# Data Bias According to Bipol: Men are Naturally Right and It is the Role of Women to Follow Their Lead

**Irene Pagliai**[*1], **Goya van Boven**[*2], **Tosin Adewumi**[*†], **Lama Alkhaled**[†], **Namrata Gurung**[3], **Isabella Södergren**[4] and **Elisa Barney**[†]
[1]University of Göttingen, Germany, [2]Utrecht University, the Netherlands, [*†]Machine Learning Group, LTU, Sweden, [3]QualityMinds GmbH, Germany, [4]Digital Services and Systems, LTU.
[1]irene.pagliai@uni-goettingen.de, [2]j.g.vanboven@students.uu.nl, [†]firstname.lastname@ltu.se,
[3]namrata.gurung@qualityminds.de [4]isasde-5@student.ltu.se | [*]Joint first authors

## Abstract

We introduce new large labeled datasets on bias in 3 languages and show in experiments that bias exists in all 10 datasets of 5 languages evaluated, including benchmark datasets on the English GLUE/SuperGLUE leaderboards. The 3 new languages give a total of almost 6 million labeled samples and we benchmark on these datasets using SotA multilingual pretrained models: mT5 and mBERT. The challenge of social bias, based on prejudice, is ubiquitous, as recent events with AI and large language models (LLMs) have shown. Motivated by this challenge, we set out to estimate bias in multiple datasets. We compare some recent bias metrics and use bipol, which has explainability in the metric. We also confirm the unverified assumption that bias exists in toxic comments by randomly sampling 200 samples from a toxic dataset population using the confidence level of 95% and error margin of 7%. Thirty gold samples were randomly distributed in the 200 samples to secure the quality of the annotation. Our findings confirm that many of the datasets have male bias (prejudice against women), besides other types of bias. We publicly release our new datasets, lexica, models, and codes.

## 1 Introduction

The problem of social bias in data is a pressing one. Recent news about social bias of artificial intelligence (AI) systems, such as Alexa[1] and ChatGPT,[2] shows that the age-old problem persists with data, which is used to train machine learning (ML) models. Social bias is the inclination or prejudice for, or against, a person, group or idea, especially in a way that is considered to be unfair, which may be based on race, religion or other factors (Bellamy et al., 2018; Antoniak and Mimno, 2021; Mehrabi et al., 2021; Alkhaled et al., 2023). It can also involve stereotypes that generalize behavior to groups (Brownstein, 2019). It can unfairly skew the output of ML models (Klare et al., 2012; Raji et al., 2020). Languages with fewer resources than English are also affected (Rescigno et al., 2020; Chávez Mulsa and Spanakis, 2020; Kurpicz-Briki, 2020). For example, in Italian, the female gender is under-represented due to the phenomena such as the “inclusive masculine" (when the masculine is over-extended to denote groups of both male and female referents) (Luccioli et al.; Vanmassenhove and Monti, 2021).

In this work, we are motivated to address the research question of *how much bias exists in the text data of multiple languages, if at all bias exists in them*? We particularly investigate 6 benchmark datasets on the English GLUE/SuperGLUE leaderboards (Wang et al., 2018, 2019) and one dataset each for the other 4 languages: Italian, Dutch, German, and Swedish. First, we train SotA multilingual Text-to-Text Transfer Transformer (mT5) (Xue et al., 2021) and multilingual Bidirectional Encoder Representations from Transformers (mBERT) models for bias classification on the multi-axes bias dataset (MAB) for each language, in a similar setup as Alkhaled et al. (2023). For the evaluations, we search through the literature to compare different metrics or evaluation methods as shown in Table 1 and discussed in Section 2. This motivates our choice of bipol, the multi-axes bias metric, which we then compare in experiments with a lexica baseline method. In addition, to confirm the unverified assumption that toxic comments contain bias (Sap et al., 2020; Alkhaled et al., 2023), we annotate 200 randomly-selected samples from the training set of the English MAB.

### Our Contributions

- We make available new large labeled datasets on bias of almost 2 million samples each for

[1]bbc.com/news/technology-66508514
[2]bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results

| Metric/Evaluator | Axis | Terms |
|---|---|---|
| Winogender (Rudinger et al., 2018) | 1 | 60 |
| WinoBias (Zhao et al., 2018) | 1 | 40 |
| StereoSet (Nadeem et al., 2021) | 4 | 321 |
| GenBiT (Sengupta et al., 2021) | 1 | - |
| CrowS-Pairs (Nangia et al., 2020) | 9 | 3,016 |
| Bipol (Alkhaled et al., 2023) | >2, <13 | >45, <466 |

Table 1: Comparison of some bias evaluation methods.

3 languages: Italian, Dutch, and German.[3]

- We make available lexica of sensitive terms for bias detection in the 3 languages.
- We confirm the unverified assumption in the underlying datasets of MAB (Social Bias Inference Corpus v2 (SBICv2) and Jigsaw) (Alkhaled et al., 2023) that toxic comments contain bias.

