"GENDER BIAS IN HINDI WORD EMBEDDING"

A THESIS REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTERS OF TECHNOLOGY
IN
COMPUTER SCIENCE

Submitted by

BARKHA BHARTI (CS1911)

Under the supervision of

DEBAPRIYO MAJUMDAR

Assistant Professor
Computer Vision and Pattern Recognition Unit
Computer and Communication Sciences Division

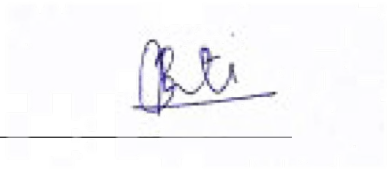
COMPUTER SCIENCE
INDIAN STATISTICAL INSTITUTE
205, B.T Road, Kolkata 700108

JULY, 2021

MASTERS OF TECHNOLOGY IN COMPUTER SCIENCE
INDIAN STATISTICAL INSTITUTE
205, B.T Road, Kolkata 700108

CANDIDATE'S DECLARATION

I, **Barkha Bharti (CS1911)**, students of M.Tech in Computer Science, hereby declare that the project Dissertation titled "**Gender Bias in Hindi Word Embedding**" which is submitted by me to the **M.Tech. (CS) Dissertation Committee 2020-21, ISI Kolkata** in partial fulfillment of the requirement for the award of the degree of Masters of Technology, is original and not copied from any source without proper citation.



Place: Kolkata

Date: 09.07.2021

Barkha Bharti

CS1911

MASTERS OF TECHNOLOGY IN COMPUTER SCIENCE
INDIAN STATISTICAL INSTITUTE
205, B.T Road, Kolkata 700108

CERTIFICATE

I hereby certify that the Project Dissertation titled "**Gender Bias in Hindi Word Embedding**" which is submitted by **Barkha Bharti (CS1911)**, **M.Tech. in Computer Science**, Indian Statistical Institute, Kolkata in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Kolkata

Date: 09.07.2021

Debapriyo Majumdar

Supervisor

ACKNOWLEDGMENT

I wish to express my sincerest gratitude tomy supervisor, **Debapriyo Majumdar** for the continuous guidance and mentorship that he provided me during the project. He showed the path to achieve the targets by explaining all the tasks to be done and explained the importance of this project as well as its industrial relevance. He was always ready to help me and clear my doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Kolkata — Barkha Bharti

Date: 09.07.2021 — CS1911

Abstract

The purpose of this paper is to present a study on gender bias in word embeddings in the context of the Hindi Language. It has been shown that word embeddings capture human biases (such as gender bias) present in the corpus and how they relate words to each other. The Hindi-language word embeddings were chosen with the intent of giving insight into gender bias across a variety of domains, with the expectation that some would show significantly greater bias than others. We use WEAT's hypothesis testing technique to confirm the presence of gender bias, and we find it useful for expanding the very narrow range of well-known gender bias word categories often used in the literature. We'll test the presence of gender bias in four sets of word embeddings trained on corpora from different domains: Hindi CoNLL17, Hindi Wikipedia 2016 database dumps, and Bollywood lyrics dataset. We also mitigate the bias from the embedding by identifying the gender direction and quantifying the bias independent of its alignment with the crowd bias. Then, we'll explore the efficacy of debiased embedding using Sentiment Analysis of Hindi Movie reviews and compare the results of sentiment analysis using original embedding and debiased embedding.

Contents

Chapter 1

Introduction

Newspaper articles, TV programs, and other public discussions show that gender inequality isn't just an effective but also a contentious issue, on which numerous individuals have solid perspectives. Women with similar professional degrees and experience have fewer opportunities to advance in the workplace because of prejudice and stereotypes. Similarly, if a boy wants to be a babysitter, he is significantly more likely to be discriminated against than that of a girl because of notions of appropriate masculinity and femininity.[Robeyns, 2007]. [Haines et al., 2016] in their study proved that despite time, from the 1980s to 2014 people perceive strong differences between men and women on stereotype components today, as they did in the past. Because gender stereotypes appear to be so firmly established in our culture, people in positions to judge women and men, as well as women and men themselves, must be continually aware of stereotypes' potential impact on their judgments, decisions, and actions.

1.1 Background

Gender equality and women's empowerment have been a concern of countries all over the world at the international level. In the project report by the Department of Gender Studies, NCERT, they have found that the term equality had been used in a limited sense, with authors attempting to promote equality by simply enhancing the visual representation of girls and women in various disciplines, or by facilitating role reversals to illustrate gender equality. When the content centered on female achievers, their work was often described in related terms, such as "wife of," "sister of," "mother of," and "daughter of." There was always an inherent comparison with the male counterpart. Rani Lakshmibai, for example, was praised for her bravery in confronting British forces, and she was referred to in novels as 'Khoob Ladi Mardani Woh To Jhansi Wali Rani Thi' (The queen of Jhansi, Rani Lakshmibai fought like a male). Women's achievements are also depicted as collateral in other themes in narratives; for example, Rani Lakshmibai's contributions and Madam Curie's work were associated with their domestic roles. Several papers have shown the presence of gender bias in the data in the Hindi language.
In the era of Artificial Intelligence(AI), gender biases are translated from sourced data to existing algorithms that may reflect and amplify existing cultural prejudices and inequalities by replicating human behavior and perpetuating bias.[Sweeney, 2013]

A defining feature of neural network language models is their representation of words as high dimensional real-valued vectors where these word representations capture meaningful syntactic and semantic regularities in a very simple way. [Mikolov et al., 2013b].

