Multilingual Bias Detection and Mitigation for Indian Languages

Ankita Maity*, **Anubhav Sharma***, **Rudra Dhar***, **Tushar Abhishek**† *
Manish Gupta† *, **Vasudeva Varma***
*IIIT Hyderabad, India
†Microsoft, India

Abstract

Lack of diverse perspectives causes neutrality bias in Wikipedia content leading to millions of worldwide readers getting exposed by potentially inaccurate information. Hence, neutrality bias detection and mitigation is a critical problem. Although previous studies have proposed effective solutions for English, no work exists for Indian languages. First, we contribute two large datasets, mWikiBias and mWNC, covering 8 languages, for the bias detection and mitigation tasks respectively. Next, we investigate the effectiveness of popular multilingual Transformer-based models for the two tasks by modeling detection as a binary classification problem and mitigation as a style transfer problem. We make the code and data publicly available.

Keywords: Neutral Point of View, Bias Detection, Bias Mitigation, Indian language NLG, Transformer Models

1. Introduction

Wikipedia has three core content policies: Neutral Point of View (NPOV), No Original Research, and Verifiability[1]. NPOV means that content should represent fairly, proportionately, and, as far as possible, without editorial bias, all the significant views that have been published by reliable sources on a topic[2]. This means (1) Opinions should not be stated as facts and vice versa. (2) Seriously contested assertions should not be stated as facts. (3) Nonjudgmental language should be preferred. (4) Relative prominence of opposing views should be indicated. This is definition of bias we follow in this paper.

Considering Wikipedia's (1) volume and diversity of content, (2) frequent updates, and (3) large and diverse userbase, automatic bias detection and suggestion of neutral alternatives is important. Bias can lead to inaccurate information or dilution of information. Particularly, lower article quality and fewer editors of Indian language Wikipedia pages makes such a system indispensable. Hence, in this work, we study how to detect sentences that violate the NPOV guidelines and convert them to more neutral sentences for Indian languages, as shown in Fig. 1.

While there exists bias detection and mitigation studies (Zhong et al., 2021; Pryzant et al., 2020; Lai et al., 2022; Liu et al., 2021) for English, there is hardly any such work for other languages. Aleksandrova et al. (Aleksandrova et al., 2019) work on bias detection for Bulgarian and French, but their method requires a collection of language-specific NPOV tags; and is therefore difficult to extend to Indian languages. Lastly, there exists no dataset for multilingual bias mitigation. We fill the gap in this paper by proposing two new multilingual bias detection and mitigation datasets, mWikiBias and mWNC, each covering 8 languages: English (en) and seven Indian languages - Hindi (hi), Marathi (mr), Bengali (bn), Gujarati (gu), Tamil (ta), Telugu (te) and Kannada (kn).

Bias detection is challenging because certain words lead to bias if they are written in some contexts, while not in other contexts. For bias detection, we perform binary classification using MuRIL (Khanuja et al., 2021), InfoXLM (Chi et al., 2021) and mDeBERTa (He et al., 2022) in zero-shot, monolingual and multilingual settings. Bias mitigation is challenging because of subjectivity and context-dependence, and the models need to strike a good balance between fairness and content preservation. For bias mitigation, we perform style transfer using IndicBART (Dabre et al., 2022), mT0 (Muennighoff et al., 2023) and mT5 (Xue et al., 2021). These models provide strong baseline results for the novel multilingual tasks.

Overall, we make the following contributions in this paper.