The rest of this paper is organized as follows. In Section 2, we discuss the literature review of related work. In Section 3, we briefly discuss the *bipol* metric. In Section 4, we explain the steps involved in the methodology and the datasets we use. In Section 5, we present our findings and discuss them. In Section 6, we end with the conclusion and possible future work.

## 2 Literature Review

Although English usually gets more support and attention in the literature, there have been attempts at measuring and mitigating bias in other languages. Testing for the presence of bias in Italian often has a contrastive perspective with English, with a focus on gender bias (Gaido et al., 2021; Rescigno et al., 2020). MuST-SHE (Bentivogli et al., 2020) and gENder-IT (Vanmassenhove and Monti, 2021) are examples of gender bias evaluation sets. Going beyond gender bias, Kurpicz-Briki and Leoni (2021) and Huang et al. (2020) also identified biases related to people's origin and speakers' age. It is essential to remember that the mentioned biases can be vehicles for misogynous and hateful discourse (El Abassi and Nisioi, 2020; Attanasio et al., 2022; Merenda et al., 2018).

Bias studies for Dutch mostly consider binary gender bias. Chávez Mulsa and Spanakis (2020) investigate gender bias in Dutch static and contextualized word embeddings by creating Dutch versions of the Word/Sentence Embedding Association Test (WEAT/SEAT) (Caliskan et al., 2017; May et al., 2019). WEAT measures bias in word embeddings and can be limited in scope, in addition to having sensitivity to seed words. McCurdy and Serbetci (2020) perform a similar evaluation in a multilingual setup to compare the effect of grammatical gender saliency across languages. Several works use different NLP techniques to evaluate bias in corpora of Dutch news articles (Wevers, 2019; Kroon et al., 2020; Kroon and van der Meer, 2021; Fokkens et al., 2018) and literary texts (Koolen and van Cranenburgh, 2017).

In Kurpicz-Briki (2020), bias is measured with regards to place of origin and gender in German word embeddings using WEAT (Caliskan et al., 2017). In Kurpicz-Briki and Leoni (2021), an automatic bias detection method (BiasWords) is presented, through which new biased word sets can be identified by exploring the vector space around the well-known word sets that show bias. In the template-based study of Cho et al. (2021), on gender bias in translations, the accuracy of gender inference was measured for multiple languages including German. It was found that, particularly for German, the inference accuracy and disparate impact were lower for female than male, implying that certain translations were wrongly performed for cases that required female inference. Since German is a grammatically gendered, morphologically rich language, Gonen and Goldberg (2019) found that debiasing methods of Bolukbasi et al. (2016) were ineffective on german word embeddings.

For Swedish, the main focus of bias research appears to be on gender. Sahlgren and Olsson (2019) show with their experiments that gender bias is present in pretrained Swedish language models. Katsarou et al. (2022) and Precenth (2019) found that the male gender tends to be associated with higher-status professions. A study with data from mainstream news corpora by Devinney et al. (2020) shows that women are associated with concepts like family, communication and relationships.

## 3 Bipol

For the purpose of this work, we summarize *bipol* here but details are discussed in Alkhaled et al. (2023). The *bipol* metric uses a two-step mechanism for estimating bias in text data: binary classification and sensitive term evaluation using lexica. It has maximum and minimum values of 1 and 0, respectively. Bipol is expressed in Equations 1b and 1c from the main Equation 1a, where $b_c$ is

[3] github.com/LTU-Machine-Learning/bipolmulti

the classification component and $b_s$ is the sensitive term evaluation component.

$$b = \begin{cases} b_c.b_s, & \text{if } b_s > 0 \\ b_c, & \text{otherwise} \end{cases} \tag{1a}$$

$$b_c = \frac{tp + fp}{tp + fp + tn + fn} \tag{1b}$$

$$b_s = \frac{1}{r}\sum_{t=1}^{r}\left(\frac{1}{q}\sum_{x=1}^{q}\left(\frac{|\sum_{s=1}^{n} a_s - \sum_{s=1}^{m} c_s|}{\sum_{s=1}^{p} d_s}\right)_x\right)_t \tag{1c}$$

In step 1, a trained model is used to classify all the samples. The ratio of the biased samples to the total samples predicted is determined. The ***tp***, ***fp***, ***tn***, and ***fn*** are values of the true positives, false positives, true negatives, and false negatives, respectively. Since there's hardly a perfect classifier, the positive error rate is usually reported. False positives are known to exist in similar classification systems like spam detection and automatic hate speech detection (Heron, 2009; Feng et al., 2018).

Step 2 is similar to *term frequency-inverse document frequency* (TF-IDF) in that it is based on term frequency (Salton and Buckley, 1988; Ramos et al., 2003), Biased samples from step 1 are evaluated token-wise along all possible bias axes, using all the lexica of sensitive terms. An axis is a domain such as gender or race. Tables 2 and 3 provide the lexica sizes. For English and Swedish, we use the same lexica released by Alkhaled et al. (2023) and Adewumi et al. (2023b), respectively. For the other 3 languages, we create new lexica of terms (e.g. she & her) associated with specific gender or stereotypes from public sources.[4] Some of the terms in the lexica were selected from the sources based on the topmost available. These may also be expanded as needed, since bias terms are known to evolve (Haemmerlie and Montgomery, 1991; Antoniak and Mimno, 2021). The non-English lexica are small because fewer terms are usually available in other languages compared to the high-resource English language and we use the same size across the languages to be able to compare performance somewhat. The Appendix lists these terms.