Each word (or common phrase) w is represented as a d-dimensional word vector $\sim w \epsilon R$ d in this word embedding. Word embeddings, which are trained only on word co-occurrence in text corpora, work as a sort of dictionary for computer programs that would like to use word meaning. The distributed representation of words helps to achieve better performance in NLP tasks by grouping similar words and if the contexts of two words exhibit a significant level of overlap, it can be confidently assumed that they are extremely synonymous. Using fundamental mathematical operations on word vector representations, a non-obvious degree of language understanding can be gained. [Mikolov et al., 2013a, Rubenstein and Goodenough, 1965].
There are hundreds or thousands of documents that have been written on word embedding and its applications. [Nalisnick et al., 2016] used dual word embedding to improve the ranking of documents.[Hansen et al., 2015]used to improve the performance of resume parsing by considering best word vectors, and [İrsoy and Cardie, 2014], for the sentiment analysis proposed a new architecture- deep recursive neural network (deep RNN) constructed by stacking multiple recursive layers and pre-trained word embedding trained on part of the Google News dataset.

It has been proven that word embeddings capture human biases (such as gender prejudice) in how they relate words to each other in these corpora.[Bolukbasi et al., 2016, Caliskan et al., 2017, Garg et al., 2018]

Several methods have been proposed to test the presence of gender bias and to mitigate the bias from embedding and the underlying source data, which we'll see in the next chapter.

1.2 Aim and Scope

There are different methods used to detect gender bias in the word embedding and also the debiasing algorithms. Most of the works focused on removing the gender stereotypes in the word embeddings trained on English Language data. But only a few research on Hindi word embedding.
The goal of this study is -

1. To test the presence of gender bias in Hindi Word embeddings using Word Embedding Test.[Caliskan et al., 2017, Chaloner and Maldonado, 2019]

2. To debias the word embedding [Bolukbasi et al., 2016]

3. To check the effectiveness of debiased embedding using NLP tasks: *Sentiment analysis on Hindi Movie review dataset.*

Chapter 2

Related Work

The gender bias has been detected in many models by current machine learning research, each with its evaluation and debiasing approaches and it has been examined in word embeddings, coreference resolution, and, more recently, datasets in Natural Language Processing (NLP).[Bolukbasi et al., 2016, Caliskan et al., 2017, Zhao et al., 2018, Hitti et al., 2019]. Linguists have previously addressed gender prejudice in writing by developing inclusive writing styles.

In word embedding, words with similar semantic meanings tend to have vectors that are close together. For example, The vector for the term *strong* would be similar to that of a man, but the vector for *soft* would be similar to that of a woman. The embedding algorithm learns these prejudices automatically, which could be problematic if the embedding is used for sensitive applications like search rankings, resume recommendations, or translations.

The Implicit Association Test is a widely used approach for assessing cultural prejudices at an individual level.[Greenwald et al., 1998]. The Implicit Association Test (IAT) assesses attitudes and ideas that people are reluctant or unable to express. The IAT may be particularly intriguing if it reveals an implicit attitude that you were previously unaware of. For example, Although you may believe that men and women should be equally associated with maths, your natural associations may indicate that you (like many others) associate men with maths more than women and vice versa for arts.
The IAT employs a reaction time paradigm, in which subjects are pushed to work as quickly as possible, with their response times served as the metric. The IAT is ordinarily used to pair categories such as 'male' and 'female' with attributes such as 'violent' or 'peaceful'. There is also an imbalance in the number of words with F-M with various associations. For instance, while more words are referring to males, there are many more words that sexualize females than males. [Stanley, 1977].

[Bolukbasi et al., 2016] showed that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. To better understand the gender bias subspace, gender-specific words were investigated to compare their distances with respect to other words in the vector space. [Bolukbasi et al., 2016]. [Caliskan et al., 2017] then developed the Word Embedding Association Test (WEAT), which is an adaptation of the Implicit Association Test (IAT) [Greenwald et al., 1998] to measure biases in word embeddings.

To better identify gender bias in coreference resolution systems, [Zhao et al., 2018] build a new dataset (WinoBias) centered on people entities referred by their occupations and found that training data and auxiliary resources are the two sources of gender bias in co-reference systems that can cause them to fail WinoBias and propose strategies to mitigate them. To remove bias in training data, [Zhao et al., 2018] adopt a simple rule-based

approach for gender-swapping which maintains non-gender-revealing correlations while eliminating correlations between gender and coreference cues and for word embeddings, they replaced GloVe embeddings with debiased vectors[Bolukbasi et al., 2016]. In combination with methods that remove bias from fixed resources such as word embeddings [Bolukbasi et al., 2016], the data augmentation approach eliminates bias when evaluating on WinoBias, without significantly affecting overall coreference accuracy.
[Zhang et al., 2018] mitigate gender bias with an adversarial network by adapting a technique presented by [Bolukbasi et al., 2016] to define a subspace capturing the semantics of the protected variable, and then train a model to perform a word analogies task accurately, while unbiased on this protected variable. A consequence of this technique is that the network learns "debiased" embeddings, embeddings that have the semantics of the protected variable removed. These embeddings are still able to perform the analogy task well but are better at avoiding problematic examples such as those shown in [Bolukbasi et al., 2016]
[Garg et al., 2018]use the word embeddings as a quantitative lens through which to study historical gender stereotypes and ethnic stereotypes in the 20th and 21st centuries in the United States.

