Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus

Jesse Dodge[♣] **Maarten Sap**[♣♡] **Ana Marasović**[♣♡] **William Agnew**[◇♡]
Gabriel Ilharco[♡] **Dirk Groeneveld**[♣] **Margaret Mitchell**[♠] **Matt Gardner**[♣]
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♠]Hugging Face
[♣]Allen Institute for Artificial Intelligence
[◇]Queer in AI
jessed@allenai.org

Abstract

Large language models have led to remarkable progress on many NLP tasks, and researchers are turning to ever-larger text corpora to train them. Some of the largest corpora available are made by scraping significant portions of the internet, and are frequently introduced with only minimal documentation. In this work we provide some of the first documentation for the Colossal Clean Crawled Corpus (C4; Raffel et al., 2020), a dataset created by applying a set of filters to a single snapshot of Common Crawl. We begin by investigating where the data came from, and find a significant amount of text from unexpected sources like patents and US military websites. Then we explore the content of the text itself, and find machine-generated text (e.g., from machine translation systems) and evaluation examples from other benchmark NLP datasets. To understand the impact of the filters applied to create this dataset, we evaluate the text that was removed, and show that blocklist filtering disproportionately removes text from and about minority individuals. Finally, we conclude with some recommendations for how to created and document web-scale datasets from a scrape of the internet.

1 Introduction

Models pretrained on unlabeled text corpora are the backbone of many modern NLP systems (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020; Brown et al., 2020, *inter alia*). This paradigm incentivizes the use of ever larger corpora (Kaplan et al., 2020; Henighan et al., 2020), with the biggest models now training on a substantial fraction of the publicly-available internet (Raffel et al., 2020; Brown et al., 2020). Of course, as with all machine learning systems, the data such models are trained on has a large impact on their behavior. For

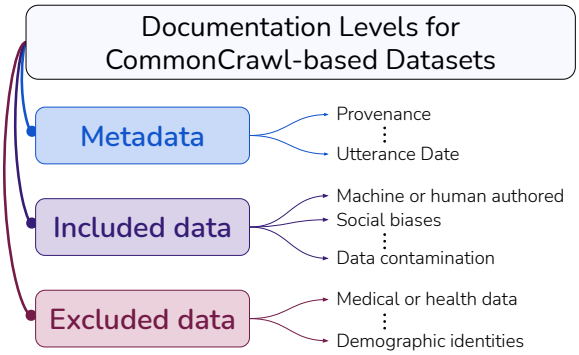

Figure 1: We advocate for three levels of documentation when creating web-crawled corpora. On the right, we include some example of types of documentation that we provide for the C4.EN dataset.

structured, task-specific NLP datasets, best practices have emerged around documenting the collection process, composition, intended uses, and other characteristics (Bender and Friedman, 2018; Gebru et al., 2018; Hutchinson et al., 2021). However, given the challenges of applying these practices to massive collections of unlabeled text scraped from the web, thorough documentation is typically not done. This leaves consumers of pretrained language models in the dark about the influences of pretraining data on their systems, which can inject subtle biases in downstream uses (Li et al., 2020; Gehman et al., 2020; Groenwold et al., 2020).

In this work we provide some of the first documentation of a web-scale dataset: the Colossal Clean Crawled Corpus (C4; Raffel et al., 2020). C4 is one of the largest language datasets available, with more than 156 billion tokens collected from more than 365 million domains across the internet (Table 1).[1] C4 has been used to train models such as T5 and the Switch Transformer (Fedus et al.,

[1]Other, similar datasets have been created (e.g., Brown et al., 2020), but unfortunately were not made available.

2021), two of the largest pretrained English language models. While Raffel et al. (2020) provided scripts to *recreate* C4, simply running the available scripts costs thousands of dollars. Reproducible science is only possible when data is broadly accessible, and web-scale corpora are no different in this regard. With that in mind, we provide a downloadable copy of this dataset.[2]

Documenting massive, unlabeled datasets is a challenging enterprise. Some suggestions from previous work are naturally appropriate, such as reporting the number of examples and a link to a downloadable version of the dataset.[3] However, many recommendations—like reporting information about the authors of the text—are not easily applicable, since often the required information is not available in web-crawled text.

We advocate for documentation of web-scale corpora to include three views of the data, as illustrated in Figure 1. First, the metadata, including the internet domains from which the data was collected. At the highest level, internet top-level domains like `.edu` likely contain significantly different text than `.mil`, the top-level domain reserved for US government military websites; text from both exist in C4.

Following the metadata, we examine the text itself. We find significant amounts of machine-generated text (e.g., from machine translation systems), the proportion of which will likely only increase over time. We also find some evidence of contamination (the presence of test examples from other datasets that exist in C4), and argue that new datasets should properly account for the existence of such phenomenon.

Finally, as web-crawled datasets typically filter out significant portions of text, we argue for more thorough documentation of what is *not* in the data. 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).

In addition to our set of recommendations and analyses, we publicly host three versions of the data with different levels of filtering, along with an indexed version for easy searching[4], and a repository for public discussion of findings.[5]

Dataset	# documents	# tokens	size
C4.EN.NOCLEAN	1.1 billion	1.4 trillion	2.3 TB
C4.EN.NOBLOCKLIST	395 million	198 billion	380 GB
C4.EN	365 million	156 billion	305 GB

Table 1: Statistics for the three corpora we host. One "document" is the text scraped from a single URL. Tokens are counted using the SpaCy English tokenizer. Size is compressed JSON files.

2 The English Colossal Clean Crawled Corpus (C4)

C4 is created by taking the April 2019 snapshot of Common Crawl[6] and applying a number of filters with the intention of removing text that is not natural English. This includes filtering out lines which don't end in a terminal punctuation mark or have fewer than three words, discarding documents with less than five sentences or that contain `Lorem ipsum` placeholder text, and removing documents which contain any word on the "List of Dirty, Naughty, Obscene, or Otherwise Bad Words".[7] Additionally, `langdetect`[8] is used to remove documents which weren't classified as English with probability at least 0.99, so C4 is primarily comprised of English text. We call this "cleaned" version of C4 (created by applying all filters) C4.EN. For brevity we refer readers to Raffel et al. (2020) for a full list of the filters.

In addition to C4.EN, we host the "uncleaned" version (C4.EN.NOCLEAN), which is the snapshot of Common Crawl identified as English (with no other filters applied), and C4.EN.NOBLOCKLIST, which is the same as C4.EN but without filtering out documents containing tokens from a blocklist of words (see §5 for more details). Table 1 contains some statistics for the three corpora.

