Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology

RISHAV HADA, Microsoft Research, India
SAFIYA HUSAIN, Karya, India
VARUN GUMMA, Microsoft Research, India
HARSHITA DIDDEE*, Carnegie Mellon University, USA
ADITYA YADAVALLI, Karya, India
AGRIMA SETH†, University of Michigan, USA
NIDHI KULKARNI, Karya, India
UJWAL GADIRAJU, Delft University of Technology, Netherlands
ADITYA VASHISTHA, Cornell University, USA
VIVEK SESHADRI, Microsoft Research, Karya, India
KALIKA BALI, Microsoft Research, India

Existing research in measuring and mitigating gender bias predominantly centers on English, overlooking the intricate challenges posed by non-English languages and the Global South. This paper presents the first comprehensive study delving into the nuanced landscape of gender bias in Hindi, the third most spoken language globally. Our study employs diverse mining techniques, computational models, field studies and sheds light on the limitations of current methodologies. Given the challenges faced with mining gender biased statements in Hindi using existing methods, we conducted field studies to bootstrap the collection of such sentences. Through field studies involving rural and low-income community women, we uncover diverse perceptions of gender bias, underscoring the necessity for context-specific approaches. This paper advocates for a community-centric research design, amplifying voices often marginalized in previous studies. Our findings not only contribute to the understanding of gender bias in Hindi but also establish a foundation for further exploration of Indic languages. By exploring the intricacies of this understudied context, we call for thoughtful engagement with gender bias, promoting inclusivity and equity in linguistic and cultural contexts beyond the Global North.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Human-centered computing**;

Additional Key Words and Phrases: Gender bias, Indic languages, Global South, India, Hindi, Community centric

ACM Reference Format:
Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. 2024. *Akal Badi ya* Bias: An Exploratory Study of Gender Bias in Hindi Language Technology. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3630106.3659017

*Work done while at Microsoft.

1 INTRODUCTION

Large Language Models (LLMs) continue to exhibit increasingly human-like precision across various tasks, leading to their integration into a wide range of real-world applications [49, 76]. As these technologies become more readily available and utilized in a multitude of languages, it becomes critical to understand, identify, and address certain critical biases that may appear. Previous research indicates that Natural Language Generation (NLG) models have the potential to generate or intensify biases and this leads to negative impacts on specific user groups and marginalized communities [14, 46, 55, 57, 83]. Gender bias, in particular, is a critical topic of concern. Gender biases that exist in language technologies can perpetuate under-representation, stereotyping, or misrepresentation of women and gender minorities [78]. Addressing gender bias in technology is crucial to bridge the digital gender gap and promote a more inclusive and equitable digital society [31].

Due to increasing adoption of language technologies in various languages, it is imperative to understand the biases these models can propagate not only in English but also in different languages and cultures. Unfortunately, however, much of the research in measuring and mitigating gender bias is in the context of the English and the Global North. Little is known about how to measure and mitigate gender bias in the context of Global South. The largely understudied dimension of gender bias in the context of Global South specifically for Hindi serves as the primary focus of this work. Filling this critical gap, we present the first comprehensive study of gender bias in Hindi. Our study highlights particularly in the context of India, it is difficult to utilize the parameters, benchmarks, and guidelines developed for identifying gender bias in English for Indic languages.[1] Figure 1 shows the pipeline of our experiments.

We conducted several experiments for mining gender-biased data in Hindi from different sources. Our experiments of mining include lexicon and heuristic-based approaches of mining, computational models for automatic classification of gender bias, and GPT-based generation of biased sentences. We explored data sources like social media comments, news media, and translation of existing gender bias datasets. Our experiments highlight several key challenges in mining gender-biased data in Hindi. We found that a large amount of data available online is Anglo-centric and hence does not serve as a good source for creating gender bias identification dataset in the Indian context. Mining social media data is extremely difficult due to growing restrictions. Heuristic-based approaches return a higher percentage of false positives. Computational models show poor performance due to limited cross-lingual and cross-domain transfer capabilities. Translations from industrial translation systems produce extremely formal and non-contextual translations. Finally, GPT generations show a limited diversity of themes.

Given the challenges faced with creating a gender bias dataset in Hindi using popular mining techniques proposed in past work for English, we conducted community centered field studies to bootstrap the collection of such sentences. Even though the prevalence of technology is growing in rural India, their opinion is often ignored in development of these technologies. For our field studies we employ rural, low income women, to include alternative voices, promoting empowerment within marginalized communities that are disproportionately affected by AI [10, 48, 70, 82]. We first aimed to understand what is the shared understanding of gender bias within a Hindi speaking community. The first field study was designed to elicit a culturally relevant definition of gender bias, and crowd-source gender biased statements via activities and plays. In the second field study we conducted a crowd-sourced annotation study to identify varying degrees of gender bias generated by GPT. Our first field study revealed variability in perceptions of gender bias. A simple gamified and interactive approach helped in gaining tacit knowledge about gender stereotypes. Our gender bias annotation workshop highlighted the importance of designing annotation tasks while keeping a variety of audiences

[1]We follow the previous works [26, 33] that state Indic languages as a superset, constituting Indo-Aryan, Dravidian, and a few low-resource languages belonging to the Austroasiatic, Sino-Tibetan, and Tai-Kadai families.

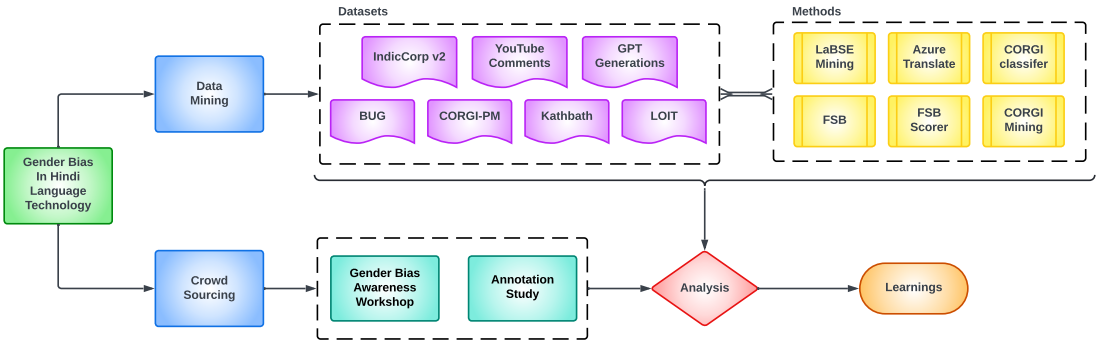

Fig. 1. Pipeline of our experiments

in mind. The Best-Worst Scaling comparative annotation framework that showed promising results with an urban audience for Hada et al. [35] was found to be complex by the rural crowd-workers employed in our study.

Our findings show that the study of gender bias involves subtle nuances making it a complex topic. Bringing this to an understudied and highly gendered context such as India is even more challenging. We navigate through the various complexities of identifying gender biased statements in Hindi. The key contributions of our work are as follows:

- We conduct in-depth experiments for mining gender biased sentences in Hindi.
- We conduct field studies adopting a community centered approach for gender bias identification, and employing rural, low-income community women to foster minority opinion.
- We highlight some of the critical challenges faced in mining sentences and field studies that other researchers and technologists should be aware of when engaging in further study of gender bias for Indic languages.
- We make recommendations on identifying and mitigating gender bias that focuses on the inclusion of communities right from the beginning.

