Bias Detection in Word Embeddings of Indian Languages

By

Karthik Varunn.S (195002059)

Manasi.R (195002070)

Guided by

Mr. B. Senthil Kumar

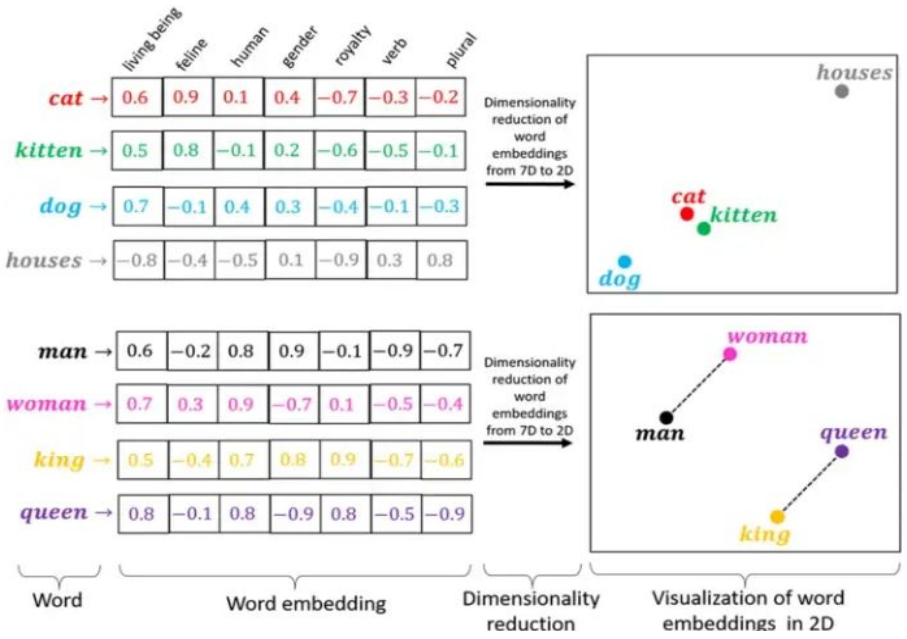
Introduction

- Natural Language Processing systems understand only numerical data, thereby making word embeddings essential for NLP downstream tasks, as embeddings represent words in numerical format which can easily be processed.
- Bias detection in word embeddings is the process of identifying and quantifying the amount of bias present in the word embeddings.
- These NLP systems capture linguistic regularities through the embeddings, which reflects human biases like gender, racial and cultural biases. Thus, there is a pressing need to address and overcome the gender bias in these systems, to avoid any unintentional discrimination and propagation of stereotypes.
- Hindi is the 3rd most spoken language in the world and the Tamil language is highly morphologically rich. This project evaluates the gender bias in the Tamil and Hindi languages.
- Word vectors trained on FastText embedding model are used with stereotyped occupational words, taken from analogy datasets for these languages. The bias is measured with the help of metrics like the Word embedding association test (WEAT) score, Relative Norm Distance (RND) and the Embedding Coherence Test (ECT). We also attempt to mitigate the gender bias recorded in these embeddings. The results show that the debiasing technique used, reduces the gender bias in these languages and this has been represented graphically.

Problem Statement

The detection of bias in word embeddings with reference to the Indian Languages such as - Tamil and Hindi.

- **Word embedding** - Word Embedding is an approach for representing words in vector form. It provides similar vector representations for words that have similar meanings. It helps the model to capture the linguistic meaning of the word.
- **Bias** - Bias in word embeddings refers to the phenomenon where the embeddings of certain words or groups of words are systematically skewed towards certain concepts or ideas, often reflecting social and cultural biases in the data used to train the embeddings.
- Bias in word embeddings can lead to the perpetuation of stereotypes, exclusion of certain groups of people, and unfair treatment in areas such as employment, criminal justice, and healthcare. This issue is particularly relevant in India, a diverse country with a rich linguistic heritage, where language plays a significant role in shaping social and cultural norms.



Gender Biased Analogies	
man → doctor	woman → nurse
woman → receptionist	man → supervisor
woman → secretary	man → principal
Racially Biased Analogies	
black → criminal	caucasian → police
asian → doctor	caucasian → dad
caucasian → leader	black → led
Religiously Biased Analogies	
muslim → terrorist	christian → civilians
jewish → philanthropist	christian → stooge
christian → unemployed	jewish → pensioners

Motivation

- NLP datasets record patterns in language that mirror human prejudices, including those related to race, gender, caste, and demographics, by utilizing text embeddings.

- These text embeddings play a significant role in shaping the information sphere and can aid in making consequential inferences about individuals.

- Also, language models trained on biased data can perpetuate and even amplify those biases. This can have negative consequences, such as reinforcing systemic discrimination, creating inaccurate predictions, and limiting opportunities for individuals or groups that are negatively affected by such biases.

- NLP use cases include CV parsing and job application assessment (to name a few), which call for a fair and just system. As NLP models are critical to real-world applications and society, there is an urgent need to detect these biases. By removing biases from text embeddings, we can promote fairness, accuracy, and equality in NLP applications.

- There already exists embedding bias detection systems for the English language, however similar research is significantly lesser for indian languages such as Hindi and Tamil, hence serving as the motivation for our chosen project.

Literature Survey

Title: ”Mitigating Gender Stereotypes in Hindi and Marathi

Authors: Neeraja Kirtane and Tanvi Anand

Published: 2022

Description:

This paper evaluates the gender stereotypes in Hindi and Marathi languages. The methodologies differ from the ones in the English language because there are masculine and feminine counterparts in the case of some words.They have measured bias with the help of Embedding Coherence Test (ECT) and Relative Norm Distance (RND). They have also created a dataset of neutral and gendered occupation words, emotion words. An attempt to mitigate the bias was also carried out in this work.

Literature Survey

Title: Morphology-Aware Meta-Embeddings for Tamil

Authors: Arjun Krishnan, Seyoon Ragavan

Published: June 2021

Description:

In this work, the generation of morphologically enhanced word embeddings for Tamil was explored, a South Indian language that is highly agglutinative and has a rich morphology but is nevertheless considered to be low-resource in terms of NLP tasks. The first word analogy dataset for Tamil is presented here, and it consists of 4499 hand-curated word tetrads across 10 semantic relation categories and 13 morphological relation types.

