Debiasing Multilingual Word Embeddings: A Case Study of Three Indian Languages

Srijan Bansal†
Dept. of ECE,
IIT Kharagpur
West Bengal, India – 721302
srijanbansal97@iitkgp.ac.in

Vishal Garimella†
Dept. of CSE,
IIT Kharagpur
West Bengal, India – 721302
vishal_g@iitkgp.ac.in

Ayush Suhane
Dept. of Mathematics,
IIT Kharagpur
West Bengal, India – 721302
ayushsuhane99@iitkgp.ac.in

Animesh Mukherjee
Dept. of CSE,
IIT Kharagpur
West Bengal, India – 721302
animeshm@cse.iitkgp.ac.in

ABSTRACT

In this paper, we advance the current state-of-the-art method for debiasing monolingual word embeddings so as to generalize well in a multilingual setting. We consider different methods to quantify bias and different debiasing approaches for monolingual as well as multilingual settings. We demonstrate the significance of our bias-mitigation approach on downstream NLP applications. Our proposed methods establish the state-of-the-art performance for debiasing multilingual embeddings for three Indian languages - Hindi, Bengali, and Telugu in addition to English. We believe that our work will open up new opportunities in building unbiased downstream NLP applications that are inherently dependent on the quality of the word embeddings used.

CCS CONCEPTS
• Computing methodologies → Machine translation.

KEYWORDS
Debiasing multilingual embeddings, Gender debias, Debiasing Indian languages

ACM Reference Format:
Srijan Bansal, Vishal Garimella, Ayush Suhane, and Animesh Mukherjee. 2021. Debiasing Multilingual Word Embeddings: A Case Study of Three Indian Languages. In Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21), August 30-September 2, 2021, Virtual Event, Ireland. ACM, 8 pages. https://doi.org/10.1145/3465336.3475118

1 INTRODUCTION

Word embeddings are ubiquitous across many downstream NLP applications. Current approaches in NLP rely on large amounts of training data. Such data is easily available for a resource-rich language like English but poses a major challenge for other languages. Multilingual word embeddings are widely used in numerous downstream NLP applications which represent words from multiple languages in a common vector space such that similar meaning words are close to each other. This allows to leverage advancement in English for improving model’s performance on low resource languages using transfer learning [1, 2, 6, 13].

However, there is an unavoidable bottleneck in the aforementioned pipeline. Lately, word embeddings have been found to be ‘blatantly sexist’ thus introducing a bias in the applications built on top of them [4]. Using Google-News embeddings for the English language the authors showed how gender-neutral profession words (e.g., programmer, homemaker) get aligned to one of the gender directions for various analogy tasks. To mitigate this problem they introduced different flavours of debiasing algorithms that can quite successfully remove the biases in the word embeddings. Later Dev and Phillips [9] showed that using simple linear projections can be more effective in attenuating bias in word vectors than complex debiasing algorithms.

An open question: While multiple efforts have focused on understanding and mitigating the bias in English word embeddings, less attention is given to understand and mitigate the bias in multilingual embeddings. A pertinent question is how well the above debiasing algorithms work for other non-English languages in a multilingual setting. This question is worth investigating for the following reasons.

• The semantics of gender words may vary from one language to another. Hassan and Alamgir [11] point out ‘তার’ can refer to both ‘he’ or ‘she’ in the sentence সুন্দর তার ছায়া ছিল গেছে। (Somebody forgot his/her umbrella).

• While Bolukbasi et al. [4] leverages the pronouns (e.g., she/he) to construct gendered directions this might not be possible for many languages (e.g., Bengali where the same pronoun is used to refer to both the male and female genders). Prates et al. [15] point out that a relatively high proportion of words in Bengali are gender-neutral.

The present work: In this paper, we aim to understand bias in multilingual embeddings. Gender bias in English can be very different from bias in other languages due to the innate structure and semantics of the languages. In addition to quantifying this bias, we advance the linear projection based bias mitigation approach to multilingual embeddings.

Our key contributions are as follows.

• We build sets of gendered words for three different Indian languages chosen from two different language families (Indo-Aryan and Dravidian) – Hindi, Bengali and Telugu. We also identify a set of gender-neutral words for these languages which include both profession words (alike Bolukbasi et al. [4]).

---

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HT ’21, August 30-September 2, 2021, Virtual Event, Ireland
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8551-0/21/08...
https://doi.org/10.1145/3465336.3475118
Figure 1: Examples of words in our corpus.

| Profession         | Adjective          | Transliterated | Gender defining pairs            |
|--------------------|--------------------|----------------|----------------------------------|
| en                 | professor, boss, poet | rich, powerful, bright | father-mother, king-queen, he-she |
| hi                 | मालूम(1), पत्रकार(2), स्न्यायाधीश(3) | कठिन(4), स्वार्थ(5), भारी(6) | लड़की-लड़के(10), आदित्य-महिला(11), पति-पत्नी(12) |
| be                 | मालूम(13), संबंधिक(14), विचारक(15) | (मानत(16), मनुष्य(17), कम(18) | पूर्व-कल्याण(22), (पति-पत्नी(23), पुत्र-पुत्री(24) |
| te                 | పిల్లత(25), పిల్లత(26) | క్రింది(28), డాక్టర్(29), రోయల్(30) | శ్రీమత్త(31), శ్రీమత్త(32), ఇందులక్ష్యం(33) |

1) gardener 2) reporter 3) judge 4) tough 5) clean 6) heavy 7) conductor 8) nurse 9) professor 10) boy-girl 11) man-woman 12) husband-wife 13) gardener 14) journalist 15) judge 16) spicy 17) spicy 18) low 19) conductor 20) nurse 21) professor 22) male-female 23) son-daughter 24) son-daughter 25) judge 26) lawyer 27) administrator 28) powerful 29) peaceful 30) careful 31) commissioner 32) consultant 33) editor 34) uncle-aunt 35) brother-sister 36) god-goddess

[4]) as well as a set of adjectives which are known to be gender-biased for many Indian languages (see Pande [14]).