We hope that our experiments and case studies can provide a strong foundation upon which to further explore the complex and critical nature of gender bias in Indic languages.[2] [3]

2 RELATED WORK

Bias Identification: Bias Identification is a crucial preliminary step in recognizing their presence in existing models. This can be done by designing special metrics or scoring frameworks to assign a "bias" score for sentences, documents, or machine-generated synthetic data [62]. Kaneko et al. [43] propose a Multilingual Bias Evaluation score, to evaluate bias using English attribute word lists and parallel corpora without requiring manually annotated data. Hada et al. [35] generate a dataset of GPT-generated sentences with normative ratings for gender bias and show that bias occurs on a spectrum. Benchmarks such as CrowS-Pairs [61] and StereoSet [60] aid in measuring various forms of social and stereotypical biases in language models. Rudinger et al. [74], Zhao et al. [92] release coreference resolution style WinoBias and WinoGender benchmarks and methods to help identify gender bias in existing co-reference resolution systems. [79] combine the two aforementioned benchmarks and devise an automatic gender bias evaluation method for

[2]The title "*Akal Badi ya Bias*" is a word play on the Hindi proverb "*Akal Badi ya Bhains*". The proverb translates to "Is wisdom greater or is the buffalo?" in English. It is used to imply that someone is behaving foolishly or lacking common sense.
[3]Code and data available at: https://aka.ms/AkalBadiyaBias

eight languages with grammatical gender, based on morphological analysis. Similarly, DisCo [89] is a metric to identify gendered correlations in publicly available pre-trained models. Dev et al. [20] present a practical framework of harms and a series of questions that practitioners can answer to guide the development of bias measures. Ramesh et al. [71] evaluate gender bias in Hindi-English Machine Translation. They evaluate Google Translate and the Hi-En OpenNMT model for gender bias using existing metrics WEAT and TGBI.

Debiasing Methods: Once the existence of a certain bias is identified in a language model (LM), it is important to address and mitigate the bias before safely deploying the model in the real world. Kaneko et al. [42] survey different debiasing methods and conclude that extrinsic evaluations, i.e., evaluations that are dependent on LMs performance on a specific task, and intrinsic evaluation measures do not have a strong correlation. Lauscher et al. [50] introduce "adapter" modules into the original LMs and train the LM on a counterfactually augmented corpus while keeping the rest of the parameters frozen. The authors evaluate this method using both intrinsic and extrinsic measures and show that it is effective in mitigating gender bias in LMs. Barikeri et al. [6] introduce a conversational dataset – REDDITBIAS – that can be used to debias LMs for dialog tasks. They show that this dataset allows for bias identification and mitigation across four dimensions: gender, race, religion, and queerness. Pujari et al. [69] propose a 2 step method for debiasing gender biased Hindi words. First step is to learn the bias space from set of definitive-gendered word pairs and second is to measure the biasness of a biased word. They show that their method is useful in decoupling the debiasing process from the word embedding process. Kirtane and Anand [47] invesitgate debiasing methods for Hindi and Marathi. They propose debiasing by using partial projection of vectors. Using partial projections overcomes the issue with linear projection where some word vectors which are gendered by definition were changed. In partial projection instead of zero magnitude along the gender direction, they project a magnitude of constant μ along with it. They show how their debiasing method works with different techniques such as RIPA and PCA.

Challenges: Capturing, measuring, and evaluating all types of biases in language models or crowdsourced data has its challenges. Orgad and Belinkov [63] find that only a few extrinsic metrics are measured in most studies and that datasets and metrics are often coupled. They discuss how their coupling hinders the ability to obtain reliable conclusions, and how one may decouple them. Draws et al. [27] argued that cognitive biases in crowdsourced data can also go unnoticed unless specifically assessed or controlled for. Sharifi Noorian et al. [75] show how crowdsourced elicitation can be inherently biased by the open and closed-world perceptions of annotators. Zhou et al. [93] show that debiasing a model trained on biased toxic language data is not as effective as simply relabeling the data to remove existing biases. Completely mitigating gender bias from models is a hard task as Bolukbasi et al. [12], Caliskan et al. [15], Dev et al. [20], Färber et al. [28], Spinde et al. [77] show that the societal and cultural prejudices are deeply embedded within the training data due to the presence of bias, and these biases, whether explicit or implicit, can significantly impact the functionality of the NLP systems [73]. We refer the readers to Stanczak and Augenstein [78] and Devinney et al. [21] for detailed surveys of Gender Bias in Natural Language Processing.

Much of the past work in measuring and mitigating gender bias focuses on English and the context of Global North. There is limited understanding of how to measure and mitigate gender bias in the context of the Global South. The emphasis of past work on understanding and mitigating gender bias in western context has resulted in a gap, as western perspectives may not fully capture the nuances of gender bias in the context and languages of Global South. In this paper, we study identification of gender bias specifically for Hindi and Indian context. We conduct several experiments to mine gender biased statements in Hindi based on existing methods proposed for English. Existing methods had several limitations when used for Hindi in the Indian context. Due to the challenges faced with mining such sentences we conduct field studies and adopt a community centered approach for gender bias identification to bootstrap the collection

of such sentences. When it comes to identifying gender bias in Indic languages, including communities that are fluent in these languages is extremely important [1]. Moreover, for the subjective task of gender bias identification it is crucial to include a diverse set of annotators as perceptions of bias depends on factors such as lived experiences, background, community, and others [35]. There is an increasing adoption of technology in rural India, however, their opinions are often unheard and overlooked while these technologies are created [68]. It is imperative to foster alternative voices and minority opinions during the development of these technologies [8, 34, 70]. The inclusion of users in the research process presents substantial potential for mitigating power imbalances and fostering empowerment within marginalized communities that experience disproportionate impacts from AI [10, 48, 70, 82]. In such a case, crowdwork, has emerged as a significant sector in the Indian labour market place [24, 25], as an effective method to include local Indic language speaking communities. Additionally, from 2018 to 2022 the portion of household in India with smartphones has doubled from 36% to 74.8% [68]. This proliferation of smartphones in India makes crowdwork platforms more accessible on mobile interfaces [17, 39, 86] allowing low-income local language speakers with only basic qualifications to benefit from participating in data annotation work.

3 SOURCING HINDI DATA FOR GENDER BIAS IDENTIFICATION

To mine gender-biased data specifically for Indian context we tried several existing methods that are used to mine gender-biased sentences in English or other languages. In this section, we describe the datasets we explored, our methods, and an analysis of our findings.

3.1 Datasets

CORGI-PM:. CORGI-PM is a Chinese corpus for gender bias probing and mitigation [90]. The corpus consists of 32.9K sentences with gender bias labels derived by following an annotation scheme specifically developed for Chinese context. To create the corpus, the authors follow an automatic method of data extraction from a raw corpus that gives them potentially gender-biased sentences. These sentences are then annotated for gender bias. Specifically, to create their corpus they follow a two-step filtering process. In the first step, they build a vocabulary of words that have gendered associations. In the second step, they use the list of gendered keywords to recall sentences from the raw corpus. These sentences are then re-ranked, and a threshold for sentence selection is determined. The selected sentences are annotated for gender bias. The corpus contains 22.5K non-biased sentences and 5.2K biased sentences.

IndicCorp v2: IndicCorp v2 is a large-scale collection of monolingual corpora for Indic languages , containing a total of 20.9 billion tokens across 23 Indian languages and English [26]. IndicCorp v2 reflects the contemporary use of Indic languages and covers a wide range of topics, primarily crawled from news articles, magazines, and blog posts. The authors source their data from popular Indian-language news websites, discovering most of their sources through online newspaper directories (e.g., w3newspaper) and automated web searches using hand-picked terms in various languages.

Kathbath: Kathbath is a read speech corpus [38]. The text used to create the corpus is derived from IndicCorp [41]. IndicCorp is a large collection of monolingual corpora consisting of 12 Indic languages collected from diverse Indic-specific sources. The authors take a subset of IndicCorp (100K sentences) for each of the 12 languages while limiting the sentence length to 8-15 words and allowing only for alphanumeric characters. In this work, we consider the Hindi transcripts of the created corpus to mine biased sentences.

