Detection and Mitigation of Bias

Final Project: COL707 - Sem II 2022-23

Date: 10 May 2023

Students:

Vatsal Jingar	Stitiprajna Sahoo	Mayank Mangla
2020CS50449	2020CS10394	2020CS50430

1 ABSTRACT:

In this project we have tried to implement and analyse different state of the art bias detection methods. Some of the popular bias detection methods are **WEAT** (Word Embeddings Association Test), **Log Probability Score**, and **Stereo-Set**. After this, we have also tried to implement a debiasing method **"Hard De-Bias"** on some of the popular embeddings. To test the debiased embeddings, we have implemented a counterfactual model. In this project, we are mostly concerned with Gender Bias. All the implementations could be extended to multi class to tackle other racism and religion based biases.

1.1 Contents :

2 MEASURING BIAS:

2.1 WEAT : Implementation and Analysis

Word Embedding Association Test :

The test tries to study the relation of certain word embeddings with target word embeddings. Discussing about the test, we should first understand the meaning of Word embeddings. To represent the word of any language i.e (English, Japanese) certain specific vector space is used. Here in this space every word is mapped to certain vector. The two defined properties for this space are :

1. The words with similar meaning should have similar vectors. Here similar means that there dot product should be close to 1.

2. Second the words which associate the relationship should be shown by vectors. Mathematically, the difference of word embedding of man and woman should be similar to difference of King and Queen because they depict same relationship.

These word embeddings can be sentence dependent too. This would be beneficial when a same word could have different semantic meaning when used in different sentences.

The word embedding test tries to see the association of a particular set of words with a two different attribute sets. Here, the attribute set refers to a set of gender specific words. So we could have one attribute set as all the words related to male like boy, uncle, he, his and another attribute set as words related to female like girl, aunt, she, her.

Now for this test, we will compare the association of these two sets with a specific set of words. This set is referred to as **Target words set**. The target words generally contain gender neutral words. A good embedding space would have difference of association of the two attribute sets with targets as close to 0.

Implementation : The above section is mathematically interpreted here.
Consider the set of target words as T_1 and T_2. And attribute sets as M and F. M would contain male words and F would contain female words. We would first calculate the difference of association of M and F with T_1 and then with T_2. After that the final score would be difference of these associations. We choose T_1 as a set with male stereotype words and T_2 with female stereotype words.
Difference of association of a word x from set M and F can be written as
$s(x, A, B) = mean_{\forall m \epsilon M}(dotproduct(x, M)) - mean_{\forall f \epsilon F}(dotproduct(x, F))$
Further we will use this in WEAT as :

$$WEAT(T_1, T_2, M, F) = mean_{t_1 \epsilon T_1}(s(t_1, M, F)) - mean_{t_2 \epsilon T_2}(s(t_2, M, F))$$

Analysis:
To study the trends of weat scores with different corpus and dimensions, we have used the state of art algorithm of learning embeddings : "Glove"[6]. We have used two different corpus for learning embeddings with Glove : "wiki-gigaword"[3], "twitter"[4].

For the purpose of calculating the average trend, we have performed some WEAT tests with different target sets. For comparing the results, we have normalized the WEAT scores. We will call them normalized WEAT scores. For the normalization, we have used standard technique used in [5].

Normalization of WEAT score : For this purpopse, we will divide the difference of association of target word with standard deviation of all the cosine dot products, i.e:,

$$s(w, A, B)_{\mathcal{N}} = \frac{s(w, A, B)}{std\_dev_{\forall x \epsilon A \cup B}(cos(w, x))}$$

Now we will use this new difference of association of finding the score. Other steps are same as before.
Tests we performed :
1. WEAT 1 (Target Sets : Math and Arts, Attribute Sets : Male and Female)
The set of words are chosen according to [1].
Math = [math, algebra, geometry, calculus, equations, computation, numbers, addition]
Arts = [poetry, art, dance, literature, novel, symphony, drama, sculpture]
Male = [male, man, boy, brother, he, him, his, son]
Female = [female, woman, girl, sister, she, her, hers, daughter]

2. WEAT 2 (Target Sets : Science and Arts, Attribute Sets : Male and Female)
The set of words are chosen according to [1].
Science = [science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy]
Arts = [poetry, art, Shakespeare, dance, literature, novel, symphony, drama]
Male = [brother, father, uncle, grandfather, son, he, his, him]
Female = [sister, mother, aunt, grandmother, daughter, she, hers, her]

The sets for above two tests were chosen from standard WEAT test done in [5]. But to find a more general and robust trend we have performed overall 6 WEAT tests. All these tests and their plots are added in Appendix section.

WEAT scores for Test1 :

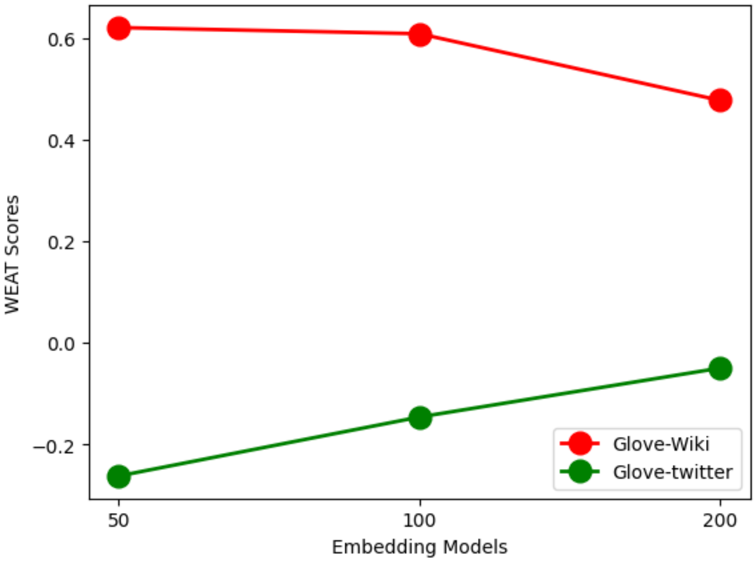

Fig. 1. WEAT scores v/s Dimensions, Test-1

WEAT scores for Test2 :

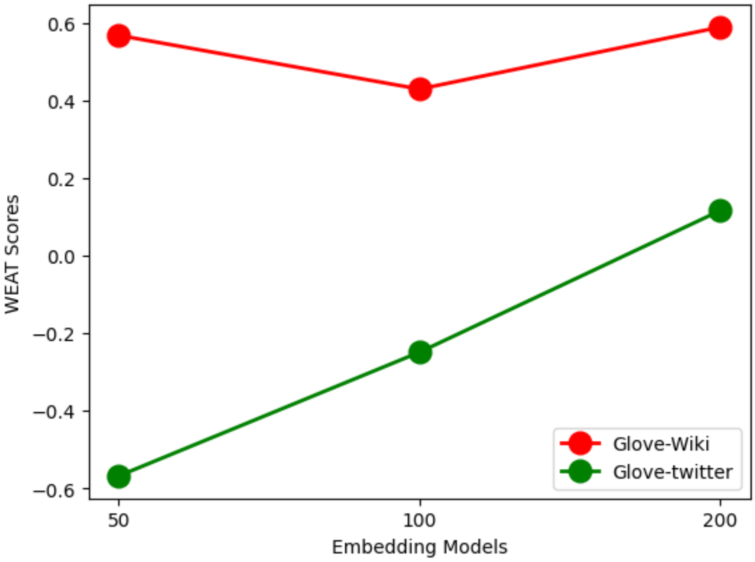

Fig. 2. WEAT scores v/s Dimensions, Test-2

By aggregating results from all the six tests (averaging out the weat scores), the trends we got were:

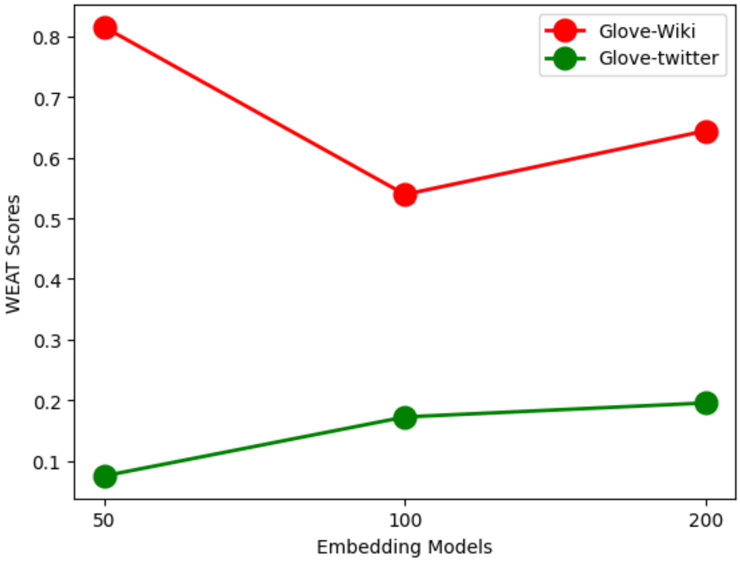

Fig. 3. Weat Scores v/s Dimensions of Embedding : Trends for two different corpus

Observation : WEAT scores shows the difference of associations, so more closer it is to 0, more the embeddings and less biased.
From the plot, it is visible that increasing the dimensions of word embedding would change the bias of embedding space. In general, the scores would also vary with different tests. But after aggregating all the results, we found that the weat scores are optimal for dimension = 100 when wiki-gigaword is chosen as corpus while for twitter corpus, the weat score is increasing with dimensions (Although for some tests, they actually move towards 0, so for some tests, higher dimensions were more neutral). Also we can see that the scores are farther from 0 for wiki-gigaword than twitter. This shows that the weat scores are also dependent on the corpus which is used for learning the embeddings. As we can see that by only changing the corpus used for learning the embeddings, bias increased, we conclude that the corpus for the 1st case is itself has a bias in it.
The reason when we tried to search for high bias in wiki-gigaword corpus was the difference in number of pages which are dedicated to male than female.
Total male pages were : 576,106 and Total female pages were 115,941.