Literature Survey

Title: Semantics derived automatically from language corpora contain human-like biases

Authors: Caliskan, A., Bryson, J. J., and Narayanan, A

Published: 2017

Description:

This study investigates biases in word embeddings derived automatically from text corpora. The authors found that such embeddings exhibit stereotypical human biases, particularly regarding gender and race and this was possible with the implementation of the WEAT(Word Embedding Association Test) score. The study concluded that these biases are likely to reflect the biases and stereotypes of the human societies that produce and use language. The findings have important implications for natural language processing applications, as biased word embeddings can perpetuate and amplify existing biases in society.

Literature Survey

Title: WEFE: The Word Embeddings Fairness Evaluation Framework

Authors: P. Badilla, F. Bravo-Marquez, and J. Perez

Published: 2020

Description:

Proposed the WEFE:Word Embeddings Fairness Evaluation Framework to gather, evaluate, and compare measures for fairness. WEFE provides a set of tools for evaluating the fairness of word embeddings based on different criteria such as gender, race and religion. It includes a set of fairness metrics that can be used to assess the extent to which word embeddings exhibit bias and discrimination against different groups of people.

Literature Survey

Title: Word embeddings quantify 100 years of gender and ethnic stereotypes.

Authors: Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou.

Published: 2018

Description:

They found that gender stereotypes were present throughout the entire time period, with words related to women more closely associated with home and family, while words related to men were more closely associated with careers and professional occupations.The work made use of the Relative Norm Distance(RND) score to quantify the extent and persistence of gender and ethnic biases in the word embeddings over time.

Literature Survey

Title: Attenuating Bias in Word vectors.

Authors: Dev, S., & Phillips, J.

Published: 2019, April

Description:

The Embedding Coherence Test (ECT) is a measure proposed in this work to evaluate the coherence and consistency of word embeddings. The ECT measures the degree of agreement between the embeddings of two sets of words that are expected to be similar based on their semantic properties. For example, the ECT was used to assess the coherence of word embeddings in capturing synonymy, by comparing the embeddings of pairs of synonyms (e.g., "happy" and "glad").

Literature Survey

Title: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

Authors: Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai

Published: 2016

Description:

The paper highlights how certain gender associations can be reinforced in word embeddings due to the biases present in the text corpora used to train them. To address this issue, Bolukbasi et al. propose a method for debiasing word embeddings - Hard Debiasing, which involves identifying the direction of the gender bias in the embeddings and projecting the word vectors onto a subspace that removes this bias. They have shown that they can successfully debias the word embeddings and improve their performance on gender-neutral tasks, such as analogical reasoning.

Objectives

- The primary objective of this project is to identify and quantify the presence of gender bias in word embeddings in the Tamil and Hindi language. Evaluation metrics such as WEAT, RND, and ECT are used to quantify the gender bias.
- To evaluate the effectiveness of these bias measurement methods on several word embedding models.
- To debias the word embeddings and optimize the scores (WEAT,RND,ECT) to a neutral point to indicate the impartiality of the same language.
- To contribute to the development of more fair and inclusive NLP applications for Indian languages by mitigating biases in the underlying word embeddings.
- To provide insights into the gender biases that may be reflected in language data and models, and to stimulate discussions around these issues.

Challenges

- Limited availability of data: The availability of large amounts of data is essential for training high-quality word embeddings. However, data in Indian languages may be limited, particularly for under-resourced languages.

- Lack of annotated data: Annotated data is crucial for identifying and quantifying biases in word embeddings. However, annotating data for Indian languages can be time-consuming and expensive. This can make it difficult to develop high-quality annotated datasets for bias detection in word embeddings.

- Indian languages have complex linguistic structures, such as compound words and morphological variations. For example, a river in Hindi has feminine gender. In contrast, words like writer have masculine and feminine counterparts. This gender association affects the pronouns, adjectives, and verb forms used during sentence construction.

- Lack of standard evaluation metrics: There is a lack of standardized evaluation metrics for measuring bias in word embeddings, particularly for Indian languages. This can make it difficult to compare, evaluate and to assess the effectiveness of bias detection and mitigation techniques.

