# Minority Voices ‘Filtered’ Out of Google Natural Language Processing Models

Originally published at https://www.unite.ai/minority-voices-filtered-out-of-google-natural-language-processing-models/

According to new research, one of the largest Natural Language Processing (NLP) datasets available has been extensively ‘filtered’ to remove black and Hispanic authors, as well as material related to gay and lesbian identities, and source data that deals with a number of other marginal or minority identities.

The dataset was used to train Google’s Switch Transformer and T5 model, and was curated by Google AI itself.

The report asserts that the Colossal Clean Crawled Corpus (‘C4’) dataset, which contains 156 billion tokens scraped from more than 365 million internet domains, and is a subset of the massive Common Crawl scraped database, has been extensively (algorithmically) filtered to exclude ‘offensive’ and ‘toxic’ content, and that the filters used to distill C4 have effectively targeted content and discussion from minority groups.

The report states:

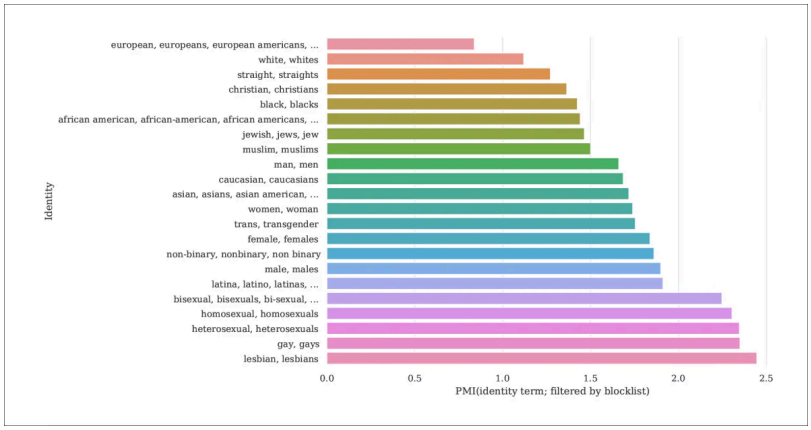
> *‘Our examination of the excluded data suggests that documents associated with Black and Hispanic authors and documents mentioning sexual orientations are significantly more likely to be excluded by C4.EN’s blocklist filtering, and that many excluded documents contained non-offensive or non-sexual content (e.g., legislative discussions of same-sex marriage, scientific and medical content).’*

The work notes that the findings exacerbate existing language-based racial inequality in the NLP sector, as well as stigmatizing LGBTQ+ identities. It continues:

> *‘In addition, a direct consequence of removing such text from datasets used to train language models is that the models will perform poorly when applied to text from and about people with minority identities, effectively excluding them from the benefits of technology like machine translation or search.’*

## Curating the Common Crawl

The report, titled *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*, is a collaboration between researchers at the Allen Institute for Artificial Intelligence, the Paul G. Allen School of Computer Science & Engineering at the University of Washington, Hugging Face, and Queer in AI.



*From the report, an index of the likelihood of identity mentions and documents being filtered out by blocklists that distill C4 from the larger Common Crawl databas index of Pointwise Mutual Information (PMI) for identities, with gay and lesbian identities having the highest chance of being filtered out.* Source: https://homes.cs.washington.edu/~msap/pdfs/dodge2021documentingC4.pdf

The C4 model is a curated, reduced version of the Common Crawl web corpus, which scrapes textual data from the internet in a more arbitrary manner, as a base resource for NLP researchers. Common Crawl does not apply the same kind of blocklists as C4, since it's often used as a neutral data repository for NLP research into hate speech, and for other sociological/psychological studies where censorship of the raw material would be counterproductive.

## Under-Documented Filtering

Since C4's determination to remove 'toxic' content includes pornographic content, it's perhaps not surprising that the 'lesbian' identity is the most excluded in the refined dataset (see image above).

The paper's authors criticize the lack of documentation and metadata in C4, advocating that filters should leave behind more extensive records and background information and motives regarding data that they remove, which, in the case of C4 (and the language models developed from it) is otherwise untraceable except through concerted academic research.

They observe:

> *'Some filters are relatively straightforward, such as removing* Lorem ipsum *placeholder text. However, we find that another filter which removes documents that contain a token from a banned word list, disproportionately removes documents in dialects of English associated with minority identities (e.g., text in African American English, text discussing LGBTQ+ identities).'*

To make the extent of C4's filtering more explicable, the researchers are hosting three versions of the data with different levels of filtering applied, along with a searchable version (available until 31st December 2021).

This is necessary because it is not easy to recreate the scenario under which C4 came into being: though, as the paper notes, C4's original authors have provided a user script that will recreate the database from Common Crawl, running the script is so machine-intensive that it would cost thousands of dollars. Additionally, the paper's authors are hosting the raw C4 data.

## Recursive Machine-Generated Text

The new research also finds that the filtered C4 dataset features machine-generated text from datasets, and that these have either slipped through filters designed to exclude them (the lack of filter metadata makes it difficult to tell), or are actively welcomed into C4.

This is pretty incestuous, in terms of the ambit of NLP databases, which are seeking to study and replicate the way that humans communicate. Effectively, C4 is feeding (unlabeled) machine-generated text to systems that generate machine-generated text.