BUG:. BUG is a large-scale gender bias dataset for coreference resolution and machine translation [51]. BUG contains 108K English sentences, sampled semi-automatically from large corpora using lexical-syntactic pattern matching. To

create the dataset, initially, a syntactic search is performed to identify sentences with challenging syntactic properties across corpora from three domains Wikipedia, Covid-19 research, and PubMed abstracts. Subsequently, the sentences are filtered to ensure they include at least one entity and a corresponding pronoun. The sentence is marked stereotypical or anti-stereotypical for gender roles. Lastly, a manual assessment of BUG is conducted. The dataset consists of 54K stereotypical sentences, $\approx 30K$ anti-stereotypical sentences, and $\approx 24.5K$ neutral sentences.

YouTube Comments: Comments or posts on social media platforms represent the thoughts of individuals and communities [36, 58]. . Therefore, to study gender bias in the context of India, we collected comments from YouTube via the YouTube API. First, we curated a list of search queries on topics where gender bias data or polarisation was expected. The list of queries are: *"Deepika Padukone Cleavage Controversy"*, *"Hijab ban controversy"*, *"Meninist' Deepika Bhardwaj Has A Few Questions For Feminists In India"*, *"Swara Bhaskar Marriage Controversy"*, *"Manipur Women Paraded"*, *"Kanika Kapoor's COVID-19 Flak"*, *"SSR and Rhea Chakraborty"*. We took the first 10 videos that appeared for the search queries and extracted up to 105 comments per video. This gave us a total of 7340 comments.

Lot of Indic Tweets (LOIT):. LOIT is a dataset of Hindi and Telugu tweets[4]. LOIT contains most of the Hindi tweets made between 13th January 2017 and 31st December 2018 and Telugu tweets made between 1st January 2010 and 25th June 2019. Twitter allows users to be their natural selves often leading them to portray the biases that they may have than otherwise.

Fifty Shades of Bias (FSB):. FSB is a dataset of 1000 GPT-generated English text with normative ratings for gender bias [35]. The dataset is created by prompting GPT systematically. The authors first create a seed set of sentences sourced from various corpora. Using the seed set, the authors prompt GPT-3.5-Turbo to either convert or complete a sentence to its gender-biased variation. The generations were then annotated in a comparative annotation setup to assign a gender bias score to each sentence.

3.2 Experiments

LaBSE Mining: Inspired by Albalak et al. [3], we employed an open-retrieval for emergent data collection. To analyze the abundance of biased sentences in existing large-scale corpora, we choose the Hindi subset of IndicCorp v2 [26] and sample 10M sentences from it. We mine biased sentences from this sample using the LaBSE model[5] [29], and the top 100 biased sentences from FSB as the queries. LaBSE is a language-agnostic embedding model based on BERT [22], which is trained with a contrastive loss and generates close embeddings for sentences that are similar across 109 languages [29]. In our experiment, we first cache the embeddings of all the 10M sentences from the sample using FAISS-DB[6], and query it using the embedding of a source English sentence from FSB. The top 5 most similar sentences (in terms of cosine similarity) for each query in Hindi for a given source were collected, leading to a total of 500 *potentially biased* sentences. From these 500 sentences, we randomly sample 200 sentences (2 per source sentence) and 2 authors of this paper go through 100 sentences each to classify them as "biased" or "not biased". We found that $\approx$ 20% of the sentences were biased.

Translating to Hindi: We took a random sample of 100 sentences each from CORGI-PM [90] and BUG [51] datasets. The sample was taken from sentences that were marked as biased. These sentences were translated into Hindi using Azure Translate. One of the authors of this paper annotated the 200 translated sentences to check if the sentence maintains gender bias upon translation and is relevant to the Indian context. We found that 24% of the translated BUG

[4]http://bpraneeth.com/projects/loit
[5]https://huggingface.co/sentence-transformers/LaBSE
[6]https://github.com/facebookresearch/faiss

sentences and 33% of the translated CORGI-PM sentences remained gender biased in Hindi, and could be used in the Indian context.

CORGI classifier: We train a binary classifier on the CORGI-PM [90] dataset to classify sentences as "biased" or "not biased". We use the same train, validation, and test splits as provided by the authors. We fine-tune mBERT[7] [22] and the corresponding hyperparameters are provided in Table 2 in the Appendix. The model achieved an accuracy of 0.81 on the test set. We used this binary classifier on the 7340 YouTube comments we extracted. Out of 7340 comments only $\approx$ 4.67%(343) comments were classified as biased. We took a random sample of 100 comments from the 343 comments and one of the authors of this paper classified them as "biased" or "not biased". We found that 36% of the 100 comments were biased.

FSB scorer: Using the 1000 English sentences from FSB [35] as a training set, we finetune the IndicBERT v2[8] model [26] using LoRA [37], to avoid over-fitting. IndicBERT v2 is the SOTA NLU model for Indic language and cross-lingual transfer to English. A regression head was appended to the model and the scores predicted were squeezed between -1 and 1, using a *tanh* activation, similar to FSB. Sweeps were conducted for optimal hyperparameters, and Table 3 in the Appendix provides the final hyperparameters chosen for training. During the final finetuning phase, the model and best hyperparameter configuration achieved an MSE of 0.057 and Pearson correlation of 0.85 on the validation set which is comparable with the score obtained by Hada et al. [35]. Using this gender bias score prediction model we obtain a gender bias score for the 7340 comments extracted from YouTube. We sample the top 100 gender bias-scored comments and one of the authors of this paper classified them as "biased" or "not biased". We found that 21% of the 100 comments were biased.

CORGI mining: We take a list 684 of Hindi adjectives from the Internet. Using these adjectives, we follow the steps as described by the authors of the CORGI-PM dataset [90] to recall sentences from raw corpora. The details are explained in Section 3.1. From the list of adjectives, we first find adjectives that have a female association. We do this by measuring the dot product of the word embedding of the adjective with the word embedding of ($\vec{Man}$ - $\vec{Woman}$) (we use Hindi words for man and woman). We conduct our experiments with IndicBERT v2 [26] and mBERT [23] to obtain the word representation. For raw corpora, we use the Hindi sentences from Kathbath and a subset of 10M Hindi tweets from the LOIT dataset. Before we filter the sentences, we normalize the scores associated with each Hindi adjective. We set a threshold to select the top female-leaning adjectives and filter sentences from the corpora that contain these adjectives. The mBERT thresholds used to filter 100 sentences each from Kathbath and LOIT are 0.45 and 0.33 respectively. Similarly, the IndicBERT threshold used to filter 100 sentences each from Kathbath and LOIT is 0.904. This gives us a total of 400 sentences. These filtered sentences are then manually inspected by one of the authors for bias. For sentences retrieved using mBERT we find that 3% of them were biased from both LOIT and Kathbath. For sentences retrieved using IndicBERT v2 we find that none of them were biased.

GPT Generation: Using the seeds and the prompts provided in FSB [35], we generate 1800 potentially gender biased sentences in English. Six of the authors of this paper went through 300 sentences each and marked if these sentences are "non-sensical" especially in the Indian context and should be removed. We found that $\approx$ 8% of these sentences were marked as "non-sensical", and the rest could be used in the Indian context, and showed a gradation of gender bias[9].

[7] https://huggingface.co/bert-base-multilingual-cased
[8] https://huggingface.co/ai4bharat/IndicBERTv2-MLM-Sam-TLM
[9] We do not annotate these sentences as biased or unbiased because Hada et al. [35] show in their work that the generations in FSB have a gradation of gender bias. Instead, upon initial examination, we found that some of the generated sentences did not make sense.

