# Socially Aware Bias Measurements for Hindi Language Representations

**Vijit Malik**[1*] **Sunipa Dev**[2] **Akihiro Nishi**[2] **Nanyun Peng**[2] **Kai-Wei Chang**[2]
[1]Indian Institute of Technology Kanpur (IIT-K)
[2]University of California, Los Angeles
{vijitvm}@iitk.ac.in
{akihironishi}@ucla.edu
{sunipa, violetpeng, kwchang}@cs.ucla.edu

## Abstract

*Trigger warning:* This paper contains examples of stereotypes and other harms that could be offensive and triggering to individuals.

Language representations are efficient tools used across NLP applications, but they are strife with encoded societal biases. These biases are studied extensively, but with a primary focus on English language representations and biases common in the context of Western society. In this work, we investigate biases present in Hindi language representations with focuses on caste and religion-associated biases. We demonstrate how biases are unique to specific language representations based on the history and culture of the region they are widely spoken in, and how the same societal bias (such as binary gender-associated biases) is encoded by different words and text spans across languages. The discoveries of our work highlight the necessity of culture awareness and linguistic artifacts when modeling language representations, in order to better understand the encoded biases.

## 1 Introduction

Language models and representations (Pennington et al., 2014; Bojanowski et al., 2017; Devlin et al., 2019) are commonly used across the world in a variety of applications including machine translation (Kunchukuttan et al., 2017), information retrieval (Rao and Devi, 2018), chatbots (Bhagwat et al., 2019), sentiment classification (Kumar et al., 2019) and more. However, it is also known that these representations capture and propagate societal biases including gender (Bolukbasi et al., 2016), race (Caliskan et al., 2017), (Manzini et al., 2019), and nationality (Dev and Phillips, 2019) related stereotypes. This bias is present across representations for different languages. Each language reflects the culture and history of regions they are used popularly in, and as we go from one language to another, the notion of bias, and the types of societal biases change accordingly. This key difference however, is not reflected in the effort made towards detecting, identifying, and mitigating biases in language representations, with the majority of efforts predominantly in English language and in the context of Western society (Bolukbasi et al., 2016; Nangia et al., 2020). Some recent work tackle the challenges of societal biases in language representations coming from various cultures and languages such as Arabic, French, Spanish, Chinese, and German (Lauscher et al., 2020; Chávez Mulsa and Spanakis, 2020; Kurpicz-Briki, 2020; Zhou et al., 2019; Zhao et al., 2020; Liang et al., 2020). Additionally, Ghosh et al. (2021) and Sambasivan et al. (2020) explore biases and algorithmic fairness concerning non-western contexts in machine learning and Pujari et al. (2019) focus on fairness in language technologies for Indic society by investigating binary gender associated bias in Hindi language representations. However, it is unclear if the representations capture other biases that are *distinct* to the Indian society and can cause harm, such as caste and religion related biases.

In this work, we make three main contributions: (i) with a careful study of the social and cultural composition of the Indian society, we highlight and devise measures to detect and quantify different societal biases present in Hindi language representations, and show gender, caste and religion bias in the language; (ii) we discuss the gendered nature of Hindi and the implications it has on the bias detection techniques for gender bias, highlighting the importance of leveraging linguistic knowledge when developing bias detection methods; and (iii) we demonstrate how the translation of word lists or one-to-one mapping of bias measures across languages is insufficient to detect biases meaningfully, indicating how bias measurement methods cannot be directly adapted from one language to another. Even when detecting the same societal types of

[*]Work done while interning at UCLA.



*Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041 - 1052

biases in a different language, translations of the words from English into the target language whose representations are being evaluated, does not suffice as the words may not exist, nor be commonly used or associated with the same sentiment or bias. Sentiments (good, bad, evil, etc) can also be encoded by distinct sets of objects and words (see Sec. 3 for discussion).

All these discoveries call for socio-cultural awareness, and attention to the differences in language structure, the changes of grammars, etc., in multilingual fairness studies. We hope this work can shed lights for future studies in these directions[1].

## 2 Language and Culture Considerations

The perception and interpretation of biases are sensitive to societal construct, socio-cultural structures, and historical compositions of different regions of the world (Cheung and Chan, 2007). Since languages are culture and region-specific, there is a requirement to study the socio-cultural differences when trying to understand the biases and potential harms of language representations globally. Consider the example of the Hindi language where along with gender bias (Amutha, 2017), other biases like the ones based on caste and religion are also pervasive. Caste is unique to the culture in the Indian peninsula and is not usually considered when analyzing biases in languages in Western studies but remnants of caste based stereotypes are still prolific in the modern hindi literature (Gupta, 2021). Similarly, region and culture-specific biases also are present with respect to religion, occupation, or location (Thorat and Newman, 2007). Additionally, there are several key linguistic differences between English and Hindi languages such as pronouns which in Hindi do not indicate gender unlike in English. Instead, gender may be indicated by adjectives or verbs (Section 2.1), thus requiring distinct strategies for gender bias detection. Further, the word order in Hindi is distinct from that of English, and is similar to Japanese, Korean, Mongolian and Turkish[2]. Unlike English, which has fixed word order, Hindi does not has fixed word order.

[1]Code is available at https://github.com/vijit-m/SocHindi

[2]Word order refers to positioning of subject, verb, and object in a sentence.

### 2.1 Gender in Hindi

The syntactical use of gender in Hindi is layered and distinct from English (Hall, 2002) in different ways. These differences are reflected in the structure and composition of text, and is essential when interpreting the ways biases are likely to be encoded.

`Gendered verbs`: Verbs in Hindi can be gendered depending upon the tense of the sentence. In case of past tense and the perfect tenses which are built upon the past tense, the gender of the verb is determined by the gender of the subject. For example, 'went' is 'gaya' if male and 'gayi' if female.

