Multilingual Bias Detection and Mitigation for Low Resource Languages

Anubhav Sharma*, Ankita Maity*, Tushar Abhishek, Rudra Dhar

Radhika Mamidi, Manish Gupta, Vasudeva Varma
IIIT Hyderabad, India

Abstract

Subjective bias in Wikipedia textual data is a significant problem and affects millions of readers worldwide. Though some monolingual work has been done in classifying and debiasing biased text in resource-rich languages, the low-resource languages with large numbers of speakers remain unattended. We present an approach for the dual problems of multilingual bias detection and its mitigation with a thorough analysis. In this work, we establish competitive baselines on our preliminary approach, which includes classification-based modelling for bias detection on a multilingual dataset curated from existing monolingual sources. For the problem of bias mitigation, we follow the style transfer paradigm and model using transformer-based seq2seq architectures. We also discuss several approaches for further improvement in both problems as a part of our ongoing work.

Keywords: Multilingual, Bias, NPOV, Classification, Style Transfer

Introduction

Wikipedia is one of the largest sources of objective information online. Its content is used for knowledge enrichment and as a data source for multiple related studies and research.

Hence Wikipedia has three core content policies, one of which is Neutral Point of View (NPOV). This policy is a set of principles, including "avoiding stating opinions as facts" and "preferring nonjudgmental language."

In this work, we study how to detect sentences that violate the NPOV guidelines and convert them to more neutral sentences in Indian languages which have low digital resources. We focus on eight Indian languages - Hindi (hi), English (en), Marathi (mr), Bengali (bn), Gujarati (gu), Tamil (ta), Telugu (te) and Kannada (kn).

Once completed, this work will enhance the quality of the Indian language Wikipedia articles and increase its credibility as the largest source of **free and fair** information.

*Both the authors have equal contribution to this work.

Data

Wiki Neutrality Corpus (WNC) (Pryzant et al., 2020) and WIKIBIAS (Zhong et al., 2021) corpus were created by looking for NPOV-related tags in the edit history of the English Wikipedia dumps. Both datasets have parallel sentence structures of more than 200k data points. We tried to replicate the data curation pipeline of these datasets. However, they did not work well with low-resource Indian languages due to lack of consistency in tag usage for edits in the revision history of the Indian language version of Wikipedia. Hence, we opted for translating the datasets using IndicTrans. After filtering out sentences with the same translation for biased and unbiased sentences and based on other heuristics, we make our own mWNC and mWIKIKIBIAS datasets with data for eight languages. More details are available in table 3.

Methodology

For our modelling approaches, we heavily rely on multilingual transformer architecture-based models, which are trained on the mWNC and mWIKIBIAS datasets as mentioned in the Data section.

For classification, we learn classifiers based on encoder-only models like InfoXLM (Chi et al., 2020), and MuRIL (Khanuja et al., 2021) to detect whether a sentence is biased. We also experimented with contrastive learning: in this approach, we aim to increase the gap between probabilities of positive to negative samples using the InfoNCE loss. For style transfer, we finetune multilingual encoder-decoder transformer-based models like mT5 (Xue et al., 2021), IndicBART (Dabre et al., 2022), and mT0 (Muennighoff et al., 2022) over the parallel corpus to perform debiasing 1.

Classifier accuracies and percentage of unchanged sentences were considered for better evaluation instead of simply content preservation metrics like BLEU scores 2 This is because comparably high BLEU scores were observed for source copy (since the source and the target sentences, in this case, are very similar). Some bias-causing words were changed from unbiased to biased form during decoding. Average attention scores for all such words were calculated as a possible metric for the quality of debiasing.

All the above experiments were done in a multilingual (all languages trained together) vs monolingual setup (one language at a time).

Results

The results are summarized in the results tables 1, 2 and 4. All the values given are averaged from the eight languages.

For classification, it was found that InfoXLM worked better for the English language, but MuRIL outperformed it in terms of performance in Indian languages. Also, the monolingual models tend to identify the sentences without bias better than the multilingual versions. However, they are worse at identifying sentences with bias. Overall, multilingual models outperform the monolingual models.

For style transfer, the models leave a large percentage of sentences unchanged for the debiasing task. In many cases, the input and output sentences are very similar, so it may be hard for the models to figure out the correct part of the sentence to change. This contributes to the difficulty of the task. The high value of the content preservation metrics like BLEU/METEOR/chrF/BERTScore only tells part of the story because the highest values are obtained for models which do not change the biased sentences in any way. Regardless, models generally report higher BLEU scores when we go from monolingual to multilingual settings.

The highest classifier accuracies were obtained for mT5 when trained with all eight languages. So even though the content preservation metrics were not the highest for this experiment, we can consider this the best experimental setting for now, going by the percentage of unchanged sentences and the classifier accuracies.

Future work

This ongoing work has provided some valuable insights into debiasing multilingual Wikipedia text, but further experimentation to harness better approaches is required as we progress. Following are some directions:

- Contextual information for sentence-level bias can be passed through category/title/citation information from Wikipedia and can be appended as prefixes.
- We pass sentences from articles on Wikipedia to our classifier. If no sentences from the article are classified as bias positive, then we say the article is unbiased. Otherwise, we note which section has the most bias positive sentences and place the section-specific NPOV tag there (narrowing down from the article-level tag to make it easier for editors to remove the biased content manually).
- We can use the bias classifier score as a reward function for reinforcement learning based training of the generative module along with next-word prediction loss.
- Additional metrics like fluency, bias and meaning can be considered for qualitative evaluation.