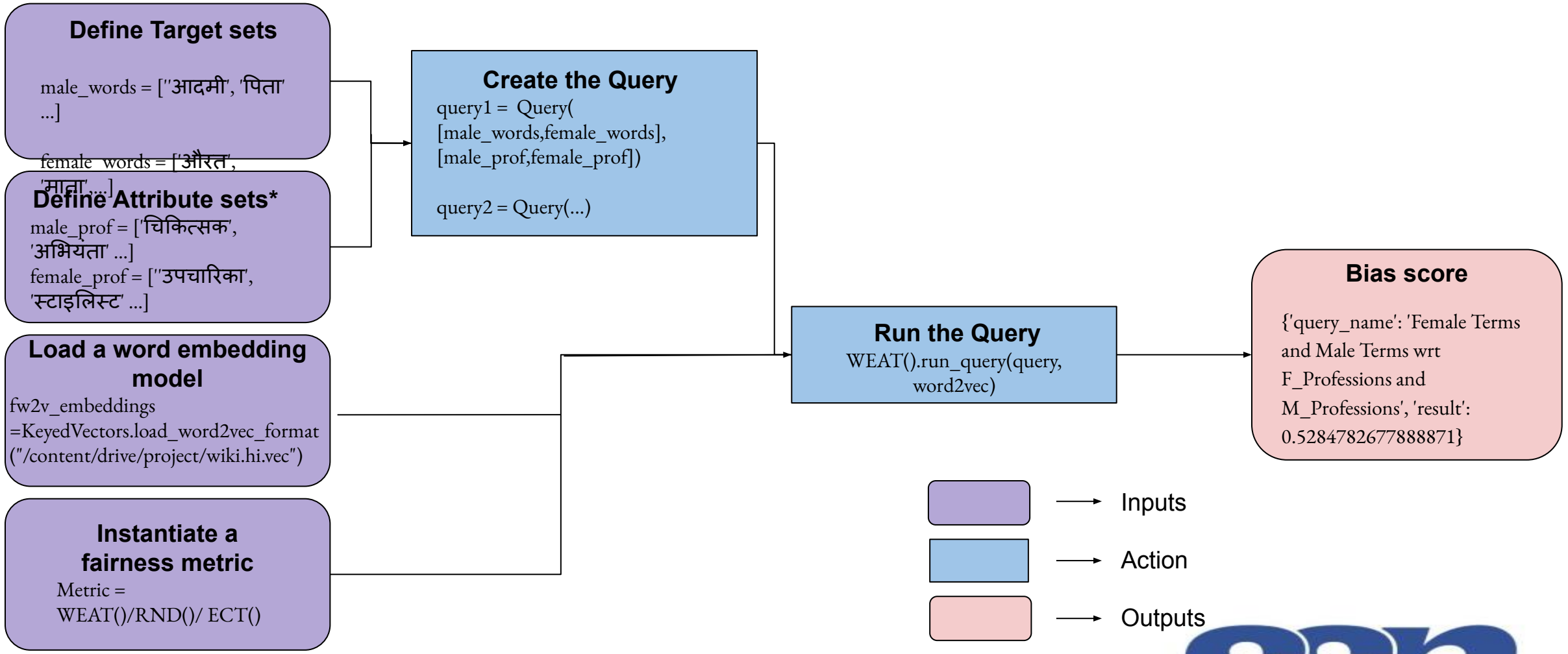
Bias Measurement and Mitigation

- WEFE (Word Embeddings Fairness Evaluation) is a Python library that provides a set of tools for measuring and mitigating bias in word embeddings.

- In bias measurement, WEFE provides a standard interface for:
  - Encapsulating existing fairness metrics.
  - Encapsulating the test words used by fairness metrics into standard objects called queries.
  - Computing a fairness metric on a given pre-trained word embedding model using user-given queries.
- Many evaluation metrics are available in WEFE to measure bias in word embeddings. The metrics chosen to evaluate bias for this project are:
  - WEAT - Word Embedding Association Test Score
  - RND - Relative Norm Distance Score
  - ECT - Embedding Coherence Test Score

- WEFE also standardizes the bias mitigation methods using a *'fit-transform'* interface
  - The first step, fit, learn the corresponding mitigation transformation.
  - The transform method applies the transformation learned in the previous step to words residing in the original embedding space.
- For this project, the Hard debias method has been employed to mitigating biases through geometric operations on embeddings.

Bias Measurement - Standard usage pattern (WEFE)



QUERIES

For the WEAT score calculation, two sets of bias attributes—A and B, are compared to two sets of target words—X and Y.

While studying gender stereotypes, it is expected that those in X would be more similar to terms in A than B, and those in Y would be the opposite.

Target set (X,Y):

- The target set contains a standard set of male terms such as 'आदमी', 'पिता' etc, and female terms such as 'औरत', 'माता'. In tamil the male terms would be 'ஆண்', 'அப்பா' etc, and female terms would be 'பெண்', 'அம்மா'.
- The association of these words with the attribute sets A and B is calculated to check for gender bias.

Attribute set:

- This contains the male and female stereotypical words like ['चिकित्सक', 'अभियंता'] , ['उपचारिका', 'स्टाइलिस्ट'] in hindi and ['மருத்துவர்', 'செவிலியர் '] , ['காவல் ', 'நடனமாடுபவர் '] in tamil.
- Different words are considered for the attribute set based on which association we consider.
- We consider tests probing for associations between gender stereotypes comparing career and family, math and arts, and intelligence and appearance.

Queries

For calculation of RND and ECT, a single bias attribute—A is compared to two sets of target words—X and Y.

While studying gender stereotypes in occupations, it is expected that those in X would be more similar to terms in A, and those in Y would be the opposite.

Target set (X,Y):

- The target set contains a standard set of male terms such as 'आदमी', 'पिता' etc, and female terms such as 'औरत', 'माता'. In tamil the male terms would be 'ஆண்', 'அப்பா' etc, and female terms would be பெண்', 'அம்மா'.
- The association of these words with the attribute set A is calculated to check for gender bias.

Attribute set:

- This contains only female stereotypical words like ['उपचारिका', 'स्टाइलिस्ट'] in hindi and ['மருத்துவர்', 'செவிலியர் '] in tamil.
- Different words are considered for the attribute set based on which association we consider.
- We consider tests probing for associations between gender stereotypes comparing family, arts, and appearance.

WEAT Score

- WEAT receives two sets of target words, and two sets of attribute words and performs a hypothesis test on the following null hypothesis: There is no difference between the two sets of target words in terms of their relative similarity to the similarity with the two sets of attribute words.
- This test measures the degree of association between target words (e.g., female names) and attribute words (e.g., career-related words) by calculating the cosine similarity between their respective embedding vectors. Higher values indicate stronger association and potential bias.
- In formal terms, 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:

$$\mathbf{WEAT}(T_1, T_2, A_1, A_2) = \sum_{x \in T_1} s(x, A_1, A_2) - \sum_{y \in T_2} s(y, A_1, A_2)$$

where

$$s(w, A, B) = \mathbf{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \mathbf{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

$s(w, A, B)$ measures the association of w with the attributes, and $\mathbf{WEAT}(T_1, T_2, A_1, A_2)$ measures the differential association of the two sets of target words with the attribute.

WEAT Score - Effect size

This metric also contains a variant: WEAT Effect Size (WEAT-ES). This variant represents a normalized measure that quantifies how far apart the two distributions of association between targets and attributes are. In practical terms, WEAT Effect Size makes the metric not dependent on the number of words used in each set.

$$\text{WEAT-ES}(T_1, T_2, A_1, A_2) = \frac{\text{mean}_{x \in T_1}\, s(x, A_1, A_2) - \text{mean}_{y \in T_2}\, s(y, A_1, A_2)}{\text{std-dev}_{w \in T_1 \cup T_2}\, s(w, A_1, A_2)}$$

RND score

- Relative Norm Distance is a score that measures the similarity between two groups of words in the embedding space.

$$\text{relative norm distance} = \sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2.$$

- It does so by :
  - Averaging the embeddings of each target set.
  - Then, for each **attribute embedding(v_m)**, calculate the distance between the attribute embedding and the **average of the target 1 (v_1)** and the distance between the embedding and **average of the target 2 (v_2)**; and subtracts them.
  - Finally, it computes the average of the differences of the distances and returns it.

- The available distances are the the difference of the normalized vectors (‘norm’) and the cosine distance (‘cos’).