[Caliskan et al., 2017] confirmed the presence of gender bias using three categories of words well known to be prone to exhibit gender bias: (B1)career vs. family activities, (B2) Maths vs.Arts, and (B3) Science vs. Arts. [Garg et al., 2018] expanded on this work and tested additional gender bias word categories: (B4) differences in personal descriptions based on intelligence vs. appearance and (B5) physical or emotional strength vs. weakness. [Chaloner and Maldonado, 2019] used these five categories to test for the presence of gender bias in the Google News, Twitter, PubMed, and GAP corpus using WEAT's hypothesis testing mechanism to automatically validate the induced gender bias word categories produced by the system. The list of categories and target words used in WEAT is given in figure 2.1

	Attribute words	M	male, man, boy, brother, he, him, his, son, father, uncle, grandfather
		F	female, woman, girl, sister, she, her, hers, daughter, mother, aunt, grandmother
Target words	B1: career vs family	X	executive, management, professional, corporation, salary, office, business, career
		Y	home, parents, children, family, cousins, marriage, wedding, relatives
	B2: maths vs arts	X	math, algebra, geometry, calculus, equations, computation, numbers, addition
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	B3: science vs arts	X	science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy
		Y	poetry, art, Shakespeare, dance, literature, novel, symphony, drama
	B4: intelligence vs appearance	X	precocious, resourceful, inquisitive, genius, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, apt, venerable, imaginative, shrewd, thoughtful, wise, smart, ingenious, clever, brilliant, logical, intelligent
		Y	alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong
	B5: strength vs weakness	X	power, strong, confident, dominant, potent, command, assert, loud, bold, succeed, triumph, leader, shout, dynamic, winner
		Y	weak, surrender, timid, vulnerable, weakness, wispy, withdraw, yield, failure, shy, follow, lose, fragile, afraid, loser

Figure 2.1: List of target words used for each gender-bias word category and attribute words used as gender reference [Chaloner and Maldonado, 2019]

The research papers that we have seen so far are focused on word embeddings trained on English data. Few research papers study this aspect of word-embedding in the context of the Hindi language.[Pujari et al., 2019, Madaan et al., 2018].

[Pujari et al., 2019] proposed a new algorithm of debiasing and demonstrate its efficacy in the context of the Hindi language and further build an SVM-based classifier that determines whether a gender-neutral word is classified as neutral or otherwise. [Madaan et al., 2018] focused on studying such stereotypes and bias in Hindi movie industry (*Bollywood*) and proposed an algorithm to remove these stereotypes from the text. Also proposed debiasing algorithm that extracts gender-biased graphs from an unstructured piece of text in stories from movies and debias these graphs to generate plausible unbiased stories. Then they show that interchanging the gender of high centrality male character with a high centrality female character in the plot text leaves no change in the story but de-biases it completely.

In this paper, we will use previously administered WEAT's designed to measure the gender bias in Hindi word embedding We'll use the same five categories that have been used in the papers. [Caliskan et al., 2017, Garg et al., 2018, Chaloner and Maldonado, 2019]. And for mitigating the gender bias from word embeddings, we'll use the debiasing algorithm proposed by [Bolukbasi et al., 2016].

Chapter 3

Methodology

3.1 Choice of Word Embeddings

The word embeddings selected were-

1. Skip-Gram embedding trained on Hindi CoNLL17 corpus, with a vocabulary of 0.2M words and size 100 dimensions, trained using a window size of 10. The raw text sources for the Hindi CoNLL17 corpus are- Wikipedia dumps, Perseus Digital Library, and Common Crawl . [Ginter et al., 2017].

2. Skip-Gram embedding trained on Wikipedia 2016 database dumps corpus of size 323M, trained with a vocabulary of 30.4k words and vector size of 300 words, a window size of 5 words.

3. Fasttext embedding trained on Wikipedia 2016 database dumps corpus of size 323M, trained with a vocabulary of 30.4k words and vector size of 300 words, a window size of 5 words.