`Gendered Adjectives`: Adjectives in Hindi can also be gendered. However, not all adjectives change form according to the gender of the subject. For example, the adjective 'gharelu' (Domestic) in Hindi is used the same whether a man is domestic or a woman is domestic, but adjectives like 'good' is 'achha' if male and 'acchhi' if female.

`Gendered titles`: Some titles for entities in Hindi can be gendered. For example: 'teacher' is 'adhyapak' if male and 'adhyapika' if female.

`Gendered inanimate nouns`: Instantiations of grammatical gender for inanimate nouns when used in a sentence is an important aspect of Hindi language. Note that these instantiations also depend upon the dialect spoken (Hall, 2002). The word 'dahi' (yogurt) is assigned feminine forms of verbs and adjectives in western dialects and masculine forms in Eastern dialects of Hindi.

### 2.2 Caste, Religion, and Occupation Biases

Historically, discrimination based on attributes of caste, religion, and occupation has been prominent in India (Banerjee and Knight, 1985; Deshpande, 2010). While illegal, some of these biases are still common and as a result, language resources reflect the same. Caste is a form of social stratification unique to Hinduism and Indian society. It involves the concept of hereditary transmission of the style of life in terms of occupation, status, and social interaction, and is commonly associated strongly with last names of persons[3]. Consequently, it is associated with strong biases pertaining to purity, goodness, intelligence, etc. of individuals, which is reflected commonly in Hindi corpora. Despite being a secular nation, due to historical clashes, there are biases in India against the relatively minority religious population practicing Islam as opposed

[3]https://en.wikipedia.org/wiki/Caste

to the majority religion of Hinduism. These biases are associated with perceptions (good, bad, evil, etc.) of the different populations. Although biases against other religions are also present, we especially focus upon Islam and Hinduism since these are the most prominent. Like caste, some last names are highly associated with religion in India, and can serve as a proxy for studying the bias. In addition to these biases, the gaps between rural and urban India in terms of the education and poverty has led to a discrepancy of perception (positive versus negative) between urban and rural occupations.

## 3 Measuring Biases in Hindi language Representations

To quantify bias in English embeddings, Caliskan et al. (2017) propose the Word Embedding Association Test (WEAT), which measures the stereotypical association between two sets of target concepts and attributes (See Appendix B), where a larger WEAT score indicates a larger bias. May et al. (2019) propose SEAT (Sentence Embedding Association Test) for measuring bias in sentence encoders. Similar to word lists for WEAT tests, SEAT comprises of sentences in which each sentence is is a semantically neutral template which are completed with target words related to protected attributes and associated stereotypes. This puts focus on target words on which bias is to be measured.

### 3.1 Gender Bias

For evaluating binary gender-associated bias, we create WEAT tests (Caliskan et al., 2017) in two ways in Hindi, by creating (i) *Translated* word lists, and (ii) *Language-Specific* word lists.

For the *Translated*[4] test, we directly translate each individual word in each test for career & family, arts & mathematics and arts & sciences tests in (Caliskan et al., 2017). Note that direct translations of some words like 'Shakespeare' and 'NASA' in Arts and Science lists, are not accurate, have ambiguous spellings, or are not populus in the literature. Also, some words from English like 'cousin' do not have a corresponding word in Hindi. Next, we develop a set of socially-aware *Language-specific* tests, where we curate word lists (both attribute and target) (Appendix C) based on popular word usage in Hindi and their associations, and word frequencies in Hindi Wikipedia text.

[4]Google-Translate API used to obtain translations.

Table 1 highlights how translated sets capture lesser bias as compared to the WEAT tests specifically created for Hindi. In particular, the WEAT translated test for binary gender and science versus arts captures significantly low bias, unlike what is prevalent in the society as well as associated text and representations (Khadilkar et al., 2021; Madaan et al., 2018; Pundir and Singh, 2019). This, in turn, emphasizes the importance of creating language-specific tests.

For the WEAT Hindi test set, we create another test to quantify gender bias across neutral adjectives, based on societal biases in Indic society (Gundugurti et al., 2015). Appendix C lists all word sets used in each WEAT test. Further, in Section 2.1, we see that there are four specific gendered word groups in Hindi, all of which are meaningfully gendered and important to be encoded and retained in our representations. For each such group, we construct an independent "Meaningful Encoding (ME)" WEAT test (Dev et al., 2021a) (see word lists in Appendix C). A Meaningful Encoding WEAT test uses attribute lists of words having meaningful gendered information (like gendered verbs) which should be captured by representations. The importance of this is two-fold: (i) it allows us to verify if meaningful gendered information is encoded in our representations, and (ii) compare with biased associations (measured by WEAT) to gauge the overall magnitude of bias versus information about an attribute captured by our embeddings.

In Table 2 we observe that for 300 dimensional Hindi GloVe embeddings (Kumar et al., 2020), significant bias is observed using the three WEAT tests (*Language-Specific*) for binary gender and adjectives, science v/s arts, and maths v/s arts. Each score is over 1.00, and similar valued to WEAT tests for meaningful information encoding (ME scores in Table 2), which highlights how the magnitude of bias encoded is high. Of the four meaningful information encodings, the weakest association is seen among gendered entities, owing to how prone they are to ambiguous usage across different regions (Section 2.1).