References

[Chi et al.2020] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

[Dabre et al.2022] Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh 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, Dublin, Ireland, May. Association for Computational Linguistics.

[Khanuja et al.2021] 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*.

[Muennighoff et al.2022] 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. 2022. Crosslingual generalization through multitask finetuning.

[Pryzant et al.2020] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489, Apr.

[Xue et al.2021] 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, June. Association for Computational Linguistics.

[Zhong et al.2021] 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, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

	% of unchanged	BLEU	METEOR	chrF	BERTScore	Classifier Acc
IndicBART (monolingual)	77.75	65.88	78.22	81.78	93.51	0.59
mT0 (monolingual)	**32.26**	60.08	72.96	79.49	92.57	0.55
mT5 (monolingual)	42.98	61.7	73.93	80.42	92.62	0.69
IndicBART (multilingual)	69.47	**65.91**	**78.26**	**81.81**	**93.59**	0.64
mT0 (multilingual)	60.99	62.13	75.84	81.63	93.35	0.64
mT5 (multilingual)	56.35	65	75.56	81.59	93.29	**0.71**

Table 1: Debiasing results on the WIKIBIAS dataset.

	% of unchanged	BLEU	METEOR	chrF	BERTScore	Classifier Acc
IndicBART (monolingual)	56.85	53.79	70	75.68	91.05	0.61
mT0 (monolingual)	63.96	51.89	69.65	74.83	91.03	0.59
mT5 (monolingual)	**29.24**	50.77	67.5	73.82	89.65	0.59
IndicBART (multilingual)	44.99	53.7	70.34	76.14	91.37	0.64
mT0 (multilingual)	67.49	**55.92**	**71.3**	**77.17**	**91.51**	0.64
mT5 (multilingual)	49.51	55.64	69.82	75.91	91.3	**0.69**

Table 2: Debiasing results on the WNC dataset.

Dataset	Split	Biased/Unbiased sentences
WNC	train	140*k* / 140*k*
	val	11.6*k* / 11.6*k*
	test	11.684*k* / 11.684*k*
WIKIBIAS	train	126*k* / 126*k*
	val	7*k* / 7*k*
	test	7*k* / 7*k*

Table 3: Per-language stats in our multilingual mWNC and mWIKIBIAS datasets. The number of samples in each language for both the datasets is consistent.

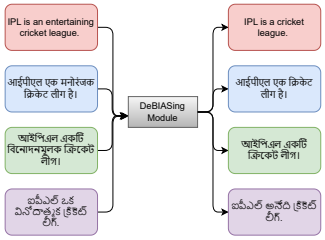

Figure 1: The debiasing module.

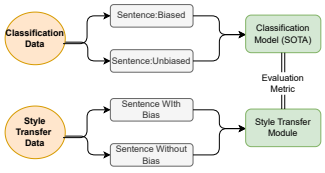

Figure 2: Using the classification module to evaluate the generated outputs of the style transfer module.

Dataset	Model	Testing Protocol	macro-F1	MCC
mWIKIBIAS	MuRIL	E.D. → E.D.	0.6672	0.3384
		E.D. → en	0.7105	0.4218
		E.D. → bn	0.6627	0.3308
		E.D. → gu	0.6533	0.3115
		E.D. → hi	0.6671	0.3367
		E.D. → kn	0.6716	0.3461
		E.D. → mr	0.6547	0.3161
		E.D. → ta	0.6601	0.3258
		E.D. → te	0.6546	0.3179
mWIKIBIAS	InfoXLM	E.D. → E.D.	0.6553	0.3115
		E.D. → en	0.7058	0.4119
		E.D. → bn	0.6301	0.2605
		E.D. → gu	0.6449	0.2928
		E.D. → hi	0.6624	0.3248
		E.D. → kn	0.6568	0.3145
		E.D. → mr	0.6533	0.3073
		E.D. → ta	0.6456	0.2945
		E.D. → te	0.6413	0.2862
mWNC	MuRIL	E.D. → E.D.	0.6676	0.3455
		E.D. → en	0.716	0.435
		E.D. → bn	0.6673	0.3457
		E.D. → gu	0.662	0.3346
		E.D. → hi	0.6728	0.3539
		E.D. → kn	0.6637	0.3387
		E.D. → mr	0.6489	0.3128
		E.D. → ta	0.6553	0.324
		E.D. → te	0.6535	0.3195
mWNC	InfoXLM	E.D. → E.D.	0.6593	0.3286
		E.D. → en	0.725	0.4538
		E.D. → bn	0.6536	0.3197
		E.D. → gu	0.6469	0.3045
		E.D. → hi	0.6671	0.3411
		E.D. → kn	0.6529	0.3156
		E.D. → mr	0.6399	0.2913
		E.D. → ta	0.6451	0.3028
		E.D. → te	0.6437	0.3002

Table 4: Classification results over both monolingual and multilingual experiments.
E.D. = Entire Data, E.D. → x specifies training on entire data, testing on x.
.