Bipol: Multi-axes Evaluation of Bias with Explainability in Benchmark Datasets

Tosin Adewumi*+, Isabella Södergren++, Lama Alkhaled+, Sana Sabah Sabry+, Foteini Liwicki+ and Marcus Liwicki+
+Machine Learning Group, EISLAB, ++Digital Services and Systems
Luleå University of Technology, Sweden
+firstname.lastname@ltu.se, ++isasde-5@student.ltu.se

Caution: This paper contains examples, from datasets, of what some may consider as stereotypes or offensive text.

Abstract

We investigate five English NLP benchmark datasets (on the superGLUE leaderboard) and two Swedish datasets for bias, along multiple axes. The datasets are the following: Boolean Question (Boolq), CommitmentBank (CB), Winograd Schema Challenge (WSC), Winogender diagnostic (AXg), Recognising Textual Entailment (RTE), Swedish CB, and SWEDN. Bias can be harmful and it is known to be common in data, which ML models learn from. In order to mitigate bias in data, it is crucial to be able to estimate it objectively. We use bipol, a novel multi-axes bias metric with explainability, to estimate and explain how much bias exists in these datasets. Multilingual, multi-axes bias evaluation is not very common. Hence, we also contribute a new, large Swedish bias-labeled dataset (of 2 million samples), translated from the English version and train the SotA mT5 model on it. In addition, we contribute new multi-axes lexica for bias detection in Swedish. We make the codes, model, and new dataset publicly available.

1 Introduction

Recent advances in artificial intelligence (AI), large language models (LLM), and chatbots have raised concerns about their potential risks to humanity (Bender et al., 2021; Adewumi et al., 2022; Yudkowsky et al., 2008).[1] One major concern is the issue of social bias, particularly with the data AI models are trained on. Bias, which can be harmful, is the unfair prejudice in favor of or against a thing, person or group (Maddox, 2004; Dhamala et al., 2021; Mehrabi et al., 2021; Antoniak and Mimno, 2021). Measuring bias in text data can be challenging because of the axes that may be involved (e.g. religious or gender bias).

In this work, our motivation is to determine whether social bias exists in NLP benchmark datasets and estimate it. After reviewing some potential bias methods, as discussed in Section 2, we settled for the recent bipol (Alkhaled et al., 2023) because of its advantages. It is a metric that estimates bias along multiple axes in text data and provides an explanation for its scores, unlike other metrics. We investigate social bias in benchmark datasets that are available on the English SuperGLUE leaderboard and two Swedish datasets. The SuperGLUE was introduced by Wang et al. (2019) and provides benchmark datasets for different NLP tasks. Benchmark datasets are datasets for comparing the performance of algorithms for specific use-cases (Dhar and Shamir, 2021; Paullada et al., 2021). Such datasets have been the foundation for some of the significant advancements in the field (Paullada et al., 2021). We investigate the following English datasets: Boolq (Clark et al., 2019), CB (De Marneffe et al., 2019), WSC (Levesque et al., 2012), AXg (Rudinger et al., 2018a), and RTE (Wang et al., 2019). The Swedish datasets are the Overlim *CB* and SWEDN. We discuss more about the datasets in Section 3.2.

Our contributions Firstly, we show quantitatively and through explainability that bias exists in the datasets. The findings correlate with characteristics of bias, such as heavy lopsidedness (Zhao et al., 2018). This work will provide researchers with insight into how to mitigate bias in text data and possibly add impetus to the conversation on whether it is even ethical to remove these social biases from data, because they represent the real world. Secondly, we create and release, possibly, the largest labeled dataset and lexica for bias de-

[1]bbc.com/news/world-us-canada-65452940

Proceedings of Recent Advances in Natural Language Processing, pages 1–10
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_001

tection in Swedish (multi-axes bias dataset (MAB)-Swedish) and train a model based on the state-of-the-art (SotA) multilingual T5 (mT5) (Xue et al., 2021). We release our codes, dataset and artefacts publicly. [2]

The rest of this paper is structured as follows. Section 2 discusses some of the previous related work. Section 3 describes the methodology, including details of the characteristics of bipol and the new MAB-Swedish dataset. Section 4 presents the results and discusses some of the qualitative results. In Section 5, we give concluding remarks.