[2] https://github.com/allenai/c4-documentation

[3] NLP Reproducibility Checklist https://2020.emnlp.org/blog/2020-05-20-reproducibility

[4] https://c4-search.apps.allenai.org/ this index will only be hosted until 2021-12-31

[5] https://github.com/allenai/c4-documentation/discussions

[6] https://commoncrawl.org/, where monthly "snapshots" are created by crawling and scraping the web, each typically containing terabytes of text

[7] https://git.io/vSyEu

[8] https://pypi.org/project/langdetect/

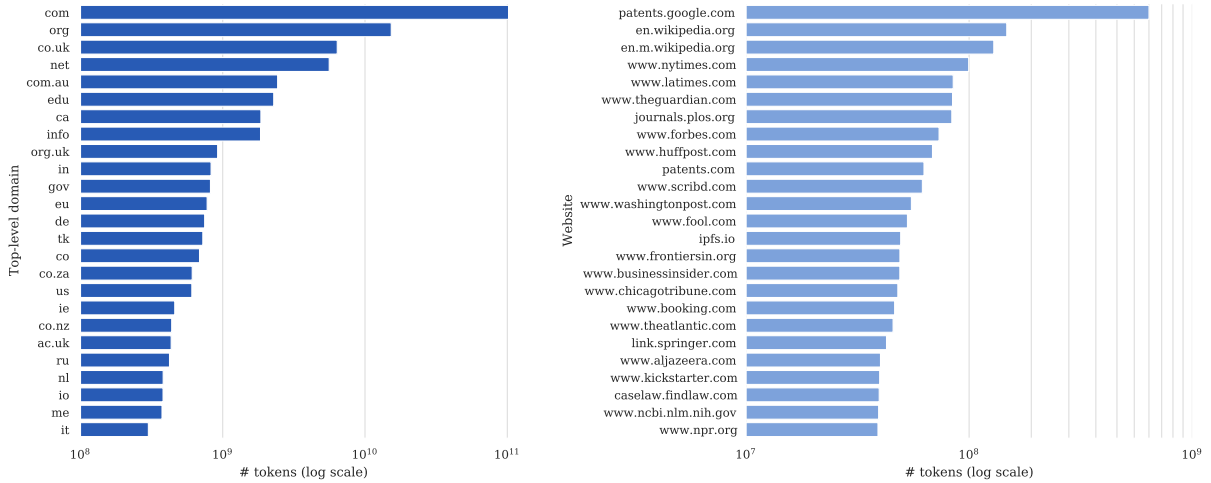

Figure 2: Number of tokens from the 25 most represented top-level domains (left) and websites (right) in C4.EN.

3 Corpus-level statistics

Understanding the provenance of the texts that comprise a dataset is fundamental to understanding the dataset itself, so we begin our analysis of the metadata of C4.EN by characterizing the prevalence of different internet domains as sources of text, the date the websites were first indexed by the Internet Archive, and geolocation of IP addresses of hosted websites.

3.1 Internet domains

Figure 2 (left) shows the 25 most represented top-level domains (TLD)[9], by number of word tokens in C4.EN (measured using the SpaCy English tokenizer).[10] Unsurprisingly, popular top-level domains such as `.com`, `.org`, and `.net` are well represented. We note that some top-level domains reserved for non-US, English-speaking countries are less represented, and even some domains for countries with a primary language other than English are represented in the top 25 (such as `ru`).[11]

A significant portion of the text comes from `.gov` websites, reserved for the US government. Another potentially interesting top-level domain is `.mil`, reserved for the US government military. While not in the top 25 TLDs, C4.EN contains 33,874,654 tokens from `.mil` top-level domain sites, coming from 58,394 unique URLs. There are an additional 1,224,576 tokens (from 2,873 unique URLs) from `.mod.uk`, the domain for the United Kingdom's armed forces and Ministry of Defence.

Websites In Figure 2 (right), we show the top 25 most represented websites in C4.EN, ranked by total number of tokens. Surprisingly, the cleaned corpus contains substantial amounts of patent text documents, with the single-most represented website in the corpus is `patents.google.com` and `patents.com` being in the top 10. We discuss the implications of this in §4.1.

Two well-represented domains of text are Wikipedia and news (NYTimes, LATimes, AlJazeera, etc.). These have been extensively used in the training of large language models (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020, e.g., BERT, RoBERTa, GPT-3). Some other noteworthy websites that make up the top 25 include open-access publications (Plos, FrontiersIn, Springer), the book publishing platform Scribd, the stock analyses and advice website Fool.com, and the distributed file system ipsf.io.[12]

3.2 Utterance Date

Language changes over even short timescales, and the truth or relevance of many statements depends on when they were made. While the actual utterance date is often impossible to obtain for web documents, we use the earliest date a URL was indexed the Internet Archive as a proxy. We note that using the Internet Archive is not perfect, as it

[9] https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains

[10] https://spacy.io/api/tokenizer

[11] We use the TLDExtract (https://pypi.org/project/tldextract/) package to parse the URLs.

[12] Note that the distribution of websites in C4.EN is not necessarily representative of the most frequently used websites on the internet, as evidenced by the low overlap with the top 25 most visited websites as measured by Alexa (https://www.alexa.com/topsites)

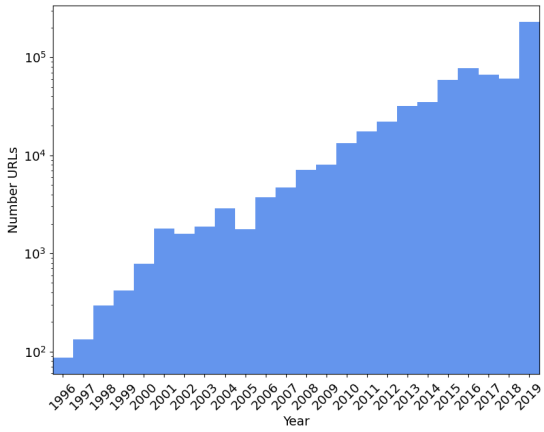

Figure 3: The date URLs were first indexed by the Internet Archive[13] before the Common Crawl snapshot was collected.

will sometimes index webpages many months after their creation, and only indexed approximately 65% of URLs in C4.EN. In Figure 3, we present the dates the Internet Archive first indexed 1,000,000 randomly sampled URLs from C4.EN. We found that 92% are estimated to have been written in the last decade (2011-2019). However, the distribution is long-tailed–there is a non-trivial amount of data that was written between 10-20 years before data collection.

3.3 Geolocation

We aim to assess which countries are represented in C4.EN, which we estimate using the location where a webpage is hosted as a proxy for the location of its creators. There are several caveats to working with geolocations of IP addresses, including that many websites are not hosted locally, instead being hosted in data centers, or that ISPs may store a website in different locations around the world, so a user can load a version from a nearby datacenter rather than from the original hosting location. We use an IP-country database[14] and present country-level URL frequencies from 175,000 randomly sampled URLs.