ECT score

- Embedding Coherence Test Score: This score measures the degree to which a word embedding exhibits equal similarity between two groups of words with respect to a given concept (e.g., occupation).

- It calculates the average target group vectors, measures the cosine similarity of each to a list of attribute words and calculates the correlation of the resulting similarity lists.

    - Embed all given target and attribute words with the given embedding model.
    - Calculate mean vectors for the two sets of target word vectors.
    - Measure the cosine similarity of the mean target vectors to all of the given attribute words.
    - Calculate the Spearman r correlation between the resulting two lists of similarities.
    - Return the correlation value as score of the metric

Interpretation Of Scores

WEAT

- The range of WEAT effect size is from -2 to 2, where a score of 0 indicates no association between the target and attribute word groups, and scores closer to -2 or 2 indicate stronger associations.
- A positive effect size confirms the hypothesis that words in X are rather stereotypical for the attributes in A and words in Y stereotypical for words in B, while a negative effect size indicates that the stereotypes would be counter-wise. The magnitude of the score reflects the strength of the association, with larger magnitudes indicating stronger association.

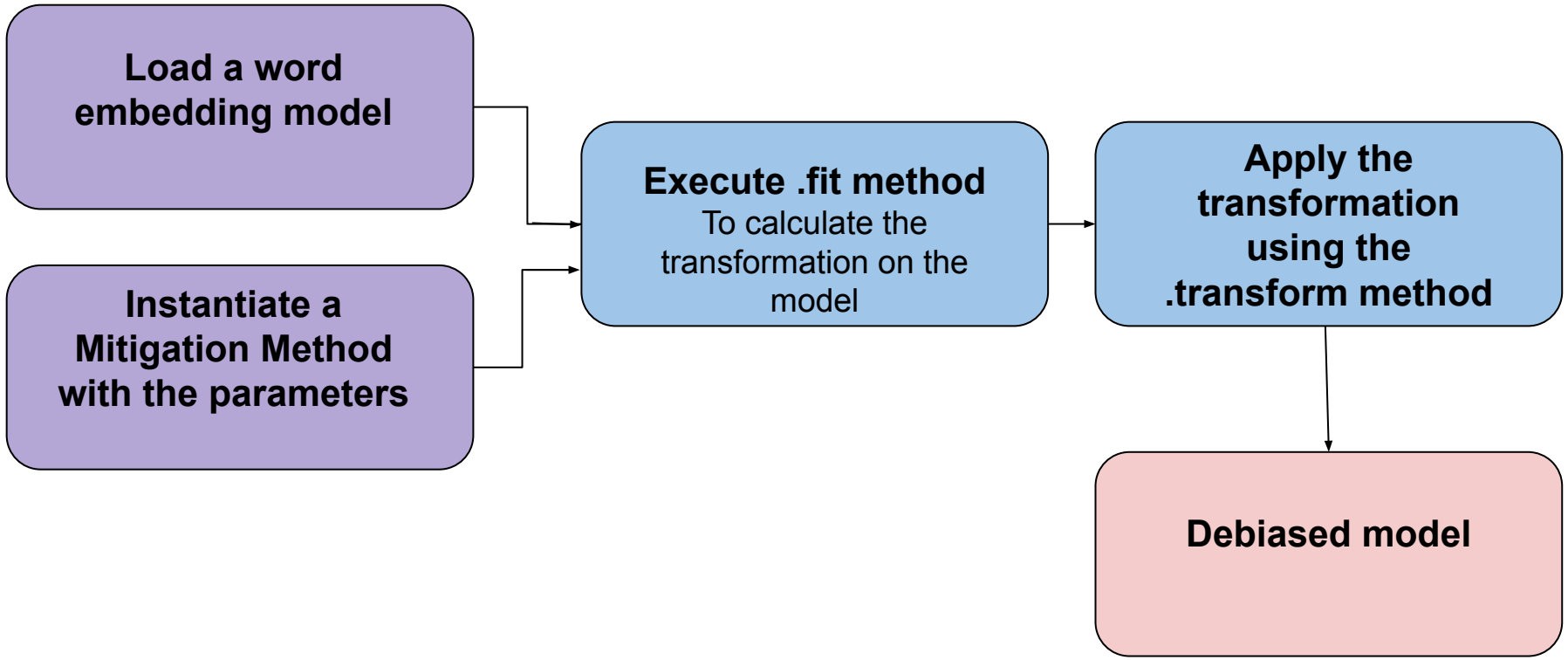
RND

- The more positive (negative) the relative distance from the norm, the more associated are the sets of attributes towards group two (one).
- A RND score of 0 indicates that there is no difference in similarity between the two groups of words in the embedding space.
- Scores greater than 0 indicate greater similarity within groups.

ECT

- The ECT score ranges from -1 to 1, where a score of 1 indicates no bias, and a score lesser than 1 indicates bias.
- A lower ECT score indicates a higher degree of bias in the embedding space, while a higher score indicates a higher degree of equalization with respect to a specific concept or attribute.
- That is values closer to 1 are better as they represent less bias.

DEBIASING PROCESS



HARD DEBIAS

Hard Debias debiasing method. This method allows mitigating biases through geometric operations on embeddings. This method is binary because it only allows 2 classes of the same bias criterion, such as male or female.

- ➢ The hard debiasing technique typically involves identifying a set of biased words, which are words that are strongly associated with certain attributes or concepts in the embedding space. These words are then modified so that their embedding vectors are adjusted in a way that reduces or eliminates the bias.
- ➢ The main idea of this method is:

1. Identify a bias subspace through the defining sets.
2. Neutralize the bias subspace of embeddings that should not be biased. First, it is defined a set of words that are correct to be related to the bias criterion: the criterion specific gender words.

   Then, it is defined that all words outside this set should have no relation to the bias criterion and thus have the possibility of being biased. Therefore, this set of words is neutralized with respect to the bias subspace found in the previous step.

   The neutralization is carried out under the following operation:

   - U : embedding
   - V : bias direction

   First calculate the projection of the embedding on the bias subspace :

$$\text{bias subspace} = \frac{v \cdot (v \cdot u)}{(v \cdot v)}$$

Then subtract the projection from the embedding :

$$u' = u - \text{bias subspace}$$

3. Equalizate the embeddings with respect to the bias direction. Given an equalization set this step executes, for each pair, an equalization with respect to the bias direction.

