SOBR: A Corpus for Stylometry, Obfuscation, and Bias on Reddit

Chris Emmery[1], **Marilù Miotto**[2], **Sergey Kramp**[1], **Bennett Kleinberg**[2,3]
[1]Department of Cognitive Science & Artifical Intelligence, Tilburg University
[2]Department of Methodology and Statistics, Tilburg University
[3]Department of Security and Crime Science, University College London
cmry@pm.me

Abstract

Sharing textual content in the form of public posts on online platforms remains a significant part of the social web. Research on stylometric profiling suggests that despite users' discreetness, and even under the guise of anonymity, the content and style of such posts may still reveal detailed author information. Studying how this might be inferred and obscured is relevant not only to the domain of cybersecurity, but also to those studying bias of classifiers drawing features from web corpora. While the collection of gold standard data is expensive, prior work shows that distant labels (i.e., those gathered via heuristics) offer an effective alternative. Currently, however, pre-existing corpora are limited in scope (e.g., variety of attributes and size). We present the SOBR corpus: 235M Reddit posts for which we used subreddits, flairs, and self-reports as distant labels for author attributes (age, gender, nationality, personality, and political leaning). In addition to detailing the data collection pipeline and sampling strategy, we report corpus statistics and provide a discussion on the various tasks and research avenues to be pursued using this resource. Along with the raw corpus, we provide sampled splits of the data, and suggest baselines for stylometric profiling. We close our work with a detailed set of ethical considerations relevant to the proposed lines of research.

Keywords: corpus, author identification, author profiling, author obfuscation, computational stylometry, bias

1. Introduction

The increasing computational capabilities of language models do not bode well for safety in online public spaces. A large variety of pre-trained Large Language Models (LLMs) made readily available through platforms such as the HuggingFace Model Hub (Wolf et al., 2020) can be used to generate (Pan et al., 2020; Carlini et al., 2021) and infer (Tesfay et al., 2019; Kleinberg et al., 2022), sensitive information. While these often deal with concrete mentions of personal information, a handful (so far) seeks to uncover latent author attributes through computational stylometry.

Stylometry posits that one's unique writing style might encode features about an author's identity, which eventually extended to sociodemographic factors such as gender and age (Schler et al., 2006; Bamman et al., 2014b), personality (Plank and Hovy, 2015), education and income (Rao et al., 2010; Volkova et al., 2014), region of origin (Bamman et al., 2014a; Tulkens et al., 2016), political or religious affiliation (Koppel et al., 2009; Pennacchiotti and Popescu, 2011), and mental health issues (Choudhury et al., 2013; Coppersmith et al., 2015). These attributes can, often with high accuracy, be inferred through computational analysis of publicly shared writing.

Computational stylometry is an example of dual-use research: despite the merits of profiling techniques in various research fields such as computational sociolinguistics (Daelemans, 2013), detecting fraud, deception, and identity theft (Badaskar et al., 2008; Ott et al., 2011; Banerjee et al., 2014; Fornaciari and Poesio, 2014), it enables malicious actors to infer potentially sensitive information unbeknownst to the user. This is particularly harmful to individuals in a vulnerable position regarding race, political affiliation, mental health, or any other personal details made explicitly unavailable.

Historically, collecting high-quality labels for stylometric classification tasks was a costly process (both in time and resources) requiring trained annotators. Collecting the data itself, and fine-tuning models, would also require expertise and computational infrastructure. Works such as Beller et al. (2014) and Emmery et al. (2017), however, showed that targeting Twitter users that post self-reported attributes ("I'm a ...") generates enough distantly labeled data to train models that match the performance of models trained on gold standard data. These pipelines run within a day on consumer hardware; implying that the barrier to entry is low, making regulation of profiling algorithms significantly more challenging. Hence, providing vulnerable Internet users with tools to mitigate such harmful inferences is an important contribution to their online privacy and security.

The current work[1] introduces a distantly annotated corpus to grant insight into the workings and

[1]All code and baseline models for reproduction are openly available via https://github.com/cmry/SOBR. The SOBR corpus is made available under fair use data sharing agreements (see Section 5). Refer to our repository for contact details to request access.

LREC-COLING 2024, pages 14967–14983
20-25 May, 2024.

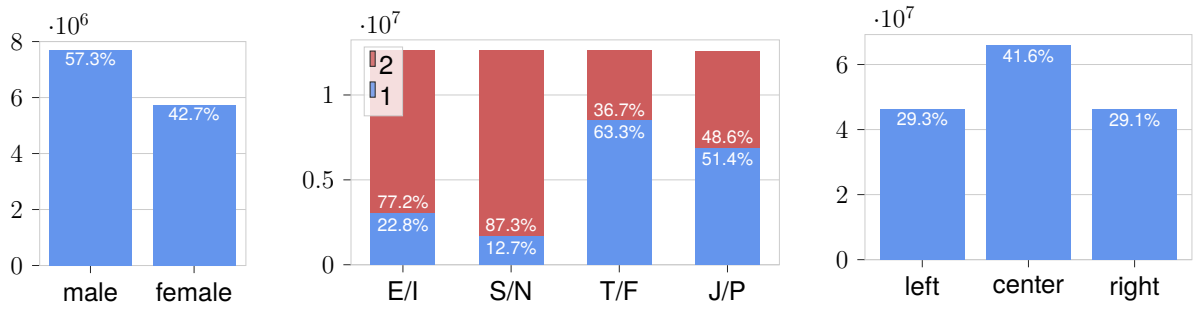

Figure 1: Gender, Personality, and Political Leaning distributions and proportions in the raw corpus. Personality labels are stacked, denoting the first of the MBTI dimension with 1, and the second with 2 (E = extrovert, I = introvert; S = sensing, N = intuitive; T = thinking, F = feeling; J = judging, P = perceiving).

effectiveness of stylometric profiling techniques, to emphasize how even inadvertently revealing personal information might harm individuals, and to develop tools aimed towards increasing one's online privacy. Accordingly, in addition to describing the collection process and characteristics of the data, we provide initial profiling baselines, and a discussion of intended purposes of the corpus.