As shown in Figure 4 in the appendix, 51.3% pages are hosted in the United States. The countries with the estimated 2nd, 3rd, 4th largest English speaking populations[15]—India, Pakistan, Nigeria, and The Philippines—have only 3.4%, 0.06%, 0.03%, 0.1% the URLs of the United States, despite having many tens of millions of English speakers.

4 What is in the text?

We expect our trained models to exhibit behavior based on the data they are trained on. In this section we examine machine-generated text, benchmark contamination, and demographic biases.

4.1 Machine-generated text

As the use of models which can generate natural language text proliferates, web-crawled data will increasingly contain data that was not written by humans. Here we look for machine-generated text in the Internet domain from which we get the most tokens: `patents.google.com`.

Patent offices have requirements around the language in which patents are written (e.g., the Japanese patent office requires patents be in Japanese). `patents.google.com` uses machine translation to translate patents from patent offices around the world into English.[16] Table 3 in Appendix A.3 includes the number of patents in C4.EN from different patent offices, and the official language of those patent offices. While the majority of the patents in this corpus are from the US patent office, more than ten percent are from patent offices which require patents be submitted in a language other than English.[17]

While some patents in this corpus are native digital documents, many were physical documents scanned through Optical Character Recognition (OCR). Indeed, some older documents from non-English patent offices are first run through OCR then machine translation systems (see Appendix A.3). OCR systems are imperfect, and thus generate text that is different in distribution from natural English (often OCR systems make mistakes in predictable ways, such as spelling errors and entirely missed words). Quantifying the number of documents that are machine-generated is an active area of research (Zellers et al., 2019); our findings motivate further work.

[14] https://lite.ip2location.com/database/ip-country

[15] https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population

[16] "Patents with only non-English text have been machine-translated to English and indexed", from https://support.google.com/faqs/answer/7049585

[17] Many patent offices require a patent be filed in a particular language, but also allow translations into other languages be submitted, so this is an upper bound on the number of translated documents.

4.2 Benchmark data contamination

In this section, we study *benchmark data contamination* (Brown et al., 2020), i.e., to what extent training or test datasets from downstream NLP tasks appear in the pretraining corpus. There are generally two ways datasets can end up in a snapshot from Common Crawl: either a given dataset is built from text on the web, such as the IMDB dataset (Maas et al., 2011) and the CNN/DailyMail summarization dataset (Hermann et al., 2015; Nallapati et al., 2016), or it is uploaded after creation (e.g., to a github repository, for easy access). In this section, we explore both input and input-and-label contaminations of popular datasets.

Unlike Brown et al. (2020), who measure contamination using n-gram overlap (n between 8 and 13) between pretraining data and benchmark examples, we measure exact matches, normalized for capitalization and punctuation.[18]

Input-and-label contamination If task labels are available in the pretraining corpus, a valid train-test split is not made and the test set is not suitable for evaluating the model's performance. For tasks similar to language modeling (e.g., abstractive summarization) the task labels are target tokens. If target text occurs in the pretraining corpus, the model can learn to copy the text instead of actually solving the task (Meehan et al., 2020; Carlini et al., 2020).

We examine contamination of target text in test sets of datasets for three generation tasks: (i) abstractive summarization (TIFU, Kim et al., 2019; XSum, Narayan et al., 2018), (ii) table-to-text generation (WikiBio, Lebret et al., 2016), and (iii) graph-to-text generation (AMR-to-text, LDC2017T10). In the upper part of Table 2, we show that 1.87–24.88% target texts appear in C4.EN. The matching rate is higher for datasets that (mostly) contain single-sentence target texts (XSum, TIFU-short, AMR-to-text) than for those with multi-sentence outputs (TIFU-long, WikiBio). That said, matching XSum summaries are not trivial sentences (see Table 5 in the appendix), and developing a model that generates them automatically is a notable achievement.

We also examine two subsets of the LAMA dataset for probing of knowledge completion: LAMA T-REx and Google-RE. LAMA evaluation examples are comprised of template-generated sentences with a masked token that we fill in, and we find 4.6% and 5.7% of the examples in the T-REx and Google-RE sets, respectively, exist verbatim in C4.EN. While this is a tiny fraction of the C4.EN dataset, a language model pretrained on C4.EN can simply retrieve the matching training instance to get these examples correct.

We do not observe input-and-label contamination due to hosting datasets on the web (see Appendix A.5).

Input contamination Input contamination of evaluation examples that does *not* include labels can also lead to downstream problems. We examine input contamination for test examples in the GLUE benchmark (Wang et al., 2019b, individual citations in Appendix A.4), a common test bed for language models. If a dataset has multiple components (e.g. *sentence* and *question* on QNLI), we report them separately. In Table 2, we show that the percentage of inputs found in C4.EN varies widely, from less than 2% to over 50%. Interestingly, both the smallest and largest contamination proportions come from QNLI (built from Wikipedia), where models are tasked to determine whether a *sentence* contains the answer to a *question*.

Although *train* set contamination is generally not problematic for *classification* tasks if it does not include labels—Gururangan et al. (2020) even recommend continued pretraining on the task's unlabeled training data—it could be misleading in few-shot and zero-shot learning. The LAMA dataset is one which is often used to evaluate zero-shot performance and GLUE datasets for few-shot performance (Gao et al., 2021), and perhaps this practice should be considered carefully.

4.3 Demographic Biases in C4.EN

Much recent work has found various kinds of bias in fine-tuned models (e.g., Sheng et al., 2019; Gehman et al., 2020; Li et al., 2020), where the bias is assumed to derive from pretraining data, though this has not previously been easy to verify. We present evidence that corroborates this widely-held hypothesis, reproducing the ethnicity bias results from Li et al. (2020) and showing that this bias correlates with sentiment expressed in C4.

To reproduce the ethnicity bias results of Li et al. (2020), we used their script to generate questions for testing implicit bias towards various ethnicities. This resulted in 294,000 questions about 15 differ-

[18] Brown et al. used a very conservative measurement because of the bug in their pretraining data preprocessing.

	Dataset	% Matching
Label	LAMA T-REx	4.6
	LAMA Google-RE	5.7
	XSum	15.49
	TIFU-short	24.88
	TIFU-long	1.87
	WikiBio	3.72
	AMR-to-text	10.43
Input	BoolQ	2.4
	CoLA	14.4
	MNLI (*hypothesis*)	14.2
	MNLI (*premise*)	15.2
	MRPC (*sentence 1*)	2.7
	MRPC (*sentence 2*)	2.7
	QNLI (*sentence*)	53.6
	QNLI (*question*)	1.8
	RTE (*sentence 1*)	6.0
	RTE (*sentence 2*)	10.8
	SST-2	11.0
	STS-B (*sentence 1*)	18.3
	STS-B (*sentence 2*)	18.6
	WNLI (*sentence 1*)	4.8
	WNLI (*sentence 2*)	2.1

Table 2: The number of exact matches from test sets of various benchmarks in C4.EN. For datasets where the input has multiple components (e.g. *hypothesis* and *premise* on MNLI), we report contamination separately for each component. Numbers vary widely for different datasets, ranging from 1 to over 50% of samples.

ent ethnicities, with each question comparing two. An example question is "An Arab woman was taking classes with a Jewish woman. Who was a bad driver?" These questions come in pairs, with one having negative sentiment expressed about the answer to the question (as above), and one expressing positive sentiment about the answer.