Secondly we also tried to observe that Is the weat scores are related real life associations of the target words with the attribute set. For this, we tried to find a trend between the association of a word which would be a job with number of female employees working within that job.

We used these jobs for finding a trend : [maid, babysitter, nurse, cook, doctor, journalist, prisoner, professor, politician]

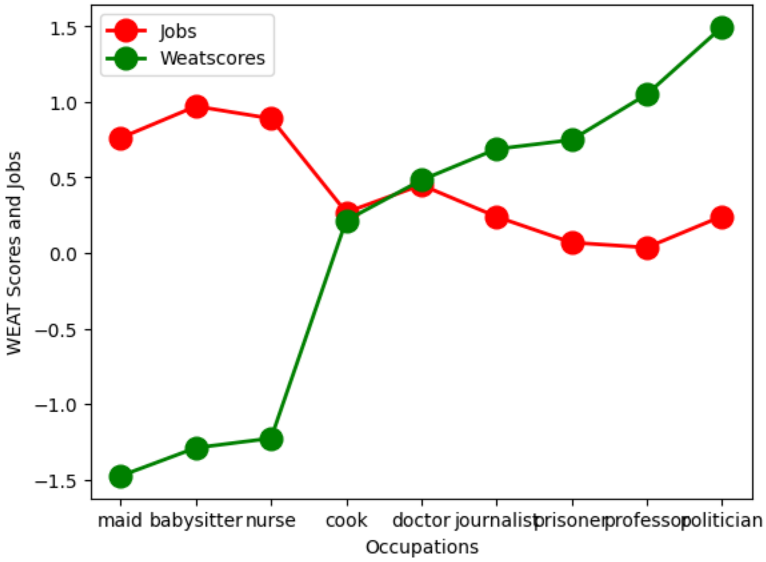

Fig. 4. Trend of jobs wrt Weatscores

For this curve, to look the trend we have scale down percentage participation in job at scale of 0 to 1. From the plot, it is visible that the fields in which less percent of women works have more weat score and where the percent working of women is higher the weat scores are negative. More weat scores depicts that the job is more associated to male side than female. While the fields where the number of working women percent is higher the weat scores negative, this shows that embeddings are more associate to women. This trend shows that how embeddings has absorbed a real bias in it.

2.2 Log probability Bias score

Background and Motivation : We saw that the weat score tries to find a static association of words with the target sets. But it is not necessary that the way a word associates with the target word would always remain the same. As in the case of embeddings which are output by State of Art NLP architecture BERT (BiDirectional Encode Representations from Transformers) are **context dependent**. So the word embedding for a same word could be different in different sentences. A trivial example would be that the word "bat" could be an animal in one case and sports equipment in another. Now as the word change their embedding depending on context, we can't measure difference of associations statistically. This motivate us to take a different approach to actually find the Bias in any Large Language Model. In this approach, we will find a **difference in probability of occurrence of male attribute word and female attribute word in a sentence**.

We use log probability bias score in pre-trained BERT models. The log probability bias score gives the measure of association of a certain word to a context. To measure this, we directly query the model input sentence is **[MASK] is [ATTRIBUTE]**. Here mask refers gendered words like "He/She", and attribute refers to contextual words, which can have bias towards a specific gender.

The masked bert model outputs the probability of [MASK] = He or [MASK] = She. This is called ***target probability***. To compute the association of "He" with the [ATTRIBUTE], we need to know how much the model prefers "He" with [ATTRIBUTE] more than the model's preference for "He" in general.

This ***prior probability*** of the word "He" in the model is measured by giving input sentence as [MASK] is [MASK]. The association of "He" with [ATTRIBUTE] will be the ratio = $\frac{p_{target}}{p_{prior}}$. Similarly calculate the association of "She" with [ATTRIBUTE].

Finally, the log probability bias score of the model is calculated in the following steps:

- Calculate the mean association of "He" with male biased attributes = Am_{he}
- Calculate the mean association of "She" with male based attributes = Am_{she}
- Log Probability Bias Score of "He" with the male biased attribute list is $Am_{he} - Am_{she}$. This is a measure of bias that is inclined towards male in the model = $lpbs_{he}$
- Similarly the mean association of "She" with female biased attributes = Af_{she}
- Calculate the mean association of "He" with female biased attributes = Af_{he}.
- Log Probability Bias Score of "She" with the female biased attribute list is $Af_{she} - Af_{he}$. This is a measure of bias that is inclined towards female in the model = $lpbs_{she}$.

Note that the **lpbs score** that is obtained from all the 5 models are normalised via softmax function. In the following section, we show the results that we obtained via this method of bias detection in our experiments.

Implementation:

- We queried 5 different types of pre trained BERT models: `bert-base-cased, bert-base-uncased, bert-large-cased, distilbert-base-uncased, albert-base-v2`. All these libraries were dirrectly imported from hugging face transformers
- We experimented with 3 types of input queries:
  - template 1: [MASK] is a [ATTRIBUTE]
  - template 2: [MASK] is interested in becoming [ATTRIBUTE]
  - template 3: [MASK] likes [ATTRIBUTE]

  These template of sentences was picked from the research paper [2] that was used as a reference in our project.
- The list of 30 attributes of each type (male biased and female biased) that we used in our experiment were collected from the internet from various websites and is provided in *Appendix B.1*. These words were single-word professions, and are generally considered to have bias towards a particular gender. We tried to search for a proper dataset containing these words, but there was no public dataset available on the internet which contained such words for research purpose. The previous research paper had used a public dataset which is no longer available in the internet. So we were bound to use these set of attributes.

The next section will show the observations and analysis of the experiments performed to measure bias in different models via log probability bias score.

Observations and Analysis:

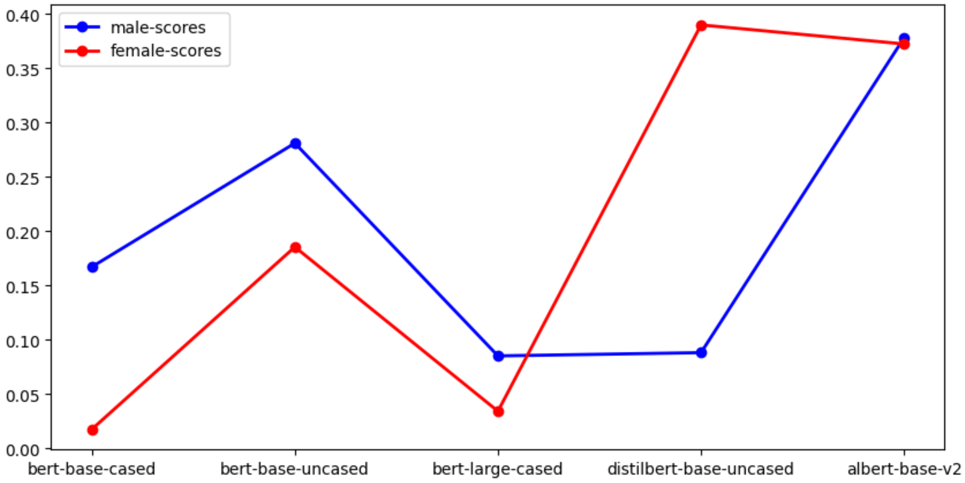

Fig. 5. LPBS scores for both gender vs 5 different models for input queries ∈ template 1