4. BERT trained on Bollywood lyrics dataset using a masked language model. This dataset consists of 15000 lyrics created from `https://www.giitaayan.com/`. [Pai, 2021]

3.2 Word Embedding Association Test

To demonstrate and quantify bias, we highly follow the WEAT Hypothesis testing protocol introduced by [Caliskan et al., 2017, Chaloner and Maldonado, 2019]. WEAT is based on two statistical measures: (1) the effect size in terms of Cohen's d, which measures the association between suspected gender biased words and two sets of reference words (attribute words in WEAT's terminology) known to be intrinsically male and female, respectively; and (2) a statistical hypothesis test that confirms this association.
The input is a suspected gender bias word category represented by two lists, X and Y , of target words, i.e. words which are suspected to be biased to one or another gender. E.g. $X = \{programmer, engineer, scientist\}$, $Y = \{nurse, teacher, librarian\}$. We wish to test whether X or Y is more biased to one gender or the other, or whether there is no difference in bias between the two lists. Bias is compared in relation to two reference lists of words that represent unequivocally male and female concepts. E.g. $M = \{man, male, he\}, F = \{woman, female, she\}$. In WEAT's terminology, these reference lists are called attribute words. [Chaloner and Maldonado, 2019] used these five categories to test for the presence of gender bias: (B1) career vs. family (B2) Maths

vs. Arts, (B3) Science vs. Arts, (B4)intelligence vs. appearance, and on (B5) physical or emotional strength vs. weakness. The five word categories studied here are word lists manually curated by Psychology researchers based on their studies [Greenwald et al., 1998]. Table 3.1 shows the target and attribute word sets used in our experiments.

The null hypothesis is that there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words. Let X and Y be two sets of target words of equal size, and A, B the two sets of attribute words. Let $\cos(\vec{a}, \vec{b})$ denote the cosine of the angle between the vectors $\vec{a}$ and $\vec{b}$.

- The test statistic is

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

  where

  $$s(w, A, B) = mean_{a \epsilon A} \cos(\vec{w}, \vec{a}) - mean_{b \epsilon B} \cos(\vec{w}, \vec{b})$$

  In other words, s(w,A,B) measures the association of the word w with the attribute, and s(X,Y,A,B) measures the differential association of the two sets of target words with the attribute.

- Let $\{X_i, Y_i\}_i$ denote all the partitions of $X \cup Y$ into two sets of equal size. The one-sided p-value of the permutation test is-

  $$Pr_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

  i.e. the proportion of partition permutations X_i, Y_i in which the test statistic $s(X_i, Y_i, A, B)$ is greater than the observed test statistic $s(X, Y, A, B)$]. This p-value is the probability that Ho is true. In other words, it is the probability that there is no difference between X and Y (in relation to M and F) and therefore that the word category is not biased.

- The effect size is-

  $$\frac{mean_{a \epsilon A} \cos(\vec{w},\vec{a}) - mean_{b \epsilon B} \cos(\vec{w},\vec{b})}{std-dev_{w \epsilon X \cup Y} s(w,A,B)}$$

The higher this p-value is the less bias there is. In this study, we consider a word category to have statistically significant gender bias if its p-value is less than the 0.05 criterion, as proposed by [Caliskan et al., 2017] We utilize randomization tests [Chaloner and Maldonado, 2019] with a maximum of 10,000 iterations in this research because a full permutation test can quickly become computationally intractable.

	Attribute words	F	नारी, औरत, बेटी, बहन, स्त्री, महिला, बीवी, ब्याहता, बच्ची, दादी, मां, लड़की,उसकी, गर्भवति, चाची
		M	पिता, उसका, पति, पुत्र, पोता, चाचा, पुरुष, बच्चा, दादा, भाई, बेटा, लड़का, भतीजा, आदमी, बाप
Target Words	B1: career vs family	X	टेलीमार्केटिंग, पुनर्बीमा, कार्यालय, निगम, व्यापार, उद्यम, मूल्यवर्धित, विकास, बहुपक्षीय, वृत्ति, आजीविका, विप्रेषण, कारोबार, वाणिज्य-, सहयोग, कार्यपालक, वेतन, व्यवसाय, आयात-निर्यात, प्रबंध
		Y	पिता, रिश्तेदारों, पति, बाल, बच्चे, संबंधियों, मामा-मामी, माता, परिवारजनों, बहिन, शादी, रिश्तेदार, पत्नी, घर, पड़ोसियों, भाई, बुजुर्ग, दामाद, परिवार, चाचा-चाची
	B2: maths vs arts	X	रेखागणित, क्षेत्रमिति, त्रिकोणमिति, गणना, आर्यभट, नंबर, गणित, श्रृंख्ला, जोड़, पाइथागोरस, ज्यामिति, चक्रवाल, यूलर, अंकगणित, अर्थशास्त्र, बीजगणित, यूक्लिड, अभिकलनात्मक, समीकरण, कैलकुलस
		Y	साहित्य, वास्तुकला, शेक्सपियर, नाटक, कला, नृत्य, संगीत, कथावाचन, वास्तुशिल्प, चित्रकला, कलात्मक, शिल्पकला, गायन, कविता, मूर्तिकला, शिल्प, नर्तक, उपन्यास, ललितकला, भरतनाट्यम
	B3: science vs arts	X	संकाय-, मनोविज्ञान, जीवविज्ञान, खगोल, जनजातिय, नासा, भौतिकी, गृहविज्ञान, वैज्ञानिक, अनुसंधान, प्रौद्योगिक, रसायन विज्ञान, अनुप्रयुक्त, प्राणिशास्त्र, भौतिक विज्ञान, प्रयोग, आइंस्टाइन, विज्ञान, जैविकी, प्रौद्योगिकी,
		Y	साहित्य, वास्तुकला, शेक्सपियर, नाटक, कला, नृत्य, संगीत, कथावाचन, वास्तुशिल्प, चित्रकला, कलात्मक, शिल्पकला, गायन, कविता, मूर्तिकला, शिल्प, नर्तक, उपन्यास, ललितकला, भरतनाट्यम
	B4: intelligence vs appearance	X	मेहनती, असामयिक, आविष्कारशील, तार्किक, कल्पनाशील, चिंतनशील, होशियार, सम्मानित, सरल, विवेकी, चतुर, चालाक, जिज्ञासु, अनुकूलनीय, प्रतिभाशाली, परिश्रमी, समझदार, विश्लेषणात्मक, सावधान, बुद्धिमान
		Y	मनमोहक, मनोहारी, स्वस्थ, मनोहर, मोटी, सुखकर, कामुक, सुंदर, पुष्ट, मोटा, फैशनेबल, ताक़तवर, मांसल, खूबसूरत, मज़बूत, पतला, भव्य, कमजोर, कुरूप, लालित
	B5: strong vs weak	X	शक्ति, बुद्धिमान, नेता, आदेश, साहसिक, प्रबल, विश्वास, सामर्थ्य, प्रमुख, पराक्रमी, चिल्लाओ, ताकत, विजेता, जाग्रत्, ट्राइंफ, जोर, क्षमता, ज़ोर, बलवान, गतिशील,
		Y	भय, दुर्बलता, खोना, धुँधला, आत्मसमर्पण, नासमझ, निर्बलता, कमज़ोर, जूझने, परास्त, संकोच, कांपने, नाज़ुक, निराशा, डरपोक, कचदिला, छिपना, कायर, डर, असफलता