- We propose multilingual bias detection and mitigation for Indian languages.
- We contribute two novel datasets, mWikiBias and mWNC, to multilingual natural language generation (NLG) community. Across 8 languages, they contain ∼568K and ∼78K samples for bias detection and mitigation resp.
- Extensive experiments show that mDeBERTa outperforms MuRIL and InfoXLM for the bias detection task. On the other hand, mT5 and

[1] https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies
[2] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

WILDRE-7 Workshop, pages 24–29

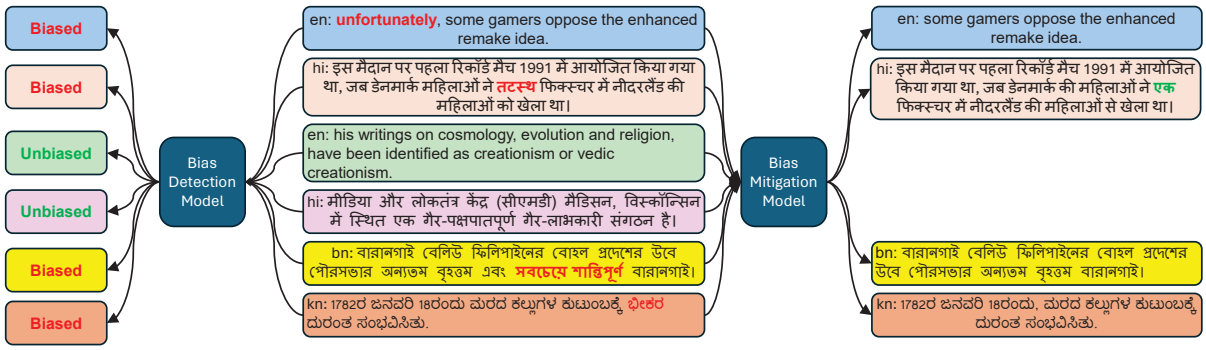

Figure 1: Bias Detection and Mitigation Examples from MWIKIBIAS Dataset

mT0 perform the best for bias mitigation on MWIKIBIAS and MWNC respectively.

2. Related Work

Several kinds of societal biases have been studied in the literature as part of responsible AI model building (Sheng et al., 2021; Badjatiya et al., 2019): promotional tone (De Kock and Vlachos, 2022), puffery (Bertsch and Bethard, 2021), political bias (Fan et al., 2019), and gender and racial bias (Field et al., 2022; Parikh et al., 2021). In this work, we focus on a more general form of bias called as neutrality bias. Earlier work on neutrality bias detection leveraged basic linguistic features (Recasens et al., 2013; Hube and Fetahu, 2018) while recent work uses Transformer based models (Pryzant et al., 2020; Zhong et al., 2021). Unfortunately, these studies (Recasens et al., 2013; Pryzant et al., 2020; Zhong et al., 2021; Hube and Fetahu, 2018) focus on English only. Aleksandrova et al. (Aleksandrova et al., 2019) work on bias detection for Bulgarian and French, but their method requires a collection of language-specific NPOV tags, making it difficult to extend to Indian languages.

Bias mitigation is under-studied even for English (Liu et al., 2021). We contribute datasets and strong initial baseline methods towards multilingual bias mitigation.

3. MWIKIBIAS and MWNC Datasets

Popular bias detection and mitigation corpora in English like Wiki Neutrality Corpus (WNC) (Pryzant et al., 2020) and WIKIBIAS (Zhong et al., 2021) were created by looking for NPOV-related tags in the edit history of the English Wikipedia dumps. Both datasets have parallel sentence structures (biased sentence linked with an unbiased version). Replication of their data curation pipeline for Indian languages did not work due to a lack of frequency and consistency in tag usage for edits in the revision history of corresponding Wikipedia pages.

Hence, we translated these datasets using IndicTrans (Ramesh et al., 2022) to create MWNC and MWIKIBIAS datasets for eight languages. To create cleaner datasets, we used the following heuristics to filter samples. (1) A biased and its corresponding unbiased sentence in English typically differ by very few words. Hence, we removed samples where translation of biased sentence and unbiased sentence were exactly the same for at least one of our target languages. (2) To reduce impact of translation errors, we removed samples where sentences contained regex matches for URLs, phone numbers, and email IDs.