- We perform all our debiasing analysis on fasttext embeddings for the monolingual setting. For multilingual settings, they are aligned to common space using a bilingual dictionary. Our choice of embedding is motivated by the need for mitigating bias in Indic languages as well as in English.

- We advance monolingual debiasing algorithms to multilingual settings and evaluate them based on intrinsic and extrinsic bias measures. Our debiasing approach reduces both intrinsic and extrinsic bias in multilingual setting achieving state-of-the-art performance without compromising the overall quality of the debiased embeddings much.

Our methods are generic and can be easily extended to any other language. The choice of the current set of languages is motivated by the knowledge of the authors in these languages. All our code is available at https://github.com/amsuhane/Debiasing-Multilingual-Word-Embeddings-A-Case-Study-of-Three-Indian-Languages.

The rest of the paper is structured as follows. Section 2 describes related works on these problems and provides context on why the problem is difficult and important to solve. Next, in sections 3 and 4, we respectively quantify bias and describe the datasets which are used to measure it. Sections 5 and 6 respectively discuss our novel multilingual debiasing algorithm and the experimental setup. We present the results in section 7 and conclude our work in section 8.

2 RELATED WORK

Bolukbasi et al. [4] provides motivation as to why debiasing word embeddings is a problem of interest and introduce concepts of gender spaces and debiasing with respect to them. Dev and Phillips [9] further advance this work by proposing simpler algorithms for debiasing based on linear projections. Caliskan et al. [5] also find the gender stereotypes in the English word embeddings based on the Word Embedding Association Test (WEAT) which cannot be adapted for other languages. Zhou et al. [21] reveal that bias exists in languages with grammatical gender and make an attempt to debias English-Spanish embeddings. Recently, Blodgett et al. [3] have raised concerns about Dev and Phillips [9] and Bolukbasi et al. [4] due to lack of downstream applications (categorized as “stereotyping”). In [8] the authors studied the advantage of word debiasing on natural language inference task. The authors showed that there is a reduction in the number of invalid inferences due to debiasing. In another recent work [17] the authors introduce the concept of iterative null space projection to reduce bias in neural representations. The last two works, which are also the most recent in this space, are again limited to the monolingual setting.

Our work is novel in various ways – we consider multiple Indic languages including samples from both the Indo-Aryan and Dravidian families, perform our analysis after the alignment of words from all the languages to a common space (unlike previous studies) and propose a single algorithm for debiasing all of them which allows for easy inclusion of many other languages. One of the earlier works on debiasing Indo-Aryan languages include [16]. Further, we address the concern of stereotyping by solving a downstream NLP task of occupation classification.

3 QUANTIFYING BIAS

In this section, we describe the notion of gender bias in word embeddings. Although in this study we consider only four languages namely English, Hindi, Bengali and Telugu the approach is general and can be easily extended to other languages. Bias in embeddings can be quantified in two major ways: intrinsic bias and extrinsic bias. While the former is used to quantify bias at the word-level the latter focuses on quantifying bias from the perspective of downstream NLP applications.
3.1 Intrinsic bias
Intrinsic bias (InBias) evaluation metric proposed by Zhao et al. [20] uses the gap between the distance of occupations and corresponding gender to show gender discrimination. Say, we have a pair of masculine and feminine words describing an occupation, such as the words actor and actress, the only difference lies in the gender information. As a result, they should have similar correlations to the corresponding gender seed words such as ‘he’ and ‘she’. If there is a gap, i.e., the distance between ‘actor’ and ‘he’ against the distance between ‘actress’ and ‘she’, it means such occupation shows discrimination against gender. We provide detailed descriptions of those words in section 4.

Definition of InBias: Given a set of masculine and feminine words, we define InBias as:

\[ \text{InBias} = \frac{1}{N} \sum_{i=1}^{N} |\text{dis}(O_M, S_M) - \text{dis}(O_F, S_F)| \]

where

\[ \text{dis}(X, Y) = \frac{1}{|Y|} \sum_{y \in Y} 1 - \cos(X, y) \]

Here (OM, OF) stand for the masculine and feminine versions of the ith occupation word, such as (actor and actress). This metric can also be generalized to languages without grammatical genders, such as English, by just using the same format of the occupation words. In our gender-neutral set (Nall), we only consider words which are same in male and female versions for all languages. SM and SF denote the set of male and female-oriented words respectively (i.e., SM contains [male, him, he, . . . ] and SF contains [female, her, she, . . . ]).

In other words, intrinsic bias is a simple geometric method to quantify the overall 'proximity' of gender-neutral words with respect to the notion of gender in a particular embedding space.

3.2 Extrinsic bias
The authors in Gonen and Goldberg [10] raise a valid concern that projection-based debiasing methods such as hard-debias and linear-projection reduce bias superficially. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly of hiding the bias, not removing it.