2. Data

Reddit[2] is a content discussion website with over 52M users. Users *post* hyperlinks, text, images, or videos, after which the core social dynamic of the website consists of up and downvoting and *commenting* on said posts. The website is topically divided into so-called *subreddits*, often covering a particular theme, or niche interest. Users can join these subreddits to receive a content timeline, or follow specific users. Posts are either communicated via a user *profile*, or under a subreddit (cross-posting is possible). When posting under a particular subreddit, or commenting under a post, users can opt to show something close to a 'status message' next to their name, called a *flair*. Flairs typically play into the theme of the subreddit, and might be funny or informative (e.g., a user's (home) country in the `r/Europe` subreddit). Users gather karma through net upvotes, and posts can be moderated if they do not abide by the rules of a particular subreddit. Depending on their size, these subreddits function as topical communities with their own norms and memes. We employ these features of Reddit to build a corpus that connects textual data (i.e., the posts) to publicly available and openly shared author attributes.

2.1. Label Collection

We considered five attributes: age, gender, nationality, personality and political leaning. Authors were labeled through related subreddits, from which we predominantly used flairs (age and gender were extracted from post-level self-reporting). Further details per attribute are noted below. Their distributions can be found in Figures 1 and 2.

Age and Gender Age and gender were obtained via posts self-reporting age and gender information. When sharing personal stories, users often follow the common practice of posting their age and gender in the format `(GAA)`[3] (e.g., "When I `(F34)` went..." indicating a female aged 34). We extracted these patterns using regular expressions; authors reporting gender `F` or `f` were labeled as female, while users reporting `M` or `m` were considered male. To avoid contradicting age labels across years, this attribute was stored as inferred year of birth (i.e., the post's year minus the self-reported age).

Nationality Nationality labels were extracted through user flairs from a set of subreddits previously used in the work of Rabinovich et al. (2018); specifically, `r/Europe`, `r/AskEurope`, `r/Eurosceptics`, `r/EuropeanCulture`, and `r/EuropeanFederalists`. We mapped flairs to country labels by hand, based on flairs that were used more than ten times in a sample month (July 2021). Note that there is no guarantee that these reports exactly map to nationality (rather than place of residency); however, similar to other labels, we assume this holds for the majority of authors. The choice for focusing on nationality labels over residency is multifacetted: residency is overall less reported on, and users may change country of residency between time of posting and time of collection. Furthermore, stylometric analysis would typically focus on traits that show through non-native English writing to predict nationality. Our assumption is residency might be picked up through content words (e.g., mentioning cities, local food, etc.), but has less effect on writing style in comparison.

[2] `https://www.reddit.com`

[3] Other less frequently used variants we observed included `(AAG)`, `AAG` and `GAA`.

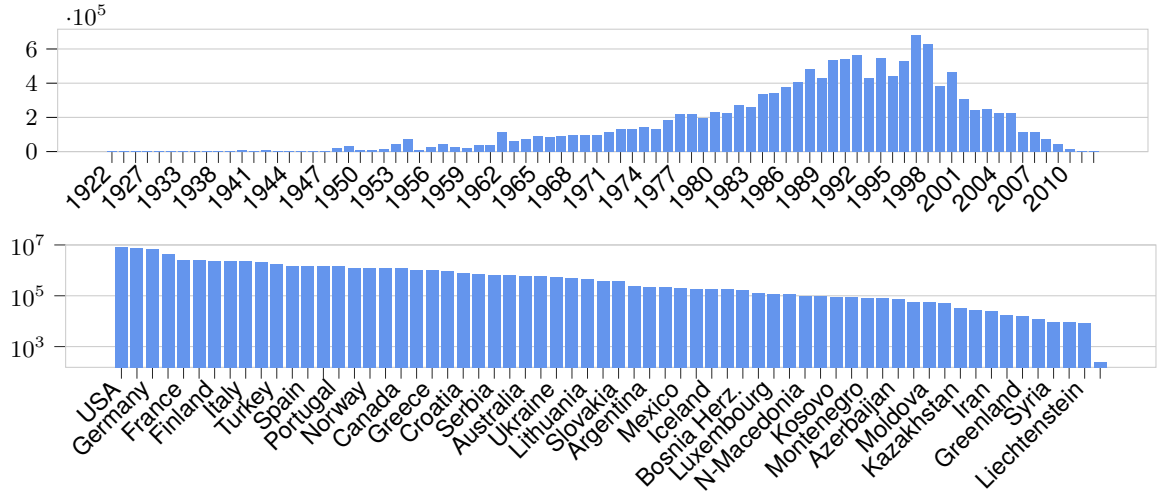

Figure 2: Age (top) and Nationality (bottom, log-scale) distributions. Not all labels are shown for readability.

Att.	Raw	Rand.	Strat.	Bal.
Age	13,441K	83,748	-	7,796
Gen.	13,441K	89,272	103,047	85,293
Nat.	65,544K	165,234	-	361
Per.	12,614K	326,520	289,860	180,800
Pol.	158,494K	114,463	120,363	97,012

Table 1: Amount instances per set per attribute (Att., Gen. = Gender, Nat. = Nationality, Per. = Personality, Pol. = Political Leaning). For the Raw part, the quantities are on post-level, for both the random sampled (Rand.), stratified (Strat.), and undersampled (Bal.) datasets, quantities are on slice-level (one instance is up to 1500 words).

Personality Personality labels were assigned through flairs for Myers-Briggs Type Indicators (MBTI) on MBTI-related subreddits: `r/entj`, `r/enfp`, `r/enfj`, `r/intp`, `r/esfj`, `/esfp`, `r/infp`, `r/intj`, `r/infj`, `r/isfj`, `r/entp`, `r/estp`, `r/estj`, `r/istj`, `r/isfp`, `r/istp`. These are split into four dimensions: extrovert/introvert, sensing/intuitive, thinking/feeling, and judging/perceiving, and broken down into four labels; e.g., for `entj`, an author would be labeled as extrovert, intuitive, thinking, and judging.[4]

Political Leaning For political leaning labels, rather than extracting users by association from political subreddits (e.g., `r/Liberal`, `r/conservatives`, etc.)—as has been shown in De Francisci Morales et al. (2021) to be more politically diverse than one would expect—we use self-reports from the Political Compass[5] in flairs on the subreddit `r/PoliticalCompass`. Test results are often reported in the form of: [social scale result (authoritarian, libertarian)]-[economic scale result (left, right or center)]; for example, `auth-left`. For our corpus, we only considered the economic scale, as a significant amount of users did not report the social scale result.

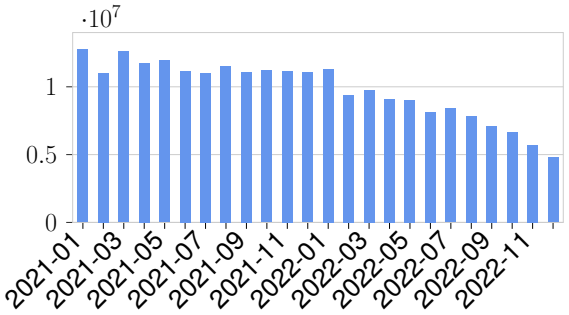

Figure 3: Raw post frequency per month.