For every parallel translated pair of (biased, unbiased) sentence in each language l, we create one sample for bias mitigation dataset, and two samples (biased and unbiased) for bias detection dataset. Overall, the total number of samples for classification are 287.6K and 280.0K for MWIKIBIAS and MWNC respectively. To reduce training compute, we took a random sample from the overall bias mitigation data, leading to 39.4K and 39.0K paired samples in the MWIKIBIAS and MWNC respectively[3]. The number of samples for each language in both the datasets is consistent. Further, both of our bias detection datasets contain an equal number of biased and unbiased samples. We divide the datasets into a train/val/test split of 90/5/5.

4. Multilingual Bias Detection and Mitigation

We train multilingual bias detection and mitigation models using train part of MWIKIBIAS and MWNC respectively. As shown in Fig. 1, these models detect whether the sentence is biased and con-

[3]Experiments with full bias mitigation dataset showed similar results.

		MuRIL				InfoXLM				mDeBERTa			
		Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
mWiki Bias	ZeroShot	59.26	61.53	59.26	57.19	59.28	59.74	59.28	58.81	60.99	61.63	60.99	60.45
	MonoLingual	62.66	65.15	62.66	60.97	60.97	62.06	60.97	60.01	64.82	65.63	64.82	64.31
	MultiLingual	65.11	**66.33**	65.11	64.41	63.42	64.55	63.42	62.64	**65.14**	65.64	**65.14**	**64.83**
mWNC	ZeroShot	63.04	64.00	63.04	62.38	62.08	62.81	62.08	61.53	63.02	64.02	63.02	62.34
	MonoLingual	64.82	65.95	64.82	64.17	63.15	63.71	63.15	62.75	66.59	66.92	66.59	66.42
	MultiLingual	66.72	**67.24**	66.72	66.46	65.49	65.75	65.49	65.34	**66.96**	67.03	**66.96**	**66.92**

Table 1: Bias Detection Results.

		MuRIL				InfoXLM				mDeBERTa			
		Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
mWikiBias	bn	64.55	**65.65**	64.55	63.92	62.17	63.38	62.17	61.30	**64.62**	65.15	**64.62**	**64.31**
	en	73.69	**74.74**	73.69	73.40	72.89	73.74	72.89	72.65	**74.19**	74.57	**74.19**	**74.09**
	gu	63.77	**64.93**	63.77	63.06	62.03	63.25	62.03	61.13	**63.91**	64.35	**63.91**	**63.63**
	hi	**65.31**	**66.37**	**65.31**	64.73	63.54	64.60	63.54	62.86	65.01	65.34	65.01	**64.82**
	kn	**64.33**	**65.63**	**64.33**	63.57	62.48	63.50	62.48	61.76	63.96	64.49	63.96	**63.64**
	mr	**62.74**	**63.98**	**62.74**	61.89	61.47	62.52	61.47	60.65	62.26	62.71	62.26	**61.92**
	ta	63.05	64.47	63.05	62.12	61.99	63.23	61.99	61.07	**63.80**	**64.55**	**63.80**	**63.33**
	te	**63.46**	**64.90**	**63.46**	62.56	60.81	62.17	60.81	59.69	63.34	63.99	63.34	**62.91**
	avg	65.11	**66.33**	65.11	64.41	63.42	64.55	63.42	62.64	**65.14**	65.64	**65.14**	**64.83**
mWNC	bn	**66.75**	**67.34**	**66.75**	**66.47**	65.01	65.32	65.01	64.83	66.46	66.53	66.46	66.42
	en	71.08	71.43	71.08	70.96	71.57	71.66	71.57	71.54	**72.92**	**72.92**	**72.92**	**72.91**
	gu	66.00	**66.48**	66.00	65.76	64.33	64.63	64.33	64.15	**66.42**	66.46	**66.42**	**66.40**
	hi	67.13	**67.53**	67.13	66.95	66.28	66.44	66.28	66.20	**67.45**	67.47	**67.45**	**67.44**
	kn	66.40	**66.92**	66.40	66.14	64.77	65.04	64.77	64.61	**66.55**	66.61	**66.55**	**66.52**
	mr	**64.90**	**65.48**	**64.90**	**64.57**	63.70	63.94	63.70	63.54	64.44	64.56	64.44	64.37
	ta	**65.70**	**66.29**	**65.70**	65.38	64.36	64.69	64.36	64.15	65.68	65.83	65.68	**65.60**
	te	**65.78**	**66.43**	**65.78**	65.44	63.93	64.30	63.93	63.69	65.75	65.87	65.75	**65.68**
	avg	66.72	**67.24**	66.72	66.46	65.49	65.75	65.49	65.34	**66.96**	67.03	**66.96**	**66.92**