Dataset	Method	# of samples annotated	Yield (in percent)
IndicCorp v2	LaBSE mining	200	20
BUG	Azure Translate	100	24
CORGI-PM	Azure Translate	100	33
YouTube comments	CORGI classifier	100	36
YouTube comments	FSB Scorer	100	21
GPT generations	FSB	1800	92
LOIT & Kathbath (mBERT)	CORGI mining	200	3
LOIT & Kathbath (IndicBERT)	CORGI mining	200	0

Table 1. Yield of potentially gender-biased text from different data sources and methods.

3.3 Results and Analysis

With the above datasets and methods, we saw varying yields of potentially gender-biased statements in the Indian context. Table 1 gives a summary of the yield for the different methods and datasets. In this section, we discuss the key challenges faced in using each of these methods and datasets:

Internet data is Anglo Centric: Gathering data from the internet for natural language processing tasks, especially when focusing on Indian languages and topics such as gender bias, poses a significant hurdle. In the context of NLP applications, data collection often involves web scraping. However, it's essential to acknowledge that the majority of data available on the internet tends to align with dominant viewpoints and consists primarily of content in the English language [8]. This prevalence of English-centric content on popular internet platforms makes it exceedingly challenging to obtain relevant data for languages other than English, particularly those spoken in regions like the global south. For our experiments, we used IndicCorp v2 [26] which was a large-scale effort to collect data in Indian languages. In our experiment of LaBSE mining of gender-biased Hindi statements, using 100 gender-biased statements from FSB [35] as source we found only 20% of the examined statements to be biased. A higher yield from this method was expected because we sampled 10M Hindi sentences from IndicCorp v2 and picked 500 sentences that were most similar to very explicitly biased statements from FSB. The low yield could be attributed to the nature of sentences in IndicCorp v2. As mentioned earlier, the corpus was collected by scraping sentences from online news and media sources where the content might already be sanitized and censored.

Translation is a bottleneck: When data for a particular task and language is not available, the NLP community has often relied on the impressive performance of translation models to translate task data from a high-resource language to the target language. In our experiments, we translated biased statements from the BUG (an English language) dataset [51] and CORGI-PM (a Chinese language) dataset [90] to Hindi. From the biased statements we examined, we found only $\approx$ 27% of them to maintain their bias after translation. Out of the unbiased sentences, many sentences did not make any grammatical sense after translation or were not contextually relevant. A known challenge in the use of translation models is that they generate excessively formal renditions, potentially diluting the original colloquial or informal nuances present in the source statements[53]. For example, "teacher" was translated to "*adhyapak*", and "wife" was translated to "*grihani*". In colloquial Hindi these words are "*teacher*" and "*bahu*" respectively. Furthermore, the nuanced contextual element in the original statements are often not transpositioned completely during the translation process, leading to potential misinterpretations [33, 84, 91]. Beyond these challenges, variations in linguistic structures, idiomatic expressions, and cultural nuances pose additional complexities, impeding the seamless transference of gender biases across linguistic boundaries [72].

Problems with collecting social media data: Researchers have traditionally turned to social media platforms like Twitter, Reddit, and Meta when studying concepts like offensive language, hate speech, and identity attacks [32, 36, 58, 88]. Social media data has several advantages like a diverse user base, and users expressing them freely and spontaneously . However, the majority of this data is still in English. For instance, a survey conducted in June 2023 by Statista indicated that approximately 50% of the desktop traffic on Reddit originates from the United States [81]. This prevalence of English-centric content on popular internet platforms makes it exceedingly challenging to obtain relevant data for languages other than English, particularly those spoken in regions like the global south [2, 40]. Data that originates in India is still majorly in English as social media platforms are mostly used by urban population [80, 85]. Moreover, recent trends in internet platforms have seen an increase in restrictions on data access, making it even more difficult to access social media data for research purposes [19]. This tightening of data accessibility exacerbates the already challenging task of procuring suitable data for gender bias. For our experiments, we extracted 7340 comments from YouTube spanning over 10 controversial/polarising search queries. Our experiment with the CORGI classifier shows that only $\approx 4.6\%$ comments were classified as biased by the classifier, and a random sample of 100 comments showed that only 36% of those comments were correctly classified as biased. Our experiment with FSB scorer shows that comments from YouTube show a distribution of gender bias score, with a skew towards non-biased or neutral comments as observed in Figure 2 in the Appendix. A sample of top-scoring 100 comments showed that only 21% of these comments are actually biased. Our analysis of random samples also revealed that the majority of the comments are in English, often target individuals over communities, and are highly profane in some cases. Therefore, collecting gender-biased Indian context data from social media platforms is a challenging task as it depends on various factors like appropriate selection of topics, choosing the right signal to boost the representation of biased comments, finding comments in a language other than English, and more.

Loss due to limitations of cross-lingual or cross-domain transfer of embeddings: Recent works [4, 26, 33, 52, 56, 66] have shown the challenges and generalizability of cross-lingual transfer in multilingual language models. We hypothesize that this could be one of the reasons for the poor performance and yield of the CORGI classifier and FSB scorer. CORGI classifier was finetuned and evaluated on Chinese and Hindi respectively. Another reason we attribute to their poor performance is the domain mismatch between the training and test data. CORGI was trained with non-social media data but was used to predict scores for YouTube comments. Similarly, the FSB scorer is finetuned with the limited FSB data, which are rudimentary and artificial generations in English created using GPT but is evaluated on the domain of social media content.

Heuristic-based methods of retrieval are challenging: In the study of gender bias, researchers have used heuristics such as retrieval from raw corpora using gendered adjectives [90], using professional words [51], and template-based sentence creation [60, 78]. Our experiment of retrieval of gender-biased statements from raw corpora using gendered adjectives returned only 3% true positive sentences when using mBERT and 0 true positives when using IndicBERT. We also tried topic-based retrieval from YouTube, coupled with classification using computational models as discussed previously. Both these methods returned a high number of false positives indicating the limitations and challenges of using heuristic-based approaches for retrieval. Additionally, a known limitation of the LLMs is their lack of cultural awareness. While models like LaBSE, IndicBERT, mBERT, and others might be linguistically aware of multiple languages, they lack cultural and social knowledge of different communities primarily because they are not exposed to topcially diverse data from different communities [16]

Data diversity: One of the main components of tasks like gender bias is to have diverse and representative data. The meaning of a sentence is perceived based on one's identity; expression of gender bias in language can often entail cultural nuances, hence, it is equally important that the data is contextually situated [9, 11, 59]. In our experiments, we tried various data sources such as IndicCorp v2 which is made primarily of news articles, comments from socia media sites, and text generation from GPT. Generated sentences from GPT showed the most promise in terms of getting a high proportion of biased statements. However, the generated data has limited diversity. We generate up to 4 sentences per seed provided in FSB. The themes are repetitive, sentences do not capture Indian cultural nuances, and are very simple hence do not have idiomatic or linguistic diversity either. The biggest challenge in generating data with this method is a careful selection of diverse seed sets, as the seed sentences can have a significant impact on the generation of biased sentences.

4 CROWDSOURCING CONTEXTUALLY RELEVANT GENDER BIAS DEFINITION AND ANNOTATIONS

Due to the difficulties encountered in developing a dataset on gender bias in Hindi using established mining techniques previously suggested for English, we conducted field studies to bootstrap the compilation of such sentences. We do this by adopting a community-centered approach for gender bias identification. This section explores two field studies we conducted in the pursuit of understanding how to define and identify gender bias in Hindi.