2.2. Datasets

For the current version of the corpus, we extracted attributes and posts from two years worth of Reddit snapshots from `Pushshift.io`,[6] totaling 235,630,014 labeled posts (see Figure 3), or 174G of JSON data. These individual posts were collected and labeled as follows:

- In addition to labels, the techniques described above provided us with a list of author IDs (or handles) and the subreddits and posts they were extracted from (see Listing 1).
- We retrieved all posts from labeled authors across the entire two years of Reddit snapshots, and individually labeled those posts with the labels of their author (see Listing 2).

[4] We discuss the psychometric limitations of this measure (Stein and Swan, 2019) in Section 4.3.

[5] `https://www.politicalcompass.org`

[6] `https://pushshift.io`

```
    ...
    author_id: 't2_xxxxxx',
    labels: {
      personality: {
        introvert: [{
          post_id: "xxxxxxxx",
          flair: "INFP",
          subreddit: "r/infp",
          database_month: "2021-07"
        }]
      },
      gender: {
        male: [{
          post_id: "xxxxxxxx",
          regex_match: "Me (32M)",
          match_index: [5, 13],
          subreddit: "r/AmItheAsshole",
          database_month: "2021-07"
        }]
      }
    }
```

Listing 1: Author database example. Attributes retrieved with flairs (here: personality) are structured differently than text reports (here: gender).

```
{
  post_id: 'xxxxxxx',
  author_id: 't2_xxxxxxxx',
  subreddit: 'r/PoliticalCompass',
  created_on: '2022-12-01 15:43:24',
  male: null,
  female: null,
  gender_source_post: null,
  birth_year: null,
  age_source_post: null,
  nationality: null,
  nationality_in_domain: null,
  political_leaning: 'right',
  political_leaning_in_domain: 1,
  personality_extrovert: null,
  personality_introvert: null,
  personality_sensing: null,
  personality_intuitive: null,
  personality_thinking: null,
  personality_feeling: null,
  personality_judging: null,
  personality_perceiving: null,
  personality_in_domain: null,
  post: 'Are they that dumb ...'
}
```

Listing 2: Post database example. Indicators if posts originate from the same subreddit as a particular label is indicated with “in_domain”, and posts a label was retrieved from with "source_post".

- Additionally, posts were labeled according to if they were either from a subreddit for which we used its flairs for labeling, or if the post itself was used for labeling.

A full breakdown of how many posts were labeled under which attributes can be found in Table 1.

After the raw post corpus was collected, we aggregated posts by author, and split the data into three types of datasets to be used for training: one random sample, one balanced sample (undersampled based on minority class), and a stratified sample according to the corpus' label distributions[7]. For the current version of SOBR, we sampled 10,000 authors. Their full post history was—similar to Rabinovich et al. (2018)—sliced per 1500 ‘words’ (delimited by whitespace) to produce instances. Any authors posting less than 1500 words, and excess words not fitting a slice, were removed.

2.3. Considerations & Recommendations

Now that we have established the SOBR corpus and the sampled datasets for training models, we will provide some background regarding our considerations in collection and recommendations for use beyond the splits we provide.

Data Statement Our corpus includes real subjects (Reddit users). As per the Reddit User Agreement, users agree not to disclose sensitive information, and consent that their comments are publicly available and accessible through an API. They retain the right to have their posts removed through a (verified) deletion request. We store post IDs, usernames, and user IDs for compliance, but anonymize these when disseminating the corpus. The data will only be made directly available under a fair use agreement. Code to reproduce data collection is available in our repository.

Distant Labeling As mentioned, part of previous profiling work relying on using distant labeling techniques (e.g., Beller et al., 2014; Emmery et al., 2017; Gjurkovic et al., 2021) have used in-text self-reports. We argue that, despite the underlying assumption of distant labeling (also referred to as weak labeling) being that these are not always accurate, the potential for error increases when using semi-structured patterns in noisy user-generated text. In Emmery et al. (2017) in particular, potential ambiguities (e.g., “Sometimes I think I’m a girl”) were partially addressed through a set of rules shown to improve inter-rater label agreement by 12.5%. Hence, for all labels, we deliberately opted for structured retrieval (as in, i.a., Gjurkovic and Snajder, 2018). This increases confidence in the labels, limiting error to user-level (i.e., some users might still lie) rather than error in noisy language.

[7] Age and Nationality are excluded in this version. Users whose labels exhibited inconsistency or ambiguity (e.g., a user reported both male and female) were excluded from the initial dataset.

Age/Gen.	Per.	Nat.	Pol.	Prev. (%)
			v	58.54
		v		21.19
	v			8.84
v				8.57
		v	v	2.18
	v		v	0.40
	v	v		0.10
v			v	0.07
v		v		0.04
v	v			0.04
	v	v	v	0.02

Table 2: Prevalence (Prev.) of different attribute (co-)occurrences (Gen. = Gender, Nat. = Nationality, Per. = Personality, Pol. = Political Leaning).

It is also worth mentioning that despite the large sample, authors reporting multiple attributes are extremely rare (see Table 2). This is certainly a strong limitation compared to data gathered through author inquiry (see e.g., Volkova et al., 2014), and affects the extent to which cross-correlations, or multi-label author predictions can be investigated.

Reddit as a Corpus According to a report from Pew Research Center,[8] Reddit users are predominantly white (70%), male (67%), aged 18-29 (64%), with a college degree (42%), in the higher income bracket ($75k+, 35%), and lean liberal (43%)—prototypical of the WEIRD group (Henrich et al., 2010). This implies that for many research purposes, Reddit may induce a bias. Additionally, our approach inherently introduces bias by utilizing self-reports since it relies on users comfortable divulging this information about themselves.

In our sample, we observe a less WEIRD distribution (see Figures 1 and 2). While the majority is still male, at 57.3% they are less represented, the majority of the users is older than 30 years (mean year of birth=1988, SD=12.4), personality types are convincingly INT(J/P), and political leaning is distributed bell-shaped to the left and right with the majority being centrist). Due to our collection strategy, the top portion of nationalities are all English-speaking or European countries; other countries are often collectively labeled as 'non-European'.

Reddit as a platform has biases too: while featuring a huge variety of topics, the largest subreddits predominantly center around news, hobbies, and memes.[9] Hence, given all considerations above, any models trained on these data should be tested on an unbiased (or at least more representative) set, preferably one that is out-of-domain.