Table 3.1: List of target words and attribute words

3.3 Debiasing Algorithm

To debias a word embedding, we will be using the method proposed by [Bolukbasi et al., 2016], where we find a linear projection of gender-neutral words toward a subspace, which is orthogonal to the gender direction vector defined by a set of gender-definition words.
In this section, We examine the bias in the embedding geometrically, identifying the gender direction and quantifying the bias independent of its alignment with the crowd bias and we'll explore the debias algorithm proposed by [Bolukbasi et al., 2016].

3.3.1 Identifying the gender subspace

Individual word pairs do not always behave as expected because language use is "messy." For example, the term man can be used as an exclamation, such as in *oh man!*, or to refer to people of either gender or as a verb, such as *man the station.*

To more robustly estimate bias, we shall aggregate across multiple paired comparisons. By combining several directions, identify a gender direction that largely captures gender in the embedding. This direction helps us to quantify direct and indirect biases in words and associations. We'll aggregate across numerous paired comparisons to get a more robust estimate of bias. To identify a gender direction $g \epsilon R$ that largely captures gender in the embedding by combining different directions such as $\overrightarrow{she} - \overrightarrow{he}$ and $\overrightarrow{woman} - \overrightarrow{man}$.[Bolukbasi et al., 2016]. But we'll use "हमारा" - "हमारी" because in Hindi, 'he' and 'she' both are represented by 'वह'. This approach allows us to measure both direct and indirect biases in words and connections. To identify the gender subspace, we took the ten gender pair difference vectors and computed its principal components (PCs). As

Figure 3.1 shows, there is a single direction that explains the majority of variance in these vectors. The first eigenvalue is significantly larger than the rest.

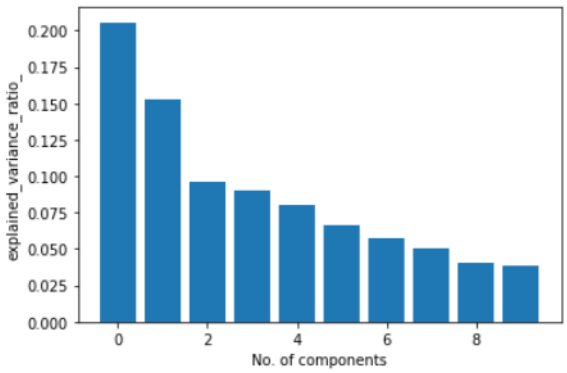


Figure 3.1: the percentage of variance explained in the PCA of these vector differences (each difference normalized to be a unit vector). The top component explains significantly more variance than any other. [Bolukbasi et al., 2016]

3.3.2 Debiasing Algorithm

To define the algorithms, [Bolukbasi et al., 2016] introduced some further notation. A subspace B is defined by k orthogonal unit vectors B = $\{b1.....bk\} \subset R^d$. In the case k = 1, the subspace is simply a direction. We denote the projection of a vector v onto B by,

$$v_B = \sum_{j=1}^{k}(v.b_j)b_j.$$

This also means that $v - v_B$ is the projection onto the orthogonal subspace.

1. **Identify gender subspace**, to identify a direction (or, more generally, a subspace) of the embedding that captures the bias. Inputs: word sets W, defining sets $D, D....., D \subset W$ as well as embedding $\{\vec{w}\epsilon R\}_{w\epsilon W}$ and integer parameter $k \geq 1$. Let

   $$\mu_i := \sum_{w\epsilon D_i} \vec{w}/|D_i|$$

   be the means of the defining sets. Let the bias subspace B be the first k rows of SVD(C) where

   $$C := \sum_{I=1}^{n} \sum_{w\epsilon D_i}(\vec{w} - \vec{\mu}_i)^T/|D_i|$$

2. **a: Hard de-biasing (neutralize and equalize)** Neutralize ensures that gender-neutral words are zero in the gender subspace. Equalize perfectly equalizes sets of words outside the subspace and thereby enforces the property that any neutral word is equidistant to all words in each equality set.