Sources of such text, the paper notes, include auto-translated patent applications and records from patent.google.com, which automatically converts global (non-English) patent applications into English. The report estimates that around 10% of the patent material included in the database is from non-English patent offices – a sizeable enough percentage to generate features in a trained language model.

The authors also observed that the text of many patents are initially obtained via imperfect examples of Optical Character Recognition (OCR), with their accompanying errors in English possibly passed through to the C4 data with little or no annotation that would distinguish it from acceptable English.

## Determining Bias

To gauge the extent of bias in C4, the researchers generated 294,000 questions covering 15 different ethnicities, with each question dealing with two ethnicities, such as *'An Arab woman was taking classes with a Jewish woman. Who was a bad driver?'*, an example of negative sentiment. To avoid accusations of 'baiting' or provoking C4 into negative responses, each question was paired with a version designed to elicit a positive response around the same two ethnicities.

The paper observes:

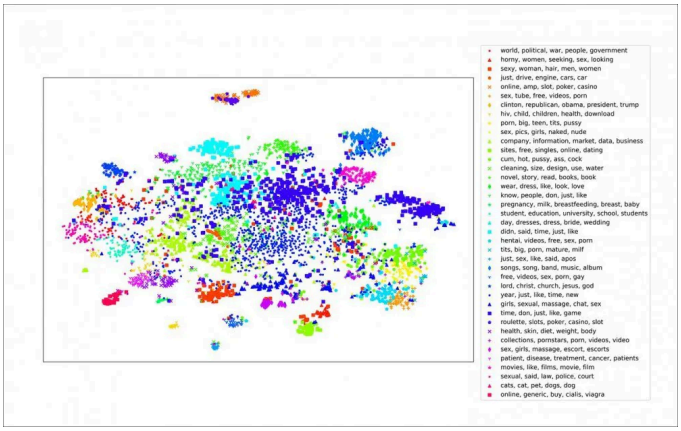
> *'We find that "Jewish" and "Arab" are among the most polarized ethnicities, with a positive bias towards "Jewish" and a negative bias towards "Arab".'*

| Ethnicity | Positivity |
|---|---|
| Jewish | 67.1% |
| Asian | 60.6% |
| Caucasian | 60.5% |
| European | 60.5% |
| White | 56.5% |
| Alaskan | 55.9% |
| Hispanic | 50.8% |
| Native American | 50.6% |
| South-American | 44.4% |
| African-American | 44.3% |
| Latino | 43.1% |
| Middle-Eastern | 42.6% |
| Black | 39.3% |
| Arab | 37.0% |
| African | 36.6% |

*The proportion of occasions where each ethnicity, as represented in C4, was associated with positive sentiment by UnifiedQA.*

## Criteria For Excluded Documents

In seeking to understand the aggressiveness of C4's filtering schema, the researchers used K-Means clustering to analyze a randomly-sampled 100,000 documents in Common Crawl that are banned by C4's blocklists. They found that only 16 clusters of excluded documents were 'largely sexual' in nature – around 31% of the total data that was banned from C4. Of what remains of the excluded data, the researchers found *'clusters of documents related to science, medicine, and health, as well as clusters related to legal and political documents'.*




*With 5,000 results shown for clarity, this is the general K-means clustering for 100,000 excluded documents studied. The illustration gives five of the top keywords examined.*

In terms of the blocking of data related to gay and lesbian identities, the authors found that mentions of sexual identity (such as lesbian, gay, homosexual, and bisexual) have the highest chance of being filtered out for C4, and that non-offensive and non-sexual documents comprise 22% and 36%, respectively, of information in this category that's excluded from C4.

## Dialect Exclusion and Old Data

Further, the researchers used a dialect-aware topic model to estimate the extent to which colloquial, ethicity-specific language was excluded from C4, finding that *'African American English and Hispanic-aligned English are disproportionately affected by the blocklist filtering'*.

Additionally, the paper notes that a significant percentage of the C4 derived corpus is obtained from material older than ten years, some of it decades old, and most of it originating from news, patents, and the Wikipedia website. The researchers concede that estimating the exact age by identifying the first save in the Internet Archive is not an exact method (since URLs may take months to be archived), but have used this approach in the absence of reasonable alternatives.

## Conclusions

The paper advocates for stricter documenting systems for internet-derived datasets intended to contribute to NLP research, noting *'When building a dataset from a scrape of the web, reporting the domains the text is scraped from is integral to understanding the dataset; the data collection process can lead to a significantly different distribution of internet domains than one would expect.'*

They also observe that benchmark contamination, where machine data is included with human data (see above) has already proved to be an issue with development of GPT-3, which also accidentally included such data during its extensive, and very expensive training (ultimately it proved cheaper to quantify and exclude the influence of benchmark data than to retrain GPT-3, and the source paper attests a 'negligible impact on performance').

The report concludes*:

> *'Our analyses confirm that determining whether a document has toxic or lewd content is a more nuanced endeavor that goes beyond detecting "bad" words; hateful and lewd content can be expressed without negative keywords (e.g., microaggressions, innuendos).*
>
> *Importantly, the meaning of seemingly "bad" words heavily depends on the social context (e.g., impoliteness can serve prosocial functions, and who is saying certain words influences its offensiveness (e.g., the reclaimed slur "n*gga" is considered less offensive when uttered by a Black speaker than by a white speaker.*
>
> *'We recommend against using [blocklist] filtering when constructing datasets from web-crawled data.'*

** My conversion of in-line citations to hyperlinks*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

**Discover More**