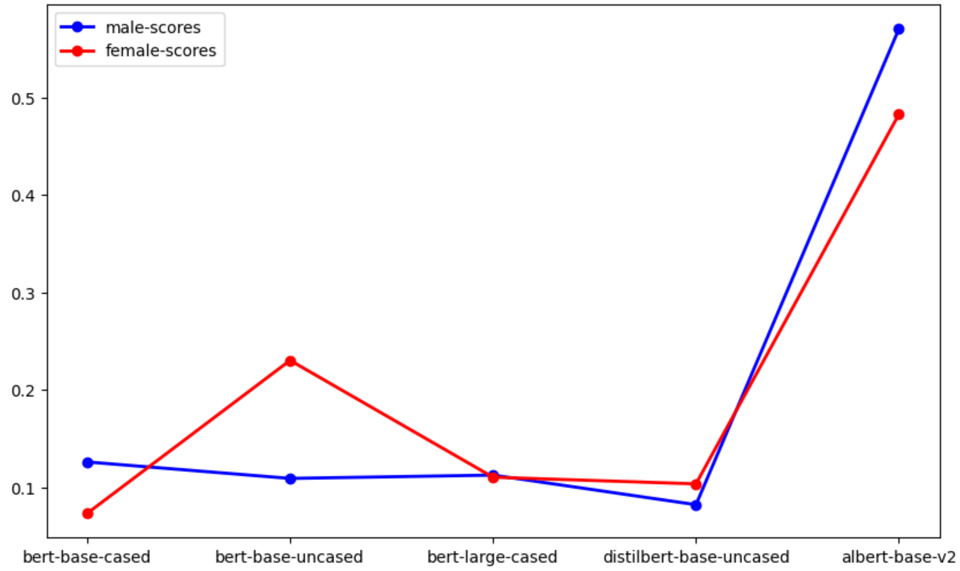

Fig. 6. LPBS scores for both gender vs 5 different models for input queries ∈ template 2

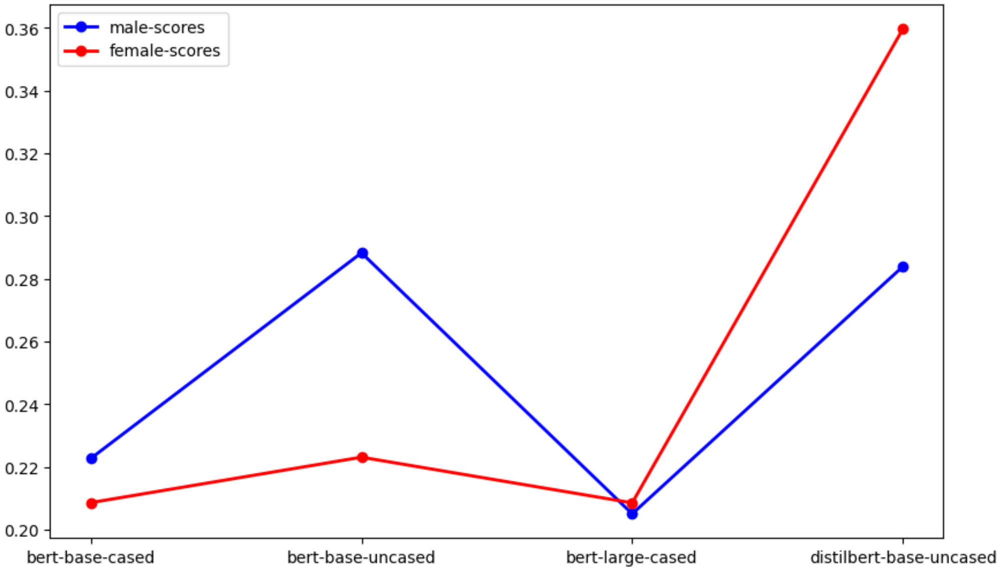

Fig. 7. LPBS scores for both gender vs 5 different models for input queries ∈ template 3

BERT models are trained on large corpus. The BERT `base-cased` and `base-uncased` models are trained on plain text corpus (namely the entirety of the English Wikipedia, and the Brown Corpus). However, the BERT model `albert-base` and `distilbert-base` are trained on the same dataset (entire English Wikipedia + the Brown Corpus) but have much lesser parameters than bert-base models.

- LPBS scores (for both male and female attributes) in `albert-base-v2` is higher for both template 1 and 2 as input sentences. `Albert-base-v2` uses parameter sharing and factorization techniques to reduce the number of parameters and make the model more efficient. In terms of accuracy, it outperforms `bert-base` models on many natural language processing tasks, while using fewer parameters and requiring less computation time. This is a small implication of the *accuracy-fairness tradeoff*, as although the model is claimed to be the best in terms of accuracy, it is highly biased.
- The figures 4 and 6 above also show a fact that uncased models (`bert-base-uncased` and `distilbert-base-uncased`) models show larger bias in the first and third template inputs. This can account to the treatment of uncased models to all words as lower-cased, which might result in difficulties in detection of common nouns or proper nouns for which the probability of the word "he" (pronoun) in place of mask can be larger. For larger sentence however there is less bias even for uncased models.
- Figures 5 and 6 show that the scores follow a similar trend across different types of input templates. Longer sentences do not have a significant effect on the amount of bias.

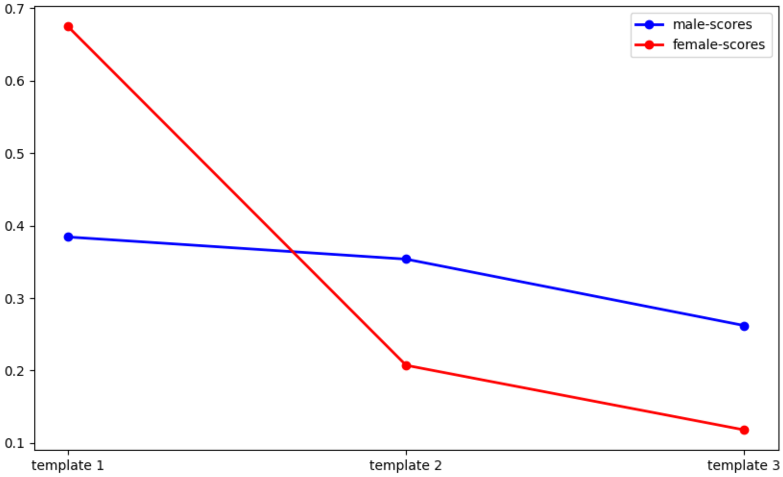

Fig. 8. LPBS scores for both gender vs 3 different templates of input query for model `bert-base-cased`

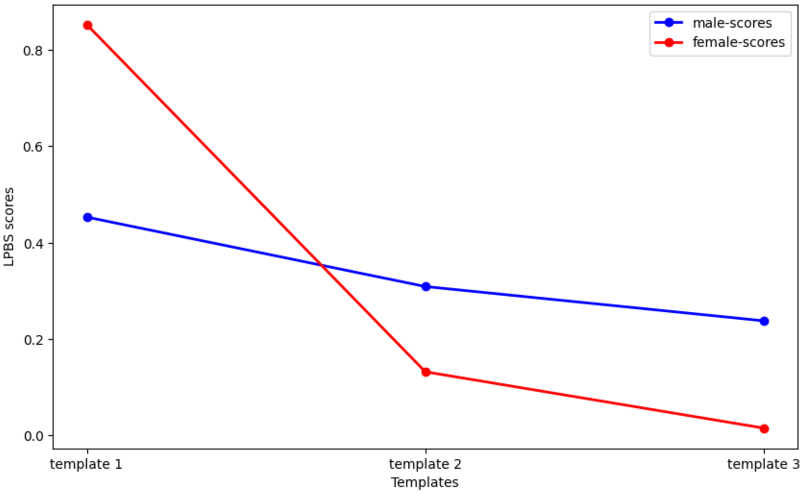

Fig. 9. LPBS scores for both gender vs 3 different templates of input query for model `bert-base-uncased`

- Plots of the LPBS scores vs template sentences given as input to the BERT models show that the longer sentences like *"He is interested in becoming Programmer"* and shorter sentences like *"He likes programmer"* do not have more bias incorporated in them.
- However the sentence template "He is programmer" has higher bias for both the models tested, as per figure 7 and 8.

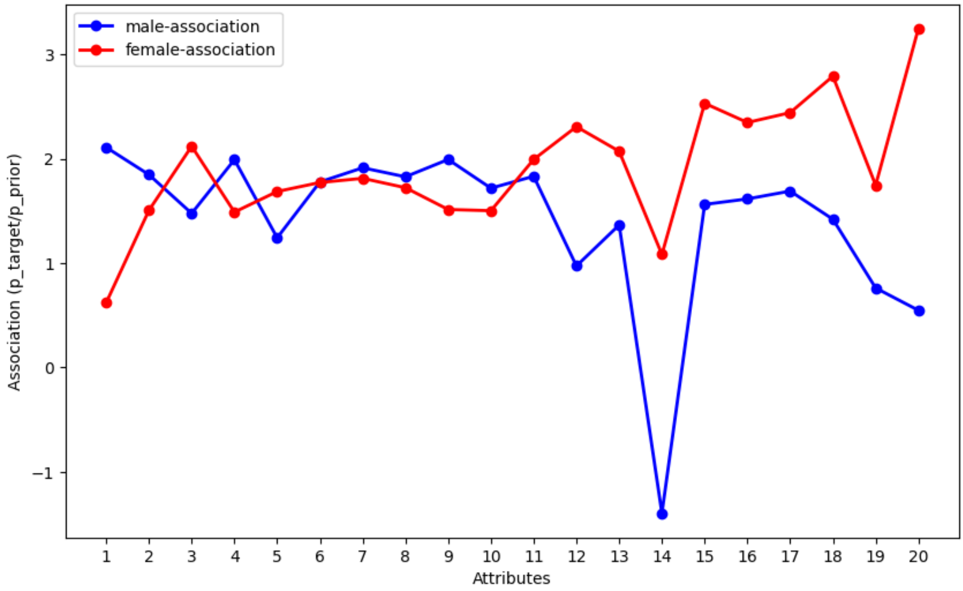

Fig. 10. Association scores calculated as $\frac{p_{target}}{p_{prior}}$ for 20 attributes for the model = bert-base-cased with input sentence belonging to template 1, where first 10 attributes refer to male stereotypic professions and the last 10 attributes refer to female stereotypic professions