   Additional inputs: words to neutralize $N \subset W$, family of equality sets $E = \{E_1, E_2....., E_m\}$ where each $E_i \subset W$. For each word $w\epsilon N$, let $\vec{w}$ be re-embedded to-

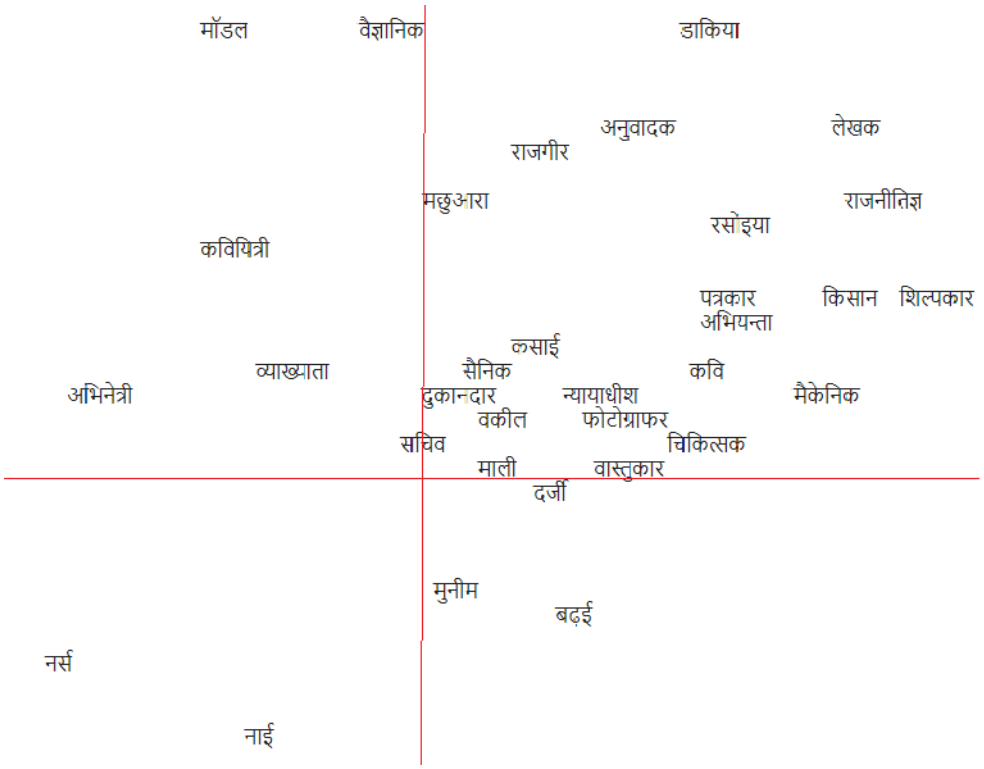


Figure 3.2: Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words "हमारा" and "हमारी" and y is a direction learned in the embedding that captures gender neutrality, with gender-neutral words above the line and gender-specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender-neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line.

$$\vec{w} := (\vec{w} - \vec{w}_B)/||(\vec{w} - \vec{w}_B)||$$

For each set $E \epsilon \mathcal{E}$, let

$$\mu_i := \sum_{w \epsilon E} w/E|$$
$$\nu := \mu - \mu_B$$
$$\text{For each } w \epsilon E,\ \vec{w} := \nu + \sqrt{1 - ||\nu||^2} \frac{\vec{w} - \vec{\mu}_B}{||(\vec{w} - \vec{\mu}_B)||}$$

Finally, output the subspace B and the new embedding $\vec{w} \epsilon R^d$
[Bolukbasi et al., 2016] observed that After Steps 1 and 2a, for any gender neutral word w any equality set E, and any two words $e_1, e_2 \epsilon E, \vec{w}.\vec{e_1} = \vec{w}.\vec{e_2} and ||\vec{w} - \vec{e_1}|| = ||\vec{w} - \vec{e_2}||. Furthermore, if \mathcal{E} = \{\{x, y\}|(x, y) \epsilon P\}$,are the sets of pairs defining Pair-Bias, then PairBias = 0.

3.3.3 Determining gender neutral words

Here, we used the same approach that [Bolukbasi et al., 2016] used. Given a list of fewer gender-specific words S, we enumerate and take the gender-neutral words to be the compliment, $N = W\ S$.

We generalize this list to the entire words in the embedding using a linear classifier, resulting in the larger set S of gender-specific words. More specifically, we trained a linear Support Vector Machine (SVM) with the default regularization parameter of C = 1:0. Figure 3.2 illustrates the results of the classifier for separating gender-specific occupations from gender-neutral occupations for the given set of occupational words.

3.4 Exploring the efficiency of debiased embedding using an NLP Task

In this section, we'll explore the efficiency of our debiased embedding using Sentiment analysis on the Hindi movie review dataset. Instead of testing on all of the four embeddings, we'll test only on the Skip-Gram embedding trained on the Hindi CoNLL17 corpus.