2 Related Work

There have been considerable effort in identifying and measuring the level of bias in datasets (Cryan et al., 2020; Dhamala et al., 2021; Stanley, 1977; Chandrabose et al., 2021). These are usually targeted at gender bias in a binary form (Zhao et al., 2018; Rudinger et al., 2018a). However, studies have shown that the biases in language models for the intersection of gender and race can be greater than those for gender and race individually and that addressing bias along only one axis can lead to more issues (Tan and Celis, 2019; Subramanian et al., 2021). To determine the level of bias in NLP datasets along multiple axes can be a significant challenge, more so that many of these methods admit their approaches may demonstrate the presence of bias but not prove its absence (Zhao et al., 2018; Rudinger et al., 2018a). Table 1 compares some of the methods that have been introduced.

Metric/Evaluator	Axis	Terms
WinoBias (Zhao et al., 2018)	1	40
Winogender (Rudinger et al., 2018a)	1	60
StereoSet (Nadeem et al., 2021)	4	321
Hurtlex (Nozza et al., 2021)	6	1,072
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. (*The upper bounds are not limited by the bipol algorithm but the dataset & lexica.)

Furthermore, Bassignana et al. (2018) proposed a multi-language approach using HurtLex to target misogyny because addressing bias in only the English language is not sufficient for addressing the potential harm to society. In the English language, there are common biases that associate female terms with subjects such as liberal arts and family while associating male terms with subjects such as science (Nosek et al., 2002). There are also more words that sexualize females more than males (Stanley, 1977). Other languages have their own peculiarities (Nozza et al., 2021).

In addition to the various methods identified in Table 1 for quantifying the extent of discrimination or bias, there is also odds ratio (OR), which compares the chance of a specific outcome happening, with a certain exposure, to the likelihood of that outcome happening without the exposure (Szumilas, 2010). Another method is the impact ratio (IR), which calculates the ratio of positive outcomes for a protected group to the general group. In Cryan et al. (2020), they compare lexicon method to model classification for gender bias in English language only. Our approach combines the strengths of both approaches and evaluates on English and Swedish data across multiple axes.

3 Methodology

3.1 Bipol

There are two stages in the implementation of bipol (see 1a) before it gives a final score between 0.0 (zero or undetected bias) and 1.0 (extreme bias). The first stage involves the classification of the data samples (into biased and unbiased categories) using a trained model (see 1b). Ideally, it is the ratio of the number of true positives (tp) to the total samples (true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn)), where fp is preferably zero. However, since the trained models will be evaluated on unseen data, the predicted biased samples are likely to have fp in the numerator as expressed in the equation. The evaluations thus come with positive error rate ($\frac{fp}{fp+tp}$) to establish the lower bound of error for the predictions. A good classifier should minimize the number of fp and maximize the number of tp but there's hardly any perfect classifier, even in other tasks such as spam detection or hate speech (Heron, 2009; Feng et al., 2018).

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

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

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

The second stage evaluates the biased samples for sensitive terms listed in the multi-axes lexica

[2] github.com/LTU-Machine-Learning/bipolswedish.git

(see 1c). It involves finding the difference between the two maximum summed frequencies in the types (e.g. female) of an axis (e.g. gender) ($|\sum_{s=1}^{n} a_s - \sum_{s=1}^{m} c_s|$), which is then divided by the summed frequencies of all the terms in that axis ($\sum_{s=1}^{p} d_s$). The average over all the axes ($\frac{1}{q}\sum_{x=1}^{q}$) is then averaged over all the biased samples ($\frac{1}{r}\sum_{t=1}^{r}$). Table 2 provides the Swedish lexica sizes. The lexica are derived from Adewumi et al. (2020a,b) and Wikipedia[3] and may be expanded as needed. They include terms that may be stereotypically associated with certain groups and specific gender (Cryan et al., 2020; Zhao et al., 2018). The English lexica contain more and are also derived from public sources (Alkhaled et al., 2023).

Axis	Axis type 1	Axis type 2
Gender	17 (female)	19 (male)
Racial	10 (black)	10 (white)

Table 2: Swedish lexica sizes. These may be expanded.