   That is, it takes both embeddings of the pair and distributes them at the same distance from the bias direction, so that neither is closer to the bias direction than the other.

Inferences:

- ➢ After Debiasing, the WEAT score should be optimised to a value closer to 0.
- ➢ After Debiasing, the RND score should be optimised to a value closer to 0.
- ➢ After Debiasing, the ECT score should be optimised to a value closer to 1.

Word Vectors - Hindi

For the Hindi language, 2 word embedding models from FastText have been considered:

(i) wiki.hi.vec

- The wiki.hi.vec dataset is a pre-trained word embedding model for the Hindi language. It is provided as part of the FastText library, which is a popular open-source library for learning word embeddings and text classification.
- The word embeddings in wiki.hi.vec were learned from a large Hindi corpus of Wikipedia articles, using FastText's skip-gram algorithm. The resulting model contains 300-dimensional vectors for 332,647 unique words in the Hindi language, including both words and subwords. Each vector represents the distributional semantics of the corresponding word, capturing its meaning and relationships with other words in the language.
- The wiki.hi.vec file is in the plain text format, where each line represents a single word and its corresponding embedding vector. The first token on each line is the word itself, followed by its embedding vector components separated by spaces.
- The file can be loaded into memory using a variety of programming languages and used as a lookup table for obtaining the embedding vectors for specific words in a given text corpus or for evaluating the bias of the word embedding model.

Word Vectors - Hindi

(ii) cc.hi.300.vec

- The cc.hi.300.vec dataset is a pre-trained word embedding model for the Hindi language provided by the FastText library.
- Unlike the wiki.hi.vec model, which was trained on a single corpus of Wikipedia articles in Hindi, the cc.hi.300.vec model was trained on a much larger corpus of text in Hindi. Specifically, it was trained on a snapshot of the Common Crawl, a large web corpus that contains data from websites all over the world.
- The model was trained using FastText's skip-gram algorithm, and each embedding vector in the model has 300 dimensions.
- The file contains embedding vectors for 2.7 million unique words in Hindi, including both words and subwords.
- The cc.hi.300.vec dataset is generally considered to be a high-quality pre-trained word embedding model for Hindi, and it has been used in a wide range of natural language processing tasks, including text classification, sentiment analysis, and machine translation.

Word Vectors - Tamil

For the Tamil language, 2 datasets have been considered:

(i)wiki.ta.vec

- The wiki.ta.vec dataset is a word embedding model trained using FastText on the Tamil Wikipedia corpus.
- The Tamil Wikipedia corpus is a collection of articles written in the Tamil language. FastText is an NLP library developed by Facebook that can learn word embeddings from large corpora of text.
- The wiki.ta.vec dataset contains word vectors for over 240,000 words in the Tamil language. Each vector has 300 dimensions and is trained using the skip-gram model with negative sampling. The dataset is publicly available and can be downloaded from the FastText website.
- Similar to the wiki.hi.vec dataset the first token on each line is the word itself, followed by its embedding vector components separated by spaces The first token on each line is the word itself, followed by its embedding vector components separated by spaces.

Word Vectors - Tamil

ii) cc.ta.300.vec

- The cc.ta.300.vec dataset is a pre-trained word embedding model for the Tamil language, trained using the FastText algorithm on the Common Crawl corpus.
- The main difference between the cc.ta.300.vec and wiki.ta.vec datasets is the corpus on which they trained .
- The cc.ta.300.vec dataset is trained on the Common Crawl corpus, which is a massive collection of web pages and documents from around the world. This means that the dataset contains a wide range of text written in Tamil from different domains and sources.
- On the other hand, the wiki.ta.vec dataset is trained on the Tamil Wikipedia corpus, which is a collection of articles written in the Tamil language. This means that the dataset is more focused on the domain of encyclopedic knowledge and may not cover as much diversity in language use as the cc.ta.300.vec dataset.

Results - Tamil

Tamil - Wiki Word Embedding Results

RESULTS							
Target Words	Attribute words	Biased Results			Debiased Results		
		WEAT	RND	ECT	WEAT	RND	ECT
Male vs Female	Career vs Family	0.327	-0.036	0.657	0.133	-0.010	1.0
Male vs Female	Professions	0.035	0.097	0.3	0.003	-0.001	0.9
Male vs Female	Maths vs Arts	0.149	0.198	0.500	0.104	-0.001	0.666

Tamil - Crawl + Wiki Word Embedding Results

RESULTS							
Target Words	Attribute words	Biased Results			Debiased Results		
		WEAT	RND	ECT	WEAT	RND	ECT
Male vs Female	Career vs Family	0.343	0.045	0.828	0.123	0.006	0.828
Male vs Female	Professions	0.295	0.017	0.8	-0.007	0.001	1.0
Male vs Female	Maths vs Arts	0.130	0.028	0.619	0.021	-0.006	0.952

Inferences

- TAMIL WIKI
  - WEAT shows bias in all 3 sets aligned with the hypothesis with highest bias for Family vs Career Attribute.
  - ECT too shows bias for all 3 attributes towards the female gender(as per hypothesis) but shows highest bias for the Professions stereotypically associated to the female gender. attribute.
  - RND scores on the other hand prove only one hypothesis right in terms of gender association but still indicate a degree of bias.

- TAMIL WIKI+CC
  - WEAT shows bias in all 3 sets aligned with the hypothesis with highest bias for Family vs Career attribute, the same previously observed for the wiki word vectors.
  - ECT on the other hand shows very negligible bias on the whole for the 3 sets chosen.
  - RND scores disprove the assumed hypothesis and indicate that words under the 3 sets are related to the male gender rather than female gender.