4.1 Case Study 1: Gender Bias Awareness Workshop - Lucknow, UP

In this field study we aimed to gain a shared understanding of gender bias within a specific community and crowd-source gender bias sentences in Hindi that could inform our data collection process. Definitions of gender bias cannot be constant across cultures and contexts. Global North definitions of gender bias in particular, should not be used to assess situations and environments in the Global South. Braff and Nelson [13] highlight how the Global North has played a critical role in shaping gender norms and structures as they are often spoken about today. These are however neither universal nor natural. It is therefore crucial to consider the cultural, political, and economic contexts of each region when addressing gender issues. In this case study where gender bias needed to be spoken about to women, a certain challenge was met in finding the right words and expressions to accurately convey messaging. As an example, the Hindi word for "bias" in the way it is understood in English does not exist; rather commonly used words for this topic instead indicate "discrimination" "favoritism" or "exclusion". To this end, we conducted a qualitative experiment to test whether definitions of gender bias in Indian languages could be generated through community-centered methods. Working in collaboration with the Milaan Foundation, a prominent non-governmental organization in India that has been working on women's empowerment and gender bias for over 20 years, we developed a Gender Bias Awareness Workshop. The workshop was held in Lucknow district of Uttar Pradesh state in Northern India with 14 women aged 18-24. Most of these women were enrolled in their bachelor's degree programs and had also been a part of previous gender-related activities held by the Milaan Foundation. All workshop facilitation and engagement from participants occurred in Hindi to ensure the capture of language-relevant expressions of gender bias. The workshops were structured around two primary components: an exploration of fundamental concepts and the implementation of four key activities to ensure inputs and involvement from the community. In the conceptual exploration participants delved into the distinctions between sex and gender, analyzed gendered language within the context of Hindi, understood the nuances of gendered and non-gendered explanations, identified prevalent instances of gender bias in the daily lives of Indian women, and shared their personal experiences of gender bias. Complementing these discussions were four interactive activities: a storytelling exercise involving two frogs to uncover backstories, skit and role-play scenarios portraying gender

dynamics in various situations, a collaborative compilation of a list highlighting bias towards women, the creation of a shared definition of gender bias.

4.2 Case Study 2: Annotation Study - Kannauj, UP

We conducted an annotation study in the Kannauj district of the Northern Indian state of Uttar Pradesh to investigate the identification of gender bias in sentences generated by GPT-3.5-Turbo. Hada et al. [35] show in their work that humans can identify gender bias to varying degrees. They use an efficient comparative annotation framework called Best–Worst Scaling (BWS) to obtain a degree of gender bias score per statement. They argue that BWS has shown to be very effective for annotation of subjective tasks such as offensive language and show it can be extended to gender bias. We followed the data generation and annotation process described in their work. Hada et al. [35] have their data annotated by an urban population, and highlight the importance of incorporating subjective judgments for this task. In contrast, we have our data annotated by a rural population, to include minority opinion for such tasks. Specifically, the study focused on employing women from rural and low-income communities to annotate a Hindi text corpus and identify sentences with gender bias. The overarching goal was to explore the feasibility of employing low-income women to generate foundational annotated datasets that could support the automation of gender bias detection and mitigation in Hindi corpora. This field study hypothesized that by employing women from Hindi-language speaking communities to highlight bias in GPT-generated corpora, identified sentences would be representative of the biases, stereotypes, and marginalization that the larger population of Hindi-speaking women may face. **Study Structure & Participants:** This study framed the annotation of gender bias sentences as digital work tasks that were deployed through the Karya App. 2000 English sentences were generated by GPT-3.5-Turbo using the prompting strategies described in [35] to create a corpus of sentences. Specifically, given a seed statement, we prompted GPT-3.5-Turbo to generate a gender-biased completion or conversion of the statement. We randomly sample sentences from COPA as our seed statements. These gender-biased generations were then translated into Hindi. We initially used machine translation systems, however these did not yield positive results as the outputs did not pass manual quality checks. The final sentences for the task were manually translated for this study to ensure higher quality. This translated corpora was used to create 2N 4-tuples for the BWS setup as described in other works [30, 35, 36, 44, 45, 54, 64, 65]. The tuples were randomly assigned to each participant. The participants were shown a tuple and asked to identify which sentence was "most biased" and which sentence was "least biased". Each participant was given 261 tuples to annotate. Participants of this study involved 15 women from low-income backgrounds who underwent a two-hour virtual training delivered by the research team. All participants were native or fluent Hindi speakers who identified as women, and 46% were aged 18-25. The majority reported being part of a marginalized caste (93%), practicing Hinduism (94%), and being married (74%). Education levels varied significantly, with 33% of women not passing the 10th grade and 30% with a High School Diploma or Bachelor's degree. Income levels were low, with 66% coming from households earning less than INR 12,000 (USD 144) per month. Most women reported being unemployed and without a steady income; when working, these women were predominantly engaged in agricultural or domestic work.

4.3 Results and Analysis

Gender Bias Awareness Workshops. During the initial stages of the workshop, participants expressed confidence in understanding gender bias, perceiving no need for discussions on the distinctions between sex and gender. As sessions progressed, however, it became evident that many participants were inadvertently conflating these concepts. Throughout the workshop, participant enthusiasm remained high, with active engagement and occasional challenges to

facilitators on the concepts they brought forth. The frog exercise worked well in exposing prevalent gender stereotypes as participants assigned stereotypical attributes to the frogs based on gender. For the "female" frog, the participants were more likely to use adjectives such as "afraid", "weak" and "unsure"; whereas for the "male" frog, words like "brave" and "curious" were assigned. Interestingly for the "male" frog, perceptions were not as constant among the participants. Some participants assigned the male frog adjectives with more negative connotations such as "stupid" and "concerned only about seeming courageous". In the skit and role-play activities, a shift towards more equitable portrayals emerged, indicating a loosening of rigid gender norms. About 75% of the paired teams decided to actively portray scenes that depicted equity and parity between men and women across various scenarios. Skits where "bias of roles" were still present were caveated through a warning that the skit represented the norms and ways their parents or people in the community would act. There became an interesting emphasis at this point on what was "right" and what was a "wrong" way to speak about people, with the overwhelming conclusion being that gender has nothing to do with why a person might behave a certain way. The creation of a shared definition posed challenges, primarily due to participant fatigue, yet the exercise yielded valuable insights, uncovering over 80 unique sentences illustrating various gender biases, phrases, and words contributing to women's marginalization in everyday scenarios. Despite the difficulty in crystallizing a definitive definition, the workshop successfully captured nuanced perspectives on gender bias.

Annotation Study. The annotation study revealed several critical insights surrounding the scalability of this type of work. Key challenges occurred in achieving agreement among annotators, with initial agreement values around 0.08, indicating random annotations by the participants. After filtering out annotations completed in less than 30 seconds, a slight improvement to 0.11 in agreement was observed. However, 30% of annotations selected the same item for both "most" and "least" biased sentences, suggesting issues with the task structure and understanding. Interviews with participants confirmed the supposition that participants had difficulty in completing the task. Most of the women highlighted that they did not understand the virtual training that occurred. While some were not prepared to be connected via a virtual mode facing challenges connecting through video conferencing, others found the language and overall explanation of the tasks too confusing and difficult to follow. Further probing highlighted sentence construction and comprehension as a significant barrier; women either found the formal Hindi used in the sentences difficult to understand or felt that the sentences were "nonsensical". A full manual review of the dataset by native Hindi speakers found that approximately 30% of sentences were identified as not aligned with the cultural and social contexts of rural India. Roughly 11% of the sentences did not have a clear meaning when translated, examples of these include "the country discovered new land". 17% included words like "dating" or "cowboy", which are not commonly found in the Indian vocabulary. A precipitating point of confusion was the formality of translation, many of the words in this grouping were transliterated in the annotated corpora which led to misunderstanding amongst the annotators. A few sentences ($\tilde{2}$%) described scenes or turns of phrase that were unfamiliar to the Indian context in general, such as "thanksgiving", "bachelorette" and "mowing the lawn". Results from this field study indicate that critical revisions in design and process are required to make this type of task understandable and accessible to participants. The key areas identified for improvement are three-fold. First, establishing more intensive training elements and administering pre-testing to ensure task completion at quality. Second, ensuring that GPT-generated sentences match the social and cultural contexts of the Hindi-speaking community. Third, creating a simpler and more intuitive task structure and expectations.