The rationale for using bipol is because of the strengths of the metric. These include 1) the relative simplicity of calculating a score, 2) it is straight-forward to implement since it is based on existing concepts like lexica and classifiers, 3) it captures semantic and term frequency (TF) aspects of data, 4) it has explainability built in, 5) it's possible to determine the error rate of predictions, and 6) it is not limited in the total number of axes that may be used. We acknowledge, however, that it has limitations that are based on the limitations of the tools that may be used to calculate it.

3.2 Datasets

3.2.1 The New MAB-Swedish Dataset

The dataset was machine-translated (from MAB (Alkhaled et al., 2023)) using the Helsinki-NLP model (Tiedemann and Thottingal, 2020), which was mostly trained with guided alignment. The automatic translation took over 48 hours on one GPU. It has 1,946,975 samples, as given in Table 3. Quality control (QC) for the MAB-Swedish involved translation verification by back-translating some random samples using Google NMT before a review by a Swedish speaker. The English version was constructed from two datasets: Jigsaw[4] and the Social Bias Inference Corpus v2 (SBICv2) by (Sap et al., 2020). Personal identifiable information (PII) were removed from the dataset. More details about the annotation of the base datasets for the MAB can be found in Alkhaled et al. (2023). Some examples in the MAB-Swedish are given in Table 4.

[3] en.wikipedia.org/wiki/Swedish_profanity

[4] medium.com/jigsaw/creating-labeled-datasets-and-exploring-the-role-of-human-raters-56367b6db298

Machine-Translation concerns Bias is a universal concern, though there can be culture-specific biases. A stereotype or degrading comment can be considered of universal concern if it is relevant across cultures. For instance, the first example in Table 4 is of universal concern because women should be respected, regardless of nationality or culture. This also applies to the other biased examples in the table. Hence, MAB-Swedish is relevant for bias detection in Swedish, though it was translated. A limitation of the dataset, however, is that it may not contain examples that are specific to the Swedish culture since the original dataset is from English, with specific references to the USA and Canada.

Set	Biased	Unbiased	Total
Training	533,544	1,209,433	1,742,977
Validation	32338	69649	101987
Test	33470	68541	102011
	599,352	1,347,623	1,946,975

Table 3: MAB-Swedish dataset split.

comment_text	label	old_id	id
Den här slynan är galen som skulle läsa en bok av en kvinna. (*This b*tch is nuts. Who would read a book by a woman.*)	biased	239612	1212584
9 av 10 människor njuter av mobbning. Jag kommer alltid att vara den 10: e (*9 out of 10 people enjoy bullying. I will always be the 10th.*)	biased	none	1517517
”Författaren säger att det finns ¿ många, många fler anledningar till att jag antar att en uppföljare finns i verken, då?” (*”The author says that there are many, many more reasons I am assuming a sequel is in the works, then?”*)	unbiased	383811	110831
Vad kallar du underkläder för araber? (*What do you call lingerie for Arabs? Socks.*)	biased	none	1618146

Table 4: Examples from the MAB-Swedish (The English in the original is in italics.)

3.2.2 Boolq

It is a question-answering (QA) task where each example has a short passage and a yes/no question about the passage (Clark et al., 2019) . These questions were provided anonymously by Google search users and afterwards paired with a paragraph from a Wikipedia article that has the answer. We evaluated the passage column of the dataset.

3.2.3 CB

This contains short texts in which, at least, one sentence has an embedded clause (De Marneffe et al., 2019). The resulting task is framed as three-class textual entailment on examples that are drawn from the following datasets: Wall Street Journal, fiction from the British National Corpus, and Switchboard. We evaluated the premise column of the dataset.

3.2.4 WSC

This is a coreference resolution dataset (Levesque et al., 2012). Examples consist of a sentence with a pronoun and a list of noun phrases from the sentence. We evaluated the text column of the dataset.

3.2.5 AXg

It is designed to measure gender bias in coreference resolution systems (Rudinger et al., 2018b). Each example consists of a premise sentence having a male or female pronoun and a hypothesis giving a possible antecedent of the pronoun. We evaluated the premise column of the dataset.

3.2.6 RTE

The datasets come from a series of annual competitions on textual entailment (Wang et al., 2019). Data from several sources were merged and converted to two-class classification: entailment and not_entailment. We evaluated the premise column of the dataset.