We address this concern by, quantifying bias from an extrinsic perspective (ExBias) based on a downstream task [20] in addition to the intrinsic metric. Extrinsic bias measures the gap in performance for a standard NLP task before and after the debiasing of the word embeddings.

Definition of ExBias: We follow the same definition of extrinsic bias evaluation as used in Zhao et al. [20], i.e., using the average performance gap between different gender groups (male and female) aggregated across all the occupations (Diff). We used the BiosBias dataset proposed by De-Arteaga et al. [7] to evaluate the bias in predicting the occupation of people using a short biography on the bio of the person written in the third person (see section 4 for more details). We split the dataset based on the gender attribute. A gender-agnostic model should have similar performance in each group. To make predictions of the occupations, we used bidirectional LSTM units followed by attention mechanism (similar to the model used in De-Arteaga et al. [7]). The predictions are generated by a softmax layer. We train such models using standard cross-entropy loss and keep the embeddings frozen during the training.

4 DATASET
In this section we describe the datasets used for measuring the InBias and ExBias.

4.1 InBias dataset
In order to evaluate intrinsic bias for embeddings, we require male-female counterparts of words or gender-neutral words and gender-oriented words1. We considered a set of gender-neutral occupational words and adjectives (Nall). For the three Indian languages, it also includes the relevant transliterated-English profession words taken from [4]. Nall consists of neutral words for each language namely Nen, Nhi, Nte, Nbn.

In addition to neutral words for each language, we considered a set of gender-defining pairs (Dall), i.e., gender specific words that are associated with a gender by definition (e.g., ([he, she], [him, her], [king, queen])). For Dall = {Dall1, Dall2, . . . , Dallm}, each Dalli represents a tuple of male-female words ([he, she], [king, queen] etc.) Dall = (δ−, δ+) where δ−, δ+ ∈ V. Dall constitutes of Den, Dhi, Dte and Dbn which are gender-defining pairs for each language.

| Lang | N_prof | N_adj | N_tr | N_lang | D_lang |
|------|--------|-------|------|--------|--------|
| en   | 59     | 50    | -    | 109    | 20     |
| hi   | 29     | 43    | 15   | 87     | 21     |
| te   | 18     | 54    | 18   | 90     | 15     |
| All  | 134    | 191   | 47   | 372    | 76     |

Table 1: Datasets statistics: |N_prof| - neutral profession words, |N_adj| - neutral adjectives, |N_tr| - neutral English transliterated words, |N_lang| - total number of neutral words for language lang, |D_lang| - total number of gender pairs for language lang.

Table 1 notes some basic statistics for the dataset used for further analysis. Our corpus is balanced with respect to different languages. Some examples of words in our dataset are noted in Figure 1. Henceforth, 'neutral' would refer to gender-neutral words.

4.2 ExBias dataset
For the investigation of ExBias we used the BiosBias dataset proposed by De-Arteaga et al. [7]. The dataset has been prepared to evaluate the bias in predicting the occupation of people using a short biography on the bio of the person written in third person. We followed the data collection procedure proposed by De-Arteaga et al. [7] for the English dataset2.

To identify bio paragraphs, we use regex patterns such as "NAME is an OCCUPATION-TITLE". In order to minimise bias due to the

1https://bit.ly/2TNKWs3
2For our work, we have removed all occurrences of author identity as well as co-references to the authors. We do not plan to release non-anonymized bio-data since it is already available for download from the commoncrawl portal.
5 DEBIASING ALGORITHMS

Bolukbasi et al. [4] propose a hard-debiasing (HD) method in which the vector difference of corresponding gender-defining pairs (e.g. (man, woman)) captures the gender direction in the embedding space. They use this gender direction to project the word embeddings of gender-neutral words into a subspace orthogonal to the gender-defining words. The HD method uses a seed set of gender-defining words to train a support vector machine classifier, and use it to expand the initial set of gender-defining words. During the subsequent debiasing process, HD focuses only on gender-neutral words (i.e., words not predicted as gender-defining by the classifier). Therefore, if the classifier erroneously predicts a stereotypical word as a gender-defining word, it would not get debiased.

Dev and Phillips [9] proposed linear projection algorithm similar to HD, which also uses a gender-direction to debias. In the rest of the paper, we shall denote the algorithm proposed by [9] as LP. Both create a vector-space $B \subset \mathbb{R}^d$ from the top PCA (Principal Component Analysis) component of gender-defining pairs and using this they remove the gender component of the word embeddings. Hard debias removes the component for gender-neutral words only and equalizes the distance of neutral words wrt each gender-defining pair. LP on the other hand achieves debiasing by removing component along $B$ for all $w \in W$, thus doing away with the need of training an SVM classifier to mine neutral words. This motivates us to use LP as our baseline.

Formally, for word vector $\vec{w} \in W$ we denote its projection onto the gender subspace $B$ as $\vec{w}_B = \sum_{i=1}^{k} (\vec{w}, \vec{b}_i)\vec{b}_i$ where $\vec{b}_i$ are basis vectors of $B$. Thus, a word vector $\vec{w}$ can be decomposed as $\vec{w} = \vec{w}_B + \vec{w}_\perp$, where $\vec{w}_\perp$ is the projection onto the orthogonal space [4]. Note that all the word vectors are normalized to unit normals.