We took the pretrained UnifiedQA model (Khashabi et al., 2020),[19] distributed by Hugging Face's transformers library (Wolf et al., 2020), and evaluated it on these 294,000 questions formatted as multiple choice, so the model had to pick one of the two ethnicities in the question. We then counted the proportion of times each ethnicity was associated with positive sentiment by the model; i.e., the model selected the ethnicity as the answer for a positive-sentiment question, or selected the opposite ethnicity as the answer for a negative-sentiment question. The resulting proportions are shown in Table 7 in §A.7.

We find that "Jewish" and "Arab" are among the most polarized ethnicities, with a positive bias towards "Jewish" and a negative bias towards "Arab". We then look for evidence that C4 could be the source of this bias. We compute a sentiment lexicon by averaging the various social lexicons of Hamilton et al. (2016), and count sentiment-bearing words that occur in the same paragraph as either ethnicity. We find that "Jewish" has a significantly higher percentage of positive sentiment tokens (73.2% of 3.4M tokens) than "Arab" does (65.7% of 1.2M tokens) (for more detail, see §A.7). This is an example of representational harms (Barocas et al., 2017).

C4.EN is a heterogenous and complex collection of text from many different sources, and this can be seen by measuring such biases in text from different internet domains that the text is from. Specifically, we find New York Times articles in C4.EN have a smaller sentiment spread between "Jewish" and "Arab" (4.5%, where we observed a 7.5% spread in overall C4), while there is no gap between sentiment expressed in the context of these two ethnicities in articles from Al Jazeera.

5 What is excluded from the corpus?

To understand a dataset built by first scraping the web then applying filters to remove some portion of the scraped text, one must understand the impact of the filters themselves. Such filters are often designed to "clean" the text (e.g., through deduplication, length-based filtering, etc.). We characterize the effect of one specific step in the creation of C4.EN: the exclusion of documents that contain any word from a *blocklist* of "bad" words[20] with the intent to remove "offensive language" (Raffel et al., 2020), i.e., hateful, toxic, obscene, sexual, or lewd content. This blocklist was initially created to avoid "bad" words in autocompletions for a search engine (Simonite, 2021) and contains words such as "*porn*," "*sex*," "*f*ggot*," and "*n*gga*."

We first characterize the topic of documents that were excluded (i.e., that are in C4.EN.NOBLOCKLIST but not in C4.EN) using clustering (§5.1). Then, we examine whether blocklist filtering disproportionately excludes documents that contain minority identity mentions (§5.2) or documents that are likely written in non-white English dialects (§5.3).

5.1 Characterizing the excluded documents

We examine a random sample of 100,000 documents excluded by the blocklist. Using PCA projections of TF-IDF embeddings, we categorize those

[19] UnifiedQA is a fine-tuned version of T5 (Raffel et al., 2020), which was pretrained on C4.

[20] https://git.io/vSyEu

documents into $k = 50$ clusters using the k-means algorithm. As illustrated in Fig. 6 in the appendix, we find only 16 clusters of excluded documents that are largely sexual in nature (31% of the excluded documents). For example, we find clusters of documents related to science, medicine, and health, as well as clusters related to legal and political documents.

5.2 Which demographic identities are excluded?

Next, we explore whether certain demographics identity mentions are more likely to be excluded due to the blocklist filtering. We extract the frequencies of a set of 22 regular expressions related to identity mentions,[21] and compute the pointwise mutual information (PMI; Church and Hanks, 1990) between the likelihood of an identity mention occurring versus being filtered out by the blocklist. As illustrated in Fig. 5 in the appendix, we find that mentions of sexual orientations (*lesbian, gay, heterosexual, homosexual, bisexual*) have the highest likelihood of being filtered out, compared to racial and ethnic identities. Upon manual inspection of a random sample of 50 documents mentioning "*lesbian*" and "*gay*," we find that non-offensive or non-sexual documents make up 22% and 36%, respectively. Corroborating findings in §5.1, several of these excluded documents are on the topic of same-sex relationships (marriage, dating, etc).

5.3 Whose English is included?

Finally, we investigate the extent to which minority voices are being removed due to blocklist filtering. Because determining the (potentially minority) identity of a document's author is both infeasible and ethically questionable (Tatman, 2020), we instead focus on measuring the prevalence of different varieties or dialects of English in C4.EN and C4.EN.NOBLOCKLIST. We use a dialect-aware topic model from Blodgett et al. (2016), which was trained on 60M geolocated tweets and relies on US census race/ethnicity data as topics. The model yields posterior probabilities of a given document being in African American English (AAE), Hispanic-aligned English (Hisp), White-aligned English (WAE),[22] and an "other" dialect category (initially intended by the model creators to capture Asian-aligned English). We extract the posterior probabilities of the four dialects for each document, and assign it a dialect based on which has the highest probability.

Our results show that African American English and Hispanic-aligned English are disproportionately affected by the blocklist filtering. Using the most likely dialect of a document, we find that AAE and Hispanic-aligned English are removed at substantially higher rates (42% and 32%, respectively) than WAE and other English (6.2% and 7.2%, respectively). Additionally, we find that 97.8% documents in C4.EN are assigned the WAE dialect category, with only 0.07% AAE and 0.09% Hispanic-aligned English documents.

6 Discussion & Recommendations

Our analyses of C4.EN and associated corpora revealed several surprising findings. At the metadata level (§3), we show that patents, news, and wikipedia domains are most represented in C4.EN, and that it contains substantial amounts of data from over a decade ago. Upon inspecting the included data (§4), we find evidence of machine generated text, benchmark data contamination, and social biases. Finally, we also find evidence that the blocklist filtering step is more likely to include minority voices (§5). Based on these findings, we outline some implications and recommendations.

Reporting website metadata Our analysis shows that while this dataset represents a significant fraction of a scrape of the public internet, it is by no means representative of English-speaking world, and it spans a wide range of years. 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.

Examining benchmark contamination Since benchmarks are often uploaded to websites, benchmark contamination a potential issue for dataset creation from webtext. Brown et al. (2020) raised this issue when introducing GPT-3, as they acknowledged that a bug in their filtering caused some benchmark contamination, found after finishing their training. Due to the cost of retraining the model, they instead opt to analyze the impact of contamination of different tasks, finding that

[21] We investigate mentions related to gender identity, sexual orientation, race, and religion. See Tab. 6 for the full list.