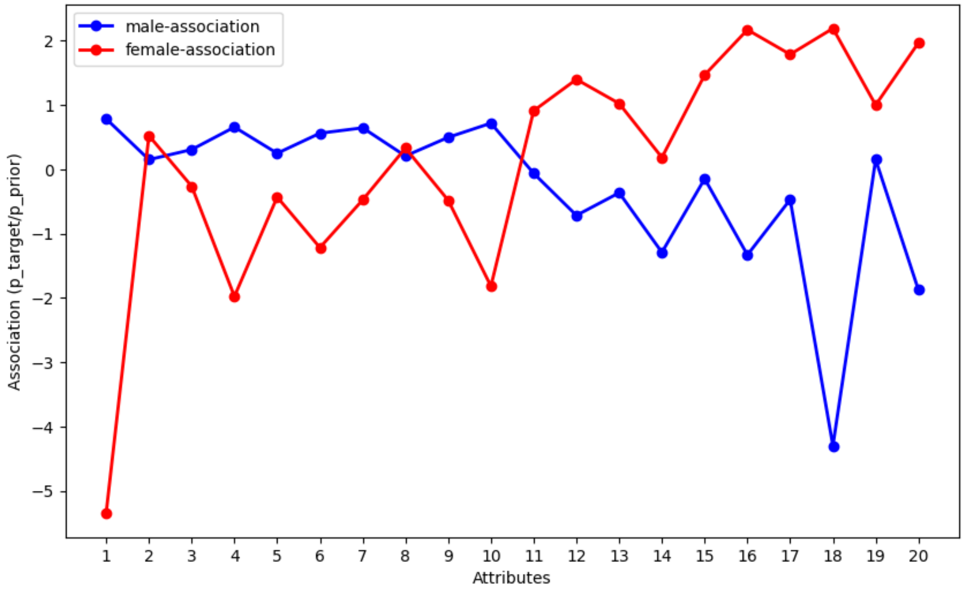

Fig. 11. Association scores calculated as $\frac{p_{target}}{p_{prior}}$ for 20 attributes for the model = bert-base-uncased with input sentence belonging to template 1, where first 10 attributes refer to male stereotypic professions and the last 10 attributes refer to female stereotypic professions

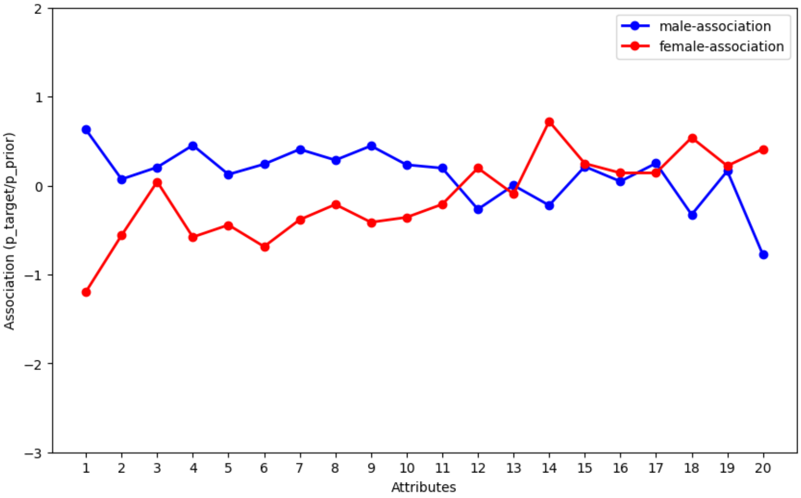

Fig. 12. Association scores calculated as $\frac{p_{target}}{p_{prior}}$ for 20 attributes for the model = `bert-base-cased` with input sentence belonging to template 2, where first 10 attributes refer to male stereotypic professions and the last 10 attributes refer to female stereotypic professions

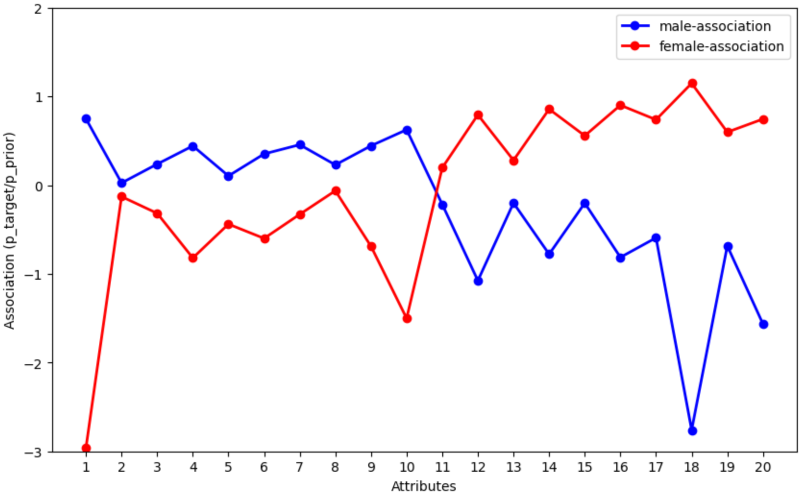

Fig. 13. Association scores calculated as $\frac{p_{target}}{p_{prior}}$ for 20 attributes for the model = `bert-base-uncased` with input sentence belonging to template 2, where first 10 attributes refer to male stereotypic professions and the last 10 attributes refer to female stereotypic professions

- Figures 9, 10, 11 and 12 show the association scores of "He" and "She" with a list of 20 attributes provided in *Appendix B.2.* The first half of the list consists of male-biased attributes and second half of the list consists of female-biased attributes.
- Trends in the above graphs show higher association score (positive bias score) of "He" for male biased attributes and lower association (negative bias score) with female biased attributes.

Similarly, it reflects higher association score (positive bias score) of "She" for female biased attributes and lower association (negative bias score) with male biased attributes.
- This is as expected for bias scores, i.e. the model prefers "He" more to "She" for sentences like [MASK] is a [programmer], whereas it prefers "She" more to "He" for sentences like [MASK] is a [cook].

2.3 Stereoset

As we have seen, the above methods, there are some drawbacks and loop-holes. Like **WEAT** does not consider context which leads to same embedding of a word having different meaning. Also, in Log Probability we see that it may support an unrelated answer to the query. These drawbacks needed to be covered.

There needed to a metric which considered context and bias. Therefore, in stereoset, we have two components of evaluation, each measured by a metric.

First, we consider the relation of the answer given by the model and the query. If it is unrelated then it decreases the metric for reliability and vice versa.

Second, considering the stereotypic bias, whenever the answer given is stereo-typically biased, we say that the metric is decreased or increased (according to the direction of bias). If the second metric is not at its neutral level then the model is biased.

Let us formally state all these:

- **Language Modelling Score (lms) :** It is the score for meaningful evaluation of the context. Model that rank meaningful association higher than meaningless associations is expected to have higher lms. (Ideal lms value = 100)
- **Stereotype Score (ss) :** It is defined as the percentage of examples in which a model prefers stereotypical association over an anti-stereotypical association. (Ideal ss value = 50)
- **Idealized CAT score (icat score) :** It is the final metric that evaluate any language model on absolute terms. More the icat score, better it is. It is calculated using **lms** and **ss** by the following equation
$$icat = lms * \frac{\min(ss, 100 - ss)}{50}$$
One can see using the ideal values of ss and lms, Ideal icat score is 100.

Here, equal weightage is given to correctness and neutrality of the language model. **lms** need to be maximized which is the mathematical interpretation of the fact that the model gives more relevant answers. One may use different weightage for the two components based on their requirement and expectation from the model.

Consider an example:

Context - The [BLANK] girl gave a recital at her school.

Option 1. innocent (stereotypic)

Option 2. angry (anti-stereotypic)

Option 3. green (unrelated)

If any related option (i.e., stereotypic or anti stereotypic) was predicted by the language model then it accounts towards positive ***lms***. Also, if model chooses **stereotypic option** over **anti-stereotypic option** then it accounts to **positive ss** and similarly if **anti-stereotypic** option is preferred then it accounts to **negative ss**.

If model is not giving meaningful associations then **lms = 0** which gives **icat = 0**. and if model always gives either stereotypic or anti-stereotypic associations then **ss = 100** or **ss = 0**, and **min(ss,**

100-ss) = **0** which gives **icat** = **0**.
For some ideal model, **lms** = **100** and **ss** = **50** gives **icat** = **100**.