LP removes $\vec{w}_B$ from all the word vectors $\vec{w} \in W$. Thus the component of neutral word along gender subspace is zero thereby, positioning the neutral words to almost equal distance from the gendered words. For a given word vector $\vec{w} \in \mathbb{R}^d$, the debiased embedding such that $\vec{w}' \in \mathbb{R}^d$ is $\vec{w}' = \vec{w} - \vec{w}_B$. If $\text{dim}(B) = k$ then resulting debiased embedding will have dimension $\text{dim}(W') = d - \text{dim}(B) = d - k$ ($k = 1$ if a single direction is used to construct $B$).

5.1 Basis vectors: PCA and PPA

Let $D$ be any set of gender-defining pairs such that $D \subseteq D_{\text{all}}$. We can represent $D$ as $D = \{D_1, D_2, \ldots, D_n\}$. For each $D_i = \{\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k\}$, difference vector can be defined as $\vec{\delta}_i = \vec{b}_2 - \vec{b}_1$. We can stack these difference vectors to form a matrix $Q = [\vec{\delta}_1, \vec{\delta}_2, \ldots, \vec{\delta}_k]^T$. Now, the gender subspace $B$ can be obtained from $Q$ using the span of the top-$k$ directions obtained from either principal component analysis (PCA) or principal polynomial analysis (PPA – Projection Pursuit Analysis) (see supplementary information). That is $B = \text{span}\{\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_k\}$ where $\vec{b}_i \in \mathbb{R}^d$ for integer parameter $k > 1$. We denote $B_{\text{lang}}$ as the space constructed using only the set $D_{\text{lang}}$.

Figure 2: Gender statistics of Bios-Bias dataset for different occupations where each occupation has at least 100 instances. X-axis here stands for the occupation names and y-axis is the number of instances for each occupation.
5.2 LP for multilingual embeddings

Using linear projection for debiasing multilingual word embeddings naturally raises some questions: LP takes the defining set \( D \) as an input to construct the gender-subspace \( B \) and gender neutral set \( N \) is needed by the evaluation metric to estimate the effect of debiasing. When there are many languages in a common space (\( N \subseteq N_{all} \)), we can have different choices of \( D = D_{en} \cup D_{hi} \cup \ldots \) etc.) for a given \( N \).

- Are the gender subspaces induced by gendered pairs for different languages \( l_1 \) and \( l_2 \) (say \( B_{l_1} \) and \( B_{l_2} \)) semantically different?
- What is the best choice for the gender-subspace \( B \) among \( B_{l_1} \) and \( B_{l_2} \) for debiasing both \( l_1 \) and \( l_2 \) words in common a space (which may be either aligned to \( l_1 \) or \( l_2 \) in the first place).
- Could there be a better alternative for constructing \( B \) as opposed to using either \( B_{l_1} \) and \( B_{l_2} \)? Will it be sufficient to take gender-defining pairs from all language (here both \( l_1 \) and \( l_2 \)) to form gender-subspace \( B \)?
- What if the gender subspace \( B \) so formed has an over-representation of vectors from one language over others.

We proceed to answer these questions in the following. Let \( B_{lang} = \text{span}\{\tilde{b}_{lang,1}, \tilde{b}_{lang,2}, \ldots, \tilde{b}_{lang,k}\} \), where \( \tilde{b}_{lang,i} \) is the \( i \)th most significant basis vector of the gender subspace \( B_{lang} \) (PCA/PPA components of matrix \( Q_{lang} \) obtained from \( D_{lang} \)). We hypothesize that for two languages \( l_1, l_2 \in \{en, hi, be, te\} \) s.t. \( l_1 \neq l_2 \), their corresponding gender subspaces \( B_{l_1} \) and \( B_{l_2} \) are inherently different from each other. These differences may arise due to different gender semantics of each language.

To further motivate our hypothesis, we use linear projection algorithm for debiasing neutral words \( N_{l_1} \subset N_{all} \) using gender direction \( \tilde{b}_{l_1,k} \) instead of \( \tilde{b}_{l_2,k} \) (\( B_{l_2} \), \( k = 1 \)). If \( B_{l_1} \) and \( B_{l_2} \) are semantically similar, then the gender direction of the choice as either \( \tilde{b}_{l_1,k} \) or \( \tilde{b}_{l_2,k} \) for debiasing \( N_{l_1} \) does not matter much since debiasing on one ensures minimal to no component along another.

We score the performance using the metric \( S_{l_1,l_2} \) defined as:

\[
S_{l_1,l_2} = \frac{1}{|N_{l_2}|} \sum_{w \in N_{l_2}} \frac{||((\tilde{w}, \tilde{b}_{l_2,1})) - ((\tilde{w}, \tilde{b}_{l_1,1}))||}{||((\tilde{w}, \tilde{b}_{l_1,1}))||}
\]

where \( \tilde{w} \) corresponds to the debiased embedding of word \( w \in N_{l_2} \) wrt \( \tilde{b}_{l_1,1} \) having original word embedding \( \tilde{w} \in \mathbb{R}^d \). This metric captures aggregate relative change in neutral-word similarity of a language wrt gender direction in that same language before and after debiasing. Note that \( (\tilde{w}, \tilde{b}_{l_1,1}) = 0 \) since embeddings are debiased wrt \( \tilde{b}_{l_1,1} \). Now, if \( B_{l_1} \) and \( B_{l_2} \) are semantically similar \( (\tilde{w}, \tilde{b}_{l_1,1}) = 0 \approx (\tilde{w}, \tilde{b}_{l_2,1}) \approx 0 \) thus \( S_{l_1,l_2} \rightarrow 1 \) (ideal case: \( S_{l_1,l_2} = 1 \)).