[8] pewresearch.org/journalism/2016/02/25

[9] reddit.com/best/communities/1/

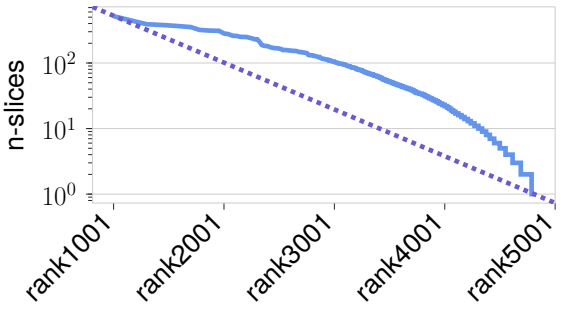

Figure 4: Rank-frequency distribution of author/n-slices in the Random dataset.

Mitigating Data Contamination Contamination may occur in setting up splits for model-training purposes. We identify a few causes here: first, and most obvious, in-text self-reports will be picked up on by text classifiers and might cause classification results to be misrepresentative and likely non-generalizable. In the current version of the corpus, we have filtered all age/gender patterns (see above), although mentions beyond that ("I'm a ...") are challenging to remove completely.

This form of contamination can be partially addressed in several ways: slicing author profiles, principled splitting, and domain filtering. Slicing (as explained in Section 2.2)—in addition to being a requirement for effective profiling (Luyckx and Daelemans, 2011)—spreads the impact that these few scattered self-reports may have. Splitting should be done at the author level; indeed, mixing authors between train and test sets creates arguably obvious issues with evaluation (not being unseen examples of the attributes they represent). Additionally, as can be observed in Figure 4, the author-to-slice ratio follows a Zipfian-like distribution, i.e., the majority is represented in only a single slice, whereas a few have the bulk of the slices. To accurately gauge generalization, the test set should consist of authors with only a few slices worth of posts. Lastly, the topical 'domain' of the subreddit might also unfairly influence classification success. As shown in Kramp et al. (2023), classifying authors gathered from the same subreddit is significantly easier, as classifiers may pick up on particular content words or other idiosyncratic cues from that domain. Explicitly excluding such subreddits (annotated in our corpus, see Listing 2) is preferable.

3. Baseline Experiments

For our text classification baselines, we opt for three tried-and-tested models with different architectures. These serve as recommendations for baselines worth exploring further, and to benchmark other profiling model architectures.

3.1. Experimental Setup

Data We will report experiments on the Random, Stratified, and Balanced sampling datasets. The instances are sorted by the most frequent author by default (descending); hence, a 90% split, only shuffling after, suffices. Known self-reports and in-domain portions were removed. We also filtered bots known to Reddit (i.e., those having received a bot tag). Tokenization for NB-LR and fastText (detailed below) was implemented using spaCy,[10] whereas for BB-LR we used the associated tokenizer. No further preprocessing was applied.

Targets All tasks were converted to binary or multi-class classification tasks. Gender used binary assignment (male or female), age was divided into categories following Pardo et al. (2015) (18-24, 25-34, 35-49, 50-64, 65-xx), nationality at country-level, personality split into four binary tasks (one per MBTI dimension), and political leaning used the original three labels (left, center, right).

LR We include a standard Logistic Regression model (implemented using `sklearn`, Pedregosa et al., 2011). As input, it uses tf-idf over n-gram features (here: token uni and bi-grams, and character tri-grams with a minimum document frequency of 3, and an occurrence rate of 90%). The idf values are smoothed, and tf values are scaled sublinearly.

fastText The `fastText` library (Joulin et al., 2017) offers a fast linear model with a single embedding layer and a hierarchical softmax function (Mikolov et al., 2013). We opted for token uni-grams and bi-grams (with a minimum occurrence of 3) as input, used an embedding size of 50, learning rate of 0.1, a bucket size of 1M, and trained for 25 epochs. It should be noted that fastText uses Hogwild (Recht et al., 2011) for parallelization; hence, our results are not exactly replicable (standard deviation is often negligible, see Emmery et al., 2017).

BB-LR Recently shown to be an effective model for profiling, Kramp et al. (2023) use Big Bird (`google/bigbird-roberta-base` in `transformers`; Wolf et al., 2020) from Zaheer et al. (2020); specifically, the `[CLS]` embeddings as input to a Logistic Regression model. Big Bird facilitates the processing of longer-form texts, where models such as BERT (Devlin et al., 2019) may underperform. Embedding extraction is relatively fast, and fine-tuning typically runs in within a day on this data.

Evaluation For our metrics, we looked at macro F_1-score averages for all sets. In the Random set, some labels are heavily skewed, and we want to keep our model evaluations directly comparable. Given that our goal is providing reasonable baselines, and not squeezing performance out of the proposed models, we did not tune any hyperparameters, and therefore did not apply cross-validation.

[10] `github.com/explosion/spaCy`

3.2. Results

Several baseline models were proposed, and their results can be found in Table 3. Generally, performance on most tasks, save for gender and nationality prediction, is below majority baseline. Note, however, that these are unoptimized models, and we have performed no preprocessing whatsoever. Given the noisy nature of the corpus, the underperforming baselines were to be expected. Even so, there are indicators of performance gain through increasing model complexity. We expect BB-LR to see significant improvements when the model is fine-tuned. The fact that there is no out-of-the-box solution offers fruitful avenues for further studies.

Additionally, these results call for further investigation of the different tasks. The MBTI dimensions from the Random set are highly skewed, and thus a priori make for a challenging task. In future versions of the corpus, this limitation would need addressing through either collecting scores or collecting more data to improve minority class representation. The nationality prediction task seems rather easy to get reasonable performance on; likely, a stricter filtering of domain-related subreddits (that were not part of the ones we extracted flairs from) is required.

Finally, to gauge the effects of different sampling methods, we selected our best performing model in the Random sample experiments (fastText) and trained it on the Balanced and Stratified data. While balancing seems to have an adverse effect on performance (both age and nationality were too sparse after undersampling), stratification seems to have a partially positive effect, in particular for the personality dimensions.

4. Further Application and Outlook

We have presented a corpus that was initially set up as one focused on author profiling. As argued in Section 1, however, this task has inherent dual-use problems. As such, we dedicate a portion of the current work to discussing potential techniques that work towards mitigation of these predictive models, recommendations for implementing them, as well as an overview of other applications that might benefit from this corpus. We end this section with a discussion of improvements intended for future versions of the SOBR corpus.