Equation 1c first finds the absolute difference between the two maximum summed frequencies in the types of an axis ($|\sum_{s=1}^{n} a_s - \sum_{s=1}^{m} c_s|$), where $n$ and $m$ are the total terms in a sentence along an axis. For example, in the sentence *´Women!!! PERSON taught you better than that. Shame on you!'*, female terms = 1 while male terms = 0. This is then divided by the summed frequencies of all the terms ($d_s$) in that axis ($\sum_{s=1}^{p} d_s$). The operation is performed for all axes ($q$) and the average taken ($\frac{1}{q}\sum_{x=1}^{q}$). It is performed for all the biased samples ($r$) and the average taken ($\frac{1}{r}\sum_{t=1}^{r}$ ).

[4] fluentu.com/blog/italian/italian-nouns, en.wiktionary.org/wiki/Category:Italian_offensive_terms, Dutch_profanity, Category:German_ethnic_slurs

| Axis | Type 1 | Type 2 | Type 3 |
|---|---|---|---|
| Racial | 84 (black) | 127 (white) | |
| Gender | 76 (female) | 46 (male) | |
| Religious | 180 (christian) | 465 (muslim) | 179 (hindu) |

Table 2: English lexica sizes (Alkhaled et al., 2023). An example of racial-black is *negro*.

| Axis | Type 1 | Type 2 |
|---|---|---|
| Gender | 17 (female) | 17 (male) |
| Racial | 10 (black) | 10 (white) |

Table 3: Italian, Dutch, German and Swedish Lexica sizes. An example of gender-female is *woman*.

## 4 Methodology

### 4.1 Datasets

#### 4.1.1 MAB

The Italian, Dutch and German datasets were machine-translated from MAB[5] with the high-quality Helsinki-NLP model (Tiedemann and Thottingal, 2020). Each translation took about 48 hours on one GPU. Examples from the data are provided in Table 4. Table 5 provides statistics about the datasets. For quality control (QC), we verified translation by back-translating some random samples using Google NMT. Personal identifiable information (PII) was removed from the MAB dataset using the spaCy library. The 3 datasets are used to train new bias classifiers. We also train on the original English and the Swedish.

**Machine-Translation issues:** Culture-specific biases may not be represented in the MAB versions for the translated languages because the original dataset is in English. This is a limitation. However, bias is also a universal concern, such that there are examples that span across cultures. For instance, the examples in Table 4 are of universal concern because individuals with non-conforming bodies and women should be respected, regardless of culture or nationality. Hence, the MAB versions

[5] The reference provides details of the annotation of the base data.

| comment_text | label | old_id | id |
|---|---|---|---|
| A fitness ranking that drops each month as more and more fat people move to Colorado. | biased | 6231681 | 167 |
| Een fitness ranking die elke maand daalt als meer en meer dikke mensen verhuizen naar Colorado. | | | |
| Ein Fitness-Ranking, das jeden Monat fällt, wie mehr und mehr fette Menschen nach Colorado zu bewegen. | | | |
| Un ranking di fitness che scende ogni mese come sempre più persone grasse si spostano in Colorado. | | | |
| En fitness ranking som sjunker varje månad som fler och fler feta människor flyttar till Colorado. | | | |
| | | | |
| Women ! ! ! PERSON taught you better than that. Shame on you! | biased | 6233025 | 628 |
| Vrouwen ! ! ! PERSOON heeft je beter geleerd dan dat. Je moet je schamen! | | | |
| Frauen!!!! PERSON lehrte Sie besser als das. Schande über Sie! | | | |
| Donne ! ! ! Person ti ha insegnato meglio di così, vergognati! | | | |
| Kvinnor ! ! !- Han lärde dig bättre än så. Skäms på dig! | | | |
| | | | |

Table 4: **English**, **Dutch**, **German**, **Italian**, and **Swedish** examples from the MAB dataset. "PERSON" is the anonymization of a piece of personal identifiable information (PII) in the dataset.

| Set | Biased | Unbiased | Total |
|---|---|---|---|
| Training | 533,544 | 1,209,433 | 1,742,977 |
| Validation | 32,338 | 69,649 | 101,987 |
| Test | 33,470 | 68,541 | 102,011 |
| | 599,352 | 1,347,623 | 1,946,975 |

Table 5: MAB dataset split

are relevant for bias detection, though they were translated.

### 4.1.2 Evaluation datasets

Ten datasets are evaluated for bias in this work. All are automatically preprocessed before evaluation, the same way the training data were preprocessed. This includes removal of IP addresses, emojis, URLs, special characters, emails, extra spaces, numbers, empty text rows, and duplicate rows. All texts are then lowercased.