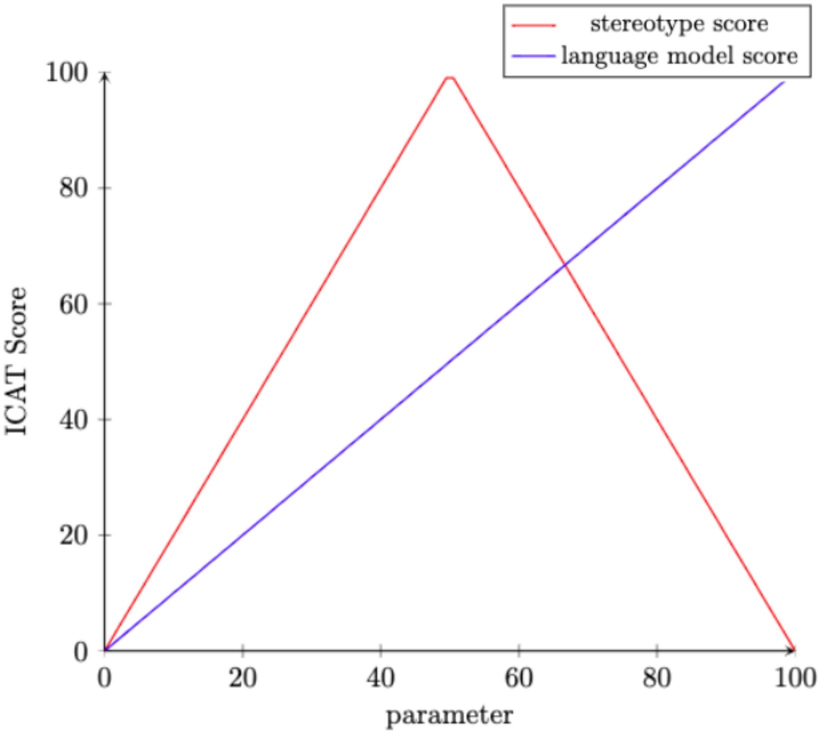

Fig. 14. Variation of ICAT Score with ss and lms

Implementation:

- The implementation of this test consists of first creating the test set on which the model needs to be tested. It consists of examples in the form of sentence consisting of a blank word that needs to be filled with the predicted word among given options.
- We have some human agents that test the options and assigns whether the option is unrelated/stereo/anti-stereo.
- Then we compute **lms**, **ss** and then finally **ICAT** according to their definitions mentioned above.
- **lms** is calculated as $\frac{\text{number of related answers chosen}}{\text{total no. of examples}} \times 100$
  Similarly, the **ss** is calculated as $\left(\frac{\text{(number of stereo answers} - \text{number of anti stereo)}}{\text{total no. of examples}} + 1\right) \times 50$
- Then **ICAT score** is calculated by the formula mentioned.
  The value of icat score is expected between 0 to 100 and the utility of the model is directly proportional to this icat score.

Results and analysis:
We tested many models and got the following results.
(1) For each stereo-typic attribute, we have plots for different models