Performance Improvement after Debiasing - Tamil

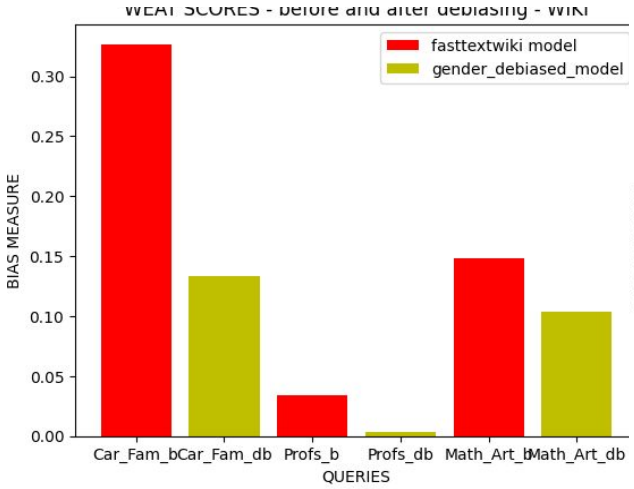
Tamil - Wiki : Percentage Decrease in bias

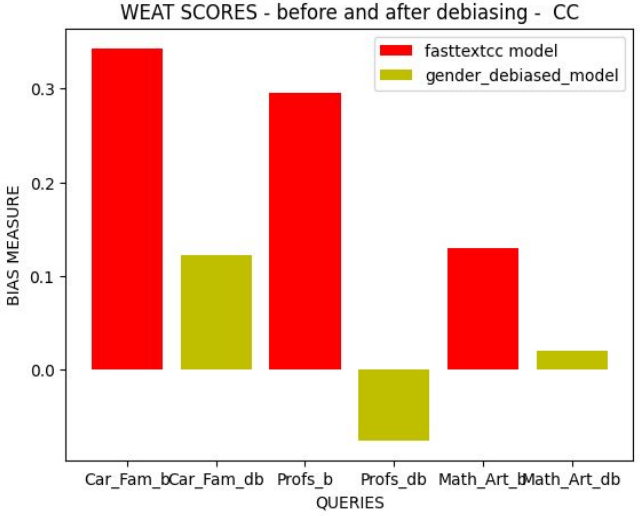
Percentage Decrease in Bias measured against each Score					
WEAT	Percentage Decrease - WEAT	RND	Percentage decrease - RND	ECT	Percentage Decrease - ECT
Career vs Family	59 %	Family	66%	Family	53%
Professions	90%	Female Professions	98%	Female Professions	200%
Math vs Arts	30%	Appearance	99%	Appearance	32%

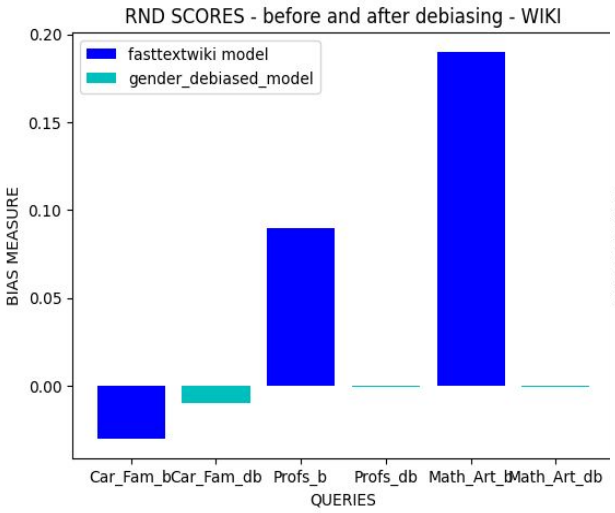
Tamil - Crawl + Wiki : Percentage Decrease in bias

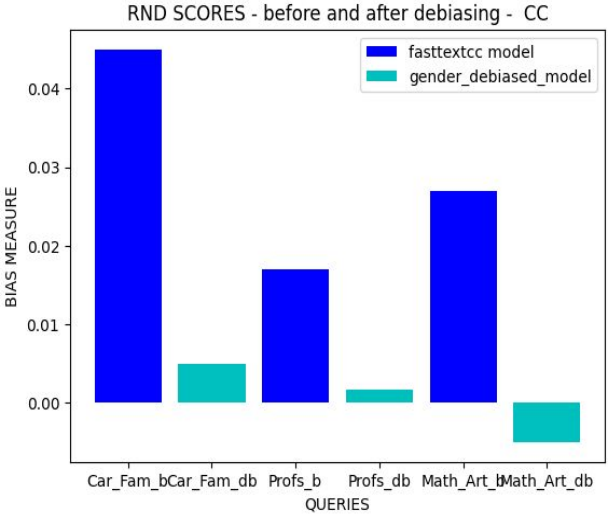
Percentage Decrease in Bias measured against each Score					
WEAT	Percentage Decrease - WEAT	RND	Percentage decrease - RND	ECT	Percentage Decrease - ECT
Career vs Family	64 %	Family	88%	Family	0%
Professions	96%	Female Professions	90%	Female Professions	25%
Math vs Arts	84%	Art	82%	Art	55%

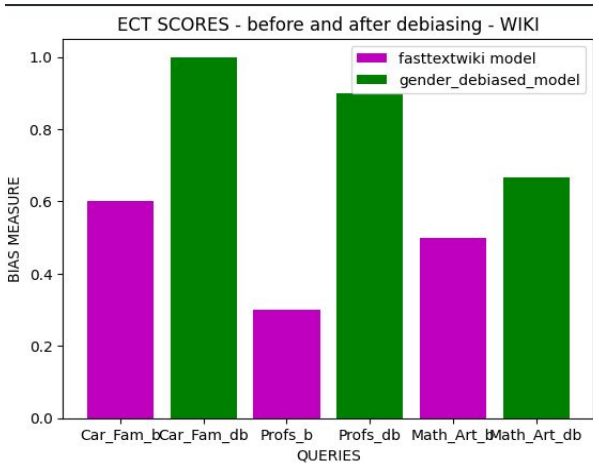
Visual Representation - Tamil

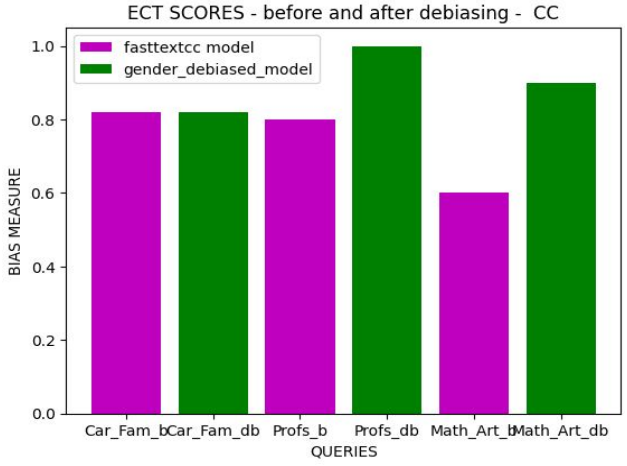












Results

Hindi - Wiki Word Embedding Results

RESULTS							
Target Words	Attribute words	Biased Results			Debiased Results		
		WEAT	RND	ECT	WEAT	RND	ECT
Male vs Female	Career vs Family	0.393	-0.023	0.771	0.182	-0.014	0.657
Male vs Female	Professions	0.528	-0.077	0.7	0.362	-0.042	1.0
Male vs Female	Intelligence vs Appearance	0.241	0.098	0.868	0.056	0.004	0.901