We selected datasets that are available on the HuggingFace (Wolf et al., 2020) Datasets. We evaluated the first 1,000 samples of each training split due to resource constraints. The understanding is that if bias is detected in these samples, then scaling over the entire dataset means there's probability of more bias. For English, we evaluated the sentence column of Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019), the sentence column of Question-Answering Natural Language Inference (QNLI) (Wang et al., 2018), the sentennce1 column of Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005), the premise column of Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018), the premise column of the CommitmentBank (CB) dataset (De Marneffe et al., 2019), and the passage column of Reading Comprehension with Commonsense Reasoning Dataset (ReCoRD) (Zhang et al., 2018). For Italian, we evaluated the context column of the Stanford Question Answering Dataset (SQuAD) (Croce et al., 2018; Rajpurkar et al., 2016); for Dutch, the sentence1 column of the Semantic Textual Similarity Benchmark (STSB) (Cer et al., 2017); for German, the text column of the German News Articles Datasets 10k (GNAD10) (Schabus et al., 2017); for Swedish, the premise of the CB.

### 4.2 Annotation for the assumption confirmation

To verify the assumption that toxic comments contain bias, we randomly selected 200 samples from the training set of MAB-English for annotation on Slack, an online platform. The selection of 200 samples is based on an error margin of 7% and a confidence level of 95%. To ensure high-quality annotation, we use established techniques for this task: 1) the use of gold (30) samples, 2) multiple (i.e. 3) annotators, and 3) minimum qualification of undergraduate study for annotators. Each annotator was paid 25 U.S. dollars and the it took about 2 hours to complete the annotation on average. We mixed the 30 gold samples with the 200, to verify the annotation quality of each annotator, as they were required to get, at least, 16 correctly for their annotation to be accepted. The 30 gold samples are samples with unanimous agreement in the original Jigsaw or SBICv2 data, which make up the MAB. We provide inter-annotator agreement (IAA) using Jaccard similarity coefficient (intersection over union) and credibility unanimous score (CUS) (Adewumi et al., 2023a) (intersection over sample size).

### 4.3 Experiments

We selected two state-of-the-art (SotA) pre-trained, multilingual models for experiments to compare their macro F1 performance: mT5-small and mBERT-base. These are from the HuggingFace hub. We further report the mT5 positive error rate of predictions. The mT5-small has 300 million parameters (Xue et al., 2021) while mBERT-Base has 110 million parameters. We trained only on the MAB datasets and evaluated using only the mT5 model, the better model of the 2, as will be observed in Section 5. For the CB and ReCoRD datasets, we evaluate all samples since they contain only about 250 and 620 entries, respectively. We used wandb (Biewald, 2020) for hyper-parameter exploration, based on Bayesian optimization. For mT5, we set the maximum and minimum learning rates as 5e-5 and 2e-5 while the maximum and minimum epochs are 20 and 4, respectively. One epoch is equivalent to the ratio of the total number of samples to the batch size (i.e. the steps). We used a batch size of 8 because higher numbers easily resulted in memory challenges.

For mBERT, we set the learning rates and epochs as with mT5. However, we explore over batch sizes of 8, 16 and 32. For both models, we set the maximum input sequence length to 512. Training took, on average, about 7.3 hours per language per epoch for mBERT while it was 6 hours for mT5. For all the experiments, we limit the run counts to 2 per language because of the long training time each takes on average. The average scores of the results are reported. The saved models with the lowest losses were used to evaluate the datasets. All the experiments were performed on two shared Nvidia DGX-1 machines that run Ubuntu 20.04 and 18.04. One machine has 8 x 40GB A100 GPUs while the other has 8 x 32GB V100 GPUs.

The lexica baseline, compared in experiments, is similar to the equation of the second step in bipol. It does not consider bias semantically and uses term frequencies, similarly to TF-IDF. It uses the same lexica as bipol. Its maximum and minimum values are 1 and 0, respectively.

## 5 Results and Discussion

From Table 6, we observe that all mT5 results are better than those of mBERT across the languages. The two-sample t-test of the difference of means between all the corresponding mT5 and mBERT scores have $p$ values $< 0.0001$ for alpha of 0.05, showing the results are statistically significant. It appears better hyper-parameter search may be required for the mBERT model to converge and achieve better performance. The best macro F1 result is for English mT5 at 0.787. This is not surprising, as English has the largest amount of training data for the pre-trained mT5 model (Xue et al., 2021). This occurred at the learning rate of 2.9e-5 and step 1,068,041.

| | **macro F1** ↑ (s.d.) | | **mT5 error** ↓ |
|---|---|---|---|
| **MAB version** | **mBERT** | **mT5** | **fp/(fp+tp)** |
| English | 0.418 (0.01) | 0.787 (0) | 0.261 |
| Italian | 0.429 (0) | 0.768 (0) | 0.283 |
| Dutch | 0.419 (0.01) | 0.768 (0) | 0.269 |
| German | 0.418 (0.01) | 0.769 (0) | 0.261 |
| Swedish | 0.418 (0.01) | 0.768 (0) | 0.274 |

Table 6: Average F1 scores on the validation sets.