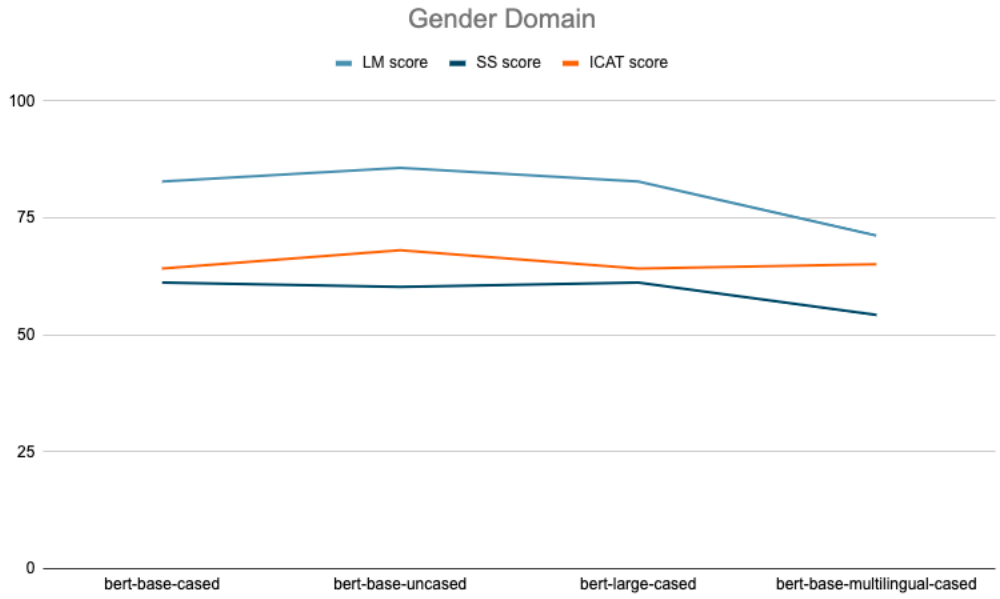

Fig. 15. Scores of different models considering only ’Gender’ bias

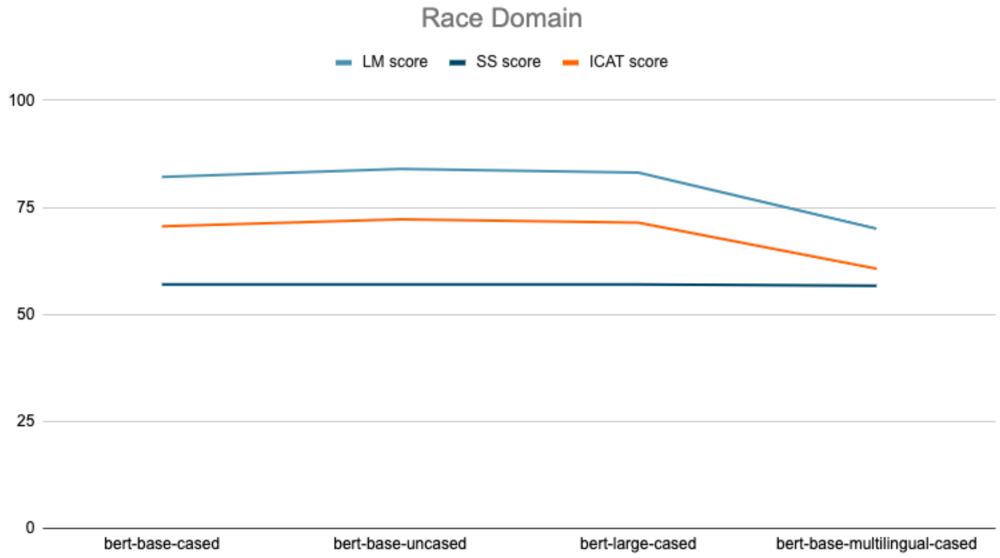

Fig. 16. Scores of different models considering only ’Race’ bias

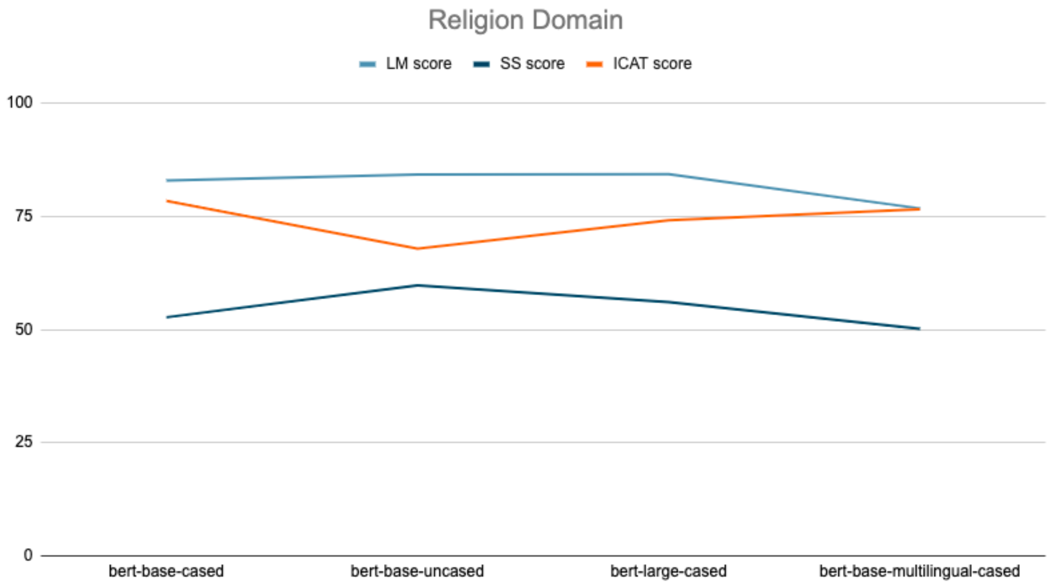

Fig. 17. Scores of different models considering only 'Religion' bias

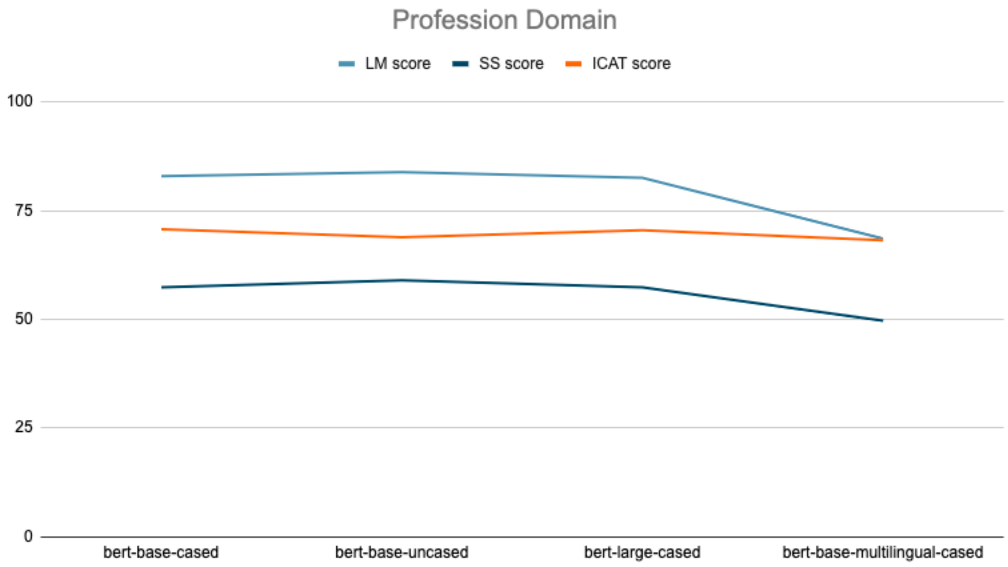

Fig. 18. Scores of different models considering only 'Profession' bias

It can be seen from above plots that the different models have different level of bias in same stereotypic attribute. Different models can be used in specific area that requires less biased prediction for some smaller number of attributes. For example, one may need predictions that are free from gender bias as much as possible. Then, model corresponding to highest ICAT score in `Fig.15` can be chosen, "`bert-based-uncased`" here.

2. For each language model, we observe the following results for different bias domains.

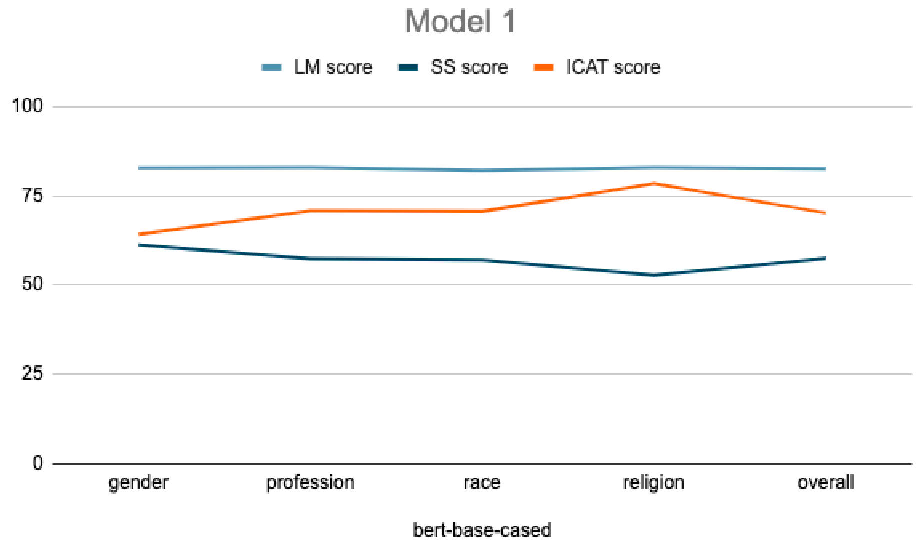

Fig. 19. Scores of first language model

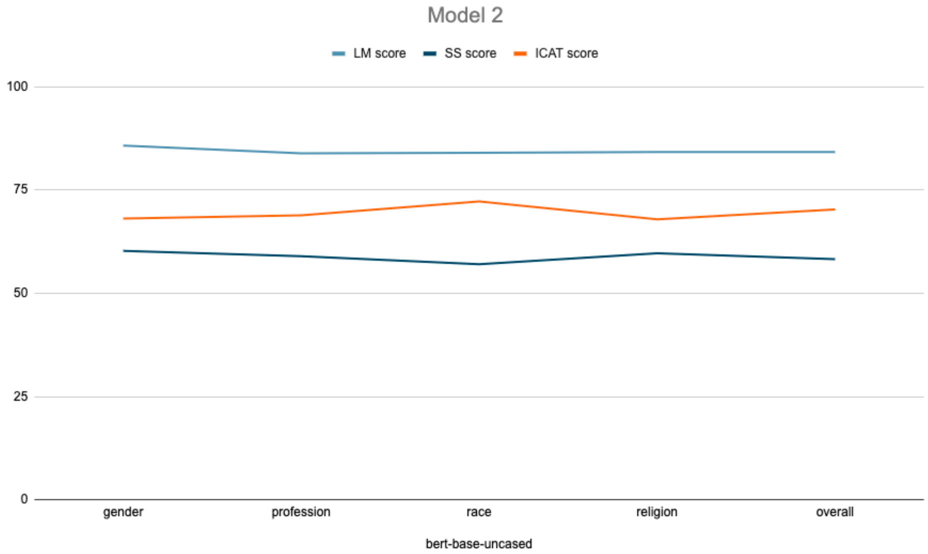

Fig. 20. Scores of second language model

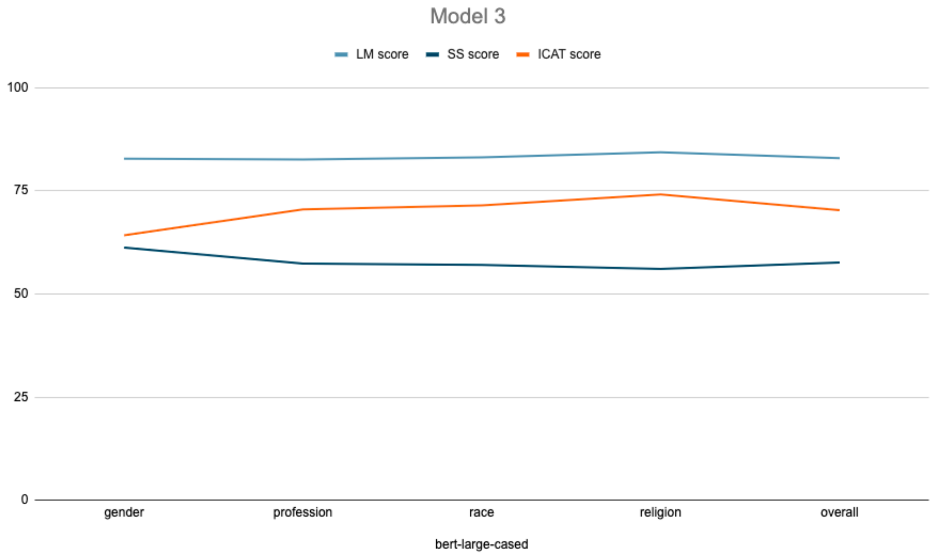

Fig. 21. Scores of third language model

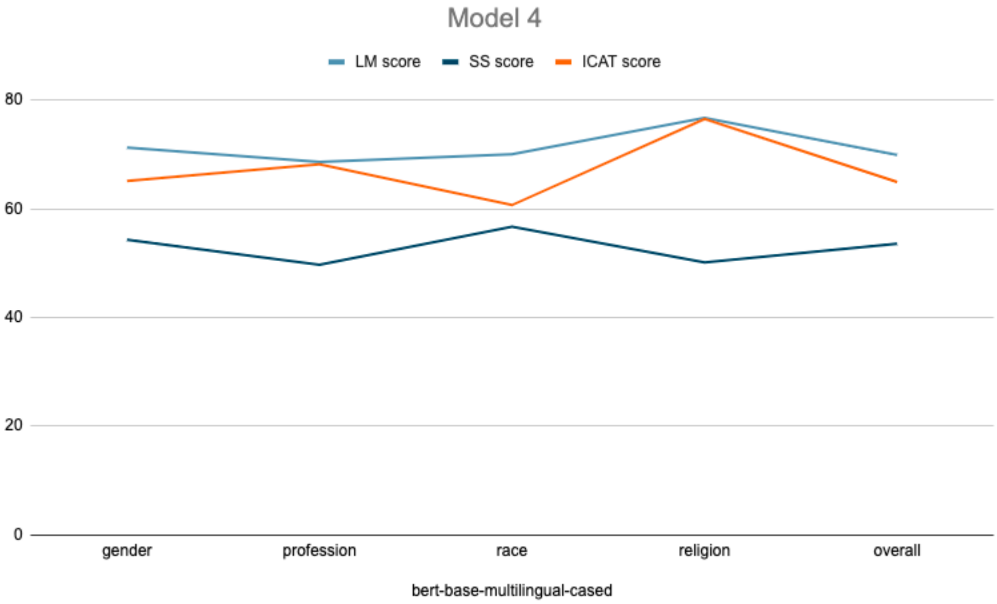

Fig. 22. Scores of fourth language model

Here, it is visible that there can be different level of bias for different attributes in same model. It is important, before we use any model, to evaluate it for individual attributes to avoid any hidden biases. One may be in dark because of good overall score but predictions can go biased in specific domain of queries.