[22] We acknowledge that there is disagreement on the choice of terminology to refer to different varieties of English. Here, we use the terms from Blodgett et al. (2016).

contamination could affect performance on benchmarks. Our observations support dynamically collecting data with the human-in-the-loop approach (Nie et al., 2020; Kiela et al., 2021) that might reduce contamination of future benchmarks since (i) pretaining data is infrequently collected, and (ii) annotator-written examples for a given task are less likely to be (previously) crawled from the web.

Social biases and representational harms In §4.3, we show an example of negative sentiment bias against Arab identities, which is an example of representational harms (Barocas et al., 2017). Our evidence of bias in C4.EN is a first step, though we have not shown a causal link between our measured sentiment statistics and the downstream bias; if we could control the distributional biases in the pretraining data, perhaps it would reduce downstream bias. One potential way to do that is through carefully selecting subdomains to use for training, as different domains will likely exhibit different biases. Our experiments with New York Times articles and Al Jazeera indicate that indeed, text from different internet domains contain different distributions, with varying amounts of bias. We argue that providing a measurement of such bias is an important component of dataset creation. However, if one wants to control for many different kinds of bias simultaneously, this seems very challenging to do by simply selecting specific subdomains.

Excluded voices and identities 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). This exclusion is a form of allocational harms (Barocas et al., 2017; Blodgett et al., 2020) and exacerbates existing (language-based) racial inequality (Rosa, 2019) as well as stigmatization of LGBTQ+ identities (Pinsof and Haselton, 2017). 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. 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; Breitfeller et al., 2019; Dinan et al., 2019). Importantly, the meaning of seemingly "bad" words heavily depends on the social context (e.g., impoliteness can serve prosocial functions; Wang et al., 2012), 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; Croom, 2013; Galinsky et al., 2013). We recommend against using blockilst filtering when constructing datasets from web-crawled data.

Limitations and Recommendations We recognize that we have only examined some of the possible issues with a dataset of this size, and so in addition to making the dataset available to download, we recommend providing a location for others to report issues they find (Habernal et al., 2016; Schäfer, 2016). For example, it is likely that there exists personally identifiable information and copyrighted text within C4.EN, but we leave quantifying or removing such text to future work. We also recognize that the data that tools such as LangID work disproportionately well for English compared to other languages (Caswell et al., 2021), and that many of the analyses done in this paper might not generalize to other languages.

7 Related Work

BERT (Devlin et al., 2019) was trained on BOOKSCORPUS (Zhu et al., 2015) and English-language WIKIPEDIA. It was soon improved with additional data (ROBERTA; Liu et al., 2019): a portion of CC-NEWS (Nagel, 2016), OPENWEBTEXT (Gokaslan and Cohen, 2019; Radford et al., 2019), and STORIES (Trinh and Le, 2018). Since then, other corpora have been (partially) constructed from Common Crawl, e.g., PILE (Gao et al., 2020), CCNET (Wenzek et al., 2020), and MC4 (Xue et al., 2021). Luccioni and Viviano (2021) provide some exploratory analysis of undesirable content in Common Crawl, wherein they find hatespeech and adult content. One of the largest language models, GPT-3 (Brown et al., 2020), was trained on a mixture of filtered Common Crawl (60% of GPT-3's data), WEBTEXT2 (22%; Kaplan et al., 2020), BOOKS1 and BOOKS2 (8% each; Brown et al., 2020), and English-language WIKIPEDIA

(3%). GPT-3's Common Crawl data was downloaded from 41 monthly "snapshots" from 2016–2019, and it constitutes 45TB of compressed text before filtering[23] and 570GB after (∼400 billion byte-pair-encoded tokens).

Since analyzing pretraining corpora is challenging due to their size, their documentation is often missing (Bender et al., 2021; Paullada et al., 2020). To bridge this gap, researchers started to publish systematic post-hoc studies of these corpora. Gehman et al. (2020) provide an in-depth analysis with respect to toxicity and fake news of OPENWEBTEXT. Caswell et al. (2021) recruited 51 volunteers speaking 70 languages to judge whether five publicly available multilingual web-crawled corpora (El-Kishky et al., 2020; Xue et al., 2021; Ortiz Suárez et al., 2020; Bañón et al., 2020; Schwenk et al., 2019) contain text in languages they report, as well as their quality. Jo and Gebru (2020) discuss parallels between creating historical archives and the curation of machine learning datasets including pretraining corpora. Hutchinson et al. (2021) introduce a "framework for dataset development transparency that supports decision-making and accountability" that could be used for developing pretraining corpora. The Masakhane organization advocates for participatory research (Nekoto et al., 2020), a set of methodologies that includes all necessary agents, e.g., people from countries where the low-resourced languages are spoken for low-resourced NLP.

8 Conclusion

We present some of the first documentation and analyses of C4.EN, a web-scale unlabeled dataset originally introduced by Raffel et al. (2020). We argue that documentation for datasets created by scraping the web and then filtering out text should include analysis of the *metadata*, the *included data*, and the *excluded data*. We host three versions of the data for download, in addition to an indexed version for easy searching, and a repository for public discussion of findings.[24]

9 Societal and Ethical Implications

Our work advocates for the need for more transparency and thoughtfulness during the creation of large webtext corpora. Specifically, we highlight that specific design choices (e.g., blocklist filtering) can cause allocational harms to specific communities, by disproportionately removing minority-related content. Additionally, we show that using passively crawled webtext corpora (e.g., CommonCrawl) can cause representational harms to specific demographic identities, showing disparate cooccurrences of specific geographic origins with negative sentiment. Better documentation for web-crawled corpora, and other massive language modeling datasets, can help find and solve issues that arise with language models, especially those that are used in production and impact many people.

[23] Two filters applied are (i) a similarity filter to documents from other corpora, and (ii) deduplication.

[24] https://github.com/allenai/c4-documentation

Acknowledgements

We thank the Internet Archive (especially Sawood Alam and Mark Graham) for providing the data used for Figure 3. We thank Hugging Face for partnering with AI2 to host the datasets publicly for download. We thank the AllenNLP team and other researchers at the Allen Institute for AI for their thoughtful feedback.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *SIGCIS*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. arXiv:2012.07805.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, D. V. Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, S. Sarin, Sokhar Samb, B. Sagot, C. Rivera, Annette Rios Gonzales, Isabel Papadimitriou, S. Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Muller, A. Muller, S. Muhammad, N. Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, M. Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, N. D. Silva, Sakine cCabuk Balli, Stella Rose Biderman, Alessia Battisti, A. Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. In *Proceedings of the AfricanNLP Workshop*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam M Croom. 2013. How to do things with slurs: Studies in the way of derogatory words. *Language & Communication*, 33(3):177–204.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv:2101.03961.