Hindi - Crawl + Wiki Word Embedding Results

RESULTS							
Target Words	Attribute words	Biased Results			Debiased Results		
		WEAT	RND	ECT	WEAT	RND	ECT
Male vs Female	Career vs Family	0.327	0.042	0.543	0.04	0.005	0.886
Male vs Female	Math vs Arts	0.045	0.036	0.857	0.002	0.012	0.964
Male vs Female	Intelligence vs Appearance	0.247	0.034	0.235	0.017	0.013	0.952

Inferences

- HINDI WIKI
  - WEAT shows bias in all 3 sets aligned with the hypothesis with highest bias for Professions attribute
  - ECT too shows a small bias for all 3 attributes towards the female gender(as per hypothesis) but shows highest bias for the Female stereotyped Professions attribute.
  - RND scores too indicate the highest bias for Female stereotyped Professions attribute. But also proves the assumed hypothesis wrong for the Appearance bias.

- HINDI WIKI+CC
  - WEAT shows a noticeable bias in Career vs Family and Intelligence vs Appearance attribute except the Professions attribute.
  - ECT scores for the words related to Arts show the least bias amongst the 3 queries with Appearance yielding a high bias.
  - The results of each of the queries are all marginally positive, which goes against our theory and the findings of the other scores and suggests that these attributes may be more biased towards the male gender than the female gender.

Performance Improvement after Debiasing - Hindi

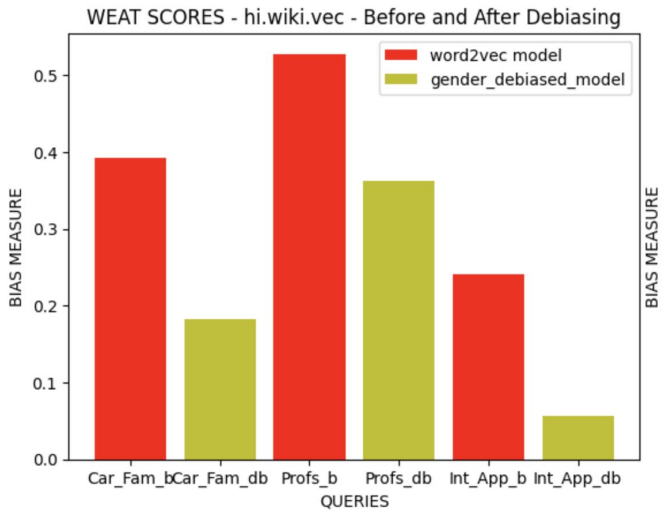
Hindi - Wiki : Percentage Decrease in bias

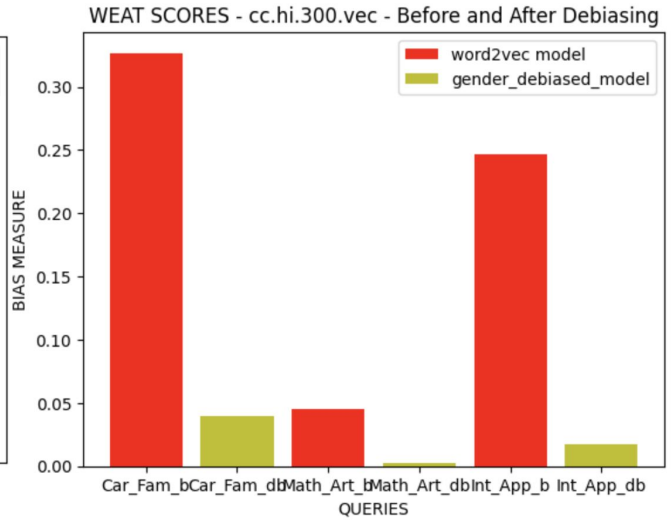
Percentage Decrease in Bias measured against each Score					
WEAT	Percentage Decrease - WEAT	RND	Percentage decrease - RND	ECT	Percentage Decrease - ECT
Career vs Family	64 %	Family	88%	Family	0%
Professions	96%	Female Professions	90%	Female Professions	25%
Math vs Arts	84%	Art	82%	Art	55%

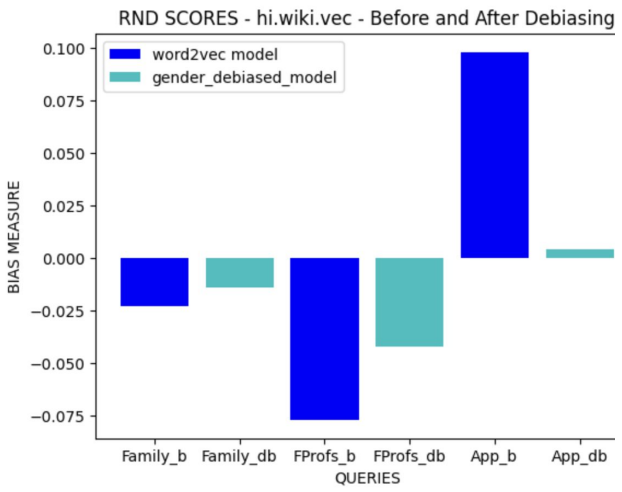
Hindi - Crawl + Wiki : Percentage Decrease in bias