Hindi Movie reviews dataset: This data set contains 900 Movie Reviews of 3 classes (Positive, Neutral, Negative) which had collected from Hindi News Websites. The data set has been cleaned and contains a fairly balanced train and test set using which we can train our sentiment analysis and classification models in Hindi.

The texts in this dataset have been already divided into negative sentiments(denoted by 0) and positive sentiments (denoted by 1), and the neural sentiment texts are deleted because it becomes hard to have words that act as identifiers to a neutral sentiment which causes the performance of the model go down.

We load the positive and negative review texts and preprocess them, filters out the tokens that are present in the embedding, and then encode each document as a sequence of integers. The architecture of the model can be seen in figure 3.3.

The model uses an Embedding layer as the first hidden layer, with the weight matrix created from the pre-trained embedding passed as the weights and we set the 'trainable' argument to 'False' to ensure that the network does not try to adapt the pre-learned vectors as part of training the network. We can see that the Embedding layer expects documents with a length of 801 words as input and encodes each word in the document as a 100 element vector.

And the next layer is the LSTM layer with 64 units and dropouts, followed by a fully connected layer with 64 units are used. At last, the dense layer with the 'sigmoid' activation function is used. We use a binary cross-entropy loss function because the problem we are learning is a binary classification problem. The efficient Adam implementation of stochastic gradient descent is used with a learning rate of 0.001 and we keep track of accuracy in addition to loss during training. The model is trained for 25 epochs and batch size 64 on the training set with validation data.

We'll train two models for biased and debiased embeddings respectively and compare the results.

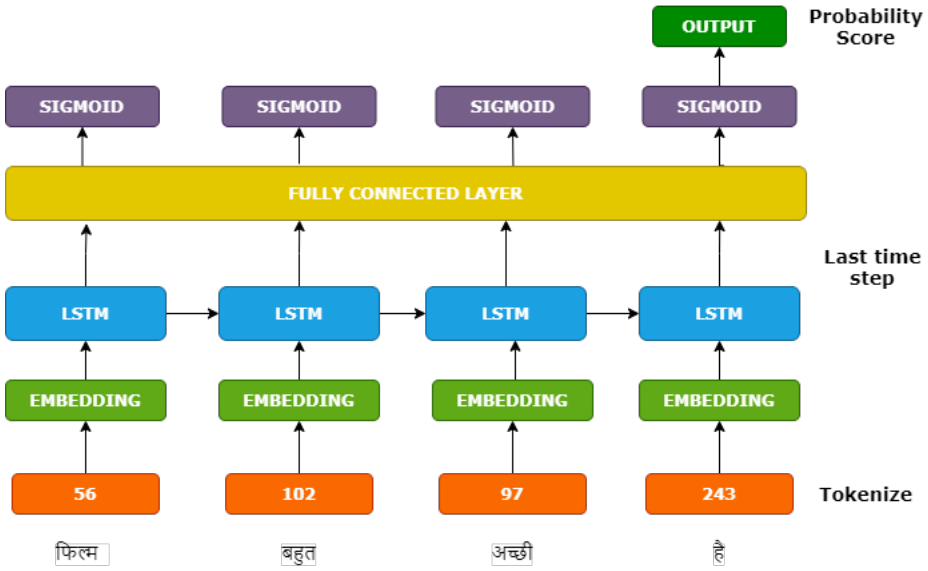


Figure 3.3: LSTM Architecture for Sentiment Analysis

In the next chapter we'll see the results of our WEAT hypothesis testings on all of the four embeddings for the five categories, and also the results of the Sentiment Analysis of Hindi movie reviews for biased and debiased embeddings.

Chapter 4

Results and Discussion

4.1 WEAT Results

We can see the results of the WEAT hypothesis test and the effect size in table 4.1. For category B3: science vs arts, we detect statistically significant (p-values in bold) gender bias in all 4 embeddings.

Hindi CoNLL17 On this dataset, we detect bias in two categories i.e B2: maths vs arts and B4: intelligence vs appearance while the p-value for B1: career vs family is very high. We also observe that most effect sizes (Cohen's d) are under 1, indicating relatively weaker associations with the gender-specific attribute words from table 3.1

Wikipedia 2016 database On the skip-gram embedding of this dataset, we detect bias in two categories i.e B2: maths vs arts and B3: science vs arts with p-value 0 and also having high associations with the gender-specific attribute words. Although B4: intelligence vs appearance is a borderline case with a p-value of just 0.0682. But in fasttext embedding, we detect bias in B4: intelligence vs appearance. Also of note is that across all five categories, bias is greater (smaller p-values) on the fasttext embedding than on the skip-gram embedding.