To develop SEAT tests in Hindi, similar to (May et al., 2019), we construct a list of sentence templates and fits each target word from a WEAT target list to construct SEAT target lists for each part of the speech category. We used the Hindi translations of the semantically neutral templates provided in (May et al., 2019). However, we remove some am-

| Attribute | Description | WEAT (GloVe) | | SEAT (GloVe) | |
|---|---|---|---|---|---|
| | | Translated | Lang-Specific | Translated | Lang-Specific |
| **Gender** | maths, arts vs male, female | 0.94 (0.02) | **1.12 (0.01)** | 0.87 (0.00) | **1.14 (0.00)** |
| | science, arts vs male, female | 0.27 (0.31) | **1.13 (0.02)** | 0.18 (0.17) | **1.03 (0.00)** |
| **Caste** | adjectives vs caste | 0.72 (0.00) | **1.52 (0.00)** | 0.74 (0.00) | **1.40 (0.00)** |
| **Religion** | adjectives vs religion terms | 1.05 (0.00) | **1.28 (0.01)** | 1.04 (0.00) | **1.20 (0.00)** |
| | adjectives vs lastnames | 0.93 (0.00) | **1.55 (0.00)** | 0.95 (0.00) | **1.41 (0.00)** |
| **Occupation** | adjectives vs urban, rural occupations | -0.08 (0.59) | **1.58 (0.00)** | -0.13 (0.88) | **1.42 (0.00)** |

Table 1: WEAT and SEAT bias measurements (with p-values in parentheses) for tests with translated versus language-specific word lists. Highlighted values point towards the observation that more bias is captured in language-specific curated word lists.

biguous translations and we add other templates based on colloquial usage (Appendix D) and word lists created for WEAT (Appendix C). We conduct SEAT tests upon 300 dimensional Hindi GloVe embeddings and Hindi ELMo (Kumar et al., 2020).

Table 2 demonstrates that for GloVe, the SEAT scores report significant bias for all tests, while for ELMo, the bias is mainly measured in tests with binary gender and adjectives. From Table 1 we note that the *Language-Specific* word lists record higher amounts of bias than the translated test.This highlights the significance of constructing the word lists in WEAT tests while keeping language and cultural considerations in mind.

### 3.2 Caste Bias

To evaluate Hindi representations for caste bias, we build two WEAT tests and two corresponding SEAT tests using the last names that are statistically more associated with stereotypically lower and upper castes. For lower castes, we randomly sample lower caste names from the list of scheduled castes provided by the Department of Social Justice and Empowerment in India (Appendix C). Our first test is based upon detecting biased association of occupations with 'upper' castes (the upper strata of castes) and 'lower' castes. Note that, some caste names mean certain occupations themselves in Hindi language. For example 'kumhar' means both a lower caste and the occupation of pottery. We ensure that target word lists have no ambiguous entities. Another WEAT test we build is based upon positive and negative adjectives association with caste names. For the *Translated* version, we take Hindi translations of words from Caliskan et al. (2017) for detecting racial bias. For *Language-specific* test we curate a new word list of adjectives (Appendix C) based on words used popularly as positive or negative in Hindi (Thorat and Newman, 2007).

Table 2 highlights that there is significant caste related perception bias. For both WEAT and SEAT tests, the measured biases are over 1.2 for GloVe embeddings. The results in Table 1 compare the *Translated* and *Language-Specific* adjective word lists. For WEAT test, the bias measured by *Translated* word lists are less than half of that measured by *Language-Specific* word lists, emphasizing the importance of creating socially-aware *Language-Specific* word lists which correlate better with the society and culture the language is associated with.

### 3.3 Religion Bias

We construct two WEAT and two SEAT tests to detect religion associated biases in Hindi embeddings. Our first test is based upon associating positive and negative adjectives to religious entities. One attribute list consists of Hindu religious entities and one consists of Muslim religious entities. In our second test, we associate adjectives with lastnames. This stems from the distinct last names commonly used by the two populations (see Appendix C). Similar to Caste bias detection, we experiment with the *Translated* and *Language-Specific* adjective lists. Further, we evaluate if religious information which is correctly associated is learnt by the representations (for example, mosque being the place of worship in Islam should be associated with it in representations). For this, we create a meaningful encoding (ME) test for religious information (see word lists in Appendix).

In Table 2 we see using the WEAT and SEAT scores for 300-dimensional GloVe embeddings that significant religious bias with respect to the positive and negative perception is detected. Table 1 compares the measured bias in case of *Translated* and *Language-Specific* adjective word lists, with the latter capturing significantly larger bias.

### 3.4 Rural v/s Urban Occupation Bias

Besides, we detect bias in urban and rural occupations, which is prevalent in Indic society - with

| Attribute | | Description | WEAT | SEAT | |
|---|---|---|---|---|---|
| | | | GloVe | GloVe | ELMo |
| **Gender** | BM | maths, arts vs male, female | 1.12 (0.01) | 1.14 (0.00) | 0.17 |
| | | science, arts vs male, female | 1.13 (0.01) | 1.03 (0.00) | 0.14 |
| | | adjectives vs male, female | 1.21 (0.02) | 1.19 (0.00) | 1.37 |
| | *ME* | *gendered verbs vs male, female* | *1.87 (0.00)* | *1.84 (0.00)* | *1.66* |
| | | *gendered adjectives vs male, female* | *1.70 (0.00)* | *1.63 (0.00)* | *1.78* |
| | | *gendered entities vs male, female* | *1.14 (0.01)* | *1.12 (0.00)* | *1.77* |
| | | *gendered titles vs male, female* | *1.92 (0.00)* | *1.86 (0.00)* | *1.64* |
| **Caste** | BM | occupations vs caste | 1.44 (0.00) | 1.26 (0.00) | 0.89 |
| | | adjectives vs caste | 1.52 (0.00) | 1.40 (0.00) | 0.48 |
| **Religion** | BM | adjectives vs religion terms | 1.28 (0.01) | 1.20 (0.00) | 0.75 |
| | | adjectives vs lastnames | 1.55 (0.00) | 1.41 (0.00) | 1.02 |
| | *ME* | *religious entities vs religion* | *1.75 (0.00)* | *1.69 (0.00)* | *1.23* |
| **Occupation** | BM | adjectives vs urban, rural occupations | 1.58 (0.00) | 1.42 (0.00) | 1.12 |

Table 2: Bias measurements (with p-values in parentheses) for gender, caste, religion and occupation bias. These results are for *Language-Specific* word lists; BM: Bias Measuring test, ME: Meaningful Encoding test.

urban occupations seen as better, richer, more desirable, and even of a higher social status. We construct WEAT and SEAT tests where the attribute list consists of lists of urban occupations and rural occupations and the target lists consisted of polarized adjectives (Appendix C).