Table 2 shows the cross-language debiasing performance. The results in Table 2 clearly support our hypothesis that gender subspace \( B_{en} \), \( B_{hi} \), \( B_{be} \) and \( B_{te} \) are significantly different (since the off-diagonal values are far from unity) and the objective of debiasing the multilingual embedding cannot be accomplished using a single language’s gender subspace. For example, \( B_{en} \) is not a good choice for \( hi \) words even while the embeddings are in the common space.

| Lang  | \( N_{en} \) | \( N_{hi} \) | \( N_{be} \) | \( N_{te} \) |
|-------|-------------|-------------|-------------|-------------|
| \( b_{en} \) | 1.0         | 0.143       | 0.137       | 0.038       |
| \( b_{hi} \) | 0.105       | 1.0         | 0.083       | 0.023       |
| \( b_{be} \) | 0.345       | 0.126       | 1.0         | 0.075       |
| \( b_{te} \) | 0.049       | 0.054       | 0.157       | 1.0         |

Table 2: \( S_{l_1,l_2} \) for debiasing \( N_{l_1} \) neutral words using gender direction \( \tilde{b}_{l_1,1} \) where \( \{l_1, l_2\} \in \{en, hi, be, te\} \).

Choice of \( k \): In the traditional setting, LP uses top-1 gender direction for experiments limited to English. However, in a multi-lingual setting \( k = 1 \) would be a bad choice as we have just observed (the gender direction of one language may be ineffective/insufficient in capturing the gender direction of another language). Hence \( k > 1 \) is a natural choice.

Hence for extending LP to a multilingual setting, the gender space \( B_{all} \) should be constructed using gender-defining training pairs drawn from all languages \( D_{all} \). PCA/PPA can then used to obtain top-k basis-vectors \( \tilde{b}_{all,1}, \tilde{b}_{all,2}, \ldots, \tilde{b}_{all,k} \) of \( B_{all} \) such that \( B_{all} = \text{span}\{\tilde{b}_{all,1}, \tilde{b}_{all,2}, \ldots, \tilde{b}_{all,k}\} \).

In addition, we also use another approach which constructs bias subspace \( B_{equal\_rep} \) ensuring equal representation of each language in the basis vectors of \( B_{equal\_rep} \). For \( L \) languages under consideration and gender subspace having \( k \) basis vectors s.t. \( k \leq |D_{all}| \), we label the PCA/PPA components of \( B_{all} \) according to their language orientation (see supplementary material for details) and choose top \( \frac{k}{L} \) components for each language. Motivation for this choice stems from the fact that if all the basis vectors are oriented along a single \( lang \) in the derived \( B_{all} \), then the gender space spanned by this subspace will behave similar to \( B_{lang} \) defeating the purpose of multilingual gender representation.

Taking top-\( k \) PCA components of the gender directions will always provide a more robust estimate of gender than using a single gender direction. We investigate that by changing the size of seed-gender directions \( D \) used to construct \( Q \). The variance in InBias metrics is considerably small. We also investigate the same when we randomly sample any subset of \( D \) (say he/she/ man-woman) and use it as gender direction, the bias-scores indicate a very low variance. In the rest of the paper, using only \( B_{lang} \) for debiasing words of a single language \( lang \) is referred to as \( LP_{mono} \). Similarly, using \( B_{all} \) or \( B_{equal\_rep} \) for debiasing is referred to as \( LP_{multi} \) or \( LP_{EQG} \) respectively.

6 EXPERIMENTAL SETUP

In our experiments to gender neutralize multilingual word embeddings, we attempt to (i) try kurtosis based projection analysis (PPA) instead of variance based (PCA) and compare them for obtaining top-\( k \) components of the gender subspace (ii) compare performances of variants of linear projection algorithm for monolingual and multilingual embeddings for both intrinsic (InBias) and extrinsic (ExBias) bias metrics.
6.1 Train test split of gender-defining pairs
For construction of the gender subspace used by LP we use 10 gender pairs (train gender pairs) of each language, i.e., \( D_{\text{train}} \subset D_\text{lang} \) such that \( |D_{\text{train}}| = 10 \) for \( \text{lang} \in \{ \text{en, hi, be, te} \} \). Remaining gender pairs (test gender pairs) are used as set \( D_{\text{test}} \subset D_\text{lang} \) such that \( D_{\text{train}} \cap D_{\text{test}} = \emptyset \) for InBias computation. Note that this split is needed since our evaluation metric is based on cosine similarities with respect to gender directions, and evaluating the embeddings on the same directions as used for constructing gender spaces would typically lead to better values of metrics (train-test leakage). We resort to train-test split strategy for gender pairs because it is intuitive that if the gender subspace obtained using the training gender pairs captures the linguistic concept of gender well, we can expect to get reliable estimate of performance with test gender pairs.

6.2 Monolingual debiasing
Monolingual setup characterizes single language applications, where the focus is on debiasing a specific language. A straightforward way is to use linear projection (LP) to debias embedding using language-specific gender subspace \( B_{\text{lang}} \) (say, we expect hi word embeddings to be best debiased when \( B_{\text{hi}} \) is used). Thus, for monolingual settings we construct language gender subspace \( B_{\text{lang}} \) using gender pairs \( D_{\text{lang}} \) for a particular language \( \text{lang} \in \{ \text{en, hi, be, te} \} \).\(^3\) k basis vectors of \( B_{\text{lang}} \) thus obtained using PCA/PPA (\( k = 4 \)) are used to debias monolingual embedding and its performance is evaluated based on InBias and ExBias.