3.2.7 Swedish CB

This is part of the OverLim dataset by the National Library of Sweden. It contains some of the GLUE and SuperGLUE tasks automatically translated to Swedish, Danish, and Norwegian, using the OpusMT models for MarianMT[5]. We evaluated its training set.

3.2.8 SWEDN

This is a text summarization corpus based on 1,963,576 news articles from the Swedish newspaper Dagens Nyheter (DN) during the years 2000 to 2020.[6] There are five categories of articles in the dataset: domestic news, economy, sports, culture, and others (Monsen and Jönsson, 2021). The training set consists of the first three categories and we evaluate the first 1,000 samples because of the computation cost of evaluation.

3.3 Experiments

The experiments were conducted on two shared Nvidia DGX-1 clusters running Ubuntu 18.04 and 20.04 with 8 × 32GB V100 and 8 x 40GB A100 GPUs, respectively. Average results are reported after running each experiment twice. To evaluate the benchmark datasets, we utilize bias-detection models (Alkhaled et al., 2023) based on RoBERTa (Liu et al., 2019), Electra (Clark et al., 2020), and DeBERTa (He et al., 2021). We train a small mT5 model with batch size of 16, due to memory constraints, on the MAB-Swedish. Wandb (Biewald, 2020), an experiment tracking tool, is run for 5 counts with bayesian optimization to suggest the best hyper-parameter combination for the learning rate (1e-3 - 2e-5) and epochs (6 - 10) before final training of the model. We use the pretrained model from the HuggingFace hub (Wolf et al., 2020). Average training time was 15 hours. Average evaluation time ranges from about 30 minutes to over 24 hours.[7]

4 Results and Discussion

From Table 5 we observe that all the datasets have bias, though little, given that they are smaller than a *bipol* score of 1. The dataset with the least amount of bias is Boolq, which is confirmed by all the three models. This is despite the dataset having the highest number of unique samples. CB has the largest amount of bias and this is also confirmed by the three models. This is also the case for the Swedish CB, when compared with SWEDN.

The average macro F1 score on the validation set of MAB-Swedish is 0.7623 with standard deviation (s.d.) of 0.0075. The resulting error rate is 0.2893. This is relatively reasonable though a bit higher than the error rate for the English RoBERTa, Electra, and DeBERTa, which are 0.198, 0.196, and 0.2, respectively (Alkhaled et al., 2023).

[5] huggingface.co/datasets/KBLab/overlim

[6] spraakbanken.gu.se/resurser/swedn

[7] particularly when cpulimit is used, in fairness to other users

		bipol level ↓ (s.d.)		
RoBERTa	samples	**corpus**	**sentence**	**bipol *(b)***
Boolq	7,929	0.0066	0.8027	0.0053 (0)
CB	250	0.08	0.8483	0.0679 (0)
WSC	279	0.0466	0.8718	0.0406 (0)
AXg	178	0.0112	1	0.0112 (0)
RTE	2,379	0.0294	0.8518	0.0251 (0)
Electra				
Boolq	7,929	0.0073	0.8089	0.0059 (0)
CB	250	0.0316	0.881	0.074 (0)
WSC	279	0.0609	0.9559	0.0582 (0)
AXg	178	0.0112	1	0.0112 (0)
RTE	2,379	0.0269	0.8593	0.0231 (0)
DeBERTa				
Boolq	7,929	0.0103	0.7212	0.0075 (0)
CB	250	0.084	0.9048	0.076 (0)
WSC	279	0.0609	1	0.0609 (0)
AXg	178	0.0112	1	0.0112 (0)
RTE	2,379	0.0366	0.8655	0.0316 (0)
mT5 on Swedish data				
CB	201	0.0796	0.7188	0.0572 (0)
SWEDN	1,000	0.053	0.9433	0.05 (0)

Table 5: Results of average bipol scores. All the datasets have bias, though little.