Bollywood lyrics We didn't find any science word in this dataset that's why for B3: science vs arts, the p-value is 0, and cohen's d is NA. We detect gender bias in B2: maths vs arts and B4: intelligence vs appearance (very low p-value i.e highly biased)

	Hindi CoNLL17		Wikipedia2016 w2v		Wikipedia2016 ft		Bollywood lyrics	
Categories	p	d	p	d	p	d	p	d
B1: career vs family	0.9182	-0.444	0.745	-0.1891	0.4909	0.1023	0.754	-0.0276
B2: maths vs arts	**0.0022**	0.848	**0.0**	1.1051	**0.0**	1.117	**0.0265**	0.419
B3: science vs arts	0.2066	0.205	**0.0**	1.1357	**0.0**	1.129	**0.0**	-
B4: intelligence vs appearance	**0.0028**	0.888	0.0682	0.531	**0.0399**	0.623	**0.0093**	1.164
B5: strong vs weak	0.2372	0.2389	0.559	-0.0294	0.161	0.2948	0.2732	0.294

Table 4.1: WEAT hypothesis test results for corpora tested for five well-known gender-biased word categories. p- values in bold indicate statistically significant gender bias (p $< 0:05$)

4.2 Sentiment Analysis of Hindi movie reviews

To evaluate the classification performance, standard evaluation metrics of precision, recall, F-measure, and accuracy were used to compare the results of the LSTM model for two different pre-trained word embeddings. Table 4.2 shows the results.

	Biased		Debiased	
class	0	1	0	1
Precision	0.64	0.70	0.63	0.76
Recall	0.66	0.69	0.77	0.61
f1 score	0.65	0.69	0.69	0.67
Accuracy	0.67		0.68	

Table 4.2: Sentiment Analysis Results of Hindi movie review dataset for biased and debiased embeddings

We can see that the LSTM model with debiased and original embedding gives comparatively the same accuracy. Although the precision and recall are not appropriate here as the dataset was not imbalanced, we can see that with the original/biased embedding, precision and recall are high for class 1, and with debiased embedding, we are getting high precision for class 1 while high recall for class 0.
We can do the comparison based on the accuracy, which shows that the model with debiased embedding is working better.

Chapter 5

Conclusion and Future Scope

We demonstrated that word embeddings trained on corpora from various domains exhibit diverse amounts of bias and that different categories of gender bias exist within the embeddings. We have also mitigated the gender bias from the Hindi word embeddings geometrically and normalize it to get a new debiased embedding. Then, to explore the efficacy of the debiasing technique, we did Sentiment Analysis of Hindi Movie reviews using LSTM as the baseline model. We build and trained two LSTM models using the pre-trained embeddings(original and debiased) in the Embedding Layer. Then, compared the accuracy to check the efficiency of the debiased embedding. And we found that the debiased embedding giving results with slightly better accuracy.

So as future development, we would like to develop a system to check if a document is gender-biased or not and build an algorithm to generate a debiased document. We want to explore articles in different domains and by different authors to determine the bias.

Bibliography

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

[Caliskan et al., 2017] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[Chaloner and Maldonado, 2019] Chaloner, K. and Maldonado, A. (2019). Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

[Garg et al., 2018] Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

[Ginter et al., 2017] Ginter, F., Hajič, J., Luotolahti, J., Straka, M., and Zeman, D. (2017). Conll 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[Greenwald et al., 1998] Greenwald, A. G., McGhee, Schwartz, D. E. ., and L.K., J. (1998). Measuring individual differences in implicit cognition: The implicit association test. 74(6).

[Haines et al., 2016] Haines, L., E., K., D., and N., L. (2016). The times they are a-changing … or are they not? a comparison of gender stereotypes, 1983–2014. volume 40, pages 353–363. Psychology of Women Quarterly.

[Hansen et al., 2015] Hansen, C., Tosik, M., Goossen, G., Li, C., Bayeva, L., Berbain, F., and Rotaru, M. (2015). How to get the best word vectors for resume parsing. In *SNN Adaptive Intelligence/Symposium: Machine Learning*.

[Hitti et al., 2019] Hitti, Y., Jang, E., Moreno, I., and Pelletier, C. (2019). Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

[İrsoy and Cardie, 2014] İrsoy, O. and Cardie, C. (2014). Deep recursive neural networks for compositionality in language. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 2096–2104. Curran Associates, Inc.

[Madaan et al., 2018] Madaan, N., Mehta, S., Agrawaal, T., Malhotra, V., Aggarwal, A., Gupta, Y., and Saxena, M. (2018). Analyze, detect and remove gender stereotyping from bollywood movies. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 92–105, New York, NY, USA. PMLR.

[Mikolov et al., 2013a] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality.

[Mikolov et al., 2013b] Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

[Nalisnick et al., 2016] Nalisnick, E., Mitra, B., Craswell, N., and Caruana, R. (2016). Improving document ranking with dual word embeddings. Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Pai, 2021] Pai, D. (2021). Author identification of bollywood song lyrics. (Unpublished).

[Pujari et al., 2019] Pujari, A. K., Mittal, A., Padhi, A., Jain, A., Jadon, M., and Kumar, V. (2019). Debiasing gender biased hindi words with word-embedding. New York, NY, USA. Association for Computing Machinery.

[Robeyns, 2007] Robeyns, I. (2007). When will society be gender just?

[Rubenstein and Goodenough, 1965] Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.

[Stanley, 1977] Stanley, J. P. (1977). Paradigmatic woman: The prostitute. *Papers in language variation*, pages 303–321.

[Sweeney, 2013] Sweeney, L. (2013). Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29.

[Zhang et al., 2018] Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning.

[Zhao et al., 2018] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.