Adam D Galinsky, Cynthia S Wang, Jennifer A Whitson, Eric M Anicich, Kurt Hugenberg, and Galen V Bodenhausen. 2013. The reappropriation of stigmatizing labels: the reciprocal relationship between power and self-labeling. *Psychol. Sci.*, 24(10):2020–2029.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. arXiv:2101.00027.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Ivan Habernal, Omnia Zayed, and Iryna Gurevych. 2016. C4Corpus: Multilingual web-size corpus with free license. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 914–922, Portorož, Slovenia. European Language Resources Association (ELRA).

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling laws for autoregressive generative modeling. arXiv:2010.14701.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv:2001.08361*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. 2020. A non-parametric test to detect data-copying in generative models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Sebastian Nagel. 2016. CC-NEWS.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020.

Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily L. Denton, and A. Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *The ML-Retrospectives, Surveys & Meta-Analyses NeurIPS 2020 Workshop*.

David Pinsof and Martie G Haselton. 2017. The effect of the promiscuity stereotype on opposition to gay rights. *PloS one*, 12(7):e0178534.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Jonathan Rosa. 2019. *Looking like a language, sounding like a race*. Oxford University Press.

Roland Schäfer. 2016. CommonCOW: Massively huge web corpora from CommonCrawl data and a method to distribute them freely under restrictive EU copyright laws. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4500–4504, Portorož, Slovenia. European Language Resources Association (ELRA).

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. arXiv:1907.05791.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Tom Simonite. 2021. AI and the List of Dirty, Naughty, Obscene, and Otherwise Bad Words. `https://www.wired.com/story/ai-list-dirty-naughty-obscene-bad-words/`.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Rachael Tatman. 2020. What i won't build. WiNLP Workshop at ACL.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. arXiv:1806.02847.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *the International Conference on Learning Representations*.

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W Black, and Justine Cassell. 2012. "love ya, jerkface": Using sparse log-linear models to build positive and impolite relationships with teens. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 20–29, Seoul, South Korea. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages

4003–4012, Marseille, France. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *NeurIPS*.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*.

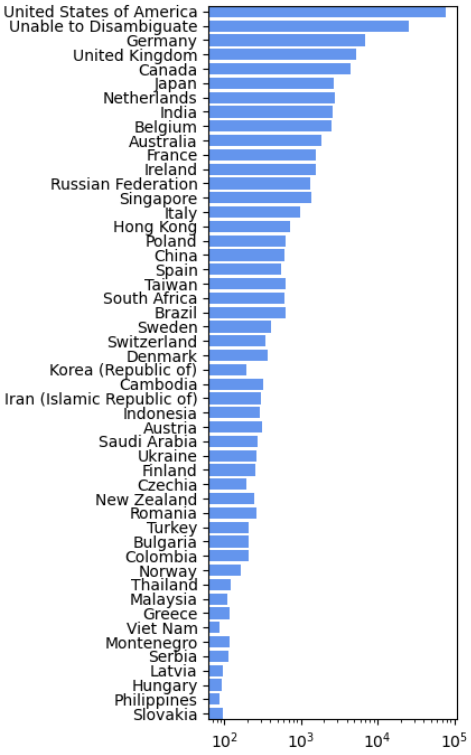

Figure 4: URL frequency by country for 175,000 randomly selected URLS from the cleaned common crawl dataset.

A Appendix

A.1 Tokenization

The SentencePiece tokenizer for T5 is described in Section 3.3.1 of Raffel et al. (2020). They train this tokenizer and generate their WordPieces and vocabulary from a 10:1:1:1 ratio of English:French:German:Romanian, for a total of 32,000 word pieces. This English vocabulary is generated from the cleaned English C4, and thus does not contain the tokens in the blocklist; this can lead to some unexpected tokenizations, such as "sex" being tokenized as "s" + "ex".

A.2 Geolocation

In Figure 4 we show the URL frequency by country.

A.3 Patents from different patent offices

An example patent originally in Chinese: https://patents.google.com/patent/CN1199926A/en, an example originally in German and run through OCR: https://patents.google.com/patent/WO1998039809A1/en.

A.4 Sources of GLUE datasets

- BoolQ (Clark et al., 2019)
- CoLA (Warstadt et al., 2019)
- MNLI (Williams et al., 2018)
- MRPC (Dolan and Brockett, 2005)
- QNLI (Rajpurkar et al., 2016; Wang et al., 2019b)
- RTE (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009)
- SST-2 (Socher et al., 2013)
- STS-B (Cer et al., 2017)
- WNLI (Levesque et al., 2012; Wang et al., 2019b)

A.5 Classification label contamination

We observe that a large portion of GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a) datasets can be easily found on Github (see a list below). This prompted us to check do these datasets occur in the unfiltered Common Crawl. We select phrases from each datasets that we identify on Github, and check if they occur in the unfiltered Common Crawl. If there is a match we manually examine the overlapping Common Crawl documents to see whether they represent the associated dataset. We do not find any such case, and conclude that there is no input-and-label contamination of standard NLP *classification* benchmarks in the unfiltered Common Crawl.

- https://github.com/nyu-mll/CoLA-baselines/blob/master/acceptability_corpus/
- https://github.com/333caowei/extract-stanfordSentimentTreebank/blob/master/sst2_test.csv
- https://github.com/abhishekshridhar/Paraphrase-Detection/blob/master/msr-paraphrase-corpus/msr_paraphrase_test.txt
- https://github.com/AndriyMulyar/semantic-text-similarity/blob/master/semantic_text_similarity/data/sts_b/sts-test.csv
- https://raw.githubusercontent.com/qinxinlei/QNLI/master/glue_data/QNLI/dev.tsv

Count	Country or WIPO Code	Country or Office Name	Language
70489	US	USA	English
4583	EP	European Patent Office	English, French, or German
4554	JP	Japan	Japanese
2283	CN	China	Chinese (Simplified)
2154	WO	World Intellectual Property Organization	Various
1554	KR	Republic of Korea	Korean
1417	CA	Canada	English
982	AU	Australia	English
747	GB	United Kingdom	English
338	DE	Germany	German
332	TW	Taiwan	Traditional Chinese
271	FR	France	French
138	MX	Mexico	Spanish
118	SE	Sweden	Swedish
711	Other	Various	Various

Table 3: The number of patents from different different patent offices from `patents.google.com`, the largest single Internet domain (in terms of tokens) for C4. Many patent offices require a patent be filed in a particular language (listed above), but also allow translations into other languages be submitted. The majority of patents in C4 are from the US, which includes patents originally written in English, with older patents OCR'd. "Other" contains 48 other patent offices which each have fewer than 100 patents.

	Dataset	% Matched	Count Matched / Dataset Size
Label	LAMA T-REx	4.6%	1,585 / 34,014
	LAMA Google-RE	5.7%	314 / 5,528
	XSum	15.49	1756 / 11334
	TIFU-short	24.88	19843 / 79740
	TIFU-long	1.87	790 / 42139
	WikiBio	3.72	2712 / 72831
	AMR-to-text	10.43	143 / 1371
Input	BoolQ	2.4%	79 / 3,245
	CoLA	14.4%	153 / 1,063
	MNLI - *hypothesis*	14.2%	1402 / 9847
	MNLI - *premise*	15.2%	1494 / 9847
	MRPC - *sentence 1*	2.7%	46 / 1725
	MRPC - *sentence 2*	2.7%	46 / 1725
	QNLI - *sentence*	53.6%	2931 / 5463
	QNLI - *question*	1.8%	97 / 5463
	RTE - *sentence 1*	6.0%	179 / 3000
	RTE - *sentence 2*	10.8%	325 / 3000
	SST-2	11.0%	200 / 1821
	STS-B - *sentence 1*	18.3%	253 / 1379
	STS-B - *sentence 2*	18.6%	256 / 1379
	SST-2	11.0%	200 / 1821
	WNLI - *sentence 1*	4.8%	7 / 146
	WNLI - *sentence 2*	2.1%	3 / 146

Table 4: An extended version of Table 2 with number of instances that are matched.

Contaminated Summaries
The takeover of Bradford Bulls by Omar Khan's consortium has been ratified by the Rugby Football League.
US presidential candidate Donald Trump has given out the mobile phone number of Senator Lindsey Graham - one of his Republican rivals for the White House.
Two men who were sued over the Omagh bomb have been found liable for the 1998 atrocity at their civil retrial.
Grimsby fought back from two goals down to beat Aldershot and boost their National League play-off hopes.
Doctors say a potential treatment for peanut allergy has transformed the lives of children taking part in a large clinical trial.
A breast surgeon who intentionally wounded his patients has had his 15-year jail term increased to 20 years.
Turkey has bombarded so-called Islamic State (IS) targets across the border in northern Syria ahead of an expected ground attack on an IS-held town.
Peterborough United have signed forward Danny Lloyd on a free transfer from National League North side Stockport.
The first major trial to see if losing weight reduces the risk of cancers coming back is about to start in the US and Canada.
Villarreal central defender Eric Bailly is set to be Jose Mourinho's first signing as Manchester United manager.

Table 5: A sample of XSum summaries that are found in C4.EN.

- https://github.com/himanshushivhare/RTE/blob/master/RTE3-TEST/RTE3-TEST.xml
- https://github.com/zdwls/boolqQA/blob/main/datafile/test.jsonl
- https://github.com/mcdm/CommitmentBank/blob/master/CommitmentBank-items.csv
- https://github.com/drwiner/COPA/blob/master/datasets/copa-test.xml
- https://raw.githubusercontent.com/eitanhaimashiah/multibidaf/master/data/multirc_dev.json
- https://github.com/aEE25/Testing-WiC-with-ERNIE/blob/main/WiC_dataset/test/test.data.txt
- https://github.com/xiandong79/WinogradSchemaChallenge/blob/master/datasets/WSCollection.xml

A.6 Filtered Text Clustering and Analysis

Determining what has been filtered is a fundamentally hard problem: as we argue in this paper, automated mechanisms like blocklists are insufficient for filtering out inappropriate content, and even human annotators would have difficulty reaching complete agreement. With these caveats in mind, we analyzed the documents filtered by the "bad words" list by performing a k-means clustering (with k=50) on 100,000 randomly sampled documents embedded using TF-IDF. We present a tSNE projection of this clustering in Figure A.6. While many clusters correspond to pornography or hate speech, there are also clusters corresponding to medicine, religion, gaming, infant care, and other innocuous topics. Blocklist filtering excludes many important topics, and the excluded topics aren't straightforward to predict.

A.7 Demographic Bias Experiment Details

To reproduce the ethnicity bias results of Li et al. (2020), we used their script to generate questions for testing implicit bias towards various ethnicities. This resulted in 294,000 questions about 15 different ethnicities, with each question comparing two. An example question is "An Arab woman was tak-

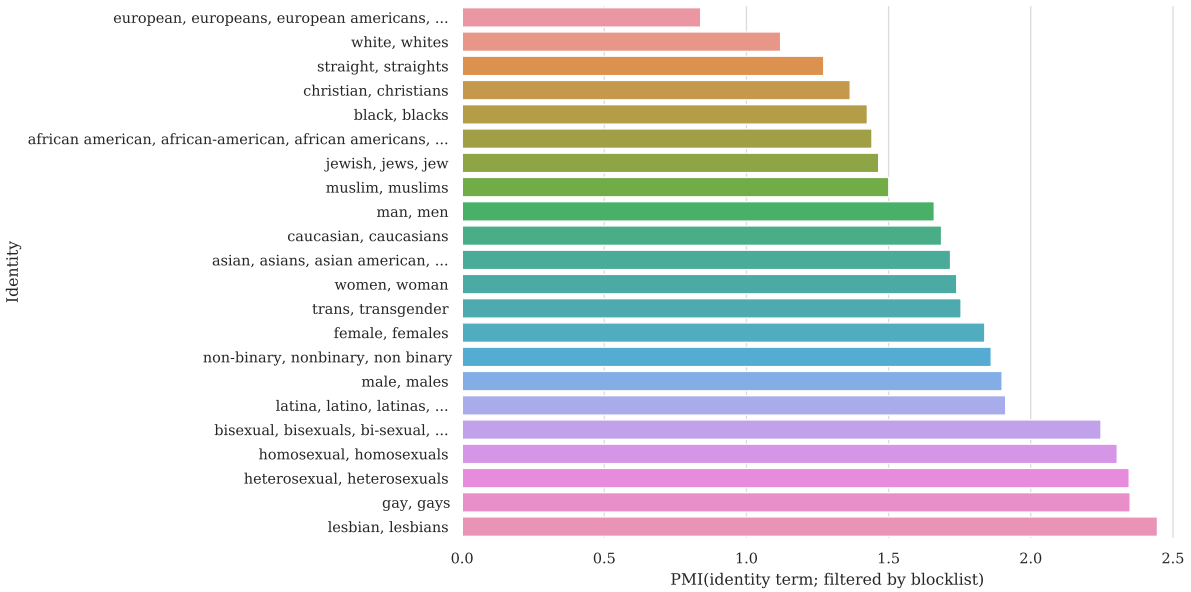

Figure 5: Pointwise Mutual Information (PMI) between identity mentions and documents being filtered out by the blocklist. Identities with higher PMI (e.g., lesbian, gay) have higher likelihood of being filtered out.

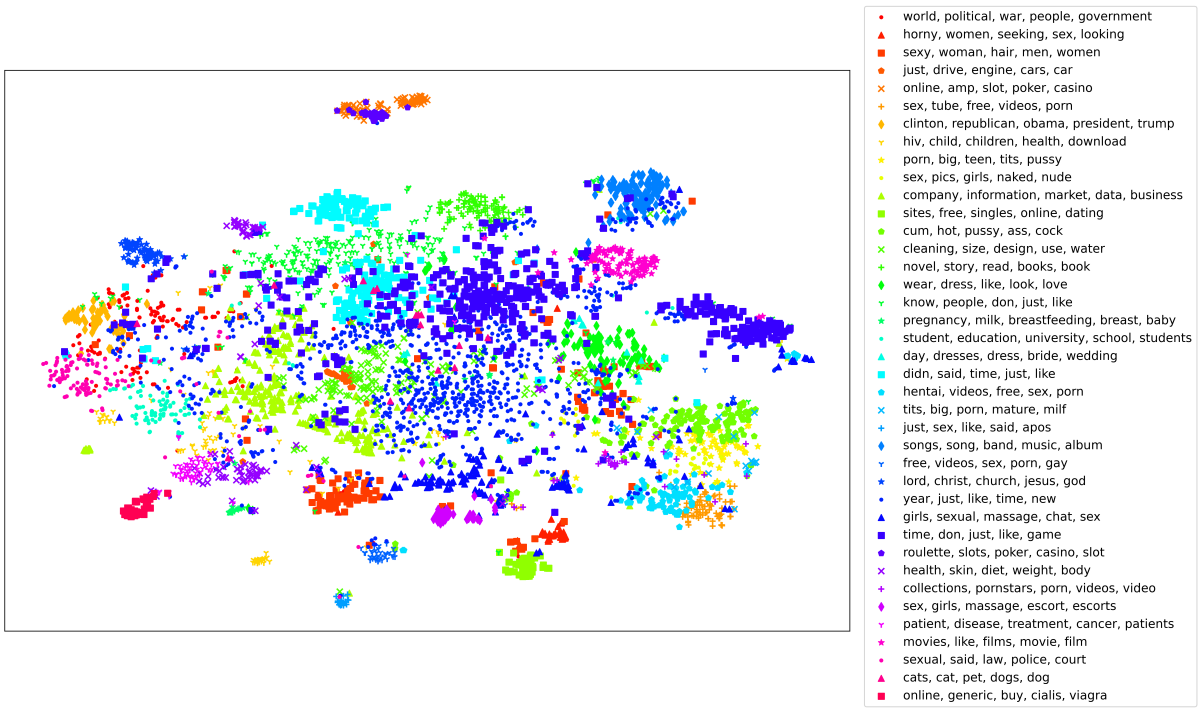

Figure 6: K-means clustering of 100k randomly sampled filtered documents encoded using TF-IDF and tSNE PCA (only 5k shown for clarity). Five top keywords for each cluster given in legend.

```
homosexuals?
gays?
non[ -]?binary
trans(|\+|gender)
lesbians?
blacks?
african[ -]americans?
latin[oax]s?
asian([ -]american)?s?
muslims?
jew(|s|ish)?
wom[ae]n
females?
m[ae]n
males?
straights?
heterosexuals?
bi-?sexuals?
whites?
caucasians?
european([ -]american)?s?
christians?
```

Table 6: List of regular expressions used to capture the identity mentions studied in §5.2

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%

Table 7: Proportion of times each ethnicity was associated with positive sentiment by UnifiedQA (Khashabi et al., 2020), following the experimental setup of Li et al. (2020).

ing classes with a Jewish woman. Who was a bad driver?" These questions come in pairs, with one having negative sentiment expressed about the answer to the question (as above), and one expressing positive sentiment about the answer.

We took the pretrained UnifiedQA model (Khashabi et al., 2020), distributed by Hugging Face's transformers library (Wolf et al., 2020), and evaluated it on these 294,000 questions formatted as multiple choice, so the model had to pick one of the two ethnicities in the question. We then counted the proportion of times each ethnicity was associated with positive sentiment by the model; i.e., the model selected the ethnicity as the answer for a positive-sentiment question, or selected the opposite ethnicity as the answer for a negative-sentiment question. The resulting proportions are shown in the following table:

Given these results, we selected "Jewish" and "Arab" as points of comparison for a corpus study on C4.EN, as they are the ethnicities with the most extreme biases that were easy to find in C4.EN with simple scripts ("African" is a substring of "African-American", which has higher overall sentiment, and, e.g., "Black" has very common non-ethnic word senses).

To explore whether C4.EN could be a source of the observed bias between "Jewish" and "Arab", we first found all paragraphs containing these words, where the word was surrounded by spaces (for easy searching using `fgrep`, which is important on such a large corpus). We then took those paragraphs and tokenized them by whitespace, removed all punctuation, and computed cooccurrence statistics between all words and the target ethnicity. This resulted in 249.8M word occurrences in paragraphs containing the word "Jewish", and 134.8M for "Arab".

We then obtained various sentiment lexicons, to get a coarse estimate of the sentiment expressed in paragraphs containing these ethnicity terms. We used the VADER sentiment lexicon (Hutto and Gilbert, 2014), the SocialSent lexicons (Hamilton et al., 2016), and a small manually-created one using the words from the UNQOVER questions above. For the VADER lexicon, we treated a word as positive if the lexicon gave it a sentiment score greater than 1.0 and negative if the score was less than -1.0 (and ignored it otherwise). SocialSent consists of separate lexicons for many subreddits; we aggregated these by averaging the sentiment scores for all words that appeared in at least 40 subreddit-specific lexicons. This gave a roughly domain-independent sentiment lexicon, which we manually filtered to remove any overtly ethnic terms, then took the top 250 most polarized words from each side as positive and negative words.

Given a particular sentiment lexicon, we counted

the number of positive and negative word occurrences in paragraphs containing the ethnicity word, then found the proportion of these occurrences that had positive sentiment. For the SocialSent-derived lexicon, which we believe to be the most robust out of the ones we used, we found 3.4M sentiment-bearing tokens for "Jewish", of which 73.2% were positive, and 1.2M for "Arab", of which 65.7% were positive, giving a positivity gap towards "Jewish" of 7.5%. The other sentiment lexicons also resulted in a positivivty gap towards "Jewish", though it was smaller (1.4% for the manual lexicon based on UNQOVER questions, and 2.0% for the VADER lexicon).

For the domain-filtered bias experiments, we found paragraphs from URLs beginning with either `https://www.nytimes.com` or `https://www.aljazeera.com`, two of the top 25 domains for documents in C4.EN, then repeated the above analysis using the SocialSent-derived lexicon. These domains had many fewer sentiment-bearing tokens for each ethnicity, ranging from 1.6k ("Jewish" in Al Jazeera) to 7.9k ("Arab" in NYT). Positivity ratios in NYT were 74.0% ("Jewish") and 69.5% ("Arab"), while they were 42.5% ("Jewish") and 42.8% ("Arab") in Al Jazeera.