4.1 Error Analysis

Figure 1 presents the confusion matrix for the mT5 on the MAB-Swedish. The tn, fp, fn, tp are 61,689, 7,960, 12,781, and 19,557, respectively, which are relevant for Eq. 1b. We observe that the model is better at predicting unbiased samples. This is expected since the training data contains more examples of unbiased samples. Table 6 presents some qualitative examples of apparently correct and incorrect predictions in two of the datasets. The first correct example in the English CB appears to have a clear stereotype that *men are naturally right and it is the role of women to follow their lead*. The second correct example, in both the English and Swedish data, may have been perceived as biased by the two different models because of the offensive term *fool* or the overgeneralization that *folk will always take advantage of weakness* or both. Overgeneralization is a characteristic of bias (Rudinger et al., 2018a; Nadeem et al., 2021).

Explaining bias type

The type of overall bias (for the gender axis) in many of the datasets is explained by the dictionary of lists produced by bipol (see Appendix .1) and represented in "top-5 frequent terms" bar graphs of Figures 2 to 13. As expected, we observed that AXg is limited to only gender, unlike Boolq, which also reflects religious bias, as explained in their bipol dictionaries of lists. We observe from Figures 2, 3, and 4 that Boolq is male-biased. Figures 5, 6, and 7 show that CB is also male-biased. This is the case also for RTE, as revealed by Figures 8, 9, and 10. On the other hand, we observe that the case of WSC is not clear-cut because Figure 11 shows only a marginal lead for female bias, Figure 12 shows the difference among the top-5 is zero and Figure 13 shows a slight overall male bias.

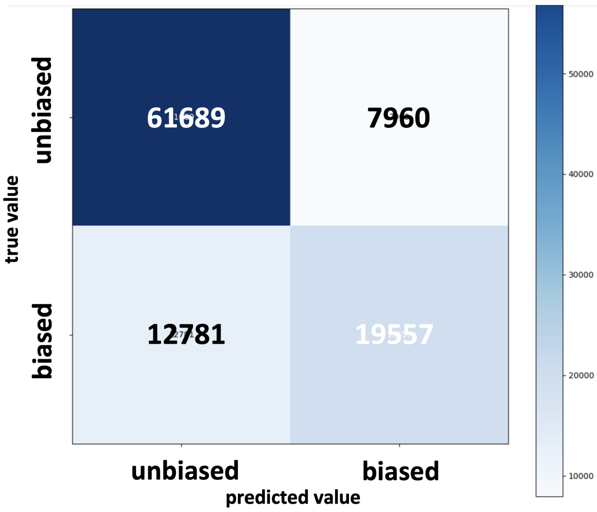

Figure 1: Confusion matrix for mT5.

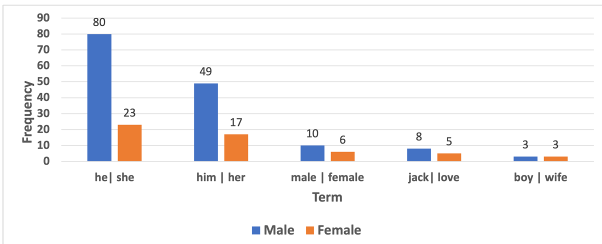

Figure 2: Top-5 gender frequent terms in Boolq by RoBERTa.

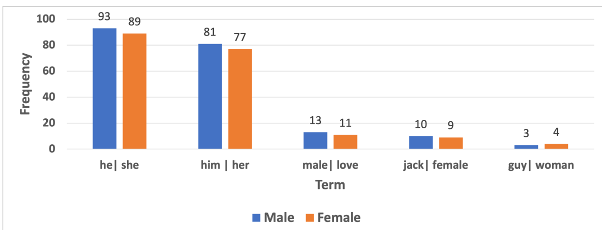

Figure 3: Top-5 gender frequent terms in Boolq by DeBERTa.