Figures 1 and 2 depict the validation sets macro F1 and loss line graphs for the 2 runs for the 5 languages, respectively. From Table 7, we observe that all the evaluated datasets have biases, though seemingly little (but important) when compared to the maximum of 1. We say important because many of the datasets contain small number of sam-

| English | bipol scores $b_c$ | $b_s$ | ↓ (s.d.) bipol *(b)* | baseline ↓ |
|---|---|---|---|---|
| CB | 0.096 | 0.875 | 0.084 (0) | 0.88 |
| CoLA | 0.101 | 0.943 | 0.095 (0) | 0.958 |
| ReCoRD | 0.094 | 0.852 | 0.025 (0) | 0.829 |
| MRPC | 0.048 | 0.944 | 0.045 (0) | 0.957 |
| MNLI | 0.063 | 0.833 | 0.053 (0) | 0.965 |
| QNLI | 0.03 | 0.933 | 0.028 (0) | 0.945 |
| Italian | | | | |
| SQuAD | 0.014 | 0 | 0.014 (0) | 0.989 |
| Dutch | | | | |
| STSB | 0.435 | 0.992 | 0.432 (0) | 0.987 |
| German | | | | |
| GNAD10 | 0.049 | 0.502 | 0.025 (0) | 1 |
| Swedish | | | | |
| CB | 0.08 | 0.938 | 0.075 (0) | 0.97 |

Table 7: Average bipol & lexica baseline scores.

ples yet they can be detected. Furthermore, a low value does not necessarily diminish the weight of the effect of bias in society or the data but we leave the discussion about what amount should be tolerated open for the NLP community. Our recommendation is to have a bias score as close to zero as possible. On the other hand, the lexica baseline appears overly confident of much more bias, which is incorrect because the method fails to exclude unbiased text in its evaluation, which is a shortcoming of methods based solely on it. The Dutch STSB is higher than the other bipol scores because of the higher bipol classifier component score of 0.435, which may be because of the nature of the dataset.

### 5.1 Error analysis & qualitative results

According to the error matrix in Figure 3, the mT5 model is better at correctly predicting unbiased samples. This is because of the higher unbiased samples in the training data of MAB. In Table 8, the first example for the English CB contains a stereotypical statement "*men are naturally right and it is the role of women to follow their lead*", leading to the correct biased prediction by the model. Similarly, this correct prediction is made in the Swedish CB. We notice over-generalization (May et al., 2019; Nadeem et al., 2021) in the correct examples for the CoLA predictions, where "*every*" is used. The table also shows some incorrect predictions.

### 5.2 Consistent prediction with perturbation

An interesting property of relative consistency that we observed with the model predictions, as demonstrated with the CoLA dataset, is that when sentences are perturbed, the model mostly maintains its predictions, as long as the grounds for prediction (in this case - over-generalization) remain the same. The perturbations are inherent in the CoLA dataset itself, as the dataset is designed that way. Some examples are provided in Table 9 in the Appendix, where 6 out of 8 are correctly predicted. This property is repeated consistently in other examples not shown here.

### 5.3 Explainability by graphs

We show explainability by visualization using graphs. Bipol produces a dictionary of lists for every evaluation and we show the *top-5 frequent terms* bar graph for the GNAD10 dataset in Figure 4, which has overall male bias. Many of the 10 evaluated datasets display overall male bias.

### 5.4 Assumption confirmation through annotation

The results of the annotation of the 200 MAB samples reveal that toxic comments do contain bias. This is shown in Figure 5. The Jaccard similarity coefficient and CUS of IAA are 0.261[6] and 0.515, respectively, given that over 50% is the intersection of unanimous decision.

## 6 Conclusion

The findings of this work show that bias besets Natural Language Processing (NLP) datasets regardless of language, including benchmark datasets on the GLUE/SuperGLUE leaderboards. We introduced MAB datasets in 3 languages for training models in bias detection. Each has about 2 million labeled samples. We also contribute lexica of bias terms for the languages. In addition, we verified the assumption that toxic comments contain bias. It may be impossible to completely remove bias from data or models, since they reflect the real world, but resources for estimating bias can provide insight into mitigation strategies for reducing bias. Future work may explore ways of minimizing false positives in classifiers to make them more effective. One may also explore how this work scales to other languages or how multilignual models compare to language-specific monolingual models or large language models (LLMs). Regarding culture-specific biases in datasets, one solution will be to

[6] Not to be interpreted using Kappa for 2 annotators on 2 classes. Ours involved 3 annotators