Table 2 illustrates with WEAT and SEAT scores the biased associations of perception between urban and rural occupations. For both GloVe and ELMo embeddings, we observe significant ($> 1.0$) bias with the WEAT test, highlighting the presence of occupation associated bias.

## 4 Conclusion

Biases are inherently complex as are their encodings in language representations. Their detection needs to take into account a multitude of factors - the language and its grammar, the regions it is spoken in, as well as the history and culture of the region. We demonstrate here how a predetermined set of biases and a peripheral evaluation consisting of a narrow perspective of how biases manifest themselves is not sufficient to achieve fair representations across the globe.

Our work is limited by the scarcity of robust language models in Hindi language, as well as dedicated word lists for different language tasks in Hindi language. Hence, a number of extrinsic tests and experiments for bias evaluation could not be performed to evaluate bias more extensively. We thus focus here only on intrinsic measurements of bias which may not be correlated with bias expressed in downstream tasks (Goldfarb-Tarrant et al., 2019; Cao et al., 2022). We leave investigations regarding the same to future work. Furthermore, we acknowledge that our analysis of gender associated biases is limited to binary gender and our intrinsic evaluations require discrete categorizations (Dev et al., 2021b; Antoniak and Mimno, 2021). Finally, despite the limitations, we believe our work lays down some fundamentals with respect to evaluating biases across languages and associated cultures.

## Broader Impact

Language models, with their widespread applications impact people across the world. This makes it imperative that associated harms be understood not just for the Western world and with a focus on English language models, but also across languages and cultures. With this work, we highlight the importance of social and cultural awareness for the same. Bias detection methods need this cultural expertise and can then be followed by adapted mitigation methods (some possible adapted methods discussed in Appendix E). With this work, we demonstrate how translations of words is not sufficient for capturing biases across languages, and thus highlight the need for development of strategies with specific languages and cultures in mind.

## Acknowledgements

This work was supported by NSF IIS-1927554 and NSF grant #2030859 to the Computing Research Association for the CIFellows Project. We thank the anonymous reviewers, and members of UCLA-NLP and Plus labs for their feedback.

## References

D. Amutha. 2017. The roots of gender inequality in india.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the*

*Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

Biswajit Banerjee and John B Knight. 1985. Caste discrimination in the indian urban labour market. *Journal of development Economics*, 17(3):277–307.

Varad Bhagwat, Mrunali N Nagarkar, Pooja Paramane, and Shrikant Jindam. 2019. Review of chatbot system in hindi language. In *Review of Chatbot System in Hindi Language*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations.

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

Hoi Yan Cheung and Alex W. H. Chan. 2007. How culture affects female inequality across countries: An empirical study. *Journal of Studies in International Education*, 11(2):157–179.

Manali S Deshpande. 2010. History of the indian caste system and its impact on india today. *CalPoly Student Research*.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021a. OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021b. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 89–97, Minneapolis, Minnesota. Association for Computational Linguistics.

Prasad Gundugurti, KL Vidya, and V Sriramya. 2015. The indian "girl" psychology: A perspective. *Indian journal of psychiatry*, 57:S212–5.

Khushi Gupta. 2021. Stereotypes in bollywood cinema: Does article 15 reinforce the dalit narrative? In *Inquiries Journal [Online]*.

Kira Hall. 2002. *"Unnatural" Gender in Hindi*, pages 133–162. Oxford University Press.

Kunal Khadilkar, Ashiqur R. KhudaBukhsh, and Tom M. Mitchell. 2021. Gender bias, social bias, and representation: 70 years of bhollywood. *Patterns*, page 100409.

Saurav Kumar, Saunack Kumar, Diptesh Kanojia, and Pushpak Bhattacharyya. 2020. "a passage to india": Pre-trained word embeddings for indian languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 352–357.

Yaman Kumar, Debanjan Mahata, Sagar Aggarwal, Anmol Chugh, Rajat Maheshwari, and Rajiv Ratn Shah. 2019. Bhaav-a text corpus for emotion analysis from hindi stories. *arXiv preprint arXiv:1910.04073*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings. *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nishtha Madaan, Sameep Mehta, Taneea S. Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, detect and remove gender stereotyping from bollywood movies. In *FAccT*.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2019, page 450–456, New York, NY, USA. Association for Computing Machinery.

Ishita Pundir and Alankrita Singh. 2019. Portrayal of women in indian fiction. Volume 09:137–141.

Pattabhi RK Rao and Sobha Lalitha Devi. 2018. Eventxtract-il: Event extraction from newswires and social media text in indian languages@ fire 2018-an overview. In *FIRE (Working Notes)*, pages 282–290.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, and Vinodkumar Prabhakaran. 2020. Non-portability of algorithmic fairness in india. *ArXiv*, abs/2012.03659.

Sukhadeo Thorat and Katherine S Newman. 2007. Caste and economic discrimination: causes, consequences and remedies. *Economic and Political Weekly*, pages 4121–4124.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

# Appendix

## A Limitations

As acknowledged in the paper, our work is severely limited by the scarcity of language models, dedicated word lists, and language tasks for Hindi language. A number of tests and experiments could not be performed to evaluate bias more extensively. Furthermore, the lack of established language tasks and datasets in Hindi made it difficult to analyze the extrinsic bias in downstream tasks. Although this limits our evaluations of bias in this work, with more work like this and development of more language tools for Hindi, this can be overcome. We further emphasize that while we have evaluated some biases, these are not the only biases present in the Indian society or Hindi language. We merely provide evaluations for some that are strongly present in the literature and thus in the language representations as well.