examples	label	csv id
Correct predictions from the English CB, based on DeBERTa		
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
she rubbed them away with an angry fist. she was a fool to let anyone get round her. how long before she learned that folk ll always take advantage of weakness	biased	73
Incorrect prediction		
he d gone. philip had to get them back. his dad would kill him if he found that he d taken them.	biased	107
Correct predictions from the Swedish CB, based on mT5		
det är en del av deras religion en religion som jag inte hånar eftersom den innehåller många element som matchar vår egen trots att den saknar vår sanning. vid en av deras stora högtider har de ritualen att driva ut djävlarna från sina kroppar. först kommer trummisarna jag kan säga att inga kvinnor tillåts delta i denna ritual och damerna här kanske håller med mig om att de är lyckligt lottade i detta utelämnande. (*it is part of their religion a religion that I do not mock because it contains many elements that match our own even though it lacks our truth. at one of their great festivals they have the ritual of casting the devils out of their bodies. first come the drummers i can say no women are allowed to participate in this ritual and the ladies here may agree with me that they are lucky in this omission.*)	biased	2
hon gnuggade bort dem med en arg knytnäve. hon var en idiot som lät vem som helst komma runt henne. hur lång tid tar det innan hon får veta att folk alltid kommer att utnyttja svaghet (*she rubbed them away with an angry fist. she was a fool to let anyone get round her. how long before she learned that folk will always take advantage of weakness*)	biased	59
Incorrect prediction		
han var borta. philip var tvungen att få tillbaka dem. hans pappa skulle döda honom om han fann att han hade tagit dem. (*he had gone. philip had to get them back. his dad would kill him if he found that he had taken them.*)	biased	86

Table 6: Qualitative examples of apparently correct and incorrect predictions in some of the datasets. The English translations appear in *italics* for the Swedish examples.

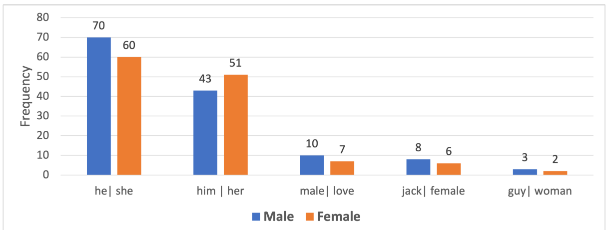

Figure 4: Top-5 gender frequent terms in Boolq by Electra.

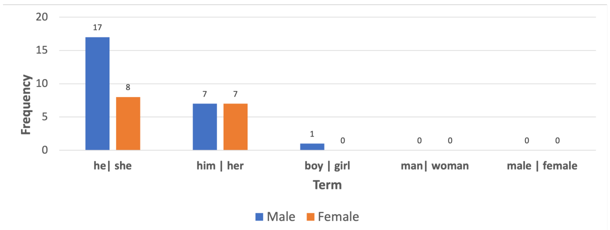

Figure 5: Top-5 gender frequent terms in CB by Roberta.

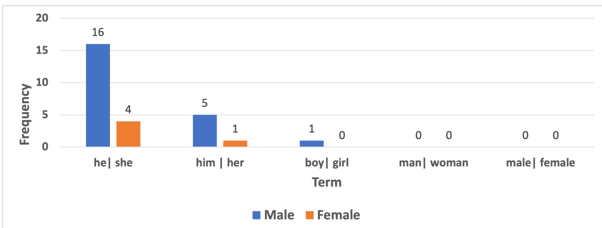

Figure 6: Top-5 gender frequent terms in CB by DeBERTa.

5 Conclusion

We show that all benchmark datasets we evaluated, including the Swedish datasets, contain bias to different degrees. This is likely the first time these datasets are evaluated in such a way that estimates the amount of bias and the type. We believe these evaluations will motivate research on how to more effectively mitigate bias along multiple axes in datasets. This work may encourage dis-

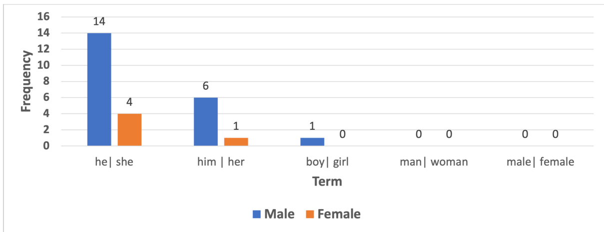

Figure 7: Top-5 gender frequent terms in CB by Electra.

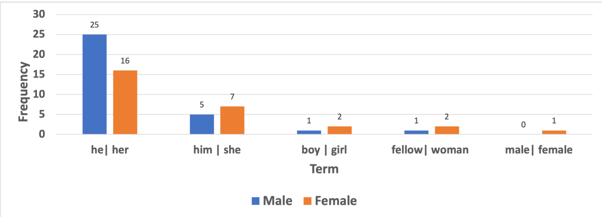

Figure 8: Top-5 gender frequent terms in RTE by RoBERTa.