5 DISCUSSION

Bring all findings together, our work suggests that studying gender bias especially in the context of the global south or more specifically India can be extremely challenging. Our extensive efforts, spanning experiments and on-field studies, showcase the intricate landscape of gender bias, revealing the need for nuanced strategies and inclusive approaches. We tried several methods of mining gender-biased data from different sources. Our experiments with mining gender-biased sentences highlighted various difficulties. Internet data being very Anglo-centric does not serve as a good source for mining gender biased data for the global south. Extracting data from social media has become increasingly difficult due to growing restrictions. The data we extracted from YouTube based on certain topics had comments majorly in English. From the random sample we annotated based on results from classifiers or gender-biased scoring, only a small portion was actually biased indicating the need for a more careful selection of topics, and domain-specific computational models. We used IndicCorp v2 to mine gender-biased sentences in Hindi. Our experiments with LaBSE embedding and explicitly gender-biased statements, a common method to retrieve similar sentences in two different languages, returned many false positives. This could be attributed to the fact that IndicCorp v2 has been sourced primarily from news outlets containing sentences that are sanitized and censored. Our experiments also highlighted the limitations of current language embeddings, showing limited performance in cross-lingual and cross-domain transfer capabilities. We found using translation systems to be a bottleneck due to their inability to translate sentences contextually and colloquially. We also found keyword-based approaches to retrieval to be ineffective. Finally, the generation of gender biased sentences from GPT showed some promise, however, had a very limited diversity.

Challenges faced with mining gender biased sentences in Hindi motivated us to conduct field studies to crowd-source such sentences. To incorporate real-world perspectives we delve into the nuanced interpretation of gender bias by conducting two on-field studies. Our case studies bridge the gap between controlled experiments and the dynamic realities of crowd-sourced, culturally relevant insights. We conducted on-field studies to crowd-source gender biased Hindi statements with culturally relevant definition of gender bias and a comparative annotation task for identifying gender-biased statements. The gender bias awareness workshop revealed several key takeaways. The initial overconfidence expressed by participants in understanding gender bias, juxtaposed with their evolving comprehension throughout the workshop, underscores the subtle complexity of the issue. An interactive gamified frog exercise helped us tap into the tacit knowledge of the workshop participants by revealing ingrained gender stereotypes, emphasizing the pervasive nature of biased attributions. The skit and role-play activities provided a platform for a positive shift, with a significant majority actively portraying scenes that challenged traditional gender norms, signaling a growing inclination towards equitable representations. Of notable interest was the acknowledgment of variability in perceptions. Discussions within the workshop revealed personalization of gender bias for participants. Overall, the workshop highlights that gender bias has ill-defined boundaries and a highly subjective interpretation of the concept based on lived experiences, background, education, and other factors. Our workshop not only unveiled the intricacies of biased language and perceptions but also highlighted the need for continued efforts in fostering inclusive dialogue with the community. Our pilot annotation study for annotating gender-biased statements in a comparative annotation setup highlighted challenges with crowdsourcing annotations for this task. One key takeaway was that low-income or rural crowd workers present an additional challenge when dealing with gender bias. While urban audiences can complete such tasks with high accuracy, it is important to think about how we can include minority voices and non-dominant viewpoints for subjective tasks such as gender bias [8, 34]. Post-annotation interviews with participants provided valuable context, unveiling a spectrum of challenges encompassing task understanding, virtual training, and language barriers. Women

participants, facing connectivity challenges and linguistic complexities, expressed their difficulties with the virtual mode and the formal Hindi used in the sentences. A manual review by native Hindi speakers identified 30% sentences as misaligned with the cultural context, indicating the necessity for culturally sensitive content. The complexity of translation as also revealed by our experiments, including transliterations and unfamiliar phrases, contributed to misunderstandings among annotators. This annotation study underscores the importance of designing annotation tasks while keeping a variety of audiences in mind. While such tasks have been tested predominantly with the WEIRD or an urban population, not including viewpoints from a rural population can lead to cultural erasure and propagation of hegemonic viewpoints [8, 67].

Overall, our study highlights several challenges faced with mining gender-biased data in the Indian context, subjective understanding of the concept with ill-defined boundaries, and including minority opinions for this task. In the future, it would be interesting to explore a more participatory approach for this task. Works leveraging "games with a purpose" (GWAP) [5, 87] can be adopted for crowdsourcing generation of gender-biased sentences in target language with cultural nuances, as these methods have shown promise to tap into the tacit knowledge of the crowd. For annotating gender-biased statements a more careful task design is required depending on the target audience, and effort should be made to include minority opinions. It would also be interesting to study the concept of intersectionality introduced by Crenshaw in her foundational work [18]. She emphasized that social categories like race, gender, and class are interlinked, thereby mutually creating unique dimensions of oppression that are not adequately addressed by frameworks that consider such social categories separately. Drawing on Crenshaw's work on intersectionality, it is pivotal that the datasets designed to capture gender bias in Indic languages contain sentences that account for bias emerging from these intersecting marginalized identities to authentically capture not just gender bias but its intersection with structural embeddings of caste, religion, and rurality, among others.

6 LIMITATIONS

Efforts to mine gender-biased data faced obstacles stemming from the dominance of Anglo-centric content on the internet. Additionally, the growing restrictions on social media further limited our ability to extract diverse and representative datasets. In future, it would be interesting to explore other social media sites like Meta and Twitter if necessary permissions can be obtained. Despite employing various methods, our experiments in mining gender-biased sentences revealed significant difficulties. The low prevalence of biased sentences in the annotated sample suggests challenges in topic selection and the need for domain-specific computational models. The use of translation systems as a bridge between languages revealed substantial limitations. Inability to translate sentences contextually and colloquially hindered our understanding of gender bias in diverse linguistic contexts. The complexities of translation, including transliterations and unfamiliar phrases, contributed to misunderstandings among annotators, underscoring the need for culturally sensitive content. Although the generation of gender bias showed promise, the method exhibited a lack of diversity. In future, it would be interesting to explore GPT generations with more contextual seeds and in-context examples, and changes in the prompt for increasing diversity and including culturally relevant information. The on-field studies faced challenges in annotating gender-biased statements due to the complex nature of the task. Connectivity issues, linguistic complexities, and cultural misalignments highlighted the difficulties faced by participants. These methodological limitations collectively point to the intricate nature of studying gender bias and emphasize the necessity for continuous refinement of our methods to better capture the complexities inherent in diverse cultural contexts, such as India.

7 ETHICAL CONSIDERATIONS

We use the framework by Bender and Friedman [7] to discuss the ethical considerations for our work.

