tsv* on **worker_id**. Some fields may be blank if left unanswered.

- **worker_id:** Anonymized crowd-worker id.
- **gender:** The gender of the crowd-worker. Takes a value in {'male', 'female', and 'other'}.
- **english_first_language:** Does the crowd-worker describe English as their first language. Takes a value in {0, 1}.
- **age_group:** The age group of the crowd-worker. Takes on values in {'Under 18', '18-30', '30-45', '45-60', 'Over 60'}.
- **education:** The highest education level obtained by the crowd-worker. Takes on values in {'none', 'some', 'hs', 'bachelors', 'masters', 'doctorate', 'professional'}. Here 'none' means no schooling, some means 'some schooling', 'hs' means high school completion, and the remaining terms indicate completion of the corresponding degree type.

License

Citation

This dataset can be cited as:

> Wulczyn, Ellery; Thain, Nithum; Dixon, Lucas (2016): Wikipedia Detox. *figshare.* doi.org/10.6084/m9.figshare.4054689 (https://doi.org/10.6084/m9.figshare.4054689)

Retrieved: 13 00, Oct 31, 2016 (GMT)