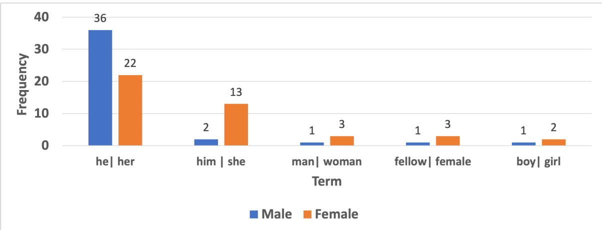

Figure 9: Top-5 gender frequent terms in RTE by DeBERTa.

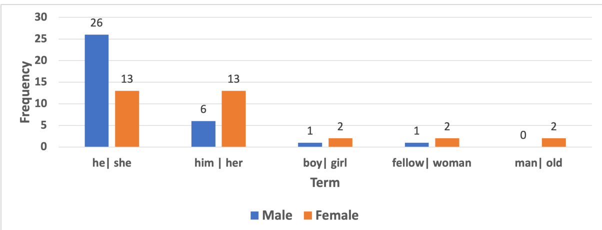

Figure 10: Top-5 gender frequent terms in RTE by Electra.

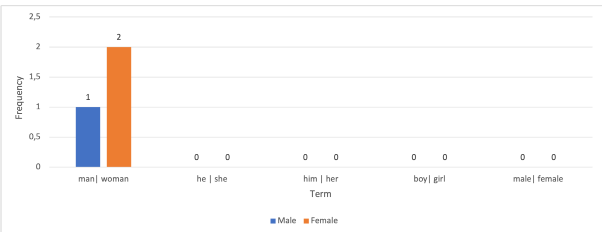

Figure 11: Top-5 gender frequent terms in WSC by RoBERTa.

cussions on whether the biased samples from the benchmark datasets should be disregarded entirely or if they should be utilized in a different manner

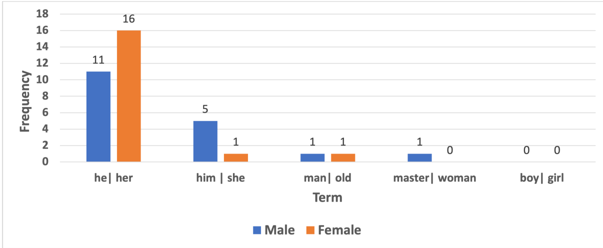

Figure 12: Top-5 gender frequent terms in WSC by DeBERTa.

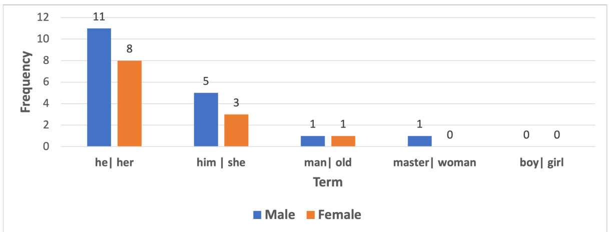

Figure 13: Top-5 gender frequent terms in WSC by Electra.

than previously done. Our public release of the new MAB-Swedish dataset, lexica and model will also facilitate future work in multilingual bias detection.

Ethics Statement

The authors made the effort to obscure offensive terms in examples that were used in this paper. We note that the models for estimating the biases in the datasets are limited in scope, as they only cover certain number of axes (12). Therefore, a result of 0 on any dataset does not necessarily indicate a bias-free dataset.

Acknowledgments

We sincerely thank the anonymous reviewers for their valuable feedback on this work.

References

Tosin Adewumi, Foteini Liwicki, and Marcus Liwicki. 2022. State-of-the-art in open-domain conversational ai: A survey. *Information*, 13(6).

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. 2020a. Corpora compared: The case of the swedish gigaword & wikipedia corpora. *arXiv preprint arXiv:2011.03281*.

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. 2020b. Exploring swedish & english fasttext embeddings for ner with the transformer. *arXiv preprint arXiv:2007.16007*.