Table 2: Detailed Language-wise Bias Detection Results for Multilingual Setup.

vert it to a more neutral sentence if bias is detected. So, for example, the Hindi sentence मीडिया और लोकतंत्र केंद्र (सीएमडी) मैडिसन, विस्कॉन्सिन में स्थित एक गैर-पक्षपातपूर्ण गैर-लाभकारी संगठन है (the Center for Media and Democracy (CMD) is a non-partisan non-profit organization based in Madison, Wisconsin) is detected as an unbiased sentence, while the Bengali sentence বারানগাই বেলিউ ফিলিপাইনের বোহল প্রদেশের উবে পৌরসভার অন্যতম বৃহত্তম এবং সবচেয়ে শান্তিপূর্ণ বারানগাই (Barangay Benliw is one of the largest and the most peaceful Barangay in the municipality of Ubay, in the province of Bohol, Philippines) is detected as biased and thus converted to a more neutral বারানগাই বেলিউ ফিলিপাইনের বোহল প্রদেশের উবে পৌরসভার অন্যতম বৃহত্তম বারানগাই (Barangay Benliw is one of the largest Barangays in the municipality of Ubay, in the province of Bohol, Philippines). Similarly, the Kannada sentence 1782ರ ಜನವರಿ 18ರಂದು ಮರದ ಕಲ್ಲುಗಳ ಕುಟುಂಬಕ್ಕೆ ಭೀಕರ ದುರಂತ ಸಂಭವಿಸಿತು (on 18 January 1782, a horrendous tragedy struck the Woodmason family) is converted to a more neutral 1782ರ ಜನವರಿ 18ರಂದು, ಮರದ ಕಲ್ಲುಗಳ ಕುಟುಂಬಕ್ಕೆ ದುರಂತ ಸಂಭವಿಸಿತು (on 18 January 1782, tragedy struck the Woodmason family).

Multilingual Bias Detection Method: For bias detection, we finetune three Transformer encoder-only multilingual models: InfoXLM (Chi et al., 2021), MuRIL (Khanuja et al., 2021), and mDeBERTa (He et al., 2022), with a twin linear layer setup to detect whether a sentence is biased. We experiment with three different training setups: (1) zero-shot (training only on English and testing on the other languages), (2) monolingual (one language at a time) and (3) multilingual (trained on all languages together). For fair comparisons, we use 12 layer models with a dimensionality of 768.

Multilingual Bias Mitigation Method: For bias mitigation, we finetune three multilingual encoder-decoder transformer-based models: mT5 (Xue et al., 2021), IndicBART (Dabre et al., 2022), and mT0 (Muennighoff et al., 2023) over the parallel corpora to perform debiasing. For fair comparisons, we use the small version of all three models for our experiments.

Metrics: We evaluate bias detection models using four popular binary classification metrics: accuracy (Acc), macro-precision (P), macro-recall (R) and macro-F1.