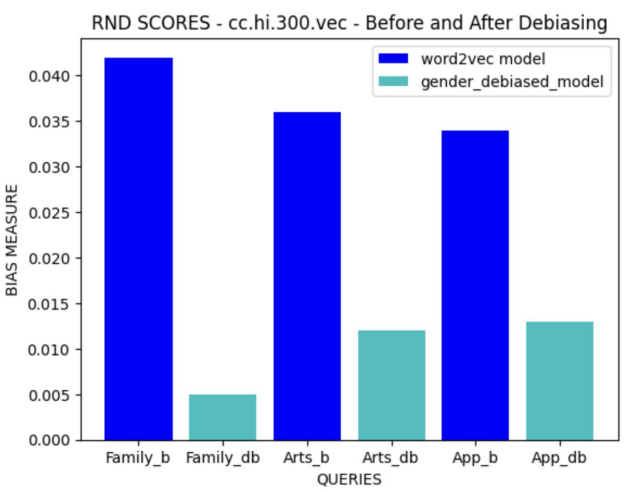
Percentage Decrease in Bias measured against each Score					
WEAT	Percentage Decrease - WEAT	RND	Percentage decrease - RND	ECT	Percentage Decrease - ECT
Career vs Family	53 %	Family	39%	Family	-14%
Professions	31%	Female Professions	45%	Female Professions	42%
Intelligence vs Appearance	76%	Appearance	42%	Appearance	3%

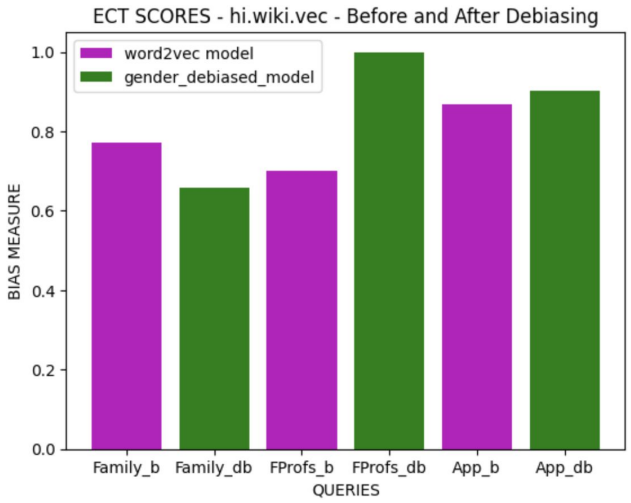
Visual Representation - Hindi

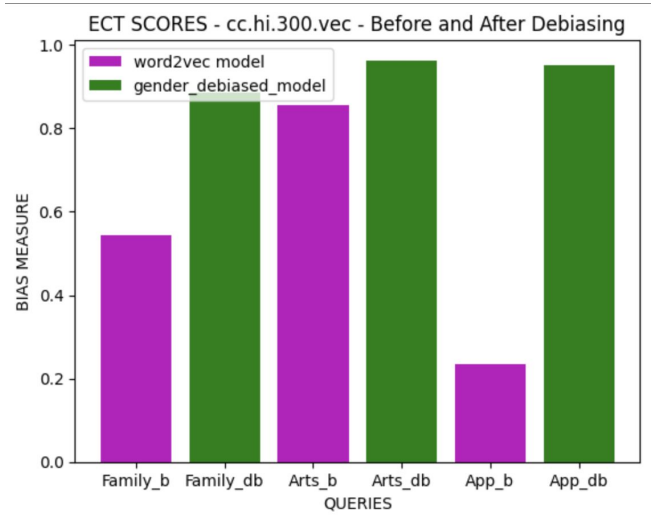












Conclusion and Future Work

- In this project, we examined gender stereotypes in the word embeddings of Tamil and Hindi languages. Evaluation metrics like the WEAT score, RND metric and ECT metric were applied in the gender subspace to quantify the bias in the embeddings of these languages. The Hard Debias method was employed as the debiasing technique. The findings indicated that both languages exhibit gender bias, and efforts to mitigate this bias has been successful as seen by a decline in bias.

- Future work could include trying out these techniques on downstream tasks and checking the performance before and after debiasing.
- An extrapolation of bias detection in word embeddings would be to measure bias in contextual data, with the help of sentence embeddings and a particular SEAT (Sentence Encoder Association Test) measure that was developed for this purpose.
- Other bias measurement and bias mitigation algorithms can be explored and implemented. Results can be compared to the existing results.
- Other known embedding models can be used to quantify bias. Metrics used with FastText embedding model, can be used on other models. The corresponding results for all metrics across all models can be recorded and compared.
- This project also has the scope to be extrapolated to all the other 121 languages spoken in India.

References

1. Bolukbasi, Tolga & Chang, Kai-Wei & Zou, James & Saligrama, Venkatesh & Kalai, Adam. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.
2. Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186.
3. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
4. Hasanuzzaman, M., Dias, G., and Way, A. (2017). Demographic word embeddings for racism detection on Twitter. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 926–936, Taipei, Taiwan, November.
5. Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.
6. Balasubramanian, Senthil Kumar & Tiwari, Pranav & Kumar, Aman & Aravindan, Chandrabose. (2022). Casteism in India, But not Racism -A Study of Bias in Word Embeddings of Indian Languages.
7. Joel Escude´ Font and Marta R. Costa-jussa`. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 147–154, Florence, Italy. Association for Computational Linguistics.
8. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin. Advances in Pre- Training Distributed Word Representations

9. Poeschko, T., & Schwarz, C. (2020). Word Embeddings as a Tool for Analyzing Gender Bias in German. In Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations (pp. 240- 244).
10. Manzini, T., Yao Chong, L., Black, A. W., and Tsvetkov, Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass 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)
11. 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), pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.
12. Bansal, S., Garimella, V., Suhane, A., and Mukherjee, A. (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, page 27–34, New York, NY, USA. Association for Computing Machinery.
13. 14. Dev, S., & Phillips, J. (2019, April). Attenuating Bias in Word vectors.
14. P . Badilla, F. Bravo-Marquez, and J. Pe´rez WEFE: The Word Embeddings Fairness Evaluation Framework In Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI 2020), Yokohama, Japan.
15. Garg et al., 2018]Nikhil Garg, Londa Schiebinger, Dan Ju-rafsky, and James Zou. Word embeddings quantify 100years of gender and ethnic stereotypes. Proceedings of theNational Academy of Sciences, 115(16):E3635–E3644,2018.
16. Arjun Sai Krishnan and Seyoon Ragavan. 2021. Morphology-Aware Meta- Embeddings for Tamil. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 94–111, Online. Association for Computational Linguistics.
17. Neeraja Kirtane and Tanvi Anand. 2022. Mitigating Gender Stereotypes in Hindi and Marathi. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 145–150, Seattle, Washington. Association for Computational Linguistics.