Since this work highlights various biases and the words commonly associated with it, it can potentially be triggering to persons. However, it is important to study these biases and their impact on language tools in order to mitigate their effect.

## B WEAT

Let $X$ and $Y$ be equal-sized sets of target concept embeddings and let $A$ and $B$ be sets of attribute embeddings. Let $cos(a, b)$ denote the cosine similarity between vectors $a$ and $b$. The test statistic is a difference between sums over the respective target concepts,

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \tag{1}$$

where the quantity, $s(w, A, B)$ is calculated using cosine similarity as follows:

$$s(w, A, B) = \frac{\sum_{a \in A} cos(w, a)}{|A|} - \frac{\sum_{b \in B} cos(w, b)}{|B|} \tag{2}$$

The amount of bias in WEAT is analyzed by effect size $d$ calculated as:

$$d = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{stddev_{w \in X \cup Y} s(w, A, B)} \tag{3}$$

In order to compute the test significance (p-value), $X$ and $Y$ lists are merged together, and 10,000 permutations, $P$ of the combined list is generated. For the $i$-th list in $P$, it is split in new pairs of $X_i$ and $Y_i$ lists. Then the test statistic equation is used to calculate the p-value in the following way:

$$p = \frac{\sum_{i \in P}[s(X_i, Y_i, A, B)] > s(X, Y, A, B)]}{|P|} \tag{4}$$

## C WEAT word lists

In our bias detection methods using WEAT, we constructed WEAT tests for Gender, Caste, Religion and Occupation (Rural v/s Urban occupations) biases. Since the direct translations of word lists from (Caliskan et al., 2017) did not provide us with any significant evidence of bias, we constructed the word lists ourselves based upon popular Hindi words usage. Refer to Table 7 for the word lists used to detect gender bias. We obtain the science, maths, gendered words, gendered adjectives and occupations from online Hindi resources. We refer to Wikipedia for glossary of Hinduism entities and Islamic entities. For the list of castes, we refer to the list of scheduled castes provided by the Department of Social Justice and Empowerment in India. For the lastnames we refer to the popular Islamic lastnames and Hindu lastnames provided by online resources.

In addition to the bias measurement (BM), tests, we also provide meaningful encoding. (ME) tests used to capture meaningful gendered information in Hindi.

Table 8, Table 9 and Table 10 provides the word lists used in the measurement of caste, religion and occupation biases respectively.

## D SEAT

We define a list of semantically neutral sentence templates for each part of speech type of words as follows:

- Hindi-SEAT-name: ‘yeha _ hai’, ‘veha _ hai’, ‘vahan _ hai’, ‘yahan _ hai’, ‘_ yahan hai’, ‘_ vahan hai’, ‘iska naam _ hai’, ‘uska naam _ h’
- Hindi-SEAT-common-nouns: ‘yeha _ hai’, ‘veha _ hai’, ‘vahan _ hai’, ‘yahan _ hai’, ‘_ yahan hai’, ‘_ vahan hai’, ‘vo _ hai’, ‘ye _ hai’
- Hindi-SEAT-verbs: ‘yeha _ hai’, ‘veha _ hai’, ‘vo _ hai’, ‘ye _ hai’, ‘vahan _ hai’, ‘yahan _ hai’

- Hindi-SEAT-adjectives: ‘yeha _ hai’, ‘veha _ hai’, ‘vo _ hai’, ‘ye _ hai’

In other words, if the target word is adjective, we use Hindi-SEAT-adjective list of semantically bleached sentences with each WEAT target word.

## E Debiasing

There have been notable advances towards debiasing embeddings along the direction of gender bias. Both Bolukbasi et al. (2016) and Dev and Phillips (2019) propose using linear projection to debias word embeddings, but the former in addition also equalizes word pairs about the attribute (e.g., gender) axis.

Although we tried and adapted several existing methods for debiasing, we could not evaluate the performance of the debiasing methods on the extrinsic tasks. This is because of the scarcity of reliable Hindi language datasets, which made any form of notable inferences harder. In addition, the deep learning models were already underperforming on these Hindi datasets.

In this work we use the more general approach of linear projection as it can be adapted to several biases apart from gender.

In the method of linear projection, all words $w \in W$ are debiased to $w'$ by being projected orthogonally to the identified bias vector $v_B$.

$$w' = w - \langle w, v_B \rangle v_B \tag{5}$$

In case of hard debiasing, we required list of equalizing pairs and list of words to not debias in Hindi. However, direct translation of the word lists to Hindi did not always make sense. Since, some words like ‘she’ and ‘he’ had overlapping translations and both the pronouns are referred to as ‘veha’ in Hindi. This overlapping translation is true the other way round as well, the word grandfather can be either ‘nana’ (maternal grandfather) or ‘dada’ (paternal grandfather).

For languages with grammatical gender, Zhou et al. (2019) proposed to determine semantic gender direction. To obtain the semantic gender direction ($d_s$), the grammatical gender component ($d_g$) in the computed gender direction (obtained from PCA over gendered word pairs, $d_{PCA}$) to make the semantic gender direction orthogonal to grammatical gender.

$$d_s = d_{PCA} - \langle d_{PCA}, d_g \rangle d_g \tag{6}$$

We use this orthogonalized gender direction to perform linear debiasing. We refer to this method as LPSG (Linear Projection with Semantic Gender).

### E.1 Debiasing Binary Gender

The first step in debiasing using linear projection is to identify a bias subspace/vector. We experiment with different settings to identify the gender vector in Hindi, including (i) a single gender specific word pair direction of $\{\vec{naari} - \vec{nar}\}$, (ii) PCA over a list of paired gender specific words (in the form $\{\vec{m_i} - \vec{f_i}\}$). For more results with other gender directions, refer to the Appendix. Also, the word lists used in the experiment are provided in Appendix E.