	Model	Age	Gen.	E/I	S/N	T/F	J/P	Nat.	Pol.	Avg
Random	Majority	**.633**	.706	**.848**	**.940**	**.784**	**.740**	.167	**.568**	.673
	LR	.548	.797	.725	.888	.720	.635	.559	.440	.664
	fastText	.577	**.825**	.756	.922	.683	.666	**.620**	.505	**.694**
	BB-LR	.522	.781	.683	.900	.648	.568	.517	.430	.631

	Model	Age	Gen.	E/I	S/N	T/F	J/P	Nat.	Pol.	Avg
Stratified	fastText	-	.819	.742	.932	.734	.666	-	.439	*.722*
Balanced	fastText	-	.793	.435	.499	.615	.635	-	.454	.572

Table 3: Test split scores (Macro F_1) per baseline on the Random and Balanced (latter excludes Age and Nationality) sets (Gen. = Gender, Nat. = Nationality, E/I = extrovert/introvert; S/N = sensing/intuitive; T/F = thinking/feeling; J/P = judging/perceiving, Pol. = Political Leaning), and their averages (Avg).

4.1. Adversarial Stylometry & Author Obfuscation

Aside from profiling, the most obvious application for the current corpus would be Adversarial Stylometry (also referred to as author obfuscation: Kacmarcik and Gamon, 2006; Brennan et al., 2012; Thi et al., 2015). This machine learning task takes a text as input and perturbs it in some way: this might be through changes in characters (Eger et al., 2019), translation (Rao and Rohatgi, 2000; Shetty et al., 2018), paraphrasing (Reddy and Knight, 2016), word substitutions (Emmery et al., 2021), or full style changes (Kabbara and Cheung, 2016). The goal of these perturbations (the 'attack') is to decrease the accuracy of a classifier used for profiling. Hence, the success of this adversarial attack is measured through its reduction in accuracy (Papernot et al., 2016b); however, an often-overlooked additional constraint—particularly different from the broader domain of adversarial Machine Learning—is that this reduction should approach chance-level performance (Emmery et al., 2018). After all, fooling a profiling classifier (e.g., for privacy reasons) to perform below chance implies we are systematically changing the labels, i.e., engaging in style transfer. This is not always desirable (e.g., writing like a liberal when one is conservative).

In relation to style, in addition to reducing performance, another soft constraint is that the perturbations should be limited to those that produce a text that is grammatically and semantically consistent with the original (Potthast et al., 2016). While language model perplexity, semantic similarity metrics, or other approximations might be used, these techniques only approximate consistency. An alternative or complementary evaluation may involve (trained) human raters. Existing annotation setup suggestions range from simple obfuscator identification tasks (Emmery et al., 2021) to complex (and more robust: Potthast et al., 2018; van der Lee et al., 2019) schemes involving linguistic analysis.

Lastly, one needs to consider if perturbations are made with the targeted profiler in a 'supervised' loop. One can assume to have access to all weights and outputs (white box), outputs only (black box), or no access at all (Papernot et al., 2016a). If one intends to design a tool for privacy purposes, the latter option is most realistic, as Internet users typically do not have access to the profiling systems they are potentially subjected to. Evaluation should, in that case, include transferability; i.e., how well does an attack fitted in isolation work on targets it has had no access to while doing so.

The SOBR corpus can be used to design obfuscation systems with a wider range of attributes than have been used before, on more data than was available before. Social media poses a challenging platform, but should also provide enough vocabulary and contemporary word usage to produce plausible output. Different (categories of) subreddits provide a novel way to measure transferability for the same author, within the same resource, but on a markedly different domain.

4.2. Investigating Bias

Representations of language have, throughout several iterations of larger scale computational techniques (e.g., static word embeddings, transformer embeddings), shown to encode and reproduce the human biases found in their training data (Manzini et al., 2019; Zhao et al., 2019; Basta et al., 2021; Elsafoury et al., 2022). With the advent of consumer-oriented conversational systems interfacing with Large Language Models (LLMs), concerns regarding their harm (Bender et al., 2021), and subsequent work on analyzing and mitigating bias (Gonen and Goldberg, 2019; Kumar et al., 2020; Wang et al., 2020) have rapidly increased (Abid et al., 2021; Felkner et al., 2023; Ghosh and Caliskan, 2023; Kolisko and Anderson, 2023).

Bias may refer to inherent skews in data, or systematic modeling errors, and the SOBR cor-

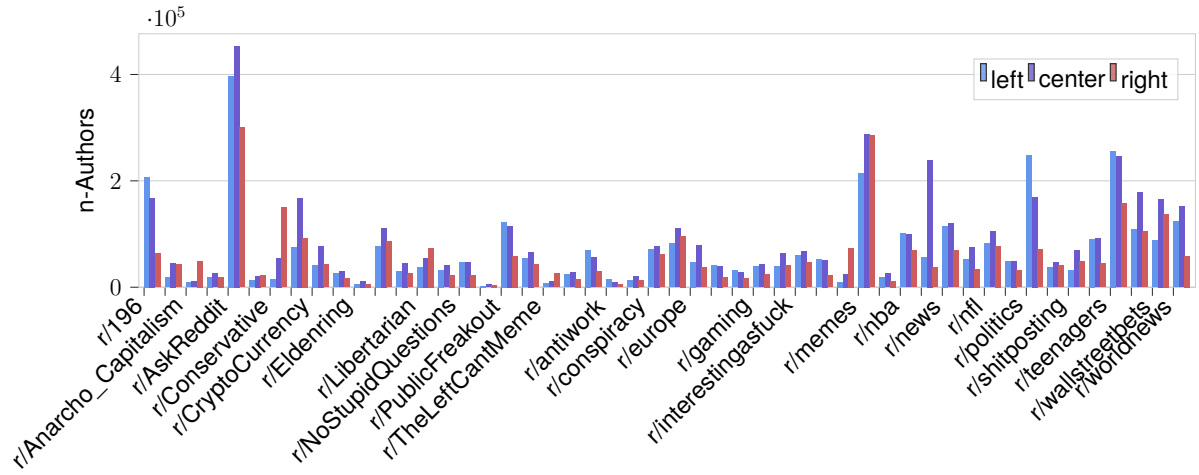

Figure 5: Political Leaning for the top 50 most represented Subreddits.

pus can be used to investigate these. While not as strong as claimed in other work, Reddit is certainly a biased resource, underlying many modern LLMs. Hence, data-sided bias (Wevers, 2019; Spinde et al., 2021), as well as societal biases related to topics that are discussed on Reddit, are prevalent in this resource. For example, Figure 5 demonstrates how our annotations of political leaning may be used to estimate bias in political representation on a particular subreddit (similar to, e.g., Gordon et al., 2020). For example, some peculiarities include that `r/politics`, `r/worldnews`, and `r/teenagers` feature predominantly left-leaning users, whereas `r/memes` and `r/wallstreetbets` have more right than left-leaning users. On the surface level, these subreddits do not voice, or are themed around, a particular leaning (in contrast to, say, `r/Anarcho_Capitalism`, `r/Libertarian`, and tangentially `r/antiwork`). More specifically, the submissions (links, text, pictures, videos) and posts (comments) on these subreddits might be evaluated in view of its general political leaning, and that of the users themselves (e.g., what type of content gets upvoted where, why, and how does that relate to bias?).