- **Institutional Review**: All aspects of this research were reviewed and approved by Microsoft Research IRB.
- **Data**: Several publicly available datasets were examined in this study. We also prompt GPT-3.5-Turbo to generate gender biased statements. No personally identifiable information (PII) was collected or distribued in this process. No PII was included in prompt or in-context examples to GPT.
- **Annotator Demographics**: Annotations for the mining experiments were done by researchers interested in studying gender bias in language technologies. These annotators were from India and had native proficiency in Hindi and English. The annotators had at least an undergraduate degree as their minimum educational qualification. The field studies were conducted with rural, low income women. The demographics are described in section 4. The participants were recruited via a data annotation platform. The pay was adjusted after discussion with the company. The company particularly ensures fair pay (20 times the minimum wage) to low-income communities in India.
- **Annotation Guidelines**: For the gender bias awareness workshop, as described in section 4.1, we partner with an organization working in the domain of women's empowerment and gender bias for a long time. All modules were created by this organization. For gender bias annotation workshop we draw from the guidelines described by Hada et al. [35]. These guidelines were written in Hindi, and the task was explained verbally during a training session conducted online. Annotators were given detailed instructions, and walk-through of the task, with examples.
- **Impact on Annotators**: For the crowd-sourced annotation study we limited the number of annotations per participant, provided a mix of explicitly, implicitly, and neutrally biased sentences, asked annotators to skip and report instances they were not comfortable annotating, and lastly, encouraged open-dialogue with the workshop facilitator.
- **Methods**: In this study we explore several methods to mine gender biased data in Hindi. We discuss the challenges and limitations of these methods. We also trained computational models for automatic classification or scoring of gender bias. While these methods can be easily misused, our intent with this study is to highlight the limitations of these methods when used in the context of Global South.

REFERENCES

[1] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2819–2826. https://aclanthology.org/2020.lrec-1.343

[2] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 4232–4267. https://doi.org/10.18653/v1/2023.emnlp-main.258

[3] Alon Albalak, Sharon Levy, and W. Wang. 2022. Addressing Issues of Cross-Linguality in Open-Retrieval Question Answering Systems For Emergent Domains. *Conference of the European Chapter of the Association for Computational Linguistics* (2022). https://doi.org/10.18653/v1/2023.eacl-demo.1

[4] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4623–4637. https://doi.org/10.18653/v1/2020.acl-main.421

[5] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. 2022. Ready Player One! Eliciting Diverse Knowledge Using A Configurable Game. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 1709–1719. https://doi.org/10.1145/3485447.3512241

[6] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1941–1955. https://doi.org/10.18653/v1/2021.acl-long.151

[7] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[8] 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* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[9] Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the Effects of Annotator Gender across NLP Tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma (Eds.). European Language Resources Association, Marseille, France, 10–19. https://aclanthology.org/2022.nlperspectives-1.2

[10] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (<conf-loc>, <city>Arlington</city>, <state>VA</state>, <country>USA</country>, </conf-loc>) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3551624.3555290

[11] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[12] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 4349–4357. https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[13] Lara Braff and Katie Nelson. 2022. Chapter 15: The Global North: Introducing the Region. *Gendered Lives* (2022).

[14] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. https://doi.org/10.1145/3411764.3445372

[15] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230 arXiv:https://www.science.org/doi/pdf/10.1126/science.aal4230

[16] Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 11370–11403. https://doi.org/10.18653/v1/2023.emnlp-main.699

[17] Manu Chopra, Indrani Medhi Thies, Joyojeet Pal, Colin Scott, William Thies, and Vivek Seshadri. 2019. Exploring Crowdsourced Work in Low-Resource Settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300611

[18] Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43, 6 (1991), 1241–1299. http://www.jstor.org/stable/1229039

[19] Brittany I Davidson, Darja Wischerath, Daniel Racek, Douglas A Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F Roscoe, Laura Ayravainen, and Alicia G Cork. 2023. Platform-controlled social media APIs threaten Open Science. *Nature Human Behaviour* (2023), 1–4.

[20] Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 246–267. https://aclanthology.org/2022.findings-aacl.24

[21] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2083–2102. https://doi.org/10.1145/3531146.3534627

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:52967399

[24] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. [n. d.]. Mechanical Turk Surveys. https://demographics.mturk-tracker.com/. (Accessed on 09/11/2023).

[25] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 135–143. https://doi.org/10.1145/3159652.3159661

[26] Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 12402–12426. https://doi.org/10.18653/v1/2023.acl-long.693

[27] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 9. 48–59.

[28] Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3007–3014. https://doi.org/10.1145/3340531.3412876

[29] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 878–891. https://doi.org/10.18653/v1/2022.acl-long.62

[30] T.N. Flynn and A.A.J. Marley. 2014. Best-worst scaling: theory and methods. In *Handbook of Choice Modelling*, Stephane Hess and Andrew Daly (Eds.). Edward Elgar Publishing, Chapter 8, 178–201. https://ideas.repec.org/h/elg/eechap/14820_8.html

[31] Organisation for Economic Co-operation and Development (OECD). 2018. Bridging the digital gender divide: Include, upskill, innovate. *OECD* (2018).

[32] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (Jun. 2018). https://doi.org/10.1609/icwsm.v12i1.14991

[33] Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=vfT4YuzAYA

[34] Rishav Hada, Amir Ebrahimi Fard, Sarah Shugars, Federico Bianchi, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. 2023. Beyond Digital "Echo Chambers": The Role of Viewpoint Diversity in Political Discussion. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (<conf-loc>, <city>Singapore</city>, <country>Singapore</country>, </conf-loc>) *(WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 33–41. https://doi.org/10.1145/3539597.3570487

[35] Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. 2023. "Fifty Shades of Bias": Normative Ratings of Gender Bias in GPT Generated English Text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1862–1876. https://doi.org/10.18653/v1/2023.emnlp-main.115

[36] Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of Offensiveness for English Reddit Comments. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2700–2717. https://doi.org/10.18653/v1/2021.acl-long.210

[37] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[38] Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. 2023. IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian Languages. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23)*. AAAI Press, Article 1452, 9 pages. https://doi.org/10.1609/aaai.v37i11.26521

[39] Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities. In *Proceedings of the 16th International Conference on Natural Language Processing*. NLP Association of India, International Institute of Information Technology, Hyderabad, India, 211–219. https://aclanthology.org/2019.icon-1.25

[40] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[41] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics,

Online, 4948–4961. https://doi.org/10.18653/v1/2020.findings-emnlp.445

[42] Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing Isn't Enough! – on the Effectiveness of Debiasing MLMs and Their Social Biases in Downstream Tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1299–1310. https://aclanthology.org/2022.coling-1.111

[43] Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender Bias in Masked Language Models for Multiple Languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 2740–2750. https://doi.org/10.18653/v1/2022.naacl-main.197

[44] Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 465–470. https://doi.org/10.18653/v1/P17-2074

[45] Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). Association for Computational Linguistics, San Diego, California, 811–817. https://doi.org/10.18653/v1/N16-1095

[46] Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. arXiv:2102.04130 [cs.CL]

[47] Neeraja Kirtane and Tanvi Anand. 2022. Mitigating Gender Stereotypes in Hindi and Marathi. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (Eds.). Association for Computational Linguistics, Seattle, Washington, 145–150. https://doi.org/10.18653/v1/2022.gebnlp-1.16

[48] Andrey Kormilitzin, Nenad Tomasev, Kevin R. McKee, and Dan W. Joyce. 2023. A participatory initiative to include LGBT+ voices in AI for mental health. *Nature Medicine* 29 (2023), 10–11. https://api.semanticscholar.org/CorpusID:255748280

[49] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Mingxuan Wang and Imed Zitouni (Eds.). Association for Computational Linguistics, Singapore, 343–352. https://doi.org/10.18653/v1/2023.emnlp-industry.33

[50] Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable Modular Debiasing of Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 4782–4797. https://doi.org/10.18653/v1/2021.findings-emnlp.411

[51] Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 2470–2480. https://doi.org/10.18653/v1/2021.findings-emnlp.211

[52] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3125–3135. https://doi.org/10.18653/v1/P19-1301

[53] Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the Threshold: Idiomatic Machine Translation through Retrieval Augmentation and Loss Weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 15095–15111. https://doi.org/10.18653/v1/2023.emnlp-main.933

[54] J. J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. Working Paper.