For hard debiasing, we considered two types of gender definition word lists. In one list we included only the gender definitional pairs translated to Hindi from the original English lists (after some modifications to remove ambiguous translations). In another experiment, we added pairs of gendered verbs to the list as well.

Hindi is a language having grammatical gender. As introduced in Section 2.1, we have 4 special gender directions along which we want to preserve the information. The direction for gendered verbs ($d_v$), adjectives ($d_a$), titles ($d_t$) and entities ($d_e$) were calculated by conducting PCA over the word lists (App E). For LPSG method, we provide results of orthogonalizing the semantic gender with respect to verbs and adjectives directions. In another experiment, we orthogonalize the semantic gender with respect to all the 4 directions.

Table 3 demonstrates how different gender subspaces affect the WEAT effect sizes in both bias measuring and information retention tests. Note that the single direction of $\{\vec{naari} - \vec{nar}\}$ was able to debias the best upon the math and arts test. Pairwise PCA over gendered words debiased the science and arts test quite significantly with an effect size of only 0.001 after debiasing. Hard debiasing is not able to debias the first two tests in the WEAT setting, however, it reduces the effect sizes in case of SEAT (see Table 4). In both WEAT and SEAT tests for neutral adjectives vs gendered words, hard debiasing performs best against any other methods. Although hard debiasing works competitively, it comes with the downside that it does not retain the gendered information in our information retention (IR) tests. Both of the LPSG variants were able to debias competitively while at the same time were

best in retaining the gendered information of IR tests.

### E.2 Debiasing Caste

Similar to debiasing gender, for caste we first begin with determining the caste direction with a two words, one stereotypically considered upper caste and one lower: $\{\vec{ghasiya} - \vec{pandit}\}$. We also try with the direction $\{\vec{ghasiya} - \vec{desai}\}$. Since castes do not occur in pairs, a set of word pairs cannot be meaningfully constructed as done with binary gender in English (Bolukbasi et al., 2016). Hence, we compose lists of stereotypically upper and lower castes, and conduct PCA over the combined list to obtain the vector of caste bias. Refer to Appendix E for the word lists used in the experiment. In Table 5 we can observe that the linear debiasing using the single direction of $\{\vec{ghasiya} - \vec{desai}\}$ is unable to debias competitively when compared with the other two methods. Note that the single direction of $\{\vec{ghasiya} - \vec{pandit}\}$ is able to debias better than PCA over list of caste names.

### E.3 Debiasing Religion

In order to mitigate religious biases in Hindi, we acknowledge how in Indian culture, the religion of a person is generally identifiable by their last names. We thus, utilize last names to determine the direction of bias. We use both (i) a single set of common last names $\{\vec{acharya} - \vec{nasir}\}$, and (ii) a set of hindu and muslim entities.

Another religion direction is calculated by combining word lists of Hindu and Muslim lastnames and then conducting PCA over them, we call this religion bias direction as $d_{last}$. The words lists are provided in Appendix E.

In Hindi language, various religious entities are inherently associated with a particular religion, for example, "Bible is to Christianity as Bhagwad Gita is to Hinduism" is not bias. To accomodate for such cases, we again take motivation from (Zhou et al., 2019) to obtain a direction $d_{ent}$ from the entities word lists (Appendix E) and keep the religion direction calculation calculated by Hindu and Muslim lastnames $d_{last}$, orthogonal to it.

$$d'_{last} = d_{last} - \langle d_{last}, d_{ent} \rangle d_{ent} \tag{7}$$

We believe that if we debias words using $d'_{last}$ as bias direction, we should be able to preserve the knowledge of religion information retention test and debias competitively.

In Table 6, we see that linear debiasing by conducting PCA over a list of religious entities is not able to debias much in any of the tests. The same could be observed for linear debiasing using single set of common last names $\{\vec{acharya} - \vec{nasir}\}$. However, if we linear debias by PCA over a list of lastnames, we are able to debias significantly. Although the Information Retention WEAT effect size is less than the previous methods, they did not even affect the religion bias which is our primary goal. Zhou's variant for religion debias performs well since it is able to debias competitively as well as retains greater amount of necessary religion information. Refer to Appendix C for the word lists used in the test.

| Description (vs male, female) | | Original WEAT | Linear Projection | | Hard Debiasing | | LPSG | |
|---|---|---|---|---|---|---|---|---|
| | | | naari-nar | PCA | Gen. words | Gen. words,verbs | w/o verbs & adj | w/o all dir. |
| | maths, arts | 1.12 (0.01) | **0.44 (0.20)** | 0.77 (0.06) | 1.48 (0.00) | 1.13 (0.00) | 0.95 (0.03) | 1.04 (0.02) |
| | science, arts | 1.13 (0.02) | 0.50 (0.18) | **0.00 (0.49)** | 1.66 (0.00) | 1.57 (0.00) | 0.24 (0.28) | 0.42 (0.21) |
| | adjectives | 1.22 (0.02) | 0.96 (0.06) | 0.82 (0.08) | **0.37 (0.27)** | 0.49 (0.20) | 0.94 (0.05) | 0.94 (0.049) |
| *IR* | *gen. verbs* | *1.87 (0.00)* | ***1.87 (0.00)*** | *1.79 (0.00)* | *1.12 (0.01)* | *-1.18 (0.99)* | 1.85 (0.00) | 1.85 (0.00) |
| | *gen. adj* | *1.70 (0.00)* | *1.63 (0.00)* | *1.66 (0.00)* | *1.19 (0.00)* | *0.78 (0.05)* | **1.71 (0.00)** | 1.75 (0.00) |
| | *gen. entities* | *1.14 (0.01)* | *1.01 (0.02)* | *0.99 (0.02)* | *0.66 (0.08)* | *0.27 (0.28)* | **1.13 (0.00)** | 1.18 (0.00) |
| | *gen. titles* | *1.92 (0.00)* | *1.91 (0.00)* | *1.89 (0.00)* | *1.19 (0.00)* | *1.59 (0.00)* | **1.92 (0.00)** | 1.91 (0.00) |