Such biases may further propagate into the models trained on this data, and manifest in the tasks they are employed for. These downstream effects of model-sided bias often prove harmful to society (Crawford, 2017), and a large body of work is dedicated to pinpointing and measuring such harmful biases (Blodgett et al., 2020; Delobelle et al., 2022; Talat et al., 2022). Dev et al. (2022) provide a broad overview of demographic dimensions that may be subject to bias, of which age, gender, nationality, and political ideology are included in SOBR. A closely related corpus is RedditBias (Barikeri et al., 2021), which provides religion, race, gender, and orientation variables, and offers a highly curated corpus specifically for studying models of language's behavior on biased statements (i.e., *a posteriori*, or *extrinsic* bias). It is therefore complementary to SOBR, which can be employed to measure a model's bias when trained, or fine-tuned on the data (i.e., *a priori*, or *intrinsic* bias).

Our corpus may also be employed to study bias mitigation (Sun et al., 2019). This can be employed data-sided (Zmigrod et al., 2019; Vanmassenhove et al., 2021; Tokpo and Calders, 2022, the latter of which use style transfer, related to the obfuscation methods discussed in Section 4.1), model-sided (Karimi Mahabadi et al., 2020), although often far from trivial (Gonen and Goldberg, 2019), and downstream (Behnke et al., 2022). Finally, our resource may also be employed for demographic-aware fine-tuning (Garimella et al., 2022); it should be noted, however, that this may in turn heighten the risk of targeted influencing attempts on a large scale (e.g., through automated generation of messages for a specific target group: Griffin et al., 2023).

4.3. Future Work

Future versions of the corpus will include more data; currently it is restricted to two years, but other than computational constraints limiting the current version, there is no reason why it cannot span more. Other attributes, such as non-binary gender, race, religion, education, income, profession, hobbies, and sexual orientation, to name a few found in other corpora (Barikeri et al., 2021; Tigunova et al., 2020) would be logical additions. Furthermore, it is worthwhile exploring options to gather multiple labels per author. Relying on flairs for more accurate labeling is a limitation in this regard; however, in the future, this portion of the corpus might be used to assess the accuracy of labels gathered through textual self-reports.

For the data splits, we looked at a sample of authors, and an undersampled variant thereof. We

did not consider varying author slice sizes, as this would increase the experimental complexity, but it is certainly worth exploring this as an experimental parameter. While we have filtered the splits according to meta-data in the corpus (domains, bot tags), inspecting the data reveals there are still quite some repeating patterns that might be introduced by moderators, or unverified bots. In the future, these should be identified and filtered using techniques from related work (Hurtado et al., 2019; Daelemans et al., 2019).

Additionally, the personality classifications employed in this study were based on the self-reported results of the Myers-Briggs personality test, which presents some psychometric limitations (Boyle, 1995; Stein and Swan, 2019). Although the popularity of the test allowed us to capture a larger set of users sharing self-reported personality assessments, potential refinements may incorporate detailed personality trait scores rather than binary categorizations, or finding user-provided information of more reliable personality evaluation instruments, such as the HEXACO test (Ashton and Lee, 2007). Gjurkovic et al. (2021) showed that lack of self-reports may be overcome through predictive modeling on weaker tests.

Lastly, an underexplored component in the current work is Reddit's dialogue context. Barikeri et al. (2021) in particular argue in favor of studying bias in this context; we believe conversational cues, and topical context to an author's posts may also prove informative for profiling and obfuscation purposes. Our corpus stored the original post ID and submission IDs, hence retrieving these across Reddit snapshots should prove an interesting addition.

5. Ethical Considerations

Our work deals with a highly sensitive topic; automatically inferring latent personal information from text data. This section provides our ethical considerations and position regarding this research.

Dual Use Computational stylometry is inherently a dual-use task (Hovy and Spruit, 2016; Emmery, 2023), as it falls within the privacy-security trade-off. The main consideration is whether the amount of sensitive information one can infer, and the errors and harms in making such inferences, outweighs potential applications for public benefit. It is thereby notably different from, for example, the encryption debate (i.e., securing all communication hiding malicious actors on a platform), and the implementation of backdoors in such algorithms (e.g., to provide CSAM detection software). We see this as being closer to debates around the use of facial recognition and biometric scanning (Smith and Miller, 2022): i) both authorship identification and profiling require data covering a large amount of individuals, ii) the collection and intended application largely takes place in 'public (online) spaces', iii) although it may be used in a targeted manner, broad application, and indexing based on individuals (e.g., Clearview AI for facial recognition) is to be expected given its various use cases.

The harms that may originate from these applications are in compromising privacy if successful, as well as in the potential errors that are made when not succesful. Computational stylometry may be used to target and ban certain demographics; aside from nefarious political and commercial applications (e.g., identifying dissidents, political microtargeting, predatory sales practices), it may be used to perpetuate discriminatory barriers within online communities, deliberately excluding certain individuals or groups and controlling their access.

Adversarial stylometry puts itself at the other side of this trade-off; through hiding author information, privacy may be preserved, and the public may protect itself from harmful inferences. It is thereby faced with the same set of ethical considerations as encryption in potentially hiding individuals with malicious intent. Author identification may be used to uncover individuals on Darknet fora (Maneriker et al., 2021), aid against deliberate spread of misinformation (Pardo et al., 2020), and for content moderation. Profiling may be used for applications such as identifying predatory behavior against minors, or preventative identification of depression (Choudhury et al., 2013). Hiding such information may affect public security and wellbeing as well.