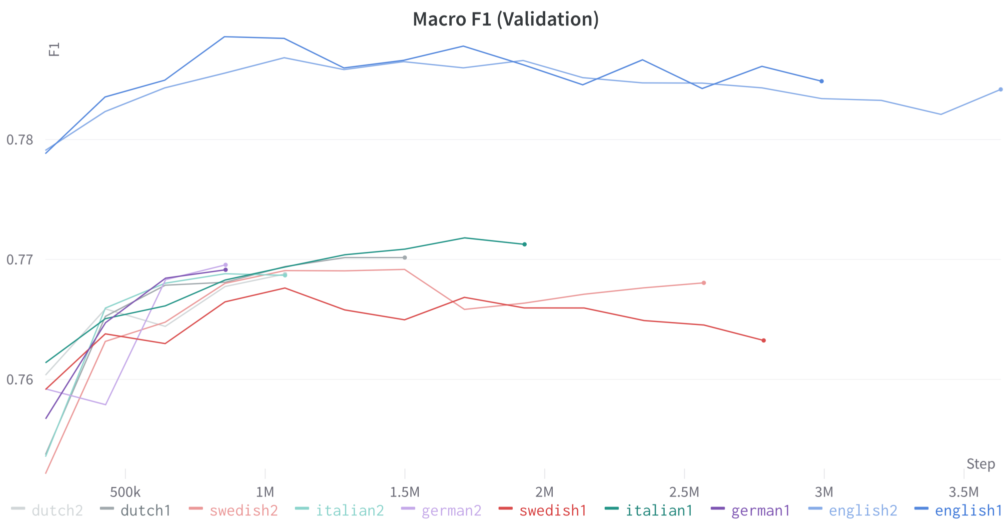


Figure 1: Macro F1 of the validation set for the 5 languages, as generated by wandb.

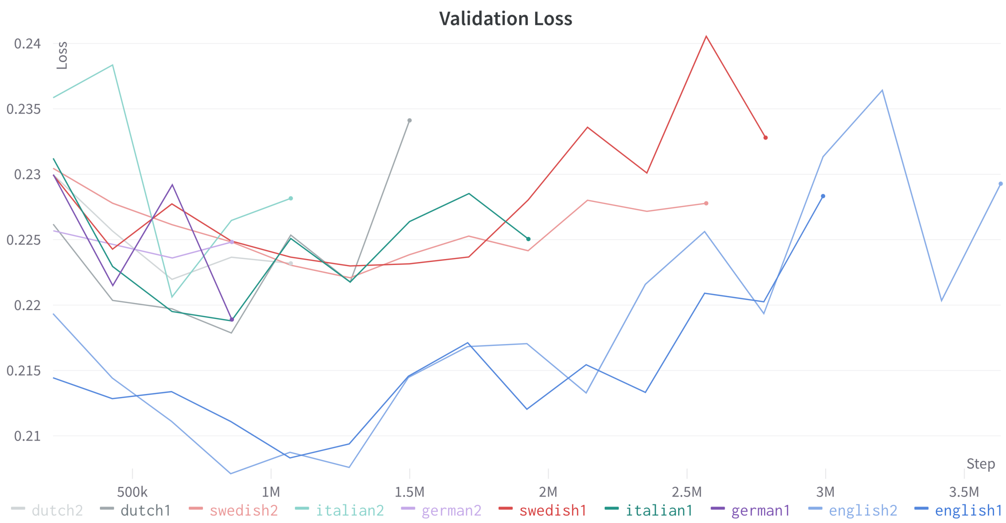


Figure 2: Loss on the validation set for the 5 languages, as generated by wandb.

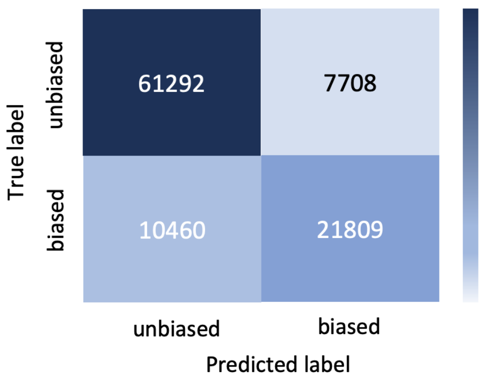


Figure 3: Error matrix of mT5 on MAB-English

collect data from the specific cultures/nationalities to capture these biases.

## Ethics statement & limitation

The authors took care while providing examples of data samples, despite some containing stereotypes or toxic content. The classifiers trained for estimating the biases in this work are limited in effectiveness, as shown in the results, hence a result of 0 on any dataset does not necessarily indicate a bias-free dataset. The original MAB was annotated by humans who may have their personal biases based on cultural or demographic backgrounds. This means the final annotations may not be seen as absolute ground truth of social biases.

| comment_text | prediction | csv id |
|---|---|---|
| **Correct predictions from the English CB** | | |
| nicky approached her with the assumption that men are naturally right and it is the role of women to follow their lead. constance whose confidence was growing daily was not prepared to give in to nickys wishes merely because of his sex. if she felt he was right then she agreed with him. | biased | 85 |
| b: thats true. a: so. b: uh the other argument is that the death penalty is a deterrent and i really dont uh agree with that. i dont think anyone who would commit uh a crime that would get them the death penalty would stop at the moment and say well i was about to kill and dismember this person but oh if they catch me theyre going to kill me so i better not do it. i just dont think uh that it works that way. a: yeah. i dont think its done. | biased | 133 |
| **Incorrect prediction** | | |
| b: yeah and the ground will filter some of it but not all of it. a: no not when you figure i didnt realize one cow produces that much manure | biased | 137 |
| | | |
| **Correct predictions from the English CoLA** | | |
| if you give him enough opportunity every senator will succumb to corruption. | biased | 266 |
| every senator becomes more corrupt the more lobbyists he talks to. | biased | 277 |
| **Incorrect prediction** | | |
| bill squeezed the puppet through the hole. | biased | 82 |
| | | |

Table 8: Qualitative examples of apparently correct and incorrect predictions in some of the English datasets, based on the mT5 model.