Effectiveness of bias mitigation models should be evaluated broadly on two aspects: match with groundtruth and debiasing accuracy. For measuring match with groundtruth unbiased sentences, we use standard NLG metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), chrF (Popović, 2015) and BERTScore (Zhang et al., 2019). We measure debiasing accuracy using "Normalized Accuracy (NAcc)" defined as follows. Let N be the percent of ground truth sentences in the test set that are classified as "unbiased" by our best bias detection model. First, given a bias mitigation model, we compute the percent of its generated outputs that are classified as "unbiased" by our best bias detection model. Sec-

		MonoLingual						MultiLingual					
		B	M	C	BS	NAcc	HM	B	M	C	BS	NAcc	HM
mWiki Bias	IndicBART	**63.67**	75.87	80.04	91.58	71.57	80.35	46.32	64.62	68.94	88.47	73.84	80.50
	mT0	61.57	77.05	80.84	93.24	76.77	84.21	60.86	77.04	80.89	93.20	77.73	84.76
	mT5	58.81	76.74	80.23	92.97	73.23	81.93	63.26	**77.41**	**81.39**	**93.40**	**79.38**	**85.82**
mWNC	IndicBART	54.98	69.25	75.27	90.99	65.52	76.18	17.58	59.67	61.15	85.54	**71.12**	77.67
	mT0	53.09	70.01	75.75	91.27	69.15	78.68	55.23	**70.61**	**76.54**	**91.50**	70.59	**79.70**
	mT5	**55.39**	70.28	76.22	91.36	66.57	77.02	55.27	70.46	76.41	91.47	69.83	79.20

Table 3: Bias Mitigation Results. B=BLEU, M=METEOR, C=chrF, BS=BERTScore, NAcc=NormAcc, HM=Harmonic Mean of BS and NAcc.

	mT5 mWikiBias						mT0 mWNC					
	B	M	C	BS	NAcc	HM	B	M	C	BS	NAcc	HM
bn	60.60	75.82	80.03	92.81	76.50	83.87	54.76	68.48	75.10	90.72	70.29	79.21
en	86.02	92.68	91.63	98.30	88.06	92.90	79.06	87.56	87.38	97.42	79.97	87.83
gu	61.35	76.39	79.54	92.85	78.02	84.79	55.47	69.56	74.73	90.79	67.34	77.32
hi	69.36	82.87	82.76	94.16	75.78	83.97	63.12	76.81	77.85	92.19	69.49	79.25
kn	60.84	75.63	82.05	93.28	78.31	85.14	54.69	68.88	77.55	91.40	67.26	77.49
mr	58.19	73.25	78.11	92.12	79.43	85.31	50.24	65.44	72.65	89.77	68.78	77.89
ta	53.03	69.70	78.15	91.57	80.26	85.54	35.66	61.91	72.51	89.37	71.66	79.54
te	56.72	72.97	78.86	92.11	78.67	84.86	48.84	66.21	74.55	90.31	70.00	78.87
avg	63.26	77.41	81.39	93.40	79.38	85.82	55.23	70.61	76.54	91.50	70.59	79.70

Table 4: Detailed Language-wise Bias Mitigation Resultsfor the best models per dataset. B=BLEU, M=METEOR, C=chrF, BS=BERTScore, NAcc=NormAcc, HM=Harmonic Mean of BS and NAcc.

ond, we normalize this quantity by N and call the ratio as Normalized Accuracy (NAcc).

A model can easily obtain high match with groundtruth by simply copying words from the input. Similarly, a model can easily obtain high NAcc score by predicting a constant highly unbiased sentence independent of the input. A good model should be able to strike a favourable tradeoff between the two aspects. Among the four metrics for computing the match, BERT-Score has been shown to be the most reliable in NLG literature, because it captures semantic match rather than just a syntactic match. Hence, we compute the harmonic mean of BERT-Score and NAcc Score and report it as HM.

Implementation Details: For MuRIL and InfoXLM, we use a learning rate of 1e-6, weight decay of 0.001, and dropout of 0.1. We trained for 15 epochs using a batch size of 320 and mixed precision training. For mDeBERTa, we use a learning rate 2e-5 with a weight decay of 0.01, keeping the other parameters the same. We use a batch size of 12 for the bias mitigation experiments and train for 10 epochs, using early stopping with a patience of 3. We use Adafactor optimizer with a learning rate of 1e-3 for mT5 and mT0 and AdamW optimiser with a learning rate of 1e-4 for IndicBART. All models use a weight decay of 0.01. All models were trained on a machine with 4 NVIDIA V100 GPUs having 32GB of RAM.