Lama Alkhaled, Tosin Adewumi, and Sana Sabah Sabry. 2023. Bipol: A novel multi-axes bias evaluation metric with explainability for nlp. *Manuscript*.

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.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

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

Aravindan Chandrabose, Bharathi Raja Chakravarthi, et al. 2021. An overview of fairness in data–illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45.

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

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. 2020. Detecting gender stereotypes: Lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–11, New York, NY, USA. Association for Computing Machinery.

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.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *ACM FAccT 2021*.

Sanchari Dhar and Lior Shamir. 2021. Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks. *Visual Informatics*, 5(3):92–101.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

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

Keith B Maddox. 2004. Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review*, 8(4):383–401.

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.

Julius Monsen and Arne Jönsson. 2021. A method for building non-english corpora for abstractive text summarization. In *Proceedings of CLARIN Annual Conference*.

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.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. Honest: Measuring hurtful sentence completion in language models. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018a. 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.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018b. 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*, New Orleans, Louisiana. 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*, pages 5477–5490, Online. Association for Computational Linguistics.

Julia Penelope Stanley. 1977. Paradigmatic woman: The prostitute. *Papers in language variation*, pages 303–321.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. *arXiv preprint arXiv:2109.10441*.

Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems*, 32.

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.

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.

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.

Eliezer Yudkowsky et al. 2008. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks*, 1(303):184.

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.

.1 Experiment

Dictionary of lists for RoBERTa on Boolq

{'gender': [' she ': 23, ' her ': 17, ' woman ': 2, ' lady ': 1, ' female ': 6, ' girl ': 1, ' skirt ': 0, ' madam ': 0, ' gentlewoman ': 0, ' madame ': 0, ' dame ': 0, ' gal ': 0, ' maiden ': 0, ' maid ': 0, ' damsel ': 0, ' senora ': 0, ' lass ': 0, ' beauty ': 0, ' ingenue ': 0, ' belle ': 0, ' doll ': 0, ' señora ': 0, ' senorita ': 0, ' lassie ': 0, ' ingénue ': 0, ' miss ': 0, ' mademoiselle ': 0, ' señorita ': 0, ' babe ': 0, ' girlfriend ': 0, ' lover ': 0, ' mistress ': 0, ' ladylove ': 0, ' inamorata ': 0, ' gill ': 0, ' old ': 2, ' beloved ': 0, ' dear ': 0, ' sweetheart ': 0, ' sweet ': 0, ' flame ': 2, ' love ': 5, ' valentine ': 0, ' favorite ': 1, ' moll ': 0, ' darling ': 0, ' honey ': 0, ' significant ': 0, ' wife ': 3, ' wifey ': 0, ' missus ':

0, ' helpmate ': 0, ' helpmeet ': 0, ' spouse ': 0, ' bride ': 1, ' partner ': 0, ' missis ': 0, ' widow ': 0, ' housewife ': 0, ' mrs ': 0, ' matron ': 0, ' soul ': 3, ' mate ': 1, ' housekeeper ': 0, ' dowager ': 0, ' companion ': 0, ' homemaker ': 0, ' consort ': 0, ' better half ': 0, ' hausfrau ': 0, ' stay-at-home ': 0, ' he ': 80, ' him ': 49, ' boy ': 3, ' man ': 1, ' male ': 10, ' guy ': 1, ' masculine ': 0, ' virile ': 0, ' manly ': 0, ' man-sized ': 0, ' hypermasculine ': 0, ' macho ': 0, ' mannish ': 0, ' manlike ': 0, ' man-size ': 0, ' hairy-chested ': 0, ' butch ': 0, ' ultramasculine ': 0, ' boyish ': 0, ' tomboyish ': 0, ' hoydenish ': 0, ' amazonian ': 0, ' gentleman ': 0, ' dude ': 0, ' fellow ': 0, ' cat ': 2, ' gent ': 0, ' fella ': 0, ' lad ': 0, ' bloke ': 0, ' bastard ': 0, ' joe ': 0, ' chap ': 0, ' chappie ': 0, ' hombre ': 0, ' galoot ': 0, ' buck ': 0, ' joker ': 3, ' mister ': 0, ' jack ': 8, ' sir ': 0, ' master ': 1, ' buddy ': 0, ' buster ': 0], 'racial':... }