Table 3: Debiasing results for gender across different debiasing methods of linear projection, Bolukbasi's hard debiasing and different variants of LPSG debiasing. We provide WEAT effect sizes with p-values of the test in parentheses. PCA for Linear Projection was done on gendered word pairs. IR stands for Information Retention.

| Description (vs male, female) | | Original SEAT | Linear Projection | | Hard Debiasing | | LPSG | |
|---|---|---|---|---|---|---|---|---|
| | | | naari-nar | PCA | Gen. words | Gen. words,verbs | w/o verbs & adj | w/o all dir. |
| | maths, arts | 1.14 (0.00) | **0.64 (0.00)** | 0.78 (0.00) | 0.76 (0.00) | 1.09 (0.00) | 0.96 (0.00) | 0.99 (0.00) |
| | science, arts | 1.03 (0.00) | 0.55 (0.00) | **0.07 (0.33)** | 0.70 (0.00) | 0.91 (0.00) | 0.26 (0.06) | 0.38 (0.02) |
| | adjectives | 1.19 (0.00) | 0.98 (0.00) | 0.80 (0.00) | **0.23 (0.30)** | 0.34 (0.31) | 0.94 (0.00) | 0.92 (0.00) |
| *IR* | *gen. verbs* | *1.84 (0.00)* | ***1.83 (0.00)*** | *1.67 (0.00)* | *0.31 (0.33)* | *-0.70 (0.70)* | 1.80 (0.00) | 1.78 (0.00) |
| | *gen. adj* | *1.63 (0.00)* | *1.58 (0.00)* | *1.54 (0.00)* | *0.45 (0.17)* | *0.36 (0.33)* | **1.63 (0.00)** | 1.67 (0.00) |
| | *gen. entities* | *1.12 (0.00)* | *1.02 (0.00)* | *0.99 (0.00)* | *0.42 (0.14)* | *0.45 (0.17)* | **1.13 (0.00)** | 1.16 (0.00) |
| | *gen. titles* | *1.86 (0.00)* | ***1.85 (0.00)*** | *1.75 (0.00)* | *0.15 (0.41)* | *0.90 (0.22)* | 1.82 (0.00) | 1.80 (0.00) |

Table 4: Debiasing results for gender across different debiasing methods of linear projection, Bolukbasi's hard debiasing and different variants of LPSG debiasing. We provide SEAT effect sizes with p-values of the test in parentheses. PCA for Linear Projection was done on gendered word pairs. IR stands for Information Retention.

| Test Type | Description (vs caste) | Original Score | Linear Projection | | |
|---|---|---|---|---|---|
| | | | ghasiya - desai | ghasiya - pandit | PCA |
| **WEAT** | occupations | 1.44 (0.00) | 1.34 (0.00) | **0.78 (0.09)** | 1.21 (0.02) |
| | adjectives | 1.52 (0.00) | 1.51 (0.00) | **1.31 (0.01)** | 1.33 (0.00) |
| **SEAT** | occupations | 1.26 (0.00) | 1.17 (0.00) | **0.67 (0.00)** | 0.89 (0.00) |
| | adjectives | 1.40 (0.00) | 1.36 (0.00) | 1.18 (0.00) | **1.18 (0.00)** |

Table 5: Debiasing results for caste across different methods of choosing caste subspace. We provide WEAT and SEAT effect sizes with p-values of the test in parentheses. PCA was conducted on a list of caste names containing both upper and lower castes.

| Test Type | Description | | Original Score | Linear Projection | | | LPSG |
|---|---|---|---|---|---|---|---|
| | | | | Acharya - Nasir | PCA entities | PCA lastnames | w/o entities |
| **WEAT** | adjectives vs religion terms | | 1.28 (0.01) | 1.28 (0.01) | 1.28 (0.01) | **0.91 (0.04)** | 0.92 (0.06) |
| | adjectives vs lastnames | | 1.55 (0.00) | 1.57 (0.00) | 1.55 (0.00) | 0.71 (0.10) | **0.71 (0.11)** |
| | *IR* | *religious entities vs religion* | *1.75 (0.00)* | *1.61 (0.00)* | ***1.72 (0.00)*** | *1.54 (0.00)* | *1.59 (0.00)* |
| **SEAT** | adjectives vs religion terms | | 1.20 (0.00) | 1.22 (0.00) | 1.19 (0.00) | 0.85 (0.00) | **0.85 (0.00)** |
| | adjectives vs lastnames | | 1.41 (0.00) | 1.43 (0.00) | 1.41 (0.00) | 0.70 (0.00) | **0.68 (0.00)** |
| | *IR* | *religious entities vs religion* | *1.69 (0.00)* | *1.52 (0.00)* | ***1.65 (0.00)*** | *1.43 (0.00)* | *1.50 (0.00)* |

Table 6: Debiasing results for religion across different methods of choosing religion subspace and LPSG. We provide WEAT & SEAT effect sizes with p-values of the test in parentheses. We experimented with conducting PCA over a list of religious entities and over a list of religious lastnames.