5. Results

Bias Detection Results: We show a summary of bias detection results, averaged across the 8 languages, in Table 1 and language-wise details in Table 2. Table 1 shows that (1) Multilingual models outperform monolingual models, which in turn outperform zero-shot approaches. (2) Across both the datasets, mDeBERTa and MuRIL, both trained in a multilingual setting, exhibit the strongest performance, with mDeBERTa slightly outperforming MuRIL.

From the language-wise results in Table 2, we observe the following: (1) As expected, for both datasets, across all models and metrics, best results are for English. We also observe that the models perform the worst for Marathi and Telugu. (2) In general, MuRIL is better in terms of precision, but mDeBERTa is better in terms of recall and also F1.

Bias Mitigation Results: We show a summary of bias mitigation results in Table 3 and language-wise bias mitigation results in Table 4 using N=73.52 and 76.17 for mWikiBias and mWNC respectively. From the results, we observe that (1) Broadly, multilingual models outperform monolingual counterparts. (2) mT5 is better for mWikiBias providing a high HM of 85.82, while mT0 is better for mWNC providing a high HM of 79.70. (3) As expected, both the models work best for English.

Human Evaluation Results:

We asked 4 Computer Science bachelors students with language expertise to evaluate the generated outputs (mT5 multilingual for mWIKIBIAS and mT0 multilingual for mWNC) on 3 criteria, each on a scale of 1 to 5: fluency, whether the bias is reduced and whether the meaning is preserved when compared to input. This was done for 50 samples per language for both datasets. Table 5 shows that automated evaluation correlates well with human judgment, with English predictions

Lang.	mWikiBias			mWNC		
	Fluency (↑)	Bias (↓)	Meaning (↑)	Fluency (↑)	Bias (↓)	Meaning (↑)
bn	4.42	3.12	4.79	3.94	2.68	4.80
en	**4.92**	2.72	**4.84**	**4.86**	**2.40**	**4.92**
hi	4.20	3.20	4.76	4.60	2.64	**4.92**
te	4.40	**2.50**	4.81	3.88	2.45	4.75

Table 5: Human Evaluation Results

showing the best results. mWNC is easier for the models to debias than mWIKIBIAS. The model outputs were generally fluent and had similar content as the input text. However, a wider variance in bias mitigation abilities was observed for the 3 Indian languages tested compared to English. Ambiguity in bias assessment and noise in the reference text made ∼20% of the samples challenging for human annotators.

6. Conclusion

In this paper, we proposed the critical problems of multilingual bias detection and mitigation for Indian languages. We also contributed two data sets, mWikiBias and mWNC. We presented baseline results using standard Transformer based models. We make our code and data set publicly available[4]. In the future, we would like to experiment with reinforcement learning based methods which could use detection based scores to enhance generation.

7. Bibliographical References

References

Desislava Aleksandrova, François Lareau, and Pierre André Ménard. 2019. Multilingual sentence-level bias detection in wikipedia. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 42–51.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The world wide web conference*, pages 49–59.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Amanda Bertsch and Steven Bethard. 2021. Detection of puffery on the english wikipedia. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 329–333.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863.

Christine De Kock and Andreas Vlachos. 2022. Leveraging wikipedia article evolution for promotional tone detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5601–5613.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349.

Anjalie Field, Chan Young Park, Kevin Z Lin, and Yulia Tsvetkov. 2022. Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Compan-*

[4] https://github.com/Ankita-Maity/Bias-Detection-Mitigation/

ion proceedings of the the web conference 2018, pages 1779–1786.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2022. Multilingual pre-training with language and task adaptation for multilingual text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 262–271.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing sexism and misogyny through neural approaches. *ACM Transactions on the Web (TWEB)*, 15(4):1–31.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, AK Raghavan, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, J Mahalakshmi, Divyanshu Kakwani, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. 2021. Wikibias: Detecting multi-span subjective biases in language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1799–1814.