6.3 Multilingual debiasing
Multilingual setup characterizes two or more language applications, where the focus is on debiasing multiple languages at once. We get multilingual embeddings by aligning monolingual embedding in a common space. This is done using bilingual dictionaries between all source-target pairs of languages by Smith et al. [19]\(^4\)’s alignment algorithm (also used in fasttext multilingual pipeline) with same hyper-parameter settings.\(^5\) We obtain all these dictionaries from MUSE.\(^6\) Since for Telugu this is not available, we construct Telugu bilingual dictionary wrt other languages using the Google Translate API.

We use the multilingual extension of LP approach, i.e., \( LP_{\text{multi}} \) and \( LP_{\text{EQR}} \) (both \( k = 4 \)) for debiasing. Multilingual gender space \( B_{\text{all}} \) is constructed using training gender pairs drawn from all languages \( D_{\text{train}} \). PCA/PPA are used to obtain top-k basis vectors of \( B_{\text{all}} \) and its performance is evaluated using test gender pairs \( D_{\text{test}} \). Specifically, for bilingual embedding denoted by \( l_1 \rightarrow l_2 \) (i.e., \( l_1 \) is aligned in \( l_2 \) space), we use both gender-defining pair sets \( D_{l_1} \) and \( D_{l_2} \) to construct vector space capturing gender in \( l_1 \rightarrow l_2 \). This is in turn used by debiasing algorithms to remove the bias component for all words (\( l_1 \) and \( l_2 \)) in the vocabulary. We experiment with PCA and PPA methods to get basis vectors for \( LP_{\text{multi}} \) and \( LP_{\text{EQR}} \).

We also experimented language-specific (monolingual) LP approaches (denoted as \( LP_{\text{mono}} \) for language \( l_2 \)) on bilingual embedding \( l_1 \rightarrow l_2 \) to show multilingual extension of LP is always better than original LP approach. We compared the performance of original and debiased embeddings using metrics InBias and ExBias.

6.4 Additional evaluation based on NLP task
Further, as a quality evaluation of the resulting debiased embeddings we choose an NLP task as well. The usefulness/purpose of word embeddings lies in its utility for a downstream task. To evaluate the quality of word embeddings we considered two downstream tasks.

For monolingual setting, we test the original and resulting embeddings on a news classification task for all the four languages.\(^7\) We used a bidirectional LSTM encoder followed by a softmax layer. Performance is reported in terms of model accuracy. For multilingual setting, we used the CVIT-Mann ki baat dataset [18] for the cross-lingual sentence retrieval task. This corpus contains parallel English sentences with Indic languages, thus limiting us to English language pairs like (hi \( \rightarrow \) en, en \( \rightarrow \) be etc). We trained a sentence encoder (bidirectional LSTM) to cross-lingual sentence embedding and used cosine distance for retrieval as suggested in Libovický et al. [12]. Performance is reported in terms of precision@10. This confirms the possibility to use such embeddings in downstream tasks.

7 RESULTS
7.1 InBias results
We compare intrinsic bias measure for both original fasttext embedding (Orig) and \( LP_{\text{mono}} \) debiased embeddings (using both PCA and PPA, \( k = 4 \)). Table 3 summarizes InBias results for monolingual setting (focus only on words in \( N_{\text{lang}} \)), where the training gender subspace is constructed by \( D_{\text{train}} \) and performance is evaluated using \( D_{\text{test}} \). We observe that \( LP_{\text{mono}} \) (PCA) outperforms (lower is better) both \( LP_{\text{multi}} \) (PPA) and original embeddings for all the four languages considered.

| \( \text{lang} \) | \( \text{Orig} \) | \( LP_{\text{mono}} \) | \( LP_{\text{PPA}} \) |
|---|---|---|---|
| en | 0.014 | 0.0 | 0.001 |
| hi | 0.015 | 0.008 | 0.019 |
| be | 0.016 | 0.011 | 0.023 |
| te | 0.040 | 0.018 | 0.029 |

Table 3: InBias measure for monolingual setting.

Table 4 summarizes the InBias results for the monolingual setting (\( N_{\text{all}} \)), i.e., the gender subspace is constructed using \( D_{\text{all}} \) and performance is evaluated using \( D_{\text{test}} \). We compare the multilingual extension of LP (\( LP_{\text{multi}} \) and \( LP_{\text{EQR}} \)) with the \( LP_{\text{mono}} \) approach. \( LP_{\text{mono}} \) approach can be applied on a bilingual embedding \( l_1 \rightarrow l_2 \) by either debiasing it only using language \( l_1 \) or \( l_2 \) (denoted as \( LP_{l_1} \) and \( LP_{l_2} \) respectively in Table 4). The interesting observations that one can make from these results are – (i) \( LP_{\text{multi}} \) and \( LP_{\text{EQR}} \) always perform better than \( LP_{\text{mono}} \) based approaches (\( LP_{l_1} \) and

\(^3\)AG news classification (en), BBC news article classification (hi), Soham news article classification (be), INLTK headline classification (te).
$LP_l$) and the original embedding, thus supporting our intuition for multilingual extension of LP. (ii) $LP_{EQR}$ is at par or slightly better than $LP_{multi}$ for all the bilingual embeddings (except $hi \rightarrow be$), and (iii) PCA based methods always outperform PPA based methods and hence we only show the PCA based results Table 4 to save space (see supplementary material for the PPA based results).