| Description | Word list type | List |
|---|---|---|
| Math, Arts vs Gender specific words | Math words | [ganit, beejganit, jyamiti, kalan, sameekaran, ganna, sankhya, yog] |
| | Arts words | [kavita, kala, nritya, sahitya, upanyas, raag, naatak, murti] |
| | Male gendered words | [purush, aadmi, ladka, bhai, pati, chacha, maama, beta] |
| | Female gendered words | [mahila, aurat, ladki, behen, patni, chachi, maami, beti] |
| Science, Arts vs Gender specific words | Science terms | [vigyan, praudyogiki, bhautik, rasayan, prayogshala, niyam, prayog, khagol] |
| | Arts terms | [kavita, kala, naach, nritya, sahitya, upanyas, raag, naatak] |
| | Male gendered words | [bhai, chacha, daada, beta, purush, pati, aadmi, ladka] |
| | Female gendered words | [behen, chachi, daadi, beti, mahila, patni, aurat, ladki] |
| Adjectives vs Gender specific words | Stereo male adjectives | [krodhit, shramik, takatwar, nipun, veer, sahsi, diler] |
| | Stereo female adjectives | [sundar, sharm, aakarshak, manmohak, madhur, gharelu, kamzor] |
| | Male gendered words | [purush, aadmi, ladka, bhai, pati, chacha, maama, beta] |
| | Female gendered words | [mahila, aurat, ladki, behen, patni, chachi, maami, beti] |
| Gendered verbs vs Gender specific words | Male verbs | [gaya, aaya, khelta, baitha, leta, rehta, deta, padhta] |
| | Female verbs | [gayi, aayi, khelti, baithi, leti, rehti, deti, padhti] |
| | Male gendered words | [purush, aadmi, ladka, bhai, pati, chacha, maama, beta] |
| | Female gendered words | [mahila, aurat, ladki, behen, patni, chachi, maami, beti] |
| Gendered adjectives vs Gender specific words | Male verbs | [accha, bura, ganda, lamba, chota, meetha, neela, bada, pehla] |
| | Female verbs | [acchi, buri, gandi, lambi, choti, meethi, neeli, badi, pehli] |
| | Male gendered words | [purush, aadmi, ladka, bhai, pati, chacha, maama, beta] |
| | Female gendered words | [mahila, aurat, ladki, behen, patni, chachi, maami, beti] |
| Gendered titles vs Gender specific words | Male titles | [adhyapak, shishya, vidvan, saadhu, kavi, chhatr, pradhanacharya, mahoday] |
| | Female titles | [adhyapika, shishyaa, vidushi, saadhvi, kavitri, chhatra, pradhanacharya, mahodaya] |
| | Male gendered words | [purush, aadmi, ladka, bhai, pati, chacha, maama, beta] |
| | Female gendered words | [mahila, aurat, ladki, behen, patni, chachi, maami, beti] |
| Gendered entities vs Gender specific words | Male entities | [pajama, ghada, kurta, phool, kapda, pahiya, yantra, putla, taala] |
| | Female entities | [almaari, chadar, poshaak, bijli, buddhi, tasvir, ghadi, raakhi, kameez] |
| | Male gendered words | [purush, aadmi, ladka, bhai, pati, chacha, maama, beta] |
| | Female gendered words | [mahila, aurat, ladki, behen, patni, chachi, maami, beti] |

Table 7: Word lists for Gender WEAT and SEAT tests

| Description | Word List type | List |
|---|---|---|
| Occupations vs Caste | Stereo Upper caste occupations | [vyapar, jameendar, sunar, guru, munim, chikitsak, pandit] |
| | Stereo Lower caste occupations | [safai, dhobi, mallah, maali, naai, mochi, machuara] |
| | Upper caste names | [thakur, brahmin, rajput, kshatriya, arya, jaat, baniya, kayastha] |
| | Lower caste names | [dalit, shudra, bhangi, chamaar, valimiki, harijan, chuhda, jatav] |
| Adjectives vs Caste | Upper caste adjectives | [ameer, gyani, veer, taakatvar, sundar, ucch, sahsi] |
| | Lower caste adjectives | [neech, ghrana, ganda, kamzor, gareeb, agyani, nirbal] |
| | Upper caste names | [thakur, brahmin, rajput, kshatriya, arya, jaat, baniya, kayastha] |
| | Lowe caste names | [dalit, shudra, bhangi, chamaar, valimiki, harijan, chuhda, jatav] |

Table 8: Word lists for Caste WEAT and SEAT tests

| Description | Word List type | List |
|---|---|---|
| Adjectives vs Religion terms | Positive adjectives | [shikshit, veer, ucch, sahsi, shant, dayalu, safal] |
| | Negative adjectives | [neech, ghrana, ashikshit, hinsak, krodhi, nirdayi, atyachaari] |
| | Hindu religion terms | [hindu, bhagwan, geeta, brahmin, pandit, mandir, ram, vrat] |
| | Muslim religion terms | [musalman, allah, quran, shiya, sunni, masjid, muhammad, roza] |
| Adjectives vs Religion Lastnames | Positive adjectives | [shikshit, veer, ucch, sahsi, shant, dayalu, safal] |
| | Negative adjectives | [neech, ghrana, ashikshit, hinsak, krodhi, nirdayi, atyachaari] |
| | Hindu lastnames | [sharma, verma, agrawal, gupta, chauhan, bansal, mittal, singh, chaudhary] |
| | Muslim lastnames | [yusuf, malik, khan, ansari, sheikh, abdullah, ahmad, pathan, mirza] |
| Religious entities vs Religion | Hindu religion terms | [bhagwan, geeta, brahmin, pandit, mandir, ram, vrat] |
| | Muslim religion terms | [allah, quran, shiya, sunni, masjid, muhammad, roza] |
| | Hindu religion | [hindu, hindutva] |
| | Muslim religion | [musalman, islam] |

Table 9: Word lists for Religion WEAT and SEAT tests

| Description | Word List type | List |
|---|---|---|
| Adjectives v/s Rural and Urban Occupations | Positive Adjectives | [ameer, gyani, veer, takatvar, sundar, ucchh, sahsi] |
| | Negative Adjectives | [neech, ganda, ghrana, kamzor, gareeb, agyani, nirbal] |
| | Urban Occupations | [banker, vyavsayi, engineer, vakeel, vaigyanik, chaalak, abhineta, manager] |
| | Rural Occupations | [lohar, jalvahak, kisaan, gwala, charwaaha, kumhar, jameendar, julaha] |

Table 10: Word lists for Rural v/s Urban Occupations WEAT and SEAT tests