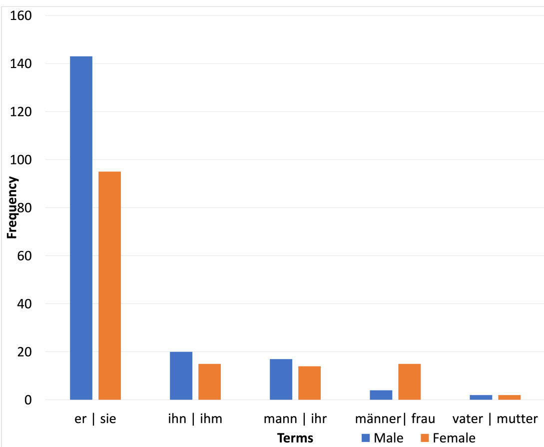


Figure 4: Top 5 frequent terms in the GNAD10 dataset (paired terms are only for comparison).

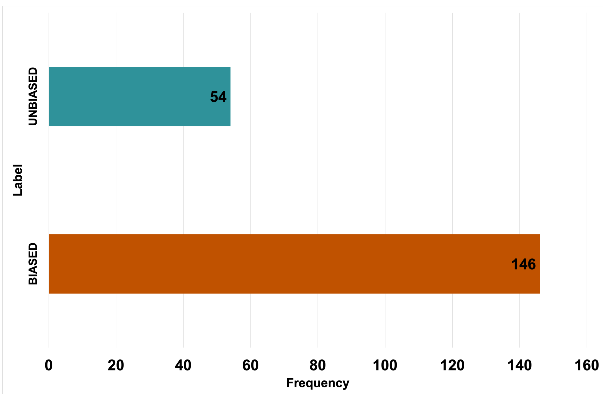


Figure 5: Annotation confirms assumption about toxic comments.


## Acknowledgments

The authors wish to thank the anonymous reviewers for their valuable feedback. This work is partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation and counterpart funding from Luleå University of Technology (LTU). The authors also thank Björn Backe for his insightful comments during the writing of this paper.

## References

Tosin Adewumi, Mofetoluwa Adeyemi, Aremu Anuoluwapo, Bukola Peters, Happy Buzaaba, Oyerinde Samuel, Amina Mardiyyah Rufai, Benjamin Ajibade, Tajudeen Gwadabe, Mory Moussou Koulibaly Traore, Tunde Oluwaseyi Ajayi, Shamsuddeen Muhammad, Ahmed Baruwa, Paul Owoicho, Tolulope Ogunremi, Phylis Ngigi, Orevaoghene Ahia, Ruqayya Nasir, Foteini Liwicki, and Marcus Liwicki. 2023a. Afriwoz: Corpus for exploiting cross-lingual transfer for dialogue generation in low-resource, african languages. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Tosin Adewumi, Isabella Södergren, Lama Alkhaled, Sana Sabah Sabry, Foteini Liwicki, and Marcus Liwicki. 2023b. Bipol: Multi-axes evaluation of bias with explainability in benchmark datasets. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria.

Lama Alkhaled, Tosin Adewumi, and Sana Sabah Sabry. 2023. Bipol: A novel multi-axes bias evaluation metric with explainability for nlp. *Natural Language Processing Journal*, 4:100030.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. 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 1889–1904.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. *arXiv preprint arXiv:2203.09192*.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Michael Brownstein. 2019. Implicit Bias. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2019 edition. Metaphysics Research Lab, Stanford University.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Won Ik Cho, Jiwon Kim, Jaeyeong Yang, and Nam Soo Kim. 2021. Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 449–457.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Hannah Devinney, 1974 Björklund, Jenny, and Henrik Björklund. 2020. Semi-supervised topic modeling for gender bias discovery in english and swedish. *EQUITBL Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79 – 92.

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

Samer El Abassi and Sergiu Nisioi. 2020. Mdd@ ami: Vanilla classifiers for misogyny identification. *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020*, page 55.

Bo Feng, Qiang Fu, Mianxiong Dong, Dong Guo, and Qiang Li. 2018. Multistage and elastic spam detection in mobile social networks through deep learning. *IEEE Network*, 32(4):15–21.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. How to split: the effect of word segmentation on gender bias in speech translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589, Online. 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. *arXiv preprint arXiv:1903.03862*.

Frances M Haemmerlie and Robert L Montgomery. 1991. Goldberg revisited: Pro-female evaluation bias and changed attitudes toward women by engineering students. *Journal of Social Behavior and Personality*, 6(2):179.

Simon Heron. 2009. Technologies for spam detection. *Network Security*, 2009(1):11–15.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *arXiv preprint arXiv:2002.10361*.