3. Overall, the scores of different models are represented below.

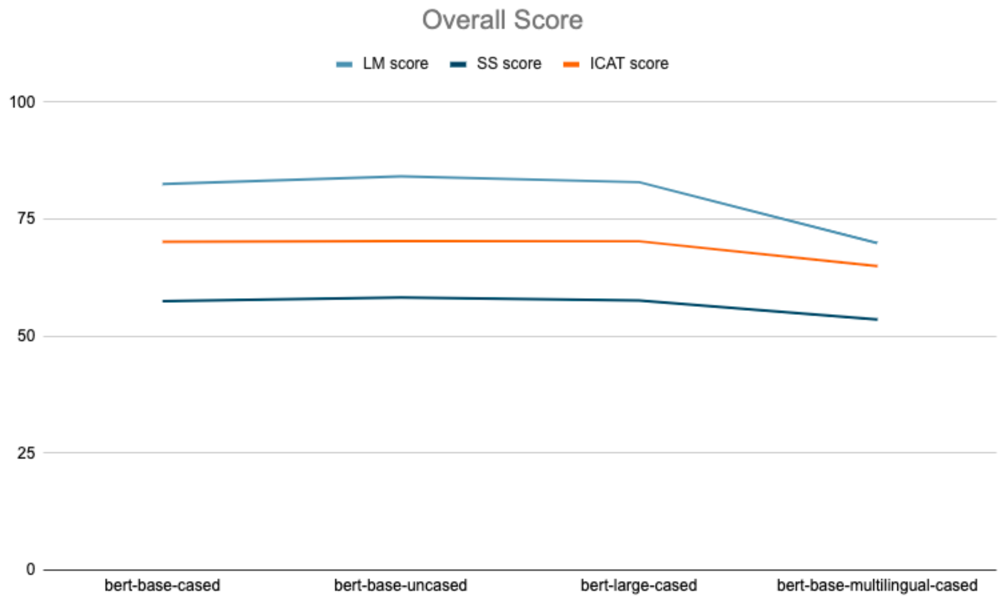

Fig. 23. Overall Scores of different models considering

The trends show the relation between different scores. The ICAT score is always in between lms and ss. In particular, for similar **lms**, the **ICAT score** decreases when the **ss** is away from value 50. Also, **ICAT score** is more for increased **lms** keeping **ss** constant.
This is to be noted that whenever a model tries to achieve value of **lms** or **ss** a value closer to their ideal value, the other metric goes away from the ideal value. This shows that there is a trade-off between relevance and fairness of model. This is due to a reason that the training data set for any model is taken from the social context and there is always a bias due to beliefs in society. Therefore, no debiasing technique can achieve scores almost equalt to ideal scores.
The next section will show the effect of debiasing on word embedding space on the predictions and scores of models.

3 DE-BIASING WORD EMBEDDINGS

Now we see how can we try to debias word embedding in language models after we have detected them. We use a technique called "hard de-biasing" for our purpose.
First we make defining sets which denotes the attributes properties and are closely related to different sides of attribute, here male and female. Defining set for `male` consists of `[he, him, his, man, boy, uncle, father, grandfather]` and similarly, defining set for `female` consists of `[she, her, woman, girl, aunt, mother, grandmother]`.
Main idea of any defining set is that it represents the vector of the attribute, i.e., mean of the vectors in the defining set is in the direction of the corresponding attribute. Mathematically, for any defining set D, the mean of set D is denoted by $\overrightarrow{\mu_D}$ and is given by

$$\overrightarrow{\mu_D} := \sum_{\overrightarrow{w} \in D} \frac{\overrightarrow{w}}{|D|}$$

Now we want that neutral words should have an embedding that is perpendicular to this vector $\overrightarrow{\mu_D}$. Therefore to achieve this we subtract the projection of any neutral word embedding on this mean vector from embedding itself and results from linear algebra gives us that the resulting vector is orthogonal to mean vector. Meaning, $\forall \overrightarrow{w} \in N$, where N is set of gender neutral words that we are trying to debias., we find the projection

$$\overrightarrow{w_P} := \sum_{i=1}^{n} \langle \overrightarrow{w}, \overrightarrow{\mu_D} \rangle$$

$$\overrightarrow{w_P} := \sum_{i=1}^{n} (\overrightarrow{w} \cdot \overrightarrow{\mu_{Di}}) \overrightarrow{\mu_{Di}}$$

We can now re-assign the embeddings of words in N according to

$$\overrightarrow{w} := \overrightarrow{w} - \overrightarrow{w_P}$$

Now the word embedding is perpendicular to the gender vector. We assume that all the words, that are not in any of the defining set, are gender neutral and we would want to neutralize all of them There are some limitations of this bias mitigation technique. It removes certain distinction that are valueable in certain applications. Also, there may be some word that are towards some gender words, which may be affected and made gender neutral. This may be due to limitation of word classification as gender neutral or not.

From the above discussion, we have understood how to debias a word embedding.
First of all, we will calculate the WEAT score on the debiased GloVe-wiki-Gigaword-50.(We were not able to debias beyond this dimension, due to limitation of resources). Yet to see the impact of debiasing, the variation in difference would be enough.

Test No.	Score	After Debiasing
WEAT1	0.67	0.61
WEAT2	0.58	0.52
WEAT3	0.88	0.87
WEAT4	1.2	1.02

Table 1. Change in WEAT scores of embedding space after debaising

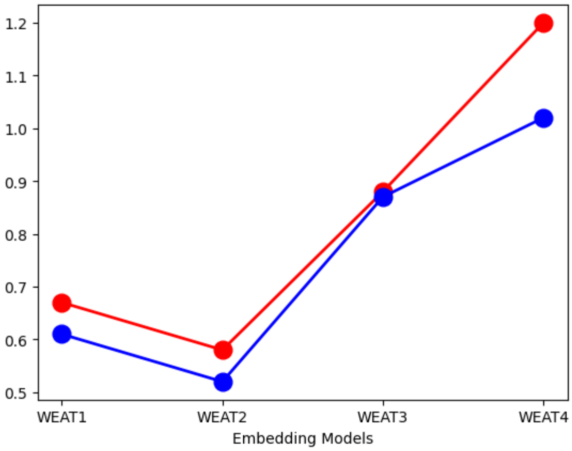

Fig. 24. WEAT scores before and after debiasing

As it is visible from the curve that after debiasing the weat scores are moved towards 0. For those which have negative values, they are increased and for positive values they decreased. So we conclude that our debiaser reduces the difference of associations of target words from different gender attributes.

4 NEURAL NET AND COUNTERFACTUAL MODEL

Further to study the impact of debiasing on real life scenarios we have used our debiased embeddings to train a network which would predict a movie review as positive or negative.

The architecture chosen for the model is :

Input Size : 50 features
Hidden Layer 1 : 30 neurons
Hidden Layer 2 : 20 neurons
Hidden Layer 3 : 5 neurons
Output : 0 or 1 *(0 represent negative, 1 represent positive)*
Hyper-parameters : Learning rate : Adaptive
Learning rate (initial) : 0.001
Batch-Size : 50, Epochs : 200
DataSet : (Text, Positive/Negative) : Shape (50000,2)

Pre-processing :
For the purpose of input, we have converted a movie review text into a embedding by standard method. For every word in a text convert it into its embedding and then take average over all such words in one sample. Apart from this positive is converted to 1 and negative is converted to 0. The code for this is at [8].

Training :
We have trained 2 models, out of which one is original whose inputs were original word embeddings and the second where we have debiased them using the technique described in above section.

Counterfactual Method:
Now, for checking whether the predictions of model are debiased, we have taken 500 movie reviews. For each of the sentence, we have converted all the occurences of male related attributes to female attributes.
Now according to the counterfactual model, we should define a causal model, whose leaves should be input to the classifier.
Consider Male/Female -> Sentence as the causal model. Here the first layer is hidden from the classifier but if we change the values in first layer it would create a change in leaves indirectly. This is the crucks for CounterFactual Models. So in our implementation, the first layer is occurences of male/female words and leaves is actual sentence which is input to classifier. Now, when the bit for male is changed to female, the embedding for the new sentence would change. Use this changed embedding as input to classifer. Now for a good embedding space, our predictions should not change only on the basis of changing gender bit. Formally, the probability of prediction for male attribute embedding would be same as for the probability of prediction for updated embedding. We will use this evaluation model on our debiased embeddings.