Demographic Attributes in NLP Several demographics used in this paper are subject to ethical concerns underscored in related work: unspecified definitions and strictly binarized representations of gender (Larson, 2017), text-based personality scoring (Fang et al., 2023, and see Section 4.3), and US-centric political leaning (Preotiuc-Pietro et al., 2017). Explicit inclusion of such (albeit limited) variables has shown to improve performance on a variety of tasks (Hovy, 2015), and these variables are therefore not uncommonly used in Natural Language Processing (NLP) research.

Furthermore, as also discussed in Section 4.2, Reddit is inherently biased. Such unbalanced sources of representation could lead to direct and more indirect forms of harm (Sap et al., 2020). It should also be mentioned that incorrect classification of these attributes—which computational stylometry, as well as obfuscation may contribute to—could lead to further harm. The role of these demographic attributes in SOBR is therefore to reproduce their framing in prior, and potentially future work. This does not mean these measures are accurate, or fair, or this status quo should be upheld.

Purpose To be as explicit as possible about the purpose of this corpus: it is intended to study the capabilities and limitations of profiling and obfuscation methods, and bias in the use and representation of demographic attributes in web corpora; here, specifically Reddit. We believe that first and foremost, the harms that originate from this technology should be studied in a controlled research environment, and by no means support broad stylometric profiling for applications beyond this.

Dissemination Our corpus is strictly intended for the purposes laid out above. We thereby only grant access under a fair use agreement and under agreement of abiding by its intents and goals. As mentioned in Section 2.3, authors are anonymized, and we will further minimize shared information through detection and cleaning of personal identifiers (using `TextWash`, Kleinberg et al., 2022),[11] username mentions, and links containing sensitive information—if not deemed necessary for the intended application of the corpus.

Future Considerations While bias receives increasing attention from the NLP community, we hope this section, and the presented resources, will also contribute to the discussion around the application of stylometric profiling within NLP and beyond. In turn, we acknowledge that the current section is limited in scope; we believe broader discussion on this topic, in particular given the current trend towards more openly accessible language technology through LLM platforms, is not only warranted, but increasingly urgent. Our future work on this resource will focus on implementing data, model, and risk cards (Gebru et al., 2021; Mitchell et al., 2019; Mohammad, 2022; Dev et al., 2022; Derczynski et al., 2023), and we hope that related work will follow suit. Finally, NLP technologies that follow a user-centered framing are not the norm, and rare in security applications (Emmery, 2023). We therefore hope the current work may in the long run contribute to increased transparency and accountability, providing the public with tools to understand and control their information-sharing practices, and the risks associated with them.

6. Acknowledgements

We would like to thank the anonymous (ethics) reviewers for their valuable feedback. This work was supported by a seed grant from Tilburg University's School of Humanities and Digital Sciences. Finally, our research strongly relied on openly available resources. We thank all whose work we build upon.

[11] https://github.com/maximilianmozes/textwash

7. Bibliographical References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 298–306. ACM.

Michael C Ashton and Kibeom Lee. 2007. Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and social psychology review*, 11(2):150–166.

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, pages 817–822. The Association for Computer Linguistics.

David Bamman, Chris Dyer, and Noah A. Smith. 2014a. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 828–834. The Association for Computer Linguistics.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014b. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Ritwik Banerjee, Song Feng, Jun Seok Kang, and Yejin Choi. 2014. Keystroke patterns as prosody in digital writings: A case study with deceptive reviews and essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1469–1473. ACL.

Soumya Barikeri, Anne Lauscher, Ivan Vulic, and Goran Glavas. 2021. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1941–1955. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2021. Extensive study on the underlying

gender bias in contextualized word embeddings. *Neural Comput. Appl.*, 33(8):3371–3384.

Hanna Behnke, Marina Fomicheva, and Lucia Specia. 2022. Bias mitigation in machine translation quality estimation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487, Dublin, Ireland. Association for Computational Linguistics.

Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'm a belieber: Social roles via self-identification and conceptual attributes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 181–186. The Association for Computer Linguistics.

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 *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. 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, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics.

Gregory J Boyle. 1995. Myers-briggs type indicator (mbti): some psychometric limitations. *Australian Psychologist*, 30(1):71–74.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.*, 15(3):12:1–12:22.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 2633–2650. USENIX Association.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, pages 1–10. The Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. Keynote at the Neural Information Processing Systems (NeurIPS) Conference. Long Beach, CA, USA.

Walter Daelemans. 2013. Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, volume 7817 of *Lecture Notes in Computer Science*, pages 451–462. Springer.

Walter Daelemans, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco M. Rangel Pardo, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Michael Tschuggnall, Matti Wiegmann, and Eva Zangerle. 2019. Overview of PAN 2019: Bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9-12, 2019, Proceedings*, volume 11696 of *Lecture Notes in Computer Science*, pages 402–416. Springer.

Gianmarco De Francisci Morales, Corrado Monti, and Michele Starnini. 2021. No echo in the chambers of political interactions on reddit. *Scientific Reports*, 11(1):2818.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1693–1706. Association for Computational Linguistics.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R.

Leiser, and Saif Mohammad. 2023. Assessing language model deployment with risk cards. *CoRR*, abs/2303.18190.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Online only, November 20-23, 2022*, pages 246–267. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Steffen Eger, Gözde Gül Sahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1634–1647. Association for Computational Linguistics.

Fatma Elsafoury, Steve R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: systematic offensive stereotyping bias in word embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 1263–1274. International Committee on Computational Linguistics.

Chris Emmery. 2023. User-centered security in natural language processing. *CoRR*, abs/2301.04230.

Chris Emmery, Grzegorz Chrupala, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text, W-NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 50–55. Association for Computational Linguistics.

Chris Emmery, Ákos Kádár, and Grzegorz Chrupala. 2021. Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2388–2402. Association for Computational Linguistics.

Chris Emmery, Enrique Arévalo Manjavacas, and Grzegorz Chrupala. 2018. Style obfuscation by invariance. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 984–996. Association for Computational Linguistics.

Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel Oberski. 2023. On text-based personality computing: Challenges and future directions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10861–10879, Toronto, Canada. Association for Computational Linguistics.

Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9126–9140. Association for Computational Linguistics.

Tommaso Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 279–287. The Association for Computer Linguistics.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 311–319. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation

and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*, pages 901–912. ACM.

Matej Gjurkovic, Mladen Karan, Iva Vukojevic, Mihaela Bosnjak, and Jan Snajder. 2021. PANDORA talks: Personality and demographics on reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, SocialNLP@NAACL 2021, Online, June 10, 2021*, pages 138–152. Association for Computational Linguistics.