| $l_1 \rightarrow l_2$ | Orig | $LP_l$ | $LP_{multi}$ | $LP_{EQR}$ |
|-----------------------|------|--------|--------------|------------|
| $en \rightarrow hi$   | 0.013| 0.015  | 0.010        | 0.007      |
| $en \rightarrow be$   | 0.013| 0.015  | 0.012        | 0.008      |
| $hi \rightarrow en$   | 0.015| 0.013  | 0.013        | 0.007      |
| $hi \rightarrow be$   | 0.016| 0.015  | 0.013        | 0.008      |
| $be \rightarrow en$   | 0.017| 0.015  | 0.012        | 0.008      |
| $be \rightarrow hi$   | 0.016| 0.014  | 0.013        | 0.008      |
| $te \rightarrow en$   | 0.040| 0.042  | 0.017        | 0.013      |
| $te \rightarrow hi$   | 0.024| 0.019  | 0.018        | 0.007      |
| $te \rightarrow be$   | 0.019| 0.023  | 0.022        | 0.007      |

Table 4: InBias measure for multilingual setting. Debiased algorithm uses PCA with value of $k = 4$.

### 7.2 ExBias results

As discussed earlier for measuring ExBias we make use of the BiosBias dataset, whereby, we classify the occupation from a gender-details-scrubbed-bio of a person. The premise for choosing this task is that the performance (accuracy) of any classification model on the BiosBias dataset must be similar across the male and female applicants, i.e., the accuracy of the classification model on the male applicant bios (denoted by $M$) must be equivalent to that of female applicant bios (denoted by $F$) as the bios are indistinguishable in all other ways. But if occupations themselves have inherent gender connotations, then an asymmetry in the performance is expected. Recall that the absolute difference of accuracy between male and female bios, averaged across all occupations quantifies ExBias (i.e., $|D\acute{e}f|\rangle$). The lesser the value of $|D\acute{e}f|\rangle$ the better it is. Also, for a more occupation-wise analysis, we use $f_i$ to denote the fraction of total occupations for which the performance gap ($|D\acute{e}f|\rangle$) decreases after debiasing.

| $l_1 \rightarrow l_2$ | Emb | M | $|D\acute{e}f|\rangle$ | $f_i$ | $P@10$ |
|-----------------------|-----|---|---------------------|-----|--------|
| $en \rightarrow hi$   | orig | 67.33 | 71.16 | 7.64 | 0.24 | 24.24 |
|                      | $LP_{multi}$ | 69.14 | 73.65 | 7.56 | 0.60 | 19.44 |
|                      | $LP_{EQR}$  | 69.74 | 73.54 | 7.28 | 0.60 | 21.95 |
| $en \rightarrow be$   | orig | 69.66 | 70.68 | 8.16 | 14.28 |
|                      | $LP_{multi}$ | 69.75 | 74.37 | 7.83 | 0.35 | 18.18 |
|                      | $LP_{EQR}$  | 69.74 | 73.33 | 7.65 | 0.65 | 20.59 |
| $hi \rightarrow en$   | orig | 67.30 | 70.55 | 7.58 | 21.21 |
|                      | $LP_{multi}$ | 70.31 | 74.20 | 8.06 | 0.35 | 22.22 |
|                      | $LP_{EQR}$  | 70.59 | 74.68 | 7.30 | 0.60 | 14.29 |
| $hi \rightarrow be$   | orig | 61.74 | 64.08 | 7.79 | 20.59 |
|                      | $LP_{multi}$ | 62.86 | 65.83 | 7.16 | 0.47 | 16.67 |
|                      | $LP_{EQR}$  | 62.81 | 65.49 | 7.22 | 0.41 | 15.63 |
| $be \rightarrow en$   | orig | 61.14 | 63.16 | 7.65 | 0.53 |
|                      | $LP_{multi}$ | 62.94 | 65.00 | 7.24 | 0.60 |
|                      | $LP_{EQR}$  | 63.15 | 65.07 | 7.19 | 0.35 |
| $be \rightarrow hi$   | orig | 61.22 | 63.49 | 7.31 | 0.60 |
|                      | $LP_{multi}$ | 63.19 | 65.14 | 7.13 | 0.60 |
|                      | $LP_{EQR}$  | 63.91 | 65.86 | 6.89 | 0.53 |
| $be \rightarrow te$   | orig | 59.64 | 61.03 | 7.73 | 21.87 |
|                      | $LP_{multi}$ | 61.87 | 63.58 | 6.96 | 0.60 | 12.53 |
|                      | $LP_{EQR}$  | 61.15 | 63.79 | 7.51 | 0.53 | 13.89 |
| $te \rightarrow en$   | orig | 59.42 | 61.22 | 6.91 | 0.53 |
|                      | $LP_{multi}$ | 61.34 | 63.79 | 7.01 | 0.35 |
|                      | $LP_{EQR}$  | 61.52 | 63.43 | 6.88 | 0.53 |
| $te \rightarrow hi$   | orig | 57.76 | 59.77 | 7.46 | 15.79 |
|                      | $LP_{multi}$ | 60.67 | 62.74 | 7.20 | 0.65 | 16.28 |
|                      | $LP_{EQR}$  | 61.38 | 62.46 | 8.08 | 0.47 | 33.33 |
| $te \rightarrow be$   | orig | 58.22 | 58.99 | 7.76 | 0.30 |
|                      | $LP_{multi}$ | 60.48 | 62.72 | 7.52 | 0.30 |
|                      | $LP_{EQR}$  | 61.15 | 63.10 | 6.98 | 0.35 |

Table 5: ExBias results in the monolingual settings.