[55] Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, and Alexandra Olteanu. 2023. "One-size-fits-all"? Observations and Expectations of NLG Systems Across Identity-Related Language Features. arXiv:2310.15398 [cs.CL]

[56] Meryem M'hamdi, Xiang Ren, and Jonathan May. 2023. Cross-lingual Continual Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3908–3943. https://doi.org/10.18653/v1/2023.acl-long.217

[57] Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 17 (apr 2021), 14 pages. https://doi.org/10.1145/3449091

[58] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling Online Abuse: A Survey of Automated Abuse Detection Methods. *CoRR* abs/1908.06024 (2019). arXiv:1908.06024 http://arxiv.org/abs/1908.06024

[59] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics* 10 (2022), 92–110. https://doi.org/10.1162/tacl_a_00449

[60] 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

[61] 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)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[62] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[63] Hadas Orgad and Yonatan Belinkov. 2022. Choose Your Lenses: Flaws in Gender Bias Evaluation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (Eds.). Association for Computational Linguistics, Seattle, Washington, 151–167. https://doi.org/10.18653/v1/2022.gebnlp-1.17

[64] B. Orme. 2009. Maxdiff analysis: Simple counting,individual-level logit, and HB. Sawtooth Software, Inc.

[65] Jiaxin Pei and David Jurgens. 2020. Quantifying Intimacy in Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 5307–5326. https://doi.org/10.18653/v1/2020.emnlp-main.428

[66] Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5877–5891. https://doi.org/10.18653/v1/2023.acl-long.323

[67] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural Incongruencies in Artificial Intelligence. arXiv:2211.13069 [cs.CY]

[68] Pratham. 2022. Annual Status of Education Report 2022. https://asercentre.org/aser-2022/ Accessed on 09/13/2023.

[69] Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2020. Debiasing Gender biased Hindi Words with Word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence* (Sanya, China) *(ACAI '19)*. Association for Computing Machinery, New York, NY, USA, 450–456. https://doi.org/10.1145/3377713.3377792

[70] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1882–1895. https://doi.org/10.1145/3593013.3594134

[71] Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating Gender Bias in Hindi-English Machine Translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, Marta Costa-jussa, Hila Gonen, Christian Hardmeier, and Kellie Webster (Eds.). Association for Computational Linguistics, Online, 16–23. https://doi.org/10.18653/v1/2021.gebnlp-1.3

[72] Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in Language Models Beyond English: Gaps and Challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2106–2119. https://doi.org/10.18653/v1/2023.findings-eacl.157

[73] Shaina Raza, Muskan Garg, Deepak John Reji, Syed Raza Bashir, and Chen Ding. 2024. Nbias: A natural language processing framework for BIAS identification in text. *Expert Syst. Appl.* 237, Part B (2024), 121542. https://doi.org/10.1016/J.ESWA.2023.121542

[74] 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)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 8–14. https://doi.org/10.18653/v1/N18-2002

[75] Shahin Sharifi Noorian, Sihang Qiu, Burcu Sayin, Agathe Balayn, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2023. Perspective: leveraging human understanding for identifying and characterizing image atypicality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 650–663.

[76] Yifan Song, Weimin Xiong, Dawei Zhu, Wenhao Wu, Han Qian, Mingbo Song, Hailiang Huang, Cheng Li, Ke Wang, Rong Yao, Ye Tian, and Sujian Li. 2023. RestGPT: Connecting Large Language Models with Real-World RESTful APIs. *arXiv preprint arXiv: 2306.06624* (2023).

[77] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 1166–1177. https://doi.org/10.18653/v1/2021.findings-emnlp.101

[78] Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *arXiv preprint arXiv: 2112.14168* (2021).

[79] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1679–1684. https://doi.org/10.18653/v1/P19-1164

[80] Statista. 2022. Use of social media platforms among people in India as of January 2022, by locality. https://www.statista.com/statistics/1388563/india-social-media-platform-usage-by-locality/ Accessed: 2024-01-02.

[81] Statista. 2023. Regional distribution of desktop traffic to Reddit.com as of April 2023 by country. https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/ Accessed: 2024-01-02.

[82] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 667–678. https://doi.org/10.1145/3531146.3533132

[83] Stojan Trajanovski, Chad Atalla, Kunho Kim, Vipul Agarwal, Milad Shokouhi, and Chris Quirk. 2021. When does text prediction benefit from additional context? An exploration of contextual signals for chat and email messages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, Young-bum Kim, Yunyao Li, and Owen Rambow (Eds.). Association for Computational Linguistics, Online, 1–9. https://doi.org/10.18653/v1/2021.naacl-industry.1

[84] Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 2203–2213. https://doi.org/10.18653/v1/2021.eacl-main.188

[85] Ajit Varghese. 2022. Celebrating Bharat's digital journey@75: The rapid increase in language-first users on social media. https://timesofindia.indiatimes.com/blogs/voices/celebrating-bharats-digital-journey75-the-rapid-increase-in-language-first-users-on-social-media/ Accessed: 2024-01-02.

[86] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2017. Respeak: A Voice-Based, Crowd-Powered Speech Transcription System. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1855–1866. https://doi.org/10.1145/3025453.3025640

[87] Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: A Game for Collecting Common-Sense Facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 75–78. https://doi.org/10.1145/1124772.1124784

[88] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou (Eds.). Association for Computational Linguistics, San Diego, California, 88–93. https://doi.org/10.18653/v1/N16-2013

[89] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. 2020. *Measuring and Reducing Gendered Correlations in Pre-trained Models.* Technical Report. https://arxiv.org/abs/2010.06032

[90] Ge Zhang, Yizhi Li, Yaoyao Wu, Linyuan Zhang, Chenghua Lin, Jiayi Geng, Shi Wang, and Jie Fu. 2023. CORGI-PM: A Chinese Corpus For Gender Bias Probing and Mitigation. arXiv:2301.00395 [cs.CL]

[91] Mike Zhang and Antonio Toral. 2019. The Effect of Translationese in Machine Translation Test Sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor (Eds.). Association for Computational Linguistics, Florence, Italy, 73–81. https://doi.org/10.18653/v1/W19-5208

[92] 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)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 15–20. https://doi.org/10.18653/v1/N18-2003

[93] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, Online, 3143–3155. https://doi.org/10.18653/v1/2021.eacl-main.274

A APPENDIX

epochs	*batch_size*	*learning_rate*	*optimizer*	*patience*
4	128	0.00001	AdamW	3

Table 2. Hyperparameters for training CORGI classifier

epochs	*lr*	*scheduler*	*optimizer*	*warmup_steps*	*batch_size*	*lora_r*	*lora_alpha*	*lora_dropout*	*lora_modules*
30	0.001	linear	AdamW	40	16	16	32	0.05	["key", "value"]

Table 3. Hyperparameters for training the FSB scorer

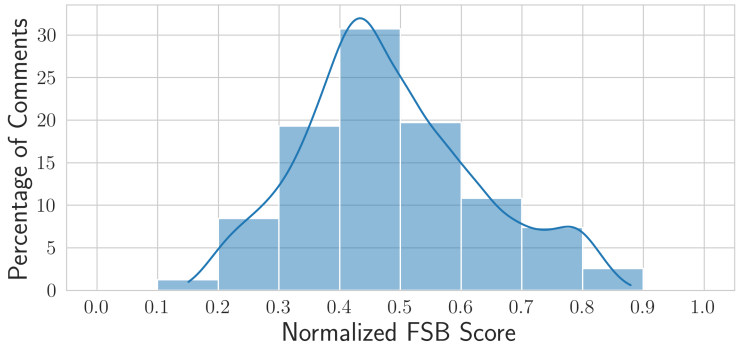

Fig. 2. A histogram of the percentage of comments–degree of gender bias. The degree of gender bias scores are grouped in bins of size 0.1.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009