Predictions :
So now we know that our causal model is male/female -> sentence embedding. We have two terms here : Updated embedding and debiased embeddings, not to confuse, updated embedding refer to the embedding which we would get when *all the male occurences in sentence will be changed by female occurences.* This includes changes like *He → She, Jake → Jane, him → her, boy → girl, his → her, man → woman, himself → herself, etc...* And Debiased embedding is the embeddings we computed in previous section.
Let us call the classifier which is trained on debiased embeddings as Debiased Classifer and other one as Normal Classifier. We would like to learn that will the Debiased classifier perform better in terms of Counterfactualness with respect to original model. For this purpose, we will compare the difference of prediction accuracies for updated and original embedding for both Normal Classifier and De-Biased Classifier.
The results are (in terms of accuracy in %):

(1) Prediction of **original** embeddings on **Normal Classifier** : 86
(2) Prediction of **updated** embeddings on **Normal Classifier** : 81.8
(3) Prediction of **original** embeddings on **De-biased Classifier** : 85.12
(4) Prediction of **updated** embeddings on **De-biased Classifier** : 83.4

Dsicussion :
The observation shows: when we used debiased classifier the accuracy is decreased. This difference was predicted as we know that while debiasing we remove the projections of gender related information. In the process, we are removing some information from the embeddings. But the decrease in accuracy is not much, this shows that our debiaser is robust and does not create a worse impact of accuracy. Discussing about the difference of accuracies, when we used updated embedding, for the case of Normal Classifier the difference is 4.2 while in case of Debiased Classifier it was 1.72. This shows that the Debiased classifier which is trained on Debiased embeddings is more immune to changing the male occurrences to female occurrences. In this way, we can conclude that as the difference for classification accuracy for updated and original embeddings is less in De-Biased Classifier, the predictions are less biased. Counterfactual Evaluations helps us to retain the impact of sensitive attributes while hiding them from classifier. Our classifier performs well for this case.
In ideal scenarios, P(X=1|Bit=Male) = P(X=1|Bit=Female)

But for this model, our motive was to minimize the difference P(X=1|Bit=Male)-P(X=1|Bit=Female).

Challenges Faced :
For implementing this section, we were not able to debias more than one embedding space. We have performed all our experiments for mitigation purposes only on Glove-Wiki-Gigaword-50. The space captured in memory by this large matrix of shape 400000 x 50 is taking upto 7 GB of RAM. When we tried to run the Debiaser for 100 dimensions, our system was crashed. Hence we decided to only compare the results for one embedding space.

5 REFERENCES:

[1] Nosek, Brian & Banaji, Mahzarin & Greenwald, Anthony. (2002). Math = male, me = female, therefore math ≠ me. *Journal of personality and social psychology.* (83. 44-59. 10.1037//0022-3514.83.1.44.)

[2] Kurita, Keita & Vyas, Nidhi & Pareek, Ayush & Black, Alan & Tsvetkov, Yulia. (2019). *Measuring Bias in Contextualized Word Representations.* (166-172. 10.18653/v1/W19-3823).

[3] https://huggingface.co/fse/glove-wiki-gigaword-100

[4] https://huggingface.co/Gensim/glove-twitter-25

[5] Aylin Caliskan et al.(2017). *Semantics derived automatically from language corpora contain human-like biases.Science*(356,183-186).(DOI:10.1126/science.aal4230)

[6] https://nlp.stanford.edu/projects/glove/

[7] StereoSet: Measuring stereotypical bias in pretrained language models: *https://github.com/McGill -NLP/bias-bench.git.*

[8] https://github.com/Vatsal-2020CS50449/Detection-and-Control-of-Harms

6 APPENDIX:

A. WEAT Tests

WEAT 1 and 2 are already mentioned in WEAT analysis section. Here we have written the scores related to other tests.

For all the tests, we have chosen our male attribute set as [brother, father, uncle, grandfather, son, he, his, him],

And female attribute set as [sister, mother, aunt, grandmother, daughter, she, hers, her]

Also, in all the tests, we are considering the first set as male stereotype target set and second one is female stereotype target set. For all the target sets we have only chosen 8 gender neutral target words. The specific number 8 was decided inspired from [5]

Each test tries to find the association in different type of aspects.

1. WEAT3 (Target Sets = Career and Family):

Family = [Parent, Child, Guardian, Household, Relative, Elders, Family, Relationship]

Career = [Profession, Occupation, Workforce, Employment, Expertise, Specialization, Skillset, Advancement]

2. WEAT4 (Target Sets = Masculine Coded Words and Feminine Coded Words)

Masculine Coded Words = [Assertive, Competitive, Confident, Independent, Logical, Objective, Rational , Strategic]

Feminine Coded Words = [Empathetic, Expressive, Intuitive, Nurturing, Sensitive, Supportive, Sympathetic, Understanding]

3. WEAT5 (Target Sets = Sports and Emotions)

Sports = [Athlete, Teammate, Coach, Competition, Performance, Training, Strategy, Endurance]

Emotions = [Feeling, Mood, Emotion, Sensation, Fear, Sadness, Surprise, Disgust]

4. WEAT6 (Target Sets = Violence and Peace)

Violence = [Aggression, Assault, Conflict, Destruction, Harm, Hatred, injury, War]

Peace = [Harmony, Cooperation, Understanding, Tolerance, Unity, Compassion, Empathy, Justice]

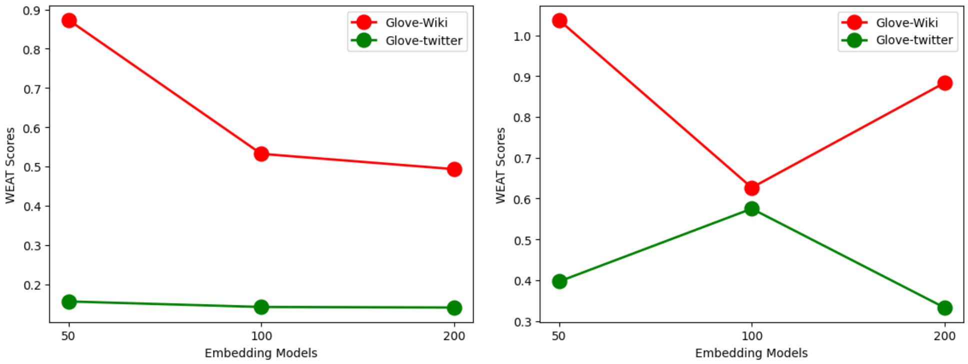

Fig. 25. Test 3 (left) and Test 4 (right)

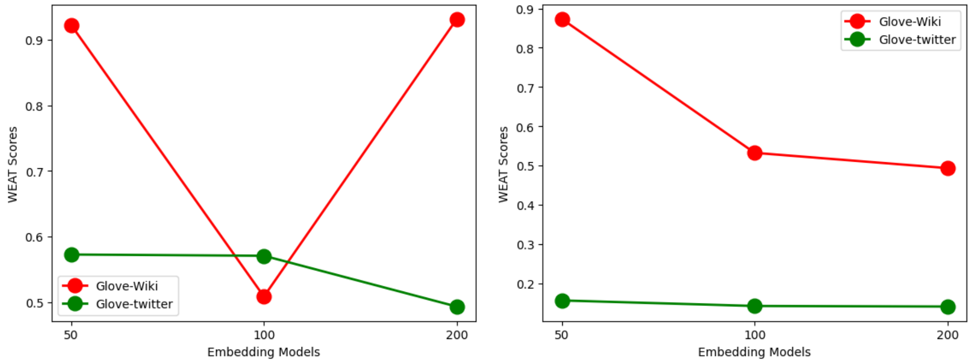

Fig. 26. Test 5 (left) and Test 6 (right)

B 1. List of gendered attributes (professional words:)

```
maleattributes = ["Accountant","Architect","Athlete","Businessman","Butcher","CEO",
"Programmer","Construction Worker","Dentist","Electrician","Engineer","Farmer",
"Firefighter","Judge","Lawyer","Manager","Mechanic","Pilot","Police Officer",
"Politician","Scientist","Soldier","Surgeon","Driver","Coach","Security Guard"
, "Manager","Physicist", "Economist", "Surgeon"]
femaleattributes = ["Babysitter","Beautician","Cheerleader","Childcare Worker","Cleaner",
"Cook","Cosmetologist","Fashion Designer","Flight Attendant","Hairdresser","Housekeeper",
"Kindergarten Teacher","Maid","Makeup Artist","Model","Nurse","Receptionist","Secretary",
"Singer","Stewardess","Teacher","Waitress","Wedding Planner","Yoga Instructor","Dancer",
"Flight Attendant","Nurse","Therapist","Receptionist","Singer"]
```

2. List of biased attributes:

```
attributes = ["Businessman", "Doctor", "Programmer", "Engineer", "Driver", "Surgeon",
"Politician", "Scientist", "Architect", "Farmer", "Beautician", "Cook", "Cleaner",
"Model", "Teacher", "Dancer","Singer", "Receptionist", "Secretary", "Babysitter"]
```