Styliani Katsarou, Borja Rodríguez-Gálvez, and Jesse Shanahan. 2022. Measuring gender bias in contextualized embeddings. *Computer Sciences and Mathematics Forum*, 3(1).

Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Anne C Kroon, Damian Trilling, Toni GLA van der Meer, and Jeroen GF Jonkman. 2020. Clouded reality: News representations of culturally close and distant ethnic outgroups. *Communications*, 45(s1):744–764.

Anne C Kroon and Toni GLA van der Meer. 2021. Who's to fear? implicit sexual threat pre and post the "refugee crisis". *Journalism Practice*, pages 1–17.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings.

Mascha Kurpicz-Briki and Tomaso Leoni. 2021. A world full of stereotypes? further investigation on origin and gender bias in multi-lingual word embeddings. *Frontiers in big Data*, 4:20.

Alessandra Luccioli, Silvia Bernardini, and Raffaella Baccolini. Stereotipi di genere e traduzione automatica dall'inglese all'italiano: uno studio di caso sul femminile nelle professioni.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Katherine McCurdy and Oguz Serbetci. 2020. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *arXiv preprint arXiv:2005.08864*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. Source-driven representations for hate speech detection. *Computational Linguistics CLiC-it 2018*, page 258.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained 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 (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Rasmus Precenth. 2019. Word embeddings and gender stereotypes in swedish and english.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 145–151, New York, NY, USA. Association for Computing Machinery.

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.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, and Andy Way. 2020. A case study of natural gender phenomena in translation. a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In *CLiC-it*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Magnus Sahlgren and Fredrik Olsson. 2019. Gender bias in pretrained swedish embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa 2019, Turku, Finland, September 30 - October 2, 2019*, pages 35–43. Linköping University Electronic Press.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

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*, pages 5477–5490, Online. Association for Computational Linguistics.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1241–1244, Tokyo, Japan.

Kinshuk Sengupta, Rana Maher, Declan Groves, and Chantal Olieman. 2021. Genbit: measure and mitigate gender bias in language datasets. *Microsoft Journal of Applied Research*, 16:63–71.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Eva Vanmassenhove and Johanna Monti. 2021. gender-it: An annotated english-italian parallel challenge set for cross-linguistic natural gender phenomena. *arXiv preprint arXiv:2108.02854*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. 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 Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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.

Melvin Wevers. 2019. Using word embeddings to examine gender bias in dutch newspapers, 1950-1990. *arXiv preprint arXiv:1907.08922*.

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.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

# A Appendix

### Swedish lexica and the English equivalent

1. Gender-female
   i) hon (she) ii) hennes (her) iii) flicka (girl) iv) mor (mother) v) kvinna (woman) vi) dotter (daughter) vii) mormor (grandmother) viii) dam (lady) ix) sondotter (son's daughter) x) dotterdotter (daughter's daughter) xi) tjej (girl) xii) tjejer (girls) xiii) gumma (old woman) xiv) fru (wife) xv) slampa (slut) xvi) slyna (slut) xvii) lebb (lesbian)

2. Gender-male
   i) han (he) ii) pojke (boy) iii) kille (boy) iv) far (father) v) farfar (grandfather) vi) hans (his) vii) man (man) viii) son (son) ix) make (husband) x) herre (lord) xi) herrar (lords) xii) sonson (son's son) xiii) dotterson (daughter's son) xiv) gubbe (old man) xv) farbro (uncle) xvi) broson (nephew) xvii) män (men)

3. Racial-black
   i) neger (nigger) ii) blåneger (nigger) iii) blåland (nigger) iv) blatte (dark immigrants) v) svartskalle (black head) vi) turk (non-Swedish) vii) sosse (political slur) viii) svarting (blackness) ix) partysvenska (political slur) x) nigga (nigger)

4. Racial-white
   i) svenne (Swedish stereotype) ii) turk (non-Swedish) iii) jugge (derogatory term) iv) sosse (political slur) v) hurrare (Finnish Swedish) vi) lapp (Sami people) vii) Ang mo (red hair) viii) partysvenska (political slur) ix) Ann (White woman) x) rutabaga (Swede)

### Example of predictions for the CoLA dataset

| comment_text | prediction | csv id |
|---|---|---|
| if you give him enough opportunity every senator will succumb to corruption. | biased | 266 |
| you give him enough opportunity and every senator will succumb to corruption. | biased | 267 |
| we gave him enough opportunity and sure enough every senator succumbed to corruption. | unbiased | 268 |
| if you give any senator enough opportunity he will succumb to corruption. | biased | 269 |
| you give any senator enough opportunity and he will succumb to corruption. | biased | 270 |
| you give every senator enough opportunity and he will succumb to corruption. | biased | 271 |
| we gave any senator enough opportunity and sure enough he succumbed to corruption. | biased | 272 |
| we gave every senator enough opportunity and sure enough he succumbed to corruption. | unbiased | 273 |

Table 9: Mostly consistent correct prediction with perturbation in the CoLA dataset.