Table 6: ExBias results in the multilingual setting. LP algorithm for PCA ($k = 4$) is used for this table.

We compare extrinsic bias (ExBias) measure for the original embeddings (orig) and $LP_{mono}$ debiased embeddings\(^6\). Table 5 summarizes the ExBias results for the monolingual setting. We observe that (i) $LP_{mono}$ always reduces extrinsic bias ($|D\acute{e}f|\rangle$) without affecting the embedding quality (i.e., the accuracy ($Acc_{task}$) of the news category classification task) much, with the only exception of Hindi. (ii) For all the languages, more than 35% of total professions saw a reduction in performance gap ($f_i$) rising up to almost 76% for Bengali after debiasing. (iii) We also observe that for all the languages, individual male and female accuracy ($M$ and $F$) increase after debiasing.

\(^6\)We always report results for the PCA approach since they outperform the PPA approach.
Availability of datasets like BiasBios. compromi
de of the embedding quality as is well established in the
s for three Indian languages in addition to English. We
depends on the target languages, as it is better to align to
glish language like Telugu and Hindi than gender-neutral lan
Gabeli and English and (v) The embedding quality (measured
P@10) remains roughly same or sometimes gets a boost due to
do biasing (P@10 scores only available for English–Indic
and (iii) For all the language pairs, around 35–60% profes
(f_i), using both $LP_{multi}$ and $LP_{EQR}$ methods, (iv) Bias
depends on the target languages, as it is.

In all other settings, $LP_{EQR}$ is the winner, thus establish-
ing the state-of-the-art performance (ii) For all the settings, male,
female accuracy (M and F) increases after debiasing, (iii) For all
the language pairs, around 35–60% professions have a reduction in per-
formance gap ($f_i$).

8 CONCLUSION

In this paper, we proposed different LP variants to debias word
embeddings for three Indian languages in addition to English. We
experiment in two different settings - monolingual and multilin-
gual; for each setting we measure two types of biases – intrinsic
and extrinsic. In both settings and for both types of evaluation
measures we observe that debiasing helps in removing the bias
from the embeddings. This comes mostly at the cost of a slight
compromise of the embedding quality as is well established in the
fairness literature.

In future, we would like to extend the framework to include
more Indian languages. We also plan to include other downstream
tasks which at this point however is difficult due to the limited
availability of datasets like BiasBios.

REFERENCES

[1] Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019. Cross-Lingu
gual Dependency Parsing with Unlabeled Auxiliary Lan
guages. In Proceedings of the 23rd Conference on Computa
tional Natural Language Learning (CoNLL). Association for Computational Linguistics, Hong Kong, China, 372–382. https://doi.org/10.18653/v1/K19-1035

[2] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah. A. Smith. 2016. Massively Multilingual Word Embeddings. ArXiv abs/1602.01925 (2016).

[3] Su Lin Blodgett, Solomon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Lang
uage (Technology) is Power: A Critical Survey of “Bias” in NLP. In Proceed
ings of the 58th Annual Meeting of the Association for Computa
tional Linguistics. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. In Proceedings of NeurIPS.

[5] Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356 (04 2017), 183–186.

[6] Xulun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-Source Cross-Lingual Model Transfer: Learning What to Share. In Proceedings of the 57th Annual Meeting of the Association for Computa
tional Linguistics. Association for Computational Linguistics, Florence, Italy, 3098–3112. https://doi.org/10.18653/v1/P19-1209

[7] Maria De Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representa
tion Bias in a High-Stakes Setting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FA’19). Association for Computing Machinery, New York, NY, USA, 120–128. https:

[8] Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srivatsa. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. Proceedings of the AAAI Conference on Artificial Intelligence 34, 05 (2020), 7659–7666.

[9] Sunipa Dev and Jeff M. Phillips. 2019. Attenuating Bias in Word Vectors. In Proceedings of ADTATS.

[10] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. 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). Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. https://doi.org/10.18653/v1/N19-1061

[11] Khandoker Hassan and Mohammad Alamgir. 2013. A Comparative Study of Gender Sensitivity between English and Bengali. Language in India 13 (11 2013), 200–208.

[12] Jindrich Libovicky, Rudolf Rosa, and Alexander M. Fraser. 2019. How Language-Neutral is Multilingual BERET? ArXiv abs/1911.03310 (2019).

[13] Tao Meng, Nanyun Peng, and Kai-Wei Chang. 2019. Target Language-Aware Constrained Inference for Cross-lingual Dependency Parsing. 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). Association for Computational Linguistics, Hong Kong, China, 1117–1128. https://doi.org/10.18653/v1/D19-1103

[14] Anjali Pande. 2004. Undoing Gender Stereotypes in Hindi. Linguistik online 21 (01 2004).

[15] Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis C. Lamb. 2020. Assessing Gender Bias in Machine Translation - A Case Study with Google Translate. Neural Computing Applications 32 (2020), 6363–6381.

[16] 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 (Sanya, China) (ACAI 2019). Association for Computing Machinery, 450–456.

[17] Shauni Ravfogel, Yana Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics, Online, 7237–7256.

[18] Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2019. Examining Gender Bias in Languages with Grammatical Gender. In Proceedings of EMNLP.

[19] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. CoRR ArXiv:1702.03859 http://arxiv.org/abs/1702.03859