Matej Gjurkovic and Jan Snajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, PEOPLES@NAACL-HTL 2018, New Orleans, Louisiana, USA, June 6, 2018*, pages 87–97. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 609–614. Association for Computational Linguistics.

Joshua Gordon, Marzieh Babaeianjelodar, and Jeanna N. Matthews. 2020. Studying political bias via word embeddings. In *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 760–764. ACM / IW3C2.

Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. 2023. Large Language Models respond to Influence like Humans. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 15–24, Toronto, Canada. Association for Computational Linguistics.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 752–762. The Association for Computer Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

Sofia Hurtado, Poushali Ray, and Radu Marculescu. 2019. Bot detection in reddit political discussion. In *Proceedings of the Fourth International Workshop on Social Sensing, SocialSens@CPSIoTWeek 2019. Montreal, QC, Canada, April 15, 2019*, pages 30–35. ACM.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomás Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 427–431. Association for Computational Linguistics.

Jad Kabbara and Jackie Chi Kit Cheung. 2016. Stylistic transfer in natural language generation systems using recurrent neural networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 43–47, Austin, TX. Association for Computational Linguistics.

Gary Kacmarcik and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Bennett Kleinberg, Toby Davies, and Maximilian Mozes. 2022. Textwash - automated open-source text anonymisation. *CoRR*, abs/2208.13081.

Skylar Kolisko and Carolyn Jane Anderson. 2023. Exploring social biases of large language models in a college artificial intelligence course. In

Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 15825–15833. AAAI Press.

Moshe Koppel, Navot Akiva, Eli Alshech, and Kfir Bar. 2009. Automatically classifying documents by ideological and organizational affiliation. In *IEEE International Conference on Intelligence and Security Informatics, ISI 2009, Dallas, Texas, USA, June 8-11, 2009, Proceedings*, pages 176–178. IEEE.

Sergey Kramp, Giovanni Cassani, and Chris Emmery. 2023. Native language identification with big bird embeddings. *CoRR*, abs/2309.06923.

Vaibhav Kumar, Tenzin Singhay Bhotia, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Trans. Assoc. Comput. Linguistics*, 8:486–503.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL, Valencia, Spain, April 4, 2017*, pages 1–11. Association for Computational Linguistics.

Kim Luyckx and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Lit. Linguistic Comput.*, 26(1):35–55.

Pranav Maneriker, Yuntian He, and Srinivasan Parthasarathy. 2021. SYSML: stylometry with structure and multitask learning: Implications for darknet forum migrant analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6844–6857. Association for Computational Linguistics.

Thomas Manzini, Yao Chong Lim, Alan W. Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 615–621. Association for Computational Linguistics.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229. ACM.

Saif M. Mohammad. 2022. Ethics sheets for AI tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8368–8379. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 309–319. The Association for Computer Linguistics.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1314–1331. IEEE.

Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. 2016a. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277.

Nicolas Papernot, Patrick D. McDaniel, Ananthram Swami, and Richard E. Harang. 2016b. Crafting adversarial input sequences for recurrent neural networks. In *2016 IEEE Military Communications Conference, MILCOM 2016, Baltimore, MD, USA, November 1-3, 2016*, pages 49–54. IEEE.

Francisco M. Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at PAN 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Francisco M. Rangel Pardo, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at PAN 2020: Profiling fake news spreaders on twitter. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 430–438. ACM.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter - or - how to get 1, 500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2015, 17 September 2015, Lisbon, Portugal*, pages 92–98. The Association for Computer Linguistics.

Martin Potthast, Matthias Hagen, and Benno Stein. 2016. Author obfuscation: Attacking the state of the art in authorship verification. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*, volume 1609 of *CEUR Workshop Proceedings*, pages 716–749. CEUR-WS.org.

Martin Potthast, Felix Schremmer, Matthias Hagen, and Benno Stein. 2018. Overview of the author obfuscation task at PAN 2018: A new approach to measuring safety. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Daniel Preotiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle H. Ungar. 2017. Beyond binary labels: Political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 729–740. Association for Computational Linguistics.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Trans. Assoc. Comput. Linguistics*, 6:329–342.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents, SMUC@CIKM 2010, Toronto, ON, Canada, October 30, 2010*, pages 37–44. ACM.

Josyula R. Rao and Pankaj Rohatgi. 2000. Can pseudonymity really guarantee privacy? In *9th USENIX Security Symposium, Denver, Colorado, USA, August 14-17, 2000*. USENIX Association.

Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 693–701.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science, NLP+CSS@EMNLP 2016, Austin, TX, USA, November 5, 2016*, pages 17–26. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pages 199–205. AAAI.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4NT: author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, pages 1633–1650. USENIX Association.

Marcus Smith and Seumas Miller. 2022. The ethical application of biometric facial recognition technology. *AI Soc.*, 37(1):167–175.

Timo Spinde, Lada Rudnitckaia, Felix Hamborg, and Bela Gipp. 2021. Identification of biased terms in news articles by comparison of outlet-specific word embeddings. In *Diversity, Divergence, Dialogue - 16th International Conference, iConference 2021, Beijing, China, March 17-31, 2021, Proceedings, Part II*, volume 12646 of *Lecture Notes in Computer Science*, pages 215–224. Springer.

Randy Stein and Alexander B. Swan. 2019. Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, 13(2):e12434.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Welderufael B. Tesfay, Jetzabel M. Serna, and Kai Rannenberg. 2019. Privacybot: Detecting privacy sensitive information in unstructured texts. In *Sixth International Conference on Social Networks Analysis, Management and Security, SNAMS 2019, Granada, Spain, October 22-25, 2019*, pages 53–60. IEEE.

Hoi Le Thi, Reihaneh Safavi-Naini, and Asadullah Al Galib. 2015. Secure obfuscation of authoring style. In *Information Security Theory and Practice - 9th IFIP WG 11.2 International Conference, WISTP 2015 Heraklion, Crete, Greece, August 24-25, 2015 Proceedings*, volume 9311 of *Lecture Notes in Computer Science*, pages 88–103. Springer.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2020. Reddust: a large reusable dataset of reddit user traits. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6118–6126. European Language Resources Association.

Ewoenam Kwaku Tokpo and Toon Calders. 2022. Text style transfer for bias mitigation using masked language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 163–171, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 355–368. Association for Computational Linguistics.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 186–196. The Association for Computer Linguistics.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5443–5453. Association for Computational Linguistics.

Melvin Wevers. 2019. Using word embeddings to examine gender bias in dutch newspapers, 1950-1990. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, LChange@ACL 2019, Florence, Italy, August 2, 2019*, pages 92